 Nucleic Acids Research, 2000, Vol. 28, No. 1 177–178

The European Large Subunit Ribosomal RNA database

Peter De Rijk, Jan Wuyts, Yves Van de Peer, Tina Winkelmans and Rupert De Wachter*

Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received October 4, 1999; Accepted October 6, 1999

ABSTRACT

The European Large Subunit (LSU) Ribosomal RNA (rRNA) database is accessible via the rRNA WWW Server at URL http://rrna.uia.ac.be/lsu/ . It is a curated database that compiles complete or nearly complete LSU rRNA sequences in aligned form, and also incorporates secondary structure information for each sequence. Taxonomic information, literature references and other information about the sequences are also available, and can be searched via the WWW interface.

LARGE SUBUNIT RIBOSOMAL RNA AND ITS STRUCTURE

The large subunit (LSU) ribosomal RNA (rRNA) structure consists of a conserved core that is interspersed with variable regions. Especially in eukaryotes these regions can vary widely in sequence as well as length, even between relatively closely related species. The secondary structure for these regions cannot always be conclusively determined for all sequences. However, they do offer interesting targets for species-specific probes. In mitochondria the core structure itself is also more variable. The LSU rRNA in the mitochondria of kinetoplastids and animals even misses helices that are present in all other molecular sequences. As a consequence, the determination of the alignment and secondary structure of the mitochondrial LSU rRNAs is less reliable.

In many species the LSU rRNA consists of several fragments of RNA that only connect through secondary structure interactions, but are clearly homologous to parts of the non-fragmented LSU rRNA of other species. This fragmentation can be quite elaborate with LSU rRNAs consisting of up to 14 separate pieces (1). In some species the LSU rRNA is even discontinuously encoded in separate gene pieces that are scrambled in order and interspersed with protein coding genes (2). LSU rRNA fragments, such as the 4.5S in plant LSU rRNA or 5.8S in animal LSU rRNA, are often ambiguously labeled as separate rRNA species. In the sequence alignment in the LSU database, the fragments are combined to the full sequence of the mature assembled rRNA. The fragmentation patterns can still be found in the reference data, where each fragment has its own entry.

Drawings of the structure for a number of species of the major taxa can be found in the supplementary material accompanying the online version of this paper. In bacteria and most Archaea the 5′ and 3′ end of the LSU rRNA are joined by a helix. From this stem helix starts a large central multi-branched loop, from which several helices emanate (3–7). While in eukaryotes the stem helix is not present, the central loop and the structures branching from it are very similar. The stem helix is labeled A, and the structures branching from the central loop are labelled B–I. Helices are numbered in the 5′ to 3′ direction. Helices get a different number when they are separated by a multi-branched loop. In the case of helices not belonging to the core structure but specific to certain taxa, an underscore and a number are appended to the name of the preceding core helix. No structure is proposed for the hypervariable regions in some taxa.

CONTENTS OF THE DATABASE

The LSU rRNA database is a curated database that compiles complete or nearly complete large subunit rRNA sequences in aligned form. The total chain length of a partially determined sequence is estimated by comparison to the complete sequence of a closely related species. Sequences are included in the database if >70% of the estimated total chain length has been determined. Secondary structure information for each sequence is incorporated by the inclusion of symbols indicating start and end of structural elements in the alignment, and ‘helix numbering’ lines indicating the name of the structural elements at a specific position. Taxonomic information, literature references and other information about the sequences are also available, and can be searched via the WWW interface.

New or updated entries containing rRNA sequences in the EMBL nucleotide sequence database (8) are detected by the ‘Current Sequence Awareness’ service of the Belgian EMBNet node (http://ben.vub.ac.be). The rRNA sequences in these entries are automatically extracted, distributed over the LSU and SSU databases, and aligned to a closely related sequence already in the database. The resulting automatic alignments are manually checked and corrected.

Sequences are labeled by the name of the species. When more than one sequence for the same species is available, former releases of the database added a number to the sequence name. From this release onwards, the accession number of the EMBL entry is added to the species name where possible instead. Also starting from this release, taxonomic specifications will be the same as those used by EMBL, contrary to taxonomic descriptions used in previous releases of the LSU rRNA database (7).

*To whom correspondence should be addressed. Tel: +32 3 820 2319; Fax: +32 3 820 2248; Email: dwachter@uia.ua.ac.be
Present address:
Yves Van de Peer, Department of Biology, University of Konstanz, D-78457 Konstanz, Germany

AVAILABILITY AND FORMAT OF THE DATABASE

The LSU rRNA database is accessible via the rRNA WWW Server at URL http://rrna.uia.ac.be/lsu/ . The data is stored on the server in files in a special distribution format. Each file contains one rRNA sequence. Each file contains information about the sequence such as accession number and taxonomic position, followed by the organism name and the sequence. If a sequence is fragmented or consists of several exons, each part is preceded by its own annotations. The sequence itself consists of a range of nucleotide symbols interspersed with gap symbols necessary for alignment, and special symbols indicating the secondary structure. Each sequence segment is terminated by an asterisk. Sequences in the distribution format can be downloaded via ftp or via the list interface on the web site (http://rrna.uia.ac.be/lsu/list/). This interface allows the user to select and download sequences one by one from a list ordered by taxonomic group.

Sequences can be selected and downloaded in a number of other formats using the forms (http://rrna.uia.ac.be/lsu/forms/) or the query (http://rrna.uia.ac.be/lsu/query/) interface. Currently supported formats are DCSE (9) alignment and reference files, EMBL, NBRF/PIR, TREECON (10), the distribution format described earlier and a printable format. Only some formats (DCSE, printable and distribution) allow information about the secondary structure of the sequences to be inserted into the alignment. If such a format is chosen, the appropriate 'Helix numbering' line that indicates the names of each helix segment will be automatically added to the downloaded sequences. In the printable format, the alignment has been sliced into blocks that fit onto a page. This format it is limited to a selection of 100 sequences.

Using the query interface, the selection of sequences is made by searching information in the database such as species name, authors, accession number, reference journal or title. Multiple search terms in the same field must be separated by spaces; if a search term includes a space, it should be surrounded by double quotes. Multiple search fields can be combined, returning sequences matching both queries. The selection can also be limited to specific taxonomic groups using the check buttons at the bottom of the query page. If one or more taxonomic groups have a check mark, only sequences from these groups will be returned. In the forms interface, sequences are simply selected by name from a list sorted by taxonomic group.

In case of problems, the authors can be contacted by Email to derijkp@uia.ua.ac.be or dwachter@uia.ua.ac.be . Users publishing results based on data retrieved from our database are requested to cite this paper.

SUPPLEMENTARY MATERIAL

The following additional information can be found in the online version of the paper.

- detailed secondary structure models of several LSU rRNA species
- nucleotide variability maps of the prokaryotic and eukaryotic LSU rRNA

ACKNOWLEDGEMENTS

Our research is supported by the Fund for Scientific Research (Flanders), by the Special Research Fund of the University of Antwerp (Belgium) and by the Faculty of Biology (Dr Axel Meyer) of the University of Konstanz. P.De R. and Y.V. de P. are Research Fellows of the Fund for Scientific Research (Flanders).

REFERENCES

1. Schnare,M.N., Cook,J.R. and Gray,M.W. (1994) *J. Mol. Biol.*, **215**, 85–91.
2. Denovan-Wright,E.M. and Lee,R.W. (1994) *J. Mol. Biol.*, **241**, 298–311.
3. Noller,H.F., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F., Herr,W., Stahl,D.A., Gupta,R. and Woese,C.R. (1981) *Nucleic Acids Res.*, **9**, 6167–6189.
4. Brimacombe,R. and Stiege,W. (1985) *Biochem. J.*, **229**, 1–17.
5. Leffers,H., Kjems,J., Østergaard,L., Larsen,N. and Garrett,A. (1987) *J. Mol. Biol.*, **195**, 43–61.
6. Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.*, **21**, 3055–3074.
7. De Rijk,P., Robbrecht,E., de Hoog,S., Caers,A., Van de Peer,Y. and De Wachter,R. (1999) *Nucleic Acids Res.*, **27**, 174–178.
8. Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) *Nucleic Acids Re s.*, **27**, 18–24. Updated article in this issue: *Nucleic Acids Res.* (2000), **28**, 19–23.
9. De Rijk,P. and De Wachter,R. (1993) *Comput. Appl. Biosci.*, **9**, 735–740.
10. Van de Peer,Y. and De Wachter,R. (1994) *Comput. Appl. Biosci.*, **10**, 569–570.