15

The RNA Folding Problem

Peter B. Moore
Departments of Chemistry, and Molecular Biophysics, and Biochemistry
Yale University
New Haven, Connecticut 06520-8107

Because the chemical and biological properties of many RNAs are determined by their conformations, an RNA equivalent exists to the protein folding problem. The RNA folding problem has both a practical side and an academic side. If the practical problem were solved, it would be possible to predict the conformations of RNAs of known sequence. The academic problem has to do with determining what happens when denatured RNAs fold in vitro, and nascent RNA chains fold in vivo. The focus of this essay is structure prediction, but it does include some commentary on what has been learned about RNA folding pathways.

Structure prediction is a matter of some urgency today because RNAs are being discovered much faster than their three-dimensional structures are being solved. Furthermore, the techniques available for determining structures experimentally, X-ray crystallography and NMR, are so time-consuming, applied to RNAs by communities so small, and so uncertain of success when applied to specific RNAs that, absent some technological and/or sociological revolution, the mismatch between supply and demand is certain to persist indefinitely.

In principle, biochemists ought to be able to do without the services of crystallographers and NMR spectroscopists. Most (all?) denatured RNAs renature spontaneously in vitro, which means that the structures of RNAs are determined by their sequences. Furthermore, the physical interactions that drive RNA folding are largely understood (see Chapter 10). Why can't the structures of RNAs of known sequence be predicted, and failing that, why can't procedures be developed that determine conformations by combining theory with nonphysical data?

RNA ARCHITECTURE

Before launching into a discussion of RNA structure prediction, it would be wise to remind ourselves exactly what it is that needs to be predicted.

The RNA World, Second Edition © 1999 Cold Spring Harbor Laboratory Press 0-87969-561-7/99

Double Helices

Like the rest of modern biology, the science of RNA conformation begins with the structure of DNA, and, as everyone knows, the typical DNA molecule consists of two polynucleotide strands arranged in an antiparallel double helix, in which As, Gs, Cs, and Ts on one strand pair with Ts, Cs, Gs, and As, respectively, on the other. Except for the (almost) trivial difference that in RNA, As pair with Us, RNA forms antiparallel double helices the same way. Under all conditions, the conformation of RNA helices resembles that adopted by DNA helices under low-humidity conditions; i.e., RNA helices are A-form, never B-form. In A-form helices, the planes of base pairs are tilted with respect to the helix axis, the major groove is narrow and deep, and the minor groove is broad and shallow (see Saenger 1984). Furthermore, the sugar pucker in A-form helices is C3′-endo, or N(orth), whereas in B-form helices, it is C2′-endo, or S(outh). The reason all RNA helices are A-form is that B-form helix formation is sterically hindered by the placement of their 2′-OH groups when their riboses are C2′-endo. Note, however, that A-form is not the only helical geometry possible for RNA. Three- and even four-strand helices exist, but they are possible only for specific sequences, and hence are seldom encountered.

Single-stranded Polynucleotides: Stem-loops

Most RNAs are single-stranded molecules that have sequences incompatible with the formation of long double helices, but, nevertheless, many are more than 60% as hyperchromic as they would be if they were perfect duplexes. In the late 1950s, Doty and coworkers explained why (Doty et al. 1959; Fresco and Alberts 1960; Fresco et al. 1960). Except for molecules with bizarre sequences, like polyuridylic acid or polycytidylic acid, RNAs are full of short sequences that are “accidentally” complementary, and RNA chains fold back on themselves to form hairpin loops or stem-loops, so that these sequences can form helices.

Many of the helical stems in RNA hairpins are interrupted by internal loops, which is to say regions where sequences are juxtaposed that cannot form GCs, AUs, or wobble GUs, which most regard as “honorary” Watson-Crick base pairs. Some form irregularly helical structures that consist of a succession of non-Watson-Crick base pairs (see, e.g., Correll et al. 1997; Dallas and Moore 1997). Others, especially those in which the number of bases contributed by the two strands is different, form more irregular structures (Puglisi et al. 1992, 1995; Battiste et al. 1994, 1996; Aboul-ela et al. 1995; Ye et al. 1995; Fan et al. 1996; Fourmy et al. 1996; Yang et al. 1996; Jiang et al. 1997; Kalurachchi et al. 1997). In some cases, the trajectory of the backbone becomes so distorted that bases project into solution, away from the body of the stem (bulged bases). In others, "extra" bases form base triples by hydrogen bonding edge-on with other base pairs, or intercalate between base pairs.

Necessarily, all stem-loops include a terminal loop where the trajectory of the RNA backbone changes direction by 180° so that the 3′ side of the sequence can pair with the 5′ side. Terminal loops vary in length from 2 nucleotides (Butcher et al. 1997), to hundreds or even thousands of bases. Some 4-base loops, or tetraloops, are unusually common and play the same role in RNA as β turns in proteins (Tuerk et al. 1988; Cheong et al. 1990; Woese et al. 1990; Allain and Varani 1995; Jucker and Pardi 1995; Jucker et al. 1996). Loops bigger than 20 or 30 nucleotides may contain internal stem-loops.

Ever since the hairpin model for RNAs was advanced, identification of the helical segments in RNAs has been a major activity, and Figure 1 shows what has been learned about the stem-loops in *Escherichia coli* 16S rRNA (Gutell 1996). The RNA literature is full of similar diagrams. Note that sequences for which conformational information is lacking for any reason are always shown as single-stranded in these diagrams.

Describing RNA Architecture

By tradition, RNA chemists describe the architecture of RNAs using terminology devised to describe proteins. It isn't very appropriate. The use of the term "secondary structure" to describe the helices in Figure 1 is a case in point. The backbone of any polymer is helical wherever there are runs of residues that have the same backbone torsion angles, and that helicity is secondary structure. Both strands of a nucleic acid double helix have secondary structure because taken in isolation, both are helical. Their antiparallel, side-by-side association, on the other hand, is not secondary structure. It is tertiary structure if two strands are part of the same sequence, and quaternary structure if they are not, because in both cases, sequences are being juxtaposed that are distantly associated, if associated at all. Thus, Figure 1 is better described as a helix diagram or a stem-loop diagram than as a "secondary structure diagram," which is the terminology commonly used. Note that the same confusion arises when the β-sheets in proteins are discussed.

The potential of most RNAs for intramolecular interaction is not exhausted by its helices. Additional interactions involving helices, the nonhelical segments, terminal loops, metal ions, water molecules, etc., can bring stem-loops together, and force RNAs to adopt globular conformations. The arrangements of helical segments that result are what most RNA chemists mean by "tertiary structure," but it is better described as suprahelical structure.

The last protein-related nomenclature issue that deserves comment is the application of the term "domain" to RNA structures. A domain is a region of a protein's structure, the conformation of which is stabilized en-

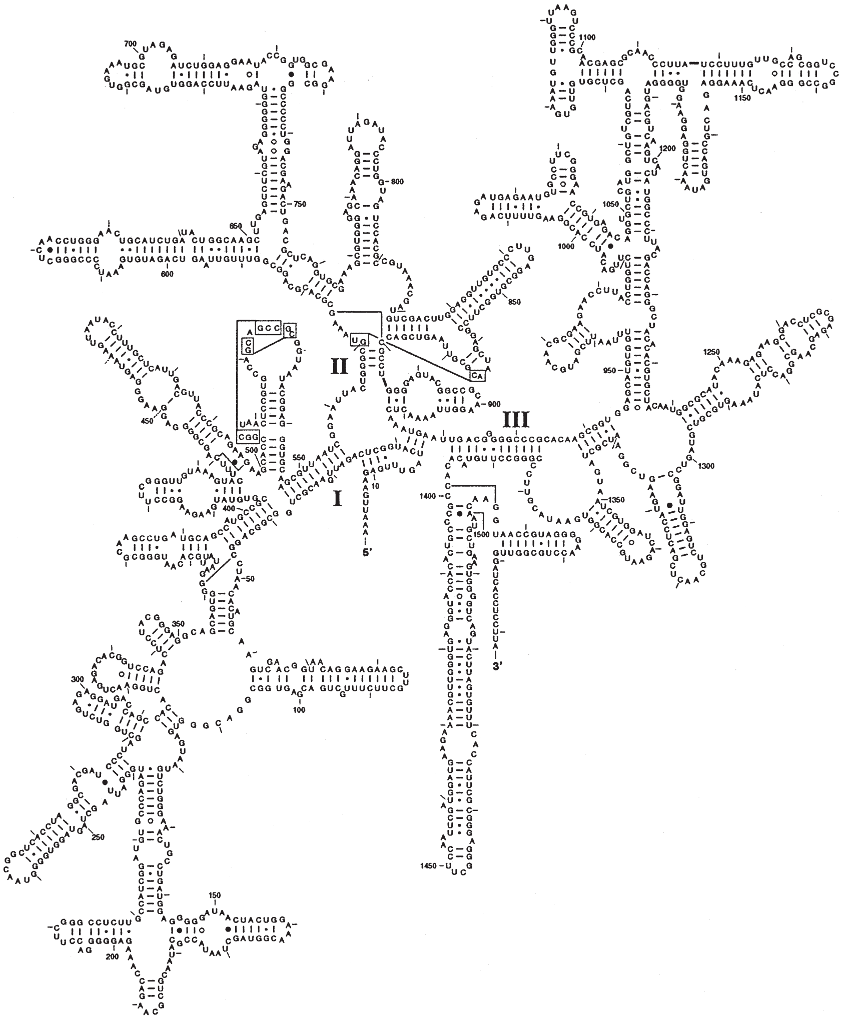

Figure 1 The helix diagram of 16S rRNA from *Escherichia coli.* This diagram was provided by Dr. Robin Gutell. (Reprinted, with permission, from Gutell 1996 [copyright CRC Press].)

tirely by internal interactions; it is a structurally autonomous unit. Domains are interesting because they retain their conformations and (often) their biological properties when they are excised from the structures to which they belong. In RNAs, there two kinds of domains. The simplest is a stem-loop that has a small terminal loop, i.e., a local hairpin loop. Many of these structures (most of them?) retain their conformations in isolation. The second type of domain is an assembly of stem-loops capable of maintaining its proper suprahelical organization in isolation, e.g., the P4-P6 domain from the *Tetrahymena* self-splicing intron (Cate et al. 1996). There are strong reasons for believing that ribosomal RNAs, for example, also contain suprahelical domains (Weitzmann et al. 1993; Samaha et al. 1994).

PREDICTING STRUCTURES FROM SEQUENCES

The essence of the conformation prediction problem was identified 30 years ago by Levinthal, who pointed out that the number of conformations possible for a random coil protein is so huge that if renaturation depended on a denatured protein finding its lowest energy conformation by a thermally driven, random walk through conformation space, renaturation would take an eon of eons (Levinthal 1968). Levinthal's argument is valid for RNA, and it follows that RNA structures cannot be predicted using techniques that rely on exhaustive searches to find lowest free-energy conformations. Thus, there are only two general strategies available that could solve the RNA folding problem. Either approaches must be found that reduce the number of conformations to be compared by astronomical factors, or methods must be developed for simulating what actually happens when RNAs refold in vitro. Both strategies have been pursued by protein chemists, and an immense literature now exists that those interested in RNA structure prediction ignore at their peril.

Ab Initio Computation of RNA Conformation

Over the past 20 years, a number of computer programs have been developed for simulating the thermal motions of biological macromolecules (e.g., CHARMM [Brooks et al. 1983], and AMBER [Weiner and Kollman 1981]). At the outset of the typical dynamics simulation, a structure is assumed for the molecule of interest, and both the energy of the molecule and the net force experienced by each of its atoms are computed. Harmonic potentials are (usually) used to represent the energetic cost of distorting bond lengths and angles away from their crystallographically determined, normal values, and classic approximations are used to represent nonbonded interactions. The trajectory of each atom is then computed for about a femtosecond, assuming force equals mass times acceleration. At that point, both the energy of the system and the forces on its atoms are re-evaluated, so that trajectories can be corrected for changes that have taken place, and another femtosecond's worth of trajectory is computed, and so on, for as many iterations as desired. The reason trajectories have to be updated so often is that molecules vibrate and collide in solution at inverse picosecond rates, and unless trajectories are updated much more rapidly than that, these motions will not be represented faithfully. Since the computer equivalent of temperature can be controlled during molecular dynamics simulations, one could imagine using these programs to simulate the thermal denaturation and renaturation of a single RNA molecule, and if that can be done, why can't RNA conformations be computed ab initio?

One reason RNA conformations cannot be computed this way is time. When RNAs renature, their stem-loops form in microseconds, but it takes seconds to minutes for suprahelical structures to develop (see below). Unfortunately, it strains the capacity of even the most powerful computers to simulate a single microsecond of the history of a modest sized macromolecule. Efforts are being made to lengthen the time interval simulated by each dynamics iteration (see Schlick et al. 1997), but still, until computer time gets many orders of magnitude cheaper than it is today, this avenue is closed.

Another reason RNA conformations cannot be computed today is accuracy. Contrary to what the chemically naive believe, the total nonbonded energy of a molecular system cannot be represented as a sum of independent atom i/atom j interaction energies. The reason is that in general, both the magnitude and direction of the force atom i exerts on atom j depends on the locations and identities of the atoms that surround them (Maitland et al. 1981). Nevertheless, because the computational cost of taking neighboring group effects into account would be prohibitive, all macromolecular dynamics programs assume pairwise additivity. At best, the pairwise potentials used are potentials of mean force, which is to say approximations that correct for the average effect of neighboring groups in solution. At worst, they are potentials that are approximately correct for groups interacting in the gas phase. It is not clear that conformations can be predicted accurately at this level of approximation.

Finally, it must be acknowledged that the RNA folding problem is intrinsically less tractable than its protein counterpart. The typical protein domain is stabilized primarily by interactions occurring in a hydrophobic core that is composed of densely packed alkyl and aromatic side chains, and polar groups, virtually all of which are involved in hydrogen bonding or other neutralizing interactions. Since the influence of solvent on intramolecular interactions in proteins can be dealt with by treating it as a continuous, dielectric medium (Sharp and Honig 1990), it is conceivable that computations based solely on the covalent structures of proteins could lead to reliable predictions of conformation.

RNAs are more complicated. The suprahelical structure of the P4-P6 domain, for example, depends on bound Mg^{++} ions, as does the conformation of an internal loop in prokaryotic 5S rRNAs, and hydrogen bonds involving water are commonplace in RNA (Cate et al. 1996; Correll et al. 1997). Since no one I know of can predict the numbers and locations of water molecules and ions tightly bound to an RNA on the basis of its sequence, ab initio computations of RNA structure must model the surrounding solvent in molecular detail, and that increases the cost of computations considerably. The only solace is that solvent effects are better evaluated this way than they are using continuum dielectric models.

RNAs are also more complicated because they have a very high charge density (one anionic charge per residue at neutral pH). As has been long understood, ionic interactions play a big role in determining RNA conformation, and unless they are dealt with properly, RNA simulations are guaranteed to fail. Ionic interactions are hard to cope with computationally because their range is much longer than that of any other kind of nonbonded interaction, but a solution to this problem may have just been found. Recent investigations suggest that the computational problems posed by Coulombic interactions can be solved using Ewald particle mesh algorithms (see Louise-May et al. 1995).

It does not follow from the above that today’s molecular dynamics programs should never be applied to RNA. If the volume of conformation space to be explored is restricted, they can be very useful. For example, experimentalists routinely use molecular dynamics programs to compute and refine RNA structures (see, e.g., Brünger 1992). The X-ray and NMR data included in their computations restrict the range of conformations available and reduce the impact that force-field inaccuracies might have. Under these circumstances, current molecular dynamics codes routinely find lowest-energy molecular conformations. Molecular dynamics computations are also providing insight into the dynamics of RNAs of known conformation, RNA ligand-binding properties, etc. (see Louise-May et al. 1996).

For completeness, I note that most issues that can be addressed by molecular dynamics methods can also be approached using Metropolis

Monte Carlo algorithms. However, it is not obvious that they are any less limited by force-field inaccuracies or the large number of states available to be explored.

Semi-empirical Approaches to the Determination of Stem-loop Diagrams

There is a lot of conformational information in a diagram like Figure 1. Suppose a certain RNA had 100,000,000 possible conformations in the Levinthal sense, and suppose its stem-loops could be identified. Since the backbone torsion angles of nucleotides in helices are fixed, if half of its nucleotides were involved in helices, the number of conformations that remained possible for the molecule would be reduced to 10,000, and if three-quarters of its bases were tied up in helices, the number of possible conformations would be 100. Obviously, it would be a good thing if helix diagrams could be deduced from sequences.

In the 1970s, it was discovered that the thermal stabilities of RNA helices can be predicted using thermodynamic data obtained from melting studies done on small, synthetic RNA oligonucleotides (Tinoco et al. 1973). The reason this works is that the stabilizing effect of a base pair is determined primarily by its identity and the identities of the bases that flank it on either side. Thus, if Watson-Crick and wobble G•U pairs are the only ones considered, ΔG°s, ΔH°s, and ΔS°s for just 21 combinations of base pairs and flanking sequences should suffice to make predictions (Serra and Turner 1995). In addition, the entropic cost of forming duplexes from separate strands has been worked out, as have thermodynamic parameters for terminal loop formation, internal loop formation, etc. Thus, the information required to predict the energetics of simple RNA hairpins exists, and their melting temperatures can be predicted with reasonable accuracy (see Chapter 10).

A number of programs have been devised that are capable of identifying all combinations of stem-loops possible for a sequence, and then estimating their free energies from thermodynamic data. MFOLD, which is a well-known example, produces the lowest-energy helix diagrams for a sequence, rank-ordered by free energy (Zuker 1989). It works well for small RNAs like tRNAs and 5S rRNAs, but its performance degrades as the size of the RNAs increases (Zuker and Jacobson 1995). The larger an RNA, the more likely it is to include structural elements, the thermodynamic properties of which are unknown. Although it is clear that the accuracy of predictions improves in parallel with the thermodynamic database, there is a practical limit to what can be done. The labor required to fill the "catalog" of thermodynamic data grows geometrically with the number of substructures it includes. In addition, it is not clear that a complete catalog would do the job. Non-nearest neighbor effects do affect conformational stability (Turner et al. 1988). This is the reason programs like SAPSSARN are interesting (Gaspin and Westhof 1995). SAPSSARN is an interactive program for constructing helix diagrams that, in addition to taking base-pairing energies into account, can be guided by nonthermodynamic constraints.

It should be noted that secondary-structure determination and prediction have long been a part of the protein field. Just as the amount of helix in an RNA can be estimated from experiments that measure hyperchromicity, the amount of helix and β-sheet in proteins can be estimated from CD/ORD spectra. In addition, in the 1970s, as the number of experimentally determined protein structures increased, it became evident that the amino acid compositions of β-sheets, on average, are not the same as those of α-helices. From this insight, it was but a short step to rules for identifying regions in protein sequences that are likely to be helix or β-sheet. The best known are the ones formulated by Chou and Fasman almost 25 years ago (Chou and Fasman 1974). Secondary-structure prediction remains a significant activity in the protein field, and the information sought is equivalent to that contained in an RNA's helix diagram.

BIOCHEMICAL METHODS FOR DETERMINING RNA CONFORMATION

Given how poorly RNA biochemists and molecular biologists have been served by their biophysical colleagues over the years, it is little wonder that many of them have devoted immense effort to the characterization of RNA structures by nonphysical means. The only protein biochemists making similar investments in the nonphysical analysis of conformation today are those interested in membrane proteins, and they also have reason to complain about the small number of crystal structures and NMR structures available to them.

The Phylogenetic Approach

In 1975, Fox and Woese proposed the three-stem helix diagram for 5S rRNA now known to be correct (Fox and Woese 1975). The reason their paper is remembered is not that their model was novel—others had already proposed similar models (Madison 1968; DuBuy and Weissman 1971; Nishikawa and Takemura 1974)—but the nature of the argument made to support it. Since 5S rRNAs from different species perform the same function, they must have similar conformations. Therefore, the correct stem-loop structure for 5S rRNAs is that which is compatible with the sequences of all of them. The same argument had been made in support of the familiar tRNA cruciform almost a decade earlier, but there was a difference. The cruciform was identified initially because it is one of three maximally base-paired structures that could be proposed for the first tRNA sequence (Holley et al. 1965). As soon as a few more tRNAs were sequenced, it was obvious that the cruciform was right. From the point of view of Fox and Woese, the number of base pairs in an RNA helix diagram is irrelevant, and when they began, it was not obvious that a unique, maximally paired structure exists for 5S rRNA. Six years later, the phylogenetic approach was applied with spectacular results to the sequence of 16S rRNA (Noller and Woese 1981).

A phylogenetic analysis has three steps, the first of which is the accumulation of the sequences of as many members of a family of homologous RNAs as possible. The second step is sequence alignment, which is important because deletions and insertions are often encountered. Alignment usually depends on the identification of sequences that are conserved within families. The third step is analysis of sequence covariations. For example, if residues x and y form a Watson-Crick pair in an RNA, what you expect is that y will be a G whenever x is a C, and y will be a U whenever x is an A, etc. Not only can regular helices be identified this way, but if the number of sequences available is large enough, noncanonical pairings can be worked out, and even some of the long-range interactions on which suprahelical structure depends (Gutell and Woese 1990). The idea that an RNA's conformation will be preserved if a base change at one position is matched by an appropriate base change made at another is the basis for the compensating mutation method for identifying RNA–RNA interactions experimentally. Indeed, one could take the view that the sequences the phylogenetic analyst compares are the products of innumerable compensating mutation experiments carried out by Nature over the course of evolution.

Biochemical Methods for Probing RNA Conformation

For over 30 years, RNA chemists have been doing chemical and enzymatic experiments to validate helix diagrams, to determine the conformation of "unstructured" sequences in helix diagrams, and to work out suprahelical structures. A remarkable variety of strategies have been developed, and examples of the application of almost every one of them can be found in the 5S rRNA literature (Moore 1995).

In the 1960s and 1970s, most of these investigations exploited the fact that the susceptibility of RNA sequences to enzymatic digestion depends on their conformations. Many sequence-specific nucleases cleave single-stranded RNA more rapidly than double-helical RNA, and less specific nucleases exist that cleave either single-stranded RNA or double-helical RNA, but not both. Nevertheless, enzymatic approaches to RNA conformation gradually fell out of favor. Enzymes are so big compared to the structural details they were being used to examine that one had to be concerned that reactivities might depend on more than just local conformation. In addition, no one could be sure how any of these enzymes would respond to RNA structures other than A-form helix. For this reason, the RNA community turned increasingly to low-molecular-weight, group-specific reagents as conformational probes.

The number of group-specific reagents that have been used to probe RNA conformation is very large (for descriptions, see Noller and Moldave 1988). The chemical probes popular today share the property that products of their reactions with RNA nucleotides do not function as templates for reverse transcriptases. Thus, the nucleotides that have reacted in a lightly modified RNA can easily be identified by primer extension. The variations in reactivity observed are attributed to differences in the degree of reagent access allowed by the conformation of the parent molecule. This type of reactivity mapping has been used to identify Watson-Crick helices, to distinguish Watson-Crick-paired bases from those paired in other ways, and has even provided the basis for detailed models of entire RNAs (see, e.g., Brunel et al. 1991).

Any protein chemist who interpreted reactivity data this way would be a laughing stock. The reactivity of protein groups can be modulated by interactions with their neighbors in ways that have nothing to do with accessibility. Indeed, the unusual reactivity of many enzymes is explained by effects of just that kind. Nevertheless, when RNAs of known conformation are probed chemically, the reactivity patterns observed are those anticipated if conformation were in fact the dominant determinant. It must be the case that RNAs seldom adopt conformations that result in significant, nongeometric alterations in reactivities, and this may be one of the reasons that protein enzymes won out over ribozymes during the course of evolution.

The most important generality to emerge from this body of work is that isolated RNA molecules usually have stem-loop structures close to those predicted phylogenetically. This is not trivial, because most RNAs perform their functions in association with proteins, and as Fox and Woese pointed out, it is their conformations in those contexts that count, not their conformations as free molecules. Association with protein must not alter the stem-loop organization of RNAs very much.

Two types of chemical probes are used to obtain information about the suprahelical organization of RNAs: cross-linking reagents and short-range labeling reagents. Cross-linking agents covalently link nucleotides in RNAs that are neighbors in space, but not necessarily neighbors in sequence (see, e.g., Harris et al. 1994), and here too the number of reagents available is large. The short-range labeling reagents favored today all contain Fe(II) complexed with EDTA. When ascorbic acid, H_2O_2, and Fe(II)-EDTA are combined, short-lived hydroxyl radicals are generated that react with almost anything. When ribose rings are attacked, chain scission often results, and that blocks reverse transcriptase (Pogozelski et al. 1995). Thus, if an Fe(II)-EDTA group is bound to a specific location in an RNA, and ascorbate and H_2O_2 are added, the RNA residues that react will be those closer to the Fe(II)-EDTA group than the distance a hydroxyl radical can diffuse before it decays (~20 Å) (see, e.g., Joseph and Noller 1996; Joseph et al. 1997).

Model-building

It is difficult to understand the conformational implications of biochemical information of the type we have been discussing without computer assistance. Programs that use the distance-geometry/simulated annealing algorithms, which NMR spectroscopists favor for their structure determinations, lend themselves to this kind of data interpretation, and YAMMP, which was developed by Harvey and his colleagues, is an example of a program of that kind that can use chemical information to construct three-dimensional RNA models (Tan and Harvey 1993).

The model-building program, MC-SYM, takes a different approach (Major et al. 1991). The conformation of any nucleotide in an RNA can be specified by defining the rotational orientations, or torsion angles, associated with 6 backbone single bonds and its glycosidic bond. Although in principle 360° rotation is possible around single bonds, in all molecules certain rotational orientations are always favored, and often heavily so: the ones in which the groups joined by the bond in question interfere the least sterically. The values for any torsion angle that have low steric energy are its “rotamers.” The maximum number of rotamers available to any one of the seven torsion angles that determine the conformation of a nucleotide is three, and the number of rotamer combinations possible for a nucleotide is about 400. Nevertheless, only 20–30 of these combinations have ever been observed in nucleic acid crystal structures (Gautheret et al. 1993).

MC-SYM contains a library of these preferred nucleotide conformations, which it uses to build RNA models sequentially. For example, if a three-nucleotide RNA is being modeled, and there are 20 conformations possible for each nucleotide, a priori, the universe of possible models has 8000 members. MC-SYM never considers that entire universe. Instead, each of the 400 two-nucleotide models produced when the second nucleotide is added to the first is tested for its compatibility with everything that is known about the conformation of the molecule being modeled. The information taken into account includes the molecule's stem-loop diagram and might also include chemical cross-linking and protection data and even distances deduced from NMR experiments. The only three-nucleotide models it produces are those that can be built by adding a nucleotide to one of the acceptable two-nucleotide structures, and they too are screened for compatibility. The pruning of incompatible conformations that occurs at each step in the construction of an RNA model in MC-SYM usually reduces the number of structures that emerge by orders of magnitude. Only the survivors need be analyzed further.

What Is a Structure?

Over the years, many RNAs have been modeled in many different ways, including tRNA (Levitt 1969), 5S rRNA (Brunel et al. 1991), M1 RNA (Harris et al. 1994; Westhof et al. 1996), the self-splicing intron from *Tetrahymena* (Michel and Westhof 1990), and 16S rRNA (Malhotra and Harvey 1994; Mueller and Brimacombe 1997; Stern et al. 1988), to name a few. The richness of the data that underlie these models, and the methods used to arrive at them, vary enormously, of course. However, if the amount of biochemical data taken into account in one of these exercises were large enough, and if the method used to analyze them was rigorous enough, there is no reason why a model might not emerge that is the truth, for all intents and purposes. A biochemical model that good would be equivalent to an X-ray or NMR structure. The problem nonspecialists confront is that whether a nonphysical RNA model is that good or not, it tends to look that good when it is displayed using an all-atom representation. How is one to know the difference? Are any of the biochemical models published to date that good?

One way to answer this question is to ask what the properties of physically determined structures are that scientists prize. I submit that there are just two. First, physical structures are reproducible, which means to say that if two groups independently determine the conformation of the same molecule by physical means, the structures they produce will be the same, to within experimental error. Second, physical structures have high explanatory power. Not only do they rationalize the data used to obtain them, they also rationalize large bodies of data that were not used. Typically they explain most of the physical, chemical, and biological properties of the molecules they represent.

A biochemical model that is reproducible and has high explanatory power would be a good model, no matter how crude the representation of the molecule it portrays, but in my estimation, none of the biochemical RNA models produced so far is “good.” When different laboratories analyze the same RNA by nonphysical means, the models that emerge are different, especially in their suprahelical organization. For example, compare Altmann and Westhof’s models for M1 RNA with that of Pace and his colleagues (Harris et al. 1994; Westhof et al. 1996) or the 16S rRNA models of Noller, Harvey, and Brimacombe (Stern et al. 1988; Malhotra and Harvey 1994; Mueller and Brimacombe 1997). This is not surprising because in almost every case, the data available are demonstrably insufficient to narrow the number of suprahelical conformations possible to just one. There is no harm in that, provided the reader is given some idea of the range of models the data permit, which has seldom been the case. Happily, methods have been developed recently for evaluating the uncertainty of such models quantitatively, and that is a huge step in the right direction (Malhotra and Harvey 1994).

The biochemical RNA models in the literature faithfully represent the stem-loop structures of their RNAs, insofar as they were known at the time they were built, and at worst, can be considered three-dimensional summaries of the data available. Most of them contain important amounts of truth, and not uncommonly they suggest explanations for some of the properties of the molecules they represent. The problem is distinguishing the wheat from the chaff.

WHAT HAPPENS WHEN RNAs FOLD?

There can be little argument that local stem-loop structures are simpler than suprahelical structure, and implicit in all chemical and semiempirical approaches to RNA structure determination is the notion that an RNA’s stem-loop structure must be determined before its suprahelical structure is addressed; i.e., the simple must be dealt with before the complex. Is that the logic that underlies the folding of real RNA molecules, or is it merely an ordering humans find attractive? This question was first addressed experimentally in the late 1960s and early 1970s using tRNAs as model systems (see Crothers and Cole 1978; Crothers 1979). The field then went into an eclipse from which it is now emerging, thanks to the development of methods for producing RNAs of defined sequence in large quantities, and thanks also to the interest in catalytic RNAs. It is worth taking a look at what has emerged (see Brion and Westhof 1997).

Early Events in RNA Folding

T-jump experiments done in the early 1970s demonstrated that in the presence of physiological concentrations of monovalent ions, the time constants for local stem-loop formation are 10s to 100s of μsec (Crothers et al. 1974). Furthermore, consistent with polymer statistics, the time constants for stem-loop formation are (roughly) proportional to the lengths of their loops raised to the 3/2 power (Crothers et al. 1974). Thus, if under some set of conditions a stem with a seven-base loop takes 80 μsec to form, a stem with a 1000-nucleotide loop will take about 135 msec. These same pioneering studies demonstrated that the suprahelical structure of tRNA is considerably less stable than the structures of its helices, and that the suprahelical structure of most tRNAs is destabilized by the absence of divalent cations, which local stem-loops generally are not.

There are parallels between local stem-loop formation in RNAs and the early events in protein folding (see Friesner and Gunn 1996; Levitt et al. 1996). When small, globular proteins are denatured, and then transferred from denaturing solvents to physiological solvents, they are almost instantaneously transformed from random coils into globular structures that are not much bigger than their final, fully folded conformations. This occurs on about the same time scale as the formation of local stem-loops in RNAs. The condensed state that results is referred to as a "molten globule," and molten globules form for the same reason that lipids form micelles. They are conformations that minimize the contact of hydrophobic groups with water while maximizing the contact of hydrophilic groups with water.

A first-year graduate student might claim that the two processes are fundamentally different; that stem-loop formation is driven by hydrogen bonding, and that molten globule formation is driven by the hydrophobic effect. A second-year graduate student would not be so sure. The most hydrophobic parts of an RNA are its bases, and significant amounts of base surface become protected from solvent when helix forms; the middle of a helix is the hydrophobic core of a nucleic acid. Furthermore, in aqueous environments, the energetic value of intramolecular hydrogen bonds is only about one kcal/mole because water molecules can always substitute for macromolecular hydrogen-bond donors and acceptors. The driving force for base-pairing in nucleic acids is thus the free-energy penalty exacted if a base donor or acceptor group fails to find a mate in the middle of a helical stem. The same thermodynamic drive is at work on the hydrogen-bond donors and acceptors buried in the cores of molten globules, and not surprisingly, molten globules contain significant amounts of α-helix and β-sheet.

Late Events in RNA Folding

Since suprahelical structure formation occurs on a seconds-to-minutes time scale, it can be studied using stopped-flow methods, and the folding pathways of many catalytic RNAs are now being addressed that way. Of the many RNAs larger than tRNAs that have been investigated recently, the one that has been most thoroughly scrutinized is the self-splicing intron from *Tetrahymena* (see Zarrinkar and Williamson 1996 and references therein). Surprisingly, significant progress has been made with this system even though the three-dimensional structure of only one of its domains is known (Cate et al. 1996).

At least for the self-splicing intron, the formation of suprahelical structures is a process, in which the n^{th} step depends on the prior completion of $(n\text{-}1)^{th}$ step, and folding intermediates accumulate because some steps are slow. It is interesting that some of the late steps in this RNA's pathway involve the formation of helix by sequences that are well-separated in the molecule. This observation validates a point made earlier, namely that helices should not be regarded as merely secondary structure elements.

The existence of folding pathways has been a matter of hot dispute in the protein field for many years. They may exist for some proteins, but almost certainly do not for all proteins. The slow step in the folding of small, globular protein domains, which is the transformation of molten globules into folded proteins, does not seem to be a strictly sequential process. The fundamental difference between a molten globule and a mature protein is that the hydrophobic core of the latter is about as densely packed as it conceivably could be, whereas the core of the former is not. The rearrangement of side chains required to achieve the densely packed, final state is slow because correlated motions are required to achieve it, and as it occurs, the helices and sheets in the molecule stabilize in their final forms.

The folding pathway of the self-splicing intron may prove to be the paradigm for all RNAs, but it would be foolish to assume that this is so. RNA folding may prove as sequence-specific and idiosyncratic as protein folding is, and the number of RNAs whose folding properties are well-characterized is so small that it would be wise not to attempt to generalize at this point.

CONCLUSION

Although significant progress has been made on the RNA folding problem, it is hard to imagine that we will be sophisticated enough to predict RNA structures ab initio any time soon, and it will be a while before biochemically derived conformational models are produced that are accurate enough to pass for structures. Those responsible for staffing academic departments and research institutes will have to keep hiring structural biologists for at least a little while longer.

ACKNOWLEDGMENTS

I thank the many colleagues who helped me locate references and answered my questions. I am particularly grateful to Professors I. Tinoco and D. Turner for their comments on the first draft of this manuscript. This work was supported by a grant from the National Institutes of Health (GM-41651).

REFERENCES

Aboul-ela F., Karn J., and Varani G. 1995. The structure of the human immunodeficiency virus type 1 TAR RNA reveals principles of RNA recognition by TAT protein. *J. Mol. Biol.* **253:** 313–332.

Allain F.H.-T. and Varani G. 1995. Structure of the P1 helix from group I self-splicing introns. *J. Mol. Biol.* **250:** 333–353.

Battiste J.L., Tan R., Fraenkel A., and Williamson J.R. 1994. Binding of an HIV Rev peptide to Rev responsive element RNA induces formation of purine-purine base pairs. *Biochemistry* **33:** 2741–2747.

Battiste J.L., Hongyuan M., Rao N.S., Tan R., Muhandiram D.R., Kay L.E., Frankel A.D., and Williamson J.R. 1996. Alpha helix-RNA major groove recognition in an HIV-1 Rev peptide-RRE RNA complex. *Science* **273:** 1547–1551.

Brion P. and Westhof E. 1997. Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **26:** 113–137.

Brooks B., Bruccoleri R., Olafson B., States D., Swaminathan S., and Karplus M. 1983. CHARMM: A program for macromolecular energy minimization and molecular dynamics calculations. *J. Computat. Chem.* **4:** 187–217.

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

Brünger A. 1992. *X-PLOR Version 3.1: A System for X-ray crystallography and NMR.* Yale University Press, New Haven, Connecticut.

Butcher S.E., Dieckmann T., and Feigon J. 1997. Solution structure of the conserved 16 S-like ribosomal RNA UGAA tetraloop. *J. Mol. Biol.* **268:** 348–358.

Cate J., Gooding A.R., Podell E., Zhou K., Golden B.L., Kundrot C.E., Cech T.R., and Doudna J.A. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273:** 1678–1685.

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

Chou P.Y. and Fasman G.D. 1974. Conformational parameters for amino acids in helical, beta sheet and random coil regions calculated from proteins. *Biochemistry* **13:** 211–221.

Correll C.C., Freeborn B., Moore P.B., and Steitz T.A. 1997. Metals, motifs and recognition in the crystal structure of a 5S rRNA domain. *Cell* **91:** 705–712.

Crothers D.M. 1979. Physical studies of tRNA in solution. In *Transfer RNA: Structure, properties and recognition* (ed. P.R. Schimmel et al.), pp. 163–176. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Crothers D.M. and Cole P.E. 1978. Conformational changes in tRNA. In *Transfer RNA* (ed. S. Altman), pp. 196–247. MIT Press, Cambridge, Massachusetts.

Crothers D.M., Cole P.E., Hilbers C.W., and Shulman R.G. 1974. The molecular mechanism of thermal unfolding of *Escherichia coli* formyl methionine transfer RNA. *J. Mol. Biol.* **87:** 63–88.

Dallas A. and Moore P.B. 1997. The loop E-loop D region of *Escherichia coli* 5S rRNA: the solution structure reveals an unusual loop that may be important for binding ribosomal proteins. *Structure* **5:** 1639–1653.

Doty P., Boedtker H., Fresco J.R., Haselkorn R., and Litt M. 1959. Secondary structure in ribonucleic acids. *Proc. Natl. Acad. Sci.* **45:** 482–499.

DuBuy B. and Weissman S.M. 1971. Nucleotide sequence of *Pseudomonas fluorescens* 5S ribonucleic acid. *J. Biol. Chem.* **246:** 747–761.

Fan P., Suri A.K., Fiala R., Live D., and Patel D. 1996. Molecular recognition in the FMN-RNA aptamer complex. *J. Mol. Biol.* **258:** 480–500.

Fourmy D., Recht M.I., Blanchard S.C., and Puglisi J.D. 1996. Structure of the A site of *Escherichia coli* 16S ribosomal RNA complexed with an aminoglycoside antibiotic. *Science* **274:** 1367–1371.

Fox G.E. and Woese C.R. 1975. 5S RNA secondary structure. *Nature* **256:** 505–507.

Fresco J.R. and Alberts B.M. 1960. The accommodation of non-complementary bases in helical polyribonucleotides and deoxyribonucleic acids. *Proc. Natl. Acad. Sci.* **46:** 311–321.

Fresco J.R., Alberts B.M., and Doty P. 1960. Some molecular details of the secondary structure of ribonucleic acid. *Nature* **188:** 98–101.

Friesner R.A. and Gunn J.R. 1996. Computational studies of protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **25:** 315–342.

Gaspin C. and Westhof E. 1995. An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J. Mol. Biol.* **254:** 163–174.

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

Gutell R.R. 1996. Comparative sequence analysis and the structure of 16S and 23S rRNA. In *Ribosomal RNA. Structure, evolution, processing and function in protein biosynthesis* (ed. A. Dahlberg and R. Zimmerman), pp. 111–128. CRC Press, Boca Raton, Florida.

Gutell R.R. and Woese C.R. 1990. Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of non-canonical pairs. *Proc. Natl. Acad. Sci.* **87:** 663– 667.

Harris M.E., Nolan J.M., Malhotra A., Brown J.W., Harvey S.C., and Pace N.R. 1994. Use of photoaffinity crosslinks and molecular modeling to analyze the global architecture of ribonuclease P RNA. *EMBO J.* **13:** 3953–3963.

Holley R.W., Apgar J., Everett G.A., Madison J.T., Marquisee M., Merrill S.H., Penswick J.R., and Zamir A. 1965. Structure of a ribonucleic acid. *Science* **147:** 1462–1465.

Jiang L., Suri A.K., Fiala R., and Patel D.J. 1997. Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex. *Chem. Biol.* **4:** 35–50.

Joseph S. and Noller H.F. 1996. Mapping the rRNA neighborhood of the acceptor end of tRNA in the ribosome. *EMBO J.* **15:** 910–916.

Joseph S., Weiser B., and Noller H.F. 1997. Mapping the inside of the ribosome with an RNA helical ruler. *Science* **278:** 1093–1098.

Jucker F.M. and Pardi A. 1995. Solution structure of the CUUG hairpin loop: A novel RNA tetraloop motif. *Biochemistry* **34:** 14416–14427.

Jucker F.M., Heus H.A., Yip P.F., Moors E.H.M., and Pardi A. 1996. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264:** 968–980.

Kalurachchi K., Uma K., Zimmermann R.A., and Nikonowicz E.P. 1997. Structural features of the binding site for ribosomal protein S8 in *Escherichia coli* 16S rRNA defined using NMR spectroscopy. *Proc. Nat. Acad. Sci.* **94:** 2139–2144.

Levinthal C. 1968. Are there pathways for protein folding? *J. Chem. Phys.* **65:** 44–47.

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

Levitt M., Gerstein M., Huang E., Subbiah S., and Tsai J. 1996. Protein folding: The endgame. *Annu. Rev. Biochem.* **66:** 549–579.

Louise-May S., Auffinger P., and Westhof E. 1995. RNA structure from molecular dynamics simulations. In *Biological structure and dynamics. Proceedings of the 9th Conversation, State University of New York* (ed. R.H. Sarma and M.H. Sarma), pp. 1– 18. Adenine Press, Albany, New York.

———. 1996. Calculations of nucleic acid conformations. *Curr. Opin. Struct. Biol.* **6:** 298–298.

Madison J.T. 1968. Primary structure of RNA. *Annu. Rev. Biochem.* **37:** 131–148.

Maitland G.C., Rigby M., Smith E.B., and Wakeham, W.A. 1981. Intermolecular forces. Their origin and determination. *Int. Ser. Monogr. Chem.*, vol. 3. Oxford University Press, Oxford, United Kingdom.

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

Malhotra A. and Harvey S.C. 1994. A quantitative model of the *Escherichia coli* 16S RNA in the 30S ribosomal subunit. *J. Mol. Biol.* **240:** 308–340.

Michel F. and Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216:** 585–610.

Moore P.B. 1995. Structure and function of 5S RNA. In *Ribosomal RNA: Structure, evolution, processing and function in protein synthesis* (ed. R.A. Zimmermann and A.E. Dahlberg), pp. 199–236. CRC Press, Boca Raton, Florida.

Mueller F. and Brimacombe R. 1997. A new model for the three-dimensional folding of *Escherichia coli* 16S ribosomal RNA. I. Fitting the RNA to a 3D electron microscopic map at 20 Å. *J. Mol. Biol.* **271:** 524–544.

Nishikawa K. and Takemura S. 1974. Nucleotide sequence of 5S RNA from *Torulopsis utilis. FEBS Lett.* **40:** 106–109.

Noller H.F., Jr., and Moldave K., eds. 1988. Ribosomes. *Methods Enzymol.* vol. 164. Academic Press, San Diego.

Noller H.F. and Woese C.R. 1981. Secondary structure of 16S ribosomal RNA. *Science* **212:** 403–411.

Pogozelski W.K., McNeese T.J., and Tullius T.D. 1995. What species is responsible for strand scission in the reaction of $[\mathrm{Fe^{II}EDTA}]^{2-}$ and H_2O_2 with DNA? *J. Am. Chem. Soc.* **117:** 6428–6433.

Puglisi J.D., Chen L., Blanchard S., and Frankel A.D. 1995. Solution structure of a bovine immunodeficiency virus Tat-TAR peptide-RNA complex. *Science* **270:** 1200–1203.

Puglisi J.D., Tan R., Calnan B.J., Frankel A.D., and Williamson J.R. 1992. Conformation of the TAR-Arginine complex by NMR spectroscopy. *Science* **257:** 76–80.

Saenger W. 1984. *Principles of nucleic acid structure.* (Springer Advanced Texts in Chemistry series). Springer-Verlag, New York.

Samaha R.R., O'Brien B., O'Brien T.W., and Noller H.F. 1994. Independent in vitro assembly of a ribonucleoprotein containing the 3′ domain of 16S rRNA. *Proc. Natl. Acad. Sci.* **91:** 7884–7888.

Schlick T., Barth E., and Mandziuk M. 1997. Biomolecular dynamics at long timesteps: Bridging the timescale gap between simulation and experimentation. *Annu. Rev. Biophys. Biomol. Struct.* **26:** 181–222.

Serra M.J. and Turner D.H. 1995. Predicting thermodynamic properties of RNA. *Methods Enzymol.* **259:** 242–261.

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

Stern S., Weiser B., and Noller H.F. 1988. Model for the 3-dimensional folding of 16S ribosomal RNA. *J. Mol. Biol.* **204:** 447–481.

Tan R.K.Z. and Harvey S.C. 1993. YAMMP: Development of a molecular mechanics program using the modular programming method. *J. Computat. Chem.* **14:** 455–470.

Tinoco I., Jr., Borer P.N., Dengler B., Levine M.D., Uhlenbeck O.C., Crothers D.M., and Gralla J. 1973. Improved estimation of secondary structure in ribonucleic acids. *Nat. New Biol.* **246:** 40–41.

Tuerk C., Gauss P., Thermes C., Groebe D.R., Gayle M., Guild N., Stormo G., d'Aubenton-Carafa Y., Uhlenbeck O.C., Tinoco I., Jr., Brody E.N., and Gold L. 1988. CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci.* **85:** 1364–1368.

Turner D., Sugimoto N., and Freier S.M. 1988. RNA structure prediction. *Annu. Rev. Biophys. Biophys. Chem.* **17:** 167–192.

Weiner P. and Kollman P. 1981. AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions. *J. Computat. Chem.* **2:** 287– 303.

Weitzmann C.J., Cunningham P.R., Nurse K., and Ofengand J. 1993. Chemical evidence for domain assembly of the *Escherichia coli* 30S ribosome. *FASEB J.* **7:** 177–180.

Westhof E., Wesolowski D., and Altman S. 1996. Mapping in three-dimensions of regions in catalytic RNA protected from attack by an Fe(II)-EDTA reagent. *J. Mol. Biol.* **258:** 600– 613.

Woese C.R., Winker S., and Gutell R.R. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops." *Proc. Natl. Acad. Sci.* **87:** 8467–8471.

Yang Y., Kochpyan M., Burgstaller P., Westhof E., and Famulok M. 1996. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. *Science* **272:** 1343–1347.

Ye X., Kumar R.A., and Patel D.J. 1995. Molecular recognition in the bovine immunodeficiency virus Tat peptide-TAR RNA complex. *Chem. Biol.* **2:** 827–840.

Zarrinkar P.P. and Williamson J.R. 1996. The kinetic folding pathway of the *Tetrahymena* ribozyme reveals possible similarities between RNA and protein folding. *Nat. Struct. Biol.* **3:** 432–438.

Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* **244:** 48–52.

Zuker M. and Jacobson A.B. 1995. Well-determined regions in RNA secondary structure prediction: Analysis of small subunit ribosomal RNA. *Nucleic Acids Res.* **23:** 2791–2798.