Nucleic Acids Research, 1997, Vol. 25, No. 1 **117–122**

Database on the structure of large ribosomal subunit RNA

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

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

Received October 15, 1996; Accepted October 21, 1996

ABSTRACT

The latest release of the large ribosomal subunit RNA database contains 429 sequences. All these sequences are aligned, and incorporate secondary structure information. The rRNA WWW Server at URL http://rrna.uia.ac.be/ provides researchers with an easily accessible resource to obtain the data in this database in a number of computer-readable formats. A new query interface has been added to the server. If necessary, the data can also be obtained by anonymous ftp from the same site.

INTRODUCTION

All new and updated entries in the EMBL database (1) are scanned for the presence of ribosomal RNA (rRNA) sequences by the Current Sequence Awareness program (a service of the Belgian EMBnet Node). On the basis of the feature table (when present) the actual ribosomal RNA sequences are extracted and sorted manually into the small subunit rRNA (SSU rRNA) and large subunit rRNA (LSU rRNA) databases. Extraction of the LSU rRNA sequence has to be done with care since it often consists of several fragments (e.g. 5.8S, 4.5S) or exons. Manual sorting is necessary because no uniform description for these types of molecules is used. The taxonomic definition of the sequences is adapted as described below. The sequences are then used to update older entries or added to the database as new entries and aligned using a combination of automatic and manual methods as provided by the program DCSE (2). During the alignment process, information about the secondary structure of each molecule is incorporated in the alignment. The structural information is also used to refine the alignment in an iterative manner.

The LSU rRNA sequences and their alignments, together with secondary structure information, literature references, accession numbers and taxonomic information are regularly made available on-line in a number of formats suitable for use in computer programs. Apart from the obvious use in structural research and phylogeny, the database can be useful for finding target sequences, for the detection of micro-organisms (e.g. 3–5) or the design of PCR primers. Careful comparison of a sequence to a set of known aligned sequences and study of its structure can also reveal potential sequencing errors.

TAXONOMIC CLASSIFICATION

The taxonomic classification of species in our database is different from that followed by the EMBL database. Therefore, the taxonomic information is adapted for all sequences. The taxonomic classification of the eukaryotic species is according to Brusca and Brusca (6) for the Animalia, according to Cronquist (7) for the higher plants, according to Ainsworth *et al.* (8) for the zygomycetes and ascomycetes, according to Moore (9) for the basidiomycetes, and according to Margulis *et al.* (10) for the remaining eukaryotes, viz. the Protoctista.

The classification of Archaea and Bacteria is based on the construction of evolutionary trees. In short, evolutionary trees are constructed by the neighbor-joining method (11) for all new sequences retrieved from the EMBL nucleotide sequence library. According to the phylogenetic position observed, the species are assigned to one of the taxa described by Woese and co-workers (12,13) and our research group (14,15). For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (16).

CONTENTS OF THE DATABASE

The database only contains complete or reasonably complete sequences. Partial sequences are excluded when <70% of the estimated chain length of the molecule has been sequenced. The chain length of a partially determined sequence is estimated by comparing it with a complete sequence of a closely related species. The latest release of the database (autumn 1996) on LSU rRNA contains a total of 429 sequences. As illustrated in Figure 1, these are not evenly distributed over the different taxonomic groups: the database holds 163 mitochondrial, 134 bacterial, 46 plastidial and 23 archaeal sequences, while only 63 eukaryotic sequences are present. The eukaryotic taxa in the database and the number of their representatives are listed in detail in Table 1.

SECONDARY STRUCTURE MODEL

The LSU rRNA molecules of Bacteria, Archaea and plastids all adopt a very similar core structure. This core structure can also be readily seen in eukaryotic sequences but is interspersed with insertion regions, which show extreme variation in length and sequence. The structure for these insertion regions has not always been conclusively determined for all sequences in the database. In mitochondria the structural variability of the core is much

* To whom correspondence should be addressed. Tel: +32 3 820 23 19; Fax: +32 3 820 22 48; Email: dwachter@uia.ua.ac.be

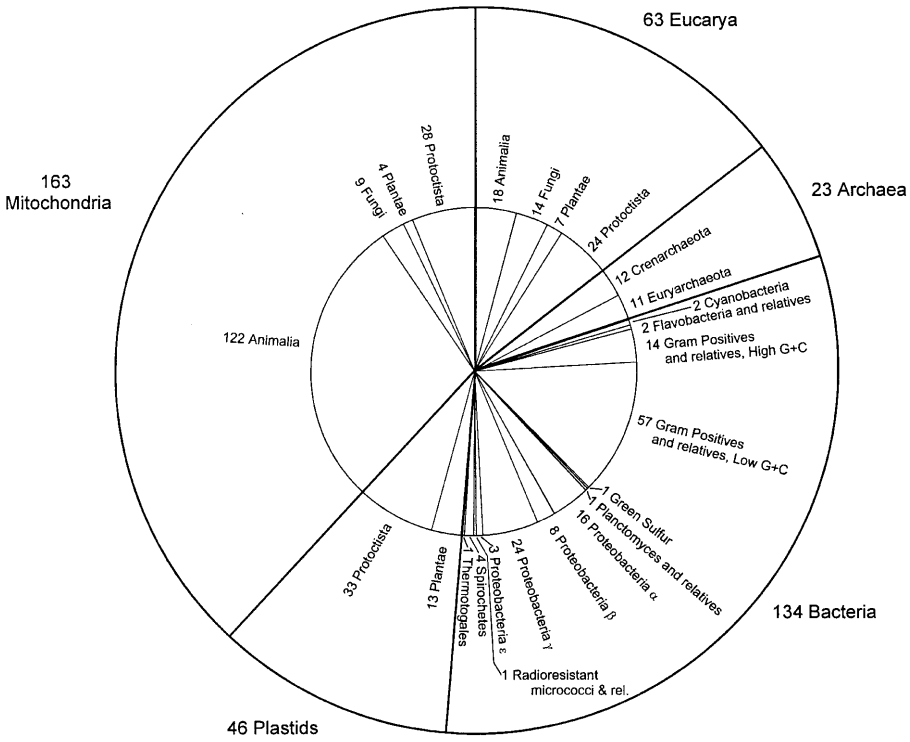

Figure 1. Distribution of representatives of the different taxonomic groups in the database. The total number of sequences is 429.

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.

The secondary structure model incorporated in the database is illustrated in Figure 2 for the LSU rRNA of the plant *Arabidopsis thaliana*. It conforms largely to the model developed in earlier studies (17–20). The basis of the structure is formed by a central multibranched loop from which several helices emanate. In Bacteria and most Archaea the central loop is closed by a stem helix joining together the 5′ and 3′ ends of the molecule.

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 from 5′ to 3′. Helices get a different number when they are separated by a multibranched loop. Helices not belonging to the core structure but specific to certain taxa are named after the preceding core helix followed by an underscore and a number. The indicated helix numbering may have to be revised if additional structural elements are identified in the future.

AVAILABILITY AND FORMAT OF THE DATABASE

All sequences in the database are stored in separate files in a very simple distribution format, which can be readily used by computer programs, or easily converted to other formats. The files start with information about the sequence such as the accession number and taxonomic position. This information is followed by the organism name and the sequence. For fragmented sequences or sequences consisting of several exons, all segments are stored in the same file, but each sequence part is preceded by its own annotations.

The sequences 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 are present for researchers who wish to use the secondary structure information. When these are incorporated into an alignment, they indicate the numbers of each helix segment.

People who wish to use the sequences, alignments or structures in their own research can obtain these easily through the World Wide Web (WWW). The LSU rRNA home page, shown in Figure 3, can be reached at http://rrna.uia.ac.be/lsu/ . It offers several methods to obtain the desired data. As described in the previous issue (21), sequences can be selected from a list using a forms interface and obtained in a number of formats. Currently supported formats are DCSE (2) alignment and reference files, EMBL, NBRF/PIR, TREECON, the distribution format and a printable format in which the alignment has been sliced into blocks that fit onto a page. The latter format is limited to a selection of 100 sequences.

Since the list of species can become quite long, downloading it might be a problem for computers with a slow connection or limited memory. Therefore a new interface was added, where the user can select the desired sequences by using a query. The query page shown in Figure 4 consists of three parts. In the first part the format in which the sequences will be downloaded can be selected from the option box labelled ‘Format’. The same formats as described above are supported. The second part contains several fields, in which a list of search terms can be typed. When only one field is filled, all sequences containing one or more of the terms in their respective annotation will be returned. The search terms are separated by spaces; if a search term has to

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

Kingdom Animalia[a]			
Phylum	Class	Number of sequences[b]	
		N	M
Nematoda	Secernentea	1	2
	Uncertain Affiliation		2
Annelida	Oligochaeta		1
Arthropoda	Malacostraca		2
	Insecta	3	13
Mollusca	Bivalvia		1
	Gastropoda		2
	Polyplacophora		1
	Pulmonata		1
Echinodermata	Asteroidea		1
	Echinoidea		3
Chordata	Ascidiacea	1	
	Agnatha		1
	Amphibia	3	3
	Aves		26
	Mammalia	3	42
	Osteichthyes	7	17
	Reptilia		4
Total		**18**	**122**

Kingdom Fungi[c]			
Subphylum	Class	Number of sequences[b]	
		N	M
Ascomycotina	Hemiascomycetes	9	4
	Plectomycetes		3
	Pyrenomycetes		2
	Uncertain Affiliation	1	
Basidiomycotina	Heterobasidiomycetes	2	
Zygomycotina	Zygomycetes	2	
Total		**14**	**9**

Kingdom Plantae				
Phylum	Class	Number of sequences[b]		
		N	M	P
Bryophyta	Marchantiopsida		1	1
Magnoliophyta	Liliopsida	1	2	3
	Magnoliopsida	6	1	8
Pinophyta	Pinopsida			1
Total		**7**	**4**	**13**

Kingdom Protoctista				
Phylum	Class	Number of sequences[b]		
		N	M	P
Apicomplexa	Coccidia	4	1	
	Hematozoa	2	3	1
Bacillariophyta	Bacillariophyceae			2
Chlorophyta	Chlorophyceae	1	4	19
Chytridiomycota			1	
Ciliophora		2	5	
Dictyostelida		1	1	
Dinoflagellata		1		
Euglenida		1		5
Eustigmatophyta	Eustigmatophyceae			2
Hypochytriomycota		1		
Oomycota		1		
Phaeophyta		1	1	1
Plasmodial slime molds	Myxomycota	2		
Rhizopoda	Lobosea	1	1	
Rhodophyta			1	3
Zoomastigina	Kinetoplastida	3	10	
	Diplomonadida	3		
Total		**24**	**28**	**33**

[a]The Metazoan taxa are listed in the same order as they appear in (6).
[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.

include a space, it should be surrounded by double quotes. If more than one field is filled, only sequences matching both queries will be returned. The third 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 groups in this section are checked, but no fields in the second part are filled in, all sequences in the checked groups will be returned when the 'Get Sequences' button is clicked on.

The files from the LSU rRNA database are also obtainable by anonymous ftp on rrna.uia.ac.be (143.169.8.11), and are made available to the EMBL nucleotide library for distribution. On the anonymous ftp server, a file called 'readme' will be present which describes the latest state of the database, giving the contents of the files and directories, and a description of the programs available for format conversion, alignment editing (2) and phylogenetic tree construction (22). 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 is supported by the BIOTECH programme of the Commission of European Communities (contract BIO2-CT94-3098), by the Programme on Interuniversity Poles of Attraction of the Office for Scientific, Cultural, and Technical Matters of the Belgian State (contract 23), by the National Fund for Scientific Research and by the Special Research Fund of the

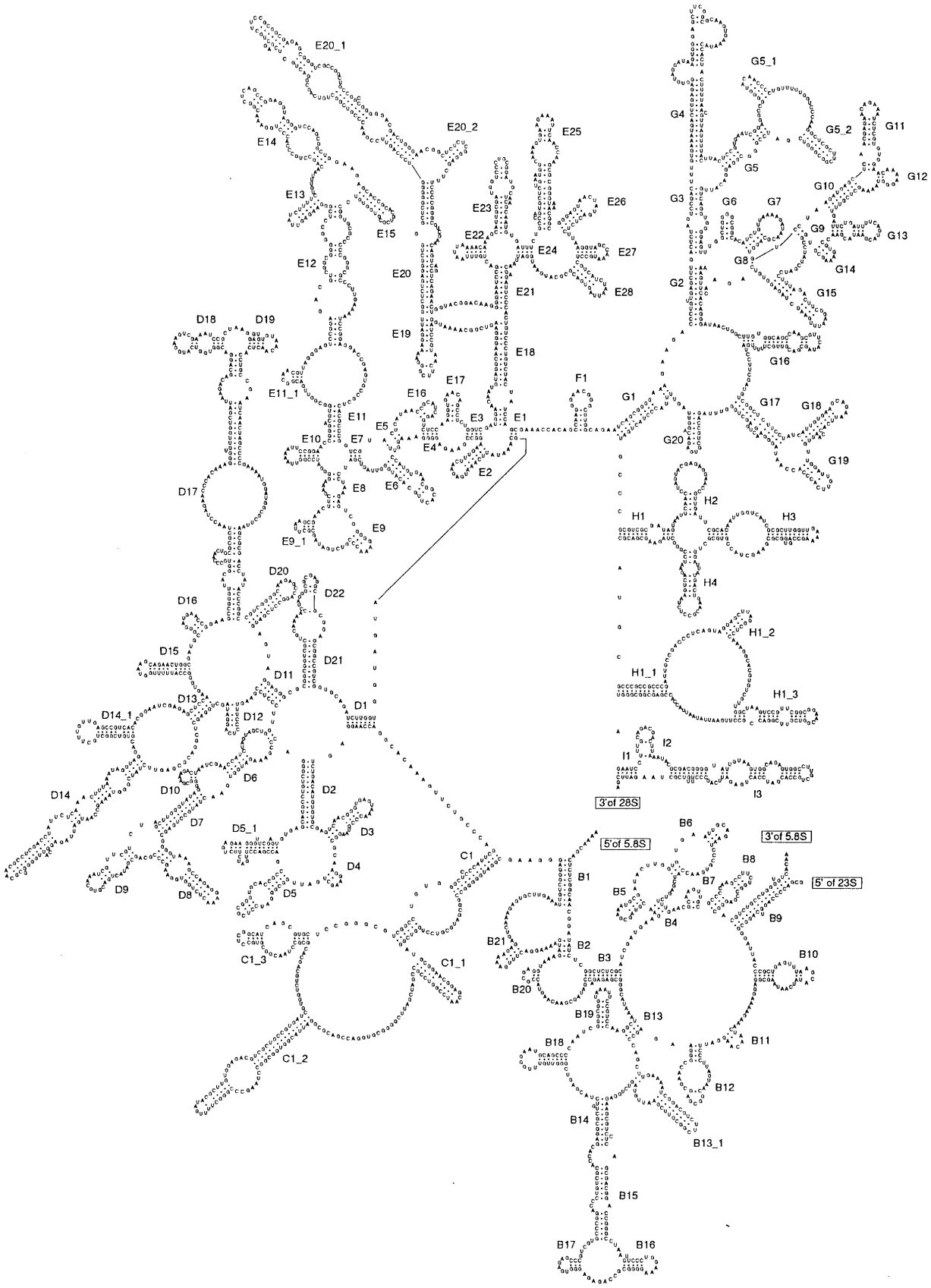

Figure 2. Secondary structure model for *A.thaliana* LSU rRNA. The sequence is written clockwise from 5′ to 3′ terminus.

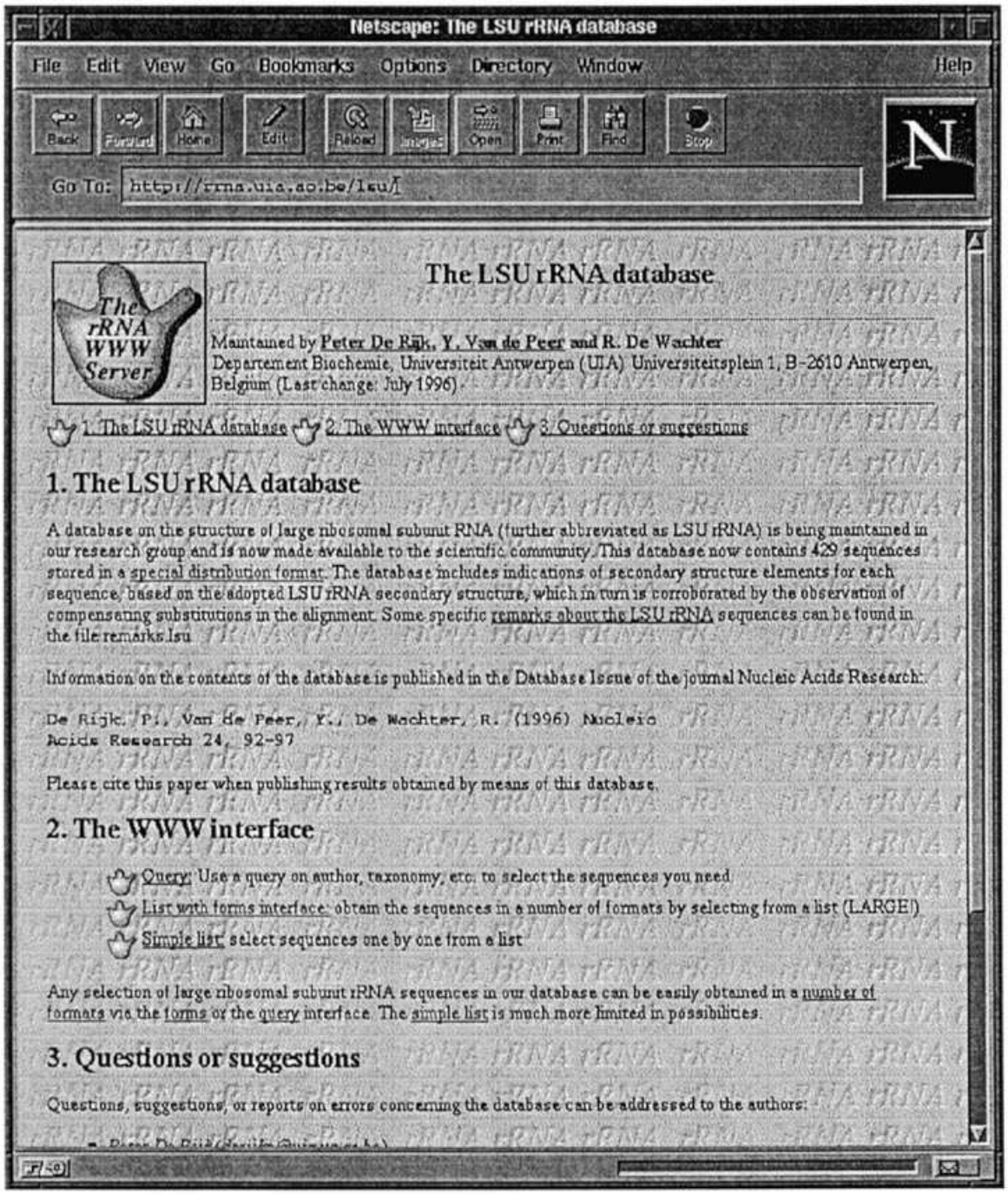

Figure 3. The LSU rRNA home page.

university. We thank Sabine Chapelle for the computer drawing of the secondary structure model. Yves Van de Peer is a Research Assistant of the National Fund for Scientific Research.

REFERENCES

1 Rodriguez-Tome,P., Stoehr,P.J., Cameron,G.N. and Flores,T.P. (1996) *Nucleic Acids Res.*, **24**, 6–12.
2 De Rijk,P. and De Wachter,R. (1993) *Comput. Appl. Biosci.*, **9**, 735–740.
3 Bastyns,K., Chapelle,S., Vandamme,P., Goossens,H. and De Wachter,R. (1994) *System. Appl. Microbiol.*, **17**, 563–568.
4 Betzl,D., Ludwig,W. and Schleifer,K.H. (1990) *Appl. Environ. Microbiol.*, **56**, 2927–2929.
5 Lew,A.E. and Desmarchelier,P.M. (1994) *J. Clin. Microbiol.*, **32**, 1326–1332.
6 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates* Sinauer Associates, Inc., Sunderland.
7 Cronquist,A. (1971) *Introductory Botany.* Harper & Row, New York.
8 Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) *The Fungi: an Advanced Treatise.* Academic Press, New York, Vol. 4A.

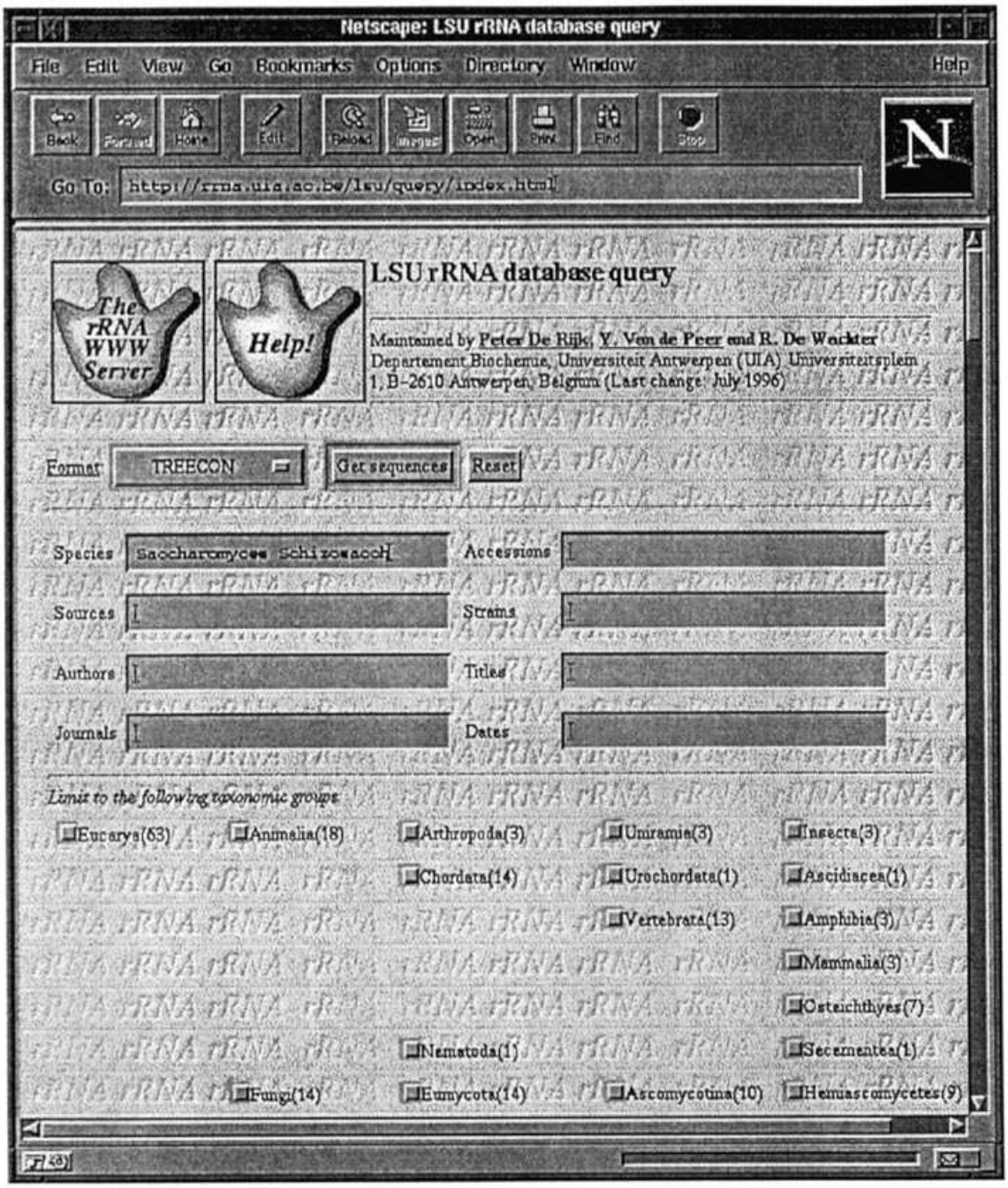

Figure 4. The query interface to the LSU rRNA database. Sequences or alignments can be obtained as described in the text.

9 Moore,R.T. (1988) in Moriarty,Ch. (ed.), *Taxonomy Putting Plants and Animals in Their Place*. Royal Irish Academy, Dublin, pp. 61–88.
10 Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds) (1990) *Handbook of Protoctista.* Jones and Bartlett Publishers, Boston.
11 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
12 Woese,C.R. (1987) *Microbiol. Rev.*, **51**, 221–271.
13 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
14 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
15 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.*, **17**, 32–38.
16 Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
17 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.
18 Brimacombe,R. and Stiege,W. (1985) *Biochem. J.*, **229**, 1–17.
19 Leffers,H., Kjems,J., Østergaard,L., Larsen,N. and Garrett,A. (1987) *J. Mol. Biol.*, **195**, 43–61.
20 Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.*, **21**, 3055–3074.
21 De Rijk,P., Van de Peer,Y. and De Wachter,R. (1996) *Nucleic Acids Res.*, **24**, 92–97.
22 Van de Peer,Y. and De Wachter,R. (1994) *Comput. Appl. Biosci.*, **10**, 569–570.