92–97 Nucleic Acids Research, 1996, Vol. 24, No. 1

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 September 27, 1995; Revised and Accepted October 31, 1995

ABSTRACT

Our database on large ribosomal subunit RNA contained 334 sequences in July, 1995. All sequences in the database are aligned, taking into account secondary structure. The aligned sequences are provided, together with incorporated secondary structure information, in several computer-readable formats. These data can easily be obtained through the World Wide Web. The files in the database are also available via anonymous ftp.

INTRODUCTION

Large ribosomal subunit RNA (further abbreviated as LSU rRNA) has proven an interesting molecule to perform phylogenetic analysis (1). It is also very useful as a target for the detection of micro-organisms (e.g. 2–4). For this type of studies, a database of aligned LSU rRNA sequences as presented in this paper can be very useful. The database can also be used for the design of PCR primers and the elucidation of the secondary structure of newly determined sequences. It can also be an invaluable tool to find sequence errors introduced during sequence analysis. These errors cause anomalies in the sequence which can often be detected by alignment to a set of known sequences and comparison of their possible secondary structure.

The database has been continuously updated by scanning the EMBL sequence database (5) for updated or new rRNA sequences using the Current Sequence Awareness program (a service of the Belgian EMBnet Node). These sequences are respectively used to update older entries, or added to the database as new entries. They are then aligned, and their secondary structure is investigated and incorporated into the alignment using the program DCSE (6).

It is our goal to offer researchers easy on-line access to these LSU rRNA sequences and their alignments, together with secondary structure information, literature references, accession numbers and taxonomic information. All the data are obtainable in a number of formats suitable for use in computer programs. Other databases on LSU rRNA are also available, offering mutation data (7), predrawn secondary structure models (8) and also sequences and alignments (9).

CONTENTS OF THE DATABASE

Only complete or reasonably complete sequences are incorporated into the database. Partial sequences are included only if the combined length of the sequenced segments amounts to at least 70% of the estimated chain length of the molecule. The chain length of a partially determined sequence is estimated by comparing it to a complete sequence of a closely related species. The database on LSU rRNA currently contains (July 1995) 334 sequences. This number comprises 46 eukaryotic, 17 archaeal, 102 bacterial, 38 plastidial and 131 mitochondrial sequences.

Table 1 lists the number of representatives for each of the eukaryotic taxa in the database. The taxonomic classification of the species is according to Brusca and Brusca (10) for the Animalia, according to Cronquist (11) for the higher plants, according to Ainsworth *et al.* (12) for the zygomycetes and ascomycetes, according to Moore (13) for the basidiomycetes, and according to Margulis *et al.* (14) for the remaining eukaryotes, viz. the Protoctista.

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
Arthropoda	Insecta	2	10
	Malacostraca		2
Mollusca	Bivalvia		1
	Pulmonata		2
	Polyplacophora		1
Echinodermata	Echinoidea		2
Chordata	Ascidiacea	1	
	Agnatha		1
	Amphibia	3	2
	Aves		18
	Mammalia	3	35
	Osteichtyes		12
	Reptilia		4
Total:		**10**	**92**

* To whom correspondence should be addressed

Kingdom Fungi				
Subphylum	Class	Number of sequences[b] N	M	
Zygomycotina	Zygomycetes	1		
Ascomycotina	Hemiascomycetes	5	3	
	Plectomycetes		3	
	Pyrenomycetes		2	
	Uncertain affiliation	1		
Basidiomycotina	Heterobasidiomycetes	2		
Total:		**9**	**8**	

Kingdom Plantae				
Phylum	Class	Number of sequences[b] N	M	P
Bryophyta	Marchantiopsida		1	1
Magnoliophyta	Liliopsida	1	2	3
	Magnoliopsida	6	1	8
Total:		**7**	**4**	**12**

Kingdom Protoctista[c]				
Phylum	Class	Number of sequences[b] N	M	P
Apicomplexa	Coccidia	3	1	
	Hematozoa		3	
Chlorophyta	Chlorophyceae	1	4	19
Ciliophora		2	5	
Dictyostelida		1	1	
Dinoflagellata		1		
Euglenida		1		5
Granuloreticulosa		1		
Oomycota		1		
Phaeophyta			1	1
Plasmodial Slime Molds	Myxomycota		2	
Rhizopoda	Lobosea	1	1	
Rhodophyta			1	1
Zoomastigina	Diplomonadida	3		
	Kinetoplastida	3	10	
Total:		**20**	**27**	**26**

[a]The Metazoan taxa are listed in the same order as they appear in (10).
[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 Protoctist phyla and classes are ordered alphabetically.

Table 2 covers the bacterial and archaebacterial LSU rRNA sequences. Their classification is based on the construction of evolutionary trees. In short, evolutionary trees are constructed by the neighbor-joining method (15) for all new sequences retrieved from the EMBL (5) nucleotide sequence library. According to the phylogenetic position observed, the species are assigned to one of the taxa described by Woese and co-workers (16,17) and our research group (18,19). For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (20). The latter division is further subdivided into seven subdivisions.

Table 2. Prokaryotic taxa represented in the database and number of their representatives

Bacteria	
Division	Number of sequences[a]
Cyanobacteria	1
Flavobacteria and relatives	2
Gram Positives and relatives, Low G+C	45
Gram Positives and relatives, High G+C	12
Green Sulfur	1
Planctomyces and relatives	1
Proteobacteria α	11
Proteobacteria β	8
Proteobacteria γ	13
Proteobacteria ε	2
Radioresistant micrococci and relatives	1
Spirochetes	4
Thermotogales	1
Total:	**102**

Archaea		
Division	Subdivision	Number of sequences[a]
Euryarchaeota	Archaeoglobales	1
	Halobacteria	5
	Methanobacteriales	1
	Methanococcales	1
	Methanomicrobium group	1
	Thermococcales	1
	Thermoplasma	1
Crenarchaeota:		6
Total:		**17**

[a]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 strains, or for different genes of the same strain.

HETEROGENEITY IN SEQUENCE AND CHAIN LENGTH

An intriguing feature of the LSU rRNA is the large variation in length between different species as illustrated in Figure 1.

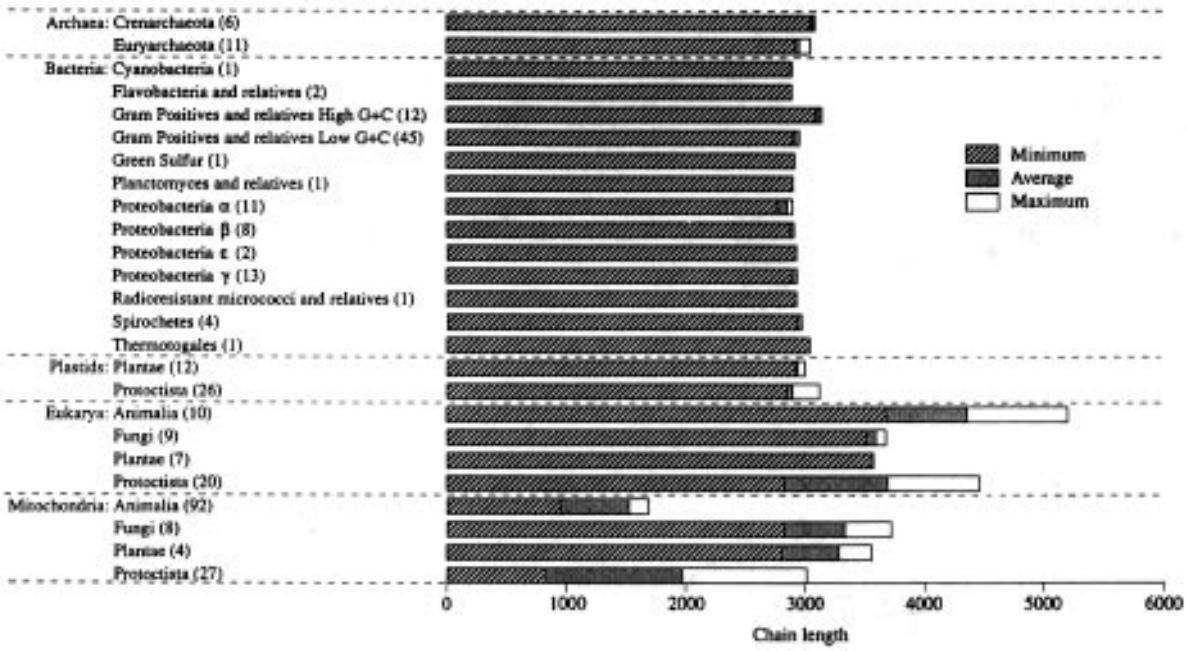

Figure 1. Length heterogeneity in LSU rRNA. For each of the groups in the figure, the number of bases in each sequence belonging to the group was counted. The bars indicate the lengths of the smallest, the average and the longest sequence in the group. The number of complete sequences in each group is indicated between brackets.

Whereas bacterial, archaeal and plastidial LSU sequences have a relatively constant length of ~2900 nt, eukaryotic sequences show a great diversity in length, ranging from sizes comparable with those of the bacteria to over 5000 bases in the *Homo sapiens* sequence. The presence of extra nucleotides seems to be restricted mainly to several extremely variable insertion regions, which occupy a constant position relative to the more conserved parts of the sequences (21,22). The variation in sequence length is even larger in mitochondria. The ribosomal RNAs found in animal and kinetoplastid mitochondria even miss large parts of the sequence conserved in other LSU rRNAs, and can be <1000 nt in size. Plant and fungal mitochondrial LSU rRNAs have chain lengths comparable with or larger than those found in bacteria.

SECONDARY STRUCTURE MODEL

The secondary structure model followed in the database conforms largely to the model developed in earlier studies (8,23–25). It is illustrated in Figure 2 for the LSU rRNA of the archaebacterium *Sulfolobus acidocaldarius*. The secondary structure of the molecule is treelike, with the helices forming branches which end either in a hairpin or in a multibranched loop. The stem of the tree joins the 5′ and 3′ ends of bacterial LSU rRNAs. From this stem emanates a central multibranched loop. In Archaea this stem is generally shorter than in Bacteria, and not always present. In Eucarya, the stem helix has completely disappeared.

Bacterial and plastidial LSU rRNA molecules adopt a structure very similar to that of the Archaea. Eukaryotic sequences also have a similar core structure, but in the variable insertion regions the structure has not always been conclusively determined. In mitochondria the structural variability of the core is much higher than in other species, and in the mitochondria of kinetoplastids and animals even many helices of the core are absent. As a consequence, the alignment and proposed secondary structure of the mitochondrial LSU rRNAs is less dependable.

The following provisional helix numbering system is used in Figure 2. Structures branching from the central loop are labelled A–I, starting with the stem helix. Within each of these structures, helices bear a different number when they are separated by a multibranched loop. All numbering is sequentially from 5′ to 3′. Helices not belonging to the core structure but specific to certain taxa are named after the preceding core helix followed by an underscore and number. The helix numbering may have to be revised if additional structural elements are identified in the future.

AVAILABILITY AND FORMAT OF THE DATABASE

Each LSU rRNA sequence, together with gaps, secondary structure information and reference information is stored in a separate file in a special distribution format. The easiest way to obtain the data is through the World Wide Web (WWW). The LSU rRNA home page can be reached at http:// rrna.uia.ac.be/ rrna/lsuform.html. Using forms, a file containing any selection of sequences can be obtained in a number of formats. The sequences can be selected either one by one using list boxes, or by whole groups using check buttons, or a combination of both (see Fig. 3). The desired format should be indicated in the appropriate selection box. Clicking on the button labelled ‘Get sequences’ will create and transfer the resulting file. Currently supported formats are DCSE (6) alignment and reference files, EMBL, NBRF/PIR, the distribution format, and a printable form in which the alignment has been cut into blocks which fit onto a page. The latter format is limited to a selection of 100 sequences.

The files from the LSU rRNA database are also obtainable by anonymous ftp on rrna.uia.ac.be (143.169.8.11), and are also made available to the EMBL nucleotide library for distribution.

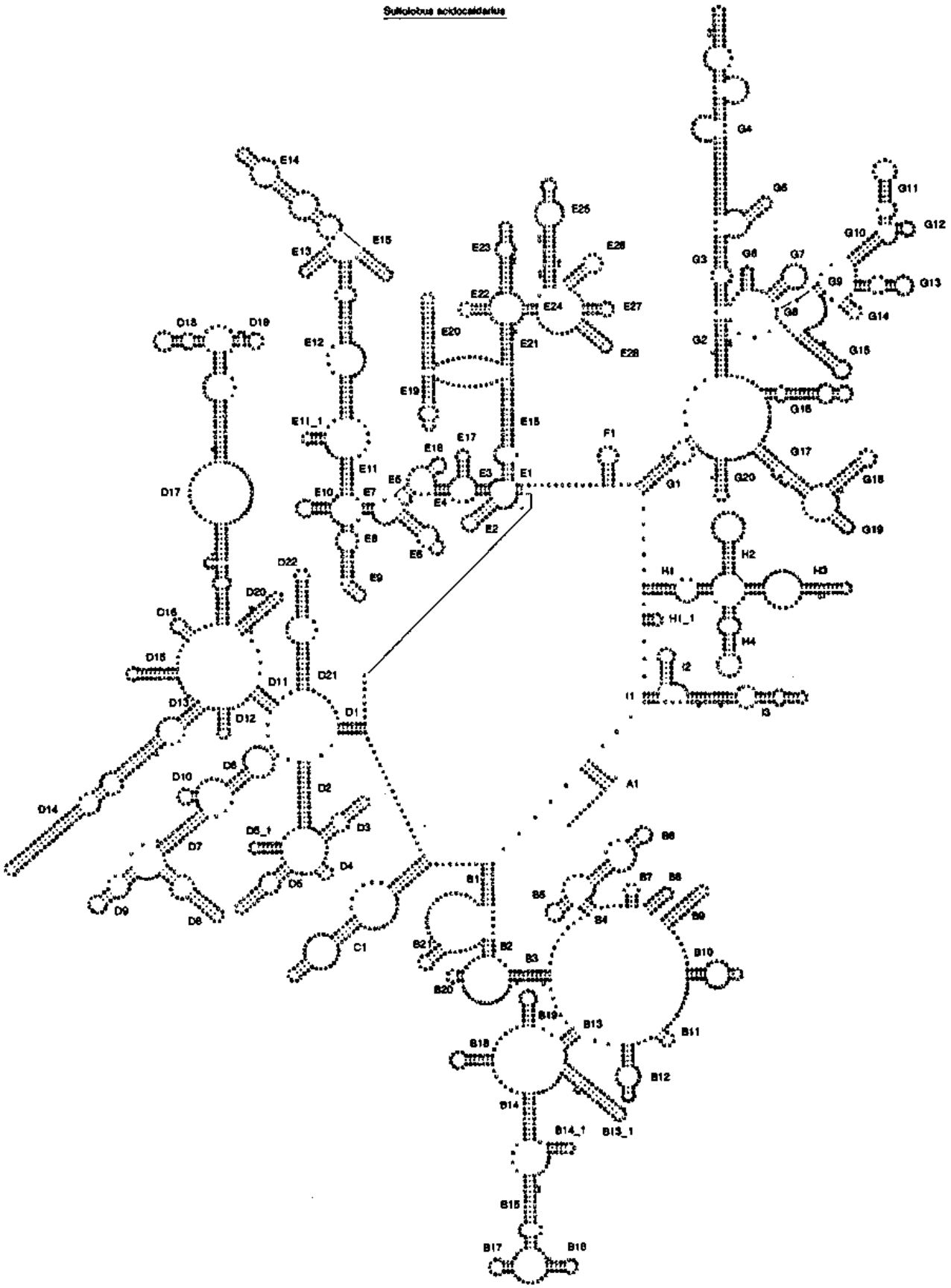

Figure 2. Secondary structure model for *Sulfolobus acidocaldarius* LSU rRNA. The sequence is written clockwise from 5′ to 3′ terminus.

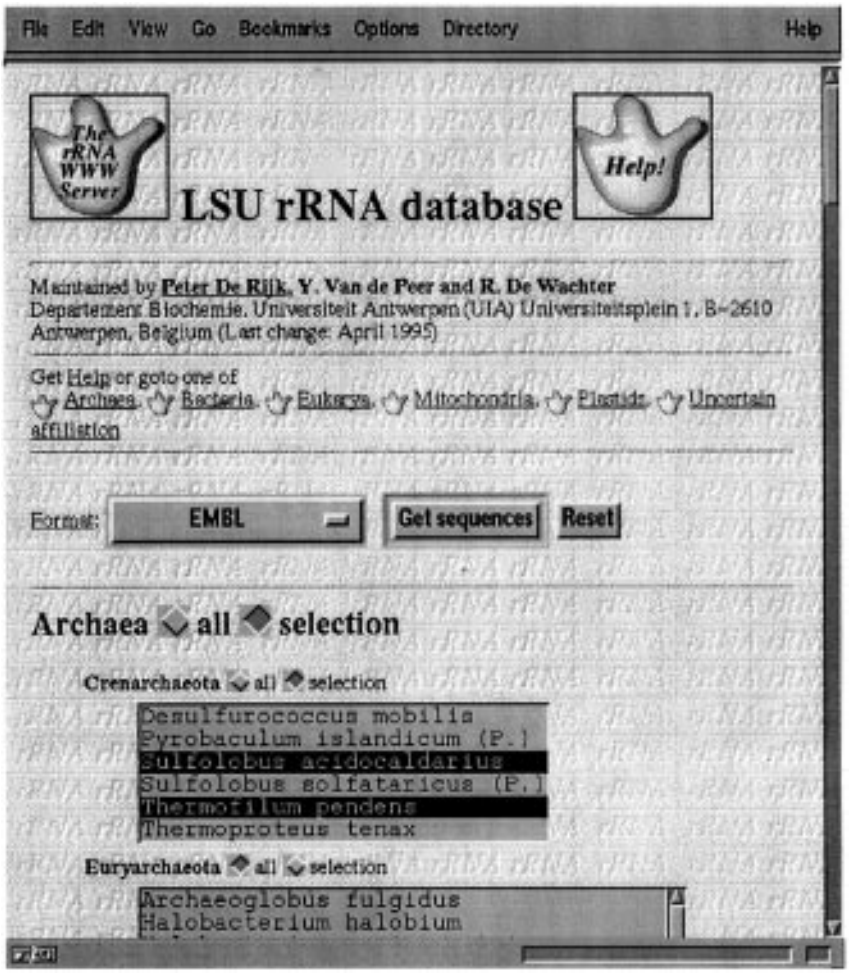

Figure 3. The WWW interface to the LSU rRNA database. Sequences or alignments can be obtained as described in the text. If the user clicks on the button labelled 'Get sequences' on this screen, he obtains the sequences of *Sulfolobus acidocaldarius, Thermofilum pendens* and all Euryarchaeota in the EMBL file format.

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 (6) and phylogenetic tree construction (26). Since each sequence is stored in a separate file, the user can also get any selection of sequences using ftp. The names of the files in the database are produced by taking characters of the genus and species names. Their extension is a code indicating the phylogenetic group to which the species belongs. This makes it possible to either retrieve specific sequences using the full file name, or to retrieve a set of sequences belonging to a phylogenetic group using wild cards. However, people using anonymous ftp will have to convert these files into a desired format themselves. A program for this purpose is available on the server.

The distribution format in which the files are stored is very simple, so that the files can be used readily by computer programs, or can easily be converted to formats used by specific programs. The files start with a few header lines which contain data about the sequence such as the accession number and literature reference. These are followed by the organism name, and the sequence. The sequence consists of a range of nucleotide symbols interspersed with gap symbols necessary for alignment. The sequence end is 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 name of each helix segment.

When a sequence consists of several fragments resulting from processing, or of several exons, the sequence of each part ends with an asterisk, and has its own header containing the accession number, literature reference and a description of the sequence segment. However, the segments are stored in the same file and have the same organism name.

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), and by the National Fund for Scientific Research. We thank Sabine Chapelle for the computer drawings of the secondary structure models. Peter De Rijk and Yves Van de Peer are Research Assistants of the National Fund for Scientific Research.

REFERENCES

1 De Rijk,P., Van de Peer,Y., Van den Broeck,I. and De Wachter,R. (1995) *J. Mol. Evol.*, **41**, 366–375.
2 Bastyns,K., Chapelle,S., Vandamme,P., Goossens,H. and De Wachter,R. (1994) *System. Appl. Microbiol.*, **17**, 563–568.
3 Betzl,D., Ludwig,W. and Schleifer,K.H. (1990) *Appl. Environ. Microbiol,* **56**, 2927–2929.
4 Lew,A.E. and Desmarchelier,P.M. (1994) *J. Clin. Microbiol.*, **32**, 1326–1332.
5 Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.H. (1994) *Nucleic Acids Res.* **22**, 3445–3449.
6 De Rijk,P. and De Wachter,R. (1993) *Comput. Applic. Biosci.*, **9**, 735–740.
7 Triman, K. L. (1996) this issue.
8 Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) *Nucleic Acids Res.*, **21**, 3055–3074.
9 Maidak,B.L., Olsen,G.J, Larsen,N, Overbeek,R, McCaughey,M. and Woese,C.W. (1996) *Nucleic Acids Res.* **24**, 82–85.
10 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates*, Sinauer Associates, Inc. Sunderland.
11 Cronquist,A. (1971) *Introductory Botany,* Harper & Row, New York.
12 Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) *The Fungi: an Advanced Treatise,* Academic Press, New York, Vol. 4A.
13 Moore,R.T. (1988) in Moriarty,Ch. (ed.) *Taxonomy Putting Plants and Animals in Their Place.* Royal Irish Academy, Dublin, pp. 61–88.
14 Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds) (1990) *Handbook of Protoctista.*, Jones and Bartlett Publishers, Boston.
15 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
16 Woese, C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
17 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
18 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **21**, 2967–2971.
19 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.* **17**, 32–38
20 Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
21 Veldman,G.M., Klootwijk,J., de Regt,V.C.H.F., Planta,R.J., Branlant,C., Krol,A. and Ebel,J.–P. (1981) *Nucleic Acids Res.*, **9**, 6935–6952
22 Michot,B., Hassouna,N. and Bachellerie,J–P. (1984) *Nucleic Acids Res.*, **12**, 4259–4279
23 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.
24 Brimacombe,R. and Stiege,W. (1985) *Biochem J.*, **229**, 1–17.
25 Leffers,H., Kjems,J. Østergaard,L., Larsen,N. and Garrett,A. (1987) *J. Mol. Biol.*, **195**, 43–61.
26 Van de Peer,Y. and De Wachter,R. (1993) *Comput. Applic. Biosci.*, **9**, 177–182.