86–91 Nucleic Acids Research, 1996, Vol. 24, No. 1

Database on the structure of small ribosomal subunit RNA

Yves Van de Peer, Stefan Nicolaï, Peter De Rijk 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

The Antwerp database on small ribosomal subunit RNA offers over 4300 nucleotide sequences (August 1995). All these sequences are stored in the form of an alignment based on the adopted secondary structure model, which in turn 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. The complete database is made available to the scientific community through anonymous ftp and World Wide Web (WWW).

CONTENTS OF THE DATABASE

The database on small ribosomal subunit RNA (further abbreviated as SSU rRNA) contained 4331 sequences in August 1995. This number comprises 1035 eukaryotic, 97 archaeal, 2988 bacterial, 64 plastid and 147 mitochondrial sequences. Partial sequences are included only if the combined length of the sequenced segments amounts to ≥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 close relative. All sequences are stored in the form of an alignment and contain the postulated secondary structure pattern in encoded form.

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

Table 2 covers the prokaryotic SSU rRNA sequences. The classification is based on the construction of evolutionary trees. In short, new sequences retrieved from the EMBL (6) and/or GenBank (7) nucleotide sequence libraries are aligned with their presumed closest relative. Evolutionary trees are then constructed by the neighbor-joining method (8), and according to the phylogenetic position observed, the species are assigned to one of the taxa described by Woese and coworkers (9,10) and our research group (11,12). 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. In particular the δ and ε subdivisions are regularly clustered separately from the other Proteobacteria (11,12). Furthermore, the γ subdivision is often found to be paraphyletic (e.g. 10,11), embracing the Proteobacteria β. In previous papers describing the Antwerp rRNA database (11,13), we also distinguished the subdivision γ*, which was formed by species attributed to the Proteobacteria group by Woese and collaborators but separated from the majority of other γ Proteobacteria by the Proteobacteria β. However, since the position of the Proteobacteria β cluster within the γ subdivision is not stable, we no longer discriminate between γ and γ* Proteobacteria, and bacteria previously ascribed to the latter taxon are now placed in the γ subdivision. For the Archaea, a distinction is made between the divisions Crenarchaeota and Euryarchaeota (14). The latter division is further subdivided into 8 subdivisions.

Other databases concerning SSU rRNA structure (15,16) and known mutations in *Escherichia coli* 16S rRNA (17) can be found in the present and the previous database issues of this journal.

SECONDARY STRUCTURE

The secondary structure models adopted for prokaryotic and eukaryotic SSU rRNAs were originally derived (18) by comparison of 6 eucaryal, 1 archaeal, 4 bacterial, 2 plastidial and 1 mitochondrial SSU rRNA sequences available in 1984 and by surveying 13 secondary structure models proposed at the time in papers listed in (18). Gradual improvements were made to the models, as reported in subsequent papers describing our database on SSU rRNA structure (19–23,11,13), taking into account compensating substitutions observed in our sequence alignments (24) and the results of studies by others (reviewed in 25). The model presently followed for bacterial SSU rRNAs is essentially identical to the models made available in graphic form by Gutell (15). It is illustrated in Figure 1 with the SSU rRNA of the Gram positive bacterium *Bacillus subtilis*. The model followed for eukaryotic SSU rRNAs includes a secondary structure pattern in certain variable areas left undefined in the models distributed by

* To whom correspondence should be addressed

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	
Porifera	Calcarea	2	
	Demospongiae	2	
Cnidaria	Anthozoa	2	
	Cubozoa	1	
Ctenophora		2	
Platyhelminthes	Trematoda	13	
	Turbellaria	3	
	Uncertain affiliation	1	
Nematoda	Secernentea	15	
Priapula		1	
Acanthocephala	Archiacanthocephala	1	
Annelida	Polychaeta	1	
Arthropoda	Branchiopoda	2	
	Chelicerata	3	
	Insecta	17	4
	Malacostraca	14	
	Maxillopoda	12	
Pentastomida	Pentastomata	1	
Mollusca	Bivalvia	13	1
	Gastropoda	2	
	Polyplacophora	1	
Phoronida		1	
Ectoprocta	Phylactolaemata	1	
Echinodermata	Echinoidea	24	3
	Asteroidea	1	
	Holothuroidea	1	
	Ophiuroidea	1	
	Crinoidea	1	
Chaetognatha		3	
Hemichordata	Enteropneusta	2	
Chordata	Agnatha	4	1
	Amphibia	18	3
	Aves	2	4
	Chondrichtyes	3	
	Mammalia	8	79
	Osteichtyes	3	8
	Reptilia	4	4
	Cephalochordata (Sub.)	1	
	Urochordata (Subphyl.)	4	
Total:		**193**	**107**

Kingdom Fungi Subphylum	Class	Number of sequences N	M
Zygomycotina	Zygomycetes	19	
Ascomycotina	Discomycetes	13	
	Hemiascomycetes	54	8
	Loculoascomycetes	16	
	Plectomycetes	25	2
	Pyrenomycetes	15	1
	Uncertain affiliation	7	
Basidiomycotina	Heterobasidiomycetes	22	
	Hymenomycetes	15	
	Uncertain affiliation	2	
Ustomycotina	Ustomycetes	11	
Uncertain affiliation		1	
Total:		**200**	**11**

Kingdom Plantae Phylum	Class	Number of sequences N	M	P
Bryophyta	Anthocerotopsida	11		
	Bryopsida	10		
	Marchantiopsida	2	1	1
Lycopodiophyta	Lycopodiopsida	8		
	Isoetopsida	1		
Magnoliophyta	Liliopsida	4	6	2
	Magnoliopsida	71	3	19
Equisetophyta		2		
Polypodiophyta		15		
Pinophyta	Cycadopsida	1		
	Gnetopsida	1		
	Ginkgoopsida	1		
	Pinopsida	7		1
Psilotophyta	Psilotopsida	3		
Total:		**137**	**10**	**23**

Kingdom Protoctista[c] Phylum	Class	Number of sequences N	M	P
Actinopoda	Heliozoa	1		
Apicomplexa	Coccidia	27		
	Hematozoa	41	3	
	Uncertain affiliation	5		
Bacillariophyta	Bacillariophyceae	5		
	Coscinopiscophyceae	4		
Chlorarachnida		7		3
Chlorophyta	Charophyceae	24		2
	Chlorophyceae	61	3	14
	Prasinophyceae	6		
	Ulvophyceae	35		
	Uncertain affiliation	4		
Chrysophyta	Chrysophyceae	9		2
	Dictyochophyceae	1		
	Uncertain affiliation	5		
Chytridiomycota		7		
Oomycota		4		
Ciliophora		52	5	
Conjugaphyta	Conjugatophyceae	8		
Cryptophyta		12		4
Dictyostelida		2		
Dinoflagellata		13		
Euglenida		1		6
Eustigmatophyta	Eustigmatophyceae	1		
Glaucocystophyta	Glaucocystophyceae	1		3
	Uncertain affiliation			1
Granuloreticulosa		5		
Haplosporidia	Haplosporea	4		
Labyrinthulomycota		3		
Microspora		21		
Myxozoa	Myxosporea	1		
Phaeophyta		4	1	1
Plasmodial Slime Molds: Myxomycota		1		
Prymnesiophyta		12		2
Rhizopoda	Filosea	1		
	Lobosea	28	2	
	Uncertain affiliation	1		
Rhodophyta		56	1	3
Xanthophyta		1		
Zoomastigina	Amebomastigota	3		
	Choanomastigotes	2		
	Diplomonadida	9		
	Kinetoplastida	17	4	
Total:		**505**	**19**	**41**

[a]The Metazoan taxa are listed in the same order as they appear in (1).

[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 Protoctist phyla and classes are ordered alphabetically.

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

Bacteria Division	Number of sequences[a]
Chlamydiae	8
Cyanobacteria	36
Fibrobacter	17
Flavobacteria and relatives	156
Fusobacterium and relatives	27
Gram Positives and relatives, Low G+C	858
Gram Positives and relatives, High G+C	494
Green Sulfur	4
Green non sulfur	5
Planctomyces and relatives	8
Proteobacteria α	444
Proteobacteria β	114
Proteobacteria γ	440
Proteobacteria δ	64
Proteobacteria ε	96
Proteobacteria, uncertain affiliation	9
Radioresistant micrococci and relatives	31
Spirochetes	137
Thermotogales	5
Uncertain affiliation[b]	35
Total:	**2988**

Archaea Division	Subdivision	Number of sequences
Euryarchaeota	Archaeoglobales	1
	Halobacteria	22
	Methanobacteriales	17
	Methanococcales	5
	Methanomicrobium group	31
	Methanopyrales	1
	Thermococcales	3
	Thermoplasma	1
Crenarchaeota		16
Total:		**97**

[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.

Gutell (15). It is illustrated in Figure 2 with the SSU rRNA of the dinoflagellate *Alexandrium tamarense.*

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. Helices 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 prokaryotic model are given composite numbers of the form Pa-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 eukaryotic model are similarly numbered Ea-b. Mitochondrial sequences show extreme variability in length and in the number of helices present. Examples of secondary structure models for mitochondrial SSU rRNAs have been given in previous compilations (11,13). Some of these have been subjected to minor changes.

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 through anonymous ftp on the server rrna.uia.ac.be or by World Wide Web at URL http://rrna.uia.ac.be/rrna/ssuform.html. Because of user friendliness, we recommend connecting to the database via WWW. Through WWW, it is very easy to select sequences either one by one, or by taxonomic group, or by a combination of both. Sequences can be retrieved in different formats. On-line information about the database is also available.

For those who choose to connect via ftp, a file called 'readme' is present under the directory 'pub' which contains information on the database contents and on how to obtain SSU rRNA sequences. We suggest to fetch and read this file first before downloading other data. The names of the files on this server are produced from the species name by taking characters of the genus and species names. Their extension is a code describing the phylogenetic group to which the species belongs. This makes it possible to either retrieve specific sequences using the full name, or to retrieve a set of sequences belonging to a phylogenetic group using wild cards. A program is available on the server which allows to create different file formats and to integrate several sequences into an alignment.

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

REFERENCES

1 Brusca,R.C. and Brusca,G.J. (1990) *Invertebrates,* Sinauer Associates, Inc. Sunderland.
2 Cronquist,A. (1971) *Introductary Botany,* Harper & Row, New York.
3 Ainsworth,G.C., Sparrow,F.K. and Sussman,A.S. (1973) *The Fungi: and Advanced Treatise,* Academic Press, New York, Vol. 4A.
4 Moore,R.T. (1988) In Moriarty,CH. (ed.) *Taxonomy putting plants and animals in their place.* Royal Irish Academy, Dublin, pp. 61–88.
5 Margulis,L., Corliss,J.O., Melkonian,M. and Chapman,D.J. (eds) (1990) *Handbook of Protoctista.* Jones and Bartlett Publishers, Boston.
6 Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.H. (1994) *Nucleic Acids Res.* **22**, 3445–3449.

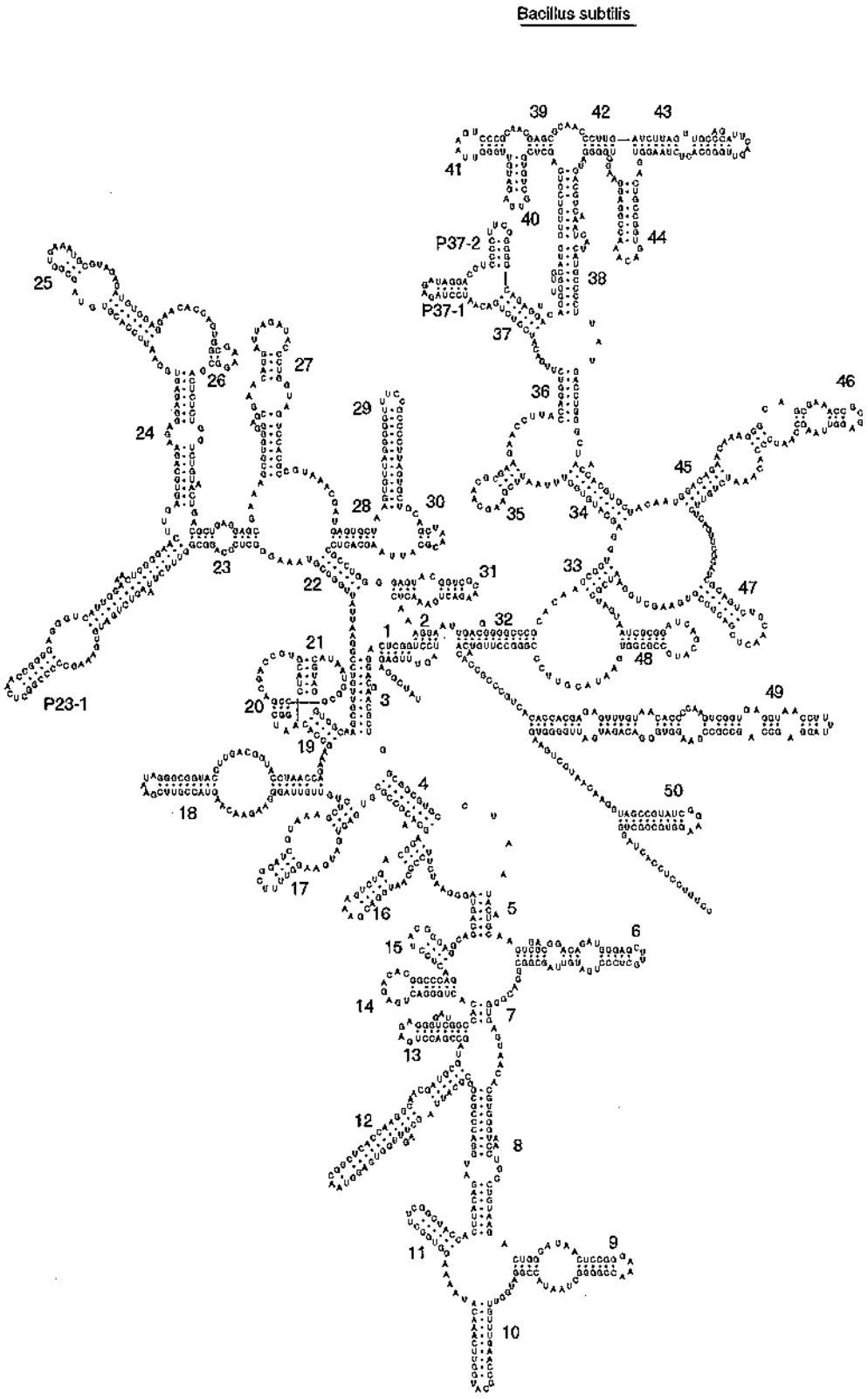

Figure 1. Secondary structure model for SSU rRNA of the gram positive bacterium *Bacillus subtilis*. The sequence is written clockwise from 5′ to 3′ terminus.

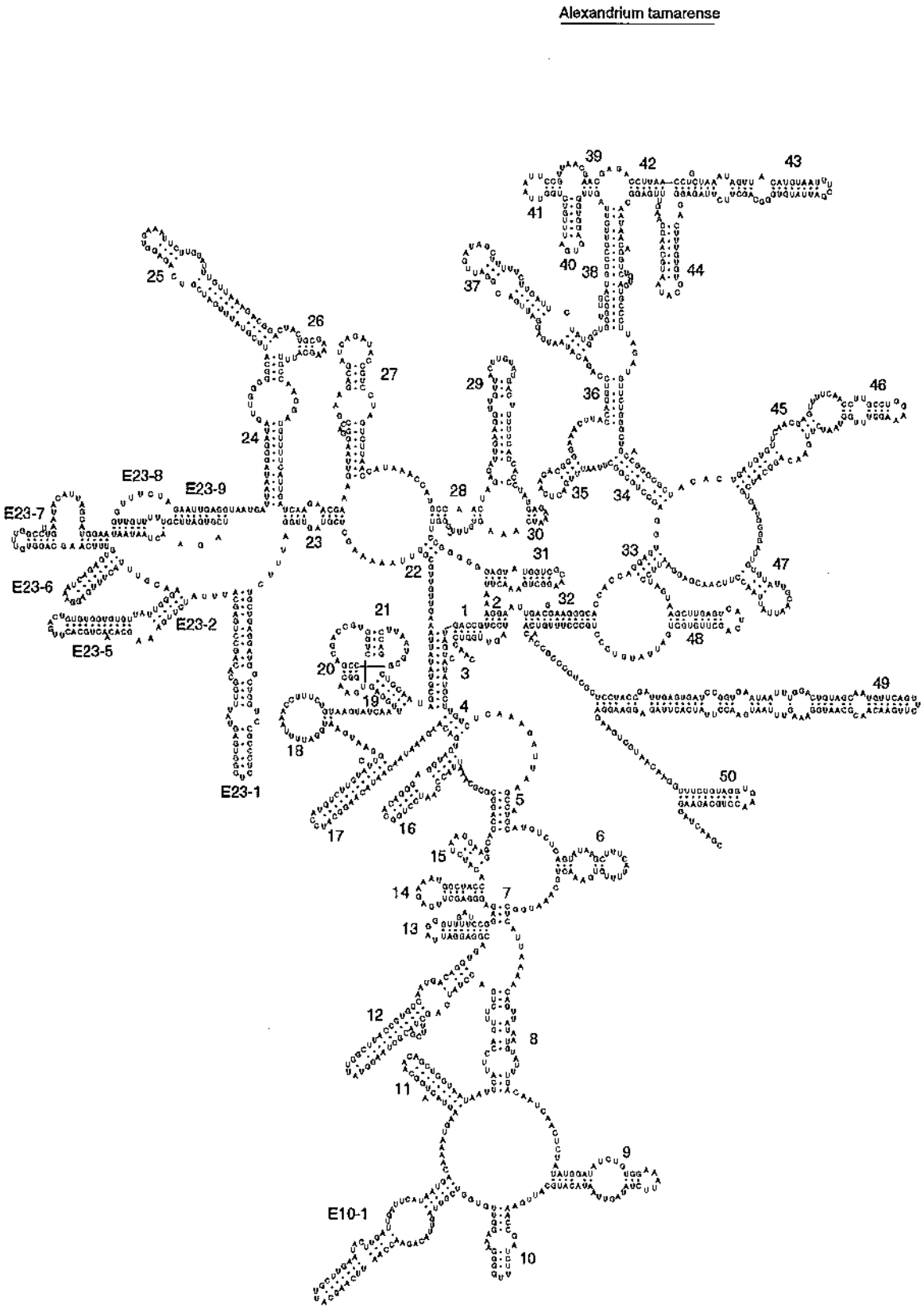

Figure 2. Secondary structure model for SSU rRNA of the dinoflagellate *Alexandrium tamarense.* The sequence is written clockwise from 5′ to 3′ terminus.

7 Benson,D.A., Boguski,M., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res.* **22**, 3441–3444.
8 Saitou,N. and Nei,M. (1978) *Mol. Biol. Evol.* **4**, 406–425.
9 Woese,C.R. (1987) *Microbiol. Rev.* **51**, 221–271.
10 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.* **176**, 1–6.
11 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.* **20**, 3025–3049.
12 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.* **17**, 32–38.
13 Van de Peer,Y., Van den Broeck,I., De Rijk,P. and De Wachter,R. (1994) *Nucleic Acids Res.* **22**, 3488–3494.
14 Olsen,G.J. and Woese,C.R. (1993) *FASEB J.* **7**, 113–123.
15 Gutell, R.R. (1994) *Nucleic Acids Res.* **22**, 3502–3507.
16 Maidak, B.L., Olsen,G.J., Larsen, N., Overbeek, R., McCaughey, M.J. and Woese, C.R. (1996) *Nucleic Acids Res.* **24**, 82–85.
17 Triman, K.L. (1996) this issue.
18 Nelles, L., Fang, B.-L., Volckaert, G., Vandenberghe, A. and De Wachter, R. (1984) *Nucleic Acids Res.* **12**, 8749–8768.
19 Huysmans, E. and De Wachter, R. (1986) *Nucleic Acids Res.* **14**, r73–r118.
20 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.
21 Neefs, J.-M., Van de Peer, Y., Hendriks, L. and De Wachter, R. (1990) *Nucleic Acids Res.* **18**, 2237–2317.
22 Neefs, J.-M.,Van de Peer, Y., De Rijk, P., Goris, A. and De Wachter, R. (1991) *Nucleic Acids Res.* **19**, 1987–2015.
23 De Rijk, P., Neefs, J.-M., Van de Peer, Y. and De Wachter, R. (1992) *Nucleic Acids Res.* **20**, 2075–2089.
24 Neefs, J.-M. and De Wachter, R. (1990) *Nucleic Acids Res.***18**, 5695–5704.
25 Gutell,R.R., Larsen, N. & Woese, C.R. (1994) *Microbiol. Rev.* **58**, 10–26.