Modeling loop structures in proteins and nucleic acids: an RNA stem-loop

I. Haneef, Simon J. Talbot* and Peter G. Stockley*

*Astbury Department of Biophysics and *Department of Genetics and Biotechnology Unit, University of Leeds, Leeds, UK*

We have used a novel modeling technique, based on combining information from several preexisting structures, to generate a three-dimensional (3D) model for the RNA stem-loop responsible for translational repression of the MS2 RNA bacteriophage replicase. Specific features of the model have been tested experimentally by chemical and enzymatic structural probes; results from these experiments have been used to "improve" the model by fixing initial assumptions. The new model and chemical modification data are in part consistent, and further predictions are being tested. The modeling algorithm has a wide range of potential applications, particularly to loop regions in proteins and nucleic acids.

Keywords: mathematical molecular modeling, homology modeling, loop regions, molecular simulations

INTRODUCTION

When the RNA bacteriophage MS2[1] infects *E. coli*, the viral RNA acts as a messenger for several rounds of protein synthesis. Yet despite the stoichiometric occurrence of the phage genes, the production of phage encoded proteins is tightly regulated by a series of control events that result in efficient production of each protein only in the amounts required physiologically. One such control mechanism is the translational repression of the replicase subunit synthesis by the phage coat protein. Repression is achieved by binding one- or two-coat protein subunits to a linear fragment of viral RNA located between the end of the coat protein gene and the start codon for replicase. In a series of elegant experiments Uhlenbeck and his colleagues,[2] working with the MS2 variant R17, have shown that all the coat protein binding activity resides in a fragment of just 19 bases, which can be stably folded into a single stem-loop structure (Figure 1). The fragment is bound in a sequence-specific fashion, and interaction with the coat protein can be monitored by filter-binding assays. These properties have made this system one of the best understood sequence-specific RNA-protein complexes.

Further experiments by Uhlenbeck and his colleagues[3] have addressed the sequence specificity of the interaction. Using many sequence variants, they have shown that all the sequence specificity lies in just four of the 19 bases. The crucial features that must be preserved are as follows:

(1) the base-pairing potential of the stem (i.e. any change in one side of the stem must be compensated by a change to the complementary base on the other side)
(2) the mismatched base on the 5′ side must be a purine
(3) the penultimate 3′ base of the four residue loop must be a pyrimidine
(4) the first and the last bases in the loop must be adenines.

Only two of the 19 bases in this fragment are base specific. The requirement of (3) can be rationalized by the evidence for formation of a covalent bond to a protein cysteine group,[4] but the other requirements are still puzzling.

We are currently using mutagenesis of the coat protein to probe the RNA recognition event. As a guide to further experiments, it was necessary to consider the structure of the RNA site, which raised several problems of modeling the loop region. In this paper we present an algorithm that can be applied to loops in both nucleic acids and proteins. We have used this algorithm and molecular simulations to generate a three-dimensional (3D) model of the RNA stem-loop. The validity of this model has been tested by biochemical structure-probing experiments.[5]

Our approach to the RNA modeling is based on a similar problem in the modeling of proteins, and the algorithm presented here is applicable to both. The modeling of homologous proteins, where the sequence homology is high, is best done using one of the known protein structures with highest sequence homology.[6] However, the modeling of proteins where only a low sequence homology exists presents a number of problems. Comparison of tertiary structures of homologous proteins shows that the 3D structures are more conserved than the protein sequences.[7] This is particularly true for core or framework regions of the molecules. The core or framework regions of an unknown structure can be constructed from structurally homologous regions

Address reprint requests to Dr. Haneef at the Astbury Department of Biophysics, University of Leeds, Leeds LS2 9JT, UK.
Received 24 May 1989; accepted 21 June 1989

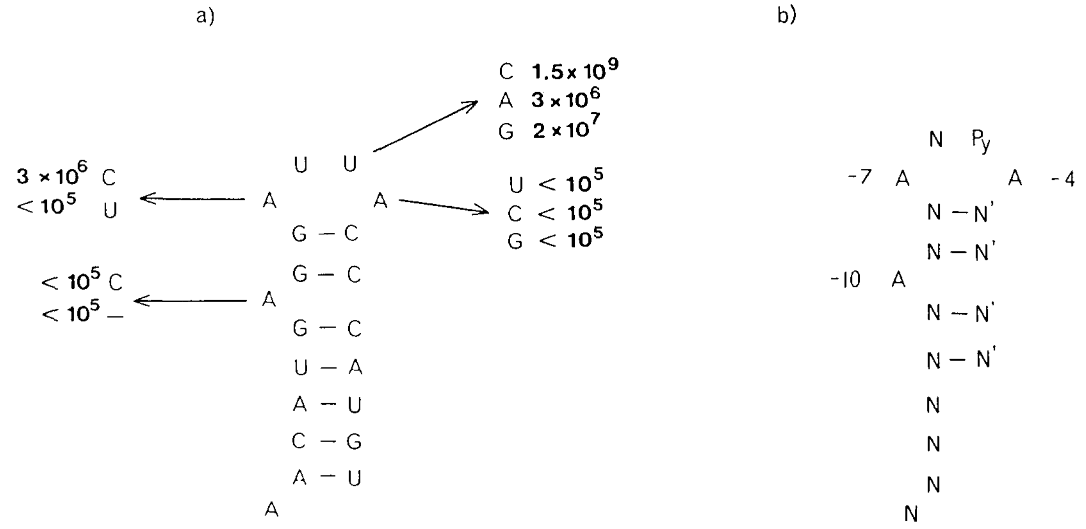

Figure 1. Proposed secondary structure of the stem-loop of MS2 RNA, which contains all the information to bind protein specifically. (a) The natural stem-loop sequence showing the results of sequence changes on the coat protein affinities. (b) The minimum sequence information required for specific binding to the protein (N is any base and N' its complementary base; Py = pyrimidine). Figures indicated are K_a (M^{-1})

from several homologous protein molecules for which the 3D structures are known. Loop or surface regions, where sequence homology is generally poor,[8] cannot in general be constructed in this manner.

Once the various fragments of the framework region(s) of an unknown structure have been built, a number of problems still exist. The correct sequence of the unknown structure needs to be built into the framework. Further, the various fragments of the framework are generally not contiguous in sequence and require modeling of, for example, loop regions that connect these fragments. Finally, the complete model needs to be either regularized or energy minimized; since the model-built structure may differ significantly from the correct structure, minimization methods are required that have a large radius of convergence. In this paper we address the problem of constructing a framework of an unknown structure using information from several known structures and obtaining a regularized structure that retains important interatomic interactions. We also describe a molecular dynamics simulation technique that can be used to impose characteristic interatomic restraints, such as those between hydrogen-bonded atoms and between Cα to Cα or C1* to C1*. The methods described have particular applications to modeling loop regions and side chains.

METHODS

Novel modeling technique

The model of the 19-mer RNA fragment was built using a locally written program, MNYFIT (Haneef, unpublished) and with the aid of the interactive computer graphics program FRODO.[9] The model was constructed on the basis of known tRNA structures in the Brookhaven Protein Data Bank.[10] We chose tRNA^{Asp} as the major basis of structural information, as this has been refined at the highest resolution.[11]

Figure 1 shows the proposed secondary structure of the 19-mer. On the basis of this, the stem region of the structure (residues −15 to −11, and −1 to 4) was built as an A-RNA-type helical structure (the stem-loop is numbered relative to the first base at the replicase gene, A, +1). Although this region can be constructed in several ways, it was built on to the 5-base helical structure composed of residues 27 to 31 and residues 39 to 43 of tRNA^{Asp} (Figure 2). Base substitutions were made by using the sugar-phosphate backbone as the frame; the bases were built by means of an automatic procedure (Haneef, unpublished), with ideal geometries and preferred conformations that maximize hydrogen-bonding capacity of the complementary bases. The two G.C. base pairs (residues −9, −8, −3 and −2) were built in a similar fashion. However, two regions of the structure, namely the adenine at −10 and the four base loop, required *de novo* modeling.

The iterative least-squares fitting and distance geometry averaging methods of Haneef[12,13] were used to construct the sugar-phosphate backbone for the loop (residues −7 to −4). The latter method involves using distance geometry techniques, where the interatomic distances are taken from several known structures, to construct an average structure. This is an elegant method for constructing a mathematically unique average of several structures such that the error function

$$E = \sum_{j=1}^{NMOLS} \sum_{i=1}^{NATOMS} (X_i - R_j Y_{ji})^2 \qquad (1)$$

is at the global minimum, where X_i represents the coordinates of the ith atom in the average structure, Y_{ij} the coordinate of the ith atom in the jth molecule, and R_j is the rotation matrix for superposing Y_j onto X. This method gives identical results to the iterative least-squares fitting method and has the advantage that the interatomic distance information can be obtained from any number of sources, and

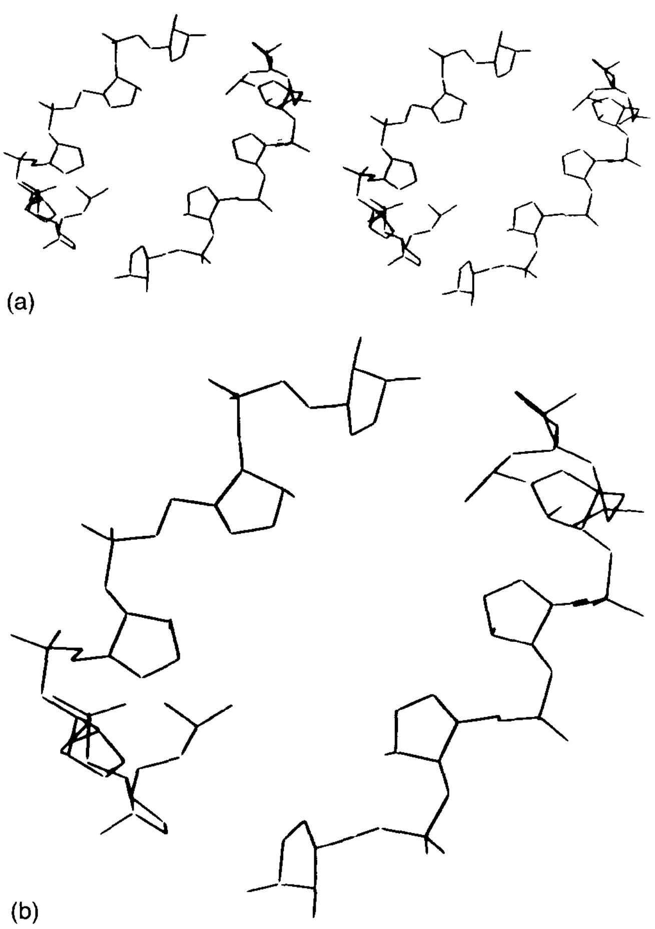

Figure 2. Sugar-phosphate backbone of residues 27–31 and 39–43 of tRNA^{Asp} used as framework for the stem region of 19-mer RNA fragment. (a) Stereo (b) mono

Table 1. Sources of interatomic distances for constructing loop region

Residue numbers	Equivalent MS2 fragment residues	Source (Brookhaven code)
55 to 56	−7 to −6	4*TNA*
55 to 56	−7 to −6	6*TNA*
55 to 56	−7 to −6	9*TNA*
33 to 36	−7 to −4	4*TNA*
33 to 36	−7 to −4	6*TNA*
33 to 36	−7 to −4	9*TNA*

that the sources of such information need not have the same number of atoms. Since the method is based upon interatomic distances, no least-squares fitting of the various molecules is required.

The interatomic distances were taken from loop regions in the known tRNA structures (Table 1); an average distance matrix D was constructed using interatomic distances from these fragments. The ijth elements of matrix D are defined by

$$D_{ij}^2 = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} d_{kij}^2 \tag{2}$$

where d_{kij} is the distance between ith and jth atom in the kth fragment, and n_{ij} is the number of sources from which ijth distance is obtained. Using this $n \times n$ distance matrix, an $n \times n$ metric matrix G was constructed, where

$$G_{ij} = \frac{(d_{io}^2 + d_{jo}^2 - D_{ij}^2)}{2} \tag{3}$$

where d_{io} is the distance to center of mass

$$d_{io}^2 = \frac{1}{n}\sum_{j=1}^{n} D_{ij}^2 - \frac{1}{n^2}\sum_{j=2}^{n}\sum_{k=1}^{j-1} D_{jk}^2 \tag{4}$$

The framework for the loop was obtained by extracting three eigensolutions of the G matrix corresponding to the three largest eigenvalues. The framework structure is, thus, defined by

$$F_{ij} = l_j^{1/2} w_{ij} \qquad 1 \le i \le n, j = 1, 2, 3 \tag{5}$$

where l_j is the jth eigenvalue of the G matrix and w_{ij} is the ith component of the corresponding eigenvector. This procedure, in general, gives a framework that has grossly distorted structure with far from ideal stereochemistry. The loop structure, thus obtained, was then regularized, subject to interatomic distance constraints. The interatomic distances defining bond lengths and bond angles were taken from the work of Weiner *et al.*[14]; the long-range distance constraints were taken from the loop regions in known tRNA structures (Table 1). These constraints comprised upper and lower interatomic distances in the known structures. The regularization algorithm used was a modified SHAKE[15] method (see Appendix). This method gave the global minimum structure for which all interatomic constraints of the form

$$d_{lij} \le d_{ij} \le d_{uij} \tag{6}$$

were satisfied, where d_{lij} and d_{uij} are lower and upper bounds, respectively, for interatomic distances. The regularized loop structure is shown in Figure 3a; superposition of the various fragments used to construct the loop is shown in Figure 3b.

We considered two possibilities in modeling the adenine at position −10. This base could be intercalated or looped out. Recent work with DNA oligonucleotides containing mismatched adenines has resulted in the determination of both types of structure (even for the same oligonucleotide sequence[16,17]), depending on whether the structure was determined in solution by NMR (intercalated) or from X-xay diffraction of single crystals (stacked out). On the basis of energetic considerations and the observation that bases in tRNA structures maximize base stacking, the former possibility was given preference in our model. Interatomic distances from known intercalation sites in tRNA structures were used to build this intercalation site (residues −11 to −9 and −2 to −1). Although there are a number of intercalation sites in the tRNA structures (Table 2), none involve two strands of the same helical region. To obtain an intercalated adenine in an A-type helical structure required embedding supplementary distances into the distance matrix. Specifically, these extra data were for interatomic dis-

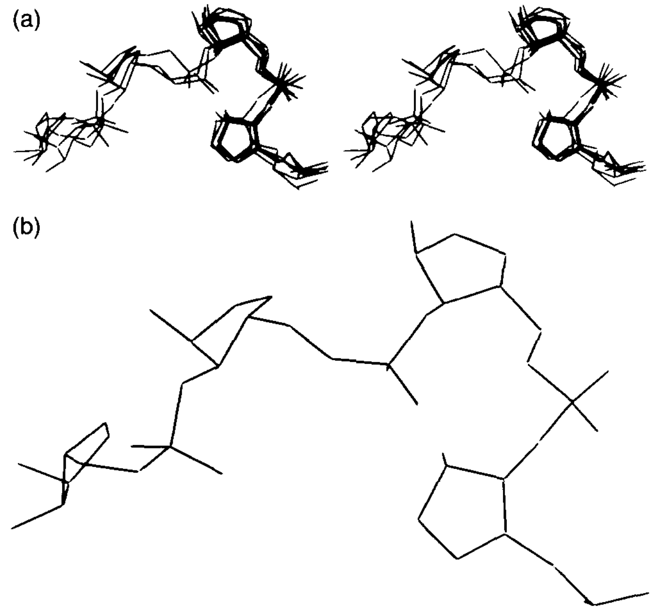

Figure 3. (a) Superposed sugar-phosphate backbone fragments from tRNA structures used to construct the loop framework. The two-residue fragments are taken from three-base-long loop regions, and the four-residue fragments are taken from five-base-long loop regions of tRNA structures (see table 1). (b) The regularized framework for the four-base loop region

Table 2. Sources of interatomic distances for construction of intercalation site

Residue numbers	Equivalent MS2 fragment residues	Source (Brookhaven code)
45 to 46	−1 to −2	4*TNA*
45 to 46	−1 to −2	6*TNA*
45 to 46	−1 to −2	9*TNA*

tances between residues −11 and −1, between −9 and −2, and between residues −11 to −9. These distances were taken to be those for an A-type helical structure. An approximate structure of this fragment was then obtained using familiar distance geometry techniques and was then regularized to have ideal stereochemistry (Figure 4). For regularization, we employed upper and lower bounds for the interatomic distances characteristic of intercalation sites and A-type helical structures; to orient the residues −1 and −2 correctly with respect to the A-type helical region −11 to −9, we also used C1* to C1* distances characteristic of G.C. base pairs.

The various fragments of the 19-mer were then fitted using the interactive computer graphics program FRODO. To achieve this we used the stem-loop region, residues 27-43, of tRNA^{Asp} as a model frame. Fitting the intercalation site into the structure was relatively easy; here we used residues −11 and −1 from the intercalation site and fitted these onto the same two residues also defined in the stem region. Similarly, the two G.C. base pairs were fitted onto the intercalation site by *superposing* one of the G.C. base pairs (residues −9 and −2) onto the same base pair also defined in the intercalation site. In constructing the complete stem region with the intercalation site, we gave preference to coordinates of residues from the framework of the intercalation site wherever these residues had been defined twice. The four-residue loop was fitted on top of the stem-loop in a manner that would maximize base stacking and provide a reasonable continuation of the phosphate backbone. Finally, the bases were built onto the four-base loop and −10 site using an automated procedure. The completed model was subjected to one further round of regularization; optimal base stacking and hydrogen bonding for complementary base pairs was achieved using the SHAKED algorithm (see Appendix) by employing upper and lower bounds for interatomic distances found in known tRNA structures.

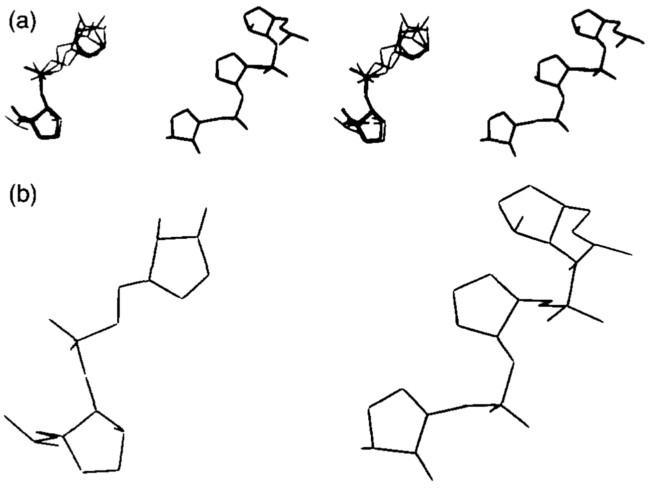

Figure 4. (a) Superposed sugar-phosphate backbone fragments of intercalation sites from known tRNA structures (see table 2) after being fitted to an A-type helical region. (b) The regularized framework for the intercalation site

Molecular dynamics in torsion angle space

To obtain a model structure with low energy, we used both Cartesian space and a novel torsion angle space simulation technique. Although most molecular dynamics techniques for macromolecules are carried out in Cartesian space, here we describe in general terms how dynamics simulations can be carried out using any set of coordinate space.

Consider the situation for a molecule whose phase space at time t is defined by displacement r and velocity v; here, r is the generalized displacement, which may be torsional or linear, and v is its associated velocity. Because the potential function $U(r)$ is known, we may calculate forces and torques and use these to predict the displacements at time $t + dt$. We use the leap-frog formulation of the Verlet algorithm,[18]

$$v\left(t + \frac{dt}{2}\right) = v\left(t - \frac{dt}{2}\right) + a(t)dt \qquad (7)$$

$$r(t + dt) = r(t) + v\left(t + \frac{dt}{2}\right) dt \qquad (8)$$

where $a(t)$ is the general form of the acceleration and is proportional to the force F

$$F = -\frac{dU(r)}{dr} \tag{9}$$

For conservative force fields, the force F is a function of the displacements only. Our analytical method for calculating the derivatives of the potential energy with respect to torsion angles is approximately four times more efficient than that of Abe *et al.*[19] on scalar machines, and it has the advantage that all computationally demanding parts of the algorithm vectorize on vector processing computers such as the Cray-1. The method involves calculating the first derivatives with respect to the Cartesian coordinates and projecting these onto the torsion angle space. This requires $3/(2N(N - 1))$ multiplications for the Cartesian derivatives, plus $9 \times M \times M(m)$ multiplications for projections onto torsion angle space (where N is the number of atoms in the molecule, M the number of torsion angles, and $M(m)$ the number of atoms dependent on a particular torsion angle). The method of Abe *et al.* requires $6N(N - 1)$ multiplications. In our quenched dynamics simulations, we used a time step, dt, of 0.002 ps for Cartesian space calculations with all bond lengths constrained, and a time step of 0.02 ps for dynamics simulations in torsion angle space.

Experimental structure probing

Various structural features of the model were tested using a number of chemical modification techniques.[5] RNA molecules were produced using either sp6 RNA polymerase and a DNA clone that contained the stem-loop sequence or by direct automated chemical synthesis using the approach of Ogilvie.[20] The full size RNA fragments were purified from acrylamide gels after radiolabelling at either the 5′ or 3′ end using $[\gamma^{32}P]$ATP or $[5'^{32}P]$pCp, respectively. Treatment with either chemical reagents or nucleases was then used to assess surface accessibility or conformation of the fragments.

RESULTS

Examination of the model showed that the loop region formed a rather tight four-residue loop; the close proximity of the two A's in the loop suggested that these bases might form a non–Watson-Crick base pair. One possible hydrogen bonding scheme for the two A's in the loop could be realized by having A-7 in the *anti* conformation and A-4 in the *syn* conformation, in which the N7 atoms of the two A's were sequestered in the non–Watson-Crick base pairing. The model was adjusted on the PS300 graphics system; to obtain good hydrogen bonds and maximal base stacking required some major changes in the backbone and ribonucleoside of the loop framework, resulting in a rather close contact between the phosphates of residues −6 and −7. The complete structure is shown in Figure 5. As expected, the model has a number of features typically found in tRNA structures. The base stem is an A-RNA helical structure and extends in an approximate fashion as far as the four-base loop. The unbase-paired A at position −10 has an approximate A-RNA structure, with the A intercalating with the G.C. base pairs on either side of it. The two A's of the loop are stacked above the G.C. base pair, with the two U's in the loop in a stacking arrangement on the 3′ side of the stem. Finally, we note that the phosphate of U-5 stacks onto the A-7, a conformation similar to that observed in the anticodon of known tRNA structures. The phosphate of A-4 lies slightly above the plane defined by the two-loop adenines in van der Waals contact with the two As, approximating to a conformation also observed in the anticodon loop of known tRNA structures.

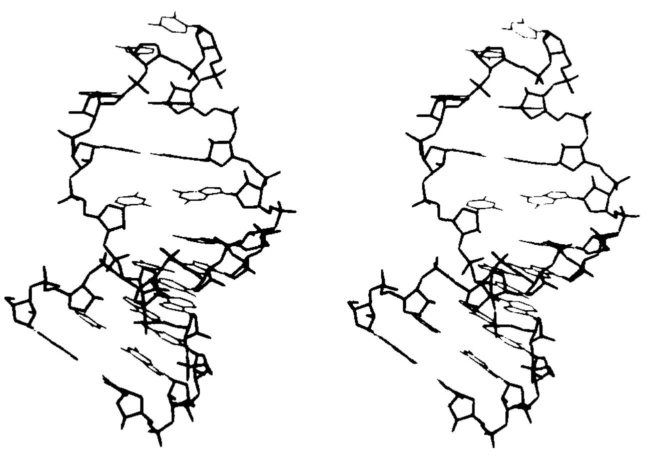

Figure 5. The complete model of 19-mer fragment constructed using the fragments shown in Figs. 2–4. This model was built using computer graphics using residues 27–43 of $tRNA^{Asp}$ as a guide. The structure has been regularized to have ideal stereochemistry

This model was then subjected to a number of energy-minimization protocols, including quenched dynamics simulations, to obtain a structure with low energy. Initial minimizations were carried out in Cartesian space, using the program EMPMDS,[21] which employs the potential function of Weiner *et al.* From these studies we were able to establish regions in the molecule that would exhibit a significant degree of strain. These regions comprise the four-residue loop and the 3′ side of the stem in the vicinity of the intercalation site. Solvent-accessibility calculations with varying probe sizes were used to establish regions of the molecule that might be accessible to chemical probes (Table 3). The various structural features of the model were then tested using chemical modification methods.

Chemical modification studies with diethyl pyrocarbonate (DEPC) were used to assess the accessibilities of the N7 atoms in the A's.[22] Contrary to the prediction from the model structure, the N7s of the A's in the loop region were not sequestered, suggesting that the two A's in the loop are not hydrogen-bonded or, possibly, that the particular non–Watson-Crick hydrogen bonding scheme we chose initially was wrong. Better agreement with the experimental accessibilities could be obtained by small changes to the model structure in which the two A's were still hydrogen bonded, with the N7's accessible from the solvent; in the new hydrogen bonding scheme the N1's are now sequestered (Figure 6).

The adenine at −10 exhibits two extreme reactivities in DEPC. At low pH, ionic strength and divalent cation con-

Table 3. Accessible surface areas and pseudo first-order rate constants of modification of N7 atoms of adenines; figures in parentheses are for the old model. Accessibilities in Å^2, probe size in Å, and rate constants in 10 000/sec. The rate constants are measured in two different buffers: TMK is 100 mM Tris at pH 8 with 80 mM KCl and 10 mM MgCl and 50 μg/ml of tRNA; D1 is 50 mM Na Cacodolate at pH 6.8 with 1 mM EDTA and 50 μg/ml tRNA

	Probe size			Rate constant	
Residue	1.4	2.8	5.6	*TMK*	*D*1
A − 10	7(11)	2(6)	0(0)	9.6	1.5
A − 7	15(0)	14(0)	3(0)	2.3	12.0
A − 4	16(0)	15(0)	14(0)	7.0	7.5

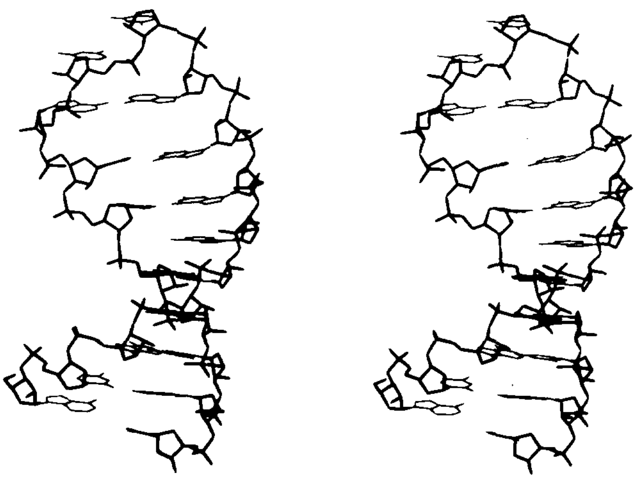

Figure 6. The new model of 19-mer after energy minimization and torsion angle space quenched dynamics simulations

centrations there is little modification at position − 10. However, at pH 8, higher ionic strength and divalent ion concentration the A at −10 is heavily modified. The degree of chemical modification by DEPC has been quantitated by assuming that modification follows first order kinetics.[5] These results show that the loop adenines and the 5′ mismatched adenine are much more reactive toward the reagent (up to fivefold) than single-stranded adenines in the test fragment. We believe that this is due to the detailed mechanism of DEPC modification, which requires the involvement of a nucleophile to make modification irreversible and detectible (by subsequent cleavage with aniline). For single-stranded adenines the nucleophile would be water. The higher reactivity of the adenines in the loop and intercalation site of the MS2 fragment can be rationalized in terms of nucleophiles supplied by neighboring bases (i.e., the adenine in the loop and adjoining guanines at the intercalation site. This hypothesis leads us to propose that the A at −10 is hyperreactive when intercalated but much less reactive when looped out or extruded from the stem). Experiments are at hand to test this idea.

We have tested for any distortions of the molecule from the A-type helical structure by digestions with the single-strand and double-strand specific nucleases S1 and V1, respectively. Cleavage by V1 occurs on the 3′ side of the stem but not on the corresponding 5′ side, suggesting that the intercalation of adenine at −10 does distort the helical conformation of the molecule.

The various experimental tests on the model clearly pointed out one major mistake in our model. In this model the hydrogen bonding arrangement for the loop adenines was chosen arbitrarily, largely dictated by the computational cost of computer simulations that would be incurred in studying all possibilities. Our *de novo* modeling of the loop region of the molecule had suggested only that the two adenines were sufficiently close in proximity that the two might hydrogen-bond. Three possibilities exist: One scheme would sequester N7's of both A's, another would sequester N7 of one and N1 of the other A and the last of these would sequester the N1's of both A's. Using the assumption that the two A's form two hydrogen bonds, and our experimental results on the accessibility of the N7s that show that both N7's are accessible, our current hydrogen bonding scheme is based on the latter arrangement in which the two N1's in the two adenines are sequestered. This prediction should also be verifiable experimentally.

Our calculations on the model structure also suggested a number of limitations in our simulation techniques. Namely, the Cartesian space molecular mechanics and quenched dynamics studies did not show significant shifts from the starting model. In particular, the two phosphate groups of residues −6 and −7 remained rather close. To alleviate these problems, we carried out all further calculations in torsion angle space using the program EMPTOR (Haneef, unpublished). The new model was energy minimized, followed by quenched dynamics simulations in torsion angle space. Due to the artificially high temperatures (ca. 1000K) used in such simulations, we employed restraining terms of the type

$$E = \sum_{ij} k(d_{ij} - d_{ij0})^2 \qquad (10)$$

where d_{ij} is the distance between atoms i and j at time t, and d_{ij0} the average "equilibrium" distance obtained from the energy-minimized structure; k is the restraining force constant. These restraints were applied to all hydrogen-bonding pairs of atoms and C1* distances between base pairs and also included all interphosphate distances. Such restraints could not be used in Cartesian space quenched dynamics simulations; our attempts to do so resulted in grossly distorted stereochemistry for the molecule. In particular, the bases were forced to be nonplanar, and major changes in torsion angles in the sugars occurred, producing structures that did not conform to sugar puckers observed in tRNA structures. The final energy-minimized structure obtained from torsion angle studies is shown in Figure 6. The model retains many of the features of our first model structure. The main areas of change are the four-base loop and the site of intercalation. The phosphate of A-4 has moved into the plane defined by the two-loop adenines, and the short contact between the phosphates of −6 and −7 has been relieved. The improved model predicts a series of surface accessibilities for phosphates in the backbone that is reasonably consistent with the biochemical data; the accessibilities for N7's of adenines suggest that they would be reactive toward DEPC as observed but do not necessarily account for the degree of reactivity. Distortions of the backbone and strain could be reflected in S1/V1 digestions. It is interesting to note that the new conformation of the loop obtained from torsion angle studies has smaller RMS difference (1.8 Å) with the loop framework than the loop conformation in our first model (2.2 Å), which was obtained after computer graphics modeling and Cartesian space simulations.

Does the model structure shed any light on the specific base requirements in the 19-mer RNA fragment? The essential requirement of -5 base to be a pyrimidine can be rationalized by the evidence for formation of a covalent link to a protein cysteine group. The energetics of base stacking strongly favors a purine at position -10; three of the four sites of intercalation in the known tRNA structures involve purines with maximal base stacking. The base specificity at -7 and -4 is also explained. The large size of the purine bases, and maximal base stacking with the A-type helical stem confers considerable stability to the structure. Due to the rather tight nature of the four-base loop, a non–Watson-Crick base pairing has a similar effect, as does the stacking of phosphate at position -5 onto the purine at position -7. The model also suggests a possible reason for the essential requirement that both bases at the top of the stem be As. In our model the distance between C1*'s of the two As is 13.2 Å; modeling Gs in place of As with similar C1* distances shows that such an arrangement would lead to a severe steric clash between the N2 group and the phosphate of -4. Adenines, therefore, help to maintain the structure of the top of the stem but may also be involved in direct interactions with protein side chains.

Further modeling studies with an A at position -4 and a G at position -7 show that optimal base stacking and hydrogen bonding between these two residues would require C1* to C1* distances in the range of 8–10 Å; such distances in the loop region result in short contacts between the phosphates of residues -6 and -7. Our modeling studies, therefore, make the prediction that the recognition event involving the coat protein requires C1* distance between the first and the last residues of the loop to be *ca.* 13 Å; this prediction is currently being tested with the aid of chemically modified bases.

DISCUSSION

Starting from the proposed secondary structure of the RNA fragment involved in translational repression of the MS2 RNA bacteriophage replicase, we have constructed a 3D model based on known tRNA structures. The modeling technique is novel and employs knowledge of atomic interactions, in terms of interatomic distances. We have tested the model experimentally and improved it by altering some of our initial assumptions. Such a technique is of wider applicability and can be used to model build protein structures based on the knowledge of side chain–side chain interactions such as those described by Singh *et al.*[23] Using distance geometry techniques, we are able to use specific information such as distances between C1*'s that are characteristic of specific base pairs; the use of the averaging technique in Cartesian space would produce a grossly distorted structure, which can require substantial regularization and remodeling to achieve specific characteristic interactions or contact distances so common in biomolecules.

A knowledge of such interactions, and the availability of information from databases such as the SERC Protein Engineering Club sequence and structural database, can be used in conjunction with distance geometry techniques to propose model structures of biomolecules; the technique is particularly suitable for constructing functionally important regions of molecules where characteristic interactions or contact distances might be expected to play a major role in determining the (local) structure. Test studies using distance geometry techniques to build several inhibitors to the active sites of aspartyl proteases (Haneef, Foundling, Cooper and Blundell, unpublished) have established the technique as a powerful and versatile tool.

However, we emphasize the need to verify such models if possible. In our study functionally important regions of the molecule required *de novo* modeling; although the distance geometry and molecular mechanics techniques provided a reasonable starting model, experimental tests of the model suggested important changes that have led to a model structure for the RNA fragment that is reasonably consistent with available experimental data on the solution structure of the 19-mer RNA. Further biochemical studies are in hand to probe the accessibilities of adenines at N1 and the functional groups of the other loop bases (i.e., the uridines). These data should allow further improvements of this initial model.

Due to the inordinate computer resources that would have been required otherwise, our molecular mechanics studies were limited to *in vacuo* simulations; such simulations are routinely used in modeling studies to give structures with good stereochemistry and low potential energy. However, it is clear from the results of our experiments that we need to carry out simulations where the solvent and the counter ions are explicitly represented if we are to account for the two differing conformations of the molecule in different pH and cation concentrations. Further, our initial simulations were based on Cartesian space techniques that have a rather small radius of convergence; therefore, these simulations did not show significant shifts from the initial proposed model. Although Cartesian space simulations are eminently suitable for most purposes, simulation techniques with much larger radius of convergence are required for models that may differ significantly from the correct structure. To alleviate some of these difficulties, we have developed a computationally efficient molecular mechanics and molecular dynamics technique in torsion angle space.

In recent years, there has been considerable interest in modeling loop regions in proteins using molecular mechanics techniques. In this regard our quenched dynamics technique in torsion angle space holds considerable promise, although here we have shown its application to RNA. Cartesian space molecular mechanics and molecular dynamics simulations caused only modest changes to the loop region in our initial model structure. However, our torsion angle quenched dynamics simulations proved a powerful approach to obtaining an energy-minimized structure that conforms better with the available experimental data; the model structure makes a number of predictions that are currently being investigated in our laboratory. Our experimental work has established the presence of two conformations of the RNA fragment, implying considerable flexibility in the molecule. Subject to verification of the various predictions made by our current model structure, we shall use molecular dynamics simulations to study the conformational flexibility. The use of torsion angle dynamics allows us to use time steps that are 10–20 times larger than those conventionally used in Cartesian space simulations. The availability of large time-scale motions from torsion angle dynamics makes possible comparison with motional behavior of the molecule studied by NMR techniques. Sufficient quantities of the RNA fragment may soon be available for NMR studies; the final test of the model is the determination of the structure of the stem-loop directly using NMR techniques or X-ray diffraction.

APPENDIX

The usual methods for constraining interatomic distances, such as those obtained from NMR studies, involve the use of pseudo-energy terms of the form

$$\begin{aligned} V(r_{ij},r_{iju}) &= K(r_{ij} - r_{iju})^2 && \textit{if } r_{ij} \geq r_{iju} \\ &= 0 && \textit{if } r_{ij} \leq r_{iju} \\ V(r_{ij},r_{ijl}) &= 0 && \textit{if } r_{ij} \geq r_{ijl} \\ &= K(r_{ij} - r_{ijl})^2 && \textit{if } r_{ij} \leq r_{ijl} \end{aligned}$$

where r_{iju} and r_{ijl} are the upper and lower limits for interatomic distances for atoms i and j, respectively. Although such methods have been used with some success in such studies, the global minimum is not attained due to entrapment in local energy minima. In our studies we use a modified form of the SHAKE algorithm to constrain all interatomic distances, where short interatomic distances are those for a structure with ideal geometry and long-range distances are taken, wherever possible, from known structures. The algorithm iteratively applies corrections to the coordinates until all distance constraints of the type

$$d_{ijl} \leq d_{ij} \leq d_{iju}$$

are satisfied. This algorithm fails under the same conditions as the original SHAKE method (i.e., when the distance constraints are inconsistent or if the trial structure is very different from the structure that will obey all the distance constraints). In our experience of applying this method to obtain a regularized average of several similar structures, such conditions do not arise.

```
C     ******************************************************************
      SUBROUTINE   S H A K E D
C     ******************************************************************
*
*     SUBROUTINE   SHAKEDistances
*
*     Based on:    Subroutine SHAKE of Berendsen & van Gunsteren
*     Ref:         Ryckaert, Ciccotti, Berendsen
*                  J Computat Phys
*                  Vol 23 (1977) 327
*
*     This version: M I J Haneef
*                  Astbury Dept Biophysics
*                  University of Leeds
*
*     SHAKED       will apply corrections to coordinates of a  molecule
*                  such that all distance constraints of the form:
*                  DLij < Dij < DUij
*                  are satisfied,  where DLij is the lower limit for
*                  distance between atoms I and J, and DUij is the
*                  upper limit.
*
*     Natoms       No. of atoms in molecule
*     Maxfn        Maximum no. of iterations that will be attempted
*     Ndists       No. of distances to be SHAKEn
*     Ijdist       Array of IJ pair of atoms defining distances
*     Skip         Work array
*     Tolerr       Return from routine if error function value drops
*                  below  TOLERR,  where error function FUN is  the
*                  sum(min(abs(dij-dijl),abs(dij-diju)))  over all
*                  ij distances
*     Xa           Input coordinates of molecule; returned with
*                  SHAKEn coordinates
*     Xb           Work array
*     Dist         Array of upper and lower bounds for distances
*                  between IJ pair of atoms
*
      PARAMETER    (MAXATS=1000,MAXDST=100000)
      PARAMETER    (ZERO=0.,TINY=1.E-6,HALF=0.5)
      LOGICAL      DONE,SKIP
      COMMON       /INTGA/
     -             NATOMS,MAXFN,NDISTS,IJDIST(MAXDST,2)
      COMMON       /LOGCA/
     -             DONE,SKIP(MAXATS,2)
      COMMON       /REALA/
     -             TOLERR,XA(3,MAXATS),XB(3,MAXATS),DIST(MAXDST,2)
      WRITE(6,1)
1     FORMAT('1',///,45X,'****** REFINING AVERAGE STRUCTURE ******',///,
     -        10X,'ITERATION NO.    ERROR VALUE',/)
      DO 2 I=1,NATOMS
      SKIP(I,1)=.TRUE.
      SKIP(I,2)=.FALSE.
      XB(1,I)=XA(1,I)
      XB(2,I)=XA(2,I)
      XB(3,I)=XA(3,I)
2     CONTINUE
```

```
      DONE=.FALSE.
      NIT=0
3     NIT=NIT+1
      IF (DONE) THEN
      ELSE
      IF (NIT.GT.MAXFN) THEN
      STOP 'MORE THAN MAXFN ITERATIONS REQUIRED '
      END IF
      DONE=.TRUE.
      FUN=ZERO
      DO 4 K=1,NDISTS
      I=IJDIST(K,1)
      J=IJDIST(K,2)
      IF (SKIP(I,2).AND.SKIP(J,2)) THEN
      ELSE
      DHI=DIST(K,1)**2
      DLO=DIST(K,2)**2
      XBX=XB(1,I)-XB(1,J)
      XBY=XB(2,I)-XB(2,J)
      XBZ=XB(3,I)-XB(3,J)
      RPIJ=XBX*XBX+XBY*XBY+XBZ*XBZ
      IF (DLO.LE.RPIJ.AND.RPIJ.LE.DHI) THEN
      ELSE
      FUN=FUN+MIN(ABS(RPIJ-DHI),ABS(RPIJ-DLO))
      XAX=XA(1,I)-XA(1,J)
      XAY=XA(2,I)-XA(2,J)
      XAZ=XA(3,I)-XA(3,J)
      RRPR=XAX*XBX+XAY*XBY+XAZ*XBZ
      IF (RRPR.LE.ZERO) THEN
      STOP 'DEVIATION TOO LARGE'
      END IF
      IF (ABS(RPIJ-DHI).GE.ABS(RPIJ-DLO)) THEN
      ACOR=HALF*(DIST(K,1)-SQRT(RPIJ))/RRPR
      ELSE
      ACOR=HALF*(DIST(K,2)-SQRT(RPIJ))/RRPR
      END IF
      XHX=XAX*ACOR
      XHY=XAY*ACOR
      XHZ=XAZ*ACOR
      XB(1,I)=XB(1,I)+XHX
      XB(2,I)=XB(2,I)+XHY
      XB(3,I)=XB(3,I)+XHZ
      XB(1,J)=XB(1,J)-XHX
      XB(2,J)=XB(2,J)-XHY
      XB(3,J)=XB(3,J)-XHZ
      SKIP(I,1)=.FALSE.
      SKIP(J,1)=.FALSE.
      DONE=.FALSE.
      END IF
      END IF
4     CONTINUE
      WRITE(6,5) NIT,FUN
5     FORMAT(13X,I10,1X,F14.2)
      DO 6 I=1,NATOMS
      SKIP(I,2)=SKIP(I,1)
      SKIP(I,1)=.TRUE.
6     CONTINUE
      DONE=FUN.LE.TOLERR
      GO TO 3
      END IF
      DO 7 I=1,NATOMS
      XA(1,I)=XB(1,I)
      XA(2,I)=XB(2,I)
      XA(3,I)=XB(3,I)
7     CONTINUE
      END
```

ACKNOWLEDGEMENTS

I.H. is a research fellow funded by the EEC. This work was supported in part by SERC grant GR/D 86331 to P.G.S. S.J.T. holds an SERC quota studentship. The authors would like to thank Drs. S.E.V. Phillips and C.W.G. Fishwick for stimulating discussions. It is a pleasure to acknowledge Dr. Jus Singh and Devon Carney for free and uninhibited exchange of ideas in the true spirit of science; thanks are also due to Dr. Steve Foundling, with whom the first pilot studies of the distance geometry techniques were initiated.

REFERENCES

1 Bernardi, A. and Spahr, P.F. *PNAS* 1972, **69**, 3033

2 Carey, J., Lowary, P.T. and Uhlenbeck, O.C. *Biochem.* 1983, **22**, 4723

3 Uhlenbeck, O.C., Carey, J., Romaniuk, P.J., Lowary, P.T. and Beckett, D.B. *J Biomol Struct Dyn* 1983, **1**, 539

4 Romaniuk, P.J. and Uhlenbeck, O.C. *Biochem.* 1985, **24**, 4239

5 Talbot, S.J., Haneef, I., Medina, G. and Stockley, P.G. Submitted to *Eur J Biochem*

6 Chothia, C. and Lesk, A.M. *EMBO J* 1986, **5**, 823

7 Lesk, A.M., and Chothia, C. *J Mol Biol* 1980, **136**, 225

8 Bajaj, M. and Blundell, T.L. *Chem Rev Biophys Bioeng* 1984, **13**, 453

9 Jones, T.A. *J Appl Crystallogr.* 1978, **11**, 268

10 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shi-

manouchi, T. and Tasumi, M. *J Mol Biol* 1977, **122**, 535

11 Moras, D., Comarmond, M.B., Fischer, J., Weiss, R., Thierry, J.C., Ebel, J.P. and Giege, R. *Nature* 1980, **288**, 669

12 Haneef, I. and Sutcliffe, M.J. *Inf Quart Protein Crystallog* 1986, **18**, 11

13 Sutcliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. *Protein Engineering* 1987, **1**, 377

14 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. *J Am Chem Soc* 1984, **106**, 765

15 Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. *J Compt Phys* 1977, **23**, 327

16 Patel, D.J., Kozlowski, S.A., Marky, L.A., Rice, J.A., Broka, C., *et al. Biochem.* 1982, **21**, 445

17 Joshua-Tor, L., Rabinovich, D., Hope, H., Frolow, F., Apella, E. and Sussman, J. *Nature* 1988, **334**, 82

18 Verlet, L. *Phys Rev* 1967, **159**, 82

19 Abe, H., Braun, W., Noguti, T. and Go, N. *Comp. and Chem.* 1984, **8**, 239

20 Usman, N., Ogilvie, K.K., Tiang, M-Y. and Cedergren, R.G. *J. Am. Chem. Soc.* 1987, **109**, 7845

21 Haneef, I. Ph. D. thesis, University of London, 1985

22 Ehrenberg *et al. Prog Nuc Acid Res* 1976, **16**, 189

23 Singh, J., Thornton, J.M., Snarey, M. and Campbell, S.F. *FEBS* 1987, **224**, 161