Nucleic Acids Research, 1997, Vol. 25, No. 1 **111–116**

Database on the structure of small ribosomal subunit RNA

Yves Van de Peer, Jan Jansen, Peter De Rijk 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 Antwerp database on small ribosomal subunit RNA now offers more than 6000 nucleotide sequences (August 1996). All these sequences are stored in the form of an alignment based on the adopted secondary structure model, which is corroborated by the observation of compensating substitutions in the alignment. Besides the primary and secondary structure information, literature references, accession numbers and detailed taxonomic information are also compiled. For ease of use, the complete database is made available to the scientific community via World Wide Web at URL http://rrna.uia.ac.be/ssu/ .

CONTENTS OF THE DATABASE

In 10 years time, the number of small ribosomal subunit RNA (further abbreviated as SSU rRNA) sequences that have been compiled by our research group in Antwerp has increased from 57 (1) to 6085. In 1986, the complete or nearly complete SSU rRNA sequence was known for 14 eukaryotes, 13 bacteria, nine archaebacteria, four plastids and 16 mitochondria. In August 1996, 1484 eukaryotic, 4155 bacterial, 142 archaeal, 88 plastid and 216 mitochondrial sequences were available. All SSU rRNA sequences included in our database are stored in the form of an alignment and contain the postulated secondary structure pattern in encoded form. Partial SSU rRNA 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 with a complete sequence of a close relative.

Table 1 lists the different eukaryotic taxa and the number of representatives in the database. The taxonomic classification of the 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 ustomycetes, and according to Margulis *et al.* (6) for the remaining eukaryotes, viz. the Protoctista.

Table 2 covers the prokaryotic SSU rRNA sequences. The classification of prokaryotes is based on the construction of evolutionary trees. In short, new sequences retrieved from the EMBL (7) and/or GenBank (8) nucleotide sequence libraries are aligned with their presumed closest relative. Evolutionary trees are then constructed by the neighbor-joining method (9), and according to the phylogenetic position observed, the species are assigned to one of the taxa described by Woese and co-workers (10,11) and our research group (12,13). In the case of the Bacteria, no hierarchical distinction is made between divisions and subdivisions such as the α, β, γ, δ and ε subdivisions of the division Proteobacteria, since these subdivisions do not always form together a monophyletic cluster in evolutionary trees. It should also be noted that tree construction shows the Protobacteria γ subdivision to be a paraphyletic taxon, comprising the Proteobacteria β subdivision. For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (14). The latter division is further subdivided into eight subdivisions.

SECONDARY STRUCTURE AND NUCLEOTIDE VARIABILITY

The secondary structure models adopted for prokaryotic and eukaryotic SSU rRNAs were originally derived (15) by comparison of six eucaryal, one archaeal, four bacterial, two plastidial and one mitochondrial SSU rRNA sequences available in 1984 and by surveying 13 secondary structure models proposed at the time in papers listed in (15). Gradual improvements were made to the models, as reported in subsequent papers describing our database on SSU rRNA structure (1,16–19,12,20,21), taking into account compensating substitutions observed in our sequence alignments (22) and the results of studies by others (reviewed in 23).

Secondary structures encoded in the sequences are based either on the prokaryotic model, which is applicable to Bacteria, Archaea, plastids and mitochondria, or on the eukaryotic model applicable to all Eucarya. The two models are slightly different, each containing a number of structural elements specific for the group. The prokaryotic model is essentially identical to those distributed by Gutell (24), but the model followed for eukaryotic SSU rRNAs includes a secondary structure pattern in certain variable areas left undefined in the models of the latter author. It is illustrated in Figure 1 with the SSU rRNA nucleotide sequence of the chlorarachniophyte *Chlorarachnion reptans*.

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

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

Kingdom Animalia[a] Phylum	Class	Number of sequences[b] N	M
Placozoa		2	
Rhombozoa		3	
Orthonectida		1	
Porifera	Calcarea	2	
	Demospongiae	2	
Cnidaria	Anthozoa	3	
	Cubozoa	1	
	Hydrozoa	2	
Ctenophora		2	
Platyhelminthes	Cestoda	1	
	Trematoda	13	
	Turbellaria	7	
Nemertea	Anopla	2	
Rotifera		1	
Gastrotricha		1	
Nematoda	Secernentea	15	
Nematomorpha		2	
Priapula		2	
Entoprocta		2	
Acanthocephala	Archiacanthocephala	1	
Annelida	Oligochaeta	1	
	Polychaeta	3	
Sipuncula	Phascolosomida	1	
Echiura		1	
Pogonophora		1	
Vestimentifera	Basibranchia	1	
Arthropoda	Branchiopoda	3	
	Chilopoda	1	
	Chelicerata	25	
	Insecta	41	6
	Malacostraca	15	
	Maxillopoda	12	
Tardigrada		2	
Pentastomida	Pentastomata	1	
Mollusca	Bivalvia	23	
	Gastropoda	2	
	Polyplacophora	1	
Phoronida		2	
Ectoprocta	Phylactolaemata	1	
Echinodermata	Asteroidea	2	
	Crinoidea	1	
	Echinoidea	24	
	Holothuroidea	1	
	Ophiuroidea	1	
Chaetognatha		3	
Hemichordata	Enteropneusta	2	
Chordata	Agnatha	4	1
	Amphibia	18	4
	Aves	3	4
	Chondrichtyes	4	
	Mammalia	10	109
	Osteichtyes	3	10
	Reptilia	4	14
	Cephalochordata (Sub.)	1	
	Urochordata (Subphyl.)	2	
Total:		**285**	**148**

Kingdom Fungi[c] Subphylum	Class	Number of sequences[b] N	M
Ascomycotina	Discomycetes	17	
	Hemiascomycetes	67	9
	Loculoascomycetes	35	
	Plectomycetes	37	25
	Pyrenomycetes	22	1
	Uncertain affiliation	21	
Basidiomycotina	Heterobasidiomycetes	47	
	Hymenomycetes	18	
Ustomycotina	Ustomycetes	30	
Zygomycotina	Zygomycetes	28	3
Uncertain affiliation		3	
Total:		**325**	**38**

Kingdom Plantae[c] Phylum	Class	Number of sequences[b] N	M	P
Bryophyta	Anthocerotopsida	11		
	Bryopsida	14		
	Marchantiopsida	8	1	1
Equisetophyta		2		1
Lycopodiophyta	Isoetopsida	1		1
	Lycopodiopsida	8		3
Magnoliophyta	Liliopsida	8	5	3
	Magnoliopsida	141	3	20
Pinophyta	Cycadopsida	1		
	Gnetopsida	3		1
	Ginkgoopsida	1		
	Pinopsida	15		1
Polypodiophyta		16		5
Psilotophyta	Psilotopsida	3		2
Total:		**232**	**9**	**38**

Kingdom Protoctista[c] Phylum	Class	Number of sequences[b] N	M	P
Actinopoda	Heliozoa	1		
Apicomplexa	Coccidia	31		
	Hematozoa	47	3	
	Uncertain affiliation	7		
Bacillariophyta	Bacillariophyceae	5		3
	Coscinopiscophyceae	4		
Chlorarachnida		17		3
Chlorophyta	Charophyceae	24		3
	Chlorophyceae	90	3	15
	Prasinophyceae	6		
	Ulvophyceae	39		
Chrysophyta	Chrysophyceae	9		2
	Dictyochophyceae	1		
Chytridiomycota		9		
Ciliophora		59	5	
Conjugaphyta	Conjugatophyceae	9		1
Cryptophyta		12		4
Dictyostelida		1	1	
Dinoflagellata		16		
Euglenida		1		6
Eustigmatophyta	Eustigmatophyceae	5		
Glaucocystophyta	Glaucocystophyceae	1		3
	Uncertain affiliation			1
Granuloreticulosa		5		
Haplosporidia	Haplosporea	7		
Hyphochytridiomycota		1		
Labyrinthulomycota		5		
Microspora		39		
Myxozoa	Myxosporea	6		
Oomycota		4		
Phaeophyta		11	1	1
Plasmodial Slime Molds: Myxomycota		1		
Prymnesiophyta		12		3
Rhizopoda	Filosea	1		
	Lobosea	31	2	
Rhodophyta		84	2	5
Xanthophyta		1		
Zoomastigina	Amebomastigota	3		
	Choanomastigotes	2		
	Diplomonadida	9		
	Kinetoplastida	24	4	
	Parabasalia	1		
	Zoomastigophora	1		
Total:		**642**	**21**	**50**

[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 SSU 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.

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

Division	Subdivision	Number of sequences[a]
Bacteria		
Chlamydiae		8
Cyanobacteria		54
Fibrobacter		17
Flavobacteria and relatives		203
Fusobacterium and relatives		30
Gram Positives and relatives, Low G+C		1073
Gram Positives and relatives, High G+C		814
Green Sulfur		4
Green non sulfur		6
Planctomyces and relatives		31
Proteobacteria α		571
Proteobacteria β		172
Proteobacteria γ		695
Proteobacteria δ		98
Proteobacteria ε		114
Proteobacteria, uncertain affiliation		2
Radioresistant micrococci and relatives		31
Spirochetes		161
Thermotogales		7
Uncertain affiliation[b]		64
Total:		4155
Archaea		
Crenarchaeota		34
Euryarchaeota	Archaeoglobales	1
	Halobacteria	31
	Methanobacteriales	17
	Methanococcales	13
	Methanomicrobium group	31
	Methanopyrales	1
	Thermococcales	13
	Thermoplasma	1
Total:		142

[a]The number of sequences listed in the database is larger than the number of species (cf. Table 1).
[b]In some cases, it cannot be decided to which taxonomic group a species should be ascribed, since the clustering of its SSU rRNA sequence is unstable and depends on the tree construction method used and on the set of sequences included in the analysis.

Helices in the SSU rRNA secondary structure model are given a different number if separated by a multibranched loop (e.g. helices 9 and 10), by a pseudoknot loop (e.g. helices 1 and 2), or by a single-stranded area that does not form a loop (e.g. helices 2 and 32). A single number is given to 50 'universal' helices, which are present in all SSU rRNAs from Archaea, Bacteria and plastids known to date. The 50 'universal' helices are also present in all known eukaryotic SSU rRNAs except in those of Microsporidia (Microspora), where some of these helices are missing. Helices specific to the eukaryotic model are numbered Ea-b, where a is the number of the preceding universal helix and b sequentially numbers all helices inserted between universal helices a and a+1. Helices specific to the prokaryotic model are similarly given composite numbers of the form Pa-b. Mitochondrial sequences show extreme variability in length and in the number of helices present. Examples of secondary structure models for prokaryotic and mitochondrial SSU rRNAs have been given in previous papers on our database (12,20,21).

Table 3. Taxonomic groups (phyla or subphyla; listed alphabetically) and the number of their representatives used for the estimation of nucleotide substitution rates

Kingdom		Number of sequences
Animalia	Acanthocephala	1
	Annelida	3
	Arthropoda	14
	Chaetognatha	1
	Chordata	18
	Cnidaria	4
	Echinodermata	6
	Hemichordata	1
	Mollusca	14
	Nematoda	6
	Phoronida	1
	Placozoa	2
	Platyhelminthes	7
	Porifera	4
	Priapula	1
Fungi		
	Ascomycotina	81
	Basidiomycotina	33
	Ustomycotina	10
	Zygomycotina	13
Plantae		
	Bryophyta	8
	Equisetophyta	2
	Lycopodiophyta	4
	Magnoliophyta	52
	Polypodiophyta	3
	Pinophyta	9
	Psilotophyta	1
Protoctista		
	Apicomplexa	39
	Bacillariophyta	9
	Chlorarachnida (nucleair)	5
	Chlorarachnida (nucleomorph)	5
	Chlorophyta	56
	Chrysophyta	9
	Chytridiomycota	6
	Ciliophora	18
	Choanoflagellata	2
	Cryptophyta	10
	Dinoflagellata	2
	Eustigmatophyta	1
	Glaucocystophyta	1
	Labyrinthulomycota	4
	Oomycota	3
	Phaeophyta	6
	Prymnesiophyta	6
	Rhizopoda	6
	Rhodophyta	13
Total		500

Recently, a new method was developed for measuring the relative substitution rate of individual sites in a nucleotide sequence alignment on the basis of the estimated evolutionary distance between sequences (25,26). By dividing nucleotides into five variability subsets, and giving a different color to each of the subsets, color maps superimposed on the secondary structure of SSU rRNAs can be constructed, as explained in detail in Van de

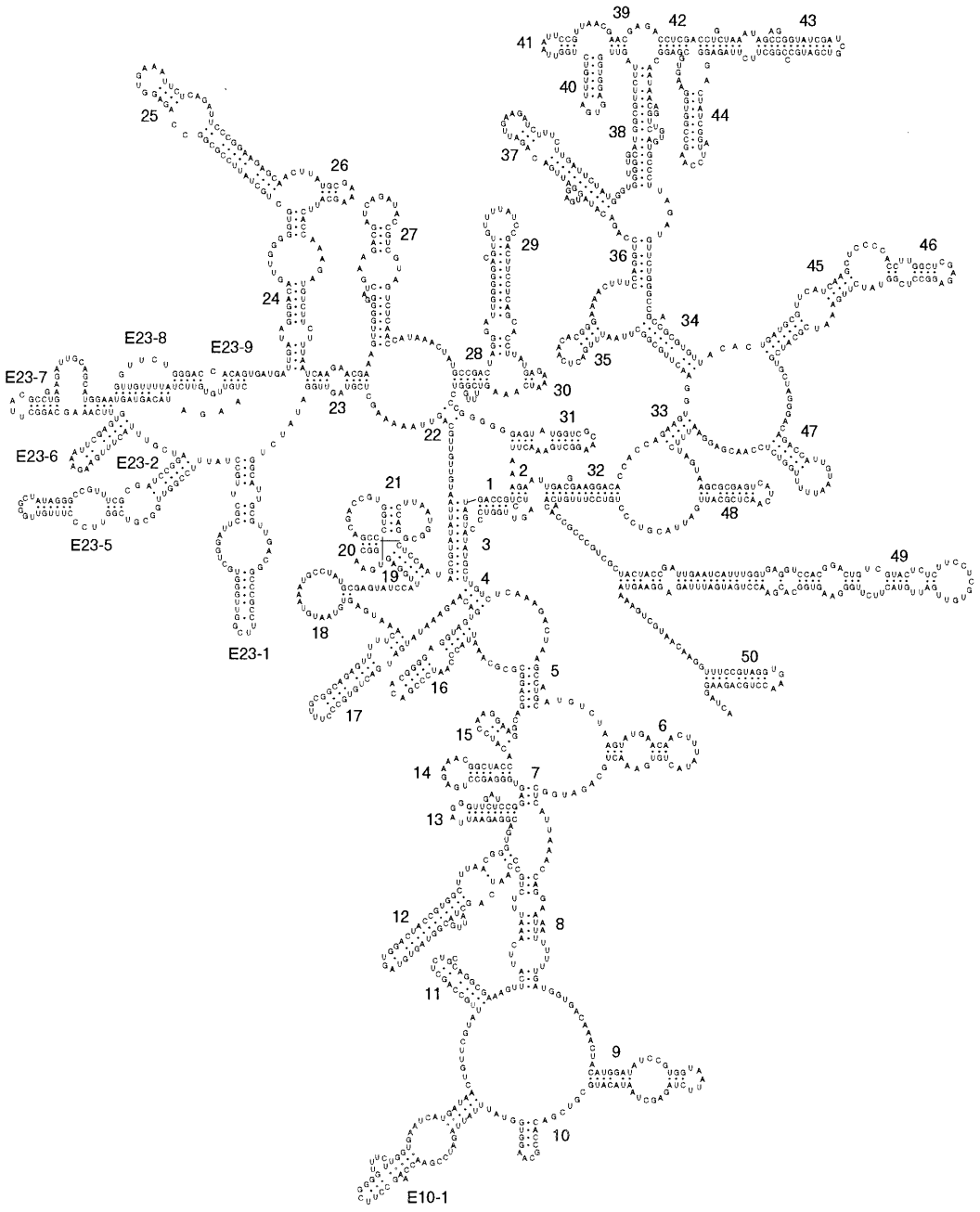

Figure 1. Secondary structure model for the nuclear SSU rRNA of the chlorarachniophyte *Chlorarachnion reptans*. The sequence is written clockwise from 5′ to 3′ terminus.

Peer *et al*. (27). In the latter study, variability color maps are presented for bacterial 5S rRNA, SSU rRNA and large ribosomal subunit rRNA (further abbreviated as LSU rRNA). The color map of Figure 2 of the current paper is superimposed on the eukaryotic SSU rRNA secondary structure model of *Saccharomyces cerevisiae* (19). Nucleotide variabilities were computed on the basis of 500 sequences of species belonging to the eukaryotic 'crown taxa' (28). The list of taxonomic groups and the number of their representatives used for this analysis is given in Table 3. Color maps for bacterial 5S rRNA, SSU rRNA, LSU rRNA (27) and eukaryotic SSU rRNA (Fig. 2) can also be consulted via internet at URL http://bioc-www.uia.ac.be/u/yvdp/ .

AVAILABILITY OF THE DATA

Each SSU rRNA sequence is stored in a separate file, in order to simplify access to the data. Each file contains primary and secondary structure information, as well as annotations such as accession number, literature reference and detailed taxonomic specifications. The SSU rRNA database is made available via

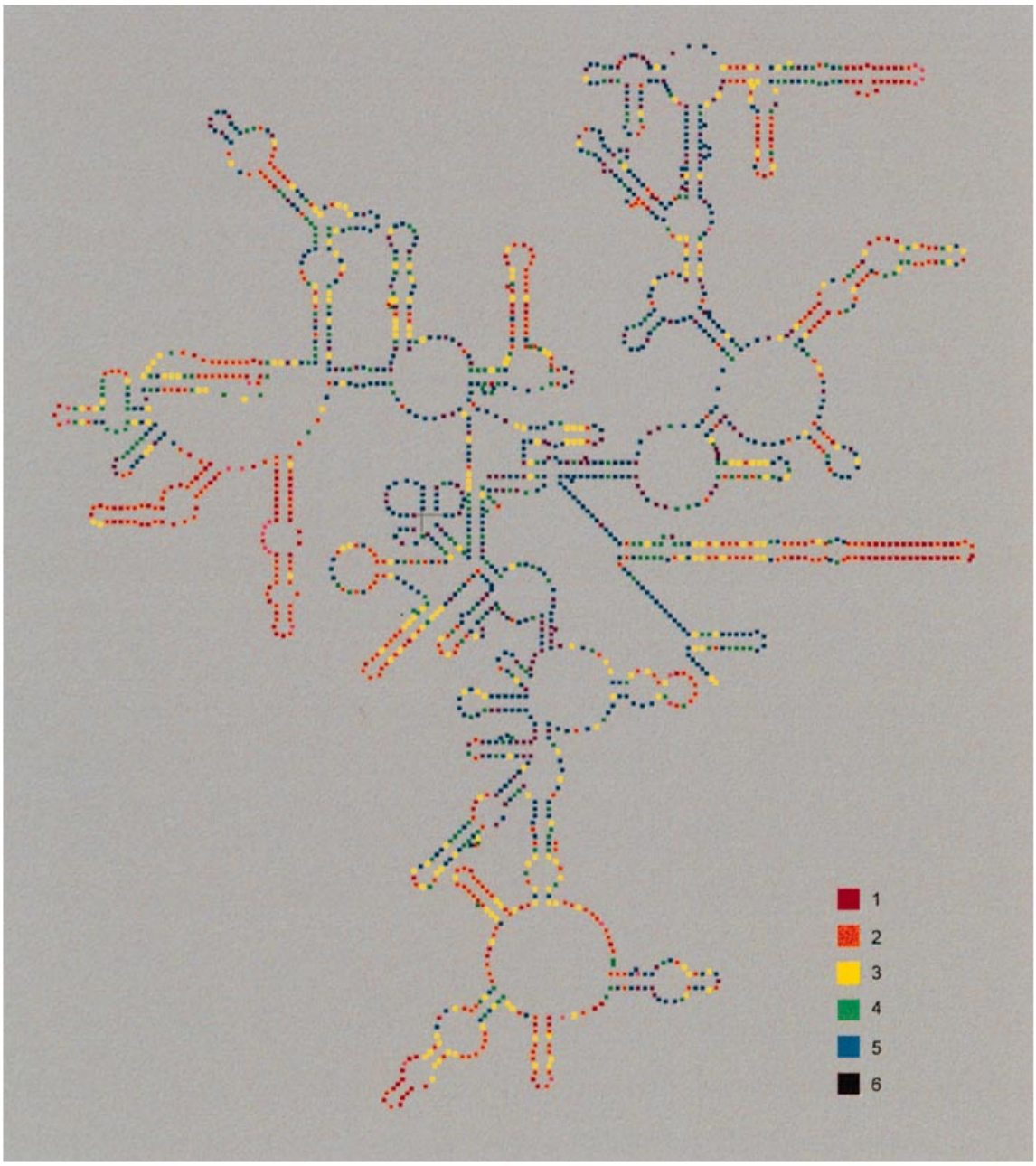

Figure 2. Color map superimposed on the SSU rRNA secondary structure model of *S.cerevisiae* (19). Nucleotides are subdivided into five groups of increasing variability [see text and (27) for details]. The most variable positions are in red, the most conserved in blue. Absolutely conserved positions are indicated in purple. Sites containing a nucleotide in *S.cerevisiae* but vacant in >75% of the aligned sequences were not considered for the variability calculation and are indicated in pink.

World Wide Web at URL http://rrna.uia.ac.be/ssu/ . Through WWW, it is very easy to select sequences either one by one, or by taxonomic group, or by a combination of both. Recently, new search tools were included to easily find specific sequences (see De Rijk *et al*., this issue). Sequences can be retrieved in different formats. On-line information about the database is also available.

If problems occur in connecting to the server or in retrieving data, the authors can be contacted by electronic mail to dwachter@uia.ua.ac.be or yvdp@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 Affairs of the

Belgian State (contract 23), by the National Fund for Scientific Research and by the Special Research Fund of the University. We thank Sabine Chapelle for the computer drawing of the secondary structure model. Yves Van de Peer is Research Assistant of the National Fund for Scientific Research.

REFERENCES

1 Huysmans,E. and De Wachter,R. (1986) *Nucleic Acids Res*. **14**, r73–r118.
2 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates.* Sinauer Associates, Inc., Sunderland.
3 Cronquist,A. (1971) *Introductary Botany.* Harper & Row, New York.
4 Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) *The Fungi: and Advanced Treatise.* Academic Press, New York, Vol. 4A.
5 Moore,R.T. (1988) in Moriarty,C.H. (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 Rodriguez-Tomé,P., Stoehr,P.J., Cameron,G.H. and Flores,T.P. (1996) *Nucleic Acids Res*. **24**, 6–12.
8 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1996) *Nucleic Acids Res.* **24**, 1–5.
9 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
10 Woese,C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
11 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
12 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **20**, 3025–3049.
13 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol*. **17**, 32–38.
14 Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
15 Nelles,L., Fang,B.-L., Volckaert,G., Vandenberghe,A. and De Wachter,R. (1984) *Nucleic Acids Res*. **12**, 8749–8768.
16 Dams,E., Hendriks,L., Van de Peer,Y., Neefs, J.-M., Smits,G., Vandenbempt,I. and De Wachter,R. (1988) *Nucleic Acids Res.* **16**, r87–r173.
17 Neefs,J.-M., Van de Peer,Y., Hendriks,L. and De Wachter,R. (1990) *Nucleic Acids Res*. **18**, 2237–2317.
18 Neefs,J.-M.,Van de Peer,Y., De Rijk,P., Goris,A. and De Wachter,R. (1991) *Nucleic Acids Res.* **19**, 1987–2015.
19 De Rijk,P., Neefs,J.-M.,Van de Peer,Y. and De Wachter,R. (1992) *Nucleic Acids Res*. **20**, 2075–2089.
20 Van de Peer,Y., Van den Broeck,I., De Rijk,P. and De Wachter,R. (1994) *Nucleic Acids Res.* **22**, 3488–3494.
21 Van de Peer,Y., Nicolaï, S., De Rijk,P. and De Wachter,R. (1994) *Nucleic Acids Res.* **24**, 86–91.
22 Neefs,J.-M. and De Wachter,R. (1990) *Nucleic Acids Res*. **18**, 5695–5704.
23 Gutell,R.R., Larsen,N. and Woese, C.R. (1994) *Microbiol. Rev.* **58**, 10–26.
24 Gutell,R.R. (1994) *Nucleic Acids Res*. **22**, 3502–3507.
25 Van de Peer,Y., Neefs,J.-M., De Rijk,P. and De Wachter,R. (1993) *J. Mol. Evol*. **37**, 221–232.
26 Van de Peer,Y., Van der Auwera,G. and De Wachter,R. (1996) *J. Mol. Evol.* **42**, 201–210.
27 Van de Peer,Y., Chapelle,S. and De Wachter,R. (1996) *Nucleic Acids Res.* **24**, 3381–3391.
28 Knoll,A.H. (1992) *Science* **256**, 622–627.