Chapter 15

An Updated Recursive Algorithm for RNA Secondary Structure Prediction with Improved Thermodynamic Parameters

David H. Mathews[1], Troy C. Andre[1], James Kim[1], Douglas H. Turner[1,3], and Michael Zuker[2]

[1]Department of Chemistry, University of Rochester, Rochester, NY 14627–0216
[2]Institute for Biomedical Computing, Washington University,
St. Louis, MO 63110

An updated recursive algorithm that minimizes free energy predicts 82.5% of phylogenetically determined base pairs from sequence in four small subunit rRNAs, four group I introns, three group II introns, and 41 tRNAs. The rRNAs and group II introns were folded in phylogenetically determined domains of no more than 500 nucleotides. The algorithm incorporates recently determined thermodynamic parameters for the free energies of internal loops of 2 by 1 and 2 by 2 nucleotides. New free energy bonuses for tetraloops and triloops have been developed by consideration of the database of phylogenetically determined structures. Finally, new rules for coaxial stacking have been applied. This new version will be available in FORTRAN for Unix machines and a C++ version is now available for use on Personal Computers with Windows 95 or Windows NT. The program was used to explore structures predicted to have a free energy near the minimum. On average, a structure with 92% of phylogentically determined base pairs is found within 2% of the minimum free energy. For a roughly 400 nucleotide RNA, this is typically 2.3 kcal/mol above the minimum free energy. Implications for determining RNA secondary structure from sequence are discussed.

Sequence information is being collected at a rate faster than a million nucleotides per day. These data are having a large impact on our understanding of biology and medicine, but to extract the maximum amount of information, additional computational analysis of the database will be necessary. One area of analysis is the prediction of RNA secondary structure from sequence.

When a large number of phylogenetically related sequences are known, sequence comparison is the standard technique for determining RNA secondary structure (*1*). Commonly, however, there are no related sequences or too few sequences for rigorous comparative methods. Thermodynamics can be used to predict

[3]Corresponding author

secondary structure when only one or a few sequences are available and to facilitate sequence comparison when many sequences are available (*2-6*).

Several approaches have been investigated for the prediction of RNA secondary structure from a single sequence. The recursive algorithm used in this study predicts a lowest free energy structure and a set of structures, called sub-optimal structures, within a given increment of free energy (*4,7*). It can also be used to generate an energy dot plot, a graphical representation of all base pairs involved in suboptimal structures. A dot plot can identify base pairs that are well determined in the predicted structure (*8*). Another recursive algorithm, by McCaskill (*9*), calculates a partition function, thus providing the probability of pairing of individual bases and the probability of individual base pairs. Another approach for RNA structure prediction is the genetic algorithm (*10,11*). The genetic algorithm has the advantage of being able to simulate possible kinetic contributions to folding. It can also predict pseudoknots, which recursive algorithms cannot. The genetic algorithm, however, takes much more time to predict a structure than the recursive algorithms.

All three approaches to predicting secondary structure from a single sequence rely upon nearest neighbor free energy parameters. Parameters for Watson-Crick helical regions are well determined experimentally (*12,13*), but the parameters for unpaired regions are an area of active research (*14-18*). It is apparent that many of the free energy models are over-simplified because they neglect the sequence of unpaired regions. Presently, it is impossible to study every possible variation for even small structural motifs. For example, a symmetric internal loop of four with two unpaired nucleotides per strand has 4,704 distinct sequences when the closing base pairs are considered. Therefore, the free energies of these motifs must be deduced by other methods. Development of new experimental approaches, however, may eventually allow measurement of most sequence variations (*19,20*).

Several methods are available for approximating the free energy parameters of structural motifs. Often, an empirical model is developed that is based on free energies determined by optical melting experiments for representative sequences (*12-18*). Another method for assigning stability to unpaired regions is to base it on the frequency of occurrence of different sequences for motifs in a database of phylogenetic structures (*3*). Lastly, parameters can be varied so as to optimize the accuracy of computer folded sequences against their phylogenetically determined structures (*2,3,21*). The last two approaches may partially simulate the effect of tertiary contacts on the stability of secondary structure. A combination of all three methods were used to update the parameters in this study. Eventually, it may be possible to calculate these parameters explicitly using methods such as free energy perturbation.

This preliminary report presents recent progress in secondary structure prediction based on free energy minimization. The following changes have been implemented: The method for forcing base pairs has been improved. A filter that removes isolated Watson-Crick or G-U base pairs (those that cannot stack on any other Watson-Crick or G-U pair) has been incorporated. Recently measured free energies for 2 by 2 internal loops (Xia, T.; McDowell, J. A.; Turner, D. H. In preparation.), 2 by 1 internal loops (*15*), and hairpin loops (*18*) have also been incorporated. Finally, a new model for coaxial stacking of helixes has been developed.

Methods

Folding Algorithm. The folding algorithm is freely available in several formats. A C++ coded version, called RNAstructure version 2.0, for personal computers running Windows 95 or Windows NT is available on the Turner lab homepage at http://rna.chem.rochester.edu. This Windows version has a graphical user interface. The computational time for a typical sequence, 433 nucleotides of the group I LSU intron in *Tetrahymena thermophila*, is 135 seconds on a Pentium 120 with 48 MB of RAM running Windows 95. The structures for this paper were predicted with the C++ code compiled for a Silicon Graphics work station.

A version of the algorithm in Fortran, called Mfold, for use on Unix machines will be at Michael Zuker's ftp directory, ftp://snark.wustl.edu/pub. This version will also be available for online folding through a World-Wide-Web interface at the Zuker homepage, http://www.ibc.wustl.edu/~zuker/rna/form1.cgi.

Thermodynamic Parameters. The thermodynamic parameters are taken from previous studies of RNA folding (*2,3*) with the exception of recently studied motifs. Changes have been made in the stabilities of internal loops of 1 by 2 nucleotides (one unpaired nucleotide opposite two unpaired nucleotides) and 2 by 2 nucleotides (two unpaired nucleotides opposite two unpaired nucleotides), hairpin loops, and multibranch loops. The complete tables of parameters are available on the Turner Lab Homepage at http://rna.chem.rochester.edu.

Internal Loops. The free energy parameters of 2 by 2 internal loops (also called tandem mismatches) have been studied for the cases of symmetric mismatches (*14*) and non-symmetric mismatches (Xia, T.; McDowell, J. A.; Turner, D. H. In preparation.). A preliminary model by Xia et al. for approximating the stabilities of 2 by 2 loops that have not been measured was used to fill a table of thermodynamic parameters for all possible 2 by 2 mismatches and closing base pairs for a total of 4,704 parameters. The algorithm now consults this table when determining the free energy of a 2 by 2 internal loop.

Adjustments to the table of 2 by 2 internal loops were made based on comparisons of predicted and known secondary structures. The stability of $\begin{array}{c}\text{GACU}\\ \text{CACA}\end{array}$ was increased from 3.7 to 2.2 kcal/mol. The parameters for $\begin{array}{c}\text{GGXU}\\ \text{CAYA}\end{array}$ with XY as AA, AC, AG, CA, CC, CG, CU, GA, GG, UC, and UU were made 0.7, 0.4, -0.7, 0.4, 0.4, 1.2, 0.4, -0.7, 0.3, 0.6, and 0.5 kcal/mol, respectively. These are within 0.5 kcal/mol of the preliminary model, with the exception of CG, which is improved in stability by 0.8 kcal/mol.

The revised free energy rules for the 2,304 possible 2 by 1 internal loops are based on the work of Schroeder et al. (*15*). A table containing a free energy for each possible 2 by 1 loop and closing base pair was added to the algorithm. The parameters for C-G closures were those measured by Schroeder et al. (*15*). The parameters for other closing pairs were estimated by adding 0.7 kcal/mol for each A-U or G-U closure.

Hairpin Loops. The free energy of hairpins is based on the model of Serra and co-workers (*16-18*). In this model, hairpin stability for loops larger than three nucleotides is independent of sequence with the exception of the first mismatch and closing pair. The free energy for a hairpin loop, ΔG°_{HL}, is:

$$\Delta G_{HL}{}^\circ = \Delta G^\circ_i + \Delta G^\circ_{stack} \qquad (1)$$

where ΔG°_i is the free energy penalty for the closure of a loop of length i and ΔG°_{stack} is the free energy for stacking of the first mismatch on the helix and is approximated by:

$$\Delta G^\circ_{stack}\ (\text{kcal/mol}) = \Delta G^\circ_{mm} + 0.6\ (\text{if closed by an A-U or U-A pair}) - 0.7\ (\text{if first mismatch is GA or UU}) \qquad (2)$$

ΔG°_{mm} is the free energy for a mismatch at the end of a helix (*12*). For the algorithm, the parameters for ΔG°_{stack} were varied within 1.0 kcal/mol from the model to maximize the number of correctly predicted basepairs in the database of structures presented below. This was done by repeatedly making changes based on comparisons of predicted structures to phylogenetic structures. The final free energies are shown in Table I. The parameter ΔG°_i is 4.9, 5.0, 5.0, 5.0, 4.9, and 5.5 kcal/mol for loops of length 4-9, respectively (*18*). This model is known to predict free energy changes for the formation of several naturally occurring hairpin loops, but underestimates the stability reported for the hexanucleotide hairpin loop ACAGUGCU (*22*).

For loops of three nucleotides, the free energy is entirely independent of loop sequence:

$$\Delta G^\circ_{37}\ (\text{kcal/mol}) = 4.8 + 0.6\ (\text{if closed by an A-U or U-A pair}) \qquad (3)$$

Closure of hairpins of less than three nucleotides are not considered in the folding algorithm.

Specific tetraloops and triloops, i.e. hairpin loops of four and three nucleotides, respectively, are given an enhanced stability. Some of these hairpins are known to be more stable than predicted by the model above (*23-27*) and it is known that these motifs are an important component of tertiary structure (*28-30*) and are therefore stabilized by tertiary contacts. Previous studies have given stability bonuses to tetraloops according to the sequence of the unpaired nucleotides alone (*2,3*). A search of the database of phylogenetic structures of small and large subunit ribosomal RNA and Group I introns (*31,32*) shows that the occurrence of tetraloops also depends on the closing base pair (*33*). Thus we tried assigning bonuses to specific tetraloops and triloops according to the sequence of the unpaired nucleotides and the closing base pair. Table II shows the bonuses given to tetraloops and triloops. The bonuses are based on phylogenetic occurrence and on optimizing the accuracy of folding. There are 50 tetraloops with varying bonus stability, whereas the study of Walter et al. (*2*) gave a flat -2.0 kcal/mol to 78 tetraloops when all the closing base pair possibilities are considered.

Table I. Stability of Closing Base pair and First Mismatch in Hairpin Loops.

Base			ΔG°_{37} (kcal/mol)		
pair	X↓	Y→ A	C	G	U
AX	A	-0.0	-0.4	+0.3	-0.3
UY	C	-0.0	-0.1	-1.5	+0.5
	G	-0.5	-1.1	-0.2	+0.1
	U	-0.3	-0.2	-0.0	-0.5
CX	A	-1.5	-1.1	-1.4	-1.8
GY	C	-1.0	-0.9	-2.9	-0.6
	G	-1.9	-2.0	-1.6	-1.2
	U	-1.7	-1.4	-1.9	-1.5
GX	A	-0.7	-1.8	-1.5	-2.1
CY	C	-1.1	-0.5	-3.0	-0.5
	G	-2.1	-2.9	-1.4	-1.4
	U	-1.9	-1.0	-2.1	-1.4
GX	A	-0.2	-0.4	-0.0	-0.3
UY	C	-0.3	-0.1	-1.5	-0.2
	G	-0.9	-1.1	-0.2	+0.1
	U	+0.1	-0.2	-0.4	-0.9
UX	A	-0.2	-0.2	-0.1	-0.5
AY	C	-0.1	-0.0	-1.2	-0.2
	G	-1.2	-1.2	-0.6	-0.0
	U	-0.3	-0.0	-0.5	-0.9
UX	A	-0.4	-0.0	-0.5	-0.5
GY	C	-0.1	-0.0	-1.7	+0.1
	G	-1.2	-1.2	-0.6	-0.7
	U	-0.6	-0.0	-0.5	-0.6

Multibranch Loops. Coaxial stacking of helixes and a Jacobson-Stockmayer function (*34*) for the free energy of multibranch loops (also called junctions) cannot be incorporated into a computationally efficient recursive algorithm for secondary structure prediction. Models for both are included in a program called efn2 (for second energy function) that re-calculates the free energy of each sub-optimal structure. These free energies are used to re-order the structures by over-all stability and the lowest free energy structure after the efn2 calculation is the predicted structure.

Table II. Tetraloop and Triloop Bonuses in kcal/mol.

Tetraloop:		Tetraloop:	
AAUCAU	-1.0	GUGAAC	-1.5
AGAAAU	-2.0	GUUCGC	-1.5
AGAGAU	-1.5	UCAGGG	-1.5
AGCAAU	-2.0	UGAAAA	-2.0
AGUAAU	-2.0	UGAGAG	-2.5
AGUGAU	-1.5	UGCAAA	-1.5
CCAAGG	-1.5	UGCAAG	-1.5
CCUUGG	-1.5	UGCCAA	-1.5
CGAAAG	-3.5	UGGAAA	-2.0
CGAGAG	-1.0	UGGAAG	-1.5
CGAGAG	-3.0	UGUAAA	-1.5
CGCAAG	-3.5	UGUGAA	-1.5
CGCAUG	-1.5	UUAGGG	-1.5
CGCCAG	-2.0	UUCCAA	-1.5
CGCGAG	-2.0	UUCCCA	-3.0
CGGAAG	-3.0	UUCCGG	-1.5
CGUAAG	-3.0	UUGAGG	-1.5
CGUGAG	-1.5	UUUAGG	-1.5
CGUGAG	-3.0	UUUCGG	-1.5
CUAAGG	-1.5		
CUACGG	-2.5		
CUUCGG	-3.5	Triloop:	
GCUUGC	-1.5	AAAAU	-1.5
GGAAAC	-3.0	AGACU	-1.5
GGAGAC	-1.0	CAAAG	-2.0
GGCAAC	-2.5	CGACG	-2.0
GGCGAC	-1.5	GAAAC	-2.5
GGGAAC	-2.0	GAUUU	-0.5
GGUAAC	-1.5	GGACC	-2.5
GGUGAC	-1.5	UAAAA	-1.0
GUGAAC	-1.5	UGACA	-1.0

In the recursive algorithm, the free energy penalty for closing a multibranch loop is given by the linear approximation:

$$\Delta G^\circ \text{ (kcal/mol)} = a + bn + ch \qquad (4)$$

where n is the number of unpaired nucleotides in the loop, h is the number of helixes that branch from the loop, and a = 4.6, b = 0.2, and c = 0.1 (*3,10*).

The Jacobson-Stockmayer function is non-linear and can therefore not be used in a recursive algorithm. In efn2, however, the Jacobson-Stockmayer function is used for multibranch loops larger than six nucleotides and the free energy penalty is given by:

$$\Delta G^\circ \text{ (kcal/mol)} = a + 6b + (1.1)\ln(n/6) + ½ + ch. \qquad (5)$$

The linear approximation of equation 1 is used for loops of six and fewer nucleotides.

In the recursive algorithm, a free energy bonus for base stacking is assigned for each unpaired nucleotide 3' or 5' to helixes exiting the loop. These parameters are dangling end free energies determined experimentally (*12*). In efn2, the base stacking bonuses are only assigned to helixes not involved in coaxial stacking.

Efn2 gives an enhanced stability for coaxial stacking of adjacent helixes with one intervening nucleotide at most. This stability is based on the work of Walter et al. (*2*) and Kim et al. (*35*). When helixes have no intervening nucleotides, a table is consulted for the stacking bonus. This bonus is the free energy parameter for stacking of base pairs in a helix made more favorable by 1.0 kcal/mol. The addition of the -1.0 kcal/mol was found to improve the accuracy of structure prediction. With one intervening nucleotide, coaxial stacking is allowed when there is a nucleotide (5' to the 5' helix or 3' to the 3' helix) that can make an intervening mismatch. There are two distinct stacks and parameters for each are contained in separate tables (Figure 1). The first stack is on the side of the continuous backbone which, in Figure 1, is a 5' A-G 3' on a U-A. The parameters for this stack are set equal to the parameter of a terminal mismatch in a hairpin loop (Table I). The second stack is on the side where the backbone opens for the entering and exiting strands. In Figure 1, this is an A-G on a C-G. This is mostly sequence independent, with a value of -1.5 kcal/mol. The exceptions to this are that A-A mismatches have a stack of -1.0 kcal/mol; A-U and C-G intervening "mismatches" of either orientation stack with -2.5 kcal/mol; and G-U wobbles are given -2.3 kcal/mol. To find the lowest free energy possible for a multibranch loop, efn2 uses a recursive algorithm to search for the most favorable combination of interactions. This is necessary because helixes involved in coaxial stacking cannot have dangling ends and cannot coaxially stack on more than one other helix.

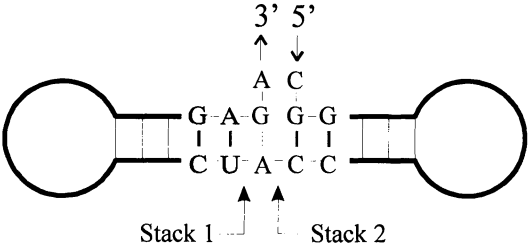

Figure 1. Coaxial Stacking of two helixes with an intervening mismatch. Stack 1 is the stack of the mismatch with a continuous backbone. Stack 2 is the stack of the mismatch with an open backbone.

Exterior Loops. Exterior loops are open loops that contain the ends of a sequence. This version of the algorithm gives bonuses for dangling ends and coaxial stacking in exterior loops using the same model as multibranch loops.

Removal of Isolated Base Pairs. To improve the accuracy of folding, isolated base pairs (those not adjacent to a possible Watson-Crick or G-U pair) are not allowed because they are rare in the database of known secondary structures. This is accomplished during the fill routine of the dynamic programming algorithm. Before calculating $V_{i,j}$, the lowest free energy for the sequence fragment from nucleotides i to j, with i and j base paired, the algorithm checks whether an adjacent canonical pair (Watson-Crick or G-U) is possible (either i+1 with j-1 or i-1 with j+1). If not, $V_{i,j}$ is set to a large integer used to represent an infinite free energy (1600 kcal/mol). This filters out isolated base pairs.

Constraining a Nucleotide to a Specific Base Pair. The folding algorithm allows the user to specify base pairs that are to occur in the final structure. In prior versions of the program, a large free energy bonus was assigned to base pairs specified by the user. This had the disadvantage of distorting the energy dot plot and making it difficult to compute a representative sample of foldings. Furthermore, for large structures, the base pair might not occur because the sum of other interactions might be more favorable than the arbitrary bonus.

The folding algorithm now forces a base pair by not allowing the two constituent nucleotides to be in the structure in any form other than the specified base pair. In other words, a large free energy (1600 kcal/mol) is given to an occurrence of the nucleotide not in the specified base pair, i.e. either single stranded or in any other base pair. This has the advantage of not complicating the determination of the free energy of a structure. The folding algorithm also uses a similar technique to force individual nucleotides to be double-stranded.

Scoring. Structures were scored against phylogenetic structures by:

$$\text{score} = (\#\text{ base pairs correctly predicted})/(\text{total bp in phylogenetic structure}) \qquad (6)$$

A phylogenetic base pair was considered correctly predicted if it was identical to a base pair in the structure or if the structure contained a base pair in which one of the two bases was slipped by one nucleotide. That is, given a phylogenetic base pair between nucleotides i and j, then a pair of i—j, (i-1)—j, (i+1)—j, i—(j-1), or i—(j+1) would be counted as a correct base pair. This allows for the common occurrence of helixes that have similar stabilities when slipped by one nucleotide.

Results

The accuracy of the folding algorithm was tested on a database of structures that contains four small subunit ribosomal RNAs (rRNA), four Group I introns, three Group II introns, and 41 transfer RNAs (tRNA). The rRNAs and Group II introns were folded in domains of less than 500 nucleotides (*3*). Table III summarizes the percent of correctly predicted base pairs for this version of the algorithm and the last

Table III. Percent of Correctly Predicted Base Pairs.

		% of phylogenetic base pairs predicted		
Structure*:	total bp	Walter et al. (2)	Current optimal after efn2	Current best suboptimal
Small Subunit rRNAs:				
E. coli	443	75.6	77.0	94.4
Rat mitochondria	216	56.0	65.3	79.2
H. volcanii	433	85.9	86.1	92.4
C. r. chloroplast	413	64.2	86.1	92.4
Group I Introns:				
LSU (I)	128	63.3	76.6	89.8
Yeast OX5α (I)	96	89.6	87.5	91.7
ND1 (I)	149	43.6	70.5	78.5
T4D (I)	75	74.7	73.3	89.3
Group II Introns:				
Yeast A1 (II)	188	75.5	86.2	94.7
Yeast A5 (II)	206	90.3	90.8	93.2
Yeast B1 (II)	219	89.5	89.5	93.2
tRNAs:	860	83.4	88.8	95.1
total†:	3426	76.6	82.5	91.6

*Structures, except tRNAs, are those used by Jaeger et al. (*3*). For 16s rRNA: *E. coli* (*32,36,37*), Rat mitochondria (*32,37*), *H.* (*Halobacterium*) *volcannii* (*32,37*), and *C. r.* (*Chlamydomonas reinhardtii*) chloroplast (*32,37*), structures are divided into four domains for folding as described by Jaeger et al. (*3*). The group I introns are LSU (from *Tetrahymena thermophila*), Yeast OX5α (from *Saccharomyces cerevisiae*), ND1 (from *Podospora anserina*), and T4D (*31,38*). Large regions of undetermined structure were replaced with four unpairing nucleotides as described in Jaeger et al. (*3*). Group II introns (*39,40*), all from *Saccharomyces cerevisiae*, are split into two domains as described by Jaeger et al. (*3*). The tRNAs (*41*) are a250, a590, c250, d250, d590, e250, e590, f250, f590, g235, g251, h250, h780, i203, i250, k590, k780, l235, l250, m235, m250, n110, n250, p235, p255, q250, q530, r235, r250, s250, s590, t250, t590, v250, v590, w250, w570, x250, x530, y250, and y590. Modified nucleotides that cannot pair or stack are forced to be single stranded except for modified nucleotides in the yeast phenylalanine tRNA, f590, which is known to adopt its secondary structure without modified nucleotides (*42*). All basepairs were correctly predicted for tRNA f590.

†Column total percentages are found by dividing the sum of correctly predicted base pairs in that column by the sum of phylogenetic base pairs.

version (*2*). Suboptimal structures were generated with a window size of zero, 10% sort in energy, and a maximum of 1000 structures for rRNAs and introns. For tRNAs, a maximum of 100 structures were generated. The window size of zero was chosen to maximize the number of subtle structural variations for efn2 to sort.

For the database of structures, 82.5% of base pairs were correctly predicted in the lowest free energy or optimal structures. When the suboptimal structures were sorted for accuracy, the most accurate structures were found to contain 91.6% of phylogenetic base pairs. For the rRNA and introns, the most accurate structure is on average 2.1% or 2.3 kcal/mol higher in free energy than the lowest free energy structure. In tRNAs, the average increase in free energy between the lowest free energy and the most accurate structures is 1.5% or 0.5 kcal/mol. In both cases, the most accurate structure is always within 6% of the free energy of the lowest free energy structure.

Another method of scoring is to examine whether a phylogenetic base pair occurs in any suboptimal structure. For this database, 98.0% of phylogenetic base pairs are found in at least one suboptimal structure.

Conclusions

Recent experimental studies show that the thermodynamic stabilities of non-Watson-Crick regions in RNA can be very sequence dependent (*14,15*). We have revised a recursive algorithm for prediction of RNA secondary structure (*2,4,7*) to allow inclusion of some of this sequence dependence. Since a limited amount of experimental data are available, the sequence dependence was initially approximated by crude models based on experimental results and frequencies of natural occurrence. These parameters were then adjusted to optimize the prediction of 3,426 base pairs contained in four small subunit rRNAs, four group I introns, three group II introns, and 41 tRNAs. A set of parameters were found that places 82.5% of the phylogenetically determined base pairs in the predicted lowest free energy structure. When a set of up to 1000 higher free energy structures is generated for each RNA and the structure most consistent with the phylogenetic structure is chosen for each RNA, 92% of known base pairs are found. This suggests that secondary structure is largely determined by free energy minimization. The small number of experimental free energy parameters and the small size of the secondary structure database used means the free energy parameters are underdetermined. Thus more experimental data and phylogenetically derived structures are required to further test this hypothesis.

For domains of about 400 nucleotides, hundreds of potential secondary structures are found within 2 kcal/mol of the lowest free energy structure. In contrast, biochemical experiments indicate that RNAs often form a single secondary structure in solution. This suggests there is much to learn about the factors determining RNA structure. For example, a single active structure may be stabilized by favorable tertiary interactions between non-Watson-Crick regions, and these interactions are not yet included in the folding algorithm. Experimental data on the unfolding of tRNA (*43,44*), group I introns (*45,46*), and a domain from 23S rRNA (*47*), however, indicate that single secondary structures form in the absence of tertiary interactions. Another possibility is that the kinetics of folding selects one of many energetically similar secondary structures. Many RNAs, however, are known to fold into their active structure both during transcription, when the molecule begins to fold before it is completely synthesized, and during renaturation from a denatured state, when the entire molecule is available for folding. There is no fundamental reason to expect the same kinetics in both cases. Possibly, the stability of the non-Watson-Crick regions, which are the subject of current study, can selectively stabilize the native structure compared to other structures. Presumably, understanding the reason for the prevalence of a single structure will allow better prediction of secondary structure from sequence. Given that an RNA of 400 nucleotides has more than 10^{100} possible secondary structures (*48*), it is encouraging that free energy minimization can provide a filter that narrows consideration to only about 10^3 structures. Selection of the native structure is then often possible from experimental data, such as chemical mapping and site directed mutagenesis, or by comparisons of sequences with similar functions (*5,6*).

Acknowledgments

The authors thank T. Xia, S. Schroeder, M. J. Serra, and T. W. Barnes for helpful discussions concerning the new thermodynamic parameters and M. E. Burkard and J. A. McDowell for discussions about computer programming. Thanks to R. R. Gutell for making phylogenetic secondary structures available. This work was supported by National Institutes of Health Grants GM22939 to D.H.T. and GM54250 to M.Z.. D.H.M. received support from National Institutes of Health Grant 5T32 GM07356.

Literature Cited

1. James, B. D.; Olsen G. J.; Pace N. R. *Methods Enzymol.* **1989**, *180*, 227-239.

2. Walter, A. E.; Turner, D. H.; Kim, J.; Lyttle, M. H.; Müller, P.; Mathews, D. H.; Zuker, M. *Proc. Natl. Acad. Sci.* **1994**, *91*, 9218-9222.

3. Jaeger, J. A,; Turner, D. H.; Zuker, M. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 7706-7710.

4. Zuker, M. *Science.* **1989**, *244*, 48-52.,

5. Mathews, D. H.; Banerjee, A. R.; Luan, D. D.; Eickbush, T. H.; Turner, D. H. *RNA.* **1997**, *3*, 1-16.

6. Lück, R; Steger, G.; Riesner D. *J. Mol. Biol.* **1996**, *258*, 813-826.

7. Zuker, M.; Stieger, P. *Nucleic Acids Res.* **1991**, *9*, 133-148.

8. Zuker, M.; Jacobson, A. B. *Nucleic Acids Res.* **1995**, *23*, 2791-2798.

9. McCaskill, J. S. *Biopolymers.* **1990**, *29*, 1105-1119.

10. Gultyaev, A. P.; Batenburg, F. H. D. van; Pleij, C. W. A. *J. Mol. Biol.* **1995**, *250*, 37-51.

11. Batenburg, F. H. D. van; Gultyaev, A. P.; Pleij, C. W. A. *J. Theor. Biol.* **1995**, *174*, 269-280.

12. Serra, M. J.; Turner, D. H. *Methods Enzymol.* **1995**, *259*, 242-261.

13. Freier, S. M.; Kierzek, R.; Jaeger, J. A.; Sugimoto, N.; Caruthers, M. H.; Neilson, T.; Turner D. H. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 9373-9377.

14. Wu, M.; McDowell, J. A.; Turner, D. H. *Biochemistry.* **1995**, *34*, 3204-3211.

15. Schroeder, S.; Kim, J.; Turner, D. H. *Biochemistry.* **1996**, *35*, 16105-16109.

16. Serra, M. J.; Lyttle, M. H.; Axenson, T. J.; Schadt, C. A.; Turner, D. H. *Nucleic Acids Res.* **1993**, *21*, 3845-3849.
17. Serra, M. J.; Axenson, T. J.; Turner, D. H. *Biochemistry.* **1994**, *33*, 14289-14296.
18. Serra, M. J.; Barnes, T. W.; Betschart, K.; Gutierrez, M. J.; Sprouse, K. J.; Riley, C. K.; Stewart, L.; Temel, R. E. *Biochemistry.* **1997**, *36*, 4844-4851.
19. Fodor, S. P. A.;Read, J. L.; Pirrung, M. C.; Stryer, L.; Lu, A. T.; Solas, D. *Science.* **1991**, *270*, 467-470.
20. O'Donnell-Maloney, M. J.; Smith, C. L.; Cantor, C. R. *Trends Biotechnol.* **1996**, *14*, 410-407.
21. Papanicolaou, C.; Gouy, M.; Ninio, J. *Nucleic Acids Res.* **1984**, *12*, 31-44.
22. Laing, L. G.; Hall, K. B. *Biochemistry.* **1996**, *35*, 13586-13596.
23. Jucker, F. M.; Pardi, A. *Biochemistry.* **1995**, *34*, 14416-14427.
24. SantaLucia, J., Jr.; Kierzek, R.; Turner, D. H. *Science.* **1992**, *256*, 217-219.
25. Antao, V. P.; Tinoco, I., Jr. *Nucleic Acids Res.* **1992**, *20*, 819-824.
26. Antao, V. P.; Lai, S. Y.; Tinoco, I., Jr. *Nucleic Acids Res.* **1991**, *19*, 5901-5905.
27. Varani, G.; Cheong, C.; Tinoco, I., Jr. *Biochemistry.* **1991**, *30*, 3280-3289.
28. Lehnert, V.; Jaeger, L.; Michel, F.; Westhof, E. *Chemistry & Biology.* **1996**, *3*, 993-1009.
29. Cate, J. H.; Gooding, A. R.; Podell, E.; Zhou, K.; Golden, B. L.; Kundrot, C. E.; Cech, T. R.; Doudna, J. A. *Science.* **1996**, *273*, 1678-1685.
30. Westhof, E.; Michel, F. *J. Mol. Biol.* **1990**, *216*, 585-610.
31. Damberger, S. H.; Gutell, R. R. *Nucleic Acids Res.* **1994**, *22*, 3508-3510.
32. Gutell, R. R. *Nucleic Acids Res.* **1994**, *22*, 3502-3507.
33. Woese, C. R.; Winker, S.; Gutell, R. R. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, 87, 8467-8471.
34. Jacobson, H.; Stockmayer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600-1606.
35. Kim, J.; Walter, A. E.; Turner, D. H. *Biochemistry.* **1996**, *35*, 13753-13761.
36. Moazed, D.; Stern, S.; Noller, H. F. *J. Mol. Biol.* **1986**, *187*, 399-416.
37. Gutell, R. R.; Weiser, B.; Woese, C. R.; Noller, H. F. *Prog. Nucleic Acid Res. Mol. Biol.* **1985**, *32*, 155-216.
38. Waring, R. B.; Davies, R. W. *Gene.* **1984**, *28*, 277-291.
39. Michel, F.; Umesono, K.; Ozeki, H. *Gene.* **1989**, *82*, 5-30.
40. Michel, F.; Jacquier, A.; Dujon, B. *Biochimie.* **1982**, *64*, 867-881.
41. Sprinzl, M.; Hartmann, T.; Weber, J.; Blank, J.; Zeidler, R. *Nucleic Acids Res.* **1989**, *17*, supplement, 1-189.
42. Hall, K. B.; Sampson, J. R.; Uhlenbeck, O. C.; Redfield, A. G. *Biochemistry.* **1989**, *28*, 5794-5801.
43. Hilbers, C. W.; Robillard, G. T.; Shulman, R. G.; Blake, R. D.; Webb, P. K.; Fresco, R.; Riesner, D. *Biochemistry.* **1976**, *15*, 1874-1882.
44. Crothers, D. M.; Cole, P. E.; Hilbers, C. W.; Shulman, R. G. *J. Mol. Biol.* **1974**, *87*, 63-88.
45. Banerjee, A. R.; Jaeger, J. A.; Turner, D. H. *Biochemistry.* **1993**, *32*, 153-163.
46. Jaeger, L.; Westhof, E.; Michel, F. *J. Mol. Biol.* **1993**, *234*, 331-346.
47. Laing, L. G.; Draper, D. E. *J. Mol. Biol.* **1994**, *237*, 560-576.
48. Zuker, M; Sankoff, D. *Bull. Math. Biol.* **1984**, *46*, 591-621.