Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA

Jan Wuyts, Yves Van de Peer[1] and Rupert De Wachter*

Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium and [1]Departement Plantengenetica, Vlaams Interuniversitair Instituut voor Biotechnologie (VIB), Universiteit Gent, Ledeganckstraat 35, B-9000 Gent, Belgium

Received September 3, 2001; Revised and Accepted November 5, 2001

ABSTRACT

The relative substitution rate of each nucleotide site in bacterial small subunit rRNA, large subunit rRNA and 5S rRNA was calculated from sequence alignments for each molecule. Two-dimensional and three-dimensional variability maps of the rRNAs were obtained by plotting the substitution rates on secondary structure models and on the tertiary structure of the rRNAs available from X-ray diffraction results. This showed that the substitution rates are generally low near the centre of the ribosome, where the nucleotides essential for its function are situated, and that they increase towards the surface. An inventory was made of insertions characteristic of the Archaea, Bacteria and Eucarya domains, and for additional insertions present in specific eukaryotic taxa. All these insertions occur at the ribosome surface. The taxon-specific insertions seem to arise randomly in the eukaryotic evolutionary tree, without any phylogenetic relatedness between the taxa possessing them.

INTRODUCTION

X-ray diffraction analysis of crystals of the small ribosomal subunit (SSU) of the bacterium *Thermus thermophilus* (1) and of the large ribosomal subunit (LSU) of the archaebacterium *Haloarcula marismortui* (2) has yielded the tertiary structures of the subunits, including those of the nearly complete ribosomal RNAs, at atomic resolution. Moreover, the complete spatial structure of the 70S ribosome of *T.thermophilus* has been determined with sufficient resolution for the coordinates of all P-atoms of the constituent RNA molecules to be known (3). On the other hand, the secondary structures of the ribosomal RNAs have been investigated for >20 years. Experimental methods played an important role in the first attempts to establish base pairing patterns. However, the currently available detailed secondary structure models for 5S rRNA (4), SSU rRNA (5,6) and LSU rRNA (7–9) were derived essentially by comparative sequence analysis and observation of compensating substitutions in sequence alignments that increased in size as ever more primary structures were published. It has been rewarding to see the predicted models, and hence the validity of the method for deriving them, confirmed by the experimental X-ray diffraction results.

Alignments of SSU and LSU rRNA sequences show alternation of conserved and variable areas. A method has been devised (10,11) for the quantitative measurement of the nucleotide substitution rate of most of the nucleotide sites in each rRNA, relative to the average substitution rate of the entire molecule, using large alignments containing hundreds of sequences. The availability of the tertiary structure of SSU and LSU rRNAs prompted us to superimpose measurements of the relative substitution rate for each site onto its spatial coordinates in order to obtain a three-dimensional variability map.

The majority of the SSU and LSU rRNA secondary structure elements are common to prokaryotic and eukaryotic ribosomes. A small number of helices are Bacteria specific. Among the mitochondrial SSU and LSU rRNAs, many are

Table 1. rRNA sequence data set

Molecule	Number of aligned sequences used[a]			
	Bacteria	Archaea	Eucarya	
			Complete	Complete + partial
SSU rRNA	3407	643	6540	7403
LSU rRNA[b]	184	37	154	278
5S rRNA	310	–	–	–

[a]The number of available sequences can be much larger than the number of species, especially for Bacteria where several strains of a species are often analysed, and to a lesser extent for Eucarya, e.g. because of the presence of multiple genes. For the Bacteria domain the numbers mentioned are those used for substitution rate calibration and correspond to one sequence per species. The Archaea and Eucarya SSU and LSU rRNA sequences were only used for secondary structure analysis and all available sequences are counted. No 5S rRNA sequences were needed for this purpose since its secondary structure is well known. In the case of Eucarya partial SSU and LSU rRNA sequences were included to study the secondary structure of certain areas.
[b]This comprises 5.8S rRNA plus 28S rRNA in the majority of eukaryotes. In a minority it consists of other processing products of the RNA polymerase I primary transcript destined for the large ribosomal subunit, as reviewed in Clark (13).

*To whom correspondence should be addressed. Tel: +32 3 8202319; Fax: +32 3 8202248; Email: dwachter@uia.ua.ac.be

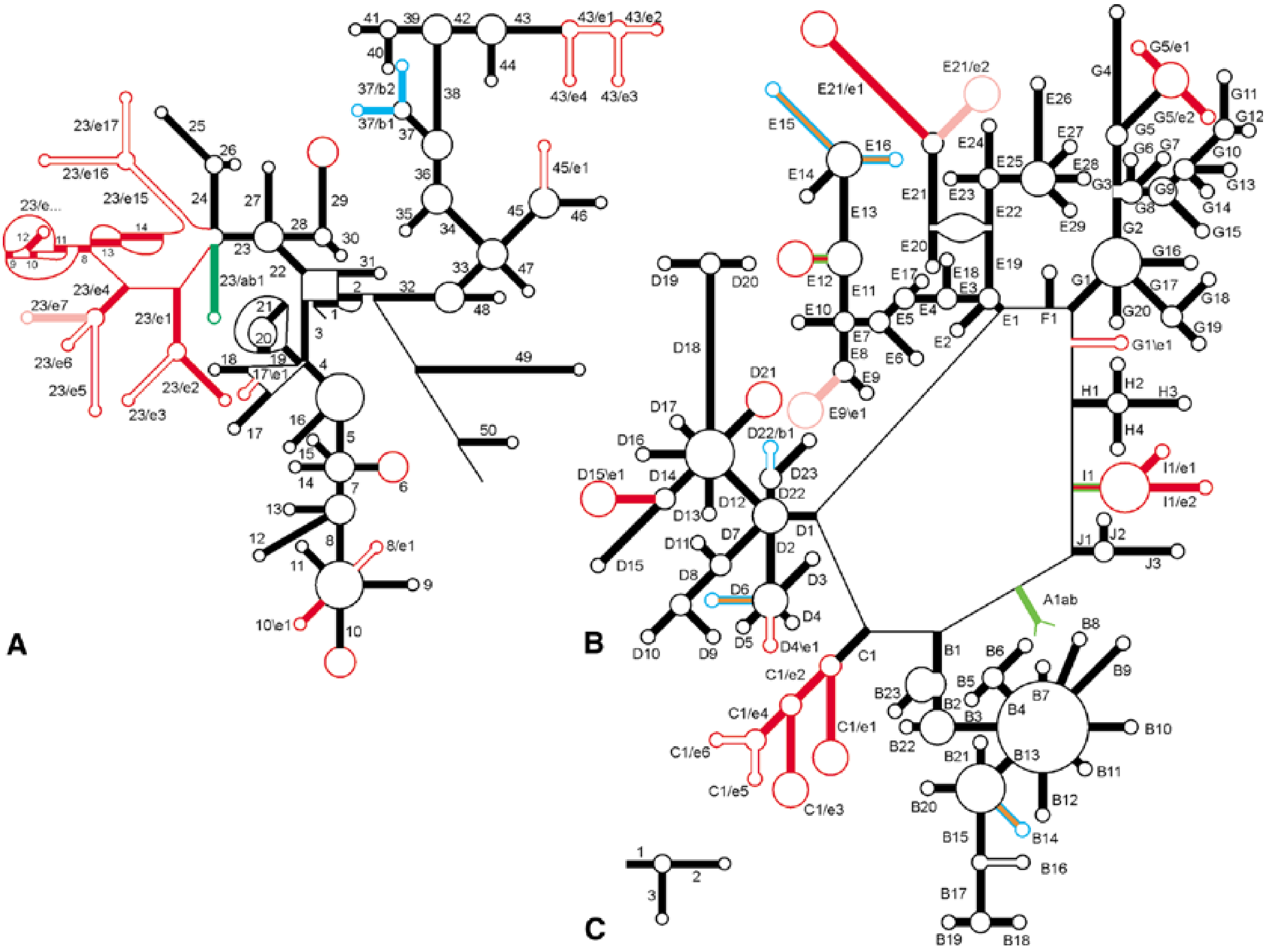


Figure 1. Secondary structure and helix numbering of SSU rRNA (**A**), LSU rRNA (**B**) and 5S rRNA (**C**). All chains run clockwise from the 5′- to 3′-terminus. The presence of helices in the three domains is indicated as follows (compare with Table 2). Bacteria only, blue; Eucarya only, red or pink; Bacteria + Archaea, green; Eucarya + Archaea, orange; all three domains, black. Solid-coloured helices occur in all species of the domain(s), those shown in outline occur only in a subset. Large red loops in hairpins such as 10 in SSU rRNA and C1/e1 and C1/e3 in LSU rRNA indicate that the loop sequence may form an additional, as yet unknown, structure, for example consisting of further branching, in certain eukaryotic species.

reduced in size with respect to the Bacteria, although insertions are present in those from certain taxa such as plants (12). The most substantial insertions occur in nuclear rRNAs of Eucarya, most of which have larger SSU and LSU rRNAs than the Bacteria and the Archaea. In addition to the insertions common to most Eucarya, extra species- or taxon-specific insertions of considerable length were discovered in variable areas of the rRNAs (see 13 for early references). Their structures, which to a certain degree can also be predicted by comparative analysis, often consist of additional hairpins originating at potential branching points in the secondary structure common to the Eucarya.

Sequence alignments and secondary structure models tell us where the Eucarya-specific helices, as well as the supernumerary helices found in certain taxa or species, are attached to the core structures. The tertiary structures of these additional helices are not yet known. However, assuming that the common cores of SSU and LSU rRNAs have a similar tertiary structure in species belonging to the three domains of the living world, it is possible to determine the spatial location of the attachment points for the eukaryotic insertions in the known bacterial tertiary structures.

MATERIALS AND METHODS

Data on rRNA structure

The number of aligned rRNA sequences used for derivation of the secondary structure models and estimation of the relative substitution rate of each nucleotide site is given in Table 1. SSU and LSU rRNA sequences are kept in our database of aligned sequences, the contents of which are regularly published (6,9). 5S rRNA sequences were obtained from the database of Szymanski *et al.* (14) with some additions from the EMBL nucleotide sequence database (15). Taxonomic classification of species is according to the NCBI taxonomy homepage at http://www.ncbi.nlm.nih.gov/Taxonomy.

Spatial coordinates of the P-atoms of SSU rRNA, LSU rRNA and 5S rRNA of *T.thermophilus* (3) were obtained from

Table 2. Helix nomenclature and colour codes used in Figure 1

Helix present in			Symbol used in Figure 1	Nomenclature[a]
Bacteria	Archaea	Eucarya[b]		
All	All	All	Solid black	N
Subset	Subset	Subset	Black outline	N
Subset	All	All	Solid orange with blue outlining[c]	N
Subset	Subset	All	Solid red with green outlining[c]	N
All	–	–	Solid blue	N/bn or N\bn
Subset	–	–	Blue outline	N/bn or N\bn
–	–	All	Solid red (exceptionally pink[d])	N/en or N\en
–	–	Subset	Red outline	N/en or N\en
All	All	–	Solid green	N/abn or N\abn

[a]N is a number in the case of SSU rRNA, a letter–number combination in LSU rRNA, and it designates a helix found in species of the three domains. N/bn designates a Bacteria-specific helix situated between the 5′- and the 3′-strand of helix N, where n is a sequential number if there are several helices between N and N+1. N\bn means that the Bacteria-specific helix follows the 3′-strand of helix N. Analogous nomenclature is used for helices specific to the Archaea (a), Eucarya (e) and combinations of domains (ab). The first helix in LSU rRNA is numbered A1ab.

[b]This applies to all Eucarya except the protist taxa Microsporidia, Parabasalidea and Diplomonadida, which have a reduced structure (see Table 5 for details).

[c]A helix numbered N is not coloured black if it occurs in all species of one or two domains and a subset of the remaining domain(s). Examples in LSU rRNA are helix B14, which occurs in all Archaea and Eucarya but in a subset of the Bacteria, and helix E12, which occurs in all Eucarya but subsets of the Bacteria and Archaea.

[d]Helix 23/e7 in SSU rRNA is absent in certain Apicomplexa, helices E9\e1 and E21/e2 in LSU rRNA are absent in red algae. These three helices are coloured solid pink rather than red.

the Protein Data Bank under accession numbers 1GIX and 1GIY.

SSU and LSU rRNA secondary structure models and helix numbering system

The secondary structure models followed for prokaryotic SSU and LSU rRNA and for 5S rRNA (6,9,16), for eukaryotic SSU rRNA (17) and for eukaryotic LSU rRNA (9) have been published previously. The models are represented in schematic form in Figure 1. In certain eukaryotic taxa, SSU and/or LSU rRNA contain insertions for which the secondary structure is not yet fully known. Examination of sequences of additional species allowed us to improve some details of the secondary structure by searching for compensating substitutions in the local alignment, using methods published previously (17). The improvements concern helices 43/e1 to 43/e4 in SSU rRNA and helices C1/e1 to C1/e6 in LSU rRNA.

Slight adjustments to the helix numbering system were made in order to eliminate potential ambiguities and to bring in line the numbering systems used for SSU rRNA and LSU rRNA. The general principle of the numbering system used in the database issues on rRNA structure published in this journal (9,18 and references cited therein) is maintained. It consists of numbering helices in their order of occurrence from the 5′- to the 3′-terminus and changing the number on each passage of a multibranched loop, a pseudoknot loop or a single-stranded area that does not form a loop. Helical segments separated only by internal (or bulge) loops are considered to belong to the same helix.

A distinction is made between two categories of helices. The first category, formerly called universal helices (18), are those encountered in rRNAs of species belonging to the three domains Bacteria, Archaea and Eucarya, but not necessarily in mitochondrial SSU rRNAs, which have a reduced structure in several eukaryotic kingdoms. They are given a single number or letter–number combination. In the case of SSU rRNA these helices are numbered from 1 to 50 and they occur in all species of the three domains. In LSU rRNA they are given a letter–number combination because the secondary structure consists of a large loop (see Figs 1B and 2B) from which depart helices A1 to J1, most of which form the basis of a branched structure. As an example, G1 branches into helices G2 to G20. The LSU rRNA helices of the first category are not necessarily present in all species of the three domains, some are missing in a fraction of them. As an example, helix B14 is present in all hitherto examined Archaea and Eucarya, but absent in Proteobacteria α, β, γ and δ subdivisions. A more extreme example is helix B16 which, although encountered in the three domains, is present in only a fraction of the species of each of them.

The second category of helices consists of those encountered in only one or two of the three domains. They are given two numbers separated by a slash or a backslash. Examples are helices 37/b1 and 37/b2 in SSU rRNA (Figs 1A and 2A). The slash indicates that the helices follow the 5′-strand, but precede the 3′-strand of helix 37 in the primary structure. The notation b indicates that the helices are found only in the Bacteria domain, notations a and e being used for the Archaea and Eucarya domains, respectively. A different example is helix G1\e1 found in some eukaryotic LSU rRNAs (Fig. 1B). The backslash indicates that the helix follows the 3′-strand of helix G1 in the primary structure. A hypothetical helix G1/e1 on the contrary would originate in the multibranched loop terminating helix G1 and would precede helix G2. This notation eliminates an ambiguity in the helix numbering used previously (9).

A special case is the helix joining the two termini of LSU rRNA, which was named A1ab because it is the first helix counting from the 5′-terminus but it is found only in Archaea and Bacteria. A synopsis of the helix numbering and the colour coding used in Figure 1 appears in Table 2.

Substitution rate calibration of bacterial rRNAs

The substitution rate or variability (v) of each nucleotide site in an rRNA alignment, relative to the average substitution rate of all sites, was estimated by substitution rate calibration (10,11). The computations were carried out by means of the programme Treecon for Windows (19), available at http://rrna.uia.ac.be/.

Ideally, the computations should be performed on the same set of species for SSU, LSU and 5S rRNA. The relative substitution rate of each site could then be measured against the average substitution rate of all the sites of the three rRNAs. However, the complete set of three rRNA sequences is available for only 84 bacterial species. If such a small set is used, a rather large fraction of the sites appear as invariant ($v = 0$) because they happen to contain the same nucleotide in all available sequences. The larger the sequence set used, the smaller becomes the fraction of invariant sites because of the

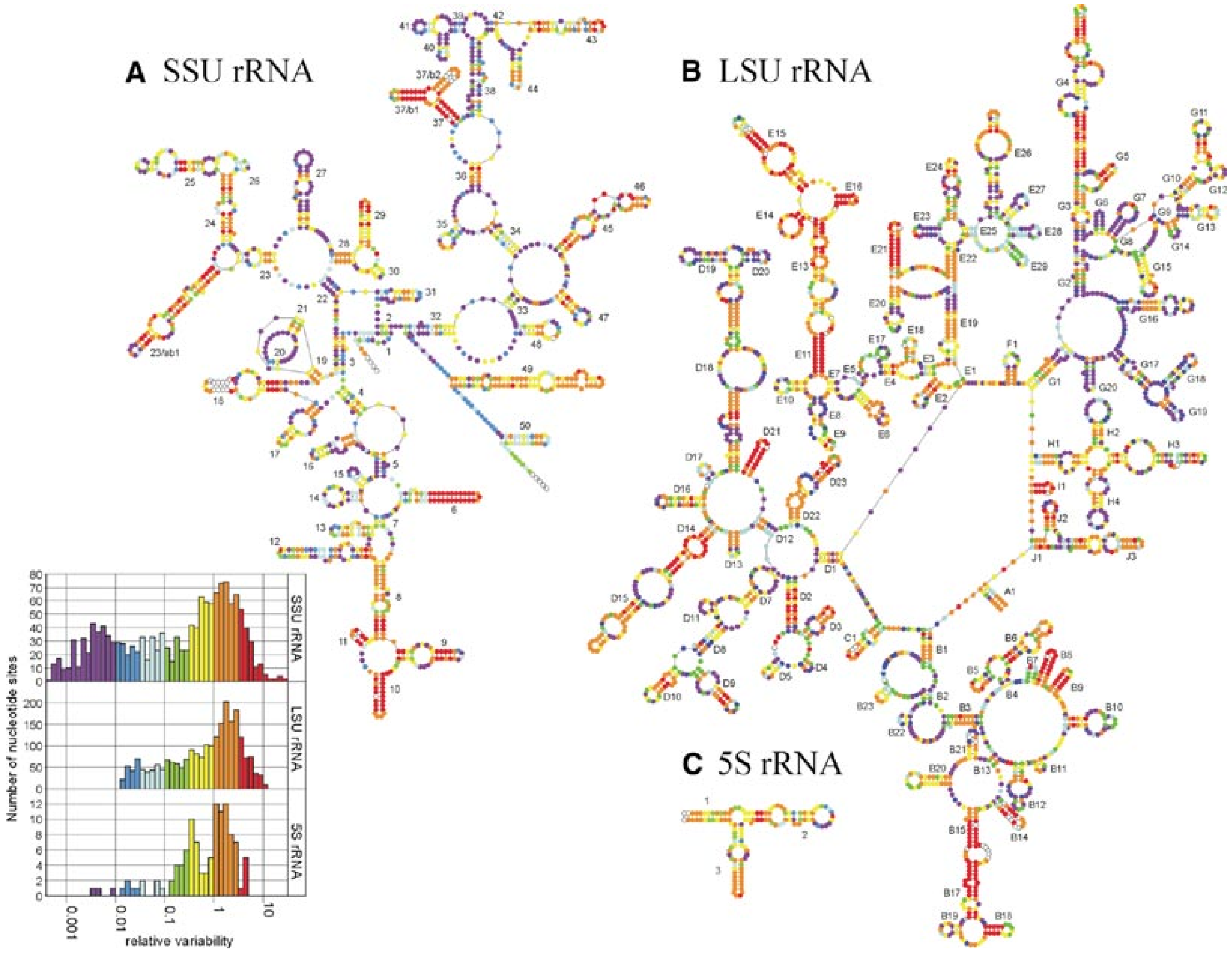


Figure 2. Variability maps of bacterial SSU rRNA (**A**), LSU rRNA (**B**) and 5S rRNA (**C**) superposed on the secondary structure models of the *T.thermophilus* molecules. Sites are subdivided into seven groups according to their relative substitution rate, coloured purple (lowest rate) to red (highest rate). The rate was not measured for sites occupied in <25% of the sequence alignment, which are shown as hollow dots. The histograms inset to (A) are the substitution rate spectra for the three molecules.

increasing probability that the set contains sufficiently divergent species to show the actual variation existing at the site. For this reason, and also because the accuracy of the variability measurements increases with the size of the sequence set, the substitution rates were measured for each molecule separately, on the complete set of sequences available for that molecule (Table 1, left column), and relative to the average substitution rate of that molecule.

Stereo models

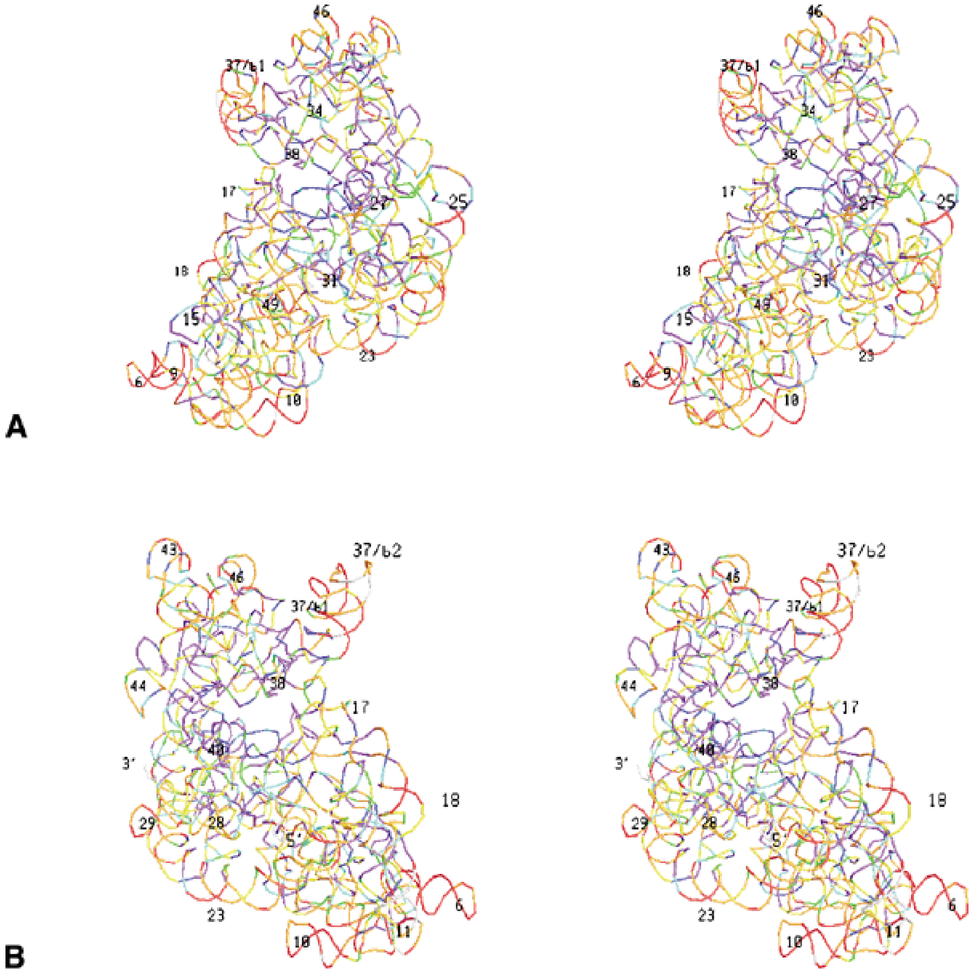
Using the openGL graphics library (20), a programme was written to combine the data on the P-atom coordinates and the nucleotide variabilities estimated by substitution rate calibration. This programme makes a three-dimensional representation of the RNA molecules by connecting the locations of the P-atoms with colour-coded cylindrical rods. Stereo models are obtained by projecting this three-dimensional model onto a plane from two suitable viewpoints. The programme is currently in an early developmental phase but may be released in the future.

RESULTS

Substitution rates plotted upon the secondary structure models of bacterial rRNAs

Figure 2 shows the relative substitution rate of each nucleotide site superimposed by means of a colour code on the secondary structure models for SSU rRNA, LSU rRNA and 5S rRNA of *T.thermophilus*. Similar 'variability maps' for the three bacterial rRNAs have been published previously (16) but they were based on much smaller numbers of sequences than those now available (Table 1) and hence are less accurate.

The inset to Figure 2 shows the distribution of substitution rates for each molecule. In each case, the maximum of the distribution falls at an approximate relative rate of 1.8 times that of the entire molecule. The same value for the maxima was found when the substitution rates were measured relative to the average rate for the three combined molecules, on the smaller set of 84 species for which all sequences are available. This means that the substitution rate measurements for the sites of the three molecules are comparable, e.g. yellow sites in SSU



rRNA possess approximately the same range of substitution rates as yellow sites in LSU rRNA and 5S rRNA. The 'tail' of purple sites in the distribution, which are those with a low rate, is longest in SSU rRNA because many more sequences were processed for this molecule than for the two others (see Table 1). The more sequences are known, the less the probability that a conservative site contains the same nucleotide in all sequences, in other words that it appears to be completely invariant. Hence, as more sequences become available, sites that were invisible on the graph because they had an apparent rate $v = 0$ shift to positions with a low but measurable rate at the left end of the spectrum.

Spatial distribution of substitution rates

In Figure 3 the colour-coded substitution rates are superimposed on stereo views of the tertiary structure of the *T.thermophilus* rRNAs in each subunit. Each site is symbolized by a small coloured bar connecting the coordinates of the two adjoining P-atoms. Due to the absence of a 3′-terminal phosphate the last nucleotide is not represented, but in SSU rRNA the substitution rates of some five nucleotides at both termini are unknown anyway because many sequences are not determined up to the termini.

It is apparent, especially when viewing the large subunit from the side of the interface with the small subunit, that sites close to the ribosome centre tend to be more conserved whereas peripheric sites tend to have higher substitution rates. This tendency is demonstrated more quantitatively by the graphs in Figure 4, where the average substitution rate of sets of sites is plotted as a function of their distance from the centre of the ribosome.

Taxon-specific helices and their sites of attachment to the core structures

In the schematic secondary structure models in Figure 1, one can distinguish a core structure of helices encountered in all ribosomes except those of mitochondria. These helices are shown in black, orange with blue outlining, or red with green outlining (compare with Table 2). They are encountered in species belonging to each of the three domains (Bacteria, Archaea, Eucarya), although not necessarily in all species of these domains. This definition of a core structure, although somewhat arbitrary, was chosen because a helix hitherto found in all species of the three domains may be found to be missing in some molecule in the future, which would require complete helix renumbering if one defined the core structure as consisting only of helices present in all species. The definition is least satisfactory for LSU rRNA where only a few hundred sequences are currently available, and where certain helices appear only sporadically in each of the three domains. In the

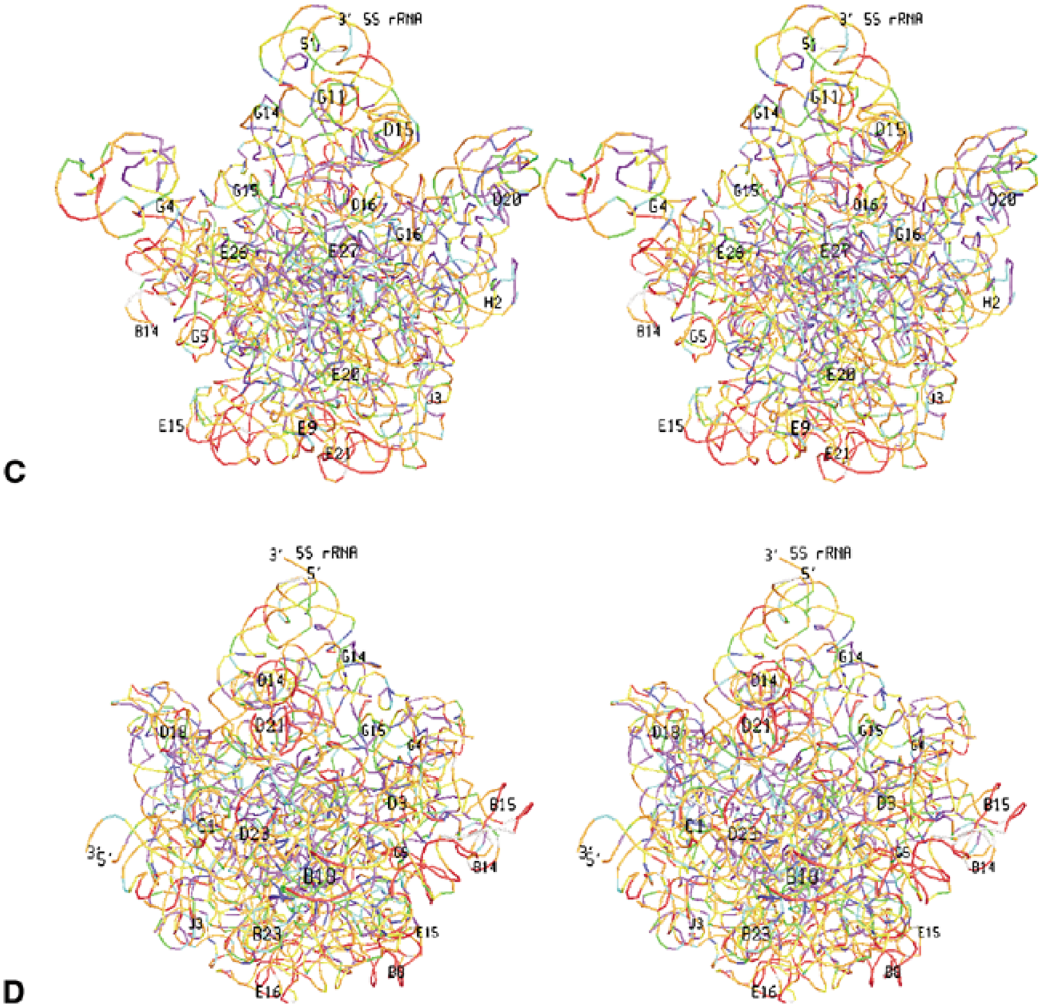


Figure 3. (Previous page and above) Variability maps superimposed on the tertiary structure of the RNAs in the *T.thermophilus* SSU and LSU. A stereo drawing of each subunit is shown from the solvent side and from the side of the interface with the other subunit. Each nucleotide is represented by a coloured bar connecting the coordinates of the two adjoining P-atoms. Colours for substitution rate intervals are as in Figure 2. The most easily recognisable helices are numbered, and the 5′- and 3′-termini are indicated. (**A**) Small subunit from interface side; (**B**) Small subunit from solvent side; (**C**) large subunit from interface side; (**D**) large subunit from solvent side.

case of SSU rRNA, on the other hand, a common core of helices is shared by thousands of Bacteria, Archaea and Eucarya. However, the concept of a core structure serves a practical purpose, namely to facilitate the description of exceptional structures resulting from the presence of extra helices or the absence of core helices.

In addition to the common core, one can also consider a eukaryotic core, consisting of the solid black helices, Eucarya-specific red helices and orange helices common to Eucarya and Archaea (Fig. 1 and Table 2). Three helices of the eukaryotic core are coloured pink in Figure 1 because each of them is absent in species of a single taxon (see Table 2, footnote d). The Bacteria core is formed by solid black helices, Bacteria-specific blue helices, and green helices common to Archaea and Bacteria. No Archaea-specific helices have been found to date. Figure 1 shows that among the domain-specific helices, those of the Eucarya are most numerous.

The Eucarya also show the largest variation in size of their SSU and LSU rRNAs, resulting from the presence of taxon-specific insertions, which are drawn as hollow red hairpins or large red hairpin loops in Figure 1. Table 3 summarises the situation in eukaryotic SSU rRNA and Table 4 that in eukaryotic LSU rRNA. Neither of these tables contain data on three protist taxa, the Microsporidia, Parabasalidea and Diplomonadida, which are treated separately in Table 5. This is because the latter taxa share the absence of several helices found in all other eukaryotes; in other words, they distinguish themselves by a reduction in the eukaryotic core structure (solid black, red, pink and orange in Fig. 1) rather than by the presence of supernumerary helices (red outline in Fig. 1). These protists are intracellular parasites devoid of mitochondria. They have been named Archezoa (21) because they were considered to have diverged from other Eucarya before the origin of mitochondria by endosymbiosis. However, this view has been challenged because genes of mitochondrial origin were found in the nuclear genome of these protists (22–24), pointing to the original presence and subsequent loss of mitochondria.

The supernumerary helices present in a limited number of eukaryotic taxa (red outline in Fig. 1) occur at potential

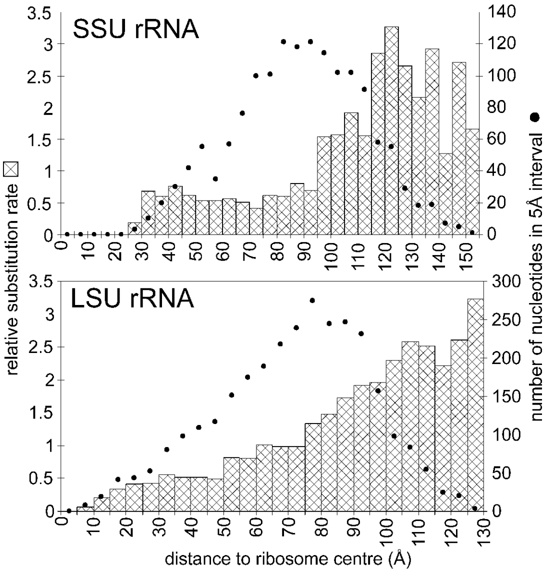


Figure 4. Substitution rate of nucleotide sites in SSU rRNA and LSU rRNA as a function of their distance from the ribosome centre. The ribosome centre is defined as the point showing the smallest sum of distances to all P-atoms of the ribosome. Nucleotide sites are identified with the 5′-P-atoms. The bars represent the average relative substitution rate for sets of sites comprised in spherical shells of 5 Å thickness, plotted with 5 Å increments. The dots represent the number of nucleotide sites in each shell.

branching points in the core structure (e.g. 23/e3 and 43/e1 to 43/e4). Variability maps of SSU rRNA and LSU rRNA from Bacteria (Fig. 2) as well as Eucarya (25,26) show that many of the potential branching points, whether situated in the common core or the Eucarya core, have a high substitution rate. This is also the case for most of the sites where Eucarya-specific and Bacteria-specific core helices (solid red, blue and green in Fig. 1) originate. As for the supernumerary helices (outlines in Fig. 1) themselves, their substitution rate can rarely be measured quantitatively because they occur in too few sequences, but visual inspection of their alignment shows extreme variability.

Apart from the supernumerary helices, the rRNAs, especially those of Eucarya, contain other hot spots for insertion, which do not result in a new branching but only in a lengthening of existing hairpins. Although a visual inspection of the sequence alignments shows that certain hairpins are more variable in length than others, no systematic attempt was made to indicate them in Figure 1 or to list them as insertion hot spots in Tables 3 and 4. The reason is that there is a continuum of hairpin length variability from near constancy to more than doubling in length. Any choice of a cut-off value would lead to an arbitrary set of variable length hairpins that would change with the composition of the available sequence alignment. There are a few sites in both molecules where certain eukaryotic species contain insertions for which the secondary structure is not yet clear. Many uncertainties consist of insertions in hairpin loops in the standard model, which may contain additional ramifications in these species. The sites of unknown structure are mentioned in Table 3 for SSU rRNA and in Table 4 for LSU rRNA. No attempt was made to include in the analysis a few sequences with up to twice the length of the average eukaryotic SSU rRNA (see, for example, 27,28), since the comparative method of deriving the secondary structure is hardly applicable to the exceptionally long insertions present in these sequences.

The spatial location of the Bacteria-specific, Eucarya-specific and Eucarya taxon-specific (supernumerary) helices is illustrated in Figure 5. The coordinates of the Bacteria-specific helices are known. Nothing is known about the shape of Eucarya-specific and supernumerary helices but we do know the coordinates of the nucleotide sites in *T.thermophilus* rRNAs homologous to the sites that serve as attachment (or branching point) for these helices in the eukaryotic rRNAs. Figure 5 shows stereo drawings where Bacteria-specific helices, as well as the loops that form extension and branching points for Eucarya-specific structures, are marked in colour. All these sites prove to be situated at the periphery and at the solvent side of each subunit. In the case of SSU rRNA, the sites where group I introns have been found (29) are also indicated for comparison. In contrast to the insertion sites, the intron sites are not found at the surface but rather in the interior of the subunit.

DISCUSSION

Three main conclusions can be drawn from this study. First, the most conserved nucleotide sites are near the ribosome centre, and variability increases towards the surface. Secondly, the structures specific for Bacteria and Archaea are situated at the periphery of the ribosome, and although the tertiary structure of the Eucarya-specific structures remains unknown, their points of attachment to the core structure are also at the periphery. Thirdly, the additional insertions characteristic of certain eukaryotic taxa always occur at the same sites, but they seem to arise haphazardly throughout the Eucarya domain without any phylogenetic link between the taxa possessing the insertions. The latter point is illustrated below, after a short comment on the former two.

The conservation of centrally located sites is easily explained in view of the fact that the functional sites of the ribosome, namely tRNA binding, mRNA decoding and peptidyl transfer, are located at the interface between the two subunits. Obviously, functional sites with a catalytic or binding activity have less freedom to mutate than those that form part of linkers maintaining the tertiary structure. As an example, the average substitution rate of nine SSU rRNA sites involved in tRNA binding (3) relative to the average rate of the entire SSU rRNA is 0.105, in other words these sites are 10 times less prone to substitution than the average site. For five tRNA binding sites in LSU rRNA the average rate is 0.073 relative to the entire molecule. The 30 sites of the multibranched loop at the end of helix G1 in LSU rRNA, which is involved in peptidyl transfer (30), have an average rate of 0.065.

The peripheric location of the insertion sites of domain-specific and eukaryotic taxon-specific insertions is clearly demonstrated in Figure 5. In the case of mammalian ribosomes, the location of the insertions at the ribosome surface has already been deduced (31) by comparison of electron microscopic images of bacterial and rabbit ribosomes. The location of two groups of insertions in the small subunit and five groups of insertions in the large subunit was postulated on the basis of three-dimensional models of the bacterial rRNAs fitted to the observed shapes of bacterial and mammalian

Table 3. Inventory of taxon-specific insertions in SSU rRNA of Eucarya[a]

Taxon	SSU rRNA helix number																		
	6	8/e1	10	10\e1	17\e1	23/e3	23/e5	23/e6	23/e7	23/e15	23/e16-e17	29	37	43	43/e1	43/e2	43/e3	43/e4	45/e1
Entamoebidae	1		1	1					1			1	1	2					1
Granuloreticulosea[b]	02	03	3	3		3	3	3	3	3	3	23	2	1	01			01	3
Heterolobosea	1		1	1	1		1		1			1	1	1	1	01	01	1	1
Euglenida	1	1	2	1			1		1	2		2	1	2					3
Kinetoplastida	1	1	1	1					1	1	1	2	1	1	1			1	1
Acanthamoeba	1	01	1	1		1	1		1			1	1	1					1
Lobosea	1	01	1	1					1			1	1	1					
Myxogastria	1	1	1	1			1		1			1	1	1					
Apicomplexa (Alveolata)	1	01	1	1					01			1	1	1					01
Ciliophora (Alveolata)	1	01	1	1					1			1	1	1					01
Labyrinthulida (Stramenopiles)	1		1	1					1			1	1	1					01
Fungi	1		1	1					1			1	1	1					01
Choanoflagellida	1	1	1	1					1			1	1	1					03
Metazoa																			
Myxozoa	1		1	1					1			1	1	1	01	01	01	01	
Annelida	1		1	1					1			1	1	1	01			01	
Neodermata (Platyhelminthes)	1		1	1			1	1	1			1	1	1					
Cladocera and Cyclestherida (Arthropoda, Crustacea)	1		1	1		1	1		1			1	1	1					
Pterygota (Arthropoda, Tracheata, Hexapoda, Insecta)	1	01	1	1		01	1	01	1	03		1	1	1	01			01	01
Chaetognatha	1	1	1	1					1			1	1	1					
Other eukaryotes	1		1	1					1			1	1	1					

[a]Insertions are indicated as follows. 0, no insertion; 1, presence of a helix; 2, presence of a hairpin with a loop large enough to potentially harbour an additional, as yet unknown, structure such as a branching into more hairpins; 3, presence of an insertion of completely unknown structure. Combinations of numbers mean that each case applies to a subset of the species of the taxon, e.g. 01 means that a subset contains a helix of known structure. Single 0s (absence of any structure in all species of the taxon) are not marked.

[b]In this taxon the area corresponding to helices 10 and 10\e1, and the area between helices 23 and 24, have an unknown structure

subunits. However, some of the inferred locations, namely those of insertion groups named L3/L32 and L1 (31), are incompatible or only partially compatible with the attachment sites for the insertions as determined in this work.

Figure 1 shows the position of Eucarya- and taxon-specific helices in the rRNAs and Tables 3–5 give a detailed inventory of which helices occur in which taxa. In order to assess the evolutionary origin of these structures, it is useful to consider the pattern of their appearance in a phylogenetic perspective. This is attempted in Figure 6 for SSU rRNA, for which the largest and most diverse sequence set is available. Figure 6A shows a phylogenetic tree of the Eucarya domain. The topology is a consensus of trees constructed from alignments of SSU rRNA on the one hand and a set of proteins comprising actin, α- and β-tubulin and elongation factor 1-α on the other hand (32,33). Figure 6B is an expansion of the Metazoan cluster, which accounts for 26% of the eukaryotic SSU rRNA sequences used in the present study. In each branch of the trees, the presence of extra helices in excess of the eukaryotic core is indicated by an appropriate symbol. The trees show that there is no phylogenetic consistency in the presence of these structures. To cite just one example, helix 23/e5, which is rather exceptional, is encountered in protists as diverse as the Myxogastria, Acanthamoebidae, Euglenida and Heterolobosea, while in the animal kingdom it is found in species belonging to the phyla Platyhelminthes and Arthropoda (full details can be found in Table 3).

Such a polyphyletic distribution of a character, in this case the presence of a helix, can be explained in two ways. Either it was present in the ancestral SSU rRNA and lost subsequently in the majority of evolutionary lineages, or it has appeared independently in several unrelated lineages. Clark (13) favoured the former hypothesis. He called the taxon-specific insertions PRLs, for progenote rRNA linkers, and assumed their presence in ancestral rRNAs, their complete (for the Eucarya core helices) or partial (for the taxon-specific helices)

Table 4. Inventory of taxon-specific insertions in LSU rRNA of Eucarya[a]

Taxon	LSU rRNA helix number[b]																
	B16	C1	C1/e1	C1/e2	C1/e3	C1/e4	C1/e5	C1/e6	D4\e1	D15\e1	D21	E9\e1	E12	E21/e1	E21/e2	G5/e1-e2	G1\e1
Entamoebidae	3	2								1	1	1	2			1	
Cryptophyta		2							1	1	1	1	1	3	1	1	
Euglenida		2								1	1	1	2	3	3	3	
Kinetoplastida	3	2							1	1	2	1	2	3	3	3	3
Myxogastria		1	1	1	2	1				1	2	1	2	3	3	3	1
Apicomplexa (Alveolata)	03	1	1	1	1	1				1	12	1	1	1	1	1	03
Stramenopiles		1	1	1	1	1			01	1	1	1	1	1	1	1	
Rhodophyta		1	1	1	1	1				1	1		1	1		1	
Embryophyta		1	1	1	1	1	1	1		1	1	1	1	1	1	1	
Insecta (Metazoa, Arthropoda, Tracheata)		1	1	1	2	1				2	2	12	2	1	1	3	
Chordata (Metazoa)		1	1	1	2	3			1	1	2	12	2	23	23	1	
Other eukaryotes[c]		1	1	1	1	1				1	1	1	1	1	1	1	

[a]Conventions as in Table 3.
[b]Hairpins G5/e1 and G5/e2 belong to the Eucarya core but the loop in between contains an insertion of unknown structure in Euglenida, Kinetoplastida, Myxogastria and Insecta, which is indicated by code 3.
[c]Among the presently known sequences, the Fungi, Chlorophyta and a subset of the stramenopiles conform to this secondary structure pattern.

Table 5. Helix occupation in Diplomonadida, Parabasalidea and Microsporidia[a]

Taxon	SSU rRNA helix number[b]														
	8/e1 +	10	10\e1	11	18	23/e1	23/e2	23/e4	23/e7	23/e8-e12	23/e13-e14	23/e15 +	23/e4-e17	43	46
Diplomonadida	01	1		01	1	1							3	1	1
Parabasalidea	1	1			1	01		01		1		1		1	1
Microsporidia				01		01								01	

Taxon[c]	LSU rRNA helix number																	
	B6	B7	B8	B16	B17-B19	B23	C1/e1-e4	D4	D5	D6	D15\e1	E9\e1	E16	E21/e1-e2	G5	G5/e1-e2	I1	I1/e1-e2
Diplomonadida	1	01	01		1	1		1		1	1		3		1		1	
Microsporidia	01				3	01		01										

[a]Conventions and symbols are as in Table 3.
[b]This table lists helices of the eukaryotic core structure that are missing in at least one of the three protist taxa or some of their species. Exceptions are helices 8/e1 and 23/e15 in SSU rRNA, marked with a plus sign, which are present in some of these taxa though they do not belong to the eukaryotic core.
[c]There are not enough data on Parabasalidea LSU rRNA sequences to derive a dependable pattern of missing helices.

maintenance in eukaryotes and their loss in prokaryotes. His explanation is that the primitive ribosome consisted of an assembly of several functional RNA chains, which in the course of evolution became connected by linkers which are dispensable for the function but useful for ensuring simultaneous transcription and equimolar synthesis of the functional and more conserved sequences. The preferential disappearance of the linkers in prokaryotes would be due to the streamlining of their rapidly replicating genome and comparable to the elimination of introns. The alternative explanation, also considered by Clark (13), is that insertions are allowable at certain sites of the molecule because they do not interfere with the function and hence they arise and disappear randomly in evolutionary history. We favour the latter hypothesis, which fits nicely with the peripheric location of the insertions, as opposed to the central location of the functional sites. As for the introns present in the rRNA precursors of certain species, there is no such constraint on their location since they are absent from the mature rRNA. An additional argument for the random and independent appearance of the supernumerary helices (red outline in Fig. 1) is that in many cases hardly any sequence similarity is observable among sequences of the same helix in distant taxa. If these areas have indeed arisen independently in different evolutionary lines, then they should not be included in sequence alignments used for the reconstruction of Eucarya phylogeny as a whole. They may be of some use for studying the phylogeny of sets of closely related species sharing the presence of a given expansion structure. Even so, it has been pointed out that some of these structures are rich in simple repetitive sequences, resulting from rapid

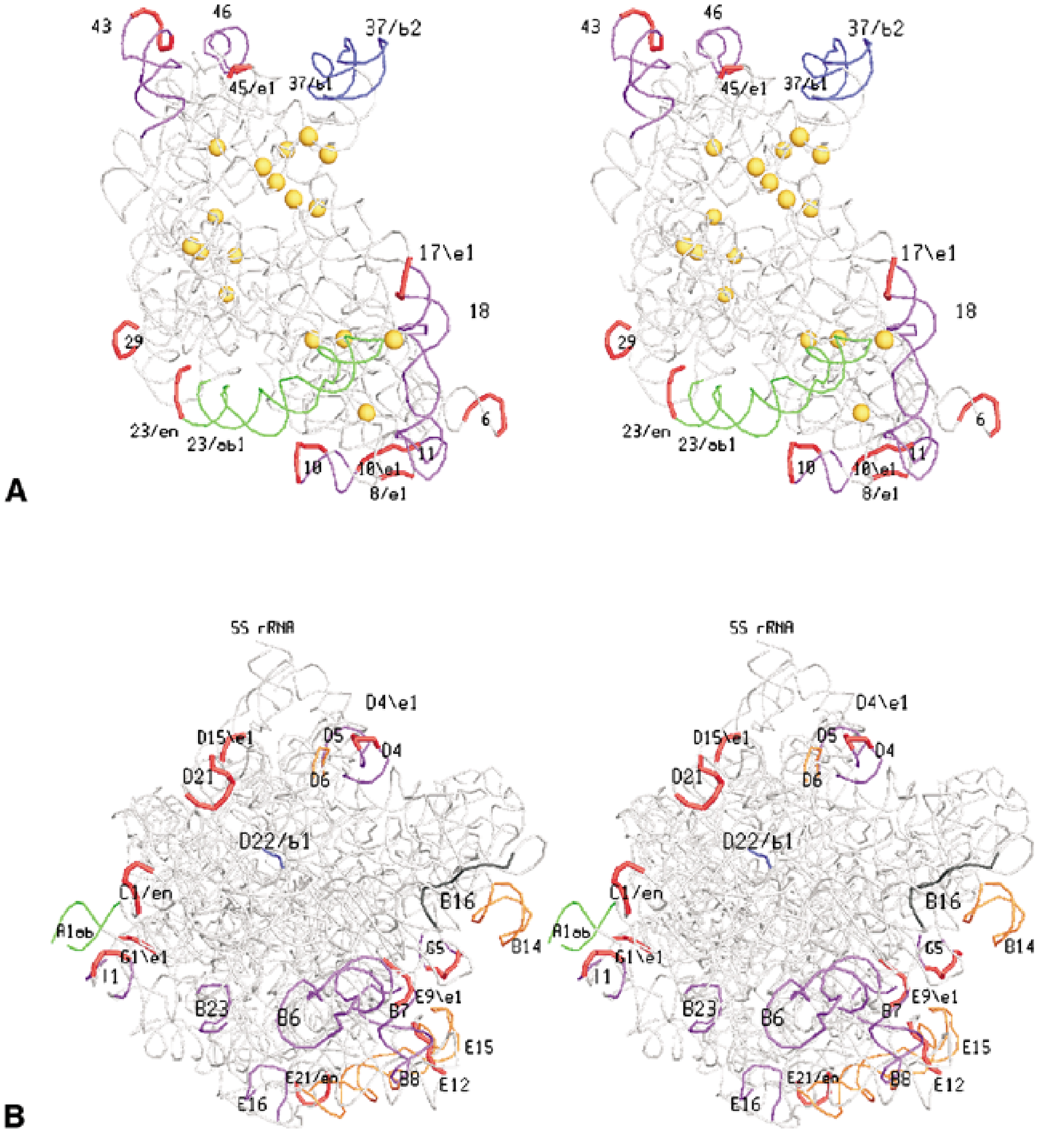


Figure 5. Domain-specific helices or their attachment points in the tertiary structures of SSU rRNA (**A**) and LSU rRNA (**B**). A stereo drawing of each subunit is shown from the solvent side. Domain-specific structures are coloured using the code in Figure 1. For the structures present only in Eucarya, the loops in bacterial rRNAs homologous to insertion points in Eucarya rRNAs are coloured red. Helices deleted in a number of Diplomonadida, Parabasalidea and/or Microsporidia according to the pattern listed in Table 5 are coloured purple. Helices 10, 43, G5 and I1 deleted in some of these protists but bearing insertions in other Eucarya, are coloured purple with a red loop. Yellow spheres in the SSU rRNA drawing indicate insertion points of group I introns in eukaryotic primary transcripts.

evolution due to replication slippage (34), a fact which requires extra caution when they are used for phylogenetic analysis (35).

SUPPLEMENTARY MATERIAL

The following supplementary material is available at NAR Online. (i) A set of six secondary structure models: SSU and LSU rRNA of the bacterium *T.thermophilus*; SSU and LSU rRNA of the yeast *Saccharomyces cerevisiae*, which conform to the eukaryotic core structures; SSU rRNA of the amoeba *Naegleria gruberi* and LSU rRNA of the moss *Funeraria hygrometrica*, both of which possess a number of taxon-specific helices. (ii) Rotating images of the tertiary structure variability maps of *T.thermophilus* SSU rRNA, LSU rRNA and total rRNA.

ACKNOWLEDGEMENTS

J.W. has a scholarship from the Special Research Fund of the University of Antwerp. Y.V.d.P. is a research fellow of the Fund for Scientific Research, Flanders.

REFERENCES

1. Wimberly,B.T., Brodersen,D.E., Clemons,W.M., Morgan-Warren,R.J., Carter,A.P., Vonrhein,C., Hartsch,T. and Ramakrishnan,V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.

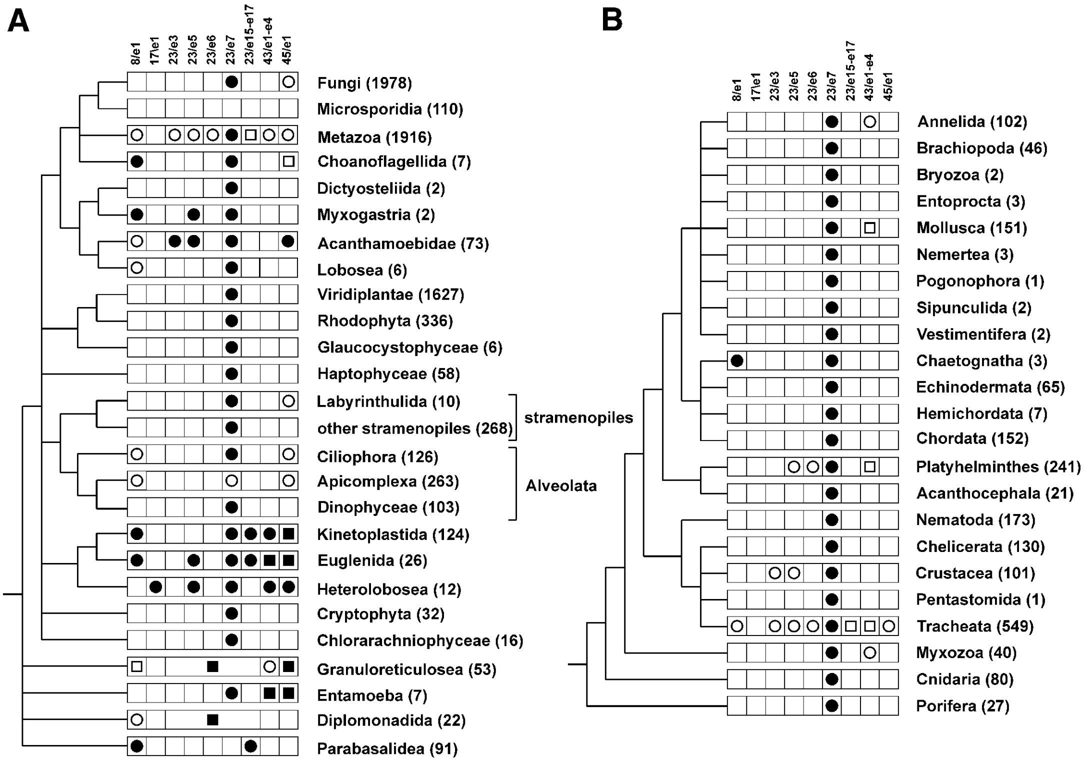


Figure 6. Distribution of SSU rRNA supernumerary helices in the phylogeny of the Eucarya domain (left) and the Metazoan kingdom (right). The presence of supernumerary helices or other structures in each taxon are indicated as follows: circle, helix of known structure; square, helix ending in an unknown structure or insertion of entirely unknown structure. A filled symbol indicates that the structure is present in all species of the taxon, an open symbol indicates that it is present in a subset of the species. The number of SSU rRNA sequences examined for each taxon is shown in parentheses after its name. In the Granuloreticulosea and the Diplomonadida the entire area between helices 23 and 24 has an unknown structure. A number of small taxa with an SSU rRNA structure conforming to the eukaryotic core are omitted from the tree.

2. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
3. Yusupov,M.M., Yusupova,G.Zh., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H.D. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science*, **292**, 883–896.
4. Schwartz,R.M. and Dayhoff,M.O. (1978) Ribosomal and other RNAs. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence*. National Biomedical Research Foundation, Silver Spring, MD, Vol. 5, supplement 3, pp. 327–337.
5. Gutell,R.R. (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res.*, **22**, 3502–3507.
6. Van de Peer,Y., De Rijk,P., Wuyts,J., Winkelmans,T. and De Wachter,R. (2000) The European small subunit ribosomal RNA database. *Nucleic Acids Res.*, **28**, 175–176.
7. Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucleic Acids Res.*, **21**, 3055–3074.
8. Schnare,M.N., Damberger,S.H., Gray,M.W. and Gutell,R.R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J. Mol. Biol.*, **256**, 701–719.
9. Wuyts,J., De Rijk,P., Van de Peer,Y., Winkelmans,T. and De Wachter,R. (2001) The European large subunit ribosomal RNA database. *Nucleic Acids Res.*, **29**, 175–177.
10. Van de Peer,Y., Neefs,J.-M., De Rijk,P. and De Wachter,R. (1993) Reconstructing evolution from eukaryotic small ribosomal subunit RNA sequences: calibration of the molecular clock. *J. Mol. Evol.*, **37**, 221–232.
11. Van de Peer,Y., Van der Auwera,G. and De Wachter,R. (1996) The evolution of stramenopiles and alveolates as derived by 'substitution rate calibration' of small ribosomal subunit RNA. *J. Mol. Evol.*, **42**, 201–210.
12. Neefs,J.-M., Van de Peer,Y., De Rijk,P., Goris,A. and De Wachter,R. (1991) Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.*, **19**, 1987–2015.
13. Clark,C.G. (1987) On the evolution of ribosomal RNA. *J. Mol. Evol.*, **25**, 343–350.
14. Szymanski,M., Barciszewska,M.Z., Barciszewski,J. and Erdmann,V.A. (2000) 5S ribosomal RNA database Y2K. *Nucleic Acids Res.*, **28**, 166–167.
15. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21.
16. Van de Peer,Y., Chapelle,S. and De Wachter,R. (1996) A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.*, **24**, 3381–3391.
17. Wuyts,J., De Rijk,P., Van de Peer,Y., Pison,G., Rousseeuw,P. and De Wachter,R. (2000) Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA. *Nucleic Acids Res.*, **28**, 4698–4708.
18. Van de Peer,Y., Robbrecht,E., de Hoog,S., Caers,A., De Rijk,P. and De Wachter,R. (1999) Database on the structure of small subunit ribosomal RNA. *Nucleic Acids Res.*, **27**, 179–183.
19. Van de Peer,Y. and De Wachter,R. (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.*, **10**, 569–570.
20. Woo,M., Neider,J., Davis,T. and Shreiner,D. (1999) *OpenGL Programming Guide*, 3rd edn. Addison-Wesley, Reading, MA.
21. Cavalier-Smith,T. (1989) Molecular phylogeny. Archaebacteria and Archezoa. *Nature*, **339**, 100–101.
22. Germot,A., Philippe,H. and Le Guyader,H. (1997) Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP70 in *Nosema locustae*. *Mol. Biochem. Parasitol.*, **87**, 159–168.

23. Roger,A.J., Svard,S.G., Tovar,J., Clark,C.G., Smith,M.W., Gillin,F.D. and Sogin,M.L. (1998) A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. *Proc. Natl Acad. Sci. USA*, **95**, 229–234.

24. Hashimoto,T., Sanchez,L.B., Shirakura,T., Muller,M. and Hasegawa,M. (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. *Proc. Natl Acad. Sci. USA*, **95**, 6860–6865.

25. Van de Peer,Y. and De Wachter,R. (1997) Evolutionary relationships among the eukaryotic crown taxa taking into account site to site rate variation in 18S rRNA. *J. Mol. Evol.*, **45**, 619–630.

26. Ben Ali,A., Wuyts,J., De Wachter,R., Meyer,A. and Van de Peer,Y. (1999) Construction of a variability map for eukaryotic large subunit ribosomal RNA. *Nucleic Acids Res.*, **27**, 2825–2831.

27. Choe,C.P., Hancock,J.M., Hwang,U.W. and Kim,W. (1999) Analysis of the primary sequence and secondary structure of the unusually long SSU RNA of the soil bug, *Armadillidium vulgare*. *J. Mol. Evol.*, **49**, 798–805.

28. Milyutina,I.A., Aleshin,V.V., Mikrjukov,K.A., Kedrova,O.S. and Petrov,N.B. (2001) The unusually long small subunit ribosomal RNA gene found in amitochondriate amoeboflagellate *Pelomyxa palustris*: its rRNA predicted secondary structure and phylogenetic implication. *Gene*, **272**, 131–139.

29. Gargas,A., DePriest,P.T. and Taylor,J.W. (1995) Positions of multiple insertions in SSU rDNA of lichen-forming fungi. *Mol. Biol. Evol.*, **12**, 208–218.

30. Vester,B. and Garrett,R.A. (1988) The importance of highly conserved nucleotides in the binding region of chloramphenicol at the peptidyl transfer centre of *Escherichia coli* 23S ribosomal RNA. *EMBO J.*, **7**, 3577–3587.

31. Dube,P., Bacher,G., Stark,H., Mueller,F., Zemlin,F., van Heel,M. and Brimacombe,R. (1998) Correlation of the expansion segments in mammalian rRNA with the fine structure of the 80 S ribosome; a cryoelectron microscopic reconstruction of the rabbit reticulocyte ribosome at 21 Å resolution. *J. Mol. Biol.*, **279**, 403–421.

32. Van de Peer,Y., Baldauf,S., Doolittle,W.F. and Meyer,A. (2000) An updated and comprehensive rRNA phylogeny of the (crown) eukaryotes based on rate calibrated evolutionary distances. *J. Mol. Evol.*, **51**, 565–576.

33. Baldauf,S.L., Roger,A.J., Wenk-Siefert,I. and Doolittle,W.F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, **290**, 972–977.

34. Hancock,J.M. (1995) The contribution of DNA slippage to eukaryotic nuclear 18S rRNA evolution. *J. Mol. Evol.*, **40**, 626–639.

35. Hancock,J.M. and Vogler,A.P. (2000) How slippage-derived sequences are incorporated into rRNA variable-region secondary structure: implications for phylogeny reconstruction. *Mol. Phylogenet. Evol.*, **14**, 366–374.