Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 1950–1954, March 1990
Biophysics

Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: A molecular mechanics approach

(ribosome/16S rRNA/energy minimization)

ARUN MALHOTRA, ROBERT K.-Z. TAN, AND STEPHEN C. HARVEY*

Department of Biochemistry, University of Alabama at Birmingham, Birmingham, AL 35294

Communicated by Peter H. von Hippel, December 20, 1989

ABSTRACT **We introduce a computer-assisted procedure for folding large RNA chains into three-dimensional conformations consistent with their secondary structure and other known experimental constraints. The RNA chain is modeled using pseudoatoms at different levels of detail—from a single pseudoatom per helix to a single pseudoatom for each nucleotide. A stepwise procedure is used, starting with a simple representation of the macromolecule that is refined and then extrapolated into higher resolution for further refinement. The procedure is capable of folding different random-walk chains by using energy minimization, allowing generation of a range of conformations consistent with given experimental data. We use this procedure to generate several possible conformations of the 16S RNA in the 30S ribosomal subunit of *Escherichia coli* by using secondary structure and the neutron-scattering map of the 21 proteins in the small subunit. The RNA chain is modeled using a single pseudoatom per helix. RNA–RNA and RNA–protein crosslinks, reported in current literature, are included in our model. Footprinting data for different ribosomal proteins in the 16S RNA are also used. Several conformations of the 16S RNA are generated and compared to predict gross structural features of the 30S subunit as well as to identify regions of the 16S RNA that cannot be well-defined with current experimental data.**

The ribosome is a complex ribonucleoprotein system responsible for translation of genetic information on mRNAs into polypeptides during protein synthesis. Because of the importance of this polymerase, it has been the subject of intense research for the past three decades. An important key to understanding the ribosomal mechanism lies in its three-dimensional structure. The smaller subunit of the *Escherichia coli* ribosome, the 30S subunit, is made up of 21 proteins and an RNA chain (16S) with 1542 nucleotides. With the availability of a complete map of the protein positions in the small subunit of the *E. coli* ribosome (1) and a large body of crosslinking (for summary, see ref. 2) and footprinting (for summary, see ref. 3) data, it is now possible to construct plausible three-dimensional models of the 30S subunit (2–5). Models for the structure of the 16S RNA chain in the protein framework of the 30S subunit have been built manually, using mechanical models (2) and interactive computer graphics (3, 5).

Manually built models of the 30S subunit incorporate a variety of data—secondary structure, crosslinking and base-protection patterns, protein maps, and results from immunoelectron microscopic analysis of the 30S subunit. Apart from being time consuming and laborious to develop, manually built models have several drawbacks. They present only a single conformation out of all the possible folding patterns that can satisfy experimental data. With such models it is difficult to quantitatively judge inconsistencies and conflicts in data as well as to take the inherent errors in experimental data into account. The model-building process can also introduce biases into mechanically constructed models, because the interior regions of such models are not easy to manipulate. In this paper we use an automated RNA folding procedure (6) that employs molecular mechanics techniques to build models of the 16S RNA in the *E. coli* 30S ribosomal subunit.

Molecular mechanics uses energy minimization and molecular dynamics to model a system of atoms (7). In a traditional all-atom model, each atom is represented by a point mass and potential energy functions are used to mimic bonds, bond angles and torsions, van der Waal interactions, electrostatics, and other constraints and forces among these atoms. Energy minimization is used to search for the global minimum energy structure in such models.

All-atom molecular modeling, though straightforward, is currently not possible for large molecules. Energy minimization is computationally demanding for macromolecules, and at present only molecules with up to a few thousand atoms can be modeled in full detail. Apart from the computational requirements, the complexity of the potential energy surface increases with the number of atoms, hindering any search for the global minimum.

The disadvantages of all-atom modeling for macromolecules can be partially overcome by reduced representations [also called succinct models (8)] where pseudoatoms are used to represent a set of related atoms with a generally invariant structure such as a nucleotide. Reduced representations were used in early simulations of protein folding (9, 10), though this approach was later abandoned because of the difficulties of treating long-range interactions, which are so important for proteins. Nucleic acids are more amenable to reduced representations, as the major forces stabilizing RNA structure are base-stacking and hydrogen-bonding, both of which are short-range interactions. Our procedure uses a reduced representation of an RNA chain to model the 16S RNA in the 30S subunit.

An Automated RNA Folding Procedure

We use several levels of reduced representations for the RNA chain in our models. The most intuitive of these is to represent each nucleotide as a single pseudoatom located at the phosphate group (P-atom) of the nucleotide (the all-phosphate representation). A coarser reduced representation (the 1H representation) involves using a single large pseudoatom (helix or H-atom) to represent each of the double-stranded helical stems and P-atoms are used for the single-stranded regions. An intermediate level of representation uses five (or more) pseudoatoms per helix (the 5H representation), with large central atom(s) for space filling and the four

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*To whom reprint requests should be addressed.

corner P-atoms in the helix to correctly orient the helix. In the 1H (and the 5H) representation, the radii of the H-pseudoatoms (or the large central pseudoatoms) are chosen to reflect the size of the helix that they represent. In all of these models, proteins are modeled as spheres, with the radii based on their anhydrous molecular weights.

Different secondary structure motifs—helices, bulges, loops, and single-stranded regions—are imposed on the pseudoatoms by using bonds, angles, and torsions, similar to the covalent bonds, angles, and torsions used in traditional all-atom molecular mechanics. Helical regions of the RNA chain are assumed to have an ideal A-RNA conformation (11), the most common form of RNA double-helix at low ionic strength (12). The double helix is imposed using bonds between the hydrogen-bonded base pairs and neighboring nucleotides. Angles are specified along the helical backbone and the correct chirality is maintained using torsions. Nucleotides in single-stranded regions are linked by bonds. Nucleotides in loop (hairpin and internal) regions are also linked by single bonds, though angles and torsions are applied to hairpin loops during initial minimization to mimic the helical characteristics of these loops. Helices with bulges are modeled as regular helices with the bulged nucleotide(s) attached to the neighboring bases by bonds and angles similar to those of a regular helix. For a single unpaired nucleotide, these constraints force the nucleotide out of the helix and cause a kink in the helix. For a bulge with more than one unpaired nucleotide, the system of constraints is undetermined and the unpaired nucleotides are free to take different conformations depending on interactions (such as van der Waals) with neighboring atoms. Colinear or stacked helices are represented by extending helical constraints to connect the two (or more) helical regions involved. Other experimental data such as crosslinks and close contacts (RNA–RNA or RNA–protein) are incorporated in the models by the use of appropriate bonds.

We use harmonic potential functions for bonds, angles, and torsions. The potential function for a bond is thus

$$E_{\beta i} = k_{\beta i}(\beta_i - \beta_{i0})^2,$$

where $E_{\beta i}$ denotes energy of ith bond, $k_{\beta i}$ is the bond force constant for the ith bond, β_i is the bond length, and β_{i0} is the equilibrium bond length. Similar expressions are used for angle and torsion potential functions. Harmonic functions are easy to minimize and have a single unique minimum. A harmonic bond potential function is also equivalent to a gaussian distribution of bond length β_i about β_{i0}, with a standard deviation equal to $RT/2k_{\beta i}$. Force constants can thus be calculated from the standard deviation expected about the equilibrium bond lengths, angles, or torsions. In our models, the force constants are chosen to mimic variability in the $tRNA^{Phe}$ crystal coordinates (13), one of the few well characterized RNA structures. For example, an examination of the helical regions of $tRNA^{Phe}$ shows that the separation of P-atoms in a given base pair has a standard deviation of about 1 Å. The force constant of the harmonic potential function for such bonds is thus 0.298 kcal/mol·Å^2 (1 cal = 4.184 J). Similar analysis is used for other structural features—a standard deviation of 0.6 Å for the phosphate–phosphate distances in single-stranded regions and a standard deviation of 0.2 radians for angles and torsions in helices (based on standard deviations seen in the corresponding regions of $tRNA^{Phe}$ crystal coordinates). Force constants for bonds representing experimental data are chosen to reflect experimental uncertainty. Crosslinking bond lengths used in modeling are based on the approximate distances spanned by the chemical crosslinker used. A standard deviation of 2 Å is assumed in the crosslinking length. RNA–protein close contacts are represented by bonds of 5 Å with a standard deviation of 2 Å.

Nonbond interactions are used to exclude volume occupied by the pseudoatoms. We use harmonic terms for nonbond interactions:

$$E_{\gamma ij} = k_{\gamma ij}(r_{ij} - r_{ij0})^2 \quad \text{if } r_{ij} \le r_{ij0}$$

$$= 0 \quad \text{if } r_{ij} > r_{ij0},$$

where $E_{\gamma ij}$ is the nonbond interaction energy between atoms i and j, $k_{\gamma ij}$ is the nonbond force constant for the atom pair ij, r_{ij} is the distance between atoms i and j, and r_{ij0} is the minimum distance (exclusion distance) allowed between the two atoms (usually the sum of their radii). The minimum separation of P-atoms in single-stranded regions of $tRNA^{Phe}$ (4.98 Å) is used to get an exclusion distance of 5 Å between P-pseudoatoms (equivalent to an exclusion radius of 2.5 Å for P-pseudoatoms). Exclusion distances between other pseudoatoms are equal to the sum of their radii.

The use of several different levels of reduced representation allows us to begin with a low-resolution 1H model of the RNA chain. The model is refined using energy minimization, and the resulting folded chain can then be extrapolated to a higher level model with more detail. Extrapolation between models is done using superimposition of ideal helices on the helical pseudoatoms (the radii of helical pseudoatoms in the 1H and the 5H models are chosen to accommodate an ideal helix). This allows modeling to proceed from a randomly oriented chain at the 1H representation, to the 5H representation, and then to the all-phosphate representation, to get a satisfactory conformation for the RNA chain backbone. In principle, this procedure can be used to extrapolate up to an all-atom model of the RNA chain; all-atom extrapolation rules are obvious for helical regions but are yet to be determined for other regions. For small RNA chains (≈100 nucleotides), such as tRNAs, modeling can be started directly at the all-phosphate representation. Fig. 1 shows an all-phosphate representation of $tRNA^{Phe}$ and illustrates folding of such a model into a three-dimensional structure. This RNA folding procedure has been tested with $tRNA^{Phe}$ by using data available before the x-ray crystal structure for tRNA (13) was determined. These tests (6) gave results consistent with the known structure of tRNAs and illustrated two aspects of our RNA folding procedure. (*i*) Our succinct models are underdetermined and several different conformations can be generated to satisfy the given secondary and tertiary constraints. (*ii*) Our procedure can be used as a powerful tool to test alternate tertiary constraints and to look for inconsistent data.

The 30S Subunit Model

Data Used. The modeling of the 16S RNA in the 30S ribosomal subunit is based on the secondary structure taken from Stern *et al.* (3), which was based on the 16S secondary structure proposed by Gutell *et al.* (14). Crosslinking data as summarized by Brimacombe *et al.* (2) is used. This consists of eight RNA–RNA crosslinks and 23 RNA–protein crosslinks. Of the reported crosslinks to the 3′ and the 5′ termini of the 16S RNA, only two have been used—crosslinks between protein S5 and the 5′ terminus and between S18 and the 3′ terminus. The 21 ribosomal proteins are modeled as spheres with radii taken as the anhydrous radii calculated from their molecular weights (15). The proteins were tethered with bonds to the positions reported in the neutron-scattering map of the small subunit (1). Force constants for these bonds are chosen to mimic the uncertainty in protein positions as reported by Capel *et al.* (1). Additionally all the 93 individual protein–protein distances used to gen-

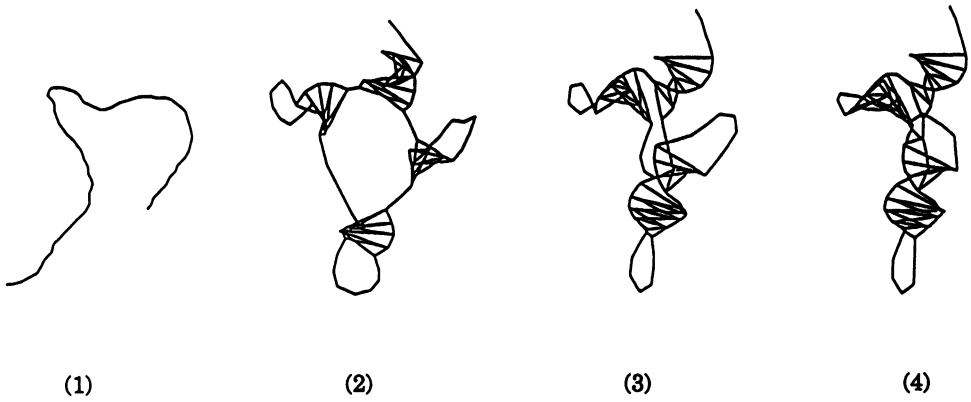

FIG. 1. Generation of a three-dimensional model for yeast $tRNA^{Phe}$ from a random chain. The above sequence shows a random chain (model 1) being folded to greater refinement as additional constraints are applied: secondary structure only (model 2), secondary structure and correct helix stacking (model 3), and secondary structure with correct helix stacking and contact between the D and T loops (model 4).

erate the neutron scattering map were included as bonds (with the reported uncertainty) to reflect asymmetry in the map's covariance matrix (1).

Footprinting data, summarized by Stern *et al.* (3), is also included in our models. This data identifies nucleotides in the 16S RNA protected from chemical and enzymatic attack on the addition of different 30S subunit ribosomal proteins. Regions of the RNA chain strongly protected from chemical attack by a protein (taken as protection of two or more neighboring bases) are assumed to be in close contact with that protein in our model, yielding a total of 63 close contacts. Footprinting data can also reflect allosteric effects (3). To diminish such effects only protection (and not reactivity enhancement) data was used and sites showing medium or weak protection or sites with protection on isolated nucleotides were ignored.

The crosslink and footprinting data used in our models was chosen to be similar to the data used to assemble the two manually built models of the 16S RNA (2, 3). This choice allows us to compare our results directly with the manually built models. Other tertiary data (for review, see ref. 4) is available for the 16S RNA and can be used in modeling studies.

Refinement Procedures. Starting structures are generated using random walk chains for the 485 pseudoatoms in the 1H model of the 16S RNA. The direction at each step of the walk is varied randomly between zero and a maximum specified angle, and the length of each step is adjusted according to the length of the bond connecting two neighboring pseudoatoms. Pseudoatoms representing the 21 proteins are given the coordinates reported by Capel *et al.* (1) and the geometric center of the RNA chain is superimposed on that of the proteins. Several chains with different random seeds and maximum angles (30°–90°) are used to get a variety of starting coordinates.

Energy minimization, using steepest descent and conjugate gradient methods, is used to refine the models. Five different starting structures were minimized to get refined 1H models. All models are initially minimized with soft nonbond interactions (16) (to give an energy of 1 kcal/mol for an overlap of 1 Å between two pseudoatoms) to permit the chain to pass through itself, allowing tangles in the starting structure to be resolved. The resulting structures are then further minimized with stiffer nonbond force constants (to give an energy of 100 kcal/mol for a 1-Å overlap between two pseudoatoms). Not all randomly generated initial structures can be untangled—these typically minimize to higher energies and are excluded from further refinement and analysis. To study the role of footprinting data (3) in defining the minimized structures, the refinement procedures were repeated without the footprinting data for the five chains.

Energy minimization was done with *yammp*, an in-house molecular mechanics package (17). Variability between the different conformations generated was examined by superimposing the refined 1H models (only the 21 protein positions are superimposed) and then calculating standard deviation of the coordinates for each nucleotide in the models. Typical time for minimizing a 1H model of the 16S RNA is about 2 hr of Cray-X/MP central processing unit time.

Modeling Results

The 1H model refinement brings different parts of the random starting structure into correct relative positions that satisfy the given secondary and tertiary structure data. Further refinement, at the 5H and all-phosphate representation, mostly changes the orientation of different parts in the chain, with no large changes in the overall shape and form of the model refined at the 1H level. Thus an analysis of different random chains refined at the 1H level can yield sufficient information about the positions of different parts of the 16S RNA chain in the minimized conformation. A typical model refined at the 1H level is shown in Fig. 2*A*.

The 16S RNA chain is not evenly constrained in our models as the complexity of the secondary structure is not uniform and the experimental footprinting and crosslinking data is unevenly spread through the chain. Thus, some parts of the RNA chain can be better characterized than others. An analysis of the variability between the different refined 1H models can be used to identify ill-defined and well-characterized regions of the RNA chain. Results of such an analysis are shown in Fig. 3 for the five refined 1H models. These five models were refined with and without the footprinting data, and the variability between the models is shown for each of the two cases. As expected, footprinting data provides crucial tertiary constraints and helps to reduce variability between the different conformations generated by our procedure. This analysis identifies four ill-defined regions in the 16S RNA that require more crosslinking or footprinting data to be properly positioned in a three-dimensional model of the 30S subunit. These regions are as follows: the 59–180 region, including the 66/103 stem and the 144/178 stem; the 437–497 stem and internal loop; the 997–1044 region; and the penultimate stem (1409–1491). These regions of the RNA chain took widely different positions in the refined conformations and very little structural prediction can be made about them.

All the 1H conformations share several characteristics that can yield important clues to the three-dimensional structure

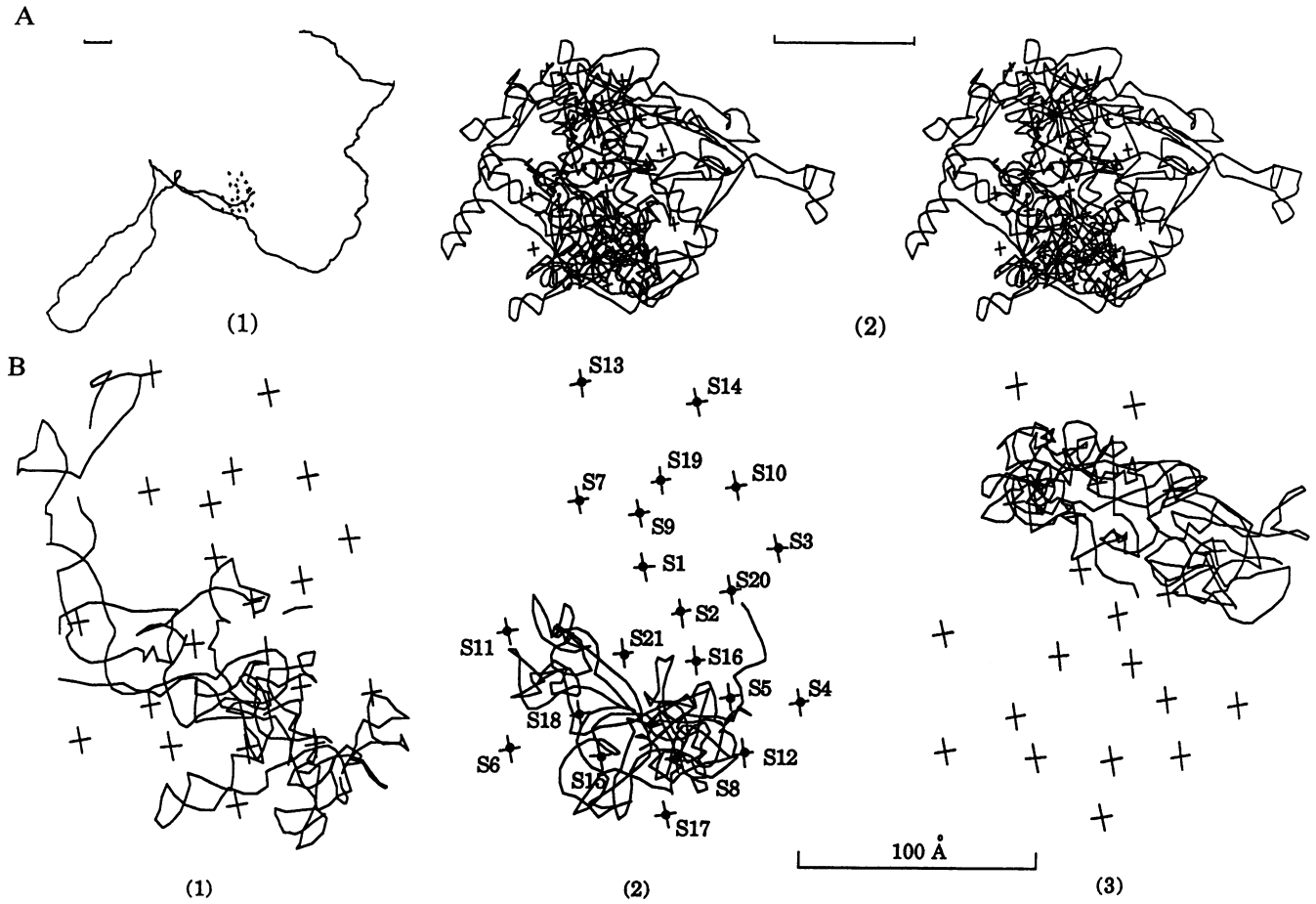

FIG. 2. (*A*) Typical 1H model of the 16S rRNA in the *E. coli* 30S ribosomal subunit: initial random walk chain (model 1) and refined 1H model (in stereo) viewed from the solvent interface (same orientation as figure 18 of ref. 3 and figure 16 of ref. 1) (model 2). The RNA backbone is shown with the protein centers indicated by the crosses (protein numbering is shown in *B*). (Scale bars = 100 Å.) (*B*) Major domains of the model shown in *A* after further refinement at the all-phosphate representation, as seen from the solvent interface: the 5′ and the pseudoknot region (model 1), the central domain (model 2), and the 3′ domain (model 3). Only well-defined regions are shown (nucleotides 1–8, 14–16, 26, 66–103, 144–178, 437–497, 916–920, 991–1045, and 1398–1542 are not shown). The RNA backbone is shown with the protein centers indicated by the crosses.

of the 30S subunit. All our models have physical dimensions similar to the 30S subunit dimensions observed using electron microscopy—230 Å high, 140 Å wide, and 115 Å thick, as reported by Lake *et al.* (18). We also see a distinct separation of the 16S RNA into two major structural units. The positions of several regions of the RNA chain—especially the

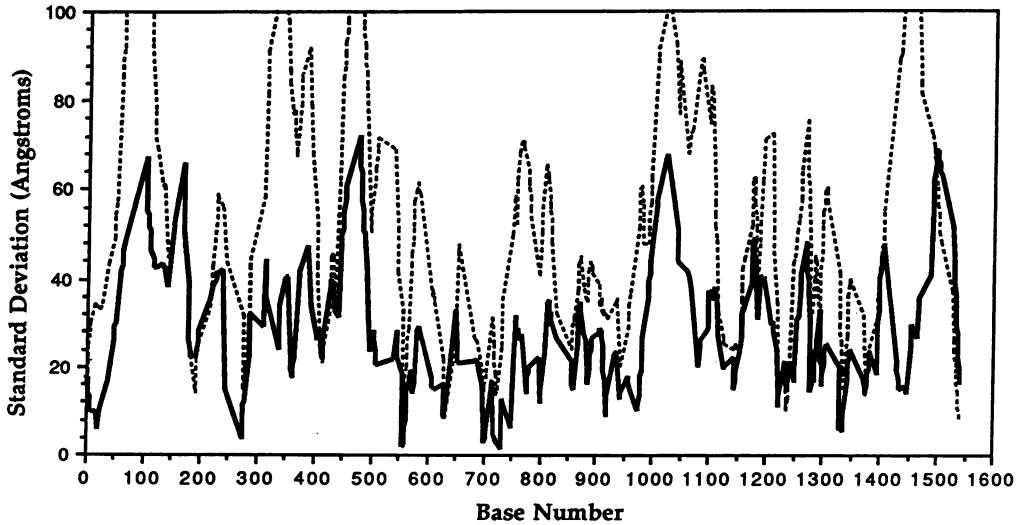

FIG. 3. Variation (measured in terms of standard deviation) between coordinates of pseudoatoms in five refined 1H models of the 16S rRNA. Solid line, variation between 1H models minimized with footprinting data; dotted line, variations between 1H models minimized in absence of footprinting data.

pseudoknot region and the central domain regions—are similar in all the conformations generated so far, indicating that the data are sufficient to define these features relatively well. As shown in Fig. 2*B*, when viewed from the cytoplasmic or the solvent side, the pseudoknot region of the 16S RNA lies in the lower part of the subunit extending from above protein S16 to S4 and S5. The 5′ domain (residues 37–546) region extends mostly from below S3 and S20 to S17 on the 50S interface side. The anomalous S13 and 189–191 link (2), which is included in our model, causes part of the 5′ domain to fold up toward S13. The orientation of that part of the 5′ domain (residues 136–227) is not well-defined in our models. The central domain (residues 557–918) is seen in the region of proteins S6, S21, and S8, and folds compactly into the bottom half of the 30S subunit in all our models. This part of the RNA chain is relatively well-defined. The top part or head of the 30S subunit, in our 1H models, is almost exclusively made up of the 3′ domain (residues 920–1396).

The 21 protein positions display an rms deviation of 4–6 Å, when our 1H models are compared to the neutron-scattering map. This is well within the range of standard errors in the protein map (1). Bonds used to tether the proteins to the protein map coordinates are very weak (because of higher experimental uncertainty), compared to the secondary structure and crosslink bonds, and so individual proteins would be easily displaced in our model if their positions were inconsistent with other constraints. Thus at the 1H level, the protein-map data appears to have no serious inconsistencies. Crosslinking and base-protection data, on the other hand, do have some inconsistencies. As described earlier, all the reported crosslinks to the 5′ and the 3′ termini cannot be simultaneously satisfied—we used only one link to each terminus. Several other crosslinks and contacts are not satisfied completely in our models.

Conclusions and Discussion

The 1H model of 16S RNA in the 30S subunit of *E. coli* ribosome is presented. The 1H model is a low-resolution model that can be used to look at the relative positions of different regions in the RNA chain.

Many possible conformations can be generated to satisfy the currently available structural data on the 16S RNA. The positions of several regions of the RNA chain relative to the major domains are similar in all the conformations generated so far, indicating that these features are reasonably well defined. The overall features of our model are similar to those of the interactive graphics model of 16S rRNA manually built by Stern *et al.* (3). Their model was based on data similar to that used for our study and it shows comparable positioning of the pseudoknot region and the three major domains of the RNA chain.

The procedure introduced in this paper is thus an effective method for folding large RNA chains for a given set of constraints. Apart from yielding information about the three-dimensional structure of the RNA chain (for a large enough set of constraints), this procedure can also be used to identify ill-defined regions of an RNA chain, to check experimental data for consistency, and to suggest which experiments would provide the most information for further limiting the range of acceptable models.

Superimposition algorithms were implemented by Martin Jones. This work was supported by a grant from the National Science Foundation (DMB-87-06551). Computer time was provided by the Alabama Supercomputer Network. Additional computer support was provided by the Atherosclerosis Research Unit, University of Alabama at Birmingham.

1. Capel, M. S., Kjeldgaard, M., Engelman, D. M. & Moore, P. B. (1988) *J. Mol. Biol.* **200,** 65–87.
2. Brimacombe, R., Atmadja, J., Stiege, W. & Schuler, D. (1988) *J. Mol. Biol.* **199,** 115–136.
3. Stern, S., Bryn, W. & Noller, H. F. (1988) *J. Mol. Biol.* **204,** 447–481.
4. Brimacombe, R. (1988) *Biochemistry* **27,** 4207–4214.
5. Nagano, K., Harel, M. & Takezawa, M. (1988) *J. Theor. Biol.* **134,** 199–256.
6. Malhotra, A., Tan, R. K.-Z. & Harvey, S. C. (1990) in *Molecular Dynamics: An Overview of Applications in Molecular Biology*, ed. Goodfellow, J. M. (Macmillan, London), in press.
7. McCammon, J. A. & Harvey, S. C. (1987) *Dynamics of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, England), p. 40.
8. Tan, R. K.-Z. & Harvey, S. C. (1989) in *Theoretical Biochemistry and Molecular Biophysics*, eds. Lavery, R. & Beveridge, D. (Adenine, New York), in press.
9. Levitt, M. & Warshel, A. (1975) *Nature (London)* **253,** 694–698.
10. Levitt, M. (1976) *J. Mol. Biol.* **104,** 59–107.
11. Arnott, S., Campbell Smith, P. J. & Chandrasekaran, R. (1976) in *Handbook of Biochemistry and Molecular Biology*, ed. Fasman, G. D. (Chemical Rubber, Cleveland), pp. 411–422.
12. Saenger, W. (1984) *Principles of Nucleic Acid Structure* (Springer, New York), p. 242.
13. Hingerty, B., Brown, R. S. & Jack, A. (1978) *J. Mol. Biol.* **124,** 523–534.
14. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985) *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 155–216.
15. Wittmann-Liebold, B. (1986) in *Structure, Function, and Genetics of Ribosomes*, eds. Hardesty, B. & Kramer, G. (Springer, New York), pp. 326–361.
16. Levitt, M. (1983) *J. Mol. Biol.* **170,** 723–764.
17. Tan, R. K.-Z. & Harvey, S. C. (1989) *J. Mol. Biol.* **205,** 573–591.
18. Lake, J. A., Sabatini, D. D. & Nonomura, Y. (1974) in *Ribosomes*, Monograph Series, eds. Nomura, M., Tissieres, A. & Lengyel, P. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), pp. 543–557.