© 1996 Oxford University Press Nucleic Acids Research, 1996, Vol. 24, No. 17 3381–3391

A quantitative map of nucleotide substitution rates in bacterial rRNA

Yves Van de Peer, Sabine Chapelle and Rupert De Wachter*

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

Received May 17, 1996; Revised and Accepted July 16, 1996

ABSTRACT

A recently developed method for estimating the variability of nucleotide sites in a sequence alignment [Van de Peer, Y., Van der Auwera, G. and De Wachter, R. (1996) *J. Mol. Evol.* 42, 201–210] was applied to bacterial 16S, 5S and 23S rRNAs. In this method, the variability of each nucleotide site is defined as its evolutionary rate relative to the average evolutionary rate of all the nucleotide sites of the molecule. Spectra of evolutionary rates were calculated for each rRNA and show the fastest evolving sites substituting at rates more than 1000 times that of the slowest ones. Variability maps are presented for each rRNA, consisting of secondary structure models where the variability of each nucleotide site is indicated by means of a colored dot. The maps can be interpreted in terms of higher order structure, function and evolution of the molecules and facilitate the selection of areas suitable for the design of PCR primers and hybridization probes. Variability measurement is also important for the precise estimation of evolutionary distances and the inference of phylogenetic trees.

INTRODUCTION

The nucleotides in rRNAs show a considerable spread in substitution rate, attributable to local differences in functional and structural constraints. Detailed information about the variability or conservation of nucleotide positions in rRNA is important for several reasons. For example, sites to which a function can be assigned are often conserved in structure (see for example 1,2). Furthermore, conserved regions are of great importance for sequence alignment and the search for homologous regions in sequences of different organisms. Oligonucleotides can be developed on the basis of the more conserved regions of the molecule and used as universal primers for the amplification of the same gene in other organisms. On the other hand, highly variable sequence regions can be used for the development of species-specific hybridization probes or PCR primers, applicable in the detection and identification of microorganisms. The measurement of site variability is also important from a phylogenetic point of view. Conserved areas can be used to unravel old relationships, while the more variable regions can be used to study evolutionary relationships between closely related organisms. Regarding phylogenetic tree construction, the study of site variability is gaining much interest lately. Newly developed tree construction methods take into account differences in nucleotide substitution rates, which leads to more consistent tree topologies (3–6).

Quantitative estimation of the substitution rates or variabilities (both terms will be used as synonyms throughout this paper) of nucleotide sites is not straightforward. For example, in previous studies, variability of sites was sometimes estimated by computing the proportion of the most common nucleotide (see for example 7). Manske and Chapman (8) pointed out that the main problem with this strategy is that it ignores the frequencies of the other less common nucleotides. Therefore, these authors suggested another method to measure the site variability, considering the relative frequencies of every nucleotide at a certain alignment position. However, this method is also not very appropriate for measuring the variability of nucleotide positions, because it still ignores the evolutionary distance necessary to achieve a substitution, as demonstrated by Van de Peer *et al*. (5).

Another way of measuring the variability of sites is by the use of maximum parsimony. Starting from a known phylogeny, the number of changes for every position is inferred from the reconstruction of nucleotide or amino acid states at the internal nodes of that particular tree topology (9–11). However, since the number of changes at each site is determined by a maximum parsimony approach (12), this method is biased and likely to give an underestimate of the number of changes that have actually occurred, especially for long branches in the tree (13,14).

Recently, a new method was developed for measuring the relative substitution rate of individual sites in a nucleotide sequence alignment on the basis of a distance approach (5,6). The main advantage of this method is that it does not depend upon a given tree topology and that nucleotide site variabilities can be estimated on the basis of several hundreds of sequences. This is important, since the more sequences taken into consideration, the more accurate the estimate.

In our research group, a database on the small and the large ribosomal subunit RNA (respectively SSU and LSU rRNA) is maintained and made available to the scientific community (15,16). We also used to maintain an alignment for the far smaller 5S rRNA (17). In this study, we determined the variability of the nucleotide sites in bacterial 5S rRNA, SSU rRNA and LSU rRNA. Bacterial sequences were chosen for two reasons. Firstly, the bacterial sequences form the most numerous subset of known rRNA sequences. Secondly, bacterial rRNAs suffer less from length heterogeneity than their eukaryotic counterparts (16,18),

[*] To whom correspondence should be addressed

which makes it easier to align them properly and to deduce their complete secondary structure.

ESTIMATION OF NUCLEOTIDE SUBSTITUTION RATES

Estimation of relative nucleotide substitution rates for bacterial 5S, 16S and 23S rRNA was as follows. For an alignment of *N* sequences, *N*(*N* – 1)/2 pairwise evolutionary distances *d* are computed according to the equation of Jukes and Cantor (19)

$$d = -\frac{3}{4}1n\left(1 - \frac{4}{3}f\right) \qquad \mathbf{1}$$

where *f* is the dissimilarity between two sequences, defined as the fraction of substitutions observed when the sequences are aligned. When all pairwise distances are computed, they are classified into a number of distance intervals, e.g. distances from 0 to 0.005, distances from 0.005 to 0.010, and so on. Next, for each alignment position and for each distance interval, the fraction of sequence pairs possessing a different nucleotide is computed. For each position, the fraction of sequences showing a difference is plotted as a function of the distance between them (5). A curve obeying equation

$$p_i = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}v_i d\right)\right] \qquad \mathbf{2}$$

is then fitted to these points by non-linear regression. Equation **2** expresses the probability p_i that alignment position *i* contains a different nucleotide in two sequences, as a function of the evolutionary distance *d* separating them. The slope of the curve in the origin yields the specific nucleotide substitution rate v_i for position *i* (5).

Actually, the estimated nucleotide substitution rates are not yet optimal, because they are derived on the basis of a distance matrix computed by means of equation **1**. This equation only gives a first approximation of the relation between dissimilarity and distance, since it starts from the unrealistic assumption that all nucleotides have the same substitution rate. Therefore, after estimation of all v_i values, alignment positions are grouped into sets of similar variability. A spectrum of relative nucleotide substitution rates is thus obtained. Such spectra are shown for the three RNA molecules in Figure 1. Once the shape of a spectrum is known, it is possible to derive the following equation for the dissimilarity, *f*, as a function of the evolutionary distance *d* (6):

$$f = \frac{3}{4}\left\{1 - \exp\left[-\frac{4}{3}p1n\left(1 + \frac{d}{p}\right)\right]\right\} \qquad \mathbf{3}$$

The value of parameter *p* depends on the shape of the substitution rate spectrum. The inverse of equation **3**

$$d = p\left[\left(1 - \frac{4}{3}f\right)^{-\frac{3}{4p}} - 1\right] \qquad \mathbf{4}$$

gives a more accurate conversion of dissimilarity into distance than equation **1**. It is then used to obtain an improved estimate of the pairwise distances from the observed dissimilarities. The relative substitution rate v_i of each alignment position is then estimated again on the basis of these new evolutionary distances, as described above, and a new spectrum of evolutionary rates is derived. This iterative process is repeated several times until the nucleotide substitution rates v_i do not change anymore. A more detailed description of nucleotide rate calibration is given elsewhere (6).

Table 1. Number of different bacterial rRNA sequences used for nucleotide substitution rate calibration

Taxon[a]	Number of sequences used[b]			
	16S rRNA	23S rRNA	5S rRNA	Combined[c]
Chlamydiae	4			
Cyanobacteria	4	1	5	
Fibrobacter	3			
Flavobacteria and relatives	50	2	13	2
Fusobacterium and relatives	18		3	
Gram-positives and relatives, low GC	50	30	42	17
Gram-positives and relatives, high GC	50	10	27	8
Green sulfur	3	1	4	1
Green non-sulfur	4		3	
Planctomyces and relatives	10	1	6	1
Proteobacteria α	50	7	46	5
Proteobacteria β	50	7	13	1
Proteobacteria γ	50	6	80	5
Proteobacteria δ	42		6	
Proteobacteria ε	50	2	2	2
Radio-resistant micrococci and relatives	6	1	10	1
Spirochetes	50	2		
Thermotogales	6	1		
Total	500	71	260	43

[a] The classification of bacteria is based on the construction of evolutionary trees (see 22). According to the phylogenetic position observed, species are assigned to one of the taxa described by Woese and collaborators (56,57) and our research group (58).

[b] When several sequences were available for the same species, only the most complete was used. Regarding 23S rRNA, all available species were taken into account. For the 5S rRNA, 260 bacterial species, represented in our sequence alignment, were considered. In the case of 16S rRNA, the number of sequences considered was restricted to 500, because of the computational time required. Therefore, species belonging to taxa comprising more than 50 sequences were selected randomly.

[c] See text for details.

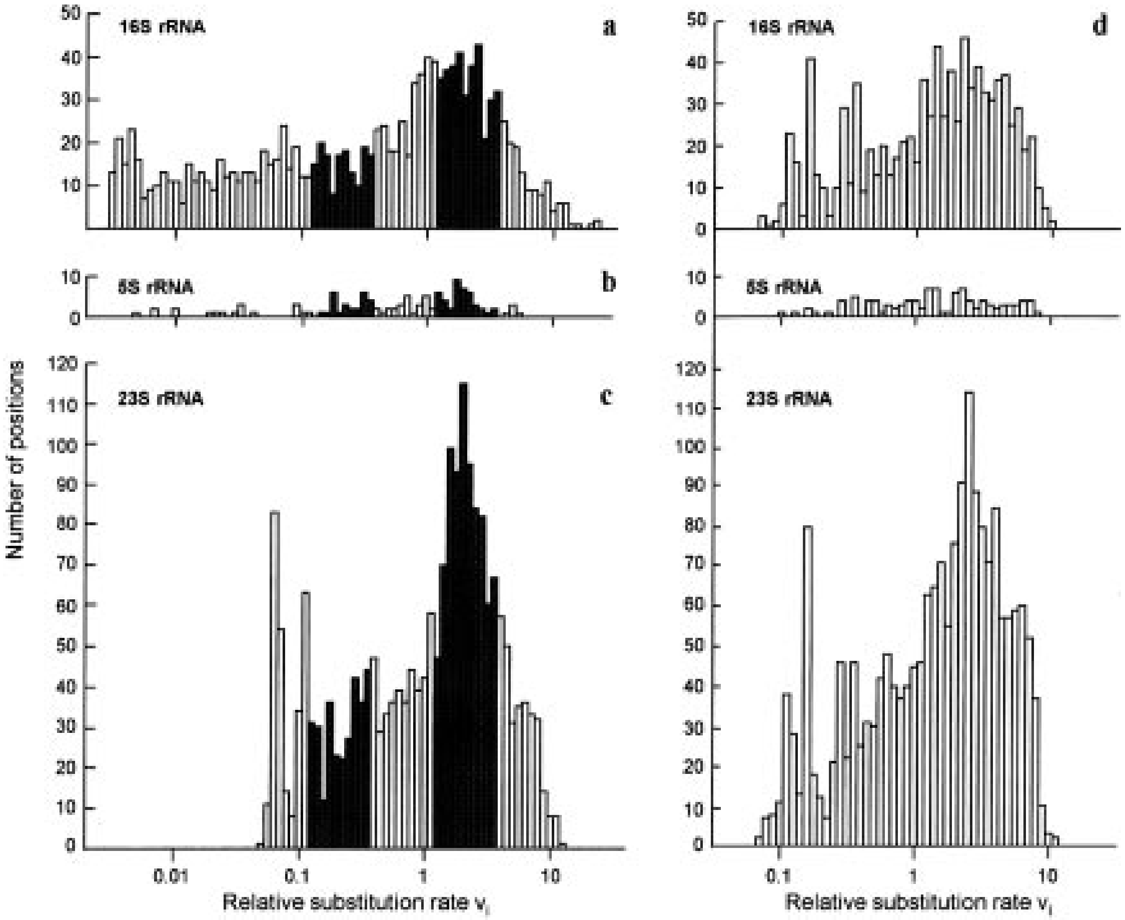

Figure 1. Distributions of relative substitution rates, estimated from an alignment of bacterial 16S (**a**), 5S (**b**) and 23S (**c**) rRNA and from a combined alignment (**d**). The number of species used in each alignment is specified in Table 1. Rates were estimated for each of the alignment positions that are not absolutely conserved and that contain a nucleotide in at least 25% of the aligned sequences. White and shaded areas in spectra (a)–(c) define sets of nucleotides indicated in different colors on the variability maps of Figures 3 and 5.

EVOLUTIONARY RATE SPECTRA OF rRNA NUCLEOTIDES

Figure 1 shows the spectra of relative evolutionary rates for the nucleotide sites of 16S rRNA, 5S rRNA and 23S rRNA. The spectra, calculated as described above, were obtained by five iterations. In the case of 16S rRNA (Fig. 1a), variabilities were estimated from an alignment of 500 16S rRNA sequences, randomly sampled among the main bacterial taxa. The number of species representing each taxon is listed in Table 1. Disregarding absolutely conserved residues, one finds that the most variable sites have a substitution rate ~7000 times higher than the least variable ones. For 5S rRNA, variabilities were estimated from an alignment of 260 sequences, distributed according to Table 1. In the 5S rRNA spectrum (Fig. 1b), the most variable nucleotide sites have a substitution rate ~1260 times that of the least variable ones. For 23S rRNA, site variabilities were estimated from an alignment of all available bacterial sequences, numbering 71, distributed as listed in Table 1. In the corresponding spectrum (Fig. 1c) the ratio of the highest to the lowest rates in 250:1.

In the 'substitution rate calibration' method described above, substitution rates of individual sites are measured for each molecule relative to the average substitution rate of all its sites (6). This implies that a position with a relative substitution rate v_i = 1 in an alignment of 5S rRNA sequences does not necessarily have the same absolute rate as a position with v_i = 1 in an alignment of 16S or 23S rRNA sequences. In order to examine whether there are considerable discrepancies in overall nucleotide site variability among 16S, 5S and 23S rRNA, a combined alignment was made for bacteria for which the three types of rRNAs are known. This resulted in an alignment of 43 sequences, distributed over the different phylogenetic groups as listed in Table 1. The resulting spectrum of relative substitution rates for a combined alignment of 16S, 5S and 23S rRNA is shown in Figure 1d. As can be seen, the range of evolutionary rates and also the shape of the distribution is rather similar for the three types of rRNAs. The spectra derived from the combined alignment are narrower and somewhat different in shape than those shown in Figure 1a–c, because they are computed on a more restricted sequence set. Evidently, a fraction of the positions that appear to be absolutely conserved in a limited set do show some variability in a more complete set, extending the spectra towards the left. In fact, some of the sites for which a very low substitution rate is measured may in fact be totally conserved, the measured variability resulting from a rare sequencing error committed in, for example, one out of 500 species at such a site. However, the overall rate measurement becomes more accurate as the sequence set becomes larger.

VARIABILITY MAPS OF rRNA

Color maps shown in Figures 3 and 5, superimposed on the respective secondary structure models of 16S, 5S and 23S rRNA, were constructed by dividing nucleotides into five variability subsets, indicated in Figure 1a–c as alternately white and shaded areas of the spectra. The relative rate limits of the subsets and the corresponding colors used in the variability maps are as follows: $<10^{-0.925}$ (blue); $10^{-0.925}$–$10^{-0.425}$ (green); $10^{-0.425}$–$10^{+0.075}$ (yellow); $10^{+0.075}$–$10^{+0.575}$ (orange); $>10^{+0.575}$ (red). Absolutely conserved positions ($v_i = 0$) are indicated in purple. Sites colored pink belong to areas that are very variable, but that are deleted in too many sequences to allow a sufficiently accurate measurement of their relative evolutionary rate. Since the distributions are not rectangular, some colors are more abundant than others. These color maps give a much more detailed and quantitative description of positional variability than the crude distinction between variable and conserved areas that is often made by visual inspection of sequence alignments.

Variability map of bacterial 16S rRNA

The models in Figures 2–5 are all drawn clockwise in the direction 5′→3′. Figure 2 shows the secondary structure model of the 16S rRNA of *Escherichia coli.* The secondary structure model adopted for eukaryotic and prokaryotic SSU rRNAs was originally derived (20) 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 that time (20). Gradual improvements were made to the models, as reported in subsequent papers describing our database on SSU rRNA structure (see for example 21,22), taking into account compensating substitutions observed in our sequence alignments and the results of studies by others (reviewed in 23). The model presently adopted for bacterial SSU rRNAs is essentially identical to the models made available in graphic form by Gutell (24). Five ‘tertiary’ interactions derived by Gutell *et al.* (25) on the basis of coordinated substitutions are indicated in Figure 2. Although described as ‘tertiary’ by the latter authors, some of these interactions would be more aptly defined as secondary interactions consisting of single base pairs, since they satisfy the principle of contiguity (26) (see also Table 2).

In Figure 3 the variability of the nucleotide sites of bacterial 16S rRNA is mapped in the shape of the secondary structure model. The variability map shown is largely congruent with a bacterial map published earlier (22). However, in the previous map, variability measurement was less precise, since it was based on Jukes and Cantor distances only and did not involve iterations (see Estimation of Nucleotide Substitution Rates) and sites were divided into five equally numerous categories of increasing variability.

Figure 3 shows that in general the two nucleotides of a base pair have the same or a neighboring color, i.e. they are about equally variable. This is as expected, since the substitution of a base paired nucleotide generally requires a compensating substitution in the opposite strand. However, there are a few exceptions. Most of these occur in sites where a particular base, usually a U or a G, seems to be required in one strand, but the complementary base can change more freely, which is possible due to the existence of G·U pairs. If the preference is for a U in one strand, the opposite base can be A or G. Such is the case with the base pair closing the hairpin loop of helix 25, where a green A or G faces a purple U. If the G must be conserved in one strand, the complementary base can be either C or U. Such a case is found in the last base pair of helix B12 of 23S rRNA (see below). There is also an example of a base pair, namely the penultimate pair of helix 28, where an A seems to be required in one strand. The opposite nucleotide is either U or C. This has been interpreted (24) as meaning that the two ultimate base pairs of helix 28 as drawn in Figure 3 do not actually exist, since a change of A·U to A·C implies the disappearance of the penultimate base pair. An alternative interpretation could be that at certain places in the molecule A·C pairs can be tolerated. The structure of the A·C pair in DNA heteroduplexes has been investigated by X-ray analysis (27) and found to be congruent with the G·T pair but containing a protonated C or possibly a rare tautomeric form of A or C. The general rule that interacting bases have similar substitution rates also applies to the tertiary interactions indicated in Figure 2. The variabilities of the nucleotides participating in each interaction are listed in Table 2, left.

Table 2. Tertiary[a] structure interactions in 16S and 23S rRNA as proposed by Gutell *et al.* (25) on the basis of comparative evidence and confirmed by our study

16S rRNA		23S rRNA	
Number[b]	Variability code[c]	Number	Variability code
1[a]	3-3	1	2-2
2	5-5	2	2,4-6,2
3	3-3	3	2-2
4[a]	6-5	4[a]	3-3
5[a]	5-6	5	2,3-3,2
		6	6,6,6,2-2,6,6,6
		7	6,5-6,6
		8	6,3-3,6
		9	2-2
		10	5-5
		11	5-5,5
		12	2-2
		13	2-2
		14	6-6
		15	2,3-3,2
		16	4-5
		17	3-3

Interactions followed by ([a]) are not actually tertiary, since they satisfy the principle of contiguity (26). On the other hand, two pseudoknots, which do not satify the principle of contiguity and therefore contain tertiary interactions, are drawn directly in the 16S rRNA nucleotide sequence in Figure 2.

[b]Interactions are listed clockwise starting from the 5′-end and are indicated by black circles in Figures 2 (16S rRNA) and 4 (23S rRNA).

[c]Variability codes are as shown in Figures 3 for 16S rRNA and 5 for 23S rRNA.

In the prokaryotic secondary structure model, nine highly variable areas can be distinguished, formed roughly by the following helices: 6; 8–11; 18; P23-1 and 24; 28 and 29; 37–P37-2; 43; 45 and 46; 49. These areas or parts thereof have been previously distinguished on a more intuitive basis by visual inspection of the 16S rRNA alignment (21,28). Much more detail is visible on the variability maps and it turns out that highly variable helices are interrupted occasionally by conserved base pairs, as in helix 24, and, conversely, a highly variable base pair is occasionally intercalated between more conserved ones, as in helix 14. More generally, sequence conservation of the 16S rRNA is found mainly in single-stranded regions. The importance of several highly conserved single-stranded regions for the structure or function of the 16S rRNA molecule is supported by a large number of studies describing specific rRNA–protein interactions or functional sites (for example 28–32). For example, it has been demonstrated that several bases in the highly conserved hairpin loop of helix 27, where eight out of nine bases are completely conserved among the 500 16S rRNA sequences, are directly involved in ribosomal subunit association (33) and initiation of protein synthesis (34). Another highly conserved region is the pseudoknot structure formed by helices 19–21. This tertiary interaction, originally proposed by Woese and Gutell (35), has been shown to be essential for ribosomal function and mild perturbations of the structure generate resistance to streptomycin, an antibiotic known to interfere with the decoding process, i.e. A-site tRNA–ribosome interaction (36).

Variability maps of bacterial 5S and 23S rRNA

Figure 4 shows secondary structure models for 23S rRNA and 5S rRNA (bottom left) of *E.coli*. The shape and the helix numbering system of the 23S rRNA model are according to De Rijk *et al.* (16,37). It conforms largely to the models developed in earlier studies (38,39).

The 5S rRNA model is slightly different from those previously published. Helices are numbered 1–3 in order to comply with the principle used for the large rRNAs, to wit that helices are given different numbers only when separated by a multibranched loop, a pseudoknot loop or a single-stranded area that does not form a loop. Helices 2 and 3 each consist of two segments separated by an internal loop. The two single base bulges on the 5′-strand of helix 2 and flanking the internal loop in this helix are absent in most 5S rRNA secondary structure models (see for example 40). The existence of the bulge on the 5′-strand upstream of the internal loop was proposed by Van den Eynde and De Wachter (41) and that on the 5′-strand downstream of the internal loop was proposed by Egebjerg *et al.* (42). Figure 5 shows the variability maps of 23S and 5S rRNA superimposed upon the secondary structure models.

Two variable double-stranded areas, namely the segment of helix 2 adjoining the bifurcation loop and the helix 3 segment adjoining the hairpin loop, can be distinguished in the 5S rRNA. The high variability of these helices was also noticed previously (see 8,40,43). In the alignment of 260 bacterial sequences, only two positions are absolutely conserved, namely $G_{44}A_{45}$ situated in the hairpin loop of helix 2. A tertiary interaction formed by a Watson–Crick pair between G_{44} and C_{28}, the second bulge on the 5′-strand of helix 2, was proposed by Egebjerg *et al.* (42). However, contrary to the statement of these authors, the bulge corresponding to position 28 in *E.coli* is not always a C. It is substituted by a U, A or G in several species belonging to the Proteobacteria α subdivision, whereas the G_{44} is absolutely conserved among the Bacteria. On the other hand, the existence of the two single base bulges on the 5′-strand of helix 2 is corroborated by compensating substitutions in the single base pairs separating each of these bulges from the internal loop half way along helix 2. This is reflected in the fact that each of these base pairs is formed by two nucleotides of similar variability (orange).

In 23S rRNA, as in 16S and 5S rRNAs, there is a general correspondence in the variabilities of nucleotides forming base pairs. The tertiary interactions identified in the former two molecules (Figs 2 and 4) also link nucleotides of corresponding varariability. This can be seen clearly in Table 2. In the case of 23S rRNA, some of these interactions consist of antiparallel base pairing between sequences of several nucleotides.

Ten highly variable areas (red and orange) can be distinguished in 23S rRNA (Fig. 5). They are formed by the following helices: B8 and B9; B14–B16; D2; D13 and D14; D20; D22; E11–E15, E20; G4 and G5; H1-1. These variable areas were also distinguishable in the *Pseudomonas cepacia* sequence conservation plot of Höpfl *et al.* (44), constructed by comparison of 20 prokaryotic and two chloroplast sequences, and in the *E.coli* sequence conservation plot of Egebjerg *et al.* (1), based on 42 sequences, of which 24 were bacterial/plastid, 11 eukaryotic and seven archaeal. Certain variable helices, such as D14 and E12, are interrupted by conservative internal loops. As in the case of 16S rRNA, the map of Figure 5 improves on previous descriptions in being based on a larger dataset and quantitative measurements and in being more easily surveyable.

Many of the variable areas are characterized by major size variations. The hairpin loop of C1, a helix which itself is only moderately variable, is also a hot-spot for extremely variable insertions in eukaryotic LSU rRNAs. To our knowledge, these insertions were first described by Hassouna *et al.* (45), who referred to them as D(ivergent)-domains. Some areas subject to intense substitution show little or no length heterogeneity, though. This is the case for helices B14–B16, D2 and E11–E13. As a rule, strong length heterogeneity seems to be most common in apical helices, i.e. those ending in a hairpin loop. Helices formed by long distance interactions, i.e. those bounded by multibranched loops, have less freedom to change in length. It should be noted that the molecule contains a number of potential branching points which bear additional helices in a limited set of species. These branching points are not visible in Figures 4 and 5 since they do not exist in *E.coli*. As an example, helices B14 and B15, though apparently forming a continuous helix, were numbered differently because they are separated by a potential branching point. Helix B14-1 branching there in some species does show length heterogeneity. Similarly potential branching points to variable helices exist between helices D13 and D14, as well as E11 and E12. Multibranched loops such as those separating D2–D5 and G3–G5 also bear additional helices subject to sequence and length heterogeneity in certain species.

Large ribosomal subunit RNAs encoded by eukaryotic nuclear genomes and plastid genomes are fragmented into at least two, and in some cases more, separate chains. The fragmentation results from post-transcriptional removal of RNA segments inserted in variable areas of the molecule (29). In most bacteria, 23S rRNA seems to occur as a contiguous molecule, but cases of fragmentation had been reported even before sequences were known (46). In species of the genera *Salmonella* (47) and *Campylobacter* (48), fragmentation has been demonstrated to be the result of post-transcriptional removal of an extension of helix D20 (helix 45 in the numbering according to Höpfl *et al.*; 44). This is one of the most variable helices in the molecule (Figs 4 and 5). To our knowledge, only one fragmented 16S rRNA sequence

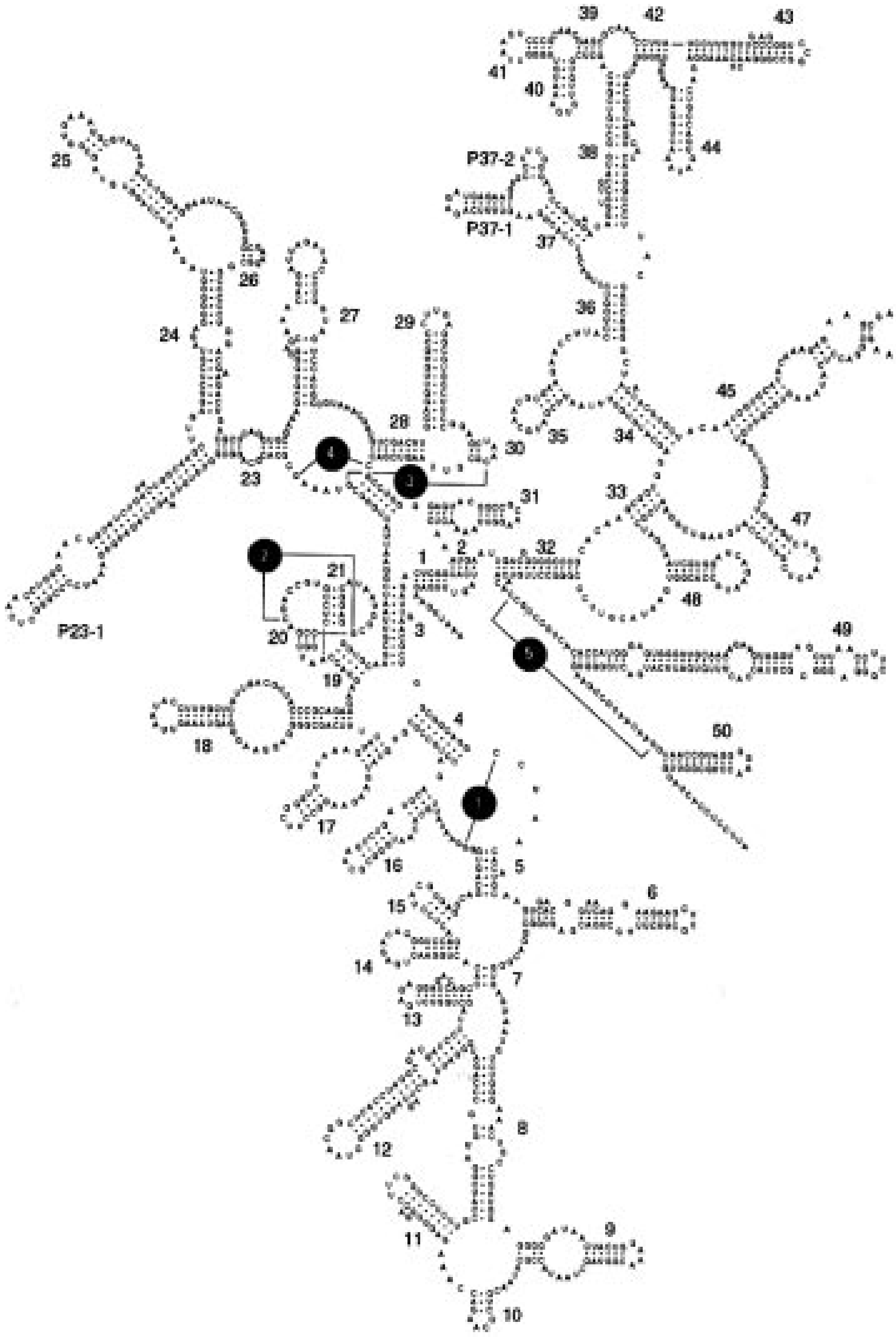

Figure 2. Secondary structure model for 16S rRNA of *E.coli*. The sequence is written clockwise 5′→3′. Tertiary interactions are indicated by solid lines and identified by a number in a black circle. Helix numbering is according to Van de Peer *et al.* (15).

has been reported so far, namely that of *Campylobacter sputorum* (48). In 16S rRNA also, the internal spacer is located in a highly variable region, namely helix 11 (Figs 2 and 3).

To explain the strong correlation between variable regions in the rRNAs and the presence of internal transcribed spacers, two conflicting views have been put forward. On the one hand, it has been proposed that these variable regions represent insertions into the ancestral RNA structure, the presence of which is tolerated since they do not disrupt its function (49). On the other hand, it was suggested that the ancestral RNA genes were discontinuous and that the variable regions are the remains of ancient internal transcribed spacer sequences that separated the functional domains of the ancestral genes. Due to strong evolutionary pressure to increase efficiency of rRNA production and assembly in fast growing organisms such as bacteria, most of these regions were eliminated in the prokaryotic lineages (50–52). Indeed, in eukaryotes variable regions are much more variable in size and fragmentation is much more abundant than in prokaryotes (29).

Beside variable regions, several regions of a highly conserved nature can be distinguished in 23S rRNA. As in the 16S rRNA

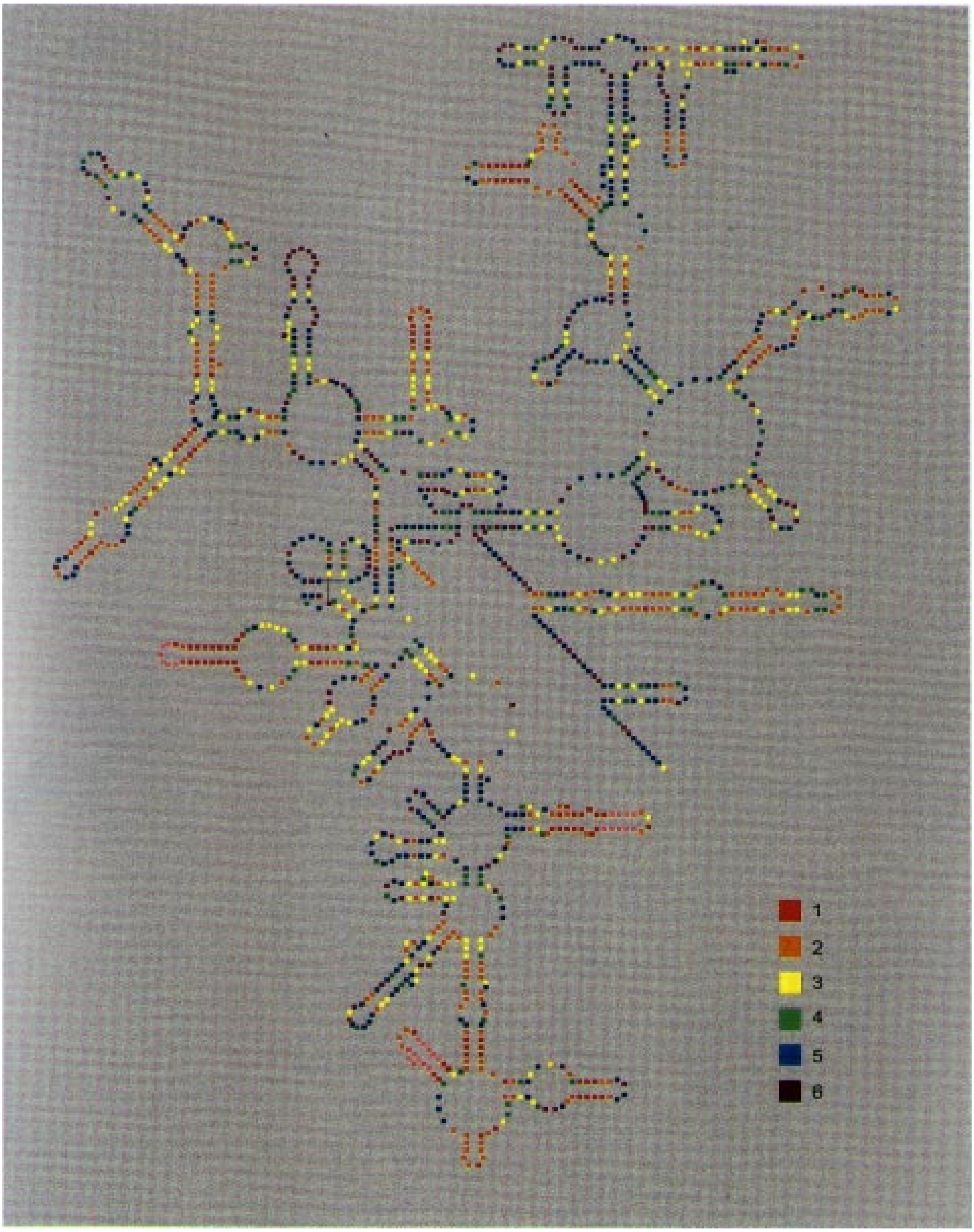

Figure 3. Color map superimposed on the 16S rRNA secondary structure model of *E.coli* shown in Figure 2. Nucleotides are subdivided into five groups of increasing variability (see text for details). The most variable positions are in red, the most conserved in blue. Absolutely conserved positions are indicated in purple. Nucleotides present in *E.coli* but absent in >75% of the bacterial sequences considered are indicated in pink.

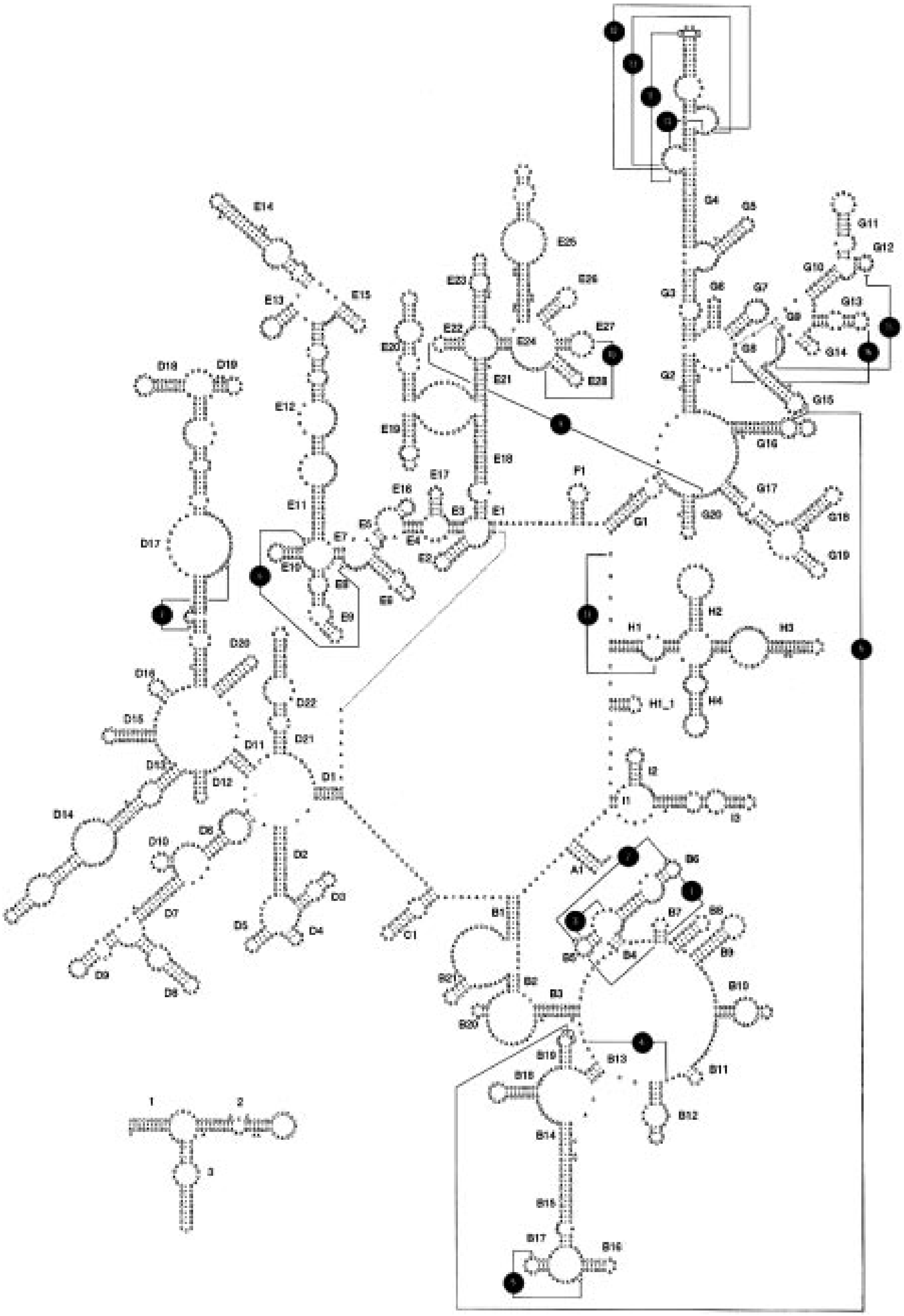

Figure 4. Secondary structure model for the 5S (bottom left) and 23S rRNA of *E.coli*. Sequences are written clockwise 5′→3′. Tertiary interactions are indicated by solid lines and identified by a number in a black circle. Those that consist of antiparallel pairing involving more than 1 bp are indicated by a single interconnecting line. Helix numbering of the 23S rRNA is according to De Rijk *et al*. (16).

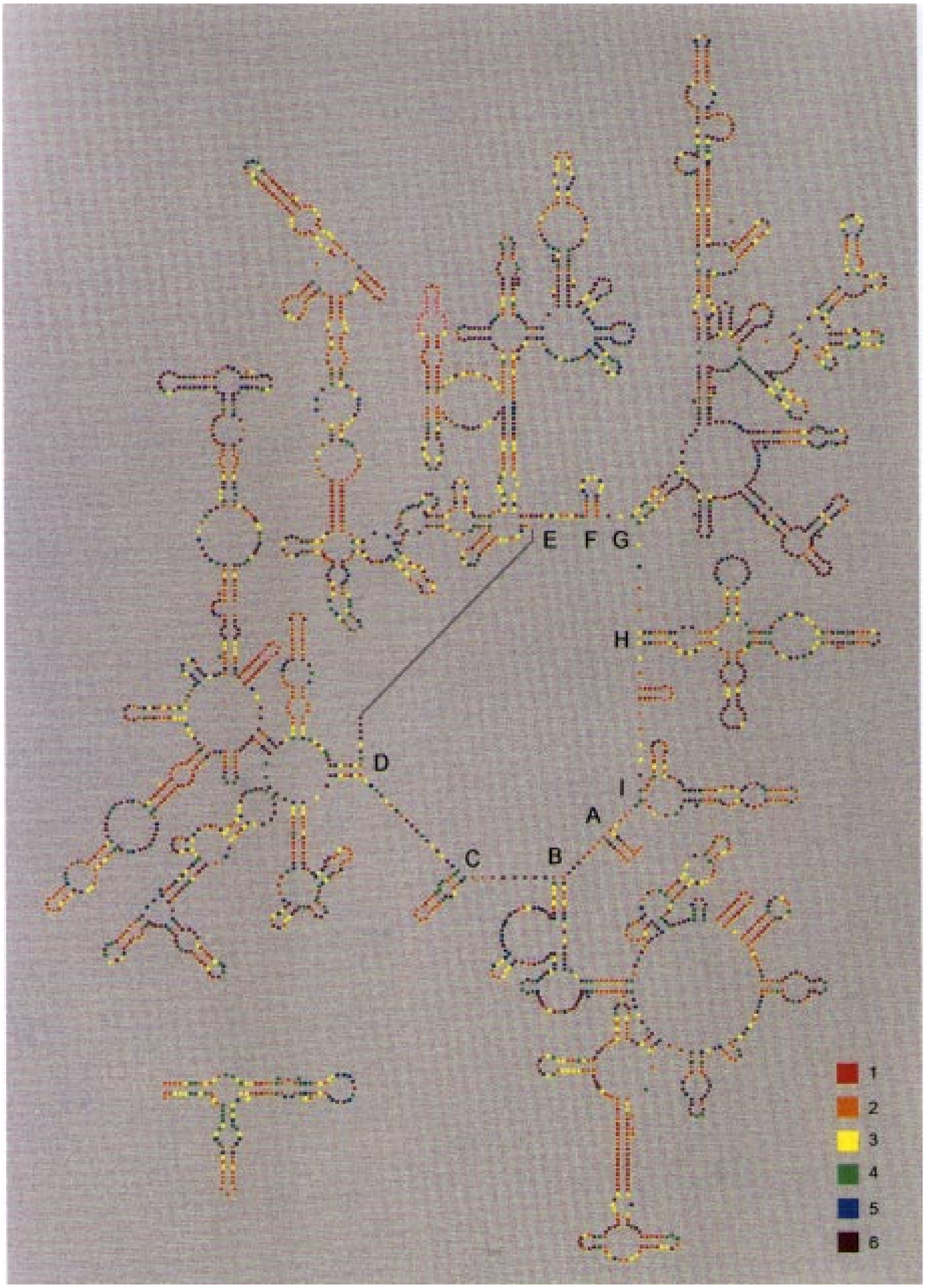

Figure 5. Color maps superimposed on the 5S and 23S rRNA secondary structure models of *E.coli* shown in Figure 4. Interpretation is as in Figure 3.

model, sequence conservation is mostly restricted to single-stranded regions. One of the most conserved areas is the multibranched loop in the G area (Figs 4 and 5). This conserved structure is generally considered to be the major element of the peptidyltransferase center of the ribosome (discussed in detail in 28). Other highly conserved structures in the 23S rRNA are helices D18 and D19, which are part of the so-called GTPase center of the ribosome, helices E22–E28 and the hairpin loop of helix H2 (see 28).

APPLICATIONS OF SUBSTITUTION RATE CALIBRATION

In the above paragraphs, the bacterial rRNA variability maps have been interpreted mainly in the context of structural interactions within the molecule, functional interactions with other molecules and evolutionary aspects of insertion hot-spots. As stated in the Introduction, the maps can also be used, more efficiently than previously published maps constructed on a more intuitive basis, for the design of primers and hybridization probes.

Perhaps the most important application of substitution rate calibration is in the construction of phylogenetic trees. Once the spectrum of relative evolutionary rates of an rRNA has been measured, this information can be used to improve the precision of tree construction by distance methods. Two approaches can be used. One is to divide alignment positions into a number of sets of increasing relative substitution rate. The conversion of sequence dissimilarity into distance can then be carried out on each set separately, taking into account the fact that substitutions in conservative areas carry more weight than those in variable ones. The final distance, found by averaging the distances computed from each set, is used for tree construction. This approach has been followed in a study of eukaryotic evolution (5). In the other approach, sequence dissimilarity is converted into distance for the entire alignment, but the conversion takes into account that dissimilarity rises more slowly as a function of distance for a set of nucleotides mutating with variable rates than for a randomly mutating set. This is achieved by using equation **4** with an appropriate parameter *p* adapted to the shape of the rate spectrum. The latter approach, used to study the evolution of eukaryotic SSU rRNA sequences of different groups of protists, yielded some significant improvements in tree topology (6,53). In particular, tree distortions due to the presence of species with an exceptionally high evolutionary rate are eliminated to a large extent by these methods. These distortions are caused by an underestimation of large distances with respect to small ones if distances are computed assuming equal variability of all nucleotides in a sequence.

In bacteria too, some taxa are characterized by an exceptionally high evolutionary rate. A well known example are the mycoplasmas, which, on the basis of oligonucleotide signatures, have been assigned earlier to the cluster of Gram-positives with low GC content (54). Construction of bacterial SSU rRNA and LSU rRNA evolutionary trees on the basis of Jukes and Cantor (19) or Kimura (55) distances regularly show the mycoplasmas as either an independent evolutionary lineage or as diverging at the base of the aforementioned cluster. In contrast, application of rate calibration to SSU and LSU rRNA sequences consistently shows mycoplasmas originating from within the cluster of Gram-positives with low GC content, as expected (54,56), and this location is supported at a high bootstrap level. This result will be discussed into more detail elsewhere.

Substitution rate calibration is a general method applicable to all genes for which a dependable alignment comprising a considerable number of species is available. It is to be expected that distance trees taking into account the shape of the evolutionary rate spectrum of the molecules used as a molecular clock will significantly improve the trustworthiness of the evolutionary trees obtained. Conceivably, parsimony methods for tree construction that take into account the shape of the evolutionary rate spectrum of genes can also be developed in the future.

ACKNOWLEDGEMENTS

Our research was supported by the Programme on Interuniversity Poles of Attraction (contract 23) of the Federal Office for Scientific, Cultural and Technical Affairs of the Belgian State and by the National Fund for Scientific Research. Yves Van de Peer is a Research Assistant of the National Fund for Scientific Research. We thank Peter De Rijk for comments on the secondary structure of the 23S rRNA.

REFERENCES

1 Egebjerg,J., Larsen,N. and Garrett,R.A. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution*. American Society of Microbiology, Washington, DC, pp. 168–179.

2 Noller,H.F., Moazed,D., Stern,S., Powers,T., Allen,P.N., Robertson,J.M., Weiser,B. and Triman,K. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution*. American Society of Microbiology, Washington, DC, pp. 73–92.

3 Olsen,G.J. (1987) *Cold Spring Harbor Symp. Quant. Biol.*, **LII**, 825–837.

4 Jin,L. and Nei,M. (1990) *Mol. Biol. Evol.*, **7**, 82–102.

5 Van de Peer,Y., Neefs,J.-M., De Rijk,P. and De Wachter,R. (1993) *J. Mol. Evol.*, **37**, 221–232.

6 Van de Peer,Y., Van der Auwera,G. and De Wachter,R. (1996) *J. Mol. Evol.*, **42**, 201–210.

7 MacDonell,M.T., Oritz-Conde,B.A., Last,G.A. and Colwell,R.R. (1986) *J. Microbiol. Methods*, **5**, 295–302.

8 Manske,C.L. and Chapman,D.J. (1987) *J. Mol. Evol.*, **26**, 226–251.

9 Uzzell,T. and Corbin,K.W. (1971) *Science*, **172**, 1089–1096.

10 Holmquist,R., Goodman,M., Conroy,T. and Czelusniak,J. (1983) *J. Mol. Evol.*, **19**, 437–448.

11 Ota,T. and Nei,M. (1994) *J. Mol. Evol.*, **38**, 642–643.

12 Fitch,W.M. (1971) *Syst. Zool.*, **20**, 406–416.

13 Wakeley,J. (1993) *J. Mol. Evol.*, **37**, 613–623.

14 Gu,X., Fu,Y.-X. and Li,W.-H. (1995) *Mol. Biol. Evol.*, **12**, 546–557.

15 Van de Peer,Y., Nicolaï,S., De Rijk,P. and De Wachter,R. (1996) *Nucleic Acids Res.*, **24**, 86–91.

16 De Rijk,P., Van de Peer,Y. and De Wachter,R. (1996) *Nucleic Acids Res.*, **24**, 92–97.

17 Erdmann,V.A., Wolters,J., Huysmans,E. and De Wachter,R. (1985) *Nucleic Acids Res.*, **13**, r105–r153.

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 Jukes,T.H. and Cantor,C.R. (1969) In Munro,H.H. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, NY, pp. 21–132.

20 Nelles,L., Fang,B.-L., Volckaert,G., Vandenberghe,A. and De Wachter,R. (1984) *Nucleic Acids Res.*, **12**, 8749–8768.

21 De Rijk,P., Neefs,J.-M., Van de Peer,Y. and De Wachter,R. (1992) *Nucleic Acids Res.*, **20**, 2075–2089.

22 Neefs,J.-M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res.*, **21**, 3025–3049.

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 Gutell,R.R. (1996) In Zimmerman,R.A. and Dahlberg,A.E. (eds), *rRNA. Structure, Evolution, Processing and Function in Protein Biosynthesis.* CRC Press, Boca Raton, FL, pp. 111–128.

26 Ninio,J. (1971) *Biochimie*, **53**, 485–494.

27 Hunter,W.N., Brown,T., Anand,N.N. and Kennard,O. (1986) *Nature*, **320**, 552–555.

28 Raué,H.A., Musters,W., Rutgers,C.A., van't Riet,J. and Planta,R.J. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution.* American Society of Microbiology, Washington, DC, pp. 217–235.

29 Raué,H.A., Klootwijk,J. and Musters,W. (1988) *Prog. Biophys. Mol. Biol.*, **51**, 77–129.

30 Powers,T. and Noller,H.F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1042–1046.

31 Noller,H.F., Moazed,D., Stern,S., Powers,T., Allen,P.N., Robertson,J.M., Weiser,B. and Triman,K. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution.* American Society of Microbiology, Washington, DC, pp. 73–92.

32 Brimacombe,R. (1995) *Eur. J. Biochem.*, **230**, 365–383.

33 Tapprich,W.E. and Hill,W.E. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 556–60.

34 Tapprich,W.E., Goss,D.J. and Dahlberg,A.E. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 4927–4931.

35 Woese,C.R. and Gutell,R.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 3119–3122.

36 Powers,T. and Noller,H.F. (1991) *EMBO J.*, **10**, 2203–2214.

37 De Rijk,P., Van de Peer,Y., Chapelle,S. and De Wachter,R. (1994) *Nucleic Acids Res.*, **22**, 3495–3501.

38 Noller,H.F., Kop,J., Wheaton,V., Brosius,J., Gutell,R.R., Kopylov,A.M., Dohme,F. and Herr,W. (1981) *Nucleic Acids Res.*, **9**, 6167–6189.

39 Brimacombe,R. and Stiege,W. (1985) *Biochem. J.*, **229**, 1–17.

40 Wolters,J. and Erdmann,V.A. (1988) *Nucleic Acids Res.*, **16**, r1–r70.

41 Van den Eynde,H. and De Wachter,R. (1987) *FEBS Lett.*, **217**, 191–196.

42 Egebjerg,J., Christiansen,J., Brown,R.S., Larsen,N. and Garrett,R.A. (1989) *J. Mol. Biol.*, **206**, 651–668.

43 Ehresmann,B., Ehresmann,C., Romby,P., Mougel,M., Baudin,F., Westhof,E. and Ebel,J.-P. (1990) In Hill,W.E., Dahlberg,A., Garrett,R.A., Moore,P.B., Schlessinger,D. and Warner,J.R. (eds), *The Ribosome. Structure, Function and Evolution.* American Society of Microbiology, Washington, DC, pp. 148–159.

44 Höpfl,P., Ludwig,W., Schleifer,K.H. and Larsen,N. (1989) *Eur. J. Biochem.*, **185**, 355–364.

45 Hassouna,N., Michot,B. and Bachellerie,J.-P. (1984) *Nucleic Acids Res.*, **12**, 3563–3581.

46 Pace,N.R. (1973) *Bacteriol. Rev.*, **37**, 562–603.

47 Burgin,A.B., Parodos,K., Lane,D.J. and Pace,N.P. (1990) *Cell*, **60**, 405–414.

48 Van Camp,G., Van de Peer,Y., Nicolaï,S., Neefs,J.-M., Vandamme,P. and De Wachter,R. (1993) *System. Appl. Microbiol.*, **16**, 361–368.

49 Ware,V.C., Renkawitz,R. and Gerbi,S.A. (1985) *Nucleic Acids Res.*, **13**, 3581–3597.

50 Clark,C.G. (1987) *J. Mol. Evol.*, **25**, 343–350.

51 Spencer,D.F., Collins,J.C., Schnare,M.N. and Gray,M.W. (1987) *EMBO J.*, **6**, 1063–1071.

52 Boer,P.H. and Gray,M.W. (1988) *Cell*, **55**, 399–411.

53 Van de Peer,Y., Rensing,S.A., Maier,U.-G. and De Wachter,R. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 7732–7736.

54 Woese,C.R., Stackebrandt,E. and Ludwig,W. (1985) *J. Mol. Evol.*, **21**, 305–316.

55 Kimura,M. (1980) *J. Mol. Evol.*, **16**, 111–120.

56 Woese,C.R. (1987) *Microbiol. Rev.*, **51**, 221–271.

57 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.*, **176**, 1–6.

58 Van de Peer,Y., Neefs,J.-M., De Rijk,P., De Vos,P. and De Wachter,R. (1994) *System. Appl. Microbiol.*, **17**, 32–38.