Nucleic Acids Research, 1998, Vol. 26, No. 1 183–186

Database on the structure of large ribosomal subunit RNA

Peter De Rijk, An Caers, Yves Van de Peer and Rupert De Wachter*

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

Received October 6, 1997; Accepted October 8, 1997

ABSTRACT

The rRNA WWW Server at URL http://rrna.uia.ac.be/ now provides a database of 496 large subunit ribosomal RNA sequences. All these sequences are aligned, incorporate secondary structure information, and can be obtained in a number of formats. Other information about the sequences, such as literature references, accession numbers and taxonomic information is also available and searchable. If necessary, the data on the server can also be obtained by anonymous ftp.

CONTENTS OF THE DATABASE

The LSU rRNA database consists of an alignment of large subunit ribosomal RNA (LSU rRNA) sequences spanning a large range of taxonomic groups. Information about the secondary structure of these molecules is also incorporated in the database, and is used to refine the alignment. The database is regularly updated with new or updated sequences available in the EMBL nucleotide sequence database (1).

Only sequences for which more than 70% of the estimated chain length of the molecule has been sequenced are included in the database. The chain length of a partially determined sequence is estimated by comparing it to the complete sequence of a closely related species.

The latest release of the database (autumn 1997) on LSU rRNA contains a total of 496 sequences. As illustrated in Figure 1, most of these are mitochondria and Bacteria, with 179 and 170 representatives respectively. Especially animal mitochondria are over represented. The database further contains 47 plastidial, 27 archaeal and 73 eukaryotic sequences. The eukaryotic taxa in the database and the number of their representatives are listed in detail in Table 1.

TAXONOMIC CLASSIFICATION

Since the taxonomic classification of species in our database is different from that followed by the EMBL database, it is adapted for all sequences. The taxonomic classification of the eukaryotic

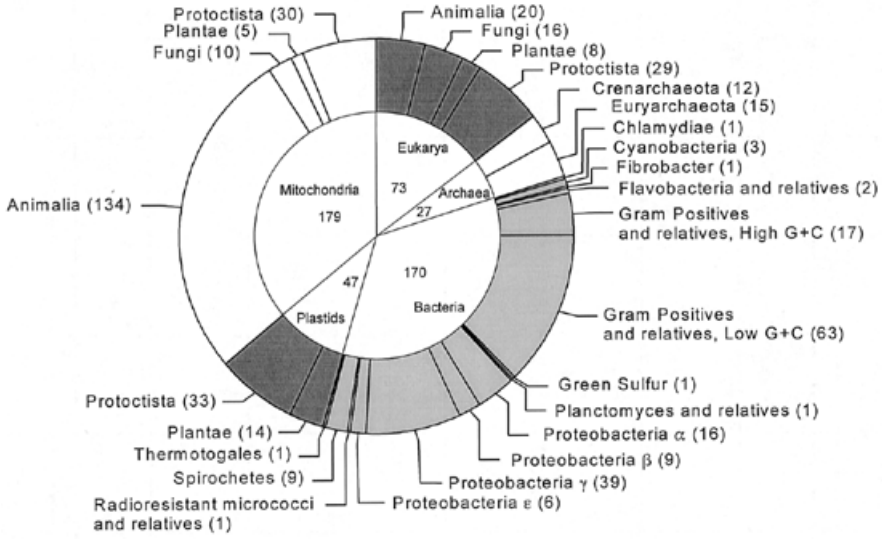

Figure 1. Distribution of representatives for the different taxonomic groups in the database. The number of sequences is mentioned between brackets after each taxon. The total number of sequences is 496.

* To whom correspondence should be addressed. Tel: +32 3 820 26 19; Fax: +32 3 820 22 48; Email: dewachter@uia.ua.ac.be

Table 1. Eukaryotic taxa represented in the database and number of their representatives

Kingdom Animalia[a]			
Phylum	Class	Number of sequences[b] N	M
Platyhelminthes	Turbellaria	1	
Nematoda	Secernentea	1	4
Annelida	Oligochaeta		2
Arthropoda	Malacostraca		2
	Insecta	3	13
Mollusca	Bivalvia		2
	Gastropoda		3
	Polyplacophora		1
Echinodermata	Asteroidea		1
	Echinoidea		3
Chordata	Ascidiacea	1	
	Agnatha		1
	Amphibia	3	3
	Aves		26
	Mammalia	4	49
	Osteichthyes	7	20
	Reptilia		4
Total		**20**	**134**

Kingdom Fungi[c]			
Subphylum	Class	Number of sequences[b] N	M
Ascomycotina	Hemiascomycetes	10	4
	Plectomycetes		4
	Pyrenomycetes		2
	Uncertain Affiliation	1	
Basidiomycotina	Heterobasidiomycetes	2	
	Hymenomycetes	1	
Zygomycotina	Zygomycetes	2	
Total		**16**	**10**

Kingdom Plantae				
Phylum	Class	Number of sequences[b] N	M	P
Bryophyta	Bryopsida	1		
	Marchantiopsida		1	1
Magnoliophyta	Liliopsida	1	2	4
	Magnoliopsida	6	2	8
Pinophyta	Pinopsida			1
Total		**8**	**5**	**14**

Kingdom Protoctista				
Phylum	Class	Number of sequences[b] N	M	P
Apicomplexa	Coccidia	4	1	
	Hematozoa	2	3	1
Bacillariophyta	Bacillariophyceae			2
Chlorarachnida		1		
Chlorophyta	Chlorophyceae	1	5	19
Chrysophyta	Chrysophyceae	1		
Chytridiomycota		1	1	
Ciliophora		2	5	
Dictyostelida		1	1	
Dinoflagellata		1		
Euglenida		1		5
Eustigmatophyta	Eustigmatophyceae	1		2
Hypochytriomycota		1		
Oomycota		1		
Phaeophyta		1	1	1
Plasmodial slime molds	Myxomycota	2		
Rhizopoda	Lobosea	1	1	
Rhodophyta			1	3
Xanthophyta		1		
Zoomastigina	Kinetoplastida	3	11	
	Diplomonadida	3		
Total		**29**	**30**	**33**

[a]The Metazoan taxa are listed in the same order as they appear in (2).
[b]The number of sequences listed in the database is larger than the number of species, because for certain species multiple LSU rRNA sequences have been determined, usually by different authors. The sequences are not necessarily identical because they may have been determined for different varieties or strains of a species, or for different genes of the same organism. The number is listed for sequences of nuclear (N), mitochondrial (M) and plastid (P) origin.
[c]The fungal, plant and protoctist phyla and classes are ordered alphabetically.

species is according to Brusca and Brusca (2) for the Animalia, according to Cronquist (3) for the higher plants, according to Ainsworth *et al.* (4) for the zygomycetes and ascomycetes, according to Moore (5) for the basidiomycetes, and according to Margulis *et al.* (6) for the remaining eukaryotes, viz. the Protoctista.

Archaea and Bacteria are classified according to the phylogenetic position observed in evolutionary trees constructed by the neighbor-joining method (7). The species are assigned to one of the taxa described by Woese and co-workers (8,9) and our research group (10,11). For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (12).

SECONDARY STRUCTURE MODEL

Figure 2 shows the secondary structure model incorporated in the database for the LSU rRNA of the animal *Xenopus laevis*. This model conforms largely to the model developed in earlier studies (13–16). All Bacteria, Archaea, plastids and Eukarya adopt a very similar core structure. However, in eukaryotes, this core is interspersed with regions which vary extremely in length and sequence, even between relatively closely related species. The structure for these regions has not always been conclusively determined for all sequences in the database, and some of the areas in Figure 2 have been left unstructured for this reason. In

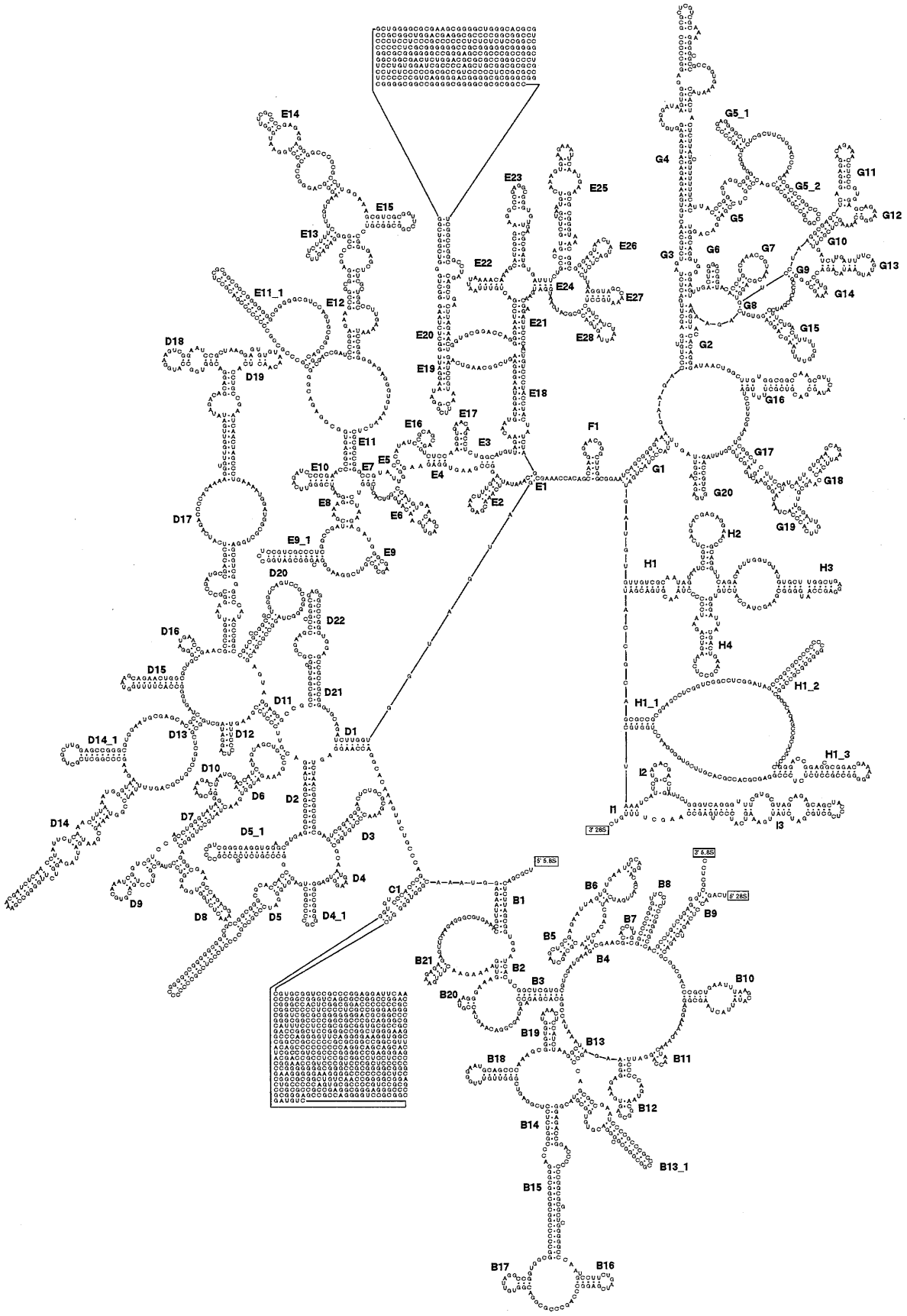

Figure 2. Secondary structure model for *X.laevis* LSU rRNA. The sequence is written clockwise from 5′ to 3′ terminus. The regions enclosed by helices C1 and E20 have been left without structure.

mitochondria the structural variability of the core is much higher than in other species, and in the mitochondria of kinetoplastids and animals many helices of the core are even absent. As a consequence, the alignment and proposed secondary structure of the mitochondrial LSU rRNAs are less reliable.

A central multibranched loop from which several helices emanate forms the basis of the structure. This central loop is closed by a stem helix joining the 5′ and 3′ ends of the molecule in Bacteria and most Archaea, but not in Eukarya. The structures branching from the central loop are labelled A–I, starting from the stem helix (not present in Fig. 2). Within each of these structures, helices are numbered in the 5′→3′ direction. Helices get a different number when they are separated by a multibranched 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.

AVAILABILITY AND FORMAT OF THE DATABASE

The sequences, alignments and structure information in the database are available from the server rrna.uia.ac.be. On the server each sequence is stored in a separate file in a simple, computer-readable distribution format. Information about the sequence such as the accession number and taxonomic position is stored at the start of each file. This information is followed by the organism name and the sequence. Parts of fragmented sequences, or sequences consisting of several exons, are stored in the same file, each part preceded by its own annotations. Each sequence consist of a range of nucleotide symbols interspersed with gap symbols necessary for alignment. The sequence ends are indicated by an asterisk. The beginning and end of secondary structure elements are indicated by insertion of special symbols. Special ‘helix numbering’ files with lines that indicate the numbers of each helix segment are present for researchers who wish to use the secondary structure information.

The easiest way to access the database is by using the World Wide Web (WWW) interface. The LSU rRNA home page can be reached at http://rrna.uia.ac.be/lsu/ . It offers several methods to obtain the desired data. The query and list interfaces use forms to let the user select and obtain sequences in a number of formats. Currently supported formats are DCSE (17) alignment and reference files, EMBL, NBRF/PIR, TREECON (18), the distribution format and a printable format. In the printable format, the alignment has been sliced into blocks that fit onto a page, but it is limited to a selection of 100 sequences. The list interface lets the user select sequences from a list of species, ordered per taxonomic group. The query interface contains several query fields, in which a list of search terms can be typed. All sequences containing one or more of the search terms in their respective annotation will be returned. The search terms are separated by spaces; if a search term must include a space, it should be surrounded by double quotes. If search terms are entered in more than one field, only sequences matching both queries will be returned. The bottom part of the query page lists all taxonomic groups with check buttons that can be used to limit the search to specific taxonomic groups. When one or more of these are checked, only matching sequences from these taxonomic groups will be returned. If taxonomic groups are checked, but all the query fields are left blank, all sequences in the checked groups will be fetched.

The files from the LSU rRNA database are also obtainable by anonymous ftp on rrna.uia.ac.be and are made available to the EMBL nucleotide library for distribution. On the anonymous ftp server, a ‘readme’ file will be present which describes the latest state of the database and listing the contents of the files and directories. Since each sequence is stored in a separate file, the user can also get any selection of sequences using ftp. However, using anonymous ftp only the distribution format can be downloaded, and users will have to convert these files into a desired format themselves.

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

ACKNOWLEDGEMENTS

Our research was supported by the BIOTECH programme of the commission of European Communities (contract BIO2-CT94-3098), by a research project of the University of Antwerp (UA), by the Fund for Scientific Research Flanders, and by the Special Research Fund of the University (UIA). Peter De Rijk and Yves Van de Peer are Research Assistants of the Fund for Scientific Research Flanders.

REFERENCES

1 Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N. (1997) *Nucleic Acids Res.* **25**, 7–13 [see also this issue (1998) *Nucleic Acids Res.* **26**, 8–15].
2 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates.* Sinauer Associates, Inc., Sunderland.
3 Cronquist,A. (1971) *Introductory Botany.* Harper & Row, New York.
4 Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) *The Fungi: an Advanced Treatise*. Academic Press, New York, Vol. 4A.
5 Moore,R.T. (1988) in Moriarty,Ch. (ed.), *Taxonomy Putting Plants and Animals in Their Place*. Royal Irish Academy, Dublin, pp. 61–88.
6 Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds) (1990) *Handbook of Protoctista*. Jones and Bartlett Publishers, Boston.
7 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
8 Woese,C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
9 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
10 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **21**, 3025–3049.
11 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.* **17**, 32–38.
12 Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
13 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.
14 Brimacombe,R. and Stiege,W. (1985) *Biochem. J.* **229**, 1–17.
15 Leffers,H., Kjems,J., Østergaard,L., Larsen,N. and Garrett,A. (1987) *J. Mol. Biol.* **195**, 43–61.
16 Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.* **21**, 3055–3074.
17 De Rijk,P. and De Wachter,R. (1993) *Comput. Applic. Biosci.* **9**, 735–740.
18 Van de Peer,Y. and De Wachter,R. (1994) *Comput. Applic. Biosci.* **10**, 569–570.