Patel, D. J., Kozlowski, S. A., Ikuta, A., & Itakura, K. (1984) *Biochemistry 23*, 3207-3217.
Plateau, P., & Guéron, M. (1982) *J. Am. Chem. Soc. 104*, 7310-7311.
Privé, G. G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M. L., & Dickerson, R. E. (1987) *Science 238*, 498-504.
Radman, M., & Wagner, R. (1986) *Annu. Rev. Genet. 20*, 523-538.
Scheek, R. R., Boelens, R., Russo, N., van Boom, J. H., & Kaptein, R. (1984) *Biochemistry 23*, 1371-1376.
Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 804-808.
Sklenář, V., & Feigon, J. (1990) *Nature 345*, 836-838.
Sowers, L. C., Fazakerley, G. V., Kim H., Dalton L., & Goodman M. F. (1986) *Biochemistry 25*, 3983-3988.
Su, S.-S., Lahue, R. S., Au, K. G., & Modrich, P. (1988) *J. Biol. Chem. 263*, 6829-6835.
Tibanyenda, N., de Bruin, S. H., Haasnoot, C. A. G., van der Marel, G. A., van Boom, J. H., & Hilbers, C. W. (1984) *Eur. J. Biochem. 139*, 19-27.
Ts'o, P. O. P., Rappaport, S. A., & Bollum, F. J. (1966) *Biochemistry 5*, 4153-4160.
van der Marel, G. A., van Boeckel, C. A. A., Wille, G., & van Boom, J. H. (1981) *Tetrahedron Lett. 22*, 3887-3890.
von Hippel, P. H., & McGhee, J. D. (1972) *Annu. Rev. Biochem. 41*, 231-236.
Webster, G. D., Sanderson, M. R., Skelly, J. V., Neidle, S., Swann, P. F., Li, B. F., & Tickle, I. J. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 6693-6697.
Woodson, S. A., & Crothers, D. M. (1988) *Biochemistry 27*, 436-445.

Predicting the Three-Dimensional Folding of Transfer RNA with a Computer Modeling Protocol†

John M. Hubbard‡ and John E. Hearst*
Department of Chemistry, University of California, Berkeley, California 94720
Received July 10, 1990; Revised Manuscript Received March 11, 1991

ABSTRACT: We have developed a computer modeling protocol that can be used to predict the three-dimensional folding of a ribonucleic acid on the basis of limited amounts of secondary and tertiary data. This protocol extends the use of distance geometry beyond the domain of NMR data in which it is usually applied. The use of this algorithm to fold the molecule eliminates operator subjectivity and reproducibly predicts the overall dimensions and shape of the transfer RNA molecule. By use of a replacement pseudoatom set based on helical substructures, a series of transfer RNA foldings have been completed that utilize only the primary structure, the phylogenetically deduced secondary structure, and five long-range interactions that were determined without reference to the crystal structure. In a control set of foldings, all the interactions suspected to exist in 1969 have been included. In all cases, the modeling process consistently predicts the global arrangement of the helical domains and to a lesser extent the general path of the backbone of transfer RNA.

The ability of single-stranded RNA[1] to form intramolecular hydrogen bonds gives it a much greater conformational variability than double-stranded DNA. This versatility combined with the large number of nucleotides in most cellular RNAs presents us with a formidable problem as we attempt to probe the form/function relationships of RNA. The sequence (primary structure), Watson-Crick base-pairing pattern (secondary structure), and compact folded conformation (tertiary structure) of RNA have been the focus of extensive research. With the vast improvements in DNA sequencing technology, the determination of the primary sequences of RNAs from the analysis of genomic DNA has become routine. With use of the thermodynamics of base stacking (Tinoco et al., 1973), it is possible to evaluate the possible secondary structures that an RNA may form. Improved empirical parameters and computer programs now make it possible to produce RNA foldings that correspond to the native hydrogen-bonding patterns with some accuracy (Jaeger et al., 1989). But the phylogenetic determination of secondary structure by comparison of the sequences of RNA molecules with common functions from different species is still the most productive technique. Phylogeny may also indicate that some bases are involved in tertiary or noncanonical base pairing (Gutell & Woese, 1990). Folded RNA molecules can also be probed for three-dimensional relationships with a variety of chemical and enzymatic techniques. Still the determination of the fully folded conformation of ribonucleic acids remains a difficult problem in spite of the rapidly increasing amount of structural information.

There are very few well-established tertiary RNA structures. The average conformational RNA A-form helix is known from fiber diffraction data (Arnott et al., 1973). Recently, the structures of two RNA oligomer duplexes have been determined (Dock-Bregeon et al., 1989; Happ et al., 1988). They generally conform to the A-form helix with local variations in stacking similar to those seen in the structures of DNA oligomers. Most significantly, the structures of phenylalanine,

† This work was supported in part by the Director, Office of Energy Research, Office of General Life Sciences, Molecular Biology Division of the U.S. Department of Energy under contract No. DE AC03-76SF00098.
* To whom correspondence should be addressed.
‡ Present address: Department of Biochemistry and Biophysics, College of Physicians and Surgeons, Columbia University, New York, NY 10032.

[1] Abbreviations: RNA, ribonucleic acid; NMR, nuclear magnetic resonance; cgr, conjugate gradient refinement; rms, root mean square; DHU, dihydrouridine; TΨC, ribothymidine-pseudouridine-cytosine; vdW, van der Waals; R_g, radius of gyration.

0006-2960/91/0430-5458$02.50/0

aspartic acid, glycine, and initiator f-methionine transfer RNAs from yeast and f-methionine tRNA from *Escherichia coli* have been determined by X-ray crystallography (Sussman et al., 1978; Moras et al., 1980; Schevitz et al., 1979; Woo et al., 1980; Wright et al., 1979). As a biologically important molecule, tRNA is the touchstone for all RNA modeling and it is reassuring to find that base stacking and the A-form helix are the predominant motifs in this RNA structure. Phylogeny also proved to be a very reliable predictor of the secondary structure of transfer RNA (Zachau et al., 1966). Given the constraints that such a hydrogen-bonding pattern would impose on the three-dimensional conformation of tRNA and some knowledge of the tertiary interactions in tRNA, several researchers tried to predict its structure (Ninio et al., 1969). One farsighted approach attempted to model the three-dimensional structure of tRNA by use of a combination of physical and computer models followed by empirical energy minimization (Levitt, 1969), but no one successfully predicted a structure for tRNA similar to the one that was revealed by X-ray crystallography (Kim et al., 1973).

Advances in computer technology have improved molecular modeling by making it possible to replace physical models with computer constructs at an earlier stage in the process. A traditional approach is to proceed up through the levels of structure, replacing a flat secondary structure model with computer-generated helices. Interactive graphics modeling can then be used to dock the helical subunits with the appropriate single-stranded connectors. Tertiary folding as dictated by long-range interactions would then produce the final model. Ideally, energy minimization might then be used to ensure a reasonable structure. As graphical modeling is not inherently different from physical modeling, it can only increase the ease of modeling, not the quality. Empirical energy modeling of DNA oligonucleotides shows that while the fine details of energy minimization and computer modeling are improving, they alone cannot yield the necessary quantitative discrimination required to distinguish among folded conformations (Srinivasan & Olson, 1987). Although some techniques (e.g., NMR or X-ray crystallography) can supply the density of information necessary to produce unique structures for small molecules, the amount of tertiary data for larger molecules remains very sparse. Therefore, a modified modeling approach must be developed if better results than those of twenty years ago are to be obtained. The introduction of distance geometry into the protocol moves the problem of folding the molecule into *N*-dimensional space, where all the primary, secondary, and tertiary constraints are simultaneously satisfied. The resultant three-dimensional structure is then the product of an objective mathematical embedding algorithm. This approach is superior to molecular mechanics or molecular dynamics approaches, which require that the underlying empirical energy algorithms be modified and recalibrated to include three-dimensional data (Brünger et al., 1986). Any attempt to predict the tertiary structure of RNA molecules must be able to predict the form of transfer RNA if any credence is to be placed in the protocol. Therefore, the modeling of tRNA with the updated protocol is used to demonstrate its feasiblity and to explore possible representations and essential variables.

Materials and Methods

The primary structure of yeast phenylalanine transfer RNA contains 76 nucleotides, including eight modified ribonucleotides and six nonstandard bases. As tRNAs can be charged with the appropriate amino acid and competently participate in translation without these special nucleotides

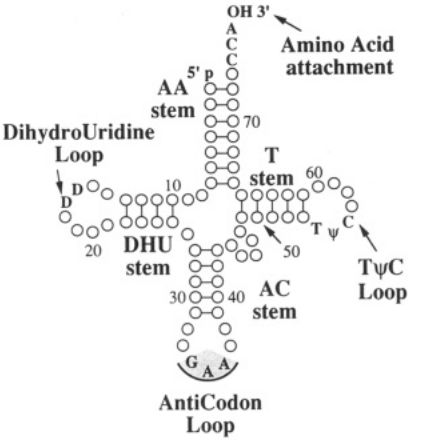

FIGURE 1: Secondary structure of yeast phenylalanine tRNA displayed as a cloverleaf. The groups responsible for the names of substructures are shown, and the anticodon bases are shaded.

(Sampson et al., 1989), standard RNA residues are used instead.

The secondary structure depiction of transfer RNA is formed by displaying the 76 residues of the yeast phenylalanine tRNA in a three-leaf cloverleaf arrangement (Figure 1). The four base-paired stems contain 42 of the nucleotides, leaving 34 single-stranded residues. The basic modeling assumption is that the double-stranded regions will form stable A-form RNA helices that persist and dominate the folded structure. The NUCGEN module of AMBER is used to create helical subunits and reference helices from which average distances and constraints are derived for secondary structures.

Transfer RNA tertiary structure information is available from a variety of sources. A total of five long-range interactions derived from three different sources other than the X-ray crystal structure were used, and constraints were devised that would reflect the character of the RNA/RNA interactions.

In some tRNAs the eighth position is occupied by the modified nucleotide thiouridine. When irradiated with ultraviolet light, this residue may become cross-linked to the cytidine at position 13 (Yaniv et al., 1969). These UV-induced cross-links are the result of bond formation between the sulfur of the thiouridine and the cytosine base. As the sulfur–carbon covalent bond cannot exceed 2.0 Å, a cross-link requires the bases to be in direct contact but does not require that the carbon–phosphate backbone be helically related.

When transfer RNA is treated with psoralen and UV radiation, five cross-links are formed (Garrett-Wheeler et al., 1984). Four of the cross-links are in the stems of the cloverleaf structure and merely confirm the phylogenetic hydrogen-bonding pattern. The fifth cross-link between the uridine in the eighth position and the cytidine in the 48th position is due to the coaxial stacking of the amino acid acceptor stem and the ribothymidine stem. Usually, psoralen cross-links are found in helical structures but other base-stacking geometries can also be cross-linked and may indicate tertiary interactions. The cyclobutane rings that it forms are highly strained and the resultant geometry fixes the distance between the C5–C6 bonds of the cross-linked bases at approximately 7 Å.

Tertiary phylogenetic relationships are deduced on the basis of the covariance of associated bases among differing species. This assumes some sort of hydrogen bonding is involved. Upper and lower constraints for these interactions are based on an A-form helix. The three tertiary links, G15–C48, G18–Ψ55, and G19–C56, were selected from the set used in early attempts to model tRNA (Levitt, 1969) on the basis of

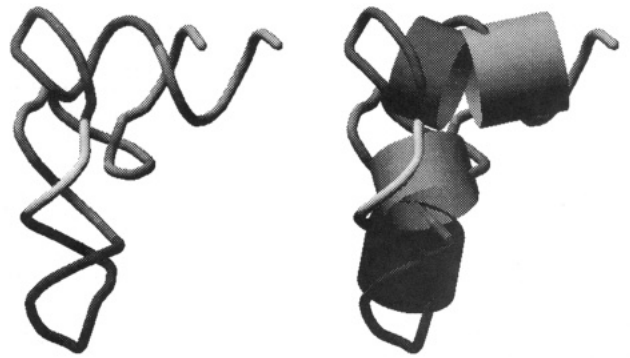

FIGURE 2: Conformation of the crystal structure of transfer RNA represented by a smoothed tube of 1-Å radius that traces the path of the phosphate atom backbone. In the right-hand display hollow cylinders 10-Å in radius have been added that approximate the helical substructures. The amino acid acceptor stem is to the right and the TΨC stem is stacked on it at the top of the structure. The anticodon loop is at the bottom of the inverted L.

the confidence placed in them by researchers at the time and the fact that they are confirmed by the crystal structure. This is justified, considering the improvement in our ability to statistically detect phylogenetic relationships (Haselman et al., 1989). The remaining seven additional relationships that were listed, A9–U12, A21–T54, C25–G57, C32–Ψ39, A44–G57, Ψ55–A58, and A73–A76, are used to generate a control group of structures.

As the standard of comparison, the crystal structure of yeast phenylalanine transfer RNA as further refined in 1978 (Sussman et al., 1978) was taken from the 1984 magnetic tape release of the Brookhaven Protein Data Bank. The structure was energetically minimized with AMBER to smooth out any inconsistencies caused by introduction of standard ribonucleotides (Figure 2).

A three-dimensional version of the cloverleaf representation of tRNA was constructed as a test of the traditional modeling procedures and provides an upper bound bench mark for unfolded conformations. Substituting A-form helices for the double-stranded regions is the first step in forming this structure. The helices and single-strand connectors were built separately with the LINK and NUCGEN modules of AMBER. The structural subunits were docked interactively on a computer graphics display. By use of the translations and rotations necessary for the visual docking, the original coordinate files were transformed and the separate files concatenated to form a single structure. No attempt was made to adjust this model to conform to the tertiary interactions, but it was minimized with AMBER to eliminate any irregularities caused by docking.

Folding the extended cloverleaf, so as to satisfy the long-range interactions, is the next step in the protocol. Folding nucleic acids should be easier than folding proteins since the helices are relatively stable and uniform subunits in which the sequence-dependent groups are hidden in the interior of the helix. Therefore, as a first approximation, constructing a three-dimensional nucleic acid model can be considered to be a packing of rigid cylinders that are linked by flexible single strands. Interactive graphical modeling is superior to physical modeling as a digital electronic representation does not have any weight or space limitations. Additionally, the changing atomic environment of small molecules can be followed quantitatively. But RNAs are not small molecules and interactive folding remains a highly subjective process that is dependent on the judgment and preferences of the modeler. It is at this stage that distance geometry will be introduced to automatically fold the molecule.

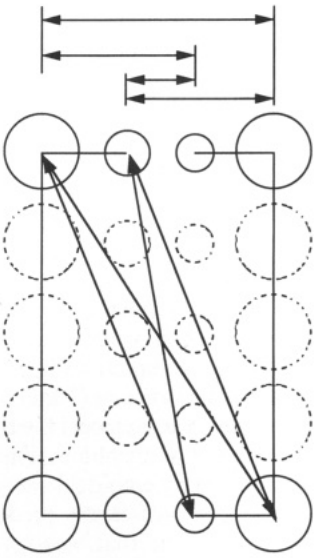

FIGURE 3: Schematic for the representation of secondary structures as pseudohelices. Only the atoms represented by solid spheres are retained as pseudoatoms. The horizontal arrows indicate constraints used to enforce double stranding, and the diagonal arrows indicate the distances that help to enforce helicity.

At the heart of the distance geometry algorithm is the distance matrix. It contains an entry for the distance from every atom of a structure to every other atom of the structure and consequently dominates the memory requirements of the computer program. As the size of the program increases as the square of the number of atoms and computation times increase at a rate proportional to the cube of the number of atoms, it is necessary to reduce the number of objects to be modeled. The replacement atom set should be chosen such that it is simple to make, reduces the size of the problem substantially, and gives added weight to helical substructures. In devising such a reduction scheme there are two major problems that must be considered. Foremost is the necessity of maintaining enough structural information so that the resultant model will still resemble the molecular folding in a significant way. Second, it must be possible to recover the full molecular structure uniquely and unambiguously. As a first step, single-stranded residues might be replaced by a single pseudophosphate. A segment of a helical strand of RNA can be represented by pseudoatoms that replace the phosphates of only the first and last residues of the strand. Pseudoatom representations for the hydrogen-bonding groups of these residues (the C4 atoms of pyrimidines or the N1 atom of purines) will also be retained. In this manner the helix length and twist will be specified and preserved in the residue definition. Double stranding is enforced by constraining the distances to the corresponding pseudophosphate and base-pairing group of the opposite strand. Implicit in such a drastic reduction scheme is the phylogenetic reasoning that a helix and not its specific sequence is most important. To ensure the proper orientation of such schematic strands, the base-pairing constraints will include not only the hydrogen-bond donor and acceptor pseudoatoms but also the distances between the pseudophosphates and the pseudophosphate to the hydrogen-bonding pseudoatom at both ends of the opposing strand (Figure 3). Of course, not all RNA helices are prefectly regular. In the crystal structure of tRNA, the 12 base pairs of the amino acid acceptor stem and the ribothymidine stem form a single, stacked unit that is clearly an A-form helix (Figure 2). But while the 3′ strand consists of a single contiguous unit (bases 61–72), the 5′ half is formed from two

unequal strands that are distant from each other in primary sequence (bases 1–7, 49–53) (Figure 1). Attempting to define the two halves of the 5′ strand of the amino acid acceptor stem as a single residue would require a very convoluted redefinition of the primary structure of tRNA and major modifications in the logic of the programs. By including the pseudophosphate and base for the residues of the 3′ strand that are involved in base pairing to a junction of some sort of the opposite strand, the necessary links for double-stranded constraints will be provided. This introduces new bonds, angles, and dihedrals that may be varied in the folding process. Therefore, it will allow irregular helices, kinks, and bulges to appear as required by the interplay of structural elements and secondary and tertiary constraints.

With the pseudohelical replacement scheme, the approximately 2500 atoms of yeast phenylalanine tRNA are reduced to a total of 100 atoms and pseudoatoms. The library of residue level replacements constructed from pseudoatoms was derived from the structure files created with the NUCGEN module of AMBER. As a means of compensating for the extreme simplicity of the replacement pseudoatoms, the number of conjugate gradient refinement cycles before restoring the all-atom representation is limited to 32 or 64 steps. With the pseudohelical residues and limited refinement it is possible to create 100 models in only 1.5 h of MicroVAX computer time. Each of the structures generated was visually examined. The structures were sorted on the basis of their distance constraint violations and how well they could be superimposed on the crystal structure. The bounds violations error equals the sum of the violations of the upper or lower bounds of all distance constraints. The fit error equals the sum over all atoms of the distance between an atom in a model structure and the same atom in the crystal structure after the two have been superimposed. The root mean square superposition deviation and the radius of gyration (R_g) were calculated in a conventional fashion.

In the final phase of the modeling process, the all-atom representation of the structure is restored. The first residue of the sequence is retained in an all-atom representation throughout the modeling process as a sort of 5′ end labeling. This makes it easy to locate the end of the molecule when it is drawn on a graphics terminal, and it also facilitates the reintroduction of the other nucleotides after the distance geometry manipulations are completed. The missing atoms of the pseudohelical residues are supplied by replacing the entire construct with a standard A-form helical strand of the appropriate sequence by simple superposition. The single-stranded pseudoatoms are replaced with complete nucleotides by AMBER. Finally, the all-atom structures are used to generate various graphical displays.

The molecular modeling package of programs, AMBER version 3.0, was obtained from Peter Kollman at UCSF. The distance geometry program, DSPACE (versions 1.3 and 2.1), written by Dennis Hare and Robert Morrison, was obtained from Hare Research, Inc. Although it was originally intended to perform the calculations on a high-performance computer, the large memory requirements restricted the modeling to a specially configured MicroVAX II. The line-drawing illustrations are screen dumps of DSPACE drawings on terminals emulating Tektronix video displays. The tube and cylinder drawings were created with the Molecular Cross Section program written by Michael Connolly obtained from the Scripps Clinic and Research Foundation.

RESULTS

The 100 trials, prefixed vt, produced 45 different structures and all have the distinctive L shape of the crystal structure. The bound violations for the set range from 85.2 to 248.0 Å, and the superposition fit on the crystal structure ranges from a minimum of 7.06 to 12.4 Å. The five with the lowest bounds violations are grouped between the low and 114.0 Å of bounds error. When displayed as line drawings the correspondence of the outlines of the models to that of the crystal structure is emphasized (Figure 4). The overall geometry, size, and shape are consistently predicted by the models and, on the basis of measurable quantities, the all-atom version of vt32 has physical characteristics that are very similar to those of the crystal structure (Table II). But at a more detailed resolution the models are less accurate. The amino acid acceptor and TΨC stems are stacked correctly, but the sharp bend of the L vertex often prevents the DHU stem from properly stacking on the anticodon stem. The correct chain path from 5′ to 3′ is approximated but is not followed in detail. Several of the conformers have problems with the positioning of the single-stranded loops, which is to be expected considering the simplicity of the pseudophosphate backbone. It is also clear that many structures are too compact and have serious van der Waals conflicts. The close approach of the 5′ and 3′ strands of the amino acid acceptor stem is a common problem and is directly attributable to poor space-filling characteristics of the pseudohelices. Simply increasing the penalty function weighting for vdW contacts did not prevent steric conflicts.

Table I: Total Bounds Violations and Superposition Fit Errors for Various Structures

structure	bounds error (Å)	fit error (Å)	no. of cgr steps
crystal	345.0	0.0	0
vt32	85.2	9.58	64
vt58	132.0	7.06	64
bad40	107.0	5.07	64
bad42	113.0	10.2	64
ext41	213.0	8.27	32
ext11	217.0	8.43	32

The model with the eighth-lowest total of bounds violations, vt58, has the best superposition fit on the crystal structure for this set of trials (Table I). An examination of the three dimensional path of the smoothed backbone reveals that the model has problems with the levels of structure below those of the helical constructs (Figure 5). In fact, because the pseudoatoms are not chiral, there are problems with the handedness of the helices in many of the models. In vt58 this can be seen in the anticodon and DHU stems, which form the vertical bar of the L. The abrupt change in chirality produces a kink at the junction between the anticodon stem and the anticodon loop at the bottom of the structure. The left-handed twist of this stem is responsible for the tight tangle in the junction between the vertical and horizontal helical stacks. But when considered as simple helical constructs, the correspondence with the conformation of the crystal structure is clear.

To see how dependent the results were on the cross-links chosen and to simulate the more likely instance where the quality of the cross-links is less certain, a new bounds matrix was constructed that included all the suspected relationships from 1969 (Levitt, 1969). The "bad" set of structures was created with a total of 12 tertiary distance constraints. Of the 100 structures created, 66 had an angular L shape while 34 had more linear conformations. Of the 48 different structures, 31 are of the bent variety. The remaining 17 structures have a rodlike or tangled conformation, and the majority of these resemble the model that Levitt constructed. The five structures with the lowest bounds violations have from 99.9 to 110.0 Å

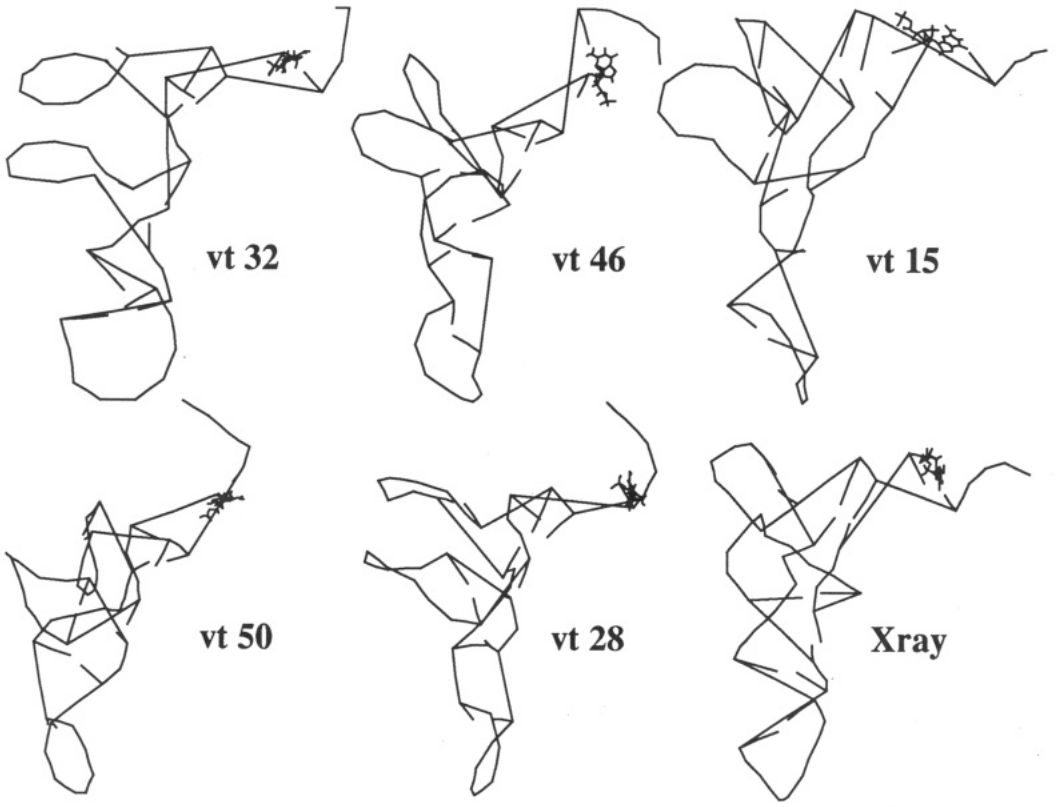

FIGURE 4: Five pseudohelical models of the vt set with the lowest error functions compared to the crystal structure in the lower right-hand corner. Only five long-range distance constraints were used to fold the molecule, and refinement was limited to 64 steps.

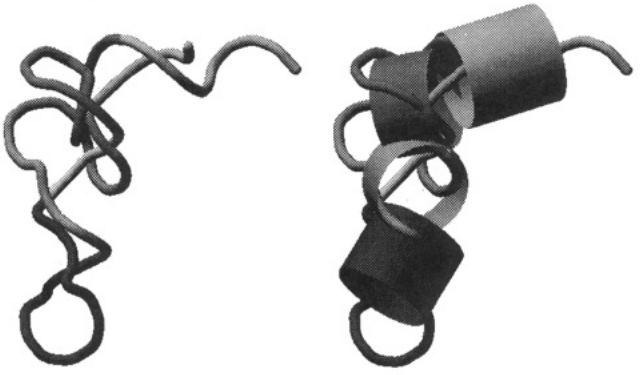

FIGURE 5: vt58 conformer shown as a smoothed tube in the same orientation and to the same scale as the crystal structure (Figure 2).

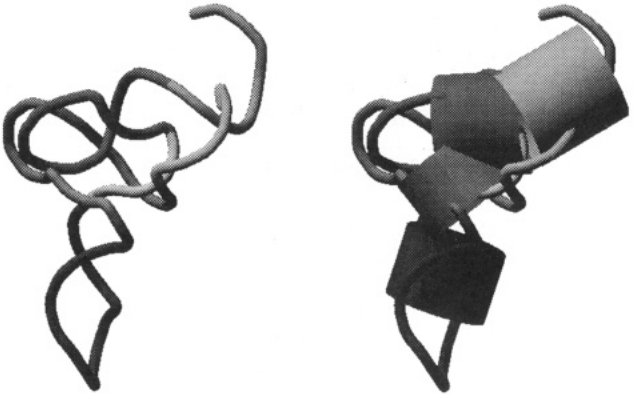

FIGURE 6: Smoothed backbone and backbone with cylinders display of the bad40 model drawn in the same orientation and to the same scale as the crystal structure (Figure 2). All the suspected long-range relationships circa 1969 were included in the set of distance constraints, and refinement was limited to 64 steps.

of error while the worst member of the set has 191.0 Å of bounds violations. Of the ten structures with low error functions, only bad42 is not a well-defined L. The worst of the these structures have error functions that are similar to the poorer bent conformers, but as with bad42, their superposition errors on the crystal structure are significantly worse (Table I). The fact that almost two-thirds of the structures evince the distinctive tRNA shape demonstrates that the use of distance geometry to fold tRNA is eliminating the subjectivity of a human modeler and that the structure is so firmly determined by the helical relationships that it can tolerate incorrect data.

The model with the fourth-lowest bounds violations total, bad40, is the structure that can best be superimposed on the reduced form of the tRNA crystal structure (Figure 6). It has a better superposition fit on the crystal structure than the best model of the previous set and the all-atom version of bad40 also has good physical dimensions (Tables I and II). The most obvious problem with bad40 is the tight tucking of the TΨC loop inside the DHU loop. Instead of sharply looping the junction between the 5′ half of the amino acid acceptor stem and the 5′ segment of the DHU stem back on itself as is found in the crystal structure, the model has the 5′ end of the molecule placed on the wrong side of the 3′ end. As a result, the 5′ segment of the amino acid stem has a left-handed twist. When considered as a collection of helices, bad40 does a better job of aligning the helices that form the vertical and

Table II: Physical Characteristics of the All-Atom Representations of Several Structures

conformer	volume (Å^3)	R_g (Å)	P/P separation (Å)
crystal	51 600	23.1	5.82
cloverleaf	70 000	25.6	6.62
vt32	52 700	23.3	5.76
bad40	49 100	22.7	5.79
ext41	53 300	23.4	5.77

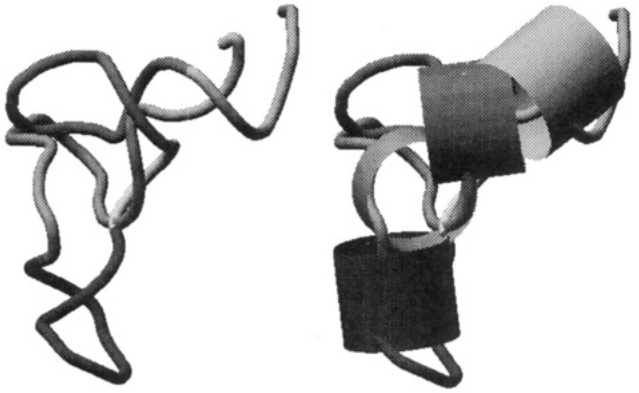

FIGURE 7: ext41 pseudohelical model. Only five tertiary distance constraints were used to fold the molecule, and conjugate gradient refinement was limited to 32 steps. The close approach of the strands in the inner vertex of the L of the molecule is exacerbated by the smoothing algorithm used to generate the display.

FIGURE 8: Smoothed backbone trace of the crystal structure shown with helical substructure cylinders of position and length obtained from a simple average of all the ext structures. The fit of the crystal structure helices on the vt58 backbone trace is shown on the right for comparison.

horizontal legs of the L. The nesting of the TΨC and DHU loops prevents a better spacing of the helices and causes an unrealistic overlap of the TΨC and DHU stems.

The continuing van der Waals overlaps suggested that perhaps even 64 steps of refinement were too many for such schematic models. Three more sets of 100-trial structures were generated with a newer issue of DSPACE (version 2.1) and the data set of five correct long-range distance constraints. The conjugate gradient refinement was done in two separate sets of 32 steps for these "ext" structures. After the first 32 steps of refinement, the computer-generated structures show similar ranges in the fit of one on another (rms = 81.3–120.0 Å) as for superposition on the crystal structure of tRNA (rms = 85.6–136.0 Å). The values of the error functions for the set of structures refined in this manner bracket the error value of 345.0 Å that the program assigns to the crystal structure when it is evaluated on the basis of pseudohelical constructs and tertiary constraints. Overall, the bounds violations for the ext set of structures varies from 213.0 to 453.0 Å while the fifth best structure has bounds errors totaling 263.0 Å. The new series of structures was similar to the previous series in that there were only 41, 47, and 50 different structures in each set. When the results for all the ext trials were collated, there was a total of only 60 independent conformations. Four of these structures are tangles that do not resemble the proper angular conformation.

The structure with the lowest bounds violations of this set, ext41, has the fourth-best fit on the crystal structure (Table I). At this stage in the refinement process, the bounds errors for ext41 are about double that of the best models from the previous series but its superposition fit and the physical characteristics of the all-atom version of ext41 have very similar values (Table II). A detailed examination of the folding of this model shows that it follows the backbone trace of the crystal structure very closely (Figure 7). The tucking of the junction between the amino acid stem and DHU stems inside the junction between the anticodon and TΨC is all that prevents the match from being nearly exact. This tangle is also responsible for the almost perpendicular orientation of the DHU and anticodon stems.

An additional 32 cycles of refinement does not substantially improve any of the structures and may actually distort them away from the ideal. For example, additional refinement of ext11 reduces its bounds violations to 108.0 Å but only marginally improves its superposition fit error to 7.63 Å. The ext11 structure is especially interesting because visual inspection revealed that the DHU and TΨC loops are knotted. Yet this serious defect does not produce the glaring numerical errors that would make it possible for the program to detect such an unrealistic structure. When the crystal structure is refined for 32 cycles against the bounds matrix, the adjusted structure shows the same distortions around the 5′ phosphate and anticodon loops as are seen in the computer generated structures when compared to the original crystal structure. In particular the loops are being forced into a rounder, less stacked conformation. The sharp kink in the junction between the amino acid stem and the DHU stem in the vertex of the L of the crystal structure is also severely changed by refinement. The two nucleotides in this junction must span the abrupt change in direction between the vertical and horizontal bars. Forcing these nucleotides to rigidly conform to the very simple nature of the pseudoatom replacements through continued refinement is a source of error in these models. After refinement, the bounds violations error function of the pseudoatom model of the crystal structure is decreased to 82.1 Å and has a superposition error of 1.89 Å when compared to the original structure.

Selecting the best conformer when the correct structure is unknown and refinement is stopped short of the convergence is not simple, although we can be confident that it will be among the 10 structures with the lowest error functions. The backbone conformation or helical placement of a single structure may be fortuitous, but the general characteristics of a set of models are more reliable. By simply averaging the positions of the purine pseudoatoms of the superimposed ext structures without regard for the quality of the conformation, it is possible to generate a display of the ensemble helices (Figure 8). When the length of the helix is calculated as the distance between the 5′ bases, the amino acid, DHU, anticodon, and TΨC stems of the crystal structure have lengths of 18.2, 11.9, 13.8, and 11.4 Å, respectively. In the ensemble average helices these lengths are 18.2, 8.56, 12.3, and 12.4 Å. The positions and sizes of amino acid, anticodon, and TΨC helical stems are quite reasonable. As in the individual models, the orientation and length of the DHU stem remain less satisfactory.

Discussion

It is an unfortunate consequence of any extreme reduction scheme that the space-filling nature of the model cannot be maintained simply by increasing the size of the modeling constructs. The pseudohelical stems used represent a compromise between a more explicit representation that would require many more distance constraints and a minimal rep-

resentation that has vdW problems and requires careful handling. Simple replacements for the residues in protein α helices are possible since the helices are single-stranded and the mass of the backbone is distributed along the center of the helix axis. B-form DNA is harder to model because the helices are double-stranded. But as the center of a Watson–Crick base pair in B-form helices is close to the helical axis and the plane of the base pair is perpendicular to the axis, it might be possible to devise simple replacements that are also space-filling. In contrast, the base-pair plane of the A-form RNA is substantially tilted with respect to the helical axis. Additionally, the helix axis is displaced almost 5 Å into the major groove of a base pair, producing a helix in which the mass is distributed on the surface of a hollow cylinder about the axis (Saenger, 1984). Therefore, a scheme that attempts to replace a helix with simple spheres centered on the helical axis cannot accurately represent the mass distribution or the bonding characteristics to the linking single strands. Replacing a double-stranded segment of an RNA with a single computer construct is also incompatible with the primary structure as the construct must be linked to both predecessors (5′ nucleotides) and successors (3′ nucleotides) at each end.

The very first models were constructed with a uniform five replacement pseudoatoms per nucleotide. While the basic bent shape and proper DHU-loop and TΨC-loop stacking of tRNA were always present, these models had problems with the coiling of the helices and the DHU- and TΨC-loop stacking orientations. These "twisting" problems result from the lack of chirality in the pseudoatoms and the low number of secondary and tertiary constraints in comparison to the total number of pseudoatoms. Attempts to improve the distance geometry generated structures of tRNA by changing the number or range of the hydrogen-bonding constraints did not significantly improve the model structures. Even adding distance constraints to the neighboring residues of a hydrogen-bonding partner produced only a gentle winding of the strands. It appears that forcing the double-stranded regions of these less schematic molecules into a helical conformation with distance constraints would require the specification of interatomic distances from each pseudoatom to at least several of the atoms on the opposite strand. Although this would be possible for transfer RNA, it would be impractical for larger RNAs and would undercut the simplicity of the protocol.

Several unsuccessful refinement schemes were attempted that treated the helical regions in differing orders or groupings before the single-stranded regions were refined. It was during this exploratory phase of the research that limited refinement was introduced as a time-saving measure. Initially, all models were refined until the bounds error function had reached a practical convergence limit. The resultant structures had the L shape of the tRNA crystal structure and had dimensions similar to the crystal structure of tRNA. But when they were visually inspected, it was apparent that there were serious van der Waals violations. Simulated annealing in which the structures were repeatedly "heated up" by the addition of small random distances to the bonds and then refined again was also tried. This procedure did not significantly improve either the compaction or the correctness of the structures, and in fact, the fit to the crystal structure was worsened in most cases. Simulated annealing is probably inappropriate for such schematic modeling constructs while the vdW problems that result from complete refinement are a consequence of the drastically reduced pseudoresidues.

In one of the exploratory runs, a distance geometry bounds matrix and transfer RNA structure were accidentally created with only primary structure constraints. The result was an extended, single-stranded helix. As was seen in the modeling of peptides, refining until convergence is achieved produces structures that exactly mimic the nature of the reference structures (Metzler et al., 1989). Since the reduced residues are derived from the Arnott parameters for A-form helices, even single-stranded residues will be forced into a helix given sufficient refinement. The positive aspect of this result lies in the lack of folding produced by this mistake. In the absence of explicit distance constraints, distance geometry foldings favor extended conformations. Therefore, any compact conformation that is consistently produced must be necessary to satisfy the long-range constraints. It is well known that, in low ionic solution, nucleic acids form extended single-stranded tangles due to the strong electronic repulsion of the negatively charged phosphate backbone. It has also been demonstrated that divalent cations or polycationic proteins are required for the proper folding of some nucleic acids (Schimmel & Redfield, 1980). Perhaps by adding in the secondary and tertiary distance constraints in stages, we can study the general folding process in those cases where the tertiary relationships can be ordered. When just the primary and secondary structure constraints are used as the basis of folding the molecule, the resulting model strongly resembles the classical cloverleaf.

The distance geometry algorithm that we used has been criticized for the manner in which it samples conformational space (Metzler et al., 1989). The random method the program uses in selecting a distance between the upper and lower bounds favors longer distances that are not characteristic of compact biological structures. In the context of our modeling protocol, this is a desirable characteristic as the subsequent steps required to restore the full structure are more likely to succeed in an open structure that minimizes the possibility of vdW overlaps. The other major criticism is that the distance geometry algorithm does not widely sample conformational space. The clustering of the more than 600 tRNA foldings close to the X-ray crystal structure and the large number of repeated structures seem to support this idea. Perhaps, as suggested by Crippen (1987), the algorithm is favoring solutions to the folding problem that are in the neighborhood of the global minimum energy conformation. On the other hand, the inability of the algorithm to distinguish between improbable tangled or knotted structures and more regular conformations is disappointing. This means that while human judgments have been removed from folding the molecule, they retain their vital role in selecting and interpreting the results.

This study demonstrates that the major determinants of the size and shape of tRNA are inherent to its basic primary and secondary structure. The types of data used are similar to that available for other RNAs, but there is no reason that data from other sources (e.g., NMR, fluorescence energy transfer, neutron diffraction) should be excluded if they are available and can be properly quantified. It will therefore be possible to use the same approach to produce low-resolution models of other RNA molecules. With the addition of only a few tertiary interactions it is possible to begin to explore the formation of three-dimensional RNA structures.

References

Arnott, S., Hukins, D. W. L., Dover, S. D., Fuller, W., & Hudgson, A. R. (1973) *J. Mol. Biol. 81*, 107–122.

Brünger, A. T., Clore, M. C., Gronenborn, A. M., & Karplus, M. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 3801–3805.

Crippen, G. M. (1987) *J. Phys. Chem. 91*, 6341–6343.

Dock-Bregeon, A. C., Dhevrire, B., Podjarny, A., Johnson, J., de Bear, J. S., Gough, G. R., Gilham, P. T., & Moras, D.

(1989) J. Mol. Biol. 209, 459-474.
Garrett-Wheeler, E., Lockard, R. E., & Kumar, A. (1984) Nucleic Acids Res. 12, 3405-3423.
Gutell, R. R. & Woese, C. R. (1990) Proc. Natl. Acad. Sci. U.S.A. 87, 663-667.
Happ, C. S., Happ, E., Nilges, M., Gronenborn, A. M., & Clore, G. M. (1988) Biochemistry 27, 1735-1743.
Haselman, T., Camp, D. G., & Fox, G. E. (1989) Nucleic Acids Res. 17, 2215-2221.
Jaeger, J. A., Turner, D. H., & Zuker, M. (1989) Proc. Natl. Acad. Sci. U.S.A. 86, 7706-7710.
Kim, S. H., Quigley, G. J., Suddath, F. L., McPherson, A., Sneden, D., Kim, J. J., Weinzierl, J., & Rich, J. (1973) Science 179, 285-288.
Levitt, M. (1969) Nature (London) 224, 759-763.
Metzler, W. J., Hare, D. R., & Pardi, A. (1989) Biochemistry 28, 7045-7052.
Moras, D., Comarmond, M. B., Fischer, J., Weiss, R., Thierry, J. C., Ebel, J. P., & Giegé, R. (1980) Nature (London) 288, 669-674.
Ninio, J., Favre, A., & Yaniv, M. (1969) Nature (London) 223, 1333-1335.
Saenger, W. (1984) in Principles of Nucleic Acid Structure, Springer-Verlag, New York.
Sampson, J. R., DiRenzo, A. B., Behlen, L. S. & Uhlenbeck, O. C. (1989) Science 243, 1363-1366.
Schevitz, R. W., Podjarny, A. D., Krishnamachari, N., Hughes, J. J., Sigler, P. B., & Sussman, J. L. (1979) Nature (London) 278, 188-190.
Shimmel, P. R., & Redfield, A. G. (1980) Annu. Rev. Biophys. Bioeng. 9, 181-221.
Srinivasan, A. R., & Olson, W. K. (1987) J. Biomol. Struct. Dyn. 4, 895-938.
Sussman, J. L., Holbrook, S. R., Warrant, R. W., Church, G. M., & Kim, S.-H. (1978) J. Mol. Biol. 123, 607-630.
Tinoco, I., Jr., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., & Gralla, J. (1973) Nature (London), New Biol. 246, 40-41.
Woo, N. H., Roe, B. A., & Rich, A. (1980) Nature (London) 286, 346-351.
Wright, H. T., Manor, P. C., Beurling, K., Karpel, R. L., & Fresco, J. (1979) in Transfer RNA: Structure, Properties, and Recognition, Cold Spring Harbor Monograph Series 9A, pp 145-160, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Yaniv, M., Favre, A., & Barrell, B. G. (1969) Nature (London) 223, 1331-1333.
Zachau, H. G., Dutting, D., Feldman, H., Melchers, F., & Karau, W. (1966) Cold Spring Harbor Symp. Quant. Biol. 31, 417-424.

Molecular Dynamics Investigation of the Interaction between DNA and Distamycin†

K. Boehncke, M. Nonella, and K. Schulten*
Beckman Institute and Department of Physics, University of Illinois, 405 North Mathews Avenue, Urbana, Illinois 61801
A. H.-J. Wang
Beckman Institute and Department of Physiology and Biophysics, University of Illinois, 524 Burrill Hall, Urbana, Illinois 61801
Received August 20, 1990; Revised Manuscript Received January 4, 1991

ABSTRACT: The complex of the minor groove binding drug distamycin and the B-DNA oligomer d-(CGCAAATTTGCG) was investigated by molecular dynamics simulations. For this purpose, accurate atomic partial charges of distamycin were determined by extended quantum chemical calculations. The complex was simulated without water but with hydrated counterions. The oligomer without the drug was simulated in the same fashion and also with 1713 water molecules and sodium counterions. The simulations revealed that the binding of distamycin in the minor groove induces a stiffening of the DNA helix. The drug also prevents a transition from B-DNA to A-DNA that was found to occur rapidly (30 ps) in the segment without bound distamycin in a water-free environment but not in simulations including water. In other simulations, we investigated the relaxation processes after distamycin was moved from its preferred binding site, either radially or along the minor groove. Binding in the major groove was simulated as well and resulted in a bound configuration with the guanidinium end of distamycin close to two phosphate groups. We suggest that, in an aqueous environment, tight hydration shells covering the DNA backbone prevent such an arrangement and thus lead to distamycin's propensity for minor groove binding.

Control of gene expression at the level of transcription is emerging as a central area of investigation in biology (Stryer, 1988). The availability of respective DNA sequences and of structures of regulatory proteins involved in this control will allow one to rationalize underlying mechanisms. For example, it is now well understood how a number of repressor proteins interact with their respective operator DNA sequences at the molecular level (Ptashne, 1986). However, the structural complexity and large variability of regulatory protein-DNA interactions make it likely that structures of regulatory protein-DNA complexes will not be readily available. Therefore, molecular modeling and molecular dynamics may play an important role in combining information from various ex-

†M.N. gratefully acknowledges financial support by the Kanton Zürich, Switzerland. This work was carried out in the Center for Parallel Computation in Molecular Dynamics funded by the National Institute of Health. Computation time was granted by the National Center for Supercomputing Applications at the University of Illinois in Urbana-Champaign supported by the National Science Foundation.
*To whom correspondence should be sent.

0006-2960/91/0430-5465$02.50/0