Volume 8 Number 10 1980 Nucleic Acids Research

Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence

C.R.Woese, L.J.Magrum, R.Gupta, R.B.Siegel and D.A.Stahl
Department of Genetics and Development, University of Illinois, Urbana, IL 61801, USA

J.Kop, N.Crawford, J.Brosius, R.Gutell, J.J.Hogan and H.F.Noller
Thimann Laboratories, University of California at Santa Cruz, Santa Cruz, CA 95064, USA

Received 2 May 1980

ABSTRACT

We have derived a secondary structure model for 16S ribosomal RNA on the basis of comparative sequence analysis, chemical modification studies and nuclease susceptibility data. Nucleotide sequences of the E. coli and B. brevis 16S rRNA chains, and of RNAse T_1 oligomer catalogs from 16S rRNAs of over 100 species of eubacteria were used for phylogenetic comparison. Chemical modification of G by glyoxal, A by m-chloroperbenzoic acid and C by bisulfite in naked 16S rRNA, and G by kethoxal in active and inactive 30S ribosomal subunits was taken as an indication of single stranded structure. Further support for the structure was obtained from susceptibility to RNases A and T_1. These three approaches are in excellent agreement. The structure contains fifty helical elements organized into four major domains, in which 46 percent of the nucleotides of 16S rRNA are involved in base pairing. Phylogenetic comparison shows that highly conserved sequences are found principally in unpaired regions of the molecule. No knots are created by the structure.

INTRODUCTION

Our understanding of the mechanics of translation today is all but rudimentary, despite two decades of intensive work on the subject. This may reflect the complexity of the mechanism. Yet, it may just as well reflect misplaced emphases. There has been a strong tendency to picture the ribosome function as somehow defined by its protein components, which is partly responsible for the heavy emphasis placed upon characterization of ribosomal proteins, protein factors, etc. Ribosomal RNAs, on the other hand, have tended to be viewed as "structural," as providing a sort of scaffolding upon which to position the function-defining proteins. However, the fact (revealed first by nucleic acid hybridization [1]) that ribosomal RNA sequence is highly conserved phylogenetically suggests these molecules to be somewhat more than mere scaffolding. This has been borne out by a variety of studies, which provide evidence for the direct participation of 16S ribosomal RNA in messenger RNA selection [2-4], tRNA binding [5-7], ribosomal subunit association [8-10], and antibiotic sensitivity/resistance [11].

In any case, this matter will soon be resolved. The capacity to determine nucleic acid sequences easily is now at hand; for this reason alone a number of ribosomal RNAs will ultimately be sequenced. The availability of the sequences of 16S and 23S rRNA from E. coli [12-14] has made is possible to begin attacking the problem of ribosomal RNA secondary structure.

Experience with the small functional RNAs, 5SrRNA and tRNA, give an indication of the problems that will be encountered in determining secondary structure for the much larger ribosomal RNAs. In both of the former cases the secondary structure of the molecule was not established until comparative sequence data was employed [15, 16]. For the larger ribosomal RNAs, which contain a bewildering array of possible double stranded structures, the difficulties in sorting out the true secondary structural elements are far greater. As we shall see, it is unsafe to make the restrictive energetic assumptions that would reduce drastically the number of possible structures; in so doing one can eliminate some of the true helices. Furthermore, one is not justified at this stage in assuming that two different helices whose sequences overlap are mutually exclusive; it is possible that both exist but at different times in the ribosomal cycle. It may even be unsafe to assume—as our computer search does—that double helical structure is determined solely by the normal Watson-Crick base pairs and G-U pairs.

The limitations of a comparative approach to secondary structure are obvious: Only secondary structural elements that are themselves conserved will be detected by the method. In a strict sense the approach demonstrates only that a pairing constraint exists, not that actual physical pairing occurs. Nevertheless, it is unlikely that actual pairing does not occur in most if not all of these instances. In practical terms, comparative evidence is not sufficient to define the extent of a double helical element; it merely reveals its existence. Because of this and the possibilities that non-Watson-Crick pairs are involved in helical structure and that not all helices are present all of the time, it is useful to provide additional evidence for the actual physical presence of the helical structures in rRNA. Therefore, we have also studied the relative reactivity of selected residues in the rRNA to various chemical modifying reagents that respect secondary structure, and have catalyzed the susceptibility of various residues to nuclease attack. Once comparative evidence has established the existence of a helical element, these criteria can then define its extent, and so on.

The present study is confined to the 16S rRNA from the true bacteria (eubacteria). The comparative data base includes the E. coli sequence [12,13],

a partial sequence (about 85 percent complete) from Bacillus brevis [17] and the catalogs of T_1 RNAse-generated oligonucleotides from over one hundred true bacteria [ref. 18 and C. R. Woese et al. unpublished]. B. brevis was chosen because the gram-positive and gram-negative organisms represent the phylogenetic extremes of the bacteria. Chemical modification of C's by bisulfite, A's by m-chloroperbenzoic acid, and G's by glyoxal (in the free 16S RNA, assayed in terms of the T_1 and pancreatic nuclease oligomer catalogs) has been used as physical evidence of secondary structure and other constraints [19]. In addition, kethoxal reactivity of 16S rRNA in the active and inactive forms of the 30S subunit [9, 10, 20, 21], is taken as evidence for single stranded structure. Further evidence for exposed regions is provided by sensitivity of residues to nuclease attack, both in naked 16S RNA and in the ribosome [13, 22-26].

During the final stages of this work, the sequence of maize chloroplast 16S rRNA [27], human mitochondrial 12S rRNA [28], and a partial sequence for yeast 18S rRNA [29] became available. These sequences were important in establishing one of the long-range interactions (see below) and furthermore supply additional comparative evidence for a number of the helices in our proposed structure. In this paper, we have restricted our discussion to the E. coli, B. brevis and RNAse T_1 catalog sequence data.

By these methods we have identified a number of helical elements in 16S rRNA. Those meeting the comparative criteria are considered to be firmly established. The comparative approach in addition distinguishes among helices as to type and reveals features of non-helical regions as well.

Space limitations do not permit a complete reporting of our results in this journal. We present here a summary of the studies together with a few representative examples of the approach. A complete documentation will be published elsewhere.

MATERIALS AND METHODS

(a) Computer analysis

Diagonal arrays of potential helical regions were generated by a computer program that will be published elsewhere. Separate runs were made for local (pairing sequences < 120 nucleotides distant) or long range (> 100 nucleotides distanct) interactions. Only helices containing four or more base pairs and having stabilities at least of the order of those seen in tRNA survived screening at this stage.

(b) Chemical modification

The details of chemical modification studies have been or will be reported elsewhere. In brief they were done as follows: bisulfite, m-chloroperbenzoic acid, and glyoxal modifications were all done on the same preparation of [^{32}P]-labelled E. coli 16S rRNA. In the case of bisulfite substitution, the labelled RNA was mixed with 20 times its volume of 3.2 $NaHSO_3$ adjusted to a pH of 5.6. After incubation at room temperature for 6 or 16 hrs, the RNA was recovered by passage over G-25 Sephadex and alcohol precipitation, the bisulfite substitution being removed by treatment with 1M ammonia. It was then mixed with an appropriate amount of [^{3}H]-labelled E. coli 16S RNA (relative specific activity of the bases A:C:U:G = 4:2:2:1) and the mixture digested with T_1 or pancreatic ribonuclease and fingerprinted by the two dimensional DEAE cellulose procedure. Individual spots on the fingerprint were located, cut out and the ^{32}P:^{3}H ratio was determined. In most cases the spots were identified or sequenced by secondary endonuclease digestion procedures, and the specific activities of the resulting secondary digestion products also determined. In this way the fraction of an oligonucleotide not modified was accurately determined and the positions of the C residues that were modified (deaminated to U) in an oligonucleotidewere ascertained in most cases.

The same double label procedure was followed for A and G substitution reactions. M-chloroperbenzoic acid was used at a final concentration of 1 mg/ml, the reaction was incubated in a pH 6 buffer at 22°C for 1 hr. In this case, it was generally not possible to ascertain which of the A residues were modified in an oligonucleotide, merely that a certain fraction of the sequence remained unmodified. Glyoxal was used at 0.3% final concentration; the reaction was incubated at 37°C for 4 hr at pH 6. In T_1 RNAse digests the glyoxal-substituted G's are cut by the enzyme under the conditions used. However, glyoxal-substituted oligonucleotides are separated from their normal counterparts in both electrophoretic dimensions, by virtue of the fact that the product of T_1 RNAse cleavage of a substituted G residue is 2'-3' cyclic phosphate, not a 3' phosphate. In pancreatic RNAse fingerprints the glyoxal-modified oligonucleotides are only partially resolved from their normal counterparts, however, making the double label approach essential. The kethoxal substitution procedures are well documented [20].

(c) Comparative sequence analysis

The data base for the comparative analysis of 16S rRNA comprises the E. coli sequence [12, 13] and about 85% of the B. brevis (ATCC 8185) sequence

[17] as well as T_1 RNAse oligonucleotide catalogs for over 100 species of true bacteria [ref. 18 and C.R. Woese et al., unpublished]. These last include the following genera: Acetobacterium, Acholeplasma, Acinetobacter, Actinomyces, Aeromonas, Alcaligenes, Aphanocapsa, Arthrobacter, Azotobacter, Bacillus, Bdellovibrio, Bifidobacterium, Brevibacterium, Cellulomonas, Chlorobium, Chloroflexus, Chromatium, Clostridium, Corynebacterium, Dactylosporangium, Escherichia, Eubacterium, Geodermatophilus, Lactobacillus, Leptospira, Leuconostoc, Microbacterium, Micrococcus, Mycobacterium, Mycoplasma, Myxococcus, Nocardia, Paracoccus, Pasteurella, Pediococcus, Peptococcus, Photobacterium, Planococcus, Propionibacterium, Pseudomonas, Rhodomicrobium, Rhodopseudomonas, Rhodospirillum, Ruminococcus, Sphaerotilus, Spirillum, Spirochaeta, Spiroplasma, Staphylococcus, Streptococcus, Streptomyces, Synechococcus, Thermoactinomyces, Treponema, and Vibrio. Comparative analysis to deduce secondary structural elements is straightforward when extensive areas of sequences are compared. When oligonucleotide catalogs are used for this purpose, the procedure is as follows: Within a defined group of organisms, the various versions of a given oligoneucltotide are generally easy to recognize; the predominant version of the sequence is missing in one or a few species, and in these is replaced by a sequence that differs by one or a few changes from it, and the various versions exist in mutually exclusive relationship to one another. In this way "oligonucleotide families" are constructed, for various locales in the 16S rRNA sequence. Given the type of conservation patterns that hold for 16S rRNA (see below), it is often the case that phylogenetically a large T_1 oligonucleotide will be quite constant in sequence in one half and somewhat variable in the other. Such oligonucleotides are particularly useful in establishing secondary structure. To establish a secondary structural element, it is required that both strands are at least partially covered by oligonucleotides that can be identified as to position in the 16S rRNA. This effectively confines consideration to T_1 oligonucleotides of size 5 or more bases. [Most of the larger T_1 oligonucleotides from any catalog can be identified with reasonable certainty as fitting at one specific locale in either the E. coli or the B. brevis sequence, but many of the pentamers and a lesser fraction of hexamers cannot be unequivocally positioned.] When a large oligonucleotide is unequivocally established as defining one strand of a putative double stranded element, it is sometimes the case that a small oligonucleotide, e.g., a pentamer, is predicted to be present in the complementary strand. In these cases it is necessary to show that when sequence in the large oligonucleotide varies, the sequence of the

predicted smaller oligonucleotide correlates with this. [In other words, it is not sufficient merely to show that the predicted smaller sequence is present in a few cases; the covariation must be established.] Several examples of this will be seen below.

Sequence of the B. brevis 16S rRNA was determined largely by traditional approaches, although the rapid gel methods are now being used to complete it. The basic strategy has been this. Pieces of 16S rRNA in the size range of 30-150 bases were generated by partial nuclease digestion (pancreatic or T_1 nuclease). After isolation, by two dimensional polyacrylamide gel procedures, each fragment was characterized in terms of its T_1 RNAse catalog (G ending oligonucleotides), pancreatic RNAse catalog (pyrimidine ending), and U_2 RNAse catalog of glyoxal blocked RNA (A ending). This information was in most cases sufficient to give sequence unequivocally, given that most of the larger fragments also existed in abbreviated versions in the set of fragments sequenced.

Alternatively, fragments generated by partial RNAse A, RNAse T_1 or Naja oxiana RNAse [30] digestion were labeled at their 3' ends with 5'-[^{32}P]-labeled pCp (New England Nuclear, 2000 Ci/mmol) and separated by 2-dimensional gel electrophoresis [31]. Eluted spots were sequenced by the chemical method of Peattie [32].

RESULTS

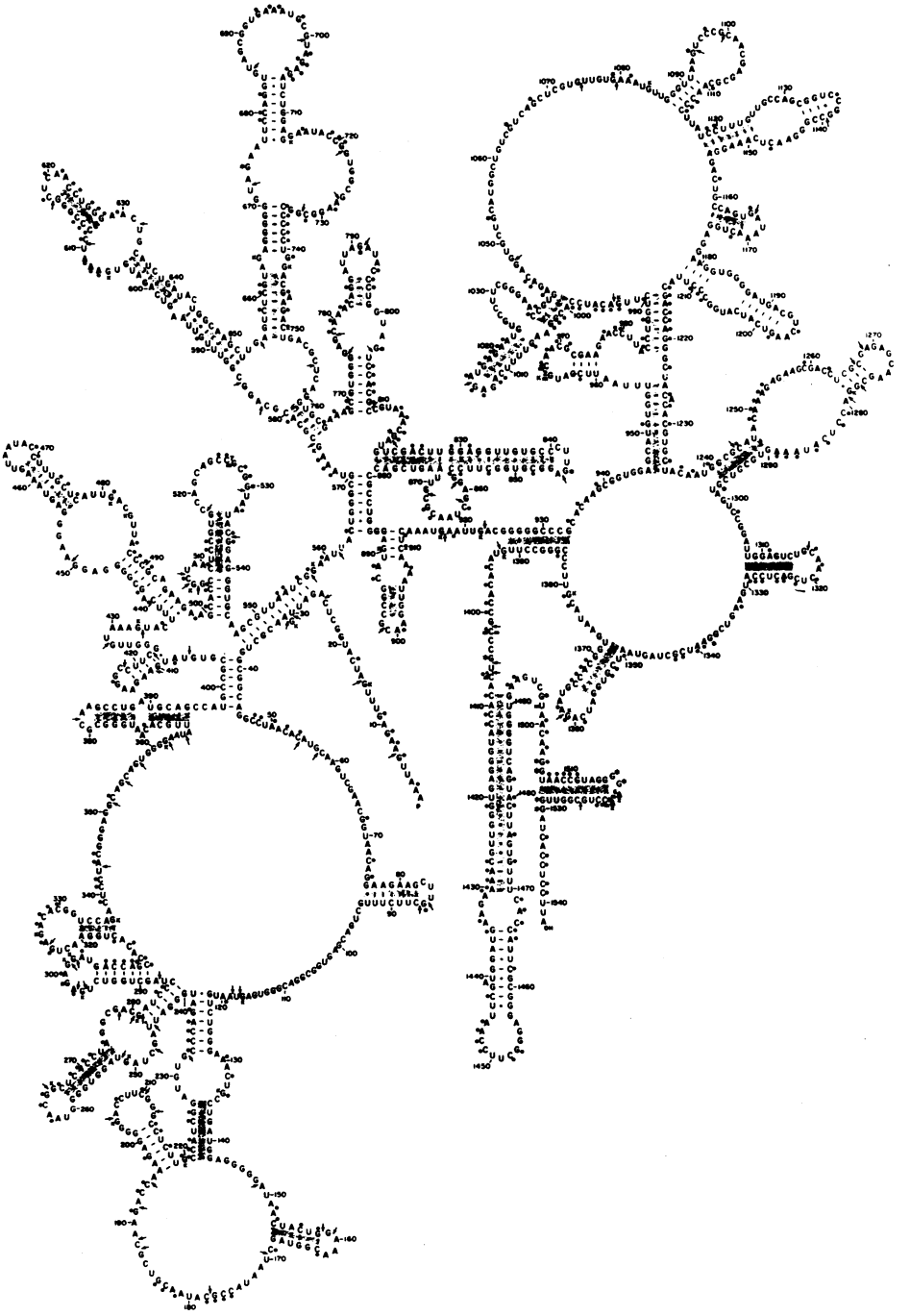
The following procedure is used to identify the secondary structural elements in 16S rRNA. A computer search generates a complete list of helices in the E. coli sequence that meet the simple and relatively non-restrictive conditions described above. A list of this sort is also generated for the B. brevis sequence to the extent this is known (about 85%). A comparison between the complete E. coli list and the incomplete B. brevis list identifies the helices common between the two and (by definition) eliminates those helices on either list that are not common. A helix is considered proven if it can be identified in both organisms, but does not have the same sequence in both cases. A combined list of "energetically likely" structures and structures common to the two organisms is also screened in terms of the 100 or so oligonucleotide catalogs for additional evidence for base pairing--either to substantiate further and define proven structures or to provide comparative evidence for helices that are possible but not established for double helical structure established by comparative criteria can then be adduced from the chemical modification experiments.

Figure 1 summarizes our findings to date; it contains fifty double helical elements. The thirty that are shaded are considered proven by the above comparative criteria. Because of the unavailability of comparative sequence data from B. brevis and RNAse T_1 catalogs bearing on the helix involving positions 27-37/547-556, the latter pairing interaction was chosen over other competing possibilities on the basis of the chloroplast 16S RNA [27], mitochondrial 12S RNA [28], and yeast 18S RNA [29] sequences. A detail discussion of the relation between the latter sequences and the structure of Figure 1 will be presented elsewhere. A preliminary version of this structure lacking the 564-570/880-886 interaction has been reviewed briefly [33]. In addition, the figure shows C, A, and G residues in 16S rRNA that are relatively susceptible (or resistant) to chemical modification [19], G residues that are susceptible to kethoxal modification in either the active or inactive 30S ribosomal subunit [9, 10, 20, 21], and points in the molecule readily cleaved by nucleases [22-26]. There is excellent agreement between these data and the secondary structure proposed: Where they have been measured, bases in double helical arrangements are almost all resistant to chemical modification or enzymatic attack; the points of enzyme attack or ready modification are located in non-double stranded regions, particularly in certain of the loops defined by helical stalks.

Fifty helices in 1542 residues amounts to one helix for every thirty bases; the helix density for 5S rRNA is the same as this (four in 120 bases [16]), while that for tRNA is about one helix per twenty bases [15]. Thus we feel that by far the majority of helices in the 16S RNA have now been discovered. Most of the remaining structures should be tertiary or quaternary. It is interesting if not unexpected that so many helices are conserved over the great phylogenetic distance represented by E. coli and B. brevis.

The following are five detailed examples of phylogenetically proven helices supported by chemical modification and nuclease cleavage data.

The structure of the region from residue 150 to residue 180 demonstrates the utility of a comparative approach both in proving and in disproving possible helices. Of the two mutually exclusive structures shown in Fig. 2, the upper one is calculated to be the more stable. However, only the lower one can be constructed from the B. brevis sequence. This is what we would call a variable sequence helix. As Table 1 shows, its sequence varies even within a single genus. In contrast, sequence in the areas surrounding the helix and in the loop defined by it are highly conserved phylogenetically. A number of other helices of this type are found in the true bacterial 16S



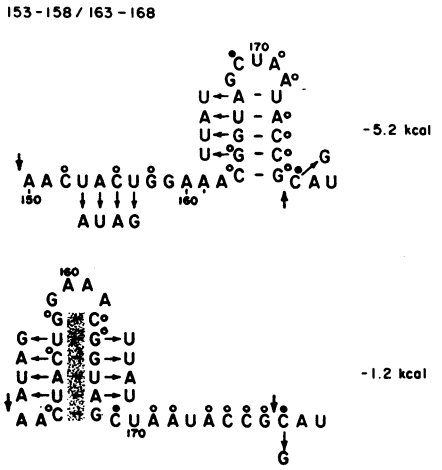


Figure 2

rRNA. Chemical modification data are in excellent agreement with the proposed structure; strong protection against bisulfite modification of C residues is seen for the T_1 oligonucleotides (in both E. coli and B. brevis) involved in the helix.

The upper helix shown in Fig. 3 (783-786/796-799) involves four pairs only. Its proof demonstrates the power of the comparative approach (see Table 2). As the figure shows, chemical modification and enzyme cleavage studies are in excellent agreement with the proposed structure: G_{791}, in the loop, is a major kethoxal-reactive site in active 30S ribosomes, and is also attacked by glyoxal and RNase T_1. RNAse U_2 cleaves at residues 792 and one or both of the sites at 787 and 790. All five of the residues in the helix proper whose reactivity could be tested are protected against chemical modification. Sequence in the helix (as well as the surrounding sequences) is highly conserved phylogenetically, unlike the previous example.

Figure 1. Secondary structure model for 16S rRNA. Shaded regions indicate helices supported by comparative phylogenetic sequence data from the B. brevis 16S RNA sequence [17] or from RNAse T_1 catalogs of 16S RNA from over 100 species of bacteria [18 and Woese et al., unpublished]. Filled circles indicate bases readily modified and open circles bases resistant to modification in naked 16S RNA [19]. Arrows show bonds readily cleaved by RNase A or T_1 [22-26]. Sites modified by kethoxal in active 30S subunits are designated by asterisks, and in inactive subunits by the letter K [9, 10].

Table 1 (153-158/163-168)

E. coli	A U A A C U A C U G G A A A C G G U A G C U A A
B. brevis[1]	A U A A C A U A G G G A A A A C U U A U G C U A A
B. alvei	A U A A C C C A C G,G A A A C G X X X G C U A A
B. polymixa	A U A A C U A C C G,G A A A C G X X X G C U A A
all other Bacillus[2]	A U A A C U U C G X G A A A C C G X X G C U A A
all Lactobacillus[3]	A U A A C A G G U G,G A A A C A G X X G C U A A

1 - AAACG (AAACCG) is found (missing) only in B. alvae, B. polymixa and B. acidocaldarius among twenty Bacillus species--expect for B. brevis which is missing both.

2 - Except B. acidocaldarius for which the analog of AUAA... has not been recognized.

3 - Except for L. viridescens, among ten species tested.

The two helices in the 820-880 region (Fig. 4), although not identical in B. brevis and E. coli, are obviously homologous. Few of the twenty corresponding base pairs in the two cases are identical. Strikingly, the upper helix contains five G-U pairs, four of them contiguous in the E. coli case! Frequent use of G-U pairs and even G-A juxtapositions is characteristic of many of the helices in the molecule, as Fig. 1 shows. The sequence variability seen in the helices proper in Figure 4 seems to extend even to the level of species within a genus (as oligonucleotide catalogs show). In contrast, the surrounding sequences (except for the small loop capping the structure) are highly constrained or conserved.

In both Escherichia and Bacillus the chemical modification data are con-

783-786 / 796-799

Figure 3

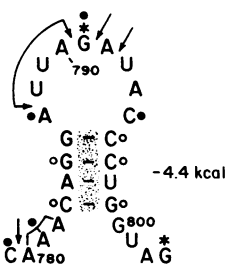


Table 2 (783-786/796-799)

E. coli	C A A A C A G G A U U A G A U A C C C U G
Acholeplasma	C A A A C A G X A U A C C C U G
other mycoplasmas[1]	C A A A U A G X A U A C C C U A G
Megasphaera	C A A A C G X X A U A C C C C G
Chloroflexus	C A A A C C G X A U A C C C G X

1 - This grouping also here includes the only two clostridia, C. innocuum and C. ramosum, that are known to be specific relatives of the mycoplasmas.

sistent with the proposed structure. Bases in the surrounding sequences, hairpin loop and bulge loop tend to be reactive, while in the helices proper they are not. [The B. brevis case affords two noteworthy examples of bisulfite reactivity. The sequence CACUCCG (covering position 875), is about 50% resistant to modification; however, the portion that does react with bisulfite is predominantly converted to CACUCUG, entirely consistent with the proposed structure. Also, the sequence UUUCAAUACCCUCAG is almost quantitatively converted to UUUUAAUACCCUCAG by bisulfite.]

It is interesting that an attractive potential helical structure in this region, 805-811/846-852, predicted to be much more stable than either

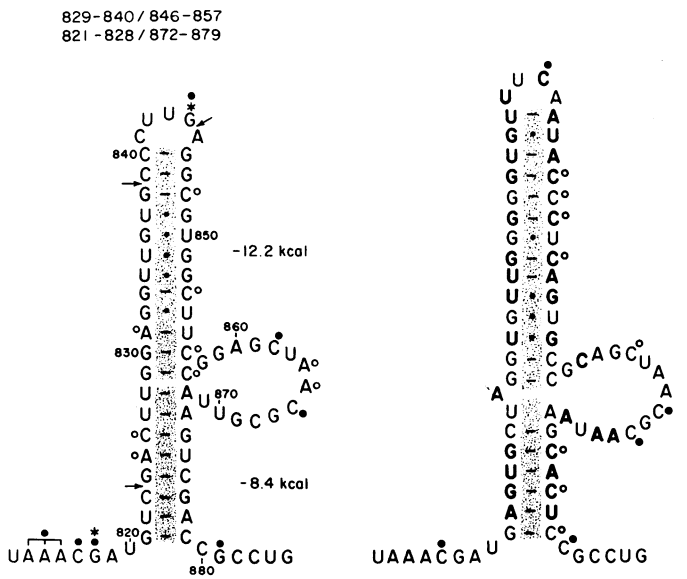
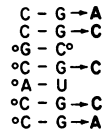


Figure 4

Figure 5



of the helices in Fig. 4 is disallowed by comparative evidence (Fig. 5).

The structure 946-955/1225-1235, Fig. 6, is particularly interesting for two reasons; it is a rather long range helix--its two strands are almost 300 residues distant--and it contains a proven example of a single base bulge, A_{1227}. Sequence in the structure and its surroundings tend to be highly conserved, making examples of base pair changes difficult to find. In all cases where the initial base of the oligonucleotide UUUAAUUCG changes (four independent examples), the corresponding residue in the sequence CUACACACG changes to an appropriate pairing complement. We consider the comparative evidence for this structure to be very strong. The single base A bulge is totally conserved phylogenetically within the true bacteria.

The final example, the structure in the region 1405-1495, is again interesting because of atypical base pairing. As Fig. 7 shows, comparative evidence definitely supports the lower of the two proposed helices. Some structure of this sort is also demanded by data from fragments of 16S rRNA produced by partial nuclease digestion. We have isolated a number of such fragments in the course of sequencing the B. brevis 16S rRNA [17]. They all begin at about position 1405 and end at about 1495, and most possess one or more internal cleavages in the 1430-1465 region. All run as an intact piece on the first (non-denaturing) dimension of the polyacrylamide gel separation system, but separate into two pieces in the second (urea-containing) dimension, implying strong secondary structure. The lower helix is noteworthy for the number of non-Watson-Crick pairings it contains. The E. coli version has only four G-U pairings but five A-G juxtapositions. In the chloroplast version,

Figure 6

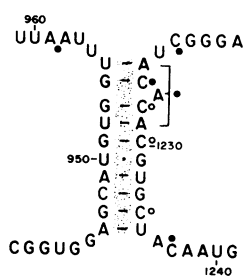
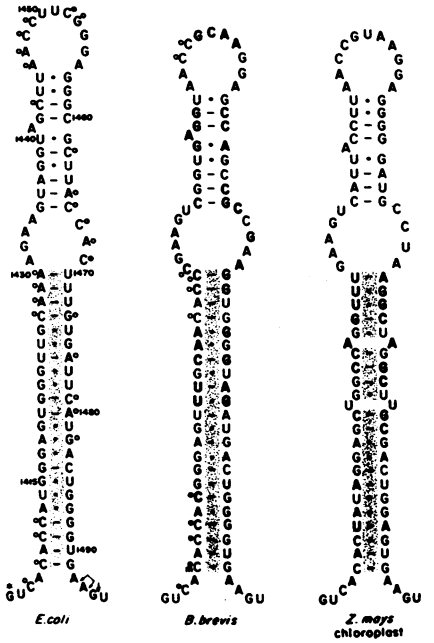


Figure 7



even A-A and U-U juxtapositions are encountered.

While chemical modification evidence is consistent with the proposed structure, all the protection observed is not accounted for by the proposed compound helical structure. For this reason in part we do not yet consider the upper of the two proposed helices to be proven. [It should also be noted that a conserved GAAGU sequence in the vicinity of position 1435 does not appear in the same relationship to secondary structure in all cases.]

DISCUSSION

A comparative analysis of RNA sequence detects far more than double helical configurations. It is not yet possible to interpret most of the conservation patterns and other constraints revealed by comparative analysis, but it is clear that they bespeak a 16S rRNA of subtle and intricate design: Helices are not simply double stranded structures; the 16S rRNA contains different classes of helical structures that are identified and distinguished to some extent by their compositions and more so by the manner in which these change phylogenetically. Non-helical regions too are under severe sequence constraints.

An RNA the size of the 16S presents an enormous number of possibilities for base paired structures of length four or larger--of the order of ten thousand such. We do not now possess the sophistication to distinguish the true helices among them through energy calculations. Many helices, some of which are discussed above, are judged stable if not preferred by energy calculation, but they fail to pass the comparative test and their bases tend not to be protected against chemical modification. However, the approach used in the present study unequivocally identifies helices; the same two regions of the 16S rRNA cannot reasonably exhibit a base pairing constraint in two organisms despite variations in sequence unless the regions are indeed physically paired at some point in the ribosome cycle.

The present study does not provide proof for all of the secondary structural elements in the 16S rRNA; it has, however, provided convincing evidence for many of these and revealed the existence of other constraints, as well. Thus, a feeling for the architecture of the molecule is beginning to emerge. Several general features are now apparent. The first of these is that the helical elements are of several different types. An unusual rather large type of helix is easily recognized. The helices involving positions 600, 830, and 1240 are proven examples. The structures in addition to being large have non-Watson-Crick pairing: G-U pairs are common; and potential A-G pairs occur (the yeast phenylalanine tRNA molecule has been shown to possess an A-G pair in the normal pairing mode [34, 35]). These appear phylogenetically interchangeable with G-U pairs (Fig. 6). Non-pairing juxtapositions also occur in these structures, as can single base bulges. The exact length (in base pairs) of any given helix of this type can vary somewhat phylogenetically. More striking is the phylogenetic variability in the sequence itself. Several examples exist in which well over half of the base pairs are different between the *E. coli* and *B. brevis* versions. Oligonucleotide cataloging shows this variability in some cases to extend even to species in the same genus.

These "variable sequence" helices could be involved in protein binding. A good example is seen at positions 587-604/634-652 in the molecule (Fig. 1); it is the putative binding site for protein S8 [24, 36]. If this be true, two characteristics of variable sequence helices--their unusual compositions and their extreme phylogenetic variability--might be explained. The irregularities in helical structure caused by the G-U and other unusual pairing arrangements could function as recognition signals for specific proteins. The sequence variability of these helices could be explained in terms of the protein's stabilizing effect *vis a vis* mutations that affect the helix; in

other words, a protein bound to the helix could stabilize a mutant in which a non-pair was created, alleviating what would otherwise be a deleterious condition; this in turn would facilitate the evolutionary replacement of base pairs.

In contrast to the variable sequence helices, helices of highly conserved sequence also occur. A good example is the 783-786/796-799 helix (Fig. 3); over 90% of the true bacteria examined exhibit the same primary structure for this helix. A potential helix, positions 960-963/972-975, whose primary structure is totally invariant among eubacteria is another possible example.

Several other characteristics of the helices in 16S rRNA deserve discussion. One is the occurrence of single base bulges in double stranded structures. So far we have encountered two convincing examples of this phenomenon, both involving single A residues, at positions 746 and 1227. Two other examples of single base bulges appear, at positions 31 and 1441. The bulged G at position 31 is seen to be uniquely accessible to chemical and enzymatic modification compared with bases in its immediate vicinity. The 5S RNAs of the archaebacteria often show a bulged-out A residue in the middle of the molecule's molecular stalk. These single base bulges would seemingly have functional significance; note the extreme phylogenetic conservation of A_{1227}.

A surprising general feature of ribosomal RNA architecture is the extreme constancy of sequence in non-paired regions of these molecules. There are numerous examples of sequences immediately preceding and succeeding double helical segments, that are highly conserved, far more so than are the paired sequences. A similar constraint seems to apply to the capping loops, bulge loops and interior loops defined by the helices. For example, the 821-879 structure is surrounded by highly conserved sequences. The preceding flanking sequence can be located unchanged in about 95% of eubacterial oligonucleotide catalogs; the few variations in this sequence that have been recorded are highly constrained--i.e., the same few have each arisen more than once. This sequence--GCCGUAAACGAUG--also seems universal in the archaebacteria [37], and a recognizable variant of it may occur in eucaryotic 18S rRNAs as well [C. R. Woese et al., unpublished results]. Interestingly, this sequence contains a kethoxal-reactive residue (G_{818}) that is known to be important for binding 50S subunits [9, 10]. Highly constrained though not universal sequence can be seen (by cataloging) in the bulge loop of the helix in question. In sharp contrast, the sequence in the helix proper is variable to the extent of changing within a single genus. These conserved, ostensibly single

stranded regions also tend to be exposed in the molecule; by far the vast majority of (identifiable) residues highly reactive with modifying reagents are located in such regions in the 16S rRNA structure.

It should be noted that the helices in the 16S rRNA tend to be "nested" into compound structures. Thirty-seven helices in the proposed structure are involved in such arrangements. Nesting is also seen in 5S rRNA [16], but not in tRNA (with the exception of its molecular stalk [15]).

Viewed in its entirety (Fig. 1) the 16S rRNA appears to be organized into domains, some of which are also defined in terms of ribosomal proteins. The initial domain, approximate positions 30-560, is structured in terms of two clusters of helices (120-335 and 365-545) all enclosed by the helix spanning positions 30 and 550. This unit is also structured by ribosomal protein S4 in that the protein protects the RNA from nuclease attack (with certain gaps) and a complex of the two is readily isolated [reviewed in ref. 38]. Several lines of evidence suggest this domain to be largely interior in the ribosome: It exhibits a higher overall inaccessibility to the chemical modifying reagents than would be expected on the basis of the assigned secondary structure, and protein S4 itself is not exposed on the surface of the 30S particle [42], as are nearly all the other 30S ribosomal proteins. Protein S20 also binds in this domain, presumably in the 240-285 region--a region absent from preparations of the S4-associated RNA unless S20 is also present during preparation [38]. A second domain is enclosed by the helix 564-570/880-886 (although we have not proven this helix, it is strongly suggested by studies on the S4-associated RNA fragment; whenever a preparation of this RNA fragment terminates at position 575 rather than 557, the additional fragment 819-887 accompanies it [22]). This domain too is involved in protein binding. The 587-605/633-652 helix binds ribosomal protein S8, while S15 is bound by the neighboring helix, 655-670/735-751 [24, 36]. It also seems likely that 821-840/845-879 binds ribosomal proteins S6 and/or S18 [38]. Moreover, the domain contains sites that are involved in 50S subunit association. The structure (position 671-734) nested by the S15 binding helix contains three G residues (at positions 674, 703, and 705) that are reactive in the 30S ribosome but unreactive in the 70S [9]. Modification of any of these sites reduces the ability of 30S subunits to form 70S couples [10]. Similarly, G positions 791, 803, and 818 also appear to be involved in 50S subunit interaction.

The next domain accounts for most of the second half of the 16S rRNA. It is defined by the long-range helix 926-933/1384-1391. There are two other

long range helices, 946-955/1225-1235 and 984-990/1215-1221, as well as a number of local helical elements in the unit. The 5' (wobble) base of the anticodon of tRNA can be photochemically crosslinked in or near this region of the 16S rRNA when it is located in the so-called "P site" of the ribosome [6]. Met_{114} of ribosomal protein S7 can be photochemically crosslinked to U_{1240} [40]. Most of the 30S subunit proteins implicated in tRNA binding [summarized in ref. 41] are dependent for assembling on protein S7 [42]--all of which makes it likely that this domain of the 16S rRNA is at very least involved in tRNA binding. One site in this region, G_{1166}, appears to make contact with the 50S ribosomal subunit [9, 10].

The terminal domain--from about position 1390 to the 3' terminus of the molecule--is coordinated about two helices. The last of these (1506-1529) has been recognized for some time; it contains the so-called "Kasugamycin sequence" (two adjacent dimethyladenosine residues in the loop)[11]. Sequence in the helix is highly constrained; recognizably similar versions of it are found in all three primary kingdoms [38, 43, 44]. In that kasugamycin has been implicated in f-met-tRNA binding [for a review, see ref. 45], this helix may somehow function in that respect. The two G residues at positions 1516 and 1517 are protected from kethoxal by 50S subunits [10]. 30S subunits lacking methylation of the two adenosine residues at positions 1518 and 1519 have reduced affinity for 50S subunits [46]. Thus the dimethyl A loop must also be central in 30S-50S subunit interaction.

The remaining helical unit in this domain is the variable sequence structure occupying positions 1408-1490. The outer, enclosing, helix comprises 22 base pairs and its sequence is somewhat variable phylogenetically; many G-U pairs are encountered, as are A-G juxtapositions. For these reasons we would suggest this helix to play a role in protein binding, either ribosomal or initiation factor.

The unpaired sequences in this domain appear important: They are all very highly conserved phylogenetically. All of them are readily accessible to chemical modifying agents. At least 3 additional G residues in this domain, at positions 1405, 1496, and 1504, make contact with the 50S subunit [9, 10]. The terminal sequence ACCUCCUUA is involved in the mRNA recognition process [2-4].

Recently, secondary structural elements for parts of the 16S RNA have been proposed by Schwarz and Kössel [27] and by Ross and Brimacombe [49]. Although there is substantial agreement between their structures and the corresponding sections of our model, many of the elements proposed by them differ signifi-

cantly from ours. It should be noted that our structure is consistent with the experimental evidence presented by the latter authors.

Nothing is now known of the overall shape of the 16S rRNA beyond the loose constraints forced upon it by the long range helices and by gross physical measurements of the ribosomal subunit. There are no "knots" [47] in the structure, a fact which may have important consequences for ribosome assembly. We do not yet have any feeling for transitions that occur within the structure during its function: Do some helices unwind and alternate helices form, or helical arms move relative to one another during this function? Or, alternatively, is all functional change of a more subtle, local nature, involving local shifting of bases within loops and other structures? What is the nature of the 16S-23S rRNA interaction; does it involve only the termini of the two molecules [48]--as has been suggested to prevent the formation of knots--or are numerous relatively short intermolecular helices formed, involving the single stranded segments in the loops and so on, of the two RNAs? Where precisely do the ribosomal proteins bind, and what is the manner of their binding? Most of the questions concerning the ribosome remain to be answered.

REFERENCES

1 Pace, N.R. (1973) Bacteriol. Rev. 37, 562-603
2 Shine, J. and Dalgarno, L. (1974) Proc. Nat. Acad. Sci. USA 71, 1342-1346
3 Steitz, J.A. and Jakes, K. (1975) Proc. Nat. Acad. Sci. USA 72, 4734-4738
4 Dunn, J.J., Buzash-Pollert, E. and Studier, W. F. (1978) Proc. Nat. Acad. Sci. USA 75, 2741-2745
5 Noller, H.F. and Chaires, J.B. (1972) Proc. Nat. Acad. Sci. USA 69, 3115-3118
6 Ofengand, J., Liou, R., Kohut, J., Schwartz, I. and Zimmermann, R.A. (1979) Biochemistry 18, 4322-4332
7 Prince, J. B., Hixson, S. S. and Zimmermann, R. A. (1979) J. Biol. Chem. 254, 4745-4749
8 Santer, M. and Shane, S. (1977) J. Bacteriol. 130, 900-910
9 Chapman, N. M. and Noller, H. F. (1977) J. Mol. Biol. 109, 131-149
10 Herr, W., Chapman, N. M. and Noller, (1979) J. Mol. Biol. 130, 433-449
11 Helser, T. L., Davies, J. E. and Dahlberg, J. E. (1972) Nature New Biol. 235, 6-8
12 Brosius, J., Palmer, M. L., Kennedy, P. J. and Noller, H. F. (1978) Proc. Nat. Acad. Sci. USA 75, 4801-4805
13 Carbon, P., Ehresmann, C., Ehresmann, B. and Ebel, J.-P. (1979) Eur. J. Biochem. 100, 399-410
14 Brosius, J., Dull, T. J. and Noller, H. F. (1980) Proc. Nat. Acad. Sci. USA 77, 201-204
15 Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R. and Zamir, A. (1965) Science 149, 1462-1465
16 Fox, G. and Woese, C. R. (1975) Nature 256, 505-506

17 Kop, J., Kopylov, A. M., Noller, H. F., Siegel, R., Gupta, R. and Woese, C. R., unpublished
18 Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T., Wolfe, R. S., Balch, W., Tanner, R., Magrum, L., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Chen, K. N. and Woese, C. R. (1980) submitted to Science
19 Magrum, L. J., Stahl, D. A. and Woese, C. R., unpublished result
20 Noller, H. F. (1974) Biochemistry 13, 4694-4703
21 Hogan, J. J. and Noller, H. F. (1978) Biochemistry 17, 587-593
22 Ehresmann, C., Stiegler, P., Carbon, P., Ungewickell, E. and Garrett, R. A. (1980) Eur. J. Biochem. 103, 439-446
23 Ehresmann, C., Stiegler, P., Fellner, P., and Ebel, J.-P. (1975) Biochimie 57, 711-748
24 Müller, R., Garrett, R. A. and Noller, H. F. (1979) J. Biol. Chem. 254, 3873-3878
25 Yuki, A. and Brimacombe, R. (1975) Eur. J. Biochem. 56, 23-34
26 Rinke, J., Yuki, A. and Brimacombe, R. (1976) Eur. J. Biochem. 64, 77-89
27 Schwarz, Z. and Kössel, H. (1980) Nature 283, 739-742
28 Eperon, I., personal communication
29 Rubtsov, P. M. and Skryabin, K. G. (1979) Dokl.Akad.Nauk USSR 248, 760-762
30 Vassilenko, S. K. and Ryte, V. C. (1975) Biokhimiya 40, 578-583
31 de Wachter, R. and Fiers, W. (1972) Anal. Biochem. 49, 184-197
32 Peattie, D. A. (1979) Proc. Natl. Acad. Sci. USA 76, 1760-1764
33 Noller, H. F. (1980) in Ribosomes, Chambliss, G. et al., Eds., Univ. Park Press, Baltimore
34 Rich, A. and RajBhandary, U. L. (1976) Ann. Rev. Biochem. 45, 805-860
35 Jack, A., Ladner, J. E. and Klug, A. (1976) J. Mol. Biol. 108, 619-649
36 Zimmermann, R. A. and Singh-Bergmann, K. (1979) Biochim. Biophys. Acta 563, 422-431
37 Fox, G. E., Magrum, L. J., Balch, W. E., Wolfe, R. S. and Woese, C. R. (1977) Proc. Nat. Acad. Sci. USA 74, 4537-4541
38 Zimmermann, R. A. (1980) in Ribosomes, Chambliss, G. et al., Eds., Univ. Park Press, Baltimore
39 Winkelmann, D. and Kahan, L. (1980) in Ribosomes, Chambliss, G. et al., Eds., Univ. Park Press, Baltimore
40 Möller, K., Zwieb, C. and Brimacombe, R. (1978) J. Mol. Biol. 126, 489-506
41 Shimizu, M. and Craven, G. R. (1976) Eur. J. Biochem. 61, 307-315
42 Nomura, M. and Held, W. A. (1974) in Ribosomes, Nomura, M., Tissières, A. and Lengyel, P., Eds., Cold Spring Harbor, N.Y.
43 Hagenbuchle, O., Santer, M., Steitz, J. A. and Mans, R. J. (1978) Cell 13, 551-563
44 Schwarz, Z. and Kössel, H. (1979) Nature 279, 520-522
45 Cundliffe, E. (1980) in Ribosomes, Chambliss, G. et al., Eds., University Park Press, Baltimore
46 Poldermans, B., Bakker, H., and Van Knippenberg, P. (1980) Nucleic Acids Res. 8, 143-151
47 Cantor, C. (1980) in Ribosomes, Chambliss, G. et al., Eds., University Park Press, Baltimore
48 Van Duin, J., Kurland, C. G., Dondon, J., Grunberg-Manago, M., Branlant, C., and Ebel, J.-P. (1976) FEBS Lett. 62, 111-114
49 Ross, A. and Brimacombe, R. (1979) Nature 281, 271-276