174–178 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

Database on the structure of large subunit ribosomal RNA

Peter De Rijk, Elmar Robbrecht[1], Sybren de Hoog[2], An Caers, Yves Van de Peer and Rupert De Wachter*

Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium, [1]Nationale Plantentuin, Domein van Bouchout, B-1860 Meise, Belgium and [2]Centraalbureau voor Schimmelcultures, PO Box 273, NL-3740 AG Baarn, The Netherlands

Received October 8, 1998; Accepted October 13, 1998

ABSTRACT

The Antwerp database on large subunit ribosomal RNA now contains 607 complete or nearly complete aligned sequences. The alignment incorporates secondary structure information for each sequence. Other information about the sequences, such as literature references, accession numbers and taxonomic information is also available. Information from the database can be downloaded or searched on the rRNA WWW Server at URL http://rrna.uia.ac.be/

CONTENTS OF THE DATABASE

The Antwerp database on the structure of large subunit ribosomal RNA (LSU rRNA) provides an alignment of 607 sequences spanning a diverse range of taxonomic groups. The alignment is regularly updated and refined using a combination of automatic and manual methods, taking into account the knowledge of the structure model. Secondary elements are indicated in each sequence.

New or updated rRNA sequences in the EMBL nucleotide sequence database (1) are scanned for rRNA sequences using the 'Current Sequence Awareness' service of the Belgian EMBNet node (http://ben.vub.ac.be). Partial sequences are not included in

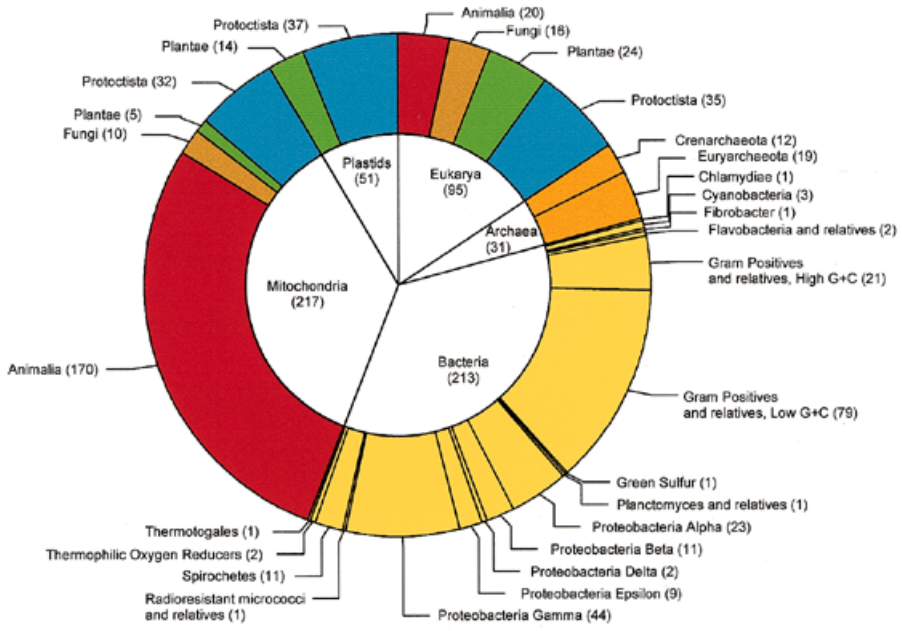


Figure 1. Distribution of representatives for the different taxonomic groups in the database. The number of sequences is mentioned between brackets after each taxon. The total number of sequences is 607.

*To whom correspondence should be addressed. Tel: +32 3 820 2319; Fax: +32 3 820 2248; Email: dwachter@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	5
Annelida	Oligochaeta		2
Arthropoda	Malacostraca		2
	Insecta	3	13
Mollusca	Bivalvia		2
	Cephalopoda		1
	Gastropoda		4
	Polyplacophora		1
Echinodermata	Asteroidea		1
	Crinoidea		1
	Echinoidea		3
Hemichordata	Enteropneusta		1
Chordata	Ascidiacea	1	
	Agnatha		1
	Amphibia	3	11
	Aves		35
	Mammalia	4	62
	Osteichthyes	7	20
	Reptilia		5
Total		**20**	**170**

Kingdom Fungi[c]			
Subphylum	Class	Number of sequences[b]	
		N	M
Ascomycota	Archiascomycetes	4	1
	Euascomycetes		6
	Hemiascomycetes	7	3
Basidiomycota	Heterobasidiomycetes	2	
	Homobasidiomycetes	1	
Zygomycota	Zygomycetes	2	
Total		**16**	**10**

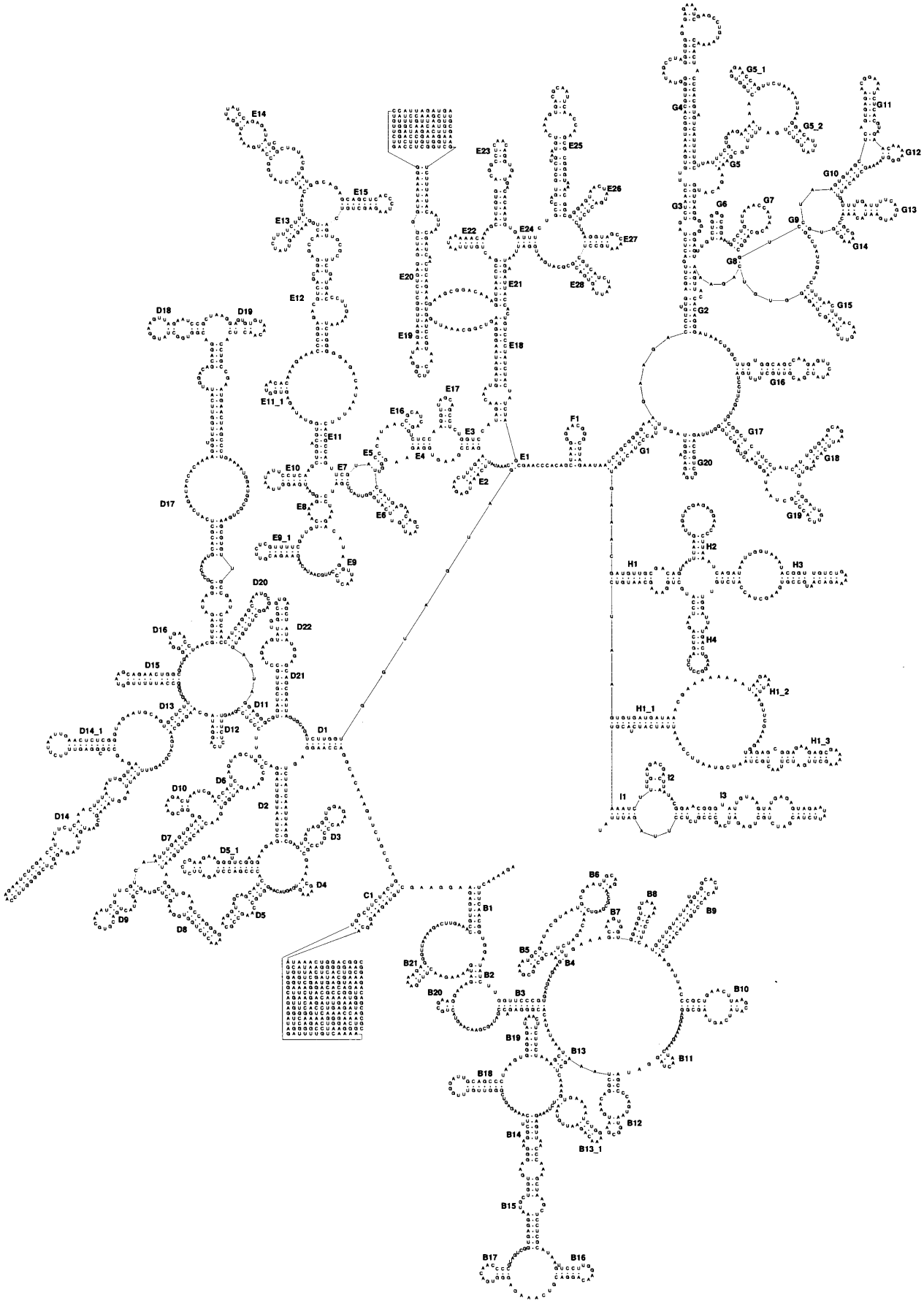
Kingdom Plantae				
Phylum	Class	Number of sequences[b]		
		N	M	P
Bryophyta	Bryopsida	2		
	Marchantiopsida		1	1
Magnoliophyta	Liliopsida	2	2	4
	Magnoliopsida	18	2	8
Pinophyta	Gnetopsida	2		
	Pinopsida			1
Total		**24**	**5**	**14**

Kingdom Protoctista				
Phylum	Class	Number of sequences[b]		
		N	M	P
Apicomplexa	Coccidia	7	1	
	Hematozoa	3	3	2
Bacillariophyta	Bacillariophyceae			2
Chlorophyta	Chlorophyceae	1	5	21
Chrysophyta	Chrysophyceae	1		
Chytridiomycota		1	1	
Ciliophora	Nassophorea		3	
	Oligohymenophorea	2	2	
Dictyostelida		1	1	
Dinoflagellata		1		
Euglenida		1		5
Eustigmatophyta	Eustigmatophyceae	1		2
Hypochytriomycota		1		
Microspora		2		
Oomycota		1		
Phaeophyta		1	1	1
Plasmodial slime molds	Myxomycota	2		
Rhizopoda	Lobosea	1	1	
Rhodophyta			2	4
Xanthophyta		1		
Zoomastigina	Kinetoplastida	3	11	
	Diplomonadida	4		
Uncertain affiliation			1	
Total		**35**	**32**	**37**

[a]The Metazoan taxa are listed in the same order as they appear in ref. 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.

the database unless >70% of the estimated chain length of the molecule has been sequenced. The total chain length of a partially determined sequence is estimated by comparison to the complete sequence of a closely related species.

Figure 1 clearly illustrates that the largest proportion of sequences in the database are from mitochondria and Bacteria, with 217 and 213 representatives respectively. The database also contains 51 plastidial, 31 archaeal and 95 eukaryotic sequences. The eukaryotic taxa in the database and the number of their representatives are listed in detail in Table 1.


G5_1
G4
G11
G5_2
G12
G5
G10
G3
G6
G7
G13
G9
G14
G8
G15
G2
G16
E14
E15
E13
E12
E23
E25
E26
E22
E24
E27
E20
E21
E28
E19
E18
D18
D19
E11_1
E17
E16
E3
F1
E11
E5
E4
E1
G1
G17
G20
G18
E10
E7
E2
D17
E8
E6
G19
E9_1
E9
H2
H1
H3
D20
H4
D16
D22
D15
D21
H1_2
D13
D11
H1_1
H1_3
D14_1
D12
D1
I2
D6
I1
I3
D10
D2
D14
D7
D3
D5_1
D4
C1
B6
B8
B1
B7
B9
B5
B4
B21
B2
B10
B20
B3
B19
B13
B11
B18
B12
B14
B13_1
B15
B17
B16

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 animal species is according to Brusca and Brusca (2). For plants and fungi, the taxonomic information has been extended up to the level of orders with regards to the previous database release, although this information is not listed in Table 1. The classification of vascular plants is according to Mabberley (3), while the classification of bryopsida is according to Crosby and Magill (4). Additional classificatory information for the terrestrial plants was excerpted from Sitte *et al*. (5) and Farr *et al*. (6). The classification of the 'true' fungi or Eumycota is according to Hawksworth *et al*. (7), Kurtzman and Fell (8), and de Hoog and Guarro (9). The remaining eukaryotes, viz. the protoctists are classified according to Margulis *et al*. (10). Overall, species are included in the database under the binomial used for the publication of the sequence. We therefore refrained from making any taxonomic name change, even where obviously needed.

Archaea and Bacteria are classified according to phylogenetic position observed in evolutionary trees. The species are assigned to one of the taxa described by Woese and co-workers (11,12) and our research group (13,14).

SECONDARY STRUCTURE MODEL

The LSU rRNA adopts a very similar core structure in all species, which consists of a central multibranched loop from which several helices emanate (15–18). In Bacteria and most Archaea the central loop is closed by a stem helix joining the 5′ and 3′ ends of the molecule. In eukaryotes, the central loop is not closed by a stem helix, and the conserved core is interspersed with hypervariable regions which vary extremely in length and sequence, even between relatively closely related species. Because of this, the structure for these regions cannot always be conclusively determined for all sequences. In 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.

Figure 2 shows the secondary structure model incorporated in the database for the LSU rRNA of the ciliate *Tetrahymena termophilus*. The structures branching from the central loop are labelled from A to I, starting from the stem helix (not present in Fig. 2). Within each of these structures, helices are numbered in the 5′ to 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. No structure is proposed for the hypervariable regions enclosed by helices C1 and E20.

AVAILABILITY AND FORMAT OF THE DATABASE

The LSU rRNA database can be accessed on the World Wide Web (WWW) at URL http://rrna.uia.ac.be/lsu/ . Besides the database, this server also offers software for working with alignments and structure, and a links section, where links to many other interesting biological and general sites can be found. The data in the LSU rRNA database can be obtained using several methods. The simple list interface lets the user select sequences from a list of species, ordered per taxonomic group. The corresponding sequence is returned in the distribution format: this format contains a list of information about the sequence such as accession number and taxonomic position, followed by the organism name and the sequence. Each part of a fragmented sequence or sequences consisting of several exons, is preceded by its own annotations. Each sequence consists of a range of nucleotide symbols interspersed with gap symbols necessary for alignment, and special symbols indicating the structure. The sequence ends are indicated by an asterisk.

Access is easier through the query or list interface: Using forms the user can select a number of sequences and obtain them in a number of formats. Currently supported formats are DCSE (19) alignment and reference files, EMBL, NBRF/PIR, TREECON (20), the distribution format and a printable format. Some formats (DCSE, printable and distribution) contain information about the secondary structure of the sequences by the insertion of special symbols indicating start and end of structure elements in the alignment. If such a format is chosen, the appropriate 'Helix numbering' line that indicates the names of each helix segment will be added to the selection. In the printable format, the alignment has been sliced into blocks that fit onto a page. This format is limited to a selection of 100 sequences.

In the list interface, sequences are selected by name from a list sorted by taxonomic group. In the query interface, sequences are selected by specifying search terms in one or more fields. All sequences containing one or more of the search terms in the respective annotation will be returned. Multiple search terms in the same field 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 selection can further be limited to specific taxonomic groups with the check buttons at the bottom of the query page. When one or more of these are checked, only matching sequences from these taxonomic groups will be returned. If taxonomic groups are indicated, but all the query fields are left blank, all sequences in these groups will be returned.

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.

ACKNOWLEDGEMENTS

Our research is supported by the Fund for Scientific Research—Flanders and by the Special Research Fund of the University.

Figure 2. (Opposite) Secondary structure model for *Tetrahymena pyriformis* 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.

P.DeR. and Y.V.deP. are Research Fellows 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.
2 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates*. Sinauer Associates, Inc. Sunderland.
3 Mabberley,D.J. (1987; reprint 1996) *The Plant Book. A Portable Dictionary of the Higher Plants*. Cambridge University Press, Cambridge, UK.
4 Crosby,M.R. and Magill,R.E. (1978) *A Dictionary of Mosses* (second printing). Monogr. Syst. Bot. **3**.
5 Sitte,P., Ziegler,H., Ehrendorfer,F. and Bresinsky,A. (1991) *Lehrbuch der Botanik für Hochschulen*. Übersicht des Pflanzenreiches: 530–828. Stuttgart, Fischer.
6 Farr,E.R., Leussink,J.A. and Stafleu,F. (1979) *Index Nominum Genericorum (Plantarum)*. Regn. Veget. 100–102.
7 Hawksworth,D.L., Kirk,P.M., Sutton,B.C. and Pegler,D.N. (1995) *Ainsworth & Bisby's Dictionary of the Fungi*, 8th ed. CAB International, Egham.
8 Kurtzman,C.P. and Fell,J.W. (1998) *The Yeasts, a Taxonomic Study*, 4th ed. Elsevier, Amsterdam.
9 de Hoog,G.S. and Guarro,J. (1995) *Atlas of Clinical Fungi*. Centraalbureau voor Schimmelcultures, Baarn; Universitat Rovira i Virgili, Reus.
10 Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds) (1990) *Handbook of Protoctista*., Jones and Bartlett Publishers, Boston.
11 Woese,C.R. (1987) *Microbiol. Rev*., **51**, 221–271.
12 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol*., **176**, 1–6.
13 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res*., **21**, 2967–2971.
14 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol*., **17**, 32–38.
15 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.
16 Brimacombe,R. and Stiege,W. (1985) *Biochem. J*., **229**, 1–17.
17 Leffers,H., Kjems,J., Østergaard,L., Larsen,N. and Garrett,A. (1987) *J. Mol. Biol*., **195**, 43–61.
18 Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res*., **21**, 3055–3074.
19 De Rijk,P. and De Wachter,R. (1993) *Comput. Applic. Biosci*., **9**, 735–740.
20 Van de Peer,Y. and De Wachter,R. (1994) *Comput. Applic. Biosci*., **10**, 569–570.