Modeling Large RNAs and Ribonucleoprotein Particles using Molecular Mechanics Techniques

Arun Malhotra, Robert K.-Z. Tan, and Stephen C. Harvey
Department of Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35294 USA

ABSTRACT There is a growing body of low-resolution structural data that can be utilized to devise structural models for large RNAs and ribonucleoproteins. These models are routinely built manually. We introduce an automated refinement protocol to utilize such data for building low-resolution three-dimensional models using the tools of molecular mechanics. In addition to specifying the positions of each nucleotide, the protocol provides quantitative estimates of the uncertainties in those positions, i.e., the resolution of the model. In typical applications, the resolution of the models is about 10–20 Å. Our method uses reduced representations and allows us to refine three-dimensional structures of systems as big as the 16S and 23S ribosomal RNAs, which are about one to two orders of magnitude larger than nucleic acids that can be examined by traditional all-atom modeling methods. Nonatomic resolution structural data—secondary structure, chemical cross-links, chemical and enzymatic footprinting patterns, protein positions, solvent accessibility, and so on—are combined with known motifs in RNA structure to predict low-resolution models of large RNAs. These structural constraints are imposed on the RNA chain using molecular mechanics-type potential functions with parameters based on the quality of experimental data. Surface potential functions are used to incorporate shape and positional data from electron microscopy image reconstruction experiments into our models. The structures are optimized using techniques of energy refinement to get RNA folding patterns. In addition to providing a consensus model, the method finds the range of models consistent with the data, which allows quantitative evaluation of the resolution of the model. The method also identifies conflicts in the experimental data. Although our protocol is aimed at much larger RNAs, we illustrate these techniques using the tRNA structure as an example and test-bed.

INTRODUCTION

The past decade has seen a rapid increase in our understanding of the role of large RNAs and ribonucleoprotein (RNP) particles (particles with RNA and proteins) in biological processes. These processes include translation (the ribosome), transcription and RNA processing (the small nuclear RNPs and the hetero-nuclear RNPs, ribozymes, etc.), and the translocation of proteins (the signal recognition particle).

Elucidation of the three-dimensional (3-D) folding of large RNAs can provide important insights into the functioning of RNPs and RNAs. Unfortunately, our understanding of the 3-D structure of large RNAs is lagging far behind that of other macromolecular systems. High-resolution structures are available only for small RNAs such as oligomers (Dock-Bregeon et al., 1989; Happ et al., 1988), several tRNAs (Hingerty et al., 1978; Sussman et al., 1978; Schevitz et al., 1979; Woo et al., 1980; Westhof et al., 1985; Basavappa and Sigler, 1991) and some RNA loops and helices (Cheong et al., 1990; Holbrook et al., 1991; Heus and Pardi, 1991). High-resolution structural techniques such as x-ray crystallography and NMR are currently not capable of handling systems as large as most RNP particles, although some initial progress is being made on crystallization of the ribosomal subunits (Yonath et al., 1990).

There is a wealth of low-resolution structural data available for several RNAs and RNPs, most notably for the ribosomal RNAs (reviewed by Brimacombe, 1988; Moore, 1988). These include results from prediction of secondary structure and some tertiary interactions based on phylogenetic studies, cross-linking and footprinting experiments, chemical accessibility, electron microscopy, mutational studies and so on, which can contribute valuable insights for building 3-D models of RNA folding. Such data have formed the basis of several manually built models for the small subunit of the *E. coli* ribosome (Expert-Bezançon and Wollenzien, 1985; Nagano et al., 1988; Brimacombe et al., 1988; Stern et al., 1988; Oakes et al., 1990a). These were developed by manipulating ideal RNA helices either physically or on a computer screen.

Manually building models of RNA folding is an excellent and intuitive approach for putting together very low-resolution structural data into a coherent model. It allows the experimenter to express his or her "feel" for the RNA chain in a concrete form. Such models have a long and colorful history of success, starting with the classic models of DNA built in the 1950s (Watson and Crick, 1953). There are, however, serious limitations to building models manually. Manual models can present only one or a few conformations, and are often restricted in the conformational space that they sample by choices made early on in the building procedure. They rarely provide reliable estimates on resolution. It is also difficult in such models to incorporate new data. Most important, with the increasing size of systems and data used, they become harder to build, manipulate, and revise.

This paper presents an automated molecular mechanics protocol for using low-resolution structural data, along with

Received for publication 17 June 1993 and in final form 25 February 1994.
Address reprint requests to Stephen C. Harvey, Department of Biochemistry, UAB Station 552 BHSP, The University of Alabama at Birmingham, 1918 University Boulevard, Birmingham, AL 35294-0005. Tel.: 205-934-5028/4753; Fax: 205-975-2547; E-mail: harvey@neptune.cmc.uab.edu.

0006-3495/94/06/1777/19 $2.00

known motifs in RNA structure, to propose 3-D models for the folding of large RNAs. (Preliminary reports of this procedure and its application to 16S RNA have appeared in Malhotra et al., 1990, 1991).

Several different computer-based approaches to the problem of 3-D RNA folding are being pursued by other researchers (Malhotra et al., 1993). These include methods based on distance geometry (Hubbard and Hearst, 1991a, b), distance matrix (Hadwiger and Fox, 1991), and conformational searching with constraint satisfaction (Major et al., 1991; Gautheret et al., 1993). Only one of these has been applied to the 16S ribosomal RNA (Hubbard and Hearst, 1991a), and, in contrast with our models (Malhotra and Harvey, 1994), those models did not include the ribosomal proteins. The distance geometry approach is handicapped by the problems of stereoisomers that cannot be distinguished with distance constraints alone. Additionally, conventional distance geometry approaches are not suitable for under-determined systems (Metzler and Hare, 1989), and we wish to tackle problems where the number of degrees of freedom exceeds the number of available experimental constraints. Exhaustive conformational searching approaches (Gautheret et al., 1993) are a promising answer to the problems of sampling conformational space but have been demonstrated only for small RNA systems where the conformational space can be severely pruned by restricting allowed conformations.

Our approach is derived from the techniques of molecular mechanics (McCammon and Harvey, 1987). In molecular mechanics, atoms, or groups of atoms, are represented by point masses whose positions are specified in 3-D space. A potential energy function (sometimes called the force field) describes the conformational dependence of the energy. In all-atom models, this includes terms from covalent interactions (bond lengths, bond angles, and torsions) and noncovalent interactions (van der Waals, electrostatics, and sometimes, hydrogen bonding). This allows the computer quantitatively to compare different model conformations, by measuring the differences in energy between them. Conformational searching methods such as energy minimization, molecular dynamics, and Monte Carlo can be used to optimize model structures or examine the pathways of conformational transitions.

It is important to distinguish between de novo modeling, where the potential function contains only information like that described in the previous paragraph, and structure refinement. Molecular mechanics is widely used to refine structural models based on experimental data, particularly data from x-ray crystallography and high-resolution NMR. In NMR, for instance, one can determine distances between specified pairs of protons. A good model is one that reproduces all of those distances, so terms are added to the potential function to penalize differences between experimental and model distances (Brünger et al., 1987). Refinement of the structure is done by optimizing the model with this special potential function. More recent NMR refinement methods directly compare the observed spectrum with that predicted by the model, rather than only using distances derived from the spectra (reviewed by James, 1991). Similarly, structures from x-ray crystallography are refined by optimizing models using a potential function containing the usual intramolecular energy terms plus terms that penalize differences between the observed diffraction pattern and the pattern predicted by the model (Brünger, 1990).

Our protocol is a refinement method. RNA folding patterns are not based on energetics of conventional potential functions, or even functions such as contact profiles, packing, surface area, etc. Instead, we use molecular mechanics in our protocol as a tool to convert existing experimental data into a coherent 3-D model. The results of such modeling, therefore, are primarily dependent on the experimental data that are used, rather than the potentials or the representations in the model. The protocol presented here is thus a procedure for structure refinement using low-resolution experimental data, rather than an attempt at de novo modeling.

One problem that we faced at the onset was the question of what level of detail to include in the models. All-atom models would be highly desirable, of course, but they would also be extremely demanding from a computational viewpoint for systems as large as the ribosome. An alternate approach is to use reduced or succinct representations (Tan and Harvey, 1990; Malhotra et al., 1993), where pseudoatoms are used to represent groups of atoms. Such representations were first used in the modeling of polypeptides and other polymers (reviewed by Flory, 1969), nucleic acids (Olson and Flory, 1972; Schellman, 1974), and proteins (Levitt and Warshel, 1975; Levitt, 1976). Similar approaches have been used for the modeling of supercoiled DNA (Tan and Harvey, 1989) where three pseudoatoms are used to represent each base pair. Vorobjev (1990a, b) has also proposed a block-unit method for studying nucleic acid chain conformations where three "blocks" are used per nucleotide (for the phosphate, ribose, and the base). Lattice approaches to folding of proteins (Covell and Jernigan, 1990; Skolnick and Kolinski, 1990; Crippen, 1991; Hinds and Levitt, 1992) and RNA (Lustig et al., 1992) also implicitly use succinct models. The use of succinct representations is appropriate for the modeling of large systems because, apart from scaling down the problem to a manageable size, these pseudoatoms also more accurately represent the level of detail appropriate for low-resolution models. For example, in our current understanding of the ribosome, experimental data provide insights on the positions of individual helices or nucleotides, so modeling should reflect this level of detail. Modeling of the ribosome in atomic detail is not yet appropriate, given the data available.

With de novo all-atom methods, the form of the potential function and the parameters are generally chosen with the goal of mimicking the physics of interatomic interactions. The additional terms needed for refinement methods have a different purpose, however. They are designed to force the models to match experimental data. The functional form of these terms is dictated by the nature of the data, and the

values of the parameters (force constants) reflect the uncertainties in the data and/or the weights of the data relative to other terms in the potential function.

We should emphasize that the low-resolution models produced by this procedure can be converted into all-atom models in regions of particular interest. For example, we have been collaborating in modeling studies on RNase P and a preliminary set of low-resolution models has been developed (Harris et al., 1993). Efforts are now underway to develop all-atom models for part of that structure, using sets of distances from the low-resolution models as input constraints for MC-SYM, an exhaustive conformational search algorithm (Major et al., 1991; Gautheret et al., 1993).

THEORY

Models and pseudoatoms

The absence of a starting structure, and the underdetermined nature of large RNAs make it necessary to start the structure refinement procedure with as few degrees of freedom (or atoms) as possible. As the refinement progresses, the level of detail (and the number of atoms) can be increased. To achieve this, several types of reduced representations are used that vary in the amount of detail or the resolution of the model. There are correspondingly several different types of pseudoatoms with different radii based on the size of the RNA chain that they represent. Models with the lowest detail use a single pseudoatom for each helical region in the RNA chain. These pseudoatoms are called 1H atoms, and the corresponding models are called 1H models. In models with a little more detail, each helical region is represented by five or more pseudoatoms (the 5H model). The highest resolution models in our protocol (the all-P models) use a single pseudoatom (the P pseudoatom) for each nucleotide in the RNA chain. This scheme is illustrated in Fig. 1, which shows the pseudoatoms used to model $tRNA^{Phe}$ (Fig. 1 *a*) in each of the three resolution levels (Fig. 1 *b*). Pseudobonds connect successive pseudoatoms together.

The 1H models use P pseudoatoms to represent nucleotides in nonhelical regions of the RNA chain. Hairpin loops are included with the hairpin stem, and each helical stem is modeled by a single 1H-atom. The radius of each 1H-atom is computed to be equal to the distance of the nucleotide farthest from the geometric center of an ideal A-RNA helix with the same number of base pairs as the helical stem. Duplexes with bulges of three or less nucleotides are treated as regular ideal helices at this resolution level. The 1H models use a very rough representation of helical segments because a sphere is used to approximate the cylindrical shape of each helix. This approximation, adequate because of the very low-resolution of the 1H models, is necessary because most molecular mechanics algorithms are aimed at spherical particles. A more accurate representation follows when the 1H model is replaced by a 5H model.

The 5H model uses five or more pseudoatoms to correctly orient helices. Nucleotides at each of the four ends of a duplex are represented by P-atoms, and these are connected to P-atoms in neighboring strands and helices as appropriate. Either one or three additional large pseudoatoms are used to fill the body of the helix. Duplexes with eight base pairs or more are longer than the diameter of an ideal A-RNA helix. For such helical stems, a single atom at the geometric center of the helix cannot fill the volume of the helix (because the space-filling atom can at the maximum have its diameter equal to the diameter of the helix), and so three space-filling atoms are used along the helix axis. In Fig. 1b, the D stem, the T stem, and the anticodon stem are each assigned three central space-filling atoms as their respective hairpin loops are included as a part of the helix. As in the 1H models, single-stranded nucleotides are represented by P-atoms, and duplexes with bulged nucleotides are modeled as ideal helices.

In all-P models each nucleotide, including those in duplexes, is represented by a single pseudoatom (P pseudoatom) placed at the phosphate atom of the nucleotide. The phosphate group, rather than the center of mass of each nucleotide, is used for pseudoatom positioning because this is independent of RNA sequence and allows easy representation of the RNA chain backbone. Thus, RNA chains of any sequence can be modeled as a string of identical P pseudoatoms. In such a representation the center of helical regions remains hollow, which would allow helices to interpenetrate. To prevent this, additional space filling atoms (X-atoms) are placed along the helix axis. The radius and placement of these X-atoms are chosen to prevent helix interpenetration while allowing for helix-helix interactions, such as helix stacking, to take place. Based on an analysis of ideal A-RNA helices and the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978), each X-atom is given a radius of 10 Å. The diameter used for the P-atoms (5 Å) is also based on the $tRNA^{Phe}$ crystal structure where the closest approach between all P-atoms pairs is approximately 5 Å (phosphate groups of residue 8 and 9 are separated by 4.98 Å. The closest approach among nonadjacent nucleotides is between 48 and 50 (5.12 Å)). Similar closest approach distances are also seen in other tRNA crystal structures.

Conversions between models of different resolutions

In our protocol, RNA chains are first modeled at the 1H level of detail. Starting with a random chain, these models are optimized to correctly position different helical regions with respect to each other based on the secondary structure and the experimental constraints imposed on the RNA chain. This low-resolution 1H model is then extrapolated to a 5H model, which serves as a starting structure for refinement at that level of detail. Refined 5H models are extrapolated to provide starting structures for modeling the RNA chain as an all-P model.

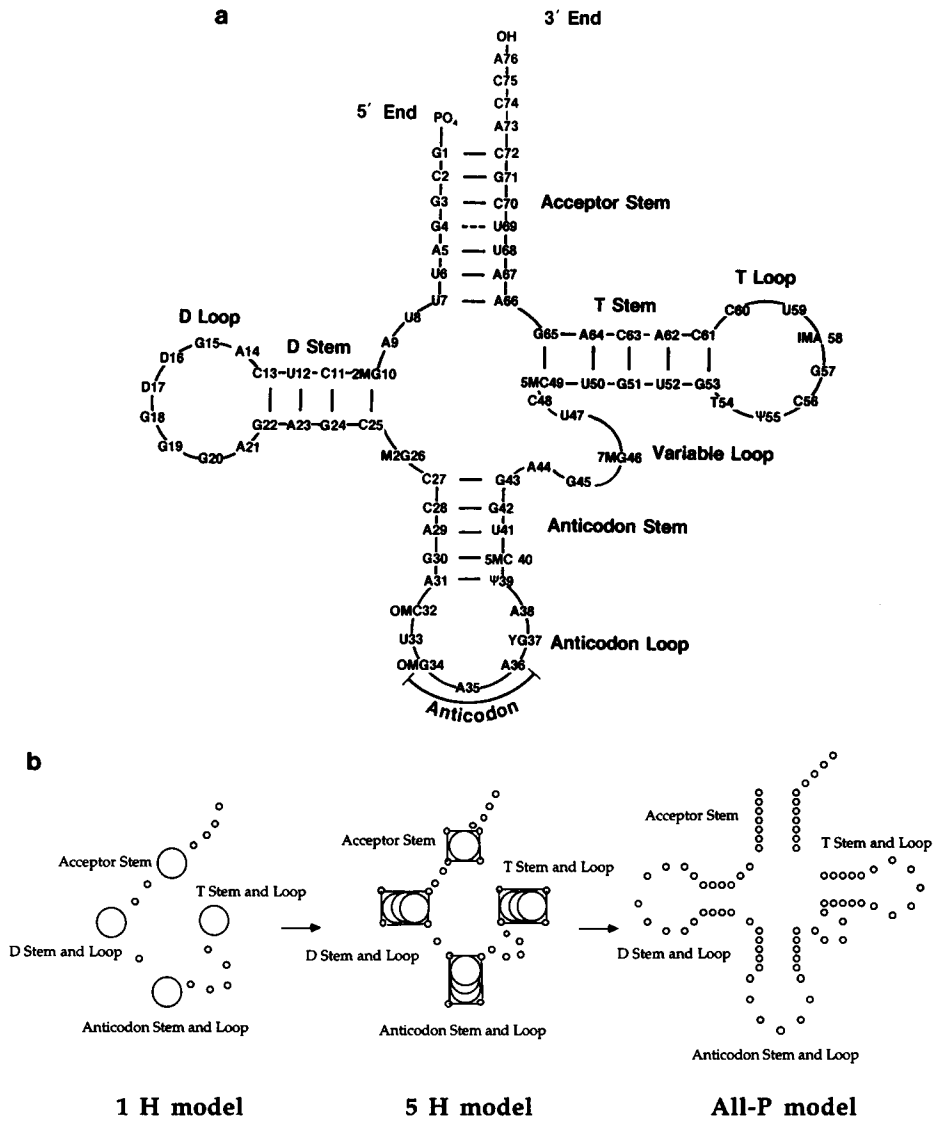

FIGURE 1 (*a*) Secondary structure of $tRNA^{Phe}$, and (*b*) Schematic representation of the 1H, 5H, and the all-P models for $tRNA^{Phe}$. Pseudoatoms are shown as circles. The small spheres represent P pseudoatoms. The large spheres in the 1H model represent the 1H atoms. The large spheres in the 5H model are used for space-filling helical stems. Hairpin loops are modeled as a part of the helix stem in the 1H and the 5H models. Space-filling pseudoatoms used in the all-P model (see text) are not shown for clarity. The pseudoatoms are not drawn to scale.

In the extrapolation to 5H models, the relative positions and orientation of the four corner P-atoms and the center space-filling atom(s) are computed for each helical stem based on the geometry of an ideal A-RNA helix of that length and total twist angle. The center of this set of atoms is then superposed on the center of the corresponding 1H atom in the 1H model. Because helical stems have directionality in the 5H models, additional information is required to properly orient the set of atoms. The positions of the nucleotides and/or helices neighboring the 1H atom in a given helix are used to get a reasonable starting orientation for that helix in the 5H model. For example, in the schematic of the 5H model of tRNA shown in Fig. 1 *b*, the orientation of the five atom set representing the acceptor stem is based on the positions of the P-atoms representing nucleotides 8, 73, and the 1H atom representing the T stem-loop in the 1H model. Single-stranded regions are treated similarly in both models, and no conversion is necessary for them.

The 5H models are extrapolated to all-P models by superimposing the center of an all-P duplex model of appropriate length on the central space-filling atom(s) for each helical region in the RNA chain; the orientation of the ideal A-RNA helix is specified by superposing the four P-atoms at the duplex vertices on the corresponding P-atoms of the 5H model. In addition, the hairpin loops are introduced into the models as an extended string of P-atoms, with structures as described below. Single-stranded regions of the RNA chain are transferred from the 5H model to the all-P model without any conversion.

Representation of RNA secondary structure features

Secondary structure motifs are imposed on the RNA chain using pseudobonds, angles, and improper torsions that enforce a particular 3-D structure. The different RNA structural motifs (reviewed recently by Chastain and Tinoco, 1991) and the corresponding constraints are discussed below; the functional form and parameters of these constraints are discussed in a later section.

RNA helices

The most important motif in RNA structure is the helix. Contiguous base-paired nucleotides adopt an A-RNA helical conformation as has been observed in all high-resolution RNA duplex structures (Arnott et al., 1976; Dock-Bregeon et al., 1989) as well as in the tRNA crystal structures (Hingerty et al., 1978; Sussman et al., 1978; Schevitz et al., 1979; Woo et al., 1980; Westhof et al., 1985; Basavappa and Sigler, 1991). This general invariance in the conformation of RNA helices prompted the use of 1H pseudoatoms to represent complete helices in the 1H models. In the 5H models, the four P-atoms at the vertices of the helix and the central atom are held together by eight bonds and one torsion in a conformation corresponding to the helix that they represent (Fig. 2 *a*). The torsion angle is required to guarantee proper chirality. For this purpose, a harmonic restoring force, usually called an "improper torsion," is used. For longer helices that require three central atoms, an additional five bonds and a valence angle of 180° are used (Fig. 2 *b*). The parameters of the bonds, angles, and torsions are determined by the length of the double helix. In the all-P models, a set of five bonds, two angles, and one torsion is used for each base pair in the helix stem (Fig. 2 *c*) (except at the helix termini). The angle constraints are imposed on each nucleotide triplet along the two backbones that form the duplex. Space-filling pseudoatoms (not shown) are anchored at the geometric center of each base pair using two bonds and one angle of 180° to the neighboring nucleotides.

The number of constraints necessary to define the conformation of N atoms in 3-D space is 3N-6. Each atom has three degrees of freedom, but the six degrees of freedom corresponding to rigid body rotation and translation do not alter the conformation. In our description of RNA helices, the number of constraints is larger than 3N-6 to increase the rate of convergence towards the desired structure. However, in tests with folding of long RNA chains into ideal RNA helices, it was observed that this over-specification was not enough to ensure a linear helical conformation after energy minimization. For example, a sample of five random chains representing a 30-base-pair RNA helix was refined to a potential energy gradient of less than 0.001 kcal/mol Å. The bonds, angles, and improper constraints described above were imposed on the all-P model chains with 60 P-atoms. The parameters and functional form of these constraints are described in a later section of this paper. Of these five random chains, only one adopted a linear global helix axis after simple energy minimization (Fig. 3 *a* shows one typical nonlinear conformation). An examination of several of these structures revealed that the nonlinear helix axis is a result of small (often less than a tenth of a percent) deviations from the specified bond lengths, angles, and torsions. This clearly illustrates the difficulty in ensuring large-scale structural integrity in models based solely on short-range constraints. In the use of short-range NMR data in distance geometry, an enmeshed network of distance constraints is necessary for long-range structural integrity (Hare and Reid, 1986). To get around the lack of overlapping constraints in our models, long-range constraints (two bonds, two angles, and one torsion) are used between the ends of duplexes to ensure that an extended linear helix can be rapidly achieved by energy refinement. If we label the P pseudoatoms at the 5′ and the 3′ ends of one strand of the duplex as i and j respectively, and if the corresponding P pseudoatoms on the other strand are designated k and l, these bond are i-j and k-l, the angles are i-j-k and j-k-l, and the improper torsion is a harmonic potential involving the torsion angle defined by the atoms i-j-k-l to ensure that the helical strands have the correct net twist. Ideal values of these constraints for a particular duplex are deduced from an ideal A-RNA helix of the same size. For helices longer than 20 base pairs, such long-range constraints are also imposed every 10 base pairs. Fig. 3 *b* shows the same random chain as in Fig. 3 *a*, refined with these long-range constraints.

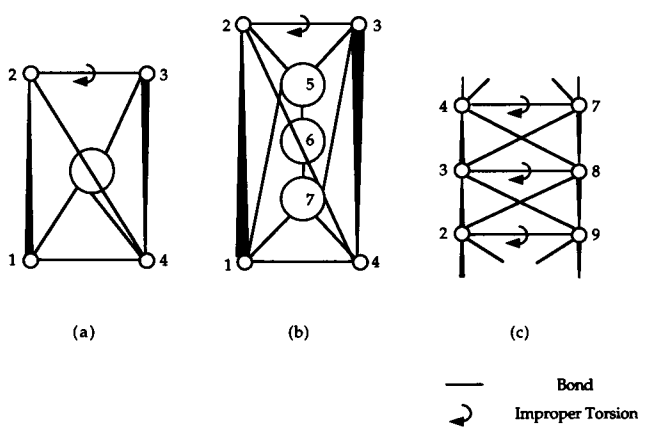

FIGURE 2 Pseudobonds and improper torsion constraints used to impose ideal A-RNA geometry on duplexes in the 5H model with (*a*) one central space-filling atom and (*b*) three central space-filling atoms, and (*c*) the all-P models. Improper torsions in the 5H models are specified by atoms 1-2-3-4. In the all-P model, improper torsions are 3-4-7-8, 2-3-8-9, etc. Long 5H model helix stems requiring three central space-filling atoms use one pseudoangle constraint (specified by atoms 5-6-7) of 180°. Pseudoangle constraints are also used in the all-P model for all atoms (except at the 5′ and the 3′ termini) along the two RNA strands forming the duplex, and these angles are specified by atoms 1-2-3, 2-3-4, 3-4-5, 6-7-8, 7-8-9, 8-9-10, and so on.

RNA helices with bulged nucleotides

Several studies have focused on the conformation adopted by bulged nucleotides in nucleic acids and their effects on the overall structure of duplexes. Unfortunately, there are no simple rules, because extra nucleotides may be stacked in the double helix or looped out into the solution, depending on sequence and solvent conditions (Morden and Maskos, 1993; Morden et al., 1990; Kalnik et al., 1990; Roy et al., 1987; Joshua-Tor et al., 1988; Miller et al., 1988; van den Hoogen et al., 1988). It is generally agreed that bulged nucleotides kink or bend the RNA helix axis or provide a point of some flexibility in the helix axis (Bhattacharyya and Lilley, 1989; Hsieh and Griffith, 1989; Bhattacharyya et al., 1990; Tang

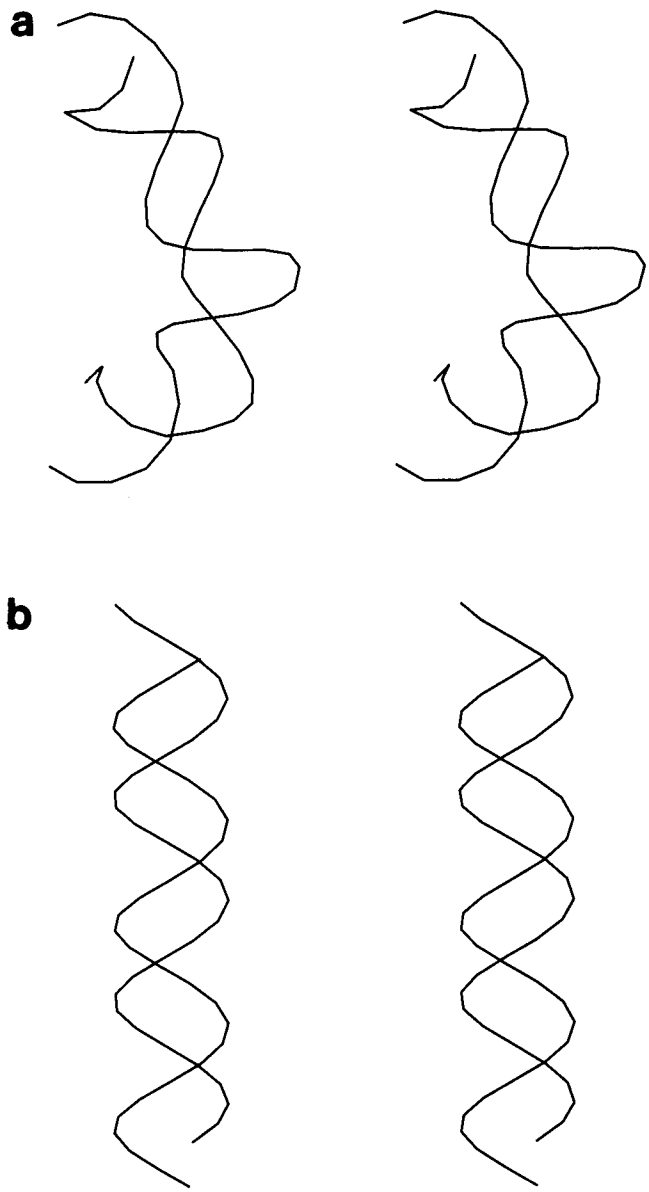

FIGURE 3 Structure of a long RNA duplex with 30 base-pairs after energy refinement (*a*) with only short-range constraints (see Fig. 2 *c*), and (*b*) with both short and long range constraints (see text), starting from a random conformation.

and Draper, 1990). To reflect this, the constraints used in all-P models for regular helices are modified at the bulge site to introduce a small kink in the helix axis for bulges with one nucleotide and to introduce flexibility for larger bulged loops. These conformational choices, although not rigorous, are appropriate to the nonatomic resolution of our models. For a bulge with one nucleotide, we add two bonds and an angle to constrain the bulged nucleotide (Fig. 4 *a*). This arrangement kinks the helix axis by about 20° at the bulge, similar to the bending observed by Woodson and Crothers (1988) and Rice and Crothers (1989) for single nucleotide bulges in DNA duplexes. For a bulge with two nucleotides, two additional bonds are used (Fig. 4 *b*), which allows one degree of freedom, giving the helix axis flexibility to bend at the bulge site. Larger bulged loops, with three or more nucleotides, are modeled as single strands that are looped out of the helix with no distortion in the helix axis. Helices with bulges are partitioned into separate helical segments for the assignment of long-range constraints; this allows for kinking at the bulge while keeping straight the two halves of the helix beyond the bulge. This is the default approach to modeling bulges, but the user can incorporate other specific distance and angular constraints if experimental information on the geometry is known, or if the user chooses to model a different geometry than that provided by the default settings.

Helix stacking

Helix stacking is an important feature in several RNAs, most notably the tRNAs where the D arm and the anticodon stem are stacked on each other, as are the T arm and the acceptor stem. In our protocol, helix stacking is imposed explicitly in all-P models by extending the constraints that are used within a helix to the interface between the two stacked helices. Additional long-range constraints are used between the two helices to ensure co-axiality. In 1H models a single bond is used to connect the 1H atoms corresponding to the stacked helices. In 5H models, stacking is imposed using two bonds and two angles between the sets of pseudoatoms representing the two helices. We rely on experimental or phylogenetic data to suggest stacking in particular RNAs. Thus, our protocol does not assume co-axiality between helical stems unless explicitly specified by the modeler.

Hairpin loops

At the low-resolution of our models, the precise atomic structure of RNA hairpin loops is not important to the global fold of large RNAs. In the 5H models, loops are considered as part of the stem structure, and the number and position of the large space-filling atoms is chosen accordingly. In the all-P models, to decrease computational complexity, rather than sample conformational space for hairpin loops, we impose geometries appropriate for RNA hairpin loops. High-resolution structures for RNA loops containing four nucleotides have been derived using NMR by several research groups (Cheong et al., 1990; Heus and Pardi, 1991). In addition, the tRNA crystal structures provide conformations for larger loops. RNA loops are characterized by extensive stacking and extension of the A form of the helix into the loop. Such observations have also led to theoretical predictions about stacking in RNA loops (Hasnoot et al., 1988). These high-resolution structures, and some all-atom modeling, were used to create a library of RNA loop structures that are used in our all-P models. The conformation for the four nucleotide loops in our models is based on the NMR structure of the UUGC loop (Cheong et al., 1990). All-atom models of loops with five and six nucleotides were constructed using the procedure described by Harvey et al. (1988) and refined with the *JUMNA 3e* nucleic acid modeling program (Lavery, 1987). The five-nucleotide loop was refined from an initial

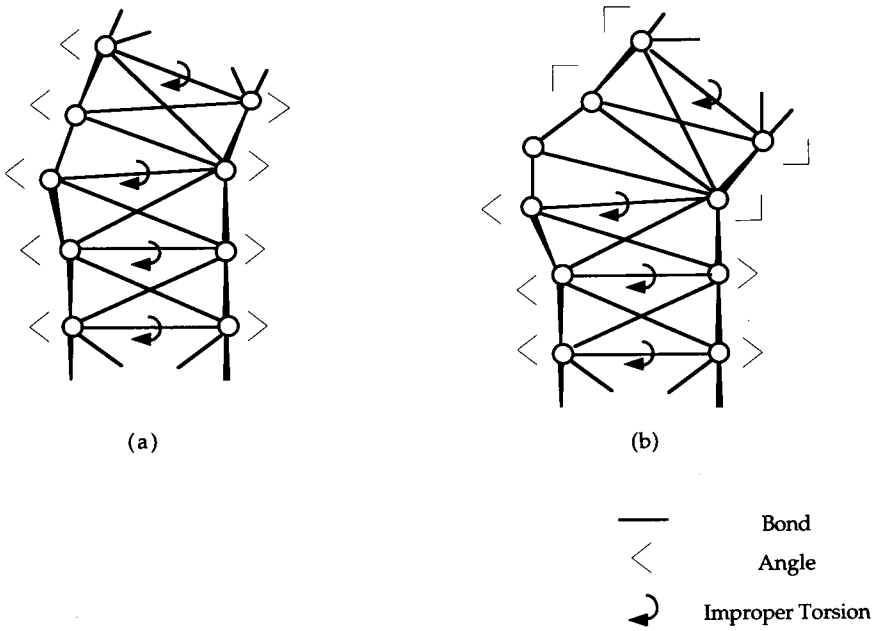

FIGURE 4 Pseudobonds, angles, and improper torsion constraints used to impose A-RNA geometry on duplexes (all-P models) (*a*) with one bulged nucleotide, and (*b*) with two bulged nucleotides.

structure with two bases stacked on the 5′ end of the helix stem and 3 bases stacked on the 3′ end. Starting structure for the lowest energy six-nucleotide loop stacked all six nucleotides on the 3′ end of the helix stem. Seven membered loops used in our models are based on the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978). These all-atom loops were converted into all-P models by placing P pseudoatoms at the phosphate positions along the RNA backbone and adding appropriate bond, angle, and torsion constraints to enforce the chosen ideal conformations from this loop library. Hairpin loops larger than seven nucleotides are "trimmed" by extending the stem.

Internal loops

Base pair mismatches have been observed in several high-resolution studies to be included in RNA stems without substantially distorting the A-form of the helix. These studies include the $tRNA^{Phe}$ and $tRNA^{Asp}$ crystal structures, which include GU wobble pairs and the NMR structure of GCGAUU(UCUG)CCCGCC, which has a six-base-pair stem with two mismatches (Puglisi et al., 1990b). Similar dodecamers have also been examined by x-ray crystallography (Holbrook et al., 1991). Based on these results, our models assume that mismatches do not distort the RNA helix backbone.

Much less is known about larger internal loops (reviewed by Chastain and Tinoco, 1991). Several studies on the 5S RNA helix 3 (Zhang and Moore, 1989; Varani et al., 1989) have shown that the symmetrical internal loop E exhibits only minor distortions from the A-form. Symmetry (or lack of it) and the sequence of internal loops are also known to be important in determining thermodynamic stability (Peritz et al., 1991), and presumably structure. For example, the asymmetric loop E has a structure quite different compared to its symmetric mutant form (Wimberly et al., 1993). In the absence of a clear structural understanding of large internal loops, we treat these as unstructured single-stranded RNA strands. In our protocol, the modeler can specify geometries of internal loops if desired; this includes incorporating internal loops into helical stems when supported by experimental data.

Pseudoknots

The high-resolution structure of only one pseudoknot has been reported in the literature (Puglisi et al., 1990a), and it indicates co-axial stacking between the two stem regions involved. In the absence of studies with larger systems, the pseudoknot loops in our models are treated as unstructured single-stranded RNAs. No co-axiality of helices is assumed, although this can be specified by the modeler.

Single-stranded RNAs

Very little is known about the structure of stretches of unpaired single-stranded RNAs. Accordingly, single-stranded RNA is left unstructured in our models. Bonds are used to maintain connectivity along the chain. The phosphate-phosphate distance along an RNA backbone depends on the backbone and sugar conformation, and varies between 6 and 7 Å (Saenger, 1984). An examination of the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978) also reveals that the average phosphate-phosphate distance in the nonstem regions is about 6 Å; this distance is used for the bonds between P-atoms in single-stranded regions of the RNA chain.

It is difficult to mimic the conformational flexibility of an RNA chain in our protocol because we represent each nucleotide by a single spherical particle. The use of only bonds and a single pseudoatom representation of the nucleotide allows

the RNA chain backbone to kink much more than is physically possible. The backbone torsions in a real RNA chain have well defined ranges, and this allows only some bending modes. Using a simple hard-sphere model for atoms, the sterically allowed and disallowed ranges of backbone torsions have been calculated by several researchers (reviewed by Olson, 1982). Using these backbone torsion ranges, and assuming that g+g− and g−g+ torsion pairs are disallowed for adjoining atoms, we examined the range of angles for successive triplets of neighboring phosphate atoms along a nucleic acid backbone. A simulation with two to three representative torsions for each backbone torsion range, yielding 11,760 different backbone conformations, showed that the most common allowed conformations had a P-P-P angle between 90° and 140°. An examination of the single-stranded regions of $tRNA^{Phe}$ also showed the angles between successive phosphate atoms to range between 120° and 170°. Based on these observations, a semi-harmonic energy penalty is imposed in our models for single-stranded RNA stretches if the angle along the phosphate backbone is less than 100°. This constraint is not meant to be rigorous, but is rather an attempt at limiting the conformational space allowed for single-stranded regions using a simple potential energy term. If the user wishes to impose greater restrictions, assuming a helical conformation, for example, the default choices described here can be revised easily in the input descriptor files.

Representation of the protein component of the RNP

In the simplest model of RNP particles, proteins can be represented as spherical particles with radii appropriate to their sizes. The approximate radius for a protein whose molecular weight is known can be computed using the commonly accepted value for the partial specific volume of proteins (0.74 cm^3/g for anhydrous proteins; 1.04 cm^3/g for the hydrated state assuming 0.3 g of water per gram of protein) (Richards, 1977).

Such a low-resolution approach is logical for several RNP particles, where much more is known about the RNA than the proteins. This is especially true for the small subunit ribosomal proteins. As more detail becomes available, it will be possible to modify our protocol and incorporate proteins into these models as either all-atom models, pseudoatom models with one pseudoatom per residue, collection of spheres, or as non-spherical surfaces.

Imposition of tertiary structure experimental data on the models

Chemical cross-links

Chemical cross-linking is a common experimental technique for obtaining structural information on RNPs. Reactive cross-linking groups are introduced into the RNA and/or proteins, the RNP is reconstituted, and the cross-linker is photoactivated. The cross-link is then localized within the RNA chain to identify the RNA nucleotides and/or protein that belong to the same neighborhood. Such tertiary information is incorporated into our models with bonds that bring the cross-linked components together. The lengths of these bonds are based on the length of the cross-linking agent plus sum of the radii of the reactive nucleotides/proteins. Force constants, or the strength of bonds representing cross-links, are chosen to be proportional to the quality of the experimental data as discussed later in the paper. Often the cross-link is localized to a stretch of RNA; in such cases the bond is placed within any helical region in that stretch. This is done because duplexes are the most structured parts of our models.

Footprinting studies

Footprinting is another technique that provides information about interactions between proteins and the RNA in RNP particles. These experiments look at the changes in accessibility of the RNA chain to chemical or enzymatic probes on the addition or deletion of the protein components of the RNP (see review in Stern et al., 1989) (Darsillo and Huber (1991) discuss the use of chemical nucleases for probing RNA-protein interactions). Inhibition of reactivity at certain nucleotides on the addition of a protein can be inferred as a direct protection (and hence contact) of the RNA by the protein at that site. Reduced reactivity can also be caused by conformational changes in the RNA chain brought about by the addition of the protein, and it is not easy to distinguish between these two modes of protection (Stern et al., 1988). Because of these problems in interpretation, footprinting data are usually less reliable than cross-linking results. On the other hand, footprinting provides extensive sets of tertiary contacts. Protein-RNA contacts identified by these techniques are incorporated in our models as bonds with force constants based on the strength of the footprints.

Cross-linking and footprinting data provide crucial tertiary constraints in our modeling protocol. Correct choice of such data is important to get useful models. It is equally important to rank correctly these tertiary contacts in terms of quality, so that appropriate weights or force constants can be assigned during the refinement of the model. Although it is often difficult to quantitatively assess such data and their quality, it is necessary that some standard scale to be used to compare data from disparate sources. This point is discussed in more detail in the section on the potential functions and force constants.

Shape and positional information from electron microscopy

Electron microscopy (EM) can provide information about the shape of RNP particles such as the ribosomal particles (reviewed by Stöffler-Meilicke and Stöffler, 1990; Oakes et al., 1990b; Frank et al., 1990). Sophisticated image analysis and 3-D reconstruction methods are now providing quantitative surface topography for the ribosomal subunits (Frank and van Heel, 1982; Frank et al., 1991).

Three-dimensional surface topography provides some constraints for the folding of an RNA chain. The usefulness

of such an approach is limited, however, unless data about the orientation of the surface with respect to the RNP components are also available. In the case of the ribosomal 30S subunit, for example, one needs to know how to orient the protein map from neutron diffraction (Capel et al., 1988) within the envelope determined by electron microscopy (Frank et al., 1991). Extensive orientation data exist for the ribosomal subunits, with both proteins (Stöffler-Meilicke and Stöffler, 1990) and RNA fragments (Oakes and Lake, 1990; Oakes et al., 1990b) localized on the subunit surfaces using immunoelectron microscopy and DNA hybridization electron microscopy.

We use spherical harmonic functions to represent EM shape data, an approach first used to approximate the solvent-accessible surfaces of molecules (Max and Getzoff, 1988). These spherical harmonic functions are used to impose the EM shape on the RNA chain as it folds by using a simple harmonic force on any part of the RNA chain that strays outside the surface. More details on this novel potential function are provided later in this paper and in Malhotra et al. (1994). Experimental data about the position of RNA fragments or proteins on the EM surface are incorporated into the models by using additional spherical surfaces within which the fragments are restrained. The radius of these spherical surfaces is based on the uncertainties of the positioning data.

Solvent accessibility data

The reactivity of nucleotides towards specific structural probes can offer insights about the accessibility and the chemical environment of different parts of an RNA chain (Ehresmann et al., 1987). Such data can be used to find the areas of interactions between an RNA chain and a protein (e.g., Romby et al., 1985) or the folding of an RNA chain by identifying helices on the surface (e.g., Celander and Cech, 1991). Chemical probe-mapped accessibility of individual atoms in nucleotides has been used to propose detailed models of RNA molecules such as the 5S RNA (Westhof et al., 1989; Brunel et al., 1991), and catalytic introns (Kim and Cech, 1987).

Solvent accessibility patterns of the phosphate groups are most relevant to our protocol, because the all-P models are based on the RNA backbone. One such probe is ethylnitrosourea (ENU), an N-nitroso alkylating agent that attacks exposed phosphate group oxygens in both helical and single-stranded nucleotides (Ehresmann et al., 1987). For large systems such as the ribosomal RNAs, which are modeled at low-resolution, it is more meaningful to compare reactivities at the level of helices rather than individual nucleotides so as to distinguish regions on the surface of the RNP particle from those buried within the core of the particle. Unfortunately, it is difficult to devise potential functions suitable for molecular modeling, which force a system of atoms towards a defined pattern of solvent accessibility. The nonatomic resolution of our models also makes it harder to incorporate data on atomic solvent accessibilities. Because of these problems, solvent accessibility data are not incorporated into the potential functions used for folding and refining the models but are used as a tool for evaluating the relative quality of different final models. As will be discussed later, most RNP systems are under-determined, and a large number of models can be proposed to fit the tertiary data. Solvent accessibility can be used to evaluate such models and choose conformations which maximize agreement with the data.

Other tertiary structure data

Apart from the specific techniques listed above, data about interactions between different regions in an RNP particle can be obtained from many other experiments—fluorescence energy transfer, mutational analysis, phylogenetic analysis, low angle neutron scattering, etc. Results from all these types of experiments can be incorporated into models as distance constraints (or pseudobonds) with force constants chosen to reflect the precision of the data.

Implementation of the modeling protocol

Potential functions and force constants

Table 1 summarizes the forms of potential functions used in our protocol for the different types of experimental data that go into low-resolution RNA models. Harmonic-type potential functions are used for all the constraints in our models, including bonds, angles, and improper torsions:

$$E_{\beta i} = k_{\beta i}(\beta_i - \beta_{io})^2 \quad \text{for bonds,} \tag{1}$$

$$E_{\alpha i} = k_{\alpha i}(\alpha_i - \alpha_{io})^2 \quad \text{for angles,} \tag{2}$$

$$E_{\tau i} = k_{\tau i}(\tau_i - \tau_{io})^2 \quad \text{for improper torsions,} \tag{3}$$

where $E_{\beta i}$, $E_{\alpha i}$, and $E_{\tau i}$ denote energy of ith bond, ith angle, and ith torsion; $k_{\beta i}$, $k_{\alpha i}$, and $k_{\tau i}$ are the force constants for the ith bond, angle, and torsion respectively; α_i, β_i, and τ_i are the ith angle, bond length, and torsion, and α_{io}, β_{io}, and τ_{io} are the corresponding equilibrium or ideal values.

Harmonic functions are used in our models because they are easy to minimize and have a unique minimum. The resulting potential function is non-negative, regardless of the conformation of the model. A potential of zero indicates that all experimental constraints are satisfied, and the value of the potential in other cases is a rough indicator of the quality of a model. A harmonic bond potential function is also equivalent to a gaussian distribution of bond length β_i about the equilibrium value β_{io}, with a variance equal to $RT/2k_{\beta i}$. Force constants can thus be chosen to reflect the uncertainties associated with bond lengths, angles, and torsions. Similar considerations apply to the force constants used for all other terms in the potential function. This provides a mechanism for appropriate weighting of different kinds of data with different levels of uncertainties. Table 1 lists typical uncertainty values and their basis for the several types of experimental data.

For some of the terms in the potential function, the determination of force constants is straightforward and fairly rigorous. This is true whenever there is a sufficiently large

TABLE 1 Functional form and parameters of the potential functions used for structure refinement of low resolution models for large RNAs and RNPs

Experimental data	Form of potential function	Typical uncertainty values	Basis for uncertainty value
Secondary structure	Harmonic pseudobonds	0.1–1 Å	$tRNA^{Phe}$ crystal structure (Hingerty et al., 1978)
	Harmonic pseudoangles	0.2 radians	
	Harmonic improper torsions	0.2 radians	
RNA-RNA or RNA-protein cross-links	Semi-harmonic pseudobonds	2–10 Å	Quality of the experimental data
RNA-protein footprinting data	Semi-harmonic pseudobonds	2–10 Å	Quality of the experimental data
Size of individual nucleotides	Semi-harmonic volume exclusion	—	Rather than use an uncertainty value for volume exclusion, force constants were arbitrarily kept low to allow tangles in RNA chains to be resolved (see text)
Size of proteins (for RNPs)	Semi-harmonic volume exclusion	5–15 Å	Nonspherical or axial nature of the protein
EM shape data	Semi-harmonic surface constraints	10–20 Å	Quality and resolution of EM image reconstruction
Immuno-EM positional data	Semi-harmonic surface constraints and/or harmonic positional constraints	25–50 Å	Quality and resolution of EM image reconstruction; probe footprint size
Protein positions (for RNPs)	Harmonic positional constraints	0–15 Å	Quality of protein localization data and experimental error estimates

database that the mean and variance of the quantity under consideration can be determined with confidence. For example, the tRNA crystal structures provide sufficient data to parameterize accurately the terms for inter-phosphate distances in double helical regions. Decisions in other parameters are more subjective. For example, a photo-activated cross-linking agent may attack a target nucleotide at any of several atoms, and there may be several single bonds in the cross-link. In this case, neither the mean nor the variance of the inter-phosphate distance can be determined very accurately. If the cross-link hits a protein, additional uncertainty arises from the protein's shape. As a consequence, parameterization of our models is an ad-hoc procedure, and it cannot be compared to the development of force fields for traditional all-atom methods.

In this regard, there are two important considerations. First, the final models are of such low-resolution that rigorous parameter values are less critical than in all-atom modeling. Our current model of the small subunit of the *E. coli* ribosome has an average resolution of about 15 Å, and the uncertainty of individual nucleotides ranges from about 5 to over 50 Å (Malhotra and Harvey, 1994). In such a model, an additional piece of information about the location of a nucleotide whose position was not well determined can be very significant; it is very important to know that this nucleotide is within cross-linking distance of some other particular nucleotide, but, at this resolution, it is much less important to know the exact inter-phosphate distance. Second, it is essential that the modeler be able to examine the effects of changing the values of force field parameters. It is for this reason that the parameters are specified in a formatted *yammp* descriptor file that is easy to read and edit (Tan and Harvey, 1993).

Force constants for secondary structure motifs are chosen to mimic variability in the $tRNA^{Phe}$ crystal coordinates (Hingerty et al., 1978), the high-resolution structure from which the duplex average "equilibrium" values for bonds, angles, and torsions are derived for our models. For example, an examination of the helical region of $tRNA^{Phe}$ shows that the interstrand separation of phosphate atoms in base pairs has a mean of 18 Å and SD of about 1 Å. The force constant of the harmonic potential function for such bonds is thus 0.3 kcal/mol $Å^2$ at $T = 300$ K. Similarly, an SD of 0.2 radians is used for angles and torsions in helices based on the variation in these in the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978). The distance between successive phosphate groups along the primary sequence in $tRNA^{Phe}$ has an SD of 0.6 Å in non-stem-loop regions, and this value is used for calculating force constants in the single-stranded regions of our models. Comparable variance and average values were seen in the other high-resolution tRNA structures in the Nucleic Acid Database (Berman et al., 1992).

Bonds used to enforce experimentally observed tertiary interactions are semi-harmonic:

$$E_{\beta i} = \begin{cases} k_{\beta i}(\beta_i - \beta_{io})^2 & \text{if } \beta_i \geq \beta_{io} \\ 0 & \text{if } \beta_i < \beta_{io}. \end{cases} \quad (4)$$

Force constants for such bonds are chosen to reflect experimental uncertainty. In absence of a quantitative error estimate, a range is chosen. For example, SDs ranging between 2 and 10 Å are used for cross-links and footprinting data, depending on the quality of the data. In general, the force constants for tertiary data are much weaker than those for secondary structure motifs.

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

$$E_{\gamma ij} = \begin{cases} k_{\gamma ij}(r_{ij} - r_{ijo})^2 & \text{if } r_{ij} \leq r_{ijo} \\ 0 & \text{if } r_{ij} > r_{ijo}, \end{cases} \quad (5)$$

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_{ijo} is the minimum distance allowed between the two atoms (usually the sum of their radii). The minimum separation of phosphate atoms in single-stranded regions of the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978) is used as a guide to get an exclusion distance of 5 Å between the P pseudoatoms. Suitable exclusion distances are used for other pseudoatoms. Rather than use experimental uncertainty, force constants for nonbond interactions are kept soft (a 5-Å overlap between pseudoatoms has an energy penalty of 1 kcal/mol) in our models to permit the chain to pass through itself. This allows tangles in the starting random structure to be resolved during energy refinement.

The potential function used to prevent kinking of the single-stranded RNA is of the form

$$E_{\theta ijk} = \begin{cases} k_\theta(\alpha_{ijk} - \alpha_o)^2 & \text{if } \alpha_{ij} \le \alpha_o \\ 0 & \text{if } \alpha_{ij} > \alpha_o, \end{cases} \quad (6)$$

where $E_{\theta ijk}$ is the bending energy between successive atoms i, j, and k along a single-stranded stretch of RNA, k_θ is the bending force constant, α_{ijk} is the angle between the atoms i, j, and k, and α_o is the minimum angle allowed (100°, as discussed earlier). A fairly stiff force constant, $k_\theta = 7.5$ kcal/mol rad^2 is used, similar to the force constants for angles and improper torsions in our model.

The potential function for surface topography uses spherical harmonics to represent surfaces (Max and Getzoff, 1988):

$$r_s(\theta, \varphi) = \sum_{n=0}^{N} \sum_{k=-n}^{+n} C_{nk} Y_{nk}(\theta, \varphi), \quad (7)$$

where $r_s(\theta, \varphi)$ is the distance to the surface from the center of the surface for the angular coordinates (θ, φ), Y_{nk} are the spherical harmonics of order n, C_{nk} are the corresponding expansion coefficients, and N is the order of the expansion. The expansion coefficients are determined by the surface integral

$$C_{nk} = \int r_s(\theta, \varphi) Y_{nk}(\theta, \varphi)\, d\Omega, \quad (8)$$

computed over the unit sphere. We determine these coefficients numerically using the *sphinx* program (Max and Getzoff, 1988), which computes C_{nk} around a center computed by averaging the coordinates of all the points on the Connolly type molecular surface.

The potential used to constrain an atom i within a surface is

$$E_{si} = \begin{cases} k_s(r_i - r_s(\theta_i \varphi_i))^2 & \text{if } r_i > r_s(\theta_i, \varphi_i) \\ 0 & \text{if } r_i \le r_s(\theta_i, \varphi_i), \end{cases} \quad (9)$$

where E_{si} is the surface topography energy for atom i, k_s is the surface topography force constant for the surface, r_i is the distance between atom i and center of the surface, and θ_i and φ_i are the polar and azimuthal coordinates of atom i with respect to the center of the surface. The spherical harmonics expansion in Eq. 7 is suitable only when $r_s(\theta, \varphi)$ is single-valued for any given value of θ and φ, i.e., the surface has no overhangs or cavities and is star-like (Max and Getzoff, 1988). Complex surfaces thus have to be divided into several convex spherical harmonics surfaces. When this is the case, the energy E_{si}, and the associated force, is computed only to the surface closest to the atom i. The details of this procedure are described elsewhere (Malhotra et al., 1994).

Modeling software

The modeling protocol was implemented using *yammp*, an in-house molecular mechanics package (Tan and Harvey, 1993). The RNA chain or RNP is described using an RNA script file (Fig. 5 shows a typical RNA script file for $tRNA^{Phe}$) where the secondary structure and the tertiary interaction data are specified. For RNP particles, details about the proteins in the system are also included in the script file. The RNA script file is converted into a molecular topology file (called a descriptor file in *yammp*) with the program *mksnad* (make succinct nucleic acids descriptor) developed for this protocol. Keywords are used in the RNA script files to direct *mksnad* to make descriptor files for modeling a system at the 1H, 5H, or the all-P level of resolution.

The program *mksnac* uses the RNA script file to create random walk chains that serve as the starting conformations for structure refinement. The direction at each step of the walk is varied randomly between zero (a perfectly straight line) and a maximum specified angle. Generally a small angle (15°) is used to get an extended random chain and to reduce tangles in the starting structure. The length of each step is based on the length of the bonds connecting the pseudoatoms that make up the RNA chain.

Modeling is started using a random walk chain in absence of any other structural models for the RNA being considered. When some reasonable starting structures are available, such coordinate data can also be used. The coordinate file (created by *mksnac* or any other reasonable starting structure) and the topology file (created using *mksnad*) are energy refined using *yammp*. After a 1H or 5H model is refined, the program *mksnacc* is used to convert the model coordinates into starting coordinate files with a higher resolution for further refinement using a new topology file.

The models are refined using energy minimization and simulated annealing with Monte Carlo (Kirkpatrick et al., 1983). Because all the potentials used in our models are harmonic, the lowest possible energy of the system is zero (i.e., all constraints are satisfied). It is thus easy to determine when a system is completely refined. In large systems, such as the ribosomal RNAs, where all the constraints cannot be fully satisfied, the potential energy is a direct measure of the extent of unsatisfied constraints in a model.

A major advantage of our procedure is that several models can be built and refined for a given set of input data. By using different initial random walk chains, we hope to span conformational space reasonably well. The comparison of a set

```
#
#  RNA script file for tRNA phe - all-P model (with space-filling X atoms)
#
 RNA_SCRIPT       1

#    Primary structure: A chain of 76 nucleotides.

 SYSTEM           76

 P_ATOM STRAND    1
  1   76

#    Secondary structure: The four double helices of the cloverleaf

 HELIX            4
  1  72  7  66
  10  25  14  21
  27  43  31  39
  49  65  53  61

#    Nucleotides not in double helices are classified as either
#    single-stranded or as loops at the end of stems

 SINGLE_STRAND    5
  7  10
  25  27
  43  49
  65  66
  72  76

 LOOP             3
  14  21
  31  39
  53  61

#    Tertiary interaction data: - Stacking of anticodon arm on D arm
#                               - Stacking of acceptor arm on T arm

 STACK            3
  7 66 49 65
  25 10 26 44
  26 44 27 43

#    Tertiary interaction data: -  Nine tertiary interactions proposed by Levitt
                                   (1969).  Actual distances between P-atoms in
                                   these nucleotides from the tRNA phe crystal
                                   structure (Hingerty et al.,1978) are used,
                                   rather than idealized tertiary interaction
                                   distances.  An uncertainty of ±4 Å is assumed
                                   for these contacts.

 CROSS_LINK       9
  8   13   4.0    9.204
  9   12   4.0    9.674
 15   48   4.0   19.096
 18   55   4.0   17.090
 19   56   4.0   15.445
 21   54   4.0   24.554
 25   57   4.0   39.555
 44   57   4.0   30.040
 73   76   4.0   15.607

#      Include space-filling atoms in the center of helices

 INCLUDE_XATOMS
```

FIGURE 5 An RNA script file describing the secondary structure of $tRNA^{Phe}$ (see Fig. 1 *a*), along with helix-stacking and tertiary interactions used for the models in Fig. 6 *b* (see figure legend for Fig. 6). Such files are used to specify secondary and tertiary structure for RNA chains.

of such models provides information on both systematic and random errors in the set. Systematic error—usually representing conflicts in the experimental data—is suspected whenever a particular constraint or set of constraints is consistently unsatisfied in all the models. Random errors occur whenever there are fewer constraints than degrees of freedom, which is always the case for our models of the ribosome. Random errors are also expected in over-determined systems whenever there are no overlapping or long-range constraints, because very small errors in the short-range constraints can accumulate to give large variability in the global structure (Fig. 3). A quantitative evaluation of random errors can be obtained from a superposition of a set of different models. The statistical fluctuations in the position of a particular nucleotide are a measure of the resolution of the model at that point, much like the Debye-Waller temperature factors in a structure determined by x-ray crystallography.

RESULTS AND DISCUSSION

Illustration of the protocol-tRNA^{Phe}

Transfer RNAs are a good test case for the modeling protocol described here, because they are the best characterized RNA molecules. Before the x-ray crystallographic structures became available for the tRNAs, the cloverleaf secondary structure and several tertiary interactions were known (Levitt, 1969). Figs. 6 and 7 show two sets of tRNA^{Phe} folding patterns generated using some of the data available before the crystal structures were derived. Fig. 6 *a* (and Fig. 7 *b*) shows several models refined using only the secondary structure and the correct helix stacking (the acceptor stem on the T arm, and the D arm on the anticodon stem), whereas Fig. 6 *b* (and Fig. 7 *c*) adds nine tertiary interactions to the constraints used in Fig. 6 *a*. The larger set of tertiary interactions in Fig. 6 *b* (8-13, 9-12, 15-48, 18-55, 19-56, 21-54, 25-57, 44-57, and 73-76) was proposed before the tRNA^{Phe} crystal structure was derived (Levitt, 1969). Two other interactions tabulated by Levitt (1969)—between nucleotides 32 and 39, and nucleotides 55 and 58—were not used because these pairs are within the same stem-loop. Rather than assume distances for these tertiary contacts, we used distances taken from the tRNA^{Phe} crystal structure (Hingerty et al., 1978). The constraints used for Fig. 6 *b* are described by the RNA script file in Fig. 5.

Fig. 6, *a* and *b* show two illustrative sets of models for tRNA, given a very limited set of experimental data. As can be seen, the secondary structure motifs—the helices, loops, and helix stacking—are imposed correctly on the RNA chains. It is also obvious that a great deal of variability exists between the different models. This is especially true in the positioning of the anticodon stem-D arm with respect to the acceptor stem-T arm. The structures shown in Fig. 6 *a* have RMS deviations in the range 12.8–14.2 Å when compared to the tRNA^{Phe} crystal structure (Hingerty et al., 1978), whereas the models in Fig. 6 *b* show a range 8.9–10.5 Å. RMS is a very gross measurement of structural differences, especially

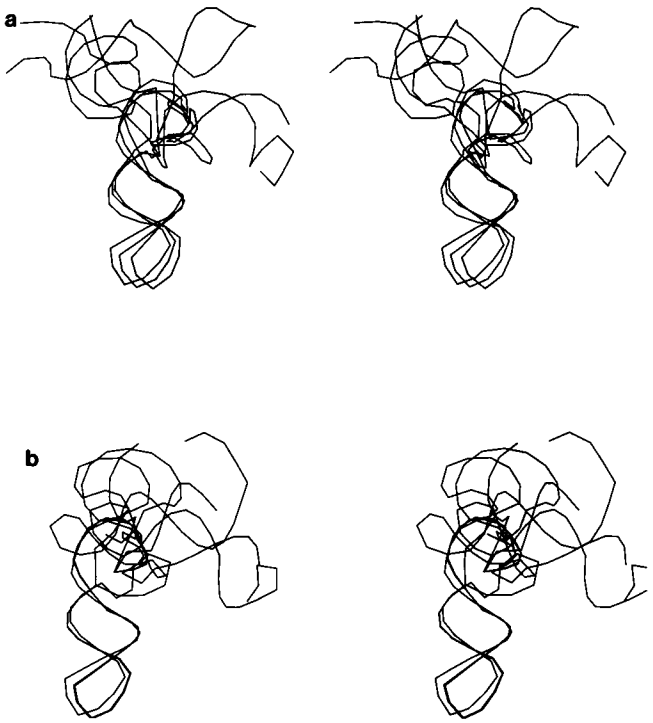

FIGURE 6 RNA folding patterns generated from random initial chains for tRNA^{Phe} using the secondary structure, helix stacking, and a few tertiary contacts as constraints. Three RNA chains are shown after energy refinement with constraints that impose (*a*) the secondary structure and stacking of the acceptor stem on the T stem and of the D stem on the anticodon stem, and (*b*) the secondary structure, helix stacking, and nine tertiary interactions proposed in 1969 before the x-ray crystal structure was solved (Levitt, 1969). Rather than use idealized tertiary interaction distances, the P-P distances from the tRNA^{Phe} crystal structure (Hingerty et al., 1978) were used. These constraints are described by the RNA script file in Fig. 5. In both panels, the RNA backbone is shown with only the P-atoms, and the chains are displayed with the D arm and the anticodon stem superimposed using a least-squares fit.

for large deviations in conformations, and does not accurately reflect the tightening in the range of conformations when the larger set of tertiary interactions are used. For example, unlike the structures in Fig. 6 *a*, all of the models with the larger set of constraints place the T-acceptor stem arm of tRNA on the same side of the anticodon-D stem arm (Fig. 6 *b*).

The variability between the conformations is a reflection of the lack of tertiary data and long-range information in these models. For an all-P model of tRNA^{Phe}, assuming that the stem-loop regions have standard geometries, there are still 54 degrees of freedom (4 stem-loops with 6 degrees of freedom each and 12 unstructured nucleotides with 3 degrees of freedom each, less 6 degrees of freedom corresponding to rigid body translation and rotation). Of these, the connectivity of the RNA chain provides 16 constraints. Thus, apart from the secondary structure, 38 additional (and independent) tertiary interactions are needed to specify completely the global folding of the RNA chain. Even ignoring the single-stranded 3′ end on the acceptor stem, there are 42 degrees of freedom that need to be specified. Clearly, even

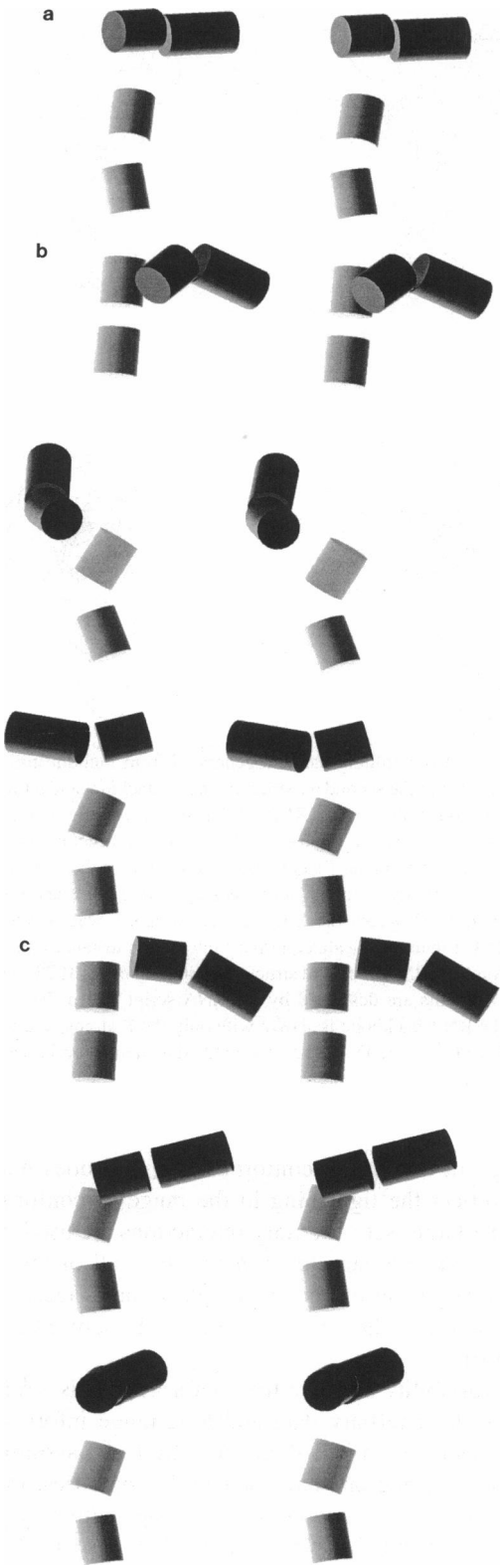

FIGURE 7 Helix positions in the $tRNA^{Phe}$ models shown in Fig. 6. The cylinders approximate the helical regions in the tRNA chain (using a best-fit superposition), and are shown with their diameters scaled by half for clarity. (*a*) $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978) in this representation; (*b*) Three tRNA models from Fig. 6 *a*; (*c*) Three tRNA chains from Fig. 6 *b*. The anticodon stem is displayed with a lighter shade of gray, and all panels show the same orientation. This figure was generated using the *ribbons* graphics program (Carson, 1987).

at the low-resolution described here, many carefully chosen tertiary contacts are needed to specify the 3-D folding of a simple RNA chain such as tRNA.

Even in under-determined systems, where no unique solution exists, meaningful information can be extracted by generating a wide range of conformations compatible with the data and comparing them. In large systems such as the ribosomal subunits, a comparison of different conformations can be used to identify quantitatively the regions of the RNA chain that are least constrained. Average positions and orientations of helices that are well constrained can also be derived (Malhotra and Harvey, 1994). In these models, very little, if anything, can be said about the nonhelical regions of the RNA chain. The cylinders used to represent helices in the RNA chain in Fig. 7 typify the resolution of these models. Of course, if single-stranded RNA were better understood, and if a great deal of structural data were available, nonhelical regions could also be better localized in these models.

Accurate comparison among the different models that satisfy the experimental data is also dependent on good sampling of conformational space. Our approach to sample a wide conformational space is to use several different random-walk chains as starting structures (typically, 10 or more). Not all of these chains can be refined, and structures with very high energies after refinement (relative to others in the group of random chains) are discarded. This approach is similar to the methods used in distance geometry where random sets of trial distances are assumed.

Another application of the RNA folding protocol is to eliminate conformations that are not sterically possible or are not compatible with experimental data. Different $tRNA^{Phe}$ stacking schemes and loop interactions were examined using this protocol to show that stacking of the acceptor stem on the anticodon stem, and the D stem on the T stem, was not a sterically acceptable arrangement (Malhotra et al., 1991).

Similarly, other conflicts in tertiary interaction data sets can be tested. In the tRNA models shown in Fig. 6 *a* with the secondary structure and helix stacking, all of the chains could be refined to a potential energy of zero, indicating that all the imposed constraints were satisfied. When the additional nine tertiary interactions were introduced as contacts separated by 5 Å or less, and with an uncertainty of 4 Å, the models could be refined to a minimum energy of only about 0.54–0.68 kcal/mol, which suggests that not all of the constraints could be satisfied simultaneously (results not shown). In these models, the constraint between nucleotides 25 and 57 was always unsatisfied, and in conflict with the helix stacking imposed between the D arm and the anticodon stem. Nucleotide 25 is involved in stacking of these helices, and our assumption of a 5-Å contact for the 25-57 interaction is clearly wrong (the phosphate atoms of these nucleotides are separated by 39.6 Å in the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978)). When the distances for these contacts were taken from the $tRNA^{Phe}$ crystal structure (Hingerty

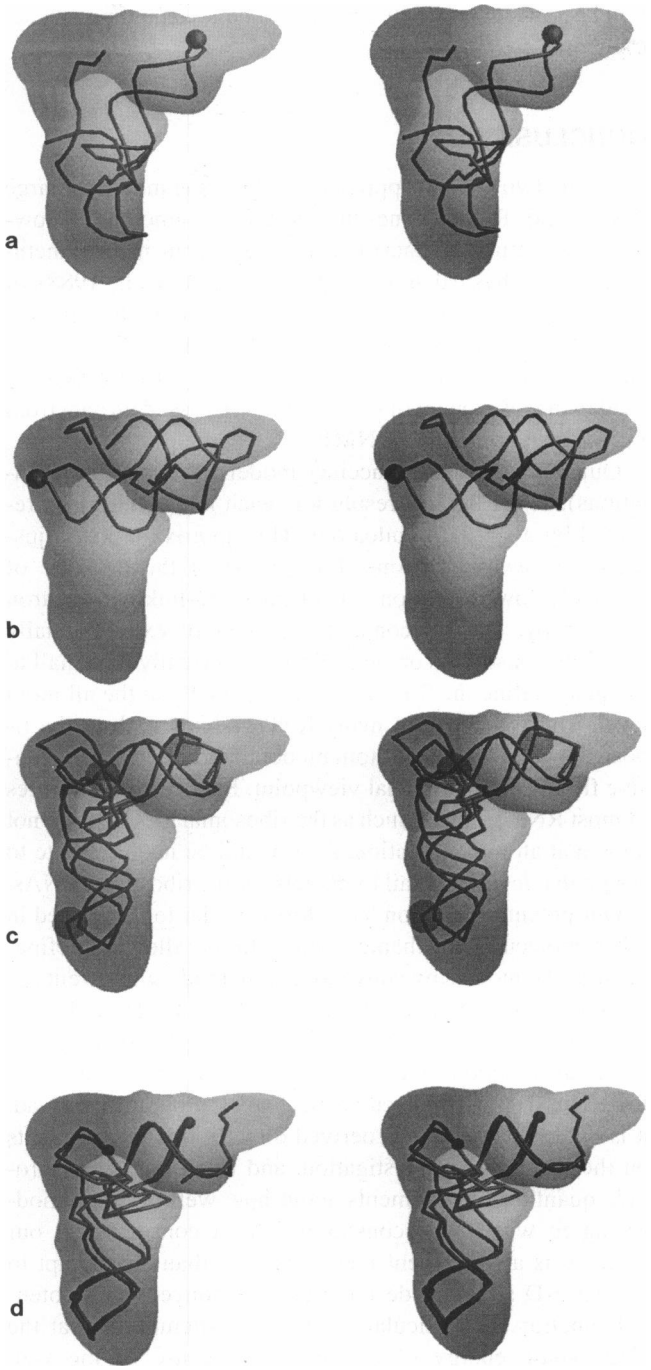

FIGURE 8 Illustration of the use of shape data and the role of orientation constraints to impose global geometry in low resolution structure refinement. The $tRNA^{Phe}$ L shape is approximated by 2 order 8 spherical harmonic surfaces (one for each arm), shown here as shaded surfaces. Random RNA chains are energy refined with constraints that impose the secondary structure, helix stacking, and nine tertiary contacts (see Fig. 5), and the RNA chains are required to be within either of the two surfaces with a special potential function (Eq. 9). In panels *a* and *b*, two RNA chains are refined without any constraints to orient the RNA chain, while being forced to conform to the $tRNA^{Phe}$ shape. The location of the anticodon loop is indicated by the sphere at nucleotide 34. Panels *c* and *d* show two RNA chains refined with constraints to orient the RNA chain, in addition to the shape constraints used in panels *a* and *b*. In panel *c*, nucleotides 1, 31, and 51 are restricted to the acceptor terminus, anticodon loop, and elbow region, respectively, by the use of three additional surface constraints that restrict these nucleotides to spheres of 5 Å radii (shown as shaded spheres). In panel *d*, nucleotides 1, 31, and 51 are constrained to their positions in the RNA^{Phe} crystal structure (Hingerty et al., 1978) by harmonic bonds. These positions are shown as spheres in panel *d*. This figure was generated using the *ribbons* graphics program (Carson, 1987).

et al., 1978), the resulting models have better anticodon-D stem stacking and refine to zero energy (Fig. 6 *b*).

This example shows that our protocol would produce models for tRNA that are of comparable quality to the best hand-built model (Levitt, 1969), if the crystal structure were not known. In addition, the method would make quantitative statements about the range of acceptable models, and it would identify conflicts in distances inferred from experiments.

Illustration of the use of shape constraints—$tRNA^{Phe}$

Fig. 8 shows the use of surface topography constraints to enforce the correct shape on the tRNA chain. The surface envelope shown is a spherical harmonics approximation of the solvent accessible surface for tRNA computed from the $tRNA^{Phe}$ crystal structure, using the *ms* program (Connolly, 1983). All atoms in the crystal structure were assigned radii of 5 Å to get a smooth surface, which is easier to approximate using spherical harmonics. The solvent accessible surface was computed separately for each arm, since spherical harmonics requires star-like surfaces with no overhangs. Coefficients for the spherical harmonic approximation of the surfaces were computed using the *sphinx* program (Max and Getzoff, 1988), and two spherical harmonics surfaces of order eight can approximate the tRNA shape quite accurately. Details of this are reported elsewhere (Malhotra et al., 1994).

Random walk chains were then refined using the $tRNA^{Phe}$ secondary structure and the tertiary interaction data used for the structures in Fig. 6 *b*, along with the shape information in the form of terms in the potential function that impose an energy penalty on atoms which stray outside the surface (Eq. 9). Two RNA chains refined to low energy are shown in Fig. 8, *a* and *b*. Fig. 8, *c* and *d* show RNA chains refined with additional constraints that tethered three nucleotides from the acceptor stem (nucleotide 1), anticodon loop (nucleotide 31), and the T arm (nucleotide 55), to fixed positions in space. These constraints help orient the RNA chain with respect to the shape being imposed.

Fig. 8 illustrates several aspects of use of surface constraints in our modeling protocol. First, orientation data (provided here by three tethering spheres or positional constraints) are important in fitting the RNA chain into the surfaces correctly. Tests without these tethering constraints positioned the RNA chain in a variety of orientations within the tRNA surface (Fig. 8, *a* and *b*). These chains were often packed along the length of the surface, leaving the ends of the surfaces mostly empty. The extreme aspherical shape of tRNA is partially to blame for these problems. The addition of positional constraints guarantees correct orientation within the surface envelope, improving the models substantially (Fig. 8, *c* and *d*).

Second, because the potential function pulls errant atoms towards the center of the surface, parts of the RNA chain tend to get pulled into the surface closest to them. This can be seen in Fig. 8 *d* where the acceptor CCA end is often bent inwards, instead of towards the correct part of the surface at the extreme right end of the upper arm. These are small errors, however, considering the low-resolution of our models. This also reinforces the importance of orientation data in correctly placing parts of the RNA chain within a surface. Lastly, our experience with the use of surfaces suggests that surface topography constraints have to be "loose" to facilitate energy refinement of the models. This requires both soft force constants and dimensions larger than the bare minimum necessary to fit the RNA chain.

The RNA chains displayed in Fig. 8 *d* have RMS deviations of 7.3 and 8 Å when compared with the $tRNA^{Phe}$ crystal structure (Hingerty et al., 1978). This range is considerably lower than that for models shown in Fig. 6 (8.9–14.2 Å). Shape constraints, therefore, are seen to improve the models substantially. Visual comparison of Figs. 6 and 8 clearly shows the improvement of the global structure of the RNA as more tertiary structure constraints are imposed. In larger RNAs, footprinting and cross-linking experiments provide similar tertiary contacts, 3-D electron microscopy image reconstructions provide shape constraints, and immunoelectron microscopy provides positional information.

Extension of low-resolution all-P models to higher resolution

Each P pseudoatom in the all-P models represents one nucleotide, and the string of P atoms traces the RNA backbone. In principle, once the RNA chain is fairly well positioned, each P atom can be extrapolated into an all-atom nucleotide. Such an extrapolation to higher resolution is fairly straightforward for helical regions, where direct superposition can be used. In other regions, extrapolation to all-atom models is hampered by the large number of possible orientations of the sugar, base, and the phosphate groups, and the current lack of structural understanding of single-stranded and loop regions of RNA chains. Work is underway in our lab to develop these extrapolation procedures. Such extrapolation procedures will have to be combined with atomic resolution experimental data to be considered a true refinement protocol. Otherwise, all-atom models must be considered somewhat speculative.

From a procedural standpoint, there are no limitations on the use of atomic resolution data in our models. Because molecular modeling techniques deal with all types of atoms, specific areas of interest such as active sites can be modeled in full atomic detail, while treating the rest of the RNA chain at a lower resolution. Other approximations in our protocol, such as the treatment of proteins as spheres, can also be replaced with atomic detail or the use of surfaces for defining protein shapes and orientations more precisely. As experimental data on large RNA systems grow, such a mixed mode approach incorporating both high and low-resolution data will become increasingly important in understanding and deciphering 3-D structures.

CONCLUSIONS

There are two general approaches to building models of large RNAs and RNPs using the increasing amount of low-resolution structural data on them. First are the manual methods, using physical models (Brimacombe et al., 1988) or computer graphics (Stern et al., 1988). An alternate approach is to use an automated computer method. Here we described an automated protocol derived from the algorithms that are used in the refinement of model structures based on data from x-ray crystallography or NMR.

Our procedure uses succinct models (reduced representations); at the highest resolution, each nucleotide is represented by a single pseudoatom. This approximation is justified for several reasons: First, most of the data are of relatively low-resolution (chemical cross-linking, electron microscopy, etc.). Second, the number of experimentally available constraints on large RNPs is currently too small to uniquely define the 3-D structure, especially at the all-atom level. Third, the size of many RNP systems such as the ribosome would make all-atom modeling prohibitively expensive from a computational viewpoint. Finally, the structures of most RNP proteins, such as the ribosomal proteins, are not known at atomic resolution, so it would be inappropriate to assign that level of detail to models for the ribosomal RNAs.

Our potential function has a form similar to those used in other molecular mechanics applications, allowing refinement of the models by conventional methods such as energy minimization, molecular dynamics, Monte Carlo, and combinations of these (e.g., simulated annealing). However, our potential function does not contain the same terms as found in all-atom molecular mechanics programs. Instead, it is based on constraints derived directly from experiments on the RNP under investigation, and it is designed to provide quantitative statements about how well the final models agree with those constraints. As a consequence, our protocol is a refinement method, rather than an attempt to predict 3-D structure de novo using a conventional potential function. In particular, there is no attempt to treat the electrostatic energy of the system, because of the well known difficulties in accurately treating electrostatic effects in molecular mechanics (Harvey, 1989). If electrostatic evaluation is required, it would have to be done post hoc, using established methods based on the numerical solution of the Poisson-Boltzmann equation (reviewed by Sharp and Honig, 1990; You and Harvey, 1993).

One important feature of our protocol is the ability to incorporate information on the shape of the RNA or RNP, when such data are available. We believe this is one of the major advantages of a molecular mechanics approach, because shape information can be included in the potential function (Malhotra et al., 1994). It is not clear how such information might be included in approaches based on distance geometry (Hubbard and Hearst, 1991a), distance ma-

trices (Hadwiger and Fox, 1991), or exhaustive conformational searching (Major et al., 1991; Gautheret et al., 1993).

The systems for which this protocol will be most useful will generally be quite under-determined. In such cases, it is important to specify quantitatively the overall resolution or the uncertainty in different regions of the model. Our protocol allows the modeler to build a range of models compatible with the data and, by comparing them, to come up with quantitative statements about model uncertainty. In addition, the method makes it easy to test alternate assumptions about different kinds of data, by varying the weights (force constants) associated with each of these. Often, the modeler may want to make manual manipulations or build an entire consensus model manually. Our protocol will facilitate this, because the resulting model can be refined with our programs. Small adjustments in the positions of some nucleotides will probably be produced, and the final model can be compared to the full set of those built automatically by comparing their energies.

In many cases, all-atom models will be ultimately desired, either for an entire RNA or for a particularly important region. The consensus low-resolution model can be extrapolated to an all-atom model manually or by using exhaustive conformational search programs, such as MC-SYM (Major et al., 1991; Gautheret et al., 1993). The set of models produced by our protocol can provide sets of inter-phosphate distance constraints (and associated uncertainties) that can be used as input to MC-SYM.

Although we have illustrated this structure refinement protocol using tRNA, it is designed to tackle much larger systems such as the ribosomal RNAs that currently lie outside the realm of molecular modeling techniques. This protocol is being used for the structure refinement of the 16S RNA in the small subunit of the ribosome (Malhotra et al., 1990; 1991; Malhotra and Harvey, 1994). It is also being applied to RNase P. It is in such systems, accessible only by low-resolution experimental techniques, that the full potential of this approach can be tapped.

Although the protocol described here is for large RNAs and RNPs, similar methods can be used to refine low-resolution models of any macromolecular system for which a body of tertiary interaction data are available. This approach should be especially attractive for structural studies of large macromolecular assemblies where much of the biology of living systems takes place, as these assemblies are often too large to be examined in their entirety using conventional structure refinement approaches such as NMR and x-ray crystallography.

Specific contributions to this work are: S. C. Harvey for inception of the protocol and overall guidance; R. K.-Z. Tan for development of the *yammp* molecular modeling package, refinement parameters and programming assistance; A. Malhotra for development, implementation, and testing of this protocol, development of the programs *mksnad, mksnac, mksnacc*, and the writing of this paper. The authors thank Dennis Sprous and Dr. Tony You for modeling of RNA loops, Martin Jones for superposition algorithms, and Dr. Tony You for some early work on 5S RNA modeling. The authors also thank Dr. Elizabeth Getzoff (Scripps Research Institute, La Jolla, CA) for the *sphinx* program, and Dr. Mike Carson (University of Alabama at Birmingham, Birmingham, AL) for the *ribbons* graphics program.

This work was supported by a grant from the National Science Foundation (DMB-90-05767). Additional grants for the purchase of Silicon Graphics workstations were provided by the National Science Foundation and the State of Alabama. Some computer time was provided by the Alabama Supercomputer Network. We thank the reviewers for useful suggestions.

REFERENCES

Arnott, S., P. J. Campbell Smith, and R. Chandrasekaran. 1976. Atomic coordinates and molecular conformations for DNA-DNA, RNA-RNA, and DNA-RNA helices. *In* Handbook of Biochemistry and Molecular Biology. G. D. Fasman, editor. Chemical Rubber Company Press, Cleveland, OH. 411–422.

Basavappa, R., and P. B. Sigler. 1991. The 3 Å crystal structure of yeast initiator transfer RNA—functional implications in initiator elongator discrimination. *EMBO J.* 10:3105–3111.

Berman, H. M., W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. R. Srinivasan, and B. Schneider. 1992. The Nucleic Acid Database: A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* 63:751–759.

Bhattacharyya, A., and D. M. J. Lilley. 1989. The contrasting structures of mismatched DNA sequences containing looped-out bases (bulges) and multiple mismatches (bubbles). *Nucleic Acids Res.* 17:6821–6840.

Bhattacharyya, A., A. I. Murchie, and D. M. J. Lilley. 1990. RNA bulges & the helical periodicity of double-stranded DNA. *Nature.* 343:484–487.

Brimacombe, R. 1988. The emerging three-dimensional structure and function of 16S ribosomal RNA. *Biochemistry.* 27:4207–4214.

Brimacombe, R., J. Atmadja, W. Stiege, and D. Schüler. 1988. A detailed model of the three-dimensional structure of Escherichia coli 16S ribosomal RNA in situ in the 30 S subunit. *J. Mol. Biol.* 199:115–136.

Brunel, C., P. Romby, E. Westhof, C. Ehresmann, and B. Ehresmann. 1991. Three-dimensional model of *Escherichia coli* ribosomal 5S RNA as derived from structure probing in solution and computer modeling. *J. Mol. Biol.* 221:293–308.

Brünger, A. T., J. Kuriyan, and M. Karplus. 1987. Crystallographic R factor refinement by molecular dynamics. *Science.* 235:458–460.

Brünger, A. T. 1990. X-PLOR Version 2.1 Manual. Yale University Press, New Haven, CT.

Capel, M. S., M. Kjeldgaard, D. M. Engelman, and P. B. Moore. 1988. Positions of S2, S13, S16, S17, S19 and S21 in the 30S ribosomal subunit of *Escherichia coli. J. Mol. Biol.* 200:65–87.

Carson, M. 1987. Ribbon models of macromolecules. *J. Mol. Graph.* 5: 103–106.

Celander, D. W. and T. R. Cech. 1991. Visualizing the higher order folding of a catalytic RNA molecule. *Science.* 251:401–407.

Chastain, M., and I. Tinoco, Jr. 1991. Structural elements in RNA. *Progr. Nucleic Acid. Res.* 41:131–177.

Cheong, C., G. Varani, and I. Tinoco, Jr. 1990. Solution structure of an unusually stable RNA hairpin, 5′ GGAC(UUCG)GUCC. *Nature.* 346: 680–682.

Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science.* 221:709–713.

Covell, D. G., and R. L. Jernigan. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry.* 29:3287–3294.

Crippen, G. M. 1991. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry.* 30:4232–4237.

Darsillo, P., and P. W. Huber. 1991. The use of chemical nucleases to analyze RNA-protein interactions. The TFIIIA-5S rRNA complex. *J. Biol. Chem.* 266:21075–21082.

Dock-Bregeon, A. C., B. Chevrier, A. Podjarny, J. Johnson, J. S. de Bear, G. R. Gough, P. T. Gilham, and D. Moras. 1989. Crystallographic structure of an RNA helix: [U(UA)6A]2. *J. Mol. Biol.* 209:459–474.

Ehresmann, C., F. Baudin, M. Mougel, P. Romby, J.-P. Ebel, and B. Ehresmann. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res.* 15:9109–9128.

Expert-Bezançon, A., and P. L. Wollenzien. 1985. Three-dimensional arrangement of the Escherichia coli 16 S ribosomal RNA. *J. Mol. Biol.* 184:53–66.

Flory, P. J. 1969. Statistical Mechanics of Chain Molecules. Interscience, New York.

Frank, J., and M. van Heel. 1982. Correspondence analysis of aligned images of biological particles. *J. Mol. Biol.* 161:124–137.

Frank, J., A. Verschoor, M. Radermacher, and T. Wagenknecht. 1990. Morphologies of eubacterial and eucaryotic ribosomes as determined by three-dimensional electron microscopy. *In* The Ribosome. Structure, Function & Evolution. W. E. Hill, A. Dahlberg, R. A. Garrett, P. B. Moore, D. Schlessinger, and J. R. Warner, editors. ASM Press, Washington, DC. 107–113.

Frank, J., P. Penczek, R. Grassucci, and S. Srivastava. 1991. Three-Dimensional reconstruction of the 70S *Escherichia coli* ribosome in ice: the distribution of ribosomal RNA. *J. Cell. Biol.* 115:597–605.

Gautheret, D., F. Major, and R. Cedergren. 1993. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol.* 229:1049–1064.

Haasnoot, C. A. G., C. W. Hilbers, G. A. van der Marel, J. H. van Boom, U. C. Singh, N. Pattabiraman, and P. A. Kollman. 1986. On loopfolding in nucleic acid hairpin-type structures. *J. Biomol. Struct. Dyn.* 3:843–857.

Hadwiger, M. A., and G. E. Fox. 1991. Explicit distance geometry: identification of all the degrees of freedom in a large RNA molecule. *J. Biomol. Struct. Dyn.* 8:759–779.

Happ, C. S., E. Happ, M. Nilges, A. M. Gronenborn, and G. M. Clore. 1988. Refinement of the solution structure of the ribonucleotide 5′r-$(GCAUGC)_2$: combined use of nuclear magnetic resonance and restrained molecular dynamics. *Biochemistry.* 27:1735–1743.

Hare, D. R., and B. R. Reid. 1986. Three-dimensional structure of a DNA hairpin in solution: two-dimensional NMR studies and distance geometry calculations on d(CGCGTTTTCGCG). *Biochemistry.* 25:5341–5350.

Harris, M., A. Malhotra, J. Brown, J. Nolan, B. K. Oh, S. C. Harvey, and N. R. Pace. 1993. Three-dimensional structure of ribonuclease P RNA: site-directed photoaffinity cross-linking and molecular mechanics computer modeling. *Abstracts 1993 Meeting on RNA Processing* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY). 144*a*.

Harvey, S. C. 1989. Treatment of electrostatic effects in macromolecular modeling. *Protein Struct. Funct. Genet.* 5:78–92.

Harvey, S. C., J. Luo, and R. Lavery. 1988. DNA stem-loop structures in oligopurine-oligopyrimidine triplexes. *Nucleic Acids Res.* 16: 11795–11809.

Heus, H. A., and A. Pardi. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science.* 253: 191–194.

Hinds, D. A., and M. Levitt. 1992. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci., USA.* 89:2536–2540.

Hingerty, B., R. S. Brown, and A. Jack. 1978. Further refinement of the structure of yeast tRNA(Phe). *J. Mol. Biol.* 124:523–534.

Holbrook, S. R., C. Cheong, I. Tinoco, Jr., and S.-H. Kim. 1991. Crystal structure of an RNA double helix incorporating a track of non-Watson-Crick base pairs. *Nature.* 353:579–581.

Hsieh, C.-H., and J. D. Griffith. 1989. Deletions of bases in one strand of duplex DNA, in contrast to single base mismatches, produces highly kinked molecules: possible relevance to the folding of single-stranded nucleic acids. *Proc. Natl. Acad. Sci. USA.* 86:4833–4837.

Hubbard, J. M., and J. E. Hearst. 1991a. Computer modeling 16S ribosomal RNA. *J. Mol. Biol.* 221:889–907.

Hubbard, J. M., and J. E. Hearst. 1991b. Predicting the three-dimensional folding of transfer RNA with a computer modeling protocol. *Biochemistry.* 30:5458–5465.

James, T. L. 1991. Relaxation matrix analysis of two-dimensional nuclear Overhauser effect spectra. *Curr. Opinion Struct. Biol.* 1:1042–1053.

Joshua-Tor, L., D. Rabinovich, H. Hope, F. Frolow, E. Appella, and J. L. Sussman. 1988. The three-dimensional structure of a DNA duplex containing looped-out bases. *Nature.* 334:82–84.

Kalnik, M. W., D. G. Norman, B. F. Li, P. F. Swann, and D. J. Patel. 1990. Conformation transitions in thymidine bulge-containing deoxytridecanucleotide duplexes. Role of flanking sequence and temperature in modulating the equilibrium between looped out and stacked thymidine bulges. *J. Biol. Chem.* 265:636–647.

Kim, S.-H., and T. R. Cech. 1987. Three-dimensional model of the active site of the self-splicing rRNA precursor of *Tetrahymena. Proc. Natl. Acad. Sci. USA.* 84:8788–8792.

Kirkpatrick, S., C. D. J. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671–680.

Lavery, R. 1987. DNA flexibility under control: the Jumna algorithm and its application to BZ junction. *In* Unusual DNA Structures. R. D. Wells and S. C. Harvey, editors. Springer-Verlag, New York. 189–206.

Levitt, M. 1969. Detailed molecular model for transfer ribonucleic acid. *Nature.* 224:759–763.

Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.

Levitt, M., and A. Warshel. 1975. Computer simulation of protein folding. *Nature.* 253:694–698.

Lustig, B., D. G. Covell, and R. L. Jernigan. 1992. Lattice method for conformation generation of RNA. *Biophys. J.* 61:150*a*. (Abstr.)

Major, F., M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. 1991. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science.* 253:1255–1260.

Malhotra, A., H. A. Gabb, and S. C. Harvey. 1993. Modeling large nucleic acids. *Curr. Opinion Struct. Biol.* 3:241–246.

Malhotra, A., and S. C. Harvey. 1994. A quantitative model of the *Escherichia coli* 16S RNA in the 30S ribosomal subunit. *J. Mol. Biol.* In press.

Malhotra, A., R. K.-Z. Tan, and S. C. Harvey. 1990. Prediction of the three dimensional structure of *Escherichia coli* 30 S ribosomal subunit—a molecular mechanics approach. *Proc. Natl. Acad. Sci. USA.* 87:1950–1954.

Malhotra, A., R. K.-Z. Tan, and S. C. Harvey. 1991. Prediction of the three-dimensional structures of ribonucleic acids: from tRNA to 16S ribosomal RNA. *In* Molecular Dynamics: applications in Molecular Biology. J. M. Goodfellow, editor. Macmillan Press, London.

Malhotra, A., R. K.-Z. Tan, and S. C. Harvey. 1994. Utilization of shape data in molecular mechanics using a potential based on spherical harmonic surfaces. *J. Comp. Chem.* 15:190–199.

Max, N. L., and E. D. Getzoff. 1988. Spherical harmonic molecular surfaces. *IEEE Comp. Graph. Appl.* 8:42–50.

McCammon, J. A., and S. C. Harvey. 1987. Dynamics of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, England.

Metzler, W. J., and D. R. Hare. 1989. Limited sampling of conformational space by the distance geometry algorithm: implications for structures generated from NMR data. *Biochemistry.* 28:7045–7052.

Miller, M., R. W. Harrison, A. Wlodawer, E. Appella, and J. L. Sussman. 1988. Crystal structure of 15-mer DNA duplex containing unpaired bases. *Nature.* 334:85–86.

Moore, P. B. 1988. The ribosome returns. *Nature.* 331:223–227.

Morden, K. M., and K. Maskos. 1993. NMR studies of an extrahelical cytosine in an A.T rich region of a deoxyribodecanucleotide. *Biopolymers.* 33:27–36.

Morden, K. M., B. M. Gunn, and K. Maskos. 1990. NMR studies of a deoxyribodecanucleotide containing an extrahelical thymidine surrounded by an oligo(dA).oligo(dT) tract. *Biochemistry.* 29:8835–8845.

Nagano, K., M. Harel, and M. Takezawa. 1988. Prediction of three-dimensional structure of *Escherichia coli* ribosomal RNA. *J. Theor. Biol.* 134:199–256.

Oakes, M. I., and J. A. Lake. 1990. DNA-Hybridization Electron Microscopy—Localization of 5 Regions of 16-S rRNA on the Surface of 30-S Ribosomal Subunits. *J. Mol. Biol.* 211:897–906.

Oakes, M. I., L. Kahan, and J. A. Lake. 1990a. DNA-hybridization electron Microscopy—tertiary structure of 16-S rRNA. *J. Mol. Biol.* 211:907–918.

Oakes, M. I., A. Scheinman, T. Atha, G. Shankweiler, and J. A. Lake. 1990b. Ribosome structure: Three-dimensional locations of rRNA and proteins. *In* The Ribosome: Structure, Function & Evolution. W. E. Hill, A. Dahlberg, R. A. Garrett, P. B. Moore, D. Schlessinger, and J. R. Warner, editors. ASM Press, Washington, DC. 123–133.

Olson, W. K., and P. J. Flory. 1972. Spatial configurations of polynucleotide chains. I. Steric interactions in polyribonucleotides: a virtual bond model. *Biopolymers.* 11:1–23.

Olson, W. K. 1982. Theoretical studies of nucleic acid conformation: potential energies, chain statistics, and model building. *In* Topics in Nucleic Acid Structures: Part 2. S. Neidle, editor. 123–133.

Peritz, A. E., R. Kierzek, N. Sugimoto, and D. H. Turner. 1991. Thermodynamic study of internal loops in oligoribonucleotides—symmetric

loops are more stable than asymmetric loops. *Biochemistry*. 30: 6428–6436.

Puglisi, J. D., J. R. Wyatt, and I. Tinoco, Jr. 1990a. Conformation of an RNA pseudoknot. *J. Mol. Biol.* 214:437–453.

Puglisi, J. D., J. R. Wyatt, and I. Tinoco, Jr. 1990b. Solution conformation of an RNA hairpin loop. *Biochemistry*. 29:4215–4226.

Rice, J. A., and D. M. Crothers. 1989. DNA bending by the bulge defect. *Biochemistry*. 28:4512–4516.

Richards, F. M. 1977. Area, volume, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* 6:151–176.

Romby, P., D. Moras, M. Bergdoll, P. Dumas, V. V. Vlassov, E. Westhof, J.-P. Ebel, and P. Giege. 1985. Yeast $tRNA^{Asp}$ tertiary structure in solution and areas of interaction of the tRNA with aspartyl-tRNA synthetase. A comparative study of the Yeast phenylalanine system by phosphate alkylation experiments with ethylnitrosourea. *J. Mol. Biol.* 184:455–471.

Roy, S., V. Sklenar, E. Appella, and J. S. Cohen. 1987. Conformational perturbation due to an extra adenosine in a self-complementary oligodeoxynucleotide duplex. *Biopolymers*. 26:2041–2052.

Saenger, W. 1984. Principles of Nucleic Acid Structure. Springer-Verlag, New York.

Schellman, J. A. 1974. Flexibility of DNA. *Biopolymers*. 13:217–226.

Schevitz, R. W., A. D. Podjarny, N. Krishnamachari, J. J. Hughes, P. B. Sigler, and J. L. Sussman. 1979. Crystal structure of a eukaryotic initiator tRNA. *Nature*. 278:188–190.

Sharp, K. A., and B. Honig. 1990. Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* 19: 301–332.

Skolnick, J., and A. Kolinski. 1990. Simulations of the folding of a globular protein. *Science*. 250:1121–1125.

Stern, S., B. Weiser, and H. F. Noller. 1988. Model for the three-dimensional folding of 16S ribosomal RNA. *J. Mol. Biol.* 204:447–481.

Stern, S., T. Powers, L.-M. Changchien, and H. F. Noller. 1989. RNA-Protein interactions in 30S ribosomal subunits: folding and function of 16S rRNA. *Science*. 244:783–790.

Stöffler-Meilicke, M., and G. Stöffler. 1990. Topology of the ribosomal proteins from *Escherichia coli* within the intact subunits as determined by immunoelectron microscopy and protein-protein cross-linking. *In* The Ribosome: Structure, Function & Evolution. W. E. Hill, A. Dahlberg, R. A. Garrett, P. B. Moore, D. Schlessinger, and J. R. Warner, editors. ASM Press, Washington, DC. 123–133.

Sussman, J. L., S. R. Holbrook, R. W. Warrant, G. M. Church, and S.-H. Kim. 1978. Crystal structure of yeast phenylalanine tRNA. I. Crystallographic refinement. *J. Mol. Biol.* 123:607–630.

Tan, R. K.-Z., and S. C. Harvey. 1989. Molecular mechanics model of supercoiled DNA. *J. Mol. Biol.* 205:573–591.

Tan, R. K.-Z., and S. C. Harvey. 1990. Succinct models: modeling of supercoiled DNA. *In* Theoretical Biochemistry and Molecular Biophysics. R. Lavery and D. Beveridge, editors. Adenine Press, New York.

Tan, R. K.-Z., and S. C. Harvey. 1993. Yammp: development of a molecular mechanics program using the modular programming method. *J. Comp. Chem.* 14:455–470.

Tang, R. S., and D. E. Draper. 1990. Bulge loops used to measure the helical twist of RNA in solution. *Biochemistry*. 29:5232–5237.

van den Hoogen, Y. T., A. A. van Beuzekom, E. de Vroom, G. A. van der Marel, J. H. van Boom, and C. Altona. 1988. Bulge-out structures in the single-stranded trimer AUA and in the duplex (CUGGUGCGG)·(CCGC-CCAG). A model-building and NMR study. *Nucleic Acids Res.* 16: 5013–5030.

Varani, G., B. Wimberly, and I. Tinoco, Jr. 1989. Conformation and dynamics of an RNA internal loop. *Biochemistry*. 28:7760–7772.

Vorobjev, Y. N. 1990a. Block-Units method for conformational calculations of large nucleic acid chains. I. Block-Units approximation of atomic structure and conformational energy of polynucleotides. *Biopolymers*. 29:1503–1518.

Vorobjev, Y. N. 1990b. Block-Units method for conformational calculations of large nucleic acid chains. II. The two-hierarchical approach and its application to conformational arrangement of the unusual TΨC loop of Rabbit tRNAVal. *Biopolymers*. 29:1519–1529.

Watson, J. D., and F. H. C. Crick. 1953. A structure for deoxyribose nucleic acid. *Nature*. 171:737–738.

Westhof, E., P. Dumas, and D. Moras. 1985. Crystallographic refinement of Yeast aspartic acid transfer RNA. *J. Mol. Biol.* 184:119–145.

Westhof, E., P. Romby, P. J. Romaniuk, J.-P. Ebel, C. Ehresmann, and B. Ehresmann. 1989. Computer modeling from solution data of Spinach Chloroplast and of Xenopus laevis Somatic and Oocyte 5 S rRNAs. *J. Mol. Biol.* 207:417–431.

Wimberly, B., G. Varani, and I. Tinoco, Jr. 1993. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry*. 32:1078–1087.

Woo, N. H., B. A. Roe, and A. Rich. 1980. Three dimensional structure of *E. coli* initiator tRNA(fMet). *Nature*. 286:346–351.

Woodson, S., and D. M. Crothers. 1988. Structural model for an oligonucleotide containing a bulged guanosine by NMR & energy minimization. *Biochemistry*. 27:3130–3141.

Yonath, A., W. Bennett, S. Weinstein, and H. G. Wittmann. 1990. Crystallography and image reconstructions of ribosomes. *In* The Ribosome: Structure, Function & Evolution. W. E. Hill, A. Dahlberg, R. A. Garrett, P. B. Moore, D. Schlessinger, and J. R. Warner, editors. ASM Press, Washington, DC. 123–133.

You, T. J., and S. C. Harvey. 1993. Finite element approach to the electrostatics of macromolecules with arbitrary geometries. *J. Comp. Chem.* 14:484–501.

Zhang, P., and P. B. Moore. 1989. An NMR study of the Helix V loop E region of the 5S RNA from *E. coli*. *Biochemistry*. 28:4607–4615.