Article No. jmbi.1999.2700 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 288, 911–940

JMB

Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure

David H. Mathews[1], Jeffrey Sabina[1], Michael Zuker[2] and Douglas H. Turner[1]*

[1]*Department of Chemistry University of Rochester Rochester, NY 14627-0216, USA*

[2]*Institute for Biomedical Computing, Washington University, St. Louis MO 63110, USA*

*Corresponding author

An improved dynamic programming algorithm is reported for RNA secondary structure prediction by free energy minimization. Thermodynamic parameters for the stabilities of secondary structure motifs are revised to include expanded sequence dependence as revealed by recent experiments. Additional algorithmic improvements include reduced search time and storage for multibranch loop free energies and improved imposition of folding constraints. An extended database of 151,503 nt in 955 structures? determined by comparative sequence analysis was assembled to allow optimization of parameters not based on experiments and to test the accuracy of the algorithm. On average, the predicted lowest free energy structure contains 73 % of known base-pairs when domains of fewer than 700 nt are folded; this compares with 64 % accuracy for previous versions of the algorithm and parameters. For a given sequence, a set of 750 generated structures contains one structure that, on average, has 86 % of known base-pairs. Experimental constraints, derived from enzymatic and flavin mononucleotide cleavage, improve the accuracy of structure predictions.

Keywords: RNA secondary structure; nearest neighbor parameters; free energy minimization; enzymatic cleavage; flavin mononucleotide cleavage

Introduction

Endeavors such as the human genome project are collecting nucleic acid sequence data at unprecedented rates of over one million nucleotides per day, and are thus providing enormous genetic detail. This wealth of data has ushered in an information age for biology. Fortunately, these advances are paralleled by a revolution in computer technology, allowing the maximum information to be extracted from the database.

One important area of database analysis is the determination of RNA structure from sequence. RNA structure is particularly suited for prediction because it is generally divided into two levels of complexity. The first level is secondary structure, involving canonical base-pairs. It is discrete in that each nucleotide is either paired or not. The interactions that govern secondary structure are generally stronger than the interactions that determine the next level of structural complexity, tertiary structure, i.e. the three-dimensional shape (Banerjee *et al.*, 1993; Jaeger *et al.*, 1993; Laing & Draper, 1994; Crothers *et al.*, 1974; Hilbers *et al.*, 1976; Mathews *et al.*, 1997). Therefore, the secondary structure can be determined largely independently of tertiary structure. The tertiary structure is then an additional level of detail for understanding structure and function (Cate *et al.*, 1996; Correl *et al.*, 1997; Michel & Westhof, 1990; Massire *et al.*, 1998; Harris *et al.*, 1997). Eventually, it may be possible to infer tertiary structure from interactions of secondary structure elements.

When many homologous RNA sequences are available, the standard technique for determining the secondary structure is comparative sequence analysis (James *et al.*, 1989; Pace *et al.*, 1999; Woese *et al.*, 1983). Comparative sequence analysis can be facilitated by predictions of secondary structures based on free energy minimization (Lück *et al.*,

Abbreviations used: FMN, flavin mononucleotide.

E-mail address of the corresponding author: turner@chem.chem.rochester.edu

1996; Mathews *et al*., 1997). When there are only one or a few known sequences for an RNA, free energy minimization can also be used to predict secondary structure models that can be tested against experimental data such as chemical modification and site directed mutagenesis (Walter *et al*., 1994a; Hofacker *et al*., 1994; Jaeger *et al*., 1989; Mathews *et al*., 1997).

Here, we present an enhanced algorithm for RNA secondary structure prediction by free energy minimization. Many recent advances are incorporated. New experimental results show that free energy parameters are more sequence-dependent than realized previously. In particular, a new model for Watson-Crick paired helices (Xia *et al*., 1998) and systematic studies of the stabilities of hairpin loops (Giese *et al*., 1998; Serra *et al*., 1993, 1994, 1997; Groebe & Uhlenbeck, 1988), small internal loops (Schroeder *et al*., 1996; Xia *et al*., 1997), and coaxially stacked helices (Kim *et al*., 1996; Walter *et al*., 1994a,b) are available. Based on these data, the thermodynamic parameters are adjusted and the sequence dependence of stability is expanded for several motifs in the algorithm. The additional sequence dependence increases the number of parameters in the program; therefore, the program and parameters were tested and refined by utilizing a large database of known secondary structures accessible largely through the World Wide Web (Gutell, 1994; Gutell *et al*., 1993; Schnare *et al*., 1996; Szymanski *et al*., 1998; Sprinzl *et al*., 1998; Larsen *et al*., 1998; Brown, 1998; Damberger & Gutell, 1994; Michel *et al*., 1989; Waring & Davies, 1984). This extensive testing is possible because of an improvement in the algorithm that accelerates multibranch loop searching and because of the revolution in computer technology. Inexpensive personal computers now have the speed and memory required for predicting RNA secondary structure for sequences of considerable length (Mathews *et al*., 1998). Additionally, the enforcement of folding constraints is improved in the algorithm and experimental constraints, derived from enzymatic and flavin mononucleotide (FMN) cleavage, are shown to further improve the accuracy of structure prediction.

Results

Thermodynamic parameters

Free energy parameters were adjusted for many motifs to include expanded sequence dependence based on new experimental data. The changes are briefly summarized here and complete details of the derivations are given in the Methods.

The stability of Watson-Crick helices is calculated using the INN-HB (individual nearest-neighbor-hydrogen bond) parameters described by Xia *et al*. (1998). This model includes a penalty term for each helix terminated by an A·U base-pair in order to account for the number of hydrogen bonds in the helix. Nearest-neighbor parameters for G·U base-pairs are derived from linear regression on a set of experimentally measured free energy contributions (He *et al*., 1991; Sugimoto *et al*., 1986; Xia *et al*., 1997; Freier *et al*., 1986a; Wu *et al*., 1995; McDowell & Turner, 1996; R. Kierzek & D.H.T., unpublished data; S.J. Schroeder & D.H.T., unpublished data;) under the assumption that the penalty for terminal A·U pairs (Xia *et al*., 1998) applies to terminal G·U pairs.

Stabilities of loop regions are known to be sequence dependent. Bulge loop stabilities are approximated with an updated parameter set derived from prior measurements (Longfellow *et al*., 1990; Groebe & Uhlenbeck, 1989; Fink & Crothers, 1972). Internal loop stabilities are revised as suggested by recent measurements. In particular, 2 × 2 (Xia *et al*., 1997), 2 × 1 (Schroeder *et al*., 1996), and 1 × 1 internal loops (R. Kierzek, M.E. Burkard & D.H.T., unpublished results) are each modeled with individual tables that contain every possible sequence variation.

Hairpin loop parameters are updated using the INN-HB nearest-neighbor model for calculation of stem stability and the recent database of hairpin stability measurements (Giese *et al*., 1998; Serra *et al*., 1997). A small penalty term previously used for hairpins closed by A·U base-pairs (Serra *et al*., 1994) is omitted because the INN-HB helical parameters sufficiently penalize terminal A·U pairs. Evidently, the prior penalty term was an artifact of an incomplete model for the sequence dependence of stem stability.

Free energy bonuses are given to certain tetraloops (hairpins of four nucleotides) because some are known to be especially stable (Tuerk *et al*., 1988; Antao & Tinoco, 1992; Antao *et al*., 1991; Varani *et al*., 1991) and others are known to be important in stabilizing tertiary structure (Costa & Michel, 1995; Butcher *et al*., 1997; Lehnert *et al*., 1996; Cate *et al*., 1996; Jucker & Pardi, 1995; Michel & Westhof, 1990; Jaeger *et al*., 1994; Murphy & Cech, 1994; Pley *et al*., 1994). The prior tetraloop bonus table (Walter *et al*., 1994a) did not include the sequence of closing base-pairs. The revised set of free energy parameters uses the closing base-pair as one of the criteria for giving a tetraloop bonus. This allows a more precise allocation of bonus free energy. The bonus assigned to a tetraloop sequence is based on the number of occurrences in a database of secondary structures. These bonuses can be omitted or changed for predicting structures of short RNAs that do not have tertiary interactions.

The sequence dependence of multibranch loops is also expanded. The initiation parameters are derived by comparing predicted and known secondary structures for a much larger set of sequences than used previously (Jaeger *et al*., 1989). Two sets of parameters are used, one to generate a set of possible structures with the dynamic programming algorithm and another to reorder the structures with free energies based on more complete energy rules, including coaxial stacking.

Coaxial stacking is expanded to include all intervening single mismatches based on the work by Kim *et al*. (1996). Coaxial stacking is also allowed outside of loops.

Accuracy

The accuracy of the algorithm is tested by predicting secondary structures of RNAs with structures determined by sequence comparison. A database of 151,503 nucleotides and 43,519 bp consisting of 22 small subunit rRNAs (Gutell, 1994), five large subunit rRNAs (Gutell *et al*., 1993; Schnare *et al*., 1996), 309 5 S rRNAs (Szymanski *et al*., 1998), 484 tRNAs (Sprinzl *et al*., 1998), 91 SRP RNAs (Larsen *et al*., 1998), 16 RNase P RNAs (Brown, 1998), 25 group I introns (Damberger & Gutell, 1994; Waring & Davies, 1984), and three group II introns (Michel *et al*., 1989) was assembled to test the algorithm and refine the thermodynamic parameters.

Table 1 shows the accuracy of the secondary structure prediction algorithm for each type of RNA. Results for two sets of energy parameters are shown. The first set is the current parameters with the expanded sequence dependence derived in this study. The second parameter set is that reported by Walter *et al*. (1994a). A third parameter set, whose accuracy is detailed in the Supplementary Material, is a crude method that counts hydrogen bonds in canonically base-paired regions, i.e. G·C pairs are given three units and each A·U or G·U pair is given two units. In this hydrogen bond parameter set, loop regions make no contribution to the count of hydrogen bonds. This parameter set serves as a control for the hypothesis that expanded sequence dependence of the thermodynamic parameters leads to more accurate structure prediction.

The results (Table 1) show that, with the revised parameters, 72.9 % of known base-pairs are predicted on average in the lowest free energy structure. This compares with 63.6 % with the parameters of Walter *et al*. (1994a). Furthermore, one structure of 750 suboptimal structures generated with free energies similar to the lowest free energy contains 86.1 % of known base-pairs on average. On average, this structure has a free energy 4.8 % higher than the minimal free energy. Finally, these 750 structures together contain 97.1 % of the known base-pairs. Standard deviations are given with the percentages to demonstrate the range of accuracy of secondary structure prediction that can be expected with a novel RNA sequence.

The group I introns and the RNase P RNAs were each split into two groups. The first group of each was included in the optimization of multibranch loop initiation parameters, while the second group (square brackets in Table 1) was withheld and only scored later. The similar accuracy of prediction for both groups of structures suggests that the multibranch loop initiation parameters are generalizable. In other words, a novel RNA, not included in the multibranch loop optimization, is expected to be predicted with roughly the same accuracy by the algorithm.

Table 1 displays the per cent of correctly predicted base-pairs for small and large subunit rRNAs when they are predicted either by known domains of less than 700 n (no parenthesis in Table 1) or by folding the entire sequence at once (parenthesis). The small and large subunit rRNA sequences are approximately 1600 and 2900 n, respectively, the longest sequences with known structures. The algorithm is significantly more predictive for known base-pairs when the sequences are broken into smaller fragments. Presumably, the folding problem is more difficult for long sequences because of the enormous increase in possible base-pairs as sequence lengthens.

The algorithm for secondary structure prediction cannot predict base-pairs in pseudoknots. A pseudoknot is formed when, given a base-pair between nucleotides i and j, another base-pair, between nucleotides m and n, exists such that $i < m < j < n$. The pseudoknot motif is encountered infrequently in the database of known secondary structures, but is known to be an important part of the structure of group I introns (Damberger & Gutell, 1994), RNase P RNA (Brown, 1998), and tmRNA which contains four pseudoknots (Williams & Bartel, 1996; Felden *et al*., 1997). Table 1 gives the per cent of pseudoknotted base-pairs (and therefore the per cent of base-pairs that cannot be predicted) for each type of RNA structure. Interestingly, over 10 % of base-pairs are pseudoknotted in the RNase P secondary structures where the predictive power of the algorithm is the least.

It is important to note that the inability to predict pseudoknots is not catastrophic to secondary structure prediction. A majority of base-pairs are correctly predicted in the lowest free energy structure for group I introns and RNase P, even though they contain 6.1 % and 12.5 % pseudoknotted base-pairs, respectively. Furthermore, the percentage of known base-pairs that occur at least once in the suboptimal structures seems to be unaffected by frequency of pseudoknots. This indicates that some suboptimal structures contain the alternative base-pairs involved in the pseudoknot. This may allow the identification of pseudoknots (Gaspin & Westhof, 1995).

The parameter set in which hydrogen bonds are maximized in canonical base-pairs gives, on average, only 20.5 % of known base-pairs correctly predicted. Furthermore, the single best suboptimal structure correctly predicts only 59.1 % of known base-pairs. Thus, accurate secondary structure prediction relies on the sequence dependence of the thermodynamic parameters.

Table 1. Accuracy of the RNA secondary structure prediction algorithm

				Current parameters[a]				Walter *et al.* (1994a)[a]		
				% correctly predicted base-pairs[b]			% energy[e] difference lowest and best	% correctly predicted base-pairs[b]		
RNA	Nucleotides	base-pairs	% pseudoknot	Lowest ΔG°	Best suboptimal	Any suboptimal		Lowest ΔG°	Best suboptimal	Any suboptimal
SSU (16 S) rRNA	33,263	8,863	1.4	66.2 ± 24.3	79.7 ± 18.9	92.6	5.4	53.5 ± 25.6	72.9 ± 20.2	91.1
				(51.1 ± 15.7)	(57.1 ± 16.6)	(77.8)	(2.1)	(42.4 ± 15.1)	(53.2 ± 15.2)	(74.2)
LSU (23 S) rRNA	13,341	3,585	0.2	70.3 ± 13.5	83.7 ± 8.4	96.3	5.2	63.5 ± 17.4	80.6 ± 12.0	95.6
				(56.7 ± 14.1)	(57.7 ± 14.6)	(77.0)	(0.7)	(50.5 ± 12.4)	(57.5 ± 9.1)	(76.9)
5 S rRNA	26,925	10,188	0.0	77.7 ± 23.1	94.6 ± 6.3	99.8	4.8	57.4 ± 27.4	94.8 ± 9.1	99.3
Group I intron-1[d]	5,518	1,532	6.0	69.9 ± 15.5	83.8 ± 10.8	98.0	7.1	67.7 ± 18.8	85.1 ± 7.8	97.4
Group I intron-2[d]	3,056	865	6.2	[60.1 ± 15.8]	[78.9 ± 11.4]	[98.0]	[9.0]	[59.0 ± 17.8]	[81.3 ± 12.6]	[98.2]
Group II intron	1,626	402	0.0	88.2 ± 9.2	92.4 ± 6.6	100.0	2.2	79.0 ± 17.3	92.4 ± 9.4	99.4
RNase P - 1[d]	2,269	694	14.4	54.5 ± 8.2	71.9 ± 7.3	94.4	6.0	50.0 ± 10.9	69.7 ± 10.8	91.0
RNase P - 2[d]	2,198	1,099	11.3	[68.1 ± 12.0]	[79.3 ± 7.0]	[96.3]	[5.3]	[51.1 ± 7.3]	[71.0 ± 9.9]	[92.1]
SRP RNA	24,383	6,273	1.9	73.0 ± 23.0	87.3 ± 12.7	96.7	3.6	57.0 ± 29.5	87.9 ± 4.2	97.3
tRNA	37,502	10,018	0.0	83.0 ± 22.2	95.7 ± 7.8	99.3	5.3	76.5 ± 24.8	94.1 ± 12.3	98.1
Total[c]	151,503	43,519	1.4	72.9 ± 10.4	86.1 ± 8.1	97.1 ± 2.7	4.8 ± 1.5	63.6 ± 11.5	84.7 ± 9.5	96.2 ± 3.5

[a] The accuracy of two sets of parameters are shown. The first set, current parameters, contains the parameters with expanded sequence dependence derived here. The second set contains the parameters reported by Walter *et al.* (1994a). A third set of parameters was used that maximizes hydrogen bonds in canonically base-paired regions only. For these parameters, 20.5 (± 6.5)% of known base-pairs were correctly predicted in the lowest free energy structure. The per cent of base-pairs in the best suboptimal and the per cent of known base-pairs that occur in at least one suboptimal structure are 59.1 (± 13.4) and 89.9 (± 6.1), respectively. The detailed results for the hydrogen bond parameter set are presented in the Supplementary Materials. Ten sets of structures determined by comparative sequence analysis were studied. The number of nucleotides, base-pairs, and the percentage of pseudoknotted base-pairs are shown for each. A complete list of the RNAs is in the Supplementary Materials.

[b] For each parameter set, the accuracy is determined for: the lowest free energy structure, the single best structure of 750 suboptimal structures generated, and the base-pairs correctly predicted in at least one suboptimal structure (any suboptimal). The accuracy is determined by counting correctly predicted base-pairs. Note that it is impossible to have 100% accuracy for structures with pseudoknots because pseudoknots are not allowed by the algorithm. It is also possible that some base-pairs determined by sequence comparison are important for a transition state rather than a ground state, which would also limit the calculated accuracy. Such a difference is suggested by comparisons between chemical modification and site directed mutagenesis results on a group I ribozyme (Jabri *et al.*, 1997). Each structure in an RNA class is scored and the average and standard deviation calculated. The 16 S and 23 S rRNAs are predicted in two ways. The first, shown without parenthesis, is structures predicted by dividing the sequence into domains of less than 700 n as described in Methods and listed in the Supplementary Material. The second, shown in parentheses, is the structure predicted for the entire sequence folded at once.

[c] For the total, each type of RNA is averaged and the standard deviation is determined. When the efn2 step is omitted, an average of 69.0 (± 7.7)% of known base-pairs are found in the lowest free energy structure when current parameters are used.

[d] Group I introns and RNase P RNAs are divided into two groups. The second group of each was not used during the optimization of the multibranch loop initiation parameters.

[e] For the current parameter set, the average % difference in free energy between the lowest free energy structure and the best suboptimal structure was calculated.

Faster calculation of minimal free energy structures

Faster multibranch loop searching and faster computers with more memory accelerate secondary structure prediction. Table 2 presents secondary structure prediction times as a function of RNA length, computer, and multibranch loop searching algorithm. The effect of the latter is greatest for long sequences. For example, the time required to fold the *Escherichia coli* small subunit rRNA was reduced from 26 to 13 minutes on a Pentium II 233 MHz personal computer. The computational details of this improvement are explained in Methods.

FMN cleavage constraint and improved enforcing of constraints

Recent studies demonstrated that flavin mononucleotide (FMN) photocleaves RNA specifically at U residues involved in G·U base-pairs (Burgstaller *et al.*, 1997; Burgstaller & Famulok, 1997). Such experimental data provide enormous constraints on the possible secondary structures for an RNA. To take advantage of this, the algorithm was modified to require specified U residues to be in G·U base-pairs.

Previous versions of the folding algorithm used bonus energies to enforce constraints that required either the formation of certain base-pairs or that designated nucleotides be double-stranded. The bonus energies were subtracted during structure generation so that a correct energy was reported for each structure. Nevertheless, the bonus energies distorted the dot plot representing the structural information (Zuker, 1989) and made it difficult to compute a representative sample of suboptimal structures. This deficiency is corrected by enforcing base-pairs through the prohibition of all possible alternatives. The details of these improvements are described in Methods.

Experimental constraints improve secondary structure prediction

To test the hypothesis that secondary structure prediction can be improved with experimental constraints from enzymatic and FMN cleavage, four test examples were drawn from the literature. These are FMN photocleavage of the T4 td intron (Burgstaller *et al.*, 1997; Jaeger *et al.*, 1993) and the enzymatic cleavage of the RNase P RNA from *Saccharomyces cerevisiae* (Tranguch *et al.*, 1994; Tranguch & Engelke, 1993), the 5 S RNA from *E. coli* (Speek & Lind; 1982; Szymanski *et al.*, 1998), and the second domain of the small subunit rRNA from *E. coli* (Kean & Draper, 1985; Gutell, 1994). Each RNA was predicted both with and without constraints. The accuracy of these predictions as compared to the known structures is shown in Table 3.

Each RNA studied with enzymatic cleavage was cleaved with enzymes that are specific to unpaired nucleotides and enzymes that are specific to paired nucleotides. For each case, a nucleotide was constrained to be single or double-stranded only if the same enzyme cleaved on both the 5′ and 3′ side of the nucleotide. For the small subunit rRNA, the intensity of each cleavage was quantified from one to four (Kean & Draper, 1985). Only cleavages of intensity two or higher were used to generate constraints for structure prediction.

Except for the *E. coli* small subunit rRNA, the accuracy of each RNA secondary structure prediction benefited significantly from the experimental constraints. The *E. coli* rRNA did not benefit from the enzymatic data because its structure was predicted with 93 % accuracy without constraints. Experimental constraints improved the predictions of the *E. coli* 5 S and T4 td intron from 26 % to 87 % and 56 % to 83 %, respectively. Figure 1 shows the predicted structures for the *E. coli* 5 S sequence both with and without experimental constraints, and Figure 2 illustrates the predictions for

Table 2. Time (minutes) for secondary structure prediction as a function of computer, RNA length, and multibranch loop searching algorithm

		RNA sequence length (nt)				
Computer	MBL search	77	268	433	631	1542
SGI	Fast	0.06	0.25	0.56	1.14	10.97
PII	Fast	0.02	0.33	0.75	1.65	12.89
PII	Slow	0.02	0.34	1.00	2.71	26.16
P90	Fast	0.05	1.29	3.01	6.72	55.00

Secondary structure prediction was timed on three platforms for five RNA sequence lengths. SGI is the World Wide Web *mfold* server, a Silicon Graphics computer with two 175 MHz IP30 Processors, 384 MB RAM, and IRIX OS 6.4. The *mfold* server executed the *mfold* code in FORTRAN using the default parameters for suboptimal structure generation. The *mfold* server was timed both for the prediction of structure by the dynamic programming algorithm and for steps involved in posting the information to the World Wide Web. The other two computers are Pentium-based personal computers using RNAstructure, written in C++, and generated suboptimal structures using the same parameters used to generate Table 1. Only the dynamic programming algorithm (Zuker, 1989) step of prediction was timed. PII is a 233 MHz Pentium II-based computer with 64 MB RAM running under Microsoft Windows 98. The predictions were timed on the Pentium II both with and without the multibranch loop searching improvement (described in Methods). This is indicated as fast or slow in the column labeled MBL search. P90 is a 90 MHz Pentium computer with 16 MB RAM and running under Microsoft Windows 95. The sequences for lengths 77, 268, 433, 631, and 1542 n are RR1664 tRNA (Sprinzl *et al.*, 1998), *Bacillus stearothermophilus* SRP RNA (Larsen *et al.*, 1998), IVS LSU group I intron from *Tetrahymena thermophila* (Damberger & Gutell, 1994), *S. cerevisiae* A5 group II intron (Michel *et al.*, 1989), and small subunit rRNA from *E. coli* (Gutell, 1994), respectively.

Table 3. Improving the accuracy of structure prediction with experimental constraints

						% prediction			
						without constraint		with constraint	
RNA	Cleavage	Ref	Nts	Base-pairs	% pseudoknot	Lowest ΔG°	Best	Lowest ΔG°	Best
E. coli 5 S rRNA	S_1 nuclease, RNase V_1	A	120	35	0	26.3	97.4	86.8	97.4
E. coli 16 S rRNA domain 2	RNase T_1, RNase T_2, RNase V_1	B	353	107	0	92.5	94.4	92.5	94.4
S. cerevisiae RNase P	RNase ONE, RNase V_1	C	369	94	5	61.7	86.2	74.5	84.0
T4 td group I intron	FMN	D	264	80	8	56.3	80	82.5	88.8

This Table summarizes the improvement in accuracy for four RNAs when experimental constraints were used. The column labeled Cleavage indicates the type of experimental constraint used. The references are: A, Speek & Lind (1982); B, Kean & Draper (1985); C, Tranguch *et al.* (1994); and D, Burgstaller *et al.* (1997). The number of nucleotides, number of base-pairs, and per cent of base-pairs in a pseudoknot are listed under Nts, Base-pairs, and % pseudoknot, respectively. The percentage of correctly predicted base-pairs for the lowest free energy structure and the best suboptimal structure both with and without experimental constraints are listed.

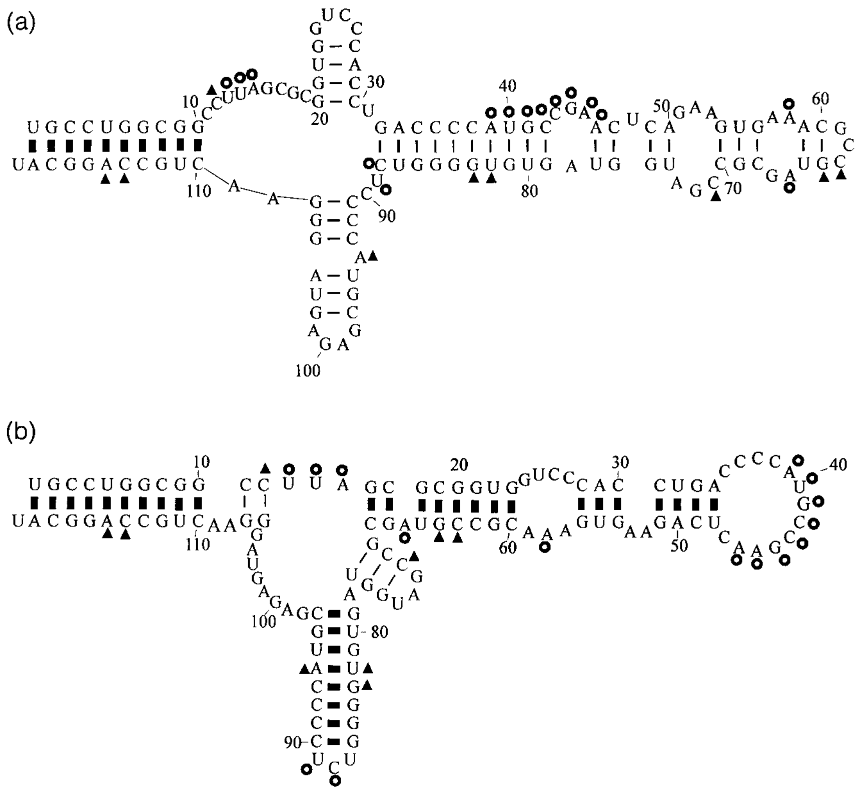

Figure 1. The predicted structures for *E. coli* 5 S rRNA without (a) and with (b) experimental constraints. Correctly predicted base-pairs are indicated with heavy lines. Circles indicate single-stranded nucleotides and triangles indicate double-stranded nucleotides as indicated by enzymatic cleavage (Speek & Lind, 1982). The constrained prediction has 86.8% of the known base-pairs (Table 3). Enzymatic cleavage indicates that nucleotide 71 is double-stranded. This is inconsistent with the structure determined by comparative sequence analysis (Szymanski *et al.*, 1998).

the T4 td group I intron with and without experimental constraints. The accuracy of the *S. cerevisiae* RNase P RNA prediction only increased from 62% to 75% for two reasons. Firstly, nucleotides in the pseudoknotted base-pairs were cleaved by RNase V_1, which is specific for double-stranded nucleotides. These constraints required those nucleotides to be double-stranded even though they could not pair correctly because the algorithm cannot predict pseudoknots. Also, two nucleotides were indicated as paired by RNase V_1 cleavage, although they are unpaired in the known structure. These constraints had to be fulfilled in the predicted structure.

Discussion

Free energy minimization is a useful tool for determining RNA secondary structure. It can help identify important regions during comparative analysis (Mathews *et al.*, 1997) and can be used when only a single sequence is available. It can facilitate interpretation of enzymatic and chemical modification data (Ehresmann *et al.*, 1987; Knapp *et al.*, 1989). The more accurate regions of predicted structure can be determined from an energy dot plot (Zuker & Jacobson, 1995) or statistical analysis (Huynen *et al.*, 1997) and this reliability information can be used to annotate predicted structures (Zuker & Jacobson, 1998).

Thermodynamic parameters for the prediction of free energy of folding are at the heart of algorithms for secondary structure prediction (Zuker, 1989; McCaskill, 1990; van Batenburg *et al.*, 1995; Gultyaev *et al.*, 1995). Parameters based on a nearest-neighbor model (Xia *et al.*, 1998) are well determined experimentally for Watson-Crick pairs, but helical regions only contain about 54% of the nucleotides in our database of known secondary structures. The remaining nucleotides are in unpaired regions, mostly loops. Recent studies have demonstrated that the stabilities of loops are highly sequence dependent (Schroeder *et al.*, 1996; Serra *et al.*, 1997; Wu *et al.*, 1995; Xia *et al.*, 1997). Currently, the stability of each possible sequence cannot be determined experimentally, although this someday may be possible with microfabricated arrays (Fodor *et al.*, 1991; O'Donnell-Maloney *et al.*, 1996; Fotin *et al.*, 1998). In the absence of a complete set of measured thermodynamic parameters, three methods were used to generate parameters to predict the stability of any possible sequence.

The first method for generating loop parameters is the extrapolation of parameters based on experimentally determined free energies of structure formation for representative molecules. For example, hairpin loop parameters were calculated in this way, based on experimental results from systematic studies of sequence dependence (Giese *et al.*, 1998; Serra *et al.*, 1993, 1994, 1997; Groebe & Uhlenbeck, 1988). A similar approach was used for internal loops.

The two other methods are knowledge based, using the database of structures determined by comparative sequence analysis. In one method, stability is based on the frequency of occurrence of a motif. This is used to determine enhanced stability of tetraloops and has the advantage of stabilizing loops that may occur frequently because of

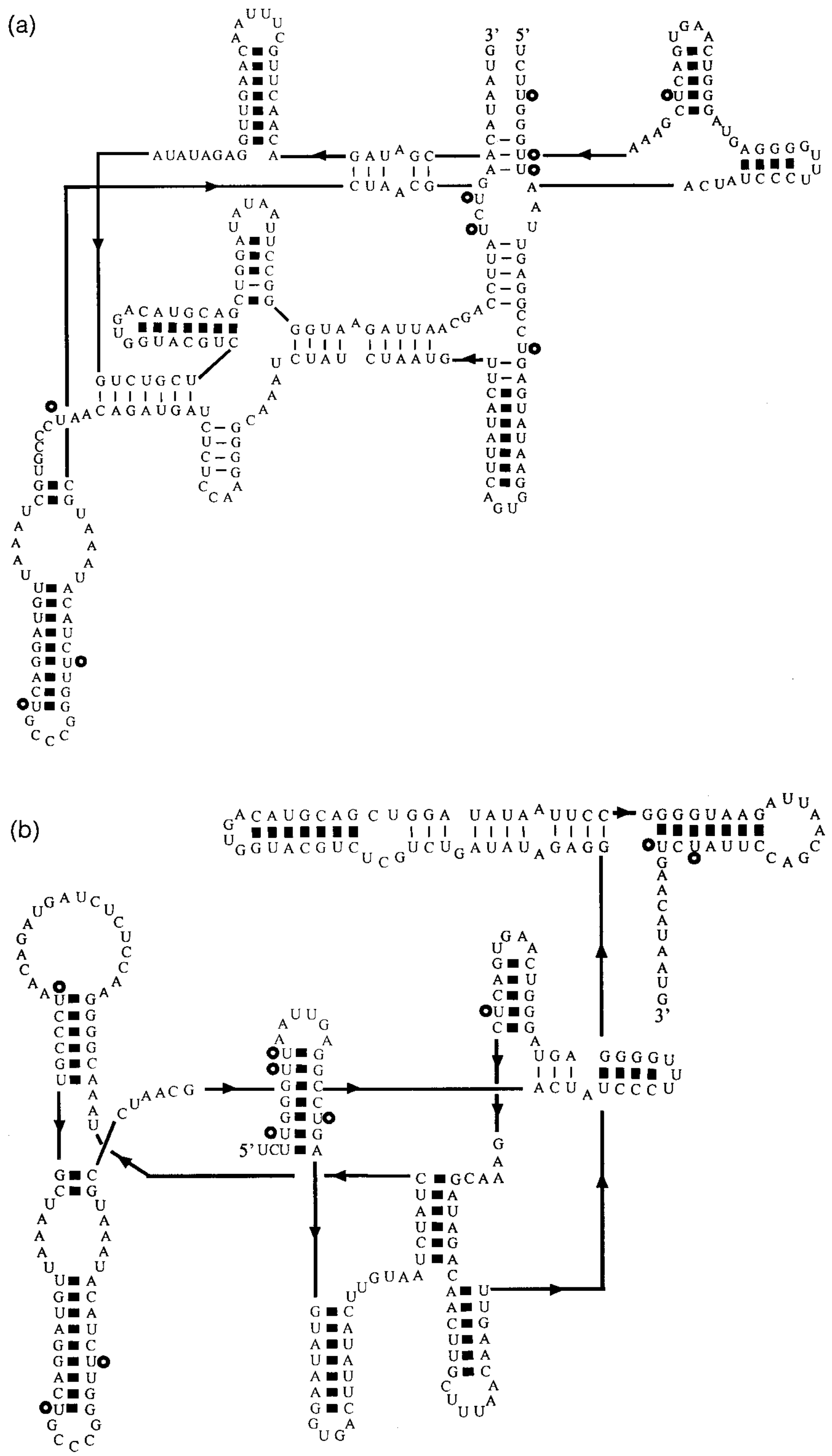

Figure 2. The predicted structures for the T4 td group I intron without (a) and with (b) experimental constraints. Correctly predicted base-pairs are indicated with heavy lines. Circles indicate U residues that are in G·U base-pairs as determined by FMN cleavage (Burgstaller *et al.*, 1997). The constrained prediction has 82.5% of the known base-pairs (Table 3).

stability imparted by tertiary contacts. When sequences are too short for tertiary contacts, these bonuses can be substituted with bonuses based on the measured or expected thermodynamic stabilities of isolated hairpins.

The other knowledge-based approach is to optimize parameters by maximizing the accuracy of the prediction algorithm. Multibranch loop initiation parameters were generated in this way. Some structures were withheld from the optimization to verify that the optimized parameters work as well on other sequences.

Comparison between the accuracy of the Walter *et al.* (1994a) parameters and the current parameters shows that the largest improvement is in the lowest free energy structure where the accuracy of predictions improved from 63.6% to 72.9%. A similar improvement is also found for sequences not included during the multibranch loop parameter optimization, suggesting the comparison is not biased because only the current parameters are optimized against the set of structures. In contrast to the improvement in lowest free energy structures, only a 1.4 % improvement is found in the most accurate suboptimal structure and only 0.9 % more known base-pairs, on average, are found in the set of 750 structures generated for each sequence (Table 1). This indicates that the parameters with expanded sequence dependence facilitate identification of the true structure from a set of possible structures. Results for four sequences with nuclease or chemical mapping data indicate that these constraints also facilitate finding the true structure (Table 3).

Prior studies of the accuracy of the folding algorithm examined only a limited number of structures (Jaeger *et al*., 1989; Walter *et al*., 1994a). The faster folding allowed by improvements in the algorithm and in computer technology, along with the availability of structures on the World Wide Web, allowed expansion of the database of test structures. The expanded database is important in two respects. Firstly, the knowledge-based parameters for tetraloops and multibranch loop initiation are more finely tuned with a large database. Secondly, a better representation of the accuracy of the algorithm is found with a large, diverse database.

The secondary structure prediction algorithm is available in three forms. The first is a C++ program, RNAstructure 3, for Windows 98, Windows 95, or Windows NT that is available at the Turner lab homepage on the World Wide Web (http://rna.chem.rochester.edu). The second is the *mfold* package, a collection of FORTRAN and C programs for predicting foldings and dot plots that is available for Unix platforms at ftp://snark.wustl.edu. This version is also available as an online web server (http://www.ibc.wustl.edu/ ~ zuker/rna/form1.cgi) linked to M. Zuker's homepage.

Methods

Method for predicting structures

Secondary structures are generated from sequences using the dynamic programming algorithm by Zuker (1989). For each sequence, a maximum of 750 suboptimal structures with up to 20 % difference in free energy from the lowest free energy are generated. The window size is set to 0 to allow predictions of structures with subtle variations. A second algorithm, efn2, is used to recalculate the free energy of each suboptimal structure with more complete energy rules for multibranch loops. These rules include coaxial stacking and a logarithmic dependence of initiation on the number of unpaired nucleotides. The efn2-calculated free energy is then used to reorder the structures by stability. The lowest free energy structure after efn2 rearrangement is used for scoring accuracy of the lowest free energy structure (Table 1).

This version of the dynamic programming algorithm filters out isolated base-pairs by assigning them a large free energy penalty (1600 kcal/mol). A base-pair between nucleotides i and j is considered isolated if pairing is not possible both between nucleotides $i + 1$ and $j - 1$ and nucleotides $i - 1$ and $j + 1$.

Database of secondary structures

The algorithm is tested with sequences of known secondary structure (Gutell, 1994; Gutell *et al*., 1993; Schnare *et al*., 1996; Szymanski *et al*., 1998; Sprinzl *et al*., 1998; Larsen *et al*., 1998; Brown, 1998; Damberger & Gutell, 1994; Michel *et al*., 1989; Waring & Davies, 1984). The Supplementary Materials list the 955 specific structures, their sources, and the domains folded.

Small (16 S) and large (23 S) subunit rRNAs are divided into domains of less than 700 n. The small subunit rRNA domains (specified in the Supplementary Material) are based on those chosen by Jaeger *et al*. (1989). The large subunit rRNAs are generally divided into six domains. For example, the *E. coli* 23 S rRNA (Gutell *et al*., 1993) is divided into domains at nucleotides: 14-526, 578-1262, 1275-1646, 1647-2010, 2022-2626, and 2629-2890. The domains for all sequences are specified in the Supplementary Material.

For tRNAs, modified nucleotides that are unable to fit in *A*-form helices are forced to be single-stranded by the algorithm. The modified nucleotides not allowed to pair are: N6-(*cis*-hydroxyisopentenyl)adenosine, lysidine, 1-methylguanosine, N2,N2,2′-*O*-trimethylguanosine, archaeosine, mannosyl-queuosine, galactosyl-queuosine, wybutosine, peroxywybutosine, 3-(3-amino-3-carboxypropyl)uridine, dihydrouridine, 5, 2′-*O*-dimethyluridine, 2-methyladenosine, 2-methylthio-N6-threonylcarbamoyladenosine, inosine, 1-methylinosine, and 2′-*O*-ribosyladenosine.

Scoring of secondary structure prediction and counting of pseudoknotted base-pairs

The predicted secondary structures are scored by comparison to base-pairs determined by comparative sequence analysis. Known base-pairs are considered correctly predicted if they occur in the predicted structure or if the predicted structure contains a base-pair shifted by at most one nucleotide on one side of the pair. For example, a known base-pair between nucleotides i and j is considered correctly predicted by base-pairs of i to j, i to $j - 1$, i to $j + 1$, $i - 1$ to j, or $i + 1$ to j. A predicted pair of $i + 1$ to $j - 1$, however, is not considered correct. The slipped helixes are rare and are considered correct because it is difficult to determine the accurate pairing scheme by comparative sequence analysis or to predict it by free energy minimization.

Base-pairs in pseudoknots in Table 1 are counted by a computer program that first identifies all occurrences of base-pairs $i - j$ and $k - l$ such that $i < k < j < l$. Then, the program determines the least number of base-pairs that can be broken to resolve the pseudoknots. This, least number of base-pairs, is used for the calculation of per cent of pseudoknotted base-pairs given in Table 1.

Derivation and implementation of thermodynamic parameters

Watson-Crick pairs

Stabilities of helical regions with Watson-Crick pairs are assigned with the INN-HB nearest-neighbor model by Xia *et al.* (1998). This model expands on prior nearest-neighbor models (Freier *et al.*, 1986b; Borer *et al.*, 1974) by including a penalty term for each terminal A·U or U·A pair in a helix. This accounts for the dependence of numbers of hydrogen bonds on base composition (Xia *et al.*, 1998).

A modification of the secondary structure prediction algorithm was necessary to implement the terminal A·U penalty term. The penalty is added to the free energy along with the free energy of the loop in which the helix terminates. For internal loops and hairpin loops, the penalty is included in the terminal mismatch stacking data tables included with the program. For bulge loops longer than one nucleotide, the algorithm explicitly checks each closing pair and applies the terminal A·U penalty if necessary. For the free energy calculation for multibranch loops and exterior loops, a modification in both the fill and traceback steps of the dynamic programming algorithm (Zuker & Stiegler, 1981; Zuker, 1989) was necessary. The penalty term, if required, is added to the penalty, c_1 or c_2, for each helix that radiates from the loop. In the efn2 algorithm, the penalty is applied even when the helixes are coaxially stacked.

G·U base-pair nearest neighbors

Nearest-neighbor free energy parameters for the stacking of G·U pairs in helical RNA, Table 4, were derived by linear regression from experimentally determined stabilities (He *et al.*, 1991; Sugimoto *et al.*, 1986; Xia *et al.*, 1997; Freier *et al.*, 1986a; Wu *et al.*, 1995; McDowell & Turner, 1996; R. Kierzek, unpublished data; S. J. Schroeder, unpublished data). The motif of

5′ GU 3′
3′ UG 5′

does not fit the nearest-neighbor model, unless duplexes containing the motif

5′ GGUC 3′
3′ CUGG 5′

are omitted (He *et al.*, 1991). Therefore, the linear regression is calculated without sequences containing the

5′ GGUC 3′
3′ CUGG 5′

motif. The value of the entire motif of

5′ GGUC 3′
3′ CUGG 5′

is found by averaging the stabilities for the four duplexes that include it.

The method and data for determining the G·U nearest-neighbors were derived largely from He *et al.* (1991). Consider the self-complementary duplex, (CGU<u>UG</u>ACG)$_2$, where the underlined nucleotides form G·U pairs. The stability component, ΔG°_{37}(component), contributed by the addition of the tandem G·U pairs is calculated by taking the difference in stabilities between duplexes with and without the G·U pair and adding back the stability of the stack broken by addition of the G·U pairs:

$$\begin{aligned}\Delta G^\circ_{37}(\text{component}) &= \Delta G^\circ_{37}(\text{CGU}\underline{\text{UG}}\text{ACG}) \\ &\quad - \Delta G^\circ_{37}(\text{CGUACG}) \\ &\quad + \Delta G^\circ_{37}\begin{pmatrix}5'\ \text{UA}\ 3' \\ 3'\ \text{AU}\ 5'\end{pmatrix} \qquad (1)\end{aligned}$$

where ΔG°_{37}(CGU<u>UG</u>ACG) is the free energy of duplex formation determined by optical melting and ΔG°_{37}(CGUACG) is the free energy of duplex formation predicted by the INN-HB nearest-neighbor parameters of Xia *et al.* (1998). Values of ΔG°_{37}(component) are equal to the sum of the free energy increments for the component nearest neighbors. For CGU<u>UG</u>ACG, three nearest neighbors are involved:

$$\begin{aligned}\Delta G^\circ_{37}(\text{component}) &= 2\Delta G^\circ_{37}\begin{pmatrix}5'\ \text{UU}\ 3' \\ 3'\ \text{AG}\ 5'\end{pmatrix} \\ &\quad + \Delta G^\circ_{37}\begin{pmatrix}5'\ \text{UG}\ 3' \\ 3'\ \text{GU}\ 5'\end{pmatrix} \qquad (2)\end{aligned}$$

Note that the neighbor

5′ UU 3′
3′ AG 5′

appears twice in the sequence, closing both sides of the tandem G·U mismatch. Terminal G·U pairs are treated like terminal A·U pairs in the INN-HB model, i.e. they are penalized 0.45 kcal/mol because they have two hydrogen bonds. An analogous approach was used to calculate nearest-neighbor parameters for enthalpy and entropy changes.

Table 5 lists the data used to fit the G·U nearest-neighbor parameters and the free energies predicted by the model. The coefficient of determination, R^2, for the regression is 0.85, indicating that about 85 % of the observed variability in stability is explained by the model. For predicting secondary structures,

5′ GU 3′
3′ UG 5′

is assigned the unfavorable value of 1.45 kcal/mol (Table 4). However, when the stack

5′ GGUC 3′
3′ CUGG 5′

occurs, it is effectively assigned a favorable free energy because the entire motif is assigned a favorable ΔG°_{37} of -4.1 kcal/mol in the separate table of 2 × 2 internal loops.

The nearest neighbor,

5′ GG 3′
3′ UU 5′,

is in only one measured duplex and therefore has a large standard error from the fit. The positive (unfavorable) ΔG°_{37} listed in Table 4 decreased the accuracy of the structure prediction algorithm compared to a favorable value. Therefore, for secondary structure prediction, the value of

5′ GG 3′
3′ UU 5′

is set to a favorable -0.5 kcal/mol, a reasonable estimate considering the size of the error.

Dangling ends and terminal mismatches

The parameters for dangling ends and mismatches are not affected by the change in nearest-neighbor model. The free energy parameters for unpaired nucleotides adjacent to Watson-Crick pairs are taken from a prior compilation (Serra & Turner, 1995). Dangling ends on terminal G·U pairs are treated like dangling ends on terminal A·U pairs with the A replacing the G.

Several stabilities are known for terminal mismatches adjacent to G·U pairs (Giese *et al*., 1998). In one case, the

5′ UA 3′
3′ GA 5′

Table 4. Nearest neighbor parameters for G·U base-pairs

Nearest-neighbor	ΔG°_{37} (kcal/mol)	Error	ΔH° (kcal/mol)	Error	ΔS° (eu)[c]	Error[c]
5′ AG 3′ 3′ UU 5′	−0.55	0.32	−3.21	2.76	−8.6	8.45
5′ AU 3′ 3′ UG 5′	−1.36	0.24	−8.81	2.10	−24.0	6.44
5′ CG 3′ 3′ GU 5′	−1.41	0.24	−5.61	2.13	−13.5	6.53
5′ CU 3′ 3′ GG 5′	−2.11	0.25	−12.11	2.22	−32.2	6.81
5′ GG 3′ 3′ CU 5′	−1.53	0.27	−8.33	2.33	−21.9	7.14
5′ GU 3′ 3′ CG 5′	−2.51	0.25	−12.59	2.18	−32.5	6.67
5′ GA 3′ 3′ UU 5′	−1.27	0.28	−12.83	2.44	−37.3	7.47
5′ GG 3′ 3′ UU 5′	+0.47 (−0.5)[b]	0.96	−13.47	8.37	−44.9	25.65
5′ GU 3′ 3′ UG 5′ [a]	+1.29	0.56	−14.59	4.92	−51.2	15.08
5′ GGUC 3′ 3′ CUGG 5′ [a]	−4.12	0.54	−30.80	8.87	−86.0	23.70
5′ UG 3′ 3′ AU 5′	−1.00	0.30	−6.99	2.64	−19.3	8.09
5′ UG 3′ 3′ GU 5′	+0.30	0.48	−9.26	4.19	−30.8	12.86
Each Terminal G·U[d]	+0.45	-	+3.72	-	+10.5	-

[a] The nearest neighbor, $\begin{array}{c}5'\text{ GU }3'\\3'\text{ UG }5'\end{array}$, is split into two environments. The first is a nearest − neighbor for the entire fragment $\begin{array}{c}5'\text{ GGUC }3'\\3'\text{ CUGG }5'\end{array}$ and the second is the nearest neighbor for $\begin{array}{c}5'\text{ GU }3'\\3'\text{ UG }5'\end{array}$ with all other closing pairs.

[b] For secondary structure prediction, $\begin{array}{c}5'\text{ GG }3'\\3'\text{ UU }5'\end{array}$ issetto − 0.5 kcal/mol.

[c] Values of ΔS° are calculated from $\Delta S^\circ = (\Delta H^\circ - \Delta G^\circ_{37})/(310.15\text{ K})$. Errors in ΔS° are from the linear regression.

[d] A terminal G·U pair is a G·U at the end of a helix, including G·U pairs adjacent to hairpins, internal loops, junctions, and bulges with more than one nucleotide. Terminal G·U parameters are assumed to be the same as the terminal A·U parameters derived by Xia *et al*. (1998).

Table 5. The compilation of helices containing G·U base-pairs

Sequence	Ref.[a]	ΔG°_{37}	ΔG°_{37}(component) (kcal/mol)	Predicted ΔG°_{37} (component) (kcal/mol)
A$\underline{G}$GC$\underline{U}$U	A	−4.1	−6.1	−5.3
A$\underline{G}$UCGA$\underline{U}$U	B	−6.0	−3.5	−3.8
A$\underline{U}$CUAG$\underline{G}$U	C	−5.9	−5.9	−5.8
A$\underline{U}$GCGC$\underline{G}$Up	A	−9.3	−5.5	−5.5
AU$\underline{G}$CG$\underline{U}$AUp	A	−5.3	−6.1	−7.0
A$\underline{U}$GC$\underline{G}$Up	A	−4.2	−6.3	−5.5
AUG$\underline{GU}$CAU	B	−5.4	−4.1	−4.1
AUG$\underline{UG}$CAU	A	−6.2	−5.1	−4.7
CCA$\underline{GU}$UGG	D	−5.7	0.5	0.2
CCA$\underline{UG}$UGG	D	−7.8	−1.6	−2.4
CCU$\underline{GU}$AGG	B	−6.8	−0.6	−0.7
CG$\underline{G}$AU$\underline{U}$CG	B	−6.6	−5.3	−5.6
C$\underline{G}$GC$\underline{U}$G	A	−5.6	−6.6	−7.0
CGG$\underline{GU}$CCG	B	−11.2	−4.7	−4.1
CGU$\underline{UG}$ACG	B	−6.9	−2.2	−2.2
C$\underline{U}$GC$\underline{G}$G	A	−4.3	−5.4	−7.0
CUG$\underline{GU}$CAG	B	−7.1	−3.4	−4.1
GAG$\underline{GU}$CUC	E	−8.8	−4.2	−4.1
GAG$\underline{GU}$GAG/	F	−7.6	−2.9	−1.6
GAG$\underline{U}$GAG/	C	−8.0	−3.0	−3.9
GAG$\underline{UG}$CUC	G	−9.4	−5.1	−4.7
GAG$\underline{UG}$GAG/	F	−9.7	−4.9	−4.3
GAG$\underline{UU}$GAG/	B	−8.2	−3.4	−3.4
$\underline{G}$AUGCAU$\underline{U}$p	A	−6.8	−3.3	−2.5
$\underline{G}$CCGG$\underline{U}$p	H	−9.2	−5.7	−5.0
GC$\underline{UG}$GC	A	−6.5	−4.2	−3.9
GGA$\underline{GU}$UCC	B	−6.4	0.3	0.2
GGA$\underline{UG}$UCC	B	−8.4	−1.7	−2.4
$\underline{G}$GCGC$\underline{U}$	H	−8.4	−4.6	−4.2
G$\underline{G}$CG$\underline{U}$C	A	−4.7	−6.8	−8.1
GGC$\underline{GU}$GCC	B	−9.7	−0.9	−1.5
GGC$\underline{UG}$GCC	G	−13.1	−4.3	−3.9
GGU$\underline{UG}$ACC	G	−8.3	−1.8	−2.2
GUC$\underline{GU}$GAC	B	−6.1	−1.4	−1.5
$\underline{G}$UCUAGA$\underline{U}$	C	−7.7	−3.8	−2.7
$\underline{U}$AUGCAU$\underline{G}$p	A	−6.4	−2.9	−2.0
$\underline{U}$CCGG$\underline{G}$p	H	−7.4	−3.9	−3.1
UC$\underline{G}$CCAGAGG/	I	−15.3	−5.6	−3.9
$\underline{U}$GGCC$\underline{G}$p	H	−8.6	−4.0	−2.8

Nucleotides in G·U base-pairs are underlined. For non-selfcomplementary sequences, only one strand is shown (/).

[a] The measured stabilities of the helices are from: A, Sugimoto *et al*. (1986); B, He *et al*. (1991); C, R.K., unpublished data; D, McDowell *et al*. (1997); E, Wu *et al*. (1995); F, Xia *et al*. (1997); G, McDowell & Turner (1996); H, Freier *et al*. (1986a); and I, S. J. Schroeder (unpublished data).

mismatch, the stability of the reference helix lacking the terminal mismatch was not measured. Therefore, the reference helix is calculated using the Watson-Crick parameters by Xia *et al*. (1998) and the G·U parameters from Table 4. The enhanced stability due to the dangling end is then calculated as:

$$\Delta G^\circ_{37}\begin{pmatrix}5'\ \mathrm{UA}3'\\ 3'\ \mathrm{GA}5'\end{pmatrix} = \frac{\Delta G^\circ_{37}\begin{pmatrix}5'\mathrm{AGCGUA}\\ 3'\mathrm{AUGCGA}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}5'\mathrm{GCGU}\\ 3'\mathrm{UGCG}\end{pmatrix}}{2} = \frac{-4.0+2.0}{2} = -1.0\ \mathrm{kcal/mol} \qquad (3)$$

where the free energy of the duplex with terminal mismatch, −4.0 kcal/mol, is from Giese *et al*. (1998). For those terminal mismatches adjacent to G·U pairs that were not measured, the stability is approximated as that of the mismatch adjacent to an A·U pair such that the A replaces the G.

Hairpin loops

Parameters for predicting the stabilities of hairpin loops are derived from experimental data for stem loop stability (Giese *et al*., 1998; Serra *et al*., 1993, 1994, 1997; Groebe & Uhlenbeck, 1988) by subtracting stabilities of stems calculated with the INN-HB nearest-neighbor model (Xia *et al*., 1998). The stabilities of hairpin loops longer than three unpaired nucleotides are approximated based on loop length and the sequences of the closing base-pair and first mismatch:

$$\begin{aligned}\Delta G^{\circ}_{37\text{loop}}(n > 3) = \Delta G^{\circ}_{37\text{ initiation}}(n) + \Delta G^{\circ}_{37}(\text{stacking of the first mismatch})\\ + \Delta G^{\circ}_{37\text{bonus}}(\text{UU or GA first mismatch, but not AG})\\ + \Delta G^{\circ}_{37\text{bonus}}(\text{special GU closure})\\ + \Delta G^{\circ}_{37\text{penalty}}(\text{oligo-C loops})\end{aligned} \tag{4}$$

where n is the number of nucleotides in the loop. Values of $\Delta G^{\circ}_{37\text{ initiation}}(n)$ are given in Table 6. Stacking of the first mismatch and the free energy bonuses are not included for loops smaller than four nucleotides because it is assumed that these loops are too constrained to allow the same stacking possible at the end of a duplex. The sequence independence of the stability of loops of three supports this assumption (Serra *et al.*, 1997). The stacking of the first mismatch is given the same free energy as a terminal mismatch parameter. A bonus is applied to loops with UU and GA (G on the 5′ side and A on 3′ side of loop) first mismatches. The bonus for G·U closed loops applies only to hairpins closed with a G on the 5′ side that is preceded by two G residues in base-pairs. The penalty for oligo-C loops is applied to hairpin loops in which all unpaired nucleotides are C and is a linear function of n for n larger than three:

$$\Delta G^{\circ}_{37\text{penalty}}(\text{oligo-C loops, } n > 3) = An + B \tag{5}$$

For loops of three C residues, the $\Delta G^{\circ}_{37\text{penalty}}$ is 1.4 kcal/mol. These bonuses and penalties are based on thermodynamic measurements (Giese *et al.*, 1998; Serra *et al.*, 1993, 1994, 1997; Groebe & Uhlenbeck, 1988). Table 6 summarizes the sequence-dependent terms used in the approximation for hairpin loop stability.

To calculate the various terms in equation (4) and Table 6, $\Delta G^{\circ}_{37\text{hairpin}}$, the sum of free energies of hairpin initiation and any stability bonuses, is defined as the $\Delta G^{\circ}_{37\text{loop}}$ with stabilities of the first mismatch subtracted. Consider the hairpin sequence rAGGA$\underline{\text{AUAAUA}}$UCCU (nucleotides in the loop are underlined) with a stability of -2.19 kcal/mol measured by optical melting (Serra *et al.*, 1993). The $\Delta G^{\circ}_{37\text{hairpin}}$ is determined by the following equation:

$$\begin{aligned}\Delta G^{\circ}_{37\text{hairpin}} &= \Delta G^{\circ}_{37\text{ stem-loop}} - \Delta G^{\circ}_{37\text{ stem}} - \Delta G^{\circ}_{37\text{mismatch}}\\ &= -2.19\text{ kcal/mol} + 6.79\text{ kcal/mol}\\ &\quad + 0.8\text{ kcal/mol}\\ &= 5.4\text{ kcal/mol}\end{aligned} \tag{6}$$

where $\Delta G^{\circ}_{37\text{ stem}}$ is estimated by the INN-HB parameters (Xia *et al.*, 1998) without an initiation term, and the $\Delta G^{\circ}_{37\text{mismatch}}$ is the value of a terminal mismatch (as defined above). Table 7 shows the complete database of stem-loop structures studied by optical melting and the $\Delta G^{\circ}_{37\text{hairpin}}$ calculated for each.

The six nucleotide hairpin loops are the most studied and are used to determine the $\Delta G^{\circ}_{37\text{bonus}}$ for UU and GA first mismatches. This is done by calculating the average $\Delta G^{\circ}_{37\text{hairpin}}$ for all hairpins without UU or GA first mismatches or 5′ G·U3′ closure preceded by two G residues.

Table 6. Hairpin loop free energy parameters

Parameter	Condition	ΔG° (kcal/mol)	Standard deviation (kcal/mol)
$\Delta G^{\circ}_{37\text{initiation}}(n)$	$n = 3$	5.7	0.47
	4	5.6	0.61
	5	5.6	0.28
	6	5.4	0.24
	7	5.9	0.23
	8	5.6	0.99
	9	6.4	0.85
$\Delta G^{\circ}_{37\text{bonus}}$	(UU or GA first mismatch)[a]	−0.8	0.22
$\Delta G^{\circ}_{37\text{bonus}}$	(special GU closure)	−2.2	0.53
$\Delta G^{\circ}_{37\text{penalty}}$(oligo-C loops).	3 nt long	1.4	-
$\Delta G^{\circ}_{37\text{penalty}}$(oligo-C loops)	A	0.3	0.052
$\Delta G^{\circ}_{37\text{penalty}}$(oligo-C loops)	B	1.6	0.34

Initiations are given as a function of length. Hairpins with less than three nucleotides are prohibited. Hairpin initiation for loops longer than nine is approximated by:

$$\Delta G^{\circ}_{37\text{initiation}}(n > 9) = \Delta G^{\circ}_{37\text{initiation}}(9) + 1.75RT\ln(n/9)$$

The free energy of hairpin loops is approximated by the equation:

$$\Delta G^{\circ}_{37\text{loop}}(n > 3) = \Delta G^{\circ}_{37\text{ initiation}}(n) + \Delta G^{\circ}_{37}(\text{stacking of the first mismatch}) + \Delta G^{\circ}_{37\text{bonus}}(\text{UU or GA first mismatch})$$
$$+ \Delta G^{\circ}_{37\text{bonus}}(\text{special GU closure}) + \Delta G^{\circ}_{37\text{penalty}}(\text{poly-C loops})$$

where the $\Delta G^{\circ}_{37\text{penalty}}$(poly-C loops) is 1.4 kcal/mol for a loop of three, or is determined by:

$$\Delta G^{\circ}_{37\text{penalty}}(\text{poly-C loops, } n > 3) = An + B$$

A special G·U closure is applied only to hairpins with a 5′ closing G that is preceded by two G residues. The bonus for GA first mismatch is applied only to hairpin loops in which the G is at the 5′ end of the hairpin loop.

[a] The $\Delta G^{\circ}_{37\text{bonus}}$ term is not applied to hairpins with AG first mismatches.

Table 7. The database of hairpin stabilities

Hairpin sequence	Ref.	ΔG°_{37} of stem loop structure (kcal/mol)	$\Delta G^\circ_{37\text{hairpin}}$: measured (kcal/mol)	$\Delta G^\circ_{37\text{hairpin}}$ predicted (kcal/mol)
GGAAUAUCC	A	0.6	5.8	5.7
GGAGAAAUUCC	B	−0.9	4.8	5.7
GGCAUAGCC	A	−0.9	5.7	5.7
GGGAAAUCC	B	0.3	5.1	3.5
GGGAUACCC	A	−0.3	6.3	5.7
GGGAUACAAAGUAUCCA	C	−6.3	5.4	5.7
GGGAUACCCCGUAUCCA	C	−4.6	7.1	7.1
GGGAUACUUUGUAUCCA	C	−7.0	5.4	5.7
GGUAUAACC	A	0.6	5.7	5.7
GGUAUAGCC	B	0.9	6.2	5.7
GCGAUUAUGC	B	−0.3	4.3	5.6
GCGGAUUAUCGC	B	−1.4	5.7	5.6
GGAAUUAUCC	A	−0.1	5.9	5.6
GGCAUUAGCC	A	−2.5	5.7	5.6
GGGACCAUCC	B	−1.9	2.7	3.4
GGGAUUACCC	A	−1.3	6.3	5.6
GGGAUACAAAAGUAUCCA	C	−7.6	5.6	5.6
GGGAUACCCCCGUAUCCA	C	−4.3	8.5	8.4
GGGAUACUUUUGUAUCCA	C	−8.2	4.7	4.9
GGUAUUAACC	A	−0.3	5.7	5.6
GGUAUUAGCC	B	0.2	6.5	5.6
GGUGCAAGCC	B	−1.4	4.4	4.9
GCGGAAGAUGC	B	−0.3	4.6	5.0
GGAAUUUAUCC	A	−0.2	5.7	5.7
GGCAUAUAGCC	A	−2.7	5.5	5.7
GGCAUUUAGCC	A	−2.6	5.6	5.7
GGGAUAUACCC	A	−1.5	6.1	5.7
GGGAUUUACCC	A	−1.7	6.0	5.7
GGGAUUUAUCC	B	−0.6	4.0	3.5
GGGAUACAAAAAGUAUCCA	C	−7.7	5.5	5.7
GGGAUACCCCCCGUAUCCA	C	−4.1	8.7	9.3
GGGAUACUUUUUGUAUCCA	C	−8.1	4.8	5.0
GGUAUUUAACC	A	−0.3	5.7	5.7
GGUAUUUAGCC	B	−0.6	5.7	5.7
ACCGACACAGGU	E	−1.6	4.9	4.7
AGGAAUAAUAUCCU	D	−2.2	5.4	5.4
AGGUAUAAUAGCCU	D	−2.2	5.7	5.4
CGGUUAAUUCCG	E	−1.9	4.4	4.7
CUCUACACCAAGGAG	E	−1.8	5.2	5.4
GCGGUGAAAUGC	B	−0.4	4.5	4.6
GCGUUAAUUUGC	B	−0.3	4.8	4.6
GGAAUAAUAUCC	D	−0.7	5.2	5.4
GGAGUAAUAUCC	E	−1.7	4.2	4.7
GGCAUAAUAGCC	D	−2.7	5.4	5.4
GGCAUAAUAGCC	E	−2.7	5.4	5.4
GGCAUAAUCGCC	E	−2.6	5.6	5.4
GGCAUAAUGGCC	E	−3.1	5.0	5.4
GGCCUAAUAGCC	E	−2.2	5.5	5.4
GGCCUAAUCGCC	E	−2.3	5.5	5.4
GGCCUAAUUGCC	E	−1.9	5.6	5.4
GGCGUAAUAGCC	D	−3.4	4.7	4.7
GGCGUAAUGGCC	E	−3.1	5.2	5.4
GGCUUAAUCGCC	E	−3.0	5.1	5.4
GGCUUAAUUGCC	E	−3.2	4.7	4.7
GGGAUAAUAUCC	B	−1.4	3.2	2.5
GGUAUAAUAACC	D	−0.3	5.7	5.4
GGUAUAAUAGCC	D	−0.5	5.8	5.4
GGUGUAAUAACC	E	−1.5	4.6	4.7
GGUGUAAUAGCC	E	−1.9	3.9	4.7
GGUGUAAUGACC	E	−1.0	5.2	5.4
GGUGUAAUGGCC	B	−1.3	4.8	5.4
GUGGUAAUACAC	E	−1.1	4.9	4.7
GUGGUAAUAUAC	E	1.1	4.4	4.6
GCGAAUAAAUAUCGC	A	−2.5	5.9	5.9
GGCAUAAAUAGCC	A	−2.2	6.0	5.9
GGGACGGACAUCC	B	−1.0	3.6	3.7
GGGAUAAAUACCC	A	−2.1	5.5	5.9
GGGAUAAAUAUCC	B	−0.3	4.3	3.7
GGGAUACAAAAAAAGUAUCCA	C	−7.2	6.0	5.9
GGGAUACCCCCCCCGUAUCCA	C	−2.9	9.9	9.6
GGGAUACUUUUUUUGUAUCCA	C	−7.4	5.5	5.2
GGUAUAAAUAACC	A	−0.4	5.6	5.9

GGU$\underline{\text{AUAAAUA}}$GCC	B	0.2	6.5	5.9
GGU$\underline{\text{GUAAAAA}}$GCC	B	−0.6	5.2	5.2
GCG$\underline{\text{AAUUCAUA}}$UGC	B	−0.6	4.0	5.6
GCU$\underline{\text{GAAUGGAA}}$GGC	B	−1.3	4.2	4.9
GGA$\underline{\text{AUAAAAUA}}$UCC	A	−2.2	3.7	5.6
GGC$\underline{\text{AUAAAAUA}}$CCG	A	−2.1	6.1	5.6
GGG$\underline{\text{AUAAAAUA}}$CCC	A	−2.3	5.3	5.6
GGG$\underline{\text{AUAAAAUA}}$UCC	B	−1.7	2.9	3.4
GGU$\underline{\text{AAUUCAUA}}$GCC	B	0.0	6.3	5.6
GGU$\underline{\text{AUAAAAUA}}$ACC	A	0.5	6.5	5.6
GGU$\underline{\text{AUAAAAUA}}$GCC	B	−0.1	6.2	5.6
GCGA$\underline{\text{AUAAAAAUA}}$UCGC	A	−2.2	6.2	6.4
GCGU$\underline{\text{AUAAAAAUA}}$ACGA	A	−1.0	7.5	6.4
GGA$\underline{\text{AUAAAAAUA}}$UCC	A	0.2	6.1	6.4
GGC$\underline{\text{AUAAAAAUA}}$GCC	A	−2.1	6.1	6.4
GGG$\underline{\text{AUAAAAAUA}}$CCC	A	−2.1	5.5	6.4
GGGAUAC$\underline{\text{AAAAAAAAA}}$GUAUCCA	C	−5.6	7.6	6.4
GGGAUAC$\underline{\text{CCCCCCCCC}}$GUAUCCA	C	−2.1	10.7	10.7
GGGAUAC$\underline{\text{UUUUUUUUU}}$GUAUCCA	C	−8.0	4.9	5.7

The hairpin loops were taken from references: A, Serra *et al*. (1997); B, Giese, *et al*. (1998); C, Groebe & Uhlenbeck (1988); D, Serra *et al*. (1993); and E, Serra *et al*. (1994). Loop nucleotides are underlined.

One sequence, GGU$\underline{\text{GUAAUG}}$GCC, is also excluded because it is 0.8 kcal/mol more stable than the average. This average is used as the $\Delta G^\circ_{37\text{initiation}}$ for hairpins of six because the $\Delta G^\circ_{37\text{hairpin}}$ equals the $\Delta G^\circ_{37\text{initiation}}$ for loops without bonuses. The average of $\Delta G^\circ_{37\text{hairpin}}$ for hairpins with UU and GA first mismatches is then calculated with the exception that the sequence, GGU$\underline{\text{GUAAUA}}$GCC, is excluded because it is 1 kcal/mol more stable than the average of other six nucleotide hairpin loops with a GA or UU mismatch. The difference between the two $\Delta G^\circ_{37\text{loop}}$ values is the value of the $\Delta G^\circ_{37\text{bonus}}$ for GA and UU first mismatches:

$$\begin{aligned}\Delta G^\circ_{37\text{bonus}}(\text{UU or GA first mismatch}) = {} & \overline{\Delta G^\circ}_{37\text{ hairpin}}(\text{hairpins of 6 nt with UU and GA first mismatch}) \\ & -\overline{\Delta G^\circ}_{37\text{ hairpin}}(\text{hairpins of 6 nt without UU and GA first mismatch and} \\ & \text{without G}\cdot\text{U closing pair preceded by two G residues}) \\ & = 4.61\text{ kcal/mol} - 5.40\text{ kcal/mol} = -0.8\text{ kcal/mol} \qquad (7)\end{aligned}$$

The error for the $\Delta G^\circ_{37\text{bonus}}$ can be approximated from the formula for error propagation as:

$$\sigma(\Delta G^\circ_{37\text{bonus}}) = \sqrt{\begin{matrix}\sigma^2(\Delta G^\circ_{37\text{hairpin}}\text{hairpins with UU and GA mismatch})+ \\ \sigma^2(\Delta G^\circ_{37\text{hairpins}}\text{hairpins without UU or GA mismatch})\end{matrix}} = 0.22\text{ kcal/mol} \qquad (8)$$

where σ is the standard deviation of the indicated data.

For loops of sizes other than six, the $\Delta G^\circ_{37\text{initiation}}$ is calculated for each studied sequence, excluding those 5′G·U 3′ closed hairpins preceded by two *G* residues and the oligo-C hairpins. For sequences with UU or GA first mismatch, the $\Delta G^\circ_{37\text{bonus}}$ is subtracted from the $\Delta G^\circ_{37\text{hairpin}}$ to determine a $\Delta G^\circ_{37\text{initiation}}$. The $\Delta G^\circ_{37\text{initiation}}$ used for predicting energies is the average of all the $\Delta G^\circ_{37\text{initiation}}$ values for hairpins of that length. Table 6 contains the $\Delta G^\circ_{37\text{initiation}}$ for hairpins from three to nine nucleotides in length along with the standard deviation of the average. Based on the treatment by Jacobson & Stockmayer (1950), $\Delta G^\circ_{37\text{initiation}}$ for hairpins longer than nine nucleotides is approximated by the equation:

$$\begin{aligned}\Delta G^\circ_{37\text{initiation}}(n > 9) = {} & \Delta G^\circ_{37\text{initiation}}(9) \\ & + 1.75RT\ln(n/9) \qquad (9)\end{aligned}$$

where n is the number of unpaired nucleotides, R is the gas constant, and T is absolute temperature (310.15 K for 37°C). Hairpins of one and two nucleotides are assumed to be too unfavorable to occur with standard geometries.

The data of hairpin loop stabilities demonstrate that 5′ G·U 3′ closed hairpins are more stable than other hairpins when the 5′ G residue is preceded by two G residues (Giese *et al*., 1998). Furthermore, in the database of secondary structures assembled here, the most frequently occurring nucleotides 5′ to a G·U closed hairpin are 5′ GG 3′. Of the 74 5′ G·U 3′ closed hairpins, 26 % are preceded by two G residues. Interestingly, another 31 % are divided between preceding 5′ GC 3′ and 5′ GU 3′. On average, 5′ G·U 3′ closed hairpins preceded by two G residues are more stable than other hairpins of the same length by −2.2 (±0.53) kcal/mol. During secondary structure prediction, these hairpins are assigned a −2.2 kcal/mol bonus.

Oligo-C hairpin loops are less stable than other hairpin loops of the same length (Groebe & Uhlenbeck, 1988). For the four oligo-C loops studied, the reduced stability increases with the number of unpaired nucleotides. The free energy penalty for oligo-C loops with more than three unpaired nucleotides fit equation 5 where n is the number of nucleotides and A and B are determined by linear regression to be 0.30 (±0.052) kcal/mol and 1.6 (±0.34) kcal/mol, respectively. The coefficient of determination, R^2, for this is 0.95. For the hairpin with three C residues, the penalty is 1.4 kcal/mol, based on one measurement. The folding algorithm applies these penalties to oligo-C loops during structure prediction.

Tetraloop bonuses

Specific tetraloops, i.e. hairpin loops with four nucleotides, are assigned enhanced stability. Some are known to be more stable than the model above would predict (Tuerk *et al*., 1988; Antao & Tinoco, 1992; Antao *et al*., 1991; Varani *et al*., 1991) and others are known to be important in stabilizing tertiary structure (Costa & Michel, 1995; Butcher *et al*., 1997; Lehnert *et al*., 1996; Cate *et al*., 1996; Jucker & Pardi, 1995; Michel & Westhof, 1990). Therefore, a table of special tetraloop sequences and the corresponding stability bonuses (Table 8) is consulted by the program for assigning stability of hairpin loops with four nucleotides.

The Table of special tetraloop sequences includes the closing base-pair. This allows a more discriminating application of tetraloop bonuses than applied previously when closing base-pair was not considered (Walter *et al*., 1994a; Jaeger *et al*., 1989). The abundance of a tetraloop sequence depends on the closing base-pair (Woese *et al*., 1990). For example, in the database assembled for this study, there are 914 tetraloops, the most common of which is GGGGAC, occurring 87 times. AGGGAU and CGGGAG occur four times and eight times, respectively, and UGGGAA does not occur at all. The magnitude of the bonus for each loop (Table 8) is based on its abundance in the database of structures assembled to test the algorithm. For this database, structures for each type of RNA were chosen from all available branches of phylogeny. Loops that occur more than 22 times in the structure database receive a bonus of −3.0 kcal/mol. Loops that occur between 16 and 18 times have a −2.5 kcal/mol bonus, −2.0 kcal/mol is assigned to loops that occur between 11 and 14 times, and a −1.5 kcal/mol bonus is assigned to loops that have between six and nine occurrences, inclusively.

With short RNA strands, tertiary interactions are not important and a second table of tetraloop stabilities can be used that are based on thermodynamic measurements (Antao & Tinoco, 1992). There are three tetraloop sequences in the literature known to have thermodynamic stability in excess of that predicted by the above model. These sequences are CUUCGG, CUACGG, and CGCUUG with bonuses of −2.1, −1.5, and −0.9 kcal/mol, respectively (Antao & Tinoco, 1992).

Bulge loops

Bulge loops, an interruption of helical structure in one strand only, destabilize RNA structure (Longfellow *et al*., 1990; Groebe & Uhlenbeck, 1989; Fink & Crothers, 1972). The free energy increments of bulge loops depend on the nearest-neighbor model used for helices. Stabilities were calculated with the INN-HB model by Xia *et al*. (1998) using the equation:

$$\Delta G^\circ_{37\text{bulge}} = \Delta G^\circ_{37\text{initiation}}(n) + \Delta G^\circ_{37\text{ bp stack}}(\text{bulges of one nucleotide only}) \qquad (10)$$

This assumes helical stacking is continuous between the adjacent helices for single bulges, but is interrupted by bulges with $n \geqslant 2$ (Jaeger *et al*., 1989; Weeks & Crothers, 1993). Therefore, the terminal A·U penalty is applied for bulge loops longer than one nucleotide only. The values, $\Delta G^\circ_{37\text{ initiation}}(n)$ for bulge loops of one to six nucleotides, are listed in Table 9. For bulges longer than six nucleotides, the following approximation is used (Jacobson & Stockmayer, 1950; Jaeger *et al*., 1989):

$$\Delta G^\circ_{37\text{initiation}}(n > 6) = \Delta G^\circ_{37\text{initiation}}(6) + 1.75\, RT \ln(n/6) \qquad (11)$$

Values of $\Delta G^\circ_{37\text{bulge}}$ for one to three unpaired nucleotides were calculated from experimental data

Table 8. Tetraloop hairpin bonuses

Sequence	Occurrence	ΔG°_{37} bonus (kcal/mol)
GGGGAC	87	−3.0
GGUGAC	76	−3.0
CGAAAG	56	−3.0
GGAGAC	47	−3.0
CGCAAG	40	−3.0
GGAAAC	36	−3.0
CGGAAG	35	−3.0
CUUCGG	28	−3.0
CGUGAG	23	−3.0
CGAAGG	18	−2.5
CUACGG	17	−2.5
GGCAAC	17	−2.5
CGCGAG	16	−2.5
UGAGAG	16	−2.5
CGAGAG	14	−2.0
AGAAAU	13	−2.0
CGUAAG	11	−2.0
CUAACG	11	−2.0
UGAAAG	11	−2.0
GGAAGC	9	−1.5
GGGAAC	9	−1.5
UGAAAA	9	−1.5
AGCAAU	8	−1.5
AGUAAU	8	−1.5
CGGGAG	8	−1.5
AGUGAU	7	−1.5
GGCGAC	6	−1.5
GGGAGC	6	−1.5
GUGAAC	6	−1.5
UGGAAA	6	−1.5

The sequences, including closing base-pair, are listed in order of frequency of occurrence. For short RNA strands without tertiary interactions, a table with only measured stabilities is used. The three sequences with measured enhanced stability are CUUCGG, CUACGG, and CGCUUG with bonuses of −2.1, −1.5, and −0.9 kcal/mol, respectively (Antao & Tinoco, 1992).

Table 9. Free energy increments for bulges up to six nucleotides

Bulge length	$\Delta G^\circ_{37\text{ bulge}}$ (kcal/mol)	Standard deviation
1	3.8	1.1
2	2.8	1.3
3	3.2	1.9
4	(3.6)	-
5	(4.0)	-
6	(4.4)	-

Note that the nearest-neighbor parameter for stacking of adjacent base-pairs is added for bulges with one nucleotide. For bulges with more than one nucleotide, calculation of the stabilities of adjacent helices includes the terminal A·U penalty terms for A·U or G·U pairs adjacent to the bulge. For longer bulges, the stability is approximated by: $\Delta G^\circ_{37\text{initiation}}(n > 6) = \Delta G^\circ_{37\text{initiation}}(6) + 1.75\ RT\ \ln(n/6)$ (Jacobson & Stockmayer, 1950). Values in parenthesis are not from measurements.

(Longfellow *et al.*, 1990; Groebe & Uhlenbeck, 1989; Fink & Crothers, 1972) with:

$$\Delta G^{\circ}_{37\text{bulge}} = \Delta G^{\circ}_{37}(\text{entire sequence with bulge}) - \Delta G^{\circ}_{37}(\text{reference sequence}) + \Delta G^{\circ}_{37\text{bp stack}}(\text{interrupted nearest neighbor for base stacking for bulges} > 1 \text{ nucleotide}) \quad (12)$$

where the reference sequence is the sequence with the bulge removed. These free energies are listed in Table 10. Values for identical numbers of nucleotides were averaged. For loops of length four, five, and six, the free energy increments are increased by 0.4 kcal/mol above the increment for the next smaller bulge. This increase was chosen to be the same as the increase in free energy between bulges of two and three nucleotides because the logarithmic increase in equation (11) is expected only for longer loops (Jacobson & Stockmayer, 1950).

The 2×2 internal loops (tandem mismatches)

The stabilities of 2 × 2 internal loops (an interruption of helical RNA by two opposing unpaired nucleotides in each strand), also called tandem mismatches, have been the subject of several prior studies (Xia *et al.*, 1997; Wu *et al.*, 1995; Walter *et al.*, 1994c, SantaLucia *et al.*, 1991a,b). Xia *et al.* (1997) developed a method of extrapolating stabilities for all 2 × 2 internal loops based upon known 2 × 2 loop stabilities. These free energy parameters have now been calculated using the INN-HB nearest-neighbor parameters. This was necessary because the stability of the loop, $\Delta G^{\circ}_{37\text{loop}}$, depends on the nearest-neighbor model for helical RNA according to:

$$\begin{aligned}\Delta G^{\circ}_{37\text{loop}} =\, & \Delta G^{\circ}_{37}(\text{entire sequence with loop}) \\ & - \Delta G^{\circ}_{37}(\text{reference sequence}) \\ & + \Delta G^{\circ}_{37}(\text{interruped nearest neighbor for base stacking})\end{aligned} \quad (13)$$

where the reference sequence is identical with the entire sequence except that the tandem mismatch is absent.

Table 11 summarizes the stabilites of symmetric 2 × 2 internal loops (Xia *et al.*, 1997; Wu *et al.*, 1995; Walter *et al.*, 1994c, SantaLucia *et al.*, 1991a,b). The sequences are listed as a periodic table of stabilities with tandem mismatch sequence on the horizontal axis and adjacent base-pair on the vertical axis (Wu *et al.*, 1995). For unmeasured sequences, the stability is predicted as the average of the adjacent-most known stabilities to the left and right, except for

5′ GGGC 3′
3′ CGGG 5′

which is set equal to

5′ CGGG 3′
3′ GGGC 5′

The free energies of all possible asymmetric tandem mismatches closed by Watson-Crick pairs were approximated from the symmetric loop stabilities according to the simple equation (Xia *et al.*, 1997):

$$\Delta G^{\circ}_{37\text{predict}}\begin{pmatrix}5'\ \text{IXYK}\ 3'\\ 3'\ \text{JWZL}\ 5'\end{pmatrix} = \left[\Delta G^{\circ}_{37\text{loop}}\begin{pmatrix}5'\ \text{IXWJ}\ 3'\\ 3'\ \text{JWXI}\ 5'\end{pmatrix} + \Delta G^{\circ}_{37\text{loop}}\begin{pmatrix}5'\ \text{LZYK}\ 3'\\ 3'\ \text{KYZL}\ 5'\end{pmatrix}\right]/2 + \Delta \quad (14)$$

where I·J and K·L are Watson-Crick pairs. The Δ is related to the size and stability of the mismatches as summarized in Table 12. In prior studies (Xia *et al.*, 1997; Mathews *et al.*, 1998), Δ values for A-U closure were estimated as less than Δ values for G·C closure. For the database of Table 1, however, structure prediction is more accurate when this difference in Δ is eliminated. For predicting the free energy of 2 × 2 loops closed by G·U pairs, each G·U pair is treated as an A-U pair. The secondary structure prediction algorithm consults a Table of free energies of all possible tandem mismatches.

Xia *et al.* (1997) measured stabilities of asymmetric tandem mismatches which are used to determine Δ (equation (14)). Table 13 gives the stabilities for 2×2 internal loops with the sequence

5′ GXYG 3′
3′ CWZC 5′

with XW and YZ as mismatches. The Δ values calculated for each experimentally measured asymmetric 2 × 2 internal loop, shown in Table 13, fit into categories of similar magnitude based on stability and mismatch size. Stable mismatches are defined as GU, GA, and UU; all other mismatches are destabilizing. The next criterion, mismatch size, is either purine-purine, purine-pyrimidine, or pyrimidine-pyrimidine. The first category is a combination of all mismatches of equal size, e.g.

5′ GAAG 3′
3′ CAGC 5′

or any combination of two unstable mismatches (excluding AC mismatches). The average Δ for these tandem mismatches is 0.0 (±0.4) kcal/mol. The next category is a combination of two stabilizing mismatches of different sizes, with an average Δ of 1.8 (±0.4) kcal/mol. Another category is composed of tandem mismatches with one stable and one unstable mismatch (excluding AC) of

Table 10. The free energy of bulged nucleotides

Bulge loop sequence	ΔG°_{37} (kcal/mol)	Reference sequence	$\Delta G^{\circ}_{37\ \text{reference}}$ (kcal/mol)	$\Delta G^{\circ}_{37\ \text{bulge}}$ (kcal/mol)
GGGACUC$\underline{\text{A}}$CG$\underline{\text{AUUA}}$CGGAGUCUAU[a]	−7.7	GGGACUCCG$\underline{\text{AUUA}}$CGGAGUCUAU	−11.1	3.4
GGGACUC$\underline{\text{U}}$CG$\underline{\text{AUUA}}$CGGAGUCUAU[a]	−8.4	GGGACUCCG$\underline{\text{AUUA}}$CGGAGUCUAU	−11.1	2.7
GGGACUC$\underline{\text{G}}$CG$\underline{\text{AUUA}}$CGGAGUCUAU[a]	−7.6	GGGACUCCG$\underline{\text{AUUA}}$CGGAGUCUAU	−11.1	3.5
GCG$\underline{\text{A}}$GCG + CGCCGC[b]	−6.9	GCGGCG+CGCCGC	−10.4	3.5
GCG$\underline{\text{U}}$GCG + CGCCGC[b]	−6.7	GCGGCG+CGCCGC	−10.4	3.7
GCGGCG + CGC$\underline{\text{A}}$CGC[b]	−6.9	GCGGCG+CGCCGC	−10.4	3.5
GCGGCG$\underline{\text{A}}$+CGC$\underline{\text{A}}$CGC$\underline{\text{A}}$[b]	−8.7	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	5.9
GCG$\underline{\text{A}}$GCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$[b]	−9.5	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	5.1
GCAACG$\underline{\text{A}}$+CGU$\underline{\text{A}}$UGC$\underline{\text{U}}$[b]	−4.6	GCAACG$\underline{\text{A}}$+CGUUGC$\underline{\text{U}}$	−9.2	4.6
GACCGC$\underline{\text{A}}$+GCG$\underline{\text{A}}$GUC$\underline{\text{A}}$[b]	−10.1	GACCGC$\underline{\text{A}}$+GCGGUC$\underline{\text{A}}$	−12.1	2.0
GCGGCG + CGC$\underline{\text{AA}}$CGC[b]	−5.4	GCGGCG+CGCCGC	−10.4	1.7
GCG$\underline{\text{UU}}$GCG + CGCCGC[b]	−5.0	GCGGCG+CGCCGC	−10.4	2.1
GCG$\underline{\text{AA}}$GCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$[b]	−6.7	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	4.6
GCG$\underline{\text{AA}}$GCG + CGCCGC[b]	−5.2	GCGGCG+CGCCGC	−10.4	1.9
GCGGCG$\underline{\text{A}}$+CGC$\underline{\text{AA}}$CGC$\underline{\text{A}}$[b]	−7.1	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	4.2
GACCGC$\underline{\text{A}}$+GCG$\underline{\text{AA}}$GUC$\underline{\text{A}}$[b]	−6.6	GACCGC$\underline{\text{A}}$+GCGGUC$\underline{\text{A}}$	−12.1	2.2
GCG$\underline{\text{AAA}}$GCG + CGCCGC[b]	−4.7	GCGGCG+CGCCGC	−10.4	2.4
GCG$\underline{\text{UUU}}$GCG + CGCCGC[b]	−4.8	GCGGCG+CGCCGC	−10.4	2.3
GCGGCG + CGC$\underline{\text{AAA}}$CGC[b]	−6.7	GCGGCG+CGCCGC	−10.4	0.4
GCG$\underline{\text{AAA}}$GCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$[b]	−5.4	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	5.9
GCGGCG$\underline{\text{A}}$+CGC$\underline{\text{AAA}}$CGC$\underline{\text{A}}$[b]	−7.3	GCGGCG$\underline{\text{A}}$+CGCCGC$\underline{\text{A}}$	−14.6	4.0
GACCGC$\underline{\text{A}}$+GCG$\underline{\text{AAA}}$GUC$\underline{\text{A}}$[b]	−5.0	GACCGC$\underline{\text{A}}$+GCGGUC$\underline{\text{A}}$	−12.1	3.8

Measurements of bulge loop stability and the corresponding reference sequence stability are from [a] Groebe & Uhlenbeck (1989) and [b] Longfellow *et al*. (1990). Unpaired nucleotides are underlined.

different sizes, with an average Δ of 1.0 (±0.4) kcal/mol. The final category is the tandem mismatches with at least one AC mismatch. These are the least well determined, with an average Δ of 0.0 (±0.7) kcal/mol. Table 12 summarizes these categories, Δ values, and standard deviations.

The Δ value is a deviation from a simple average of the two symmetric tandem mismatches containing the mismatches and adjacent base-pairs in the asymmetric mismatch. It is a stability penalty paid when differently sized mismatches are adjacent in a loop. Evidently, when the helical backbone accommodates differently sized mismatches, the helix is destabilized. This effect is more pronounced with a combination of stabilizing mismatches (Xia *et al*., 1997).

The 2×1 internal loops

The stability of internal loops of the form

5′ GXG 3′
3′ CZYC 5′

and

5′ CXC 3′
3′ GZYG 5′

were studied by Schroeder *et al.* (1996). Table 14 summarizes the stability of G·C closed 2 × 1 internal loops. Values based on experiment are derived from the measurements by Schroeder *et al.* (1996) and the INN-HB parameters by Xia *et al.* (1998) using:

$$\begin{aligned}\Delta G^{\circ}_{37\text{loop}} =\, &\Delta G^{\circ}_{37}(\text{entire sequence with } 2\times 1 \text{ loop}) - \Delta G^{\circ}_{37}(\text{reference sequence})\\ &+\Delta G^{\circ}_{37}(\text{interrupted nearest neighbor for base stacking})\end{aligned} \qquad (15)$$

The stabilities of G·C closed 2 × 1 loops are used to predict the stabilities of unmeasured 2 × 1 loops by:

$$\begin{aligned}\Delta G^{\circ}_{37\text{loop}} = &\,\Delta G^{\circ}_{37\text{loop}}(\text{G}\cdot\text{C closing pairs})\\ &+ \Delta G^{\circ}_{37\text{penalty}}(\text{for each A}\cdot\text{U closing pair})\end{aligned}$$

The $\Delta G^{\circ}_{37\text{loop}}$(G·C closing pairs) is the average of

5′ GXG 3′
3′ CZYC 5′

and

5′ CXC 3′
3′ GZYG 5′

if both were measured. To complete the table of G·C closed loops, three approximations are used (Schroeder *et al*., 1996). The first two approximations are that

A
3′ CG 5′

is approximated by the average of

5′ GAG 3′
3′ CAGC 5′

and

5′ CAC 3′
3′ GAGG 5′

and

A
3′ GC 5′

is approximated by the average of

5′ GAG 3′
3′ CGAC 5′

Table 11. The periodic table of tandem mismatches

mismatch: closing BP	UG GU	GU UG	GA AG	AG GA	UU UU	GG GG	CA AC	CU UC	UC CU	CC CC	AC CA	AA AA
G C	−4.9[e] −4.8[f] **−4.9**	−4.7[a] −4.2[b] −4.1[a] −3.4[a] **−4.1**	−2.9[c] −2.5[c] −2.4[c] **−2.6**	−1.3[c]	−0.5[b]	(0.8)	1.0[b]	0.9[b] 1.3[b] **1.1**	(1.0)	(1.0)	0.9[b]	1.5[b]
C G	−4.2[e] −4.1[f] **−4.2**	−1.3[a] −0.8[a] **−1.1**	−0.7[d]	−1.0[d] −0.5[g] **−0.7**	−0.4[d]	0.8[d]	1.1[d]	1.4[d]	1.4[d]	1.7[d]	2.0[d]	1.3[d]
U A	−2.9[a] −2.3[f] **−2.6**	−0.3[a]	0.7[c]	(0.7)	1.1[b]	(1.5)	1.9[b]	2.2[b]	2.8[b]	(2.8)	(2.8)	2.8[b]
A U	−2.1[a] −1.6[f] **−1.9**	−0.1[a] 0.5[f] **0.2**	−0.1[c] 0.7[c] **0.3**	(0.3)	0.6[b]	(1.4)	2.3[b]	(2.2)	(2.2)	(2.2)	2.5[b]	2.8[b]

The stabilities (kcal/mol) are drawn from: [a] He *et al.* (1991); [b] Wu *et al.* (1995); [c] Walter *et al.* (1994b); [d] SantaLucia *et al.* (1991a,b); [e] Sugimoto *et al.* (1986); [f] McDowell *et al.* (1997); and [g] Xia *et al.* (1997). Boldfaced numbers are averages of multiple measurements on the same tandem mismatch and numbers in parenthesis are predicted stabilities. The free energies of reference helices were taken from several sources: Freier *et al.* (1985, 1986a,b), He *et al.* (1991), Sugimoto *et al.* (1986), Walter *et al.* (1994c), Wu *et al.* (1995), and Xia *et al.*, 1997.

Table 12. Values of Δ listed by category of tandem mismatch

Category	Δ (kcal/mol)	Standard deviation
Same size mismatches or two destabilizing mismatches of different sizes	0.0	0.4
Any tandem mismatch with at least one AC	0.0	0.7
Two stabilizing mismatches of different sizes	1.8	0.4
One stabilizing and one destabilizing mismatch of different sizes (no A·C)	1.0	0.4

Stabilizing mismatches are GU, GA, or UU. All other mismatches are considered destabilizing. These values can be used with Table 11 to predict the free energy of any asymmetric 2 × 2 internal loop according to equation (14):

$$\Delta G^{\circ}_{37\text{predict}}\begin{pmatrix}5'\ \text{IXYK}\ 3'\\ 3'\ \text{JWZL}\ 5'\end{pmatrix} = \left[\Delta G^{\circ}_{37\text{loop}}\begin{pmatrix}5'\ \text{IXWJ}\ 3'\\ 3'\ \text{JWXI}\ 5'\end{pmatrix} + \Delta G^{\sigma}_{37\text{loop}}\begin{pmatrix}5'\ \text{LZYK}\ 3'\\ 3'\ \text{KYZL}\ 5'\end{pmatrix}\right]/2 + \Delta$$

Table 13. The free energy data (kcal/mol) for 2 × 2 internal loops of the form $\begin{matrix}5'\ \text{GXYC}\ 3'\\ 3'\ \text{CWZC}\ 5'\end{matrix}$, where X·W and Y·Z are mismatches

5′GXYG3′ 3′CWZC5′		Y Z⇒											
X⇓		G	U	A	G	U	G	A	U	C	C	C	A
W		U	G	G	A	U	G	C	C	U	C	A	A
U	exp	−4.4	−3.0	−0.9	−0.8	−1.0	(−0.2)	(−1.9)	(−0.8)	(−0.8)	−0.3	−0.7	0.1
G	ave	−4.5	−3.0	−2.8	−2.9	−2.6	*−2.0*	−1.9	−1.8	−1.8	−1.6	−1.5	−1.8
	Δ	0.1	0.0	1.9	2.1	1.7	(1.8)	(0.0)	(1.0)	(1.0)	1.3	0.7	1.9
G	exp	(−4.2)	−2.4	−1.3	(−0.8)	0.0	(−0.6)	(−1.5)	(−0.4)	−0.1	(−0.2)	−0.6	−0.5
U	ave	−4.2	−2.6	−2.4	−2.6	−2.3	*−1.6*	−1.5	−1.4	−1.4	−1.2	−1.1	−1.4
	Δ	(0.0)	0.2	1.2	(1.8)	2.3	(1.0)	(0.0)	(1.0)	1.3	(1.0)	0.5	0.9
G	exp	(−1.6)	0.3	−1.7	(−1.8)	0.5	(−0.9)	(−0.8)	(0.4)	(0.4)	0.7	0.1	−0.3
A	ave	−3.4	−1.8	−1.7	−1.8	−1.5	*−0.9*	−0.8	−0.6	−0.6	−0.5	−0.3	−0.6
	Δ	(1.8)	2.1	0.0	(0.0)	2.0	(0.0)	(0.0)	(1.0)	(1.0)	1.2	0.4	0.3
A	exp	−0.6	0.6	−0.5	−0.7	1.4	(−0.2)	0.7	(1.0)	0.5	0.7	0.5	0.4
G	ave	−2.7	−1.2	−1.0	−1.1	−0.9	*−0.2*	−0.1	0.0	0.0	0.2	0.3	0.0
	Δ	2.1	1.7	0.5	0.4	2.3	(0.0)	0.8	1.0	0.5	0.6	0.1	0.3
U	exp	−1.0	0.6	0.9	0.9	−0.6	(1.2)	−0.2	0.2	(0.4)	−0.1	0.0	1.5
U	ave	−2.4	−0.8	−0.6	−0.8	−0.5	*0.2*	0.3	0.4	0.4	0.6	0.7	0.4
	Δ	1.4	1.3	1.6	1.7	−0.2	(1.0)	−0.5	−0.2	0.0	−0.6	−0.7	1.1
G	exp	(−1.0)	(0.6)	(−0.3)	(−0.4)	(0.9)	(0.5)	(0.7)	(1.4)	(1.4)	(1.5)	(1.7)	(0.8)
G	ave	−2.0	−0.4	−0.3	−0.4	−0.1	*0.5*	0.7	0.8	0.8	0.9	1.1	0.8
	Δ	(1.0)	(1.0)	(0.0)	(0.0)	(1.0)	(0.0)	(0.0)	(0.6)	(0.6)	(0.6)	(0.6)	(0.0)
C	exp	(−1.6)	(0.0)	−0.4	−0.7	(1.3)	(1.5)	(1.0)	(1.2)	(1.2)	(1.3)	(1.5)	(1.1)
A	ave	−1.6	0.0	0.1	0.0	0.3	*0.9*	1.0	1.2	1.2	1.3	1.5	1.1
	Δ	(0.0)	(0.0)	−0.5	−0.7	(1.0)	(0.6)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
C	exp	(−0.6)	(1.0)	(1.2)	(1.1)	(0.3)	(1.6)	(1.1)	(1.2)	(1.2)	(1.4)	(2.1)	(1.8)
U	ave	−1.6	0.0	0.2	0.1	0.3	*1.0*	1.1	1.2	1.2	1.4	1.5	1.2
	Δ	(1.0)	(1.0)	(1.0)	(1.0)	(0.0)	(0.6)	(0.0)	(0.0)	(0.0)	(0.0)	(0.6)	(0.6)
U	exp	(−0.6)	(1.0)	(1.1)	(1.0)	0.1	(1.5)	(1.1)	1.7	(1.2)	(1.3)	(2.1)	(1.8)
C	ave	*−1.6*	*0.0*	*0.1*	*0.0*	*0.3*	*0.9*	*1.1*	*1.2*	*1.2*	*1.3*	*1.5*	*1.2*
	Δ	(1.0)	(1.0)	(1.0)	(1.0)	*−0.2*	(0.6)	(0.0)	0.5	(0.0)	(0.0)	(0.6)	(0.6)
C	exp	(−0.6)	(1.0)	(1.1)	−0.6	(0.3)	(1.5)	(1.1)	(1.2)	(1.2)	(1.3)	(2.1)	(1.8)
C	ave	*−1.6*	*0.0*	*0.1*	*0.0*	*0.3*	*0.9*	*1.1*	*1.2*	*1.2*	*1.3*	*1.5*	*1.2*
	Δ	(1.0)	(1.0)	(1.0)	*−0.7*	(0.0)	(0.6)	(0.0)	(0.0)	(0.0)	(0.0)	(0.6)	(0.6)
A	exp	(−1.6)	(−0.1)	0.7	−1.0	0.3	(1.5)	(1.0)	(1.7)	(1.7)	1.9	1.1	(1.7)
C	ave	−1.6	−0.1	0.1	0.0	0.2	*0.9*	1.0	1.1	1.1	1.3	1.4	1.1
	Δ	(0.0)	(0.0)	0.6	−0.9	0.1	(0.6)	(0.0)	(0.6)	(0.6)	0.6	−0.4	0.6
A	exp	−0.1	0.9	−0.2	0.0	1.4	(1.2)	(1.3)	(2.0)	(2.0)	2.2	0.6	0.5
A	ave	−1.3	0.2	0.4	0.3	0.6	*1.2*	1.3	1.4	1.4	1.6	1.8	1.4
	Δ	1.3	0.7	−0.6	−0.2	0.9	(0.0)	(0.0)	(0.6)	(0.6)	0.6	−1.1	−0.9

Rows labeled exp give the experimental free energy if available. Otherwise, predicted values are given in parenthesis. Rows labeled ave give the average of two symmetric loops containing the mismatches in the asymmetric loop. Italicized free energies are derived from predicted stabilities in the periodic table of tandem mismatches. Rows labeled Δ give the difference between the value of exp and ave for a given sequence. Values in parenthesis are predicted Δ values from Table 12 and those without parenthesis are calculations.

Table 14. Free energy of 2 × 1 internal loops closed by G·C pairs

$\Delta G^\circ_{37\text{loop}}$ for $\begin{array}{c}5'\text{C}\ \ \text{X C}3'\\3'\text{GZYG}5'\end{array}$ and $\begin{array}{c}5'\text{G X G}3'\\3'\text{CZYC}5'\end{array}$ in kcal/mol

ZY	X = A	ZY	X = C	ZY	X = G
AA	2.3[a]	AA	2.3[a]	AA	1.7[a]
	2.5[b]				
AC	2.1[b]	AC	(2.2)		
AG	0.8[a]	AU	2.5[a]	AG	0.8[a]
	1.2[b]			GA	0.8[a]
				GG	(2.2)
CA	(2.2)	CA	(2.2)		X = U
CC	1.7[a]	CC	2.5[a]	CC	2.2[a]
CG	(0.6)	CU	1.9[a]	CU	1.7[a]
GA	1.1[a]	UA	(2.2)		
	2.1[b]				
GC	(1.6)	UC	(2.2)	UC	1.5[a]
GG	0.4[b]	UU	(2.2)	UU	1.2[a]

Values in parenthesis are predicted values.

[a] Indicates loops of $\begin{array}{c}5'\text{ CXC }3'\\3'\text{ GYZG }5'\end{array}$ and[b] indicates loops of $\begin{array}{c}5'\text{ GXG }3'\\3'\text{ CYZC }5'\end{array}$. Experimental values are derived from

\rlap\object="link_rf74"-Schroeder *et al.* (1996). Values for $\begin{array}{c}5'\text{ CUC }3'\\3'\text{ GCUG }5'\end{array}$ and $\begin{array}{c}5'\text{ GAG }3'\\3'\text{ CAGC }5'\end{array}$ are derived from recent studies (S. Schroder

and

$$\begin{array}{c}5'\text{ CAC }3'\\3'\text{ GGAG }5'\end{array}$$

These approximations were made because of the similarity in position of AG mismatches between the unmeasured and measured sequences. The last approximation is that the average of all loops without stable mismatches (GA or UU) is used to predict stabilities of unmeasured loops without stable mismatches. This average is 2.2 (±0.3) kcal/mol. The penalty for A·U closure is extrapolated from the data on symmetric 2 × 2 internal loops not including G·U mismatches. On average, the A·U closed symmetric 2 × 2 loops are 1.3 (±0.4) kcal/mol less stable than the G·C closed loops of the same sequence (Wu *et al.*, 1995). Therefore, a 0.65 kcal/mol penalty is applied per closing A·U in a 2 × 1 internal loop. This penalty is 0.2 kcal/mol larger than expected for simply a terminal A·U base-pair. The estimated stability of all 2 × 1 loops are contained in a table that is consulted by the algorithm during secondary structure prediction.

Single mismatches (1×1 internal loops)

Recent studies show that the stabilities of single mismatches are more sequence dependent than previously realized (Morse & Draper, 1995; Zhu & Wartell, 1997; Bevilacqua & Bevilacqua, 1998; Meroueh & Chow, 1999; R. Kierzek, M. E. Burkard & D.H.T., unpublished results). To take this sequence dependence into account, the new version of the algorithm consults a table of all possible 1 × 1 loops and closing base-pair combinations for the free energies of single mismatches.

The table of 1 × 1 loops contains experimentally determined values from our lab (R.K. *et al.*, unpublished results; Peritz *et al.*, 1991). For all other mismatches except GG, the average of 0.4 kcal/mol for all experimentally determined stabilities of single mismatches with two adjacent GC pairs, excluding the GG mismatches, are used. The sequence,

$$\begin{array}{c}5'\text{ GUC }3'\\3'\text{ CUG }5'\end{array}$$

was also excluded from the average because it is nearly 1 kcal/mol more stable than other non-GG single mismatches. For GG mismatches, the average stability, −1.7 kcal/mol, is used. For each A·U or G·U closure, a 0.65 kcal/mol penalty (determined from the 2 × 2 internal loops) is applied. This penalty is consistent with experimental results (R.K.., M.E.B. & D.H.T., unpublished results).

Other internal loops

A simple approximation is used to model the free energies of other internal loops (Serra & Turner, 1995):

$$\Delta G^\circ_{37\text{loop}} = \Delta G^\circ_{\text{initiation}}(n1+n2) + \Delta G^\circ_{\text{asymm.}}|n1-n2| + \Delta G^\circ_{\text{AU/GU closure penalty}} + \Delta G^\circ_{\text{UU/GA/AG bonus}} \quad (17)$$

where $n1$ is the number of nucleotides on one side of the loop and $n2$ is the number of nucleotides on the other side of the loop. $\Delta G^\circ_{\text{initiation}}(n1+n2)$ is a penalty term for closing the loop (see Table 17), $\Delta G^\circ_{\text{asymm.}}|n1-n2|$ is a term that penalizes asymmetry in the loop, $\Delta G^\circ_{\text{AU/GU closure penalty}}$ is a penalty for loops closed by either A·U or G·U base-pairs, and $\Delta G^\circ_{\text{UU/GA/AG bonus}}$ is a favorable bonus for loops with a UU or GA first mismatch in either orientation. Table 15

Table 15. Sequence-dependent free energy terms for internal loop stability

$\Delta G^\circ_{\text{asymm.}}$ (kcal/mol)	0.48±0.04
$\Delta G^\circ_{\text{AU/GU closure penalty}}$ (kcal/mol)	0.2
$\Delta G^\circ_{\text{GA/AG bonus}}$ (GA or AG mismatch) (kcal/mol)	−1.1
$\Delta G^\circ_{\text{UU bonus}}$ (UU mismatch) (kcal/mol)	−0.7

These terms are used to predict the free energy of internal loops according to equation (17):

$$\Delta G^\circ_{37\text{loop}} = \Delta G^\circ_{\text{initiation}}(n1+n2) + \Delta G^\circ_{\text{asymm.}}|n1-n2| + \Delta G^\circ_{\text{AU/GU closure penalty}} + \Delta G^\circ_{\text{UU/GA/AG bonus}}$$

summarizes these sequence dependent free energy terms. Internal loops that are $1 \times n$ are not given a $\Delta G^{\circ}_{\text{UU/GA/AG bonus}}$ (S. Schroeder & D.H.T., unpublished results). This model for internal loop stability neglects sequence in the loop apart from the terminal mismatches.

To determine $\Delta G^{\circ}_{\text{initiation}}$, the $\Delta G^{\circ}_{37\text{loop}}$ values were recalculated for the sequences studied by Peritz *et al.* (1991) using the nearest neighbor Watson-Crick parameters by Xia *et al.* (1998) according to equation (13). Table 16 shows the loop free energies calculated. For even values of $(n1 + n2)$, $\Delta G^{\circ}_{\text{initiation}}(n1 + n2)$ is the average value of $\Delta G^{\circ}_{37\text{loop}}$ for the loops in which $n1$ equals $n2$. For odd values of $(n1 + n2)$, the $\Delta G^{\circ}_{\text{initiation}}(n1 + n2)$ is interpolated as the average of $\Delta G^{\circ}_{\text{initiation}}(n1 + n2 + 1)$ and $\Delta G^{\circ}_{\text{initiation}}(n1 + n2 - 1)$. Table 17 gives the average values and their errors. For even $(n1 + n2)$, this reported error is the standard deviation, and for odd $(n1 + n2)$, the error is estimated from error propagation as:

$$\sigma(n1 + n2) = \sqrt{\sigma^2(n1 + n2 + 1) + \sigma^2(n1 + n2 - 1)} \quad (18)$$

Internal loops in which the number of nucleotides are equal for the two strands are more favorable than loops with asymmetry in the number of nucleotides per strand. To model this effect, the term $\Delta G^{\circ}_{\text{asymm.}}|n1 - n2|$ appears in the approximation for internal loop free energy. To determine the $\Delta G^{\circ}_{\text{asymm.}}$ term, a $\Delta\Delta G^{\circ}$ of asymmetry is calculated for each asymmetric loop according to:

$$\Delta\Delta G^{\circ}_{37\text{loop}} - \Delta G^{\circ}_{\text{initiation}}(n1 + n2) \quad (19)$$

Table 16 gives the $\Delta\Delta G^{\circ}$ for each asymmetric loop and Figure 3 shows a plot of $\Delta\Delta G^{\circ}$ as a function of $|n1 - n2|$. This effect is sequence dependent, but it increases roughly linearly with $|n1 - n2|$. The data, excluding 2×1 internal loops, are fit to a line with intercept at the origin as shown in Figure 3. The slope of the line, 0.48 ($\pm$0.04) kcal/mol, is used as $\Delta G^{\circ}_{\text{asymm.}}$. The 2×1 loop penalties are larger than the 3×2 loop penalties and are excluded from the linear fit of $\Delta\Delta G^{\circ}$. The 2×1 loops are treated correctly by the secondary structure prediction algorithm using the separate model described above.

Table 16. Free energies for internal loops

Sequence 1	ΔG°_{37} (kcal/mol)	$\Delta G^{\circ}_{37\text{loop}}$ (kcal/mol)	$\Delta\Delta G^{\circ}$ asymmetry (kcal/mol)	Predicted $\Delta G^{\circ}_{37\text{loop}}$(kcal/mol)
$\mathrm{UGAC^{A}CUCA}$ $\mathrm{ACUG_{A}GAGU}$	−7.94	1.14	-	1.1
$\mathrm{UGAC\ ^{A}\ CUCA}$ $\mathrm{ACUG_{AA}GAGU}$	−6.42	2.66	1.17	2.0[a]
$\mathrm{UGAC^{AA}CUCA}$ $\mathrm{ACUG\ _{A}\ GAGU}$	−6.59	2.49	1.20	1.8[a]
$\mathrm{UGAC^{AA}CUCA}$ $\mathrm{ACUG_{AA}GAGU}$	−7.04	2.04	-	1.9
$\mathrm{UGAC\ ^{A}\ CUCA}$ $\mathrm{ACUG_{AAA}GAGU}$	−6.55	2.53	0.65	2.9
$\mathrm{UGAC^{AAA}CUCA}$ $\mathrm{ACUG\ _{A}\ GAGU}$	−6.16	2.92	1.24	2.7
$\mathrm{UGAC^{AAA}CUCA}$ $\mathrm{ACUG\ _{AA}\ GAGU}$	−6.59	2.49	0.60	2.4
$\mathrm{UGAC\ ^{A}\ CUCA}$ $\mathrm{ACUG_{AAAA}GAGU}$	−6.11	2.97	1.08	3.4
$\mathrm{UGAC^{AAAA}CUCA}$ $\mathrm{ACUG\ A\ GAGU}$	−5.56	3.52	1.83	3.2
$\mathrm{UGAC^{AAA}CUCA}$ $\mathrm{ACUG_{AAA}GAGU}$	−6.67	2.41	-	1.9
$\mathrm{UGAC\ ^{AA}\ CUCA}$ $\mathrm{ACUG_{AAAA}GAGU}$	−6.07	3.01	1.11	2.9
$\mathrm{UGAC^{AAAA}CUCA}$ $\mathrm{ACUG\ _{AA}\ GAGU}$	−6.00	3.08	1.18	2.9
$\mathrm{UGAC\ ^{A}\ CUCA}$ $\mathrm{ACUG_{AAAAA}GAGU}$	−5.71	3.37	1.47	3.9
$\mathrm{UGAC^{AAAAA}CUCA}$ $\mathrm{ACUG\ _{A}\ GAGU}$	−5.30	3.78	2.08	3.7
$\mathrm{UGAC^{CAA}CUCA}$ $\mathrm{ACUG_{AAA}GAGU}$	−7.14	1.94	-	2.4
$\mathrm{UGAC^{AAA}CUCA}$ $\mathrm{ACUG_{AAC}GAGU}$	−7.17	1.91	-	1.9
$\mathrm{CGC^{A}GCG}$ $\mathrm{GCG_{A}CGC}$	−6.09	0.67	-	0.7
$\mathrm{CGC^{AA}GCG}$ $\mathrm{GCG_{AA}CGC}$	−5.44	1.32	-	1.5[b]
$\mathrm{CGC^{AAA}GCG}$ $\mathrm{GCG_{AAA}CGC}$	−4.88	1.88	-	1.5
$\mathrm{CGG^{AAA}CCG}$ $\mathrm{GCC_{AAA}GGC}$	−4.64	1.84	-	2.3

ΔG°_{37} The values are from Peritz *et al.* (1991). Reference helices used to calculate $\Delta G^{\circ}_{37\text{loop}}$ are from Peritz *et al.* (1991) except for CGGCCG (Freier *et al.*, 1985).

[a] Predicted free energy is according to rules for 2×1 internal loops.

[b] Predicted free energy is according to rules for 2×2 internal loops.

Table 17. Initiation free energies for internal loop formation

Internal loop length ($n1 + n2$)	$\Delta G^\circ_{\text{initiation}}$ (kcal/mol)	Error (kcal/mol)
4	1.7	0.51
5	1.8	0.56
6	2.0	0.23

For loops longer than six nucleotides, initiation is approximated by a Jacobson & Stockmayer (1950) function: $\Delta G^\circ_{37\text{initiation}}(n > 6) = \Delta G^\circ_{37\text{initiation}}(6) + 1.75\ RT\ \ln(n/6)$. Note that free energies for internal loops of 1 × 1, 2 × 1, or 2 × 2 nucleotides are predicted with separate rules.

The 2 × 2 internal loops closed by A·U base-pairs are less favorable than loops closed by G·C base-pairs by 0.65 kcal/mol per A·U closure. This is 0.2 kcal/mol less stable than the 0.45 kcal/mol term alone for terminal A·U base-pairs (Xia *et al.*, 1998). This 0.2 kcal/mol, called $\Delta G^\circ_{\text{AU/GU closure penalty}}$, is added to the terminal A·U penalty of 0.45 kcal/mol for each A·U or G·U closure of a large internal loop.

On average, GA and UU mismatches are more stable than CA mismatches in symmetric 2 × 2 internal loops by −1.1 and −0.7 kcal/mol, respectively, per mismatch. For large internal loops, these free energies, $\Delta G^\circ_{\text{UU/GA bonus}}$, are applied if the first mismatch in the loop is UU, GA, or AG regardless of closing base-pair. Thermodynamic studies of 1 × *n* internal loops show that GA and UU mismatches do not impart extra stability (S.S. & D.H.T., unpublished results). Therefore, no GA or UU bonus is applied to 1 × *n* internal loops.

Multibranch loops

The free energies of multibranch loops (junctions) are composed of an unfavorable term for initiation and favorable terms for stacking. The rules for multibranch loop stability differ between the dynamic programming algorithm and efn2.

In the dynamic programming algorithm, the stability of a multibranch loop is approximated by the equation:

$$\Delta G^\circ_{\text{loop}} = a_1 + b_1 n + c_1 h + \Delta G^\circ_{\text{dangle}} \tag{20}$$

where *n* is the number of unpaired nucleotides, and *h* is the number of branching helices. Each unpaired nucleotide adjacent to a helix contributes a favorable free energy of a dangling end. Table 18 gives the values of parameters a_1, b_1, and c_1. Note that b_1 is set to zero in the dynamic programming algorithm.

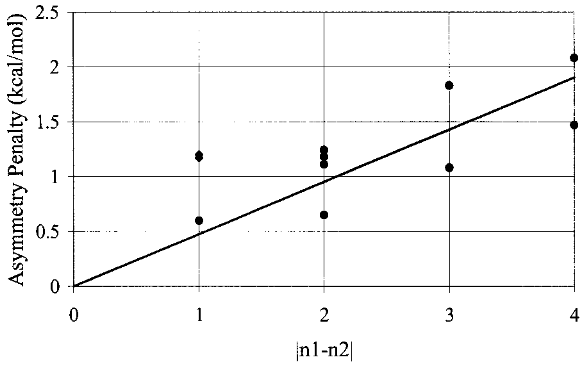

Figure 3. Plot of asymmetry penalty as a function of asymmetry. Penalties used for the fit are plotted with circles. The 2 × 1 internal loop derived asymmetry penalties, plotted with diamonds, were not used in the fit. The line is the best fit of the data to $y = mx$. The value of m is 0.48 kcal/mol.

In efn2, the form of the equation for approximating the initiation term of a multibranch loop depends on the number of unpaired nucleotides. For less than seven unpaired nucleotides, the form is:

$$\Delta G^\circ_{\text{loop}} = a_2 + b_2 n + c_2 h + \Delta G^\circ_{\text{stacking}} \tag{21}$$

where $\Delta G^\circ{}_{\text{stacking}}$ is the favorable free energy of coaxial stacking and terminal mismatch or dangling end stacking as described below. With seven or more nucleotides the form of the equation is:

$$\Delta G^\circ_{\text{loop}} = a_2 + 6b_2 + (1.1\ \text{kcal/mol}(\ln(n/6)) + c_2 h + \Delta G^\circ_{\text{stacking}} \tag{22}$$

Parameters a_2, b_2, and c_2 are given in Table 18.

Multibranch loops are potential sites of coaxial stacking, a favorable interaction of two helices stacked end to end. Stability increments for coaxial stacking have been measured in a model system composed of a short oligomer bound to a single stranded end of a stem-loop structure, creating a helical interface (Walter *et al.*, 1994a,b; Kim *et al.*, 1996). Coaxial stacking is observed as enhancement of stability of the duplex formed by the short helix above its stability without the interface. This is quantified as:

$$\begin{aligned}\Delta G^\circ_{\text{coaxial}} = {} & \Delta G^\circ \text{ (helix in context of stem loop structure)} \\ & - \Delta G^\circ \text{ (helix without stem loop structure)} \\ & + \Delta G^\circ_{\text{correction}}\end{aligned} \tag{23}$$

where $\Delta G^\circ_{\text{correction}}$ is the free energy for displacing a 3′ dangling end on the stem loop structure if one is present. Table 19 gives the free energy increments of coaxial stacking for interfaces without intervening mismatches when calculated with the INN-HB nearest-neighbor model for helices (Xia *et al.*, 1998). Table 20 gives the free energy increments for coaxial stacking with an intervening mismatch.

Efn2 gives an enhanced stability for coaxial stacking of adjacent helixes with at most one intervening mismatch. When helixes have no intervening mismatches, the stability bonus is the free energy parameter for stack-

Table 18. Initiation parameters for multibranch loops

Program		(kcal/mol)
Dynamic algorithm	a_1	3.4
	b_1	0.0
	c_1	0.4
efn2	a_2	10.1
	b_2	−0.3
	c_2	−0.3

The initiation of multibranch loops is approximated by:

$$\Delta G^\circ_{\text{initiation}} = a + bn + ch$$

where *n* is the number of unpaired nucleotides and *h* is the number of exiting helices. The exception to this is in the efn2 program when there are more than six unpaired nucleotides. For this case:

$$\Delta G^\circ_{\text{initiation}} = a_2 + 6b_2 + (1.1\ \text{kcal/mol}) \ln(n/6) + c_2 h$$

Table 19. Free energy increments for coaxial stacking without an intervening mismatch

Interface	Ref[a]	ΔG°_{37} helix (stem loop)[b] (kcal/mol)	ΔG°_{37} helix[c] (kcal/mol)	$\Delta G^\circ_{\text{coaxial}}$ (kcal/mol)	$\Delta G^\circ_{\text{coaxial}} - \Delta G^\circ_{\text{NN}}$ (kcal/mol)
5′GGAG-C- CCUC/G-	A	−7.8	−3.6	−4.2	−0.78
5′GGAG-G- CCUC/C-	B	−7.61	−3.6	−4.01	−0.75
5′GGAC-C- CCUG/G-	B	−7.99	−3.76	−4.23	−0.97
5′GGAG-A- CCUC/U-	B	−7.44	−3.6	−3.84	−1.49
5′GGUG-U- CCAC/A-	B	−6.86	−3.52	−3.34	−1.1
5′GGAC-G- CCUG/C-	A	−6.92	−3.76	−3.16	−0.8
5′GGUA-G- ACCAU/C-	B	−6.94	−3.99	−2.95	−0.87
5′GGAU-G- ACCUA/C-	B	−6.79	−3.87	−2.92	−0.81
5′GGAU-A- ACCUA/U-	B	−6.61	−3.87	−2.74	−1.41
5′GGAC/G- CCUG-C-	B	−5.52	−3.76	−3.46	−1.1
5′GGAC-G- CCUG/C- A	A	−6.65	−3.76	−2.89	−0.53
5′GGAC-G- CCUG/C- U	A	−6.5	−3.76	−2.74	−0.38
5′GGAC-G- CCUG/C- A	A	−5.08	−3.76	−3.02	−0.66
5′GGAC-G- CCUG/C- U	A	−5.37	−3.76	−2.81	−0.45
5′GGAC-G- ACCUG/C- U	A	−7.19	−5.67	−2.72	−0.36
5′GGAC-G- ACCUG/C- A A	A	−6.11	−5.67	−2.14	0.22
5′GGAC-G- ACCUG/C- A U	A	−6.59	−5.67	−2.12	0.24
5′GGAC-G- ACCUG/C- U A	A	−5.76	−5.67	−1.79	0.57
5′GGAC-G- ACCUG/C- U U	A	−5.77	−5.67	−1.3	1.06
5′GGAC-G- ACCUG/C- C A	A	−6.58	−5.67	−2.61	−0.25
5′GGAC-G- ACCUG/C- C U	A	−6.66	−5.67	−2.19	0.17
5′GGAG-C- CCUC/G- A	A	−7.29	−3.6	−3.69	−0.27
5′GGAG-C- CCUC/G- U	A	−7.56	−3.6	−3.96	−0.54
5′GGAG-C- CCUC/G- A	A	−5.8	−3.6	−3.3	0.12
5′GGAG-C- CCUC/G- U A	A	−5.03	−3.6	−2.53	0.89

The interface is shown with a dash indicating a continuation of the phosphate backbone and a slash indicating the discontinuity in the backbone. The hairpin, indicated by the parallel dashes, is not shown.

[a] Free energies are taken from the $1/T_m$ *versus* $\ln(C_T/4)$ plots of: A, Walter *et al.* (1994a), and B, Walter *et al.* (1994b).

[b] ΔG°_{37} measured for helices in the context of the stem-loop structure interface.

[c] ΔG°_{37} predicted by the INN-HB model (Xia *et al.*, 1998) for the helices without enhanced stability from coaxial stacking.

ing of base-pairs in a helix. This number was determined by calculating the excess stability above the helical stacking nearest-neighbor from Xia *et al.* (1998), $\Delta G^\circ_{\text{coaxial}} - \Delta G^\circ_{\text{NN}}$, for each measured interface. With flush interfaces, i.e. no intervening mismatch, and no strand extensions beyond the interface, the average excess stability is -1.0 (±0.27) kcal/mol. Each interface sequence has roughly the same excess stability indicating that the sequence dependence is similar to that of the nearest neighbor parameters. For interfaces followed by strand extensions, the excess stability is 0.0 (±0.54) kcal/mol. It is assumed that the environment inside a larger structure will be similar to the models with strand extensions and, therefore, coaxial stacking of helices with no intervening nucleotides is set to the nearest-neighbor parameter for those sequences within a helix.

With one intervening nucleotide, coaxial stacking is allowed when there is another nucleotide (5′ to the 5′ helix or 3′ to the 3′ helix) that can make an intervening mismatch. There are two distinct stacks (Figure 4). The first stack is on the side of the continuous backbone which, in Figure 4, is a 5′ A-G 3′ on a U-A. The parameters for this stack are the free energies of terminal

Table 20. Stability increments for coaxial stacking with an intervening mismatch

Interface	ΔG°_{37} helix (stem-loop)[a] (kcal/mol)	ΔG°_{37} helix (kcal/mol)	$\Delta G^\circ_{\mathrm{coaxial}}$ (kcal/mol)
5'GGACG-C- CCUGA/G-	−7.8	−5.31	−2.49
5'GGAC-GC- CCUG/AG-	−6.28	−3.76	−2.52
5'GGACG-A- CCUGA/U-	−7.41	−5.31	−2.1
5'GGAC-GA- CCUG/AU-	−5.86	−3.76	−2.1
5'GGAC-GA- CCUG/AG-	−5.70	−3.76	−1.94
5'GGACG-A- CCUGA/G-	−6.67	−5.31	−1.36
5'GGACA-C- CCUGG/G-	−7.79	−5.16	−2.63
5'GGAGC-CA- CCUCC/GU-	−5.64	−4.3	−1.34
5'GCACC-CA- CGUGC/GU-	−6.41	−4.78	−1.63
5'GGACG-C- CCUGA/G- A P	−6.34	−5.31	−2.13
5'GGACG-C- CCUGA/G- A A	−6.18	−5.31	−2.67
5'GGACG-C- CCUGA/G- P A	−6.5	−5.31	−2.29
5'GGAC-GC- CCUG/AG- P A	−5.63	−3.76	−2.57
5'GGAC-GC- CCUG/AG- G G	−5.43	−3.76	−2.37
5'GGACG-A- CCUGA/U- A A	−6.08	−5.31	−2.17
5'GGACG-A- CCUGA/G- A P	−6.22	−5.31	−1.51
5'GGAGC-CA- CCUCC/GU- A A	−5.12	−4.3	−1.92
5'GCACC-CA- CGUGC/GU- A A	−5.66	−4.78	−1.98

The interface sequence is shown with a dash indicating a continuation of the phosphate backbone and a slash indicating the discontinuity in the backbone. P indicates purine riboside. The hairpin, indicated by the parallel dashes, is not shown.

[a] The free energies are from $1/T_m$ *versus* $\ln(C_T/4)$ plots (Kim *et al.*, 1996).

mismatches. The second stack is on the side where the backbone opens for the entering and exiting strands. In Figure 4, this is an A-G on a C-G. This is made sequence independent, with a value of -2.1 kcal/mol, the average of the stability increments for intervening mismatches contained in Table 20.

Unpaired nucleotides adjacent to helices in multibranch loops can stack onto the end of the helix. When one nucleotide can stack on the end of a helix (3′ or 5′ end), the free energy given by efn2 is the dangling end free energy (above). When a 5′ and 3′ unpaired nucleotide can stack on the same helix, the more favorable free energy of either the 3′ dangling end or the terminal mismatch free energy is given (above).

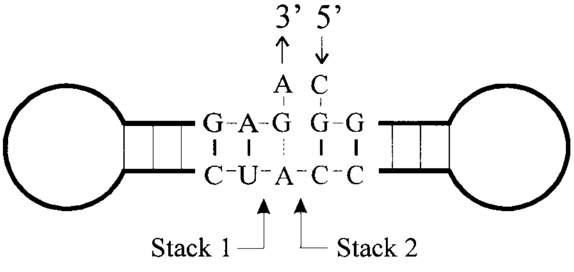

Figure 4. The two stacks involved in coaxial stacking with one mismatch pair intervening. Stack 1 is across a continuous backbone whereas stack 2 spans a break in the helix.

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 stacked dangling ends and cannot coaxially stack on more than one helix.

For example, consider the multibranch loop of the yeast phenylalanine tRNA structure shown in Figure 5 (Sprinzl *et al.*, 1998). Two coaxial stacks are possible in this structure. Helices I and IV can stack without intervening nucleotides and helices II and III can stack with an intervening mismatch. This mismatch is between nucleotides G26 and either A9 or A44. To determine which stacking is optimal, the free energy of each alternative is calculated. The coaxial stacking of helices I and IV contributes −2.4 kcal/mol or the dangling of nucleotides U8 and C48 on helices I and IV, respectively, would contribute a total of −0.4 kcal/mol. Therefore, efn2 predicts that helices I and IV are coaxially stacked with a free energy of −2.4 kcal/mol. For helices II and

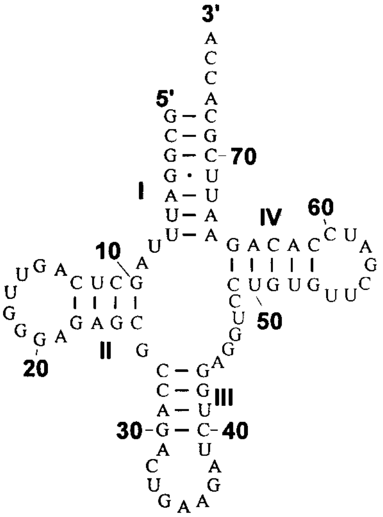

Figure 5. The yeast phenylalanine tRNA secondary structure. The helices are labeled with roman numerals and nucleotides are numbered.

III, the best predicted combination is to coaxially stack with nucleotides G26 and A9 intervening.

The initiation parameters a_1, b_1, c_1, a_2, b_2, and c_2 are found by optimizing the accuracy of the algorithm for the database of known secondary structures. To search the domain of these six variables, a genetic algorithm was written. Its starting point was suggested by stabilities determined by optical melting for an RNA multibranch loop with three branching helices (J.M. Diamond, D.H.M. & D.H.T., unpublished experiments).

A genetic algorithm works by random mutation and selection of the most fit set of variables (Forrest, 1993; Holland, 1975). In this case, the six initiation parameters are randomly mutated at each iterative step. Then, these parameters are used to predict secondary structures and the predicted structures are scored. The five most fit sets of parameters, those that result in the highest accuracy, are kept for the next iterative step for mutation.

Every fifth step in the genetic algorithm, the five sets of parameters are crossed to produce five new sets of parameters. To do this, the dynamic algorithm parameters, a_1, b_1, and c_1, and the efn2 parameters, a_2, b_2, and c_2, are chosen randomly from any of the five sets taken from the prior selection to make a new total set of parameters. For each other step, the five sets of parameters selected in the prior step are each randomly adjusted to produce a new set of parameters. Each initiation parameter had a 50 % chance of being adjusted up or down. Parameters c_1 and c_2 changed in magnitude as much as 1 kcal/mol in each step; all other parameters changed as much as 0.3 kcal/mol.

Reducing the search time and storage for multibranch loops

In versions 2.1 and earlier of *mfold*, three large triangular arrays were filled recursively. For each segment from nucleotides i to j inclusive, the quantities $W1(i,j)$, $W2(i,j)$ and $V(i,j)$ are computed. $W1(i,j)$ is the minimum folding energy of the fragment from i to j inclusive, conditional on this fragment being in a multibranch loop. Thus "free" bases and base-pairs are treated and penalized as though they were in a multibranch loop. $W2(i,j)$ is simply the unconditional minimum folding energy on i to j. Finally, $V(i,j)$ is the minimum folding energy from i to j assuming that the base-pair $i{\cdot}j$ forms. Since version 2.2, the array $W1$ is called W and the $W2$ array has been replaced by two linear arrays, $W5$ and $W3$. $W5(i)$ contains what used to be stored in $W2(1,j)$ and $W3(j)$ contains what used to be stored in $W2(j,n)$, where n is the number of nucleotides in the sequence. Both $W5$ and $W3$ can be computed recursively, and they are sufficient for finding the optimal folding on the entire sequence.

In searching for the best multibranch loop closed by base-pair $i{\cdot}j$, earlier versions of the algorithm explicitly computed:

$$VM = \min[VM_1(i,j),\quad VM_2(i,j),\quad VM_3(i,j),\quad VM_4(i,j)]$$

where:

$$\begin{aligned}
VM_1(i,j) &= \min[W(i+1,k) + W(k+1,j-1)] \quad \text{with} \quad i < k < j-1,\\
VM_2(i,j) &= b + \min[Ed(i,j,i+1) + W(i+2,k) + W(k+1,j-1)] \quad \text{with} \quad i+1 < k < j-1,\\
VM_3(i,j) &= b + \min[Ed(i,j,j-1) + W(i+1,k) + W(k+1,j-2)] \quad \text{with} \quad i < k < j-2,\\
VM_4(i,j) &= 2b + \min[Ed(i,j,i+1) + Ed(i,j,j-1) + W(k+1,j-2)] \quad \text{with} \quad i+1 < k < j-2
\end{aligned}$$

$Ed(x, y, z)$ is the free energy of base z dangling on the base-pair $x{\cdot}y$ and b is the constant (above) for the free energy of a multibranch loop.

The new version of the folding algorithm takes advantage of the fact that $\min[W(i,k) + W(k+1,j)]$, with $i \leqslant k < j$, is computed as part of the minimum folding energy from nucleotides i to j, inclusive. These numbers are now stored in an array, $WM(i,j)$, allowing VM to be computed as:

$$\begin{aligned}
VM_1(i,j) &= WM(i+1,j-1),\\
VM_2(i,j) &= b + Ed(i,j,i+1) + WM(i+2,j-1),\\
VM_3(i,j) &= b + Ed(i,j,j-1) + WM(i+1,j-2),\\
VM_4(i,j) &= 2b + Ed(i,j,i+1) + Ed(i,j,j-1) + W(i+2,j-2)
\end{aligned}$$

Because only values for j, $j-1$, and $j-2$ are addressed, it suffices to store only $WM(i,j(\text{mod}3))$ in an $n \times 3$ array. Thus storage increases only linearly with sequence length.

Improved enforcing of base-pairing constraints

A triangular array, $Fce(i,j)$, is defined for $1 \leqslant i \leqslant j \leqslant n$ where n is the number of nucleotides in the sequence. $Fce(i,j)$ is 1 if and only if there is at least one base, k, that must pair, where $i < k < j$. This array is defined before the execution of the fill algorithm by a simple recursive algorithm that executes in time order of n^2. During the fill algorithm, any loop containing a base that must pair is given a large free energy penalty (1600 kcal/mol) by checking the value of Fce for the end points of the single stranded region(s) in the loop. In multibranch and external loops, the energy penalty is included as soon as a single stranded nucleotide is added to the loop. In this way, bases that must pair are prevented from being in loops.

If the base-pair between bases i and j is forced, then both i and j are forced to pair as described above. In addition, all base-pair stacks involving either i or j are given the 1600 kcal/mol penalty if they do not contain the $i \cdot j$ base-pair.

Acknowledgements

The authors thank T. Xia, S. Schroeder, T. W. Barnes, and R. Kierzek for helpful discussions about the thermodynamic parameters. Thanks also go to M. E. Burkard for programming portions of RNAstructure. S. M. Freier, S. M. Testa, and J. M. Diamond gave invaluable help in debugging RNAstructure. Thanks to D. J. Williams and to Y. Ding for helpful suggestions. This work was supported by National Institutes of Health Grants GM22939 to D.H.T. and GM54250 to M.Z. D.H.M. is a Messersmith fellow and also received partial support from National Institutes of Health Grant 5T32 GM07356.

References

Antao, V. P. & Tinoco, I., Jr (1992). Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucl. Acids Res.* **20**, 819-824.

Antao, V. P., Lai, S. Y. & Tinoco, I., Jr (1991). A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucl. Acids Res.* **19**, 5901-5905.

Banerjee, A. R., Jaeger, J. A. & Turner, D. H. (1993). Thermal unfolding of a group I ribozyme: the low temperature transition is primarily a disruption of tertiary structure. *Biochemistry,* **32**, 153-163.

Bevilacqua, J. M. & Bevilacqua, P. C. (1998). Thermodynamic analysis of an RNA combinatorial library contained in a short hairpin. *Biochemistry,* **37**, 15877-15884.

Borer, P. N., Dengler, B., Tinoco, I., Jr & Uhlenbeck, O. C. (1974). Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.* **86**, 843-853.

Brown, J. W. (1998). The ribonuclease P database. *Nucl. Acids Res.* **26**, 351-352.

Burgstaller, P. & Famulok, M. (1997). Flavin-dependent photocleavage of RNA at G·U base pairs. *J. Am. Chem. Soc.* **119**, 1137-1138.

Burgstaller, P., Hermann, T., Huber, C., Westhof, E. & Famulok, M. (1997). Isoalloxazine derivatives promote photocleavage of natural RNAs at G·U base pairs embedded within helices. *Nucl. Acids Res.* **25**, 4018-4027.

Butcher, S. E., Dieckmann, T. & Feigon, J. (1997). Solution structure of a GAAA tetraloop receptor RNA. *EMBO J.* **16**, 7490-7499.

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

Correl, C. C., Freeborn, B., Moore, P. B. & Steitz, T. A. (1997). Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell,* **91**, 705-712.

Costa, M. & Michel, F. (1995). Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.* **14**, 1276-1285.

Crothers, D. M., Cole, P. E., Hilbers, C. W. & Schulman, R. G. (1974). The molecular mechanism of thermal unfolding of *Escherichia coli* formylmethionine transfer RNA. *J. Mol. Biol.,* **87**, 63-88.

Damberger, S. H. & Gutell, R. R. (1994). A comparative database of group I intron structures. *Nucl. Acids Res.* **22**, 3508-3510.

Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J. & Ehresmann, B. (1987). Probing the structure of RNAs in solution. *Nucl. Acids Res.* **15**, 9109-9128.

Felden, B., Himeno, H., Muto, A., McCutcheon, J. P., Atkins, J. F. & Gesteland, R. F. (1997). Probing the structure of the *Escherichia coli* 10Sa (tmRNA). *RNA,* **3**, 89-104.

Fink, T. R. & Crothers, D. M. (1972). Free energy of imperfect nucleic acid helices. I. The bulge defect. *J. Mol. Biol.* **66**, 1-12.

Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science,* **251**, 767-773.

Forrest, S. (1993). Genetic algorithms: principles of natural selection applied to computation. *Science,* **261**, 872-878.

Fotin, A. V., Drobyshev, A. L., Proudnikov, D. Y., Perov, A. N. & Mirzabekov, A. D. (1998). Parallel thermodynamic analysis of duplexes on oligodeoxyribonucleotide microchips. *Nucl. Acids Res.* **26**, 1515-1521.

Freier, S. M., Burger, B. J., Alkema, D., Neilson, T. & Turner, D. H. (1983). Effects of 3′ dangling end stacking on the stability of GGCC and CCGG double helices. *Biochemistry,* **22**, 6198-6206.

Freier, S. M., Sinclair, A., Neilson, T. & Turner, D. H. (1985). Improved free energies for G-C base-pairs. *J. Mol. Biol.* **185**, 645-647.

Freier, S. M., Kierzek, R., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986a). Free energy contributions of G·U and other terminal mismatches to helix stability. *Biochemistry,* **25**, 3209-3213.

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986b). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. USA,* **83**, 9373-9377.

Freier, S. M., Sugimoto, N., Sinclair, A., Alkema, D., Neilson, T., Kierzek, R., Caruthers, M. H. & Turner, D. H. (1986c). Stability of XGCGCp, GCGCYp, and XGCGCYp helixes: an empirical estimate of the energetics of hydrogen bonds in nucleic acids. *Biochemistry,* **25**, 3214-3219.

Gaspin, C. & Westhof, E. (1995). An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J. Mol. Biol.* **254**, 163-174.

Giese, M. R., Betschart, K., Dale, T., Riley, C. K., Rowan, C., Sprouse, K. J. & Serra, M. J. (1998). Stability of RNA hairpins closed by wobble base-pairs. *Biochemistry,* **37**, 1094-1100.

Groebe, D. R. & Uhlenbeck, O. C. (1988). Characterization of RNA hairpin loop stability. *Nucl. Acids Res.* **16**, 11725-11735.

Groebe, D. R. & Uhlenbeck, O. C. (1989). Thermal stability of RNA hairpins containing a four-membered loop and a bulge nucleotide. *Biochemistry,* **28**, 742-747.

Gultyaev, A. P., van Batenburg, F. H. D. & Pleij, C. W. A. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250**, 37-51.

Gutell, R. R. (1994). Collection of small subunit (16 S- and 16 S-like) ribosomal RNA structures. *Nucl. Acids Res.* **22**, 3502-3507.

Gutell, R. R., Gray, M. W. & Schnare, M. N. (1993). A compilation of large subunit (23 S- and 23 S-like) ribosomal RNA structures. *Nucl. Acids Res.* **21**, 3055-3074.

Harris, M. E., Kazantchev, A. V., Chen, J. L. & Pace, N. R. (1997). Analysis of the tertiary structure of the ribonuclease P ribozyme-substrate complex by photoaffinity crosslinking. *RNA,* **3**, 561-574.

He, L., Kierzek, R., SantaLucia, J., Jr, Walter, A. E. & Turner, D. H. (1991). Nearest-neighbor parameters for G·U mismatches. *Biochemistry,* **30**, 11124-11132.

Hilbers, C. W., Robillard, G. T., Shulman, R. G., Blake, R. D., Webb, P. K., Fresco, R. & Riesner, D. (1976). Thermal unfolding of yeast glycine transfer RNA. *Biochemistry,* **15**, 1874-1882.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167-188.

Holland, J. H. (1975). *Adaption in Natural and Artificial Systems*, University of Michigan Press.

Huynen, M., Gutell, R. & Konings, D. (1997). Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* **267**, 1104-1112.

Jabri, E., Aigner, S. & Cech, T. R. (1997). Kinetic and secondary structure analysis of *Naegleria anderson*: GIR1, a group I ribozyme whose putative biological function is site-specific hydrolysis. *Biochemistry,* **36**, 16345-16354.

Jacobson, H. & Stockmayer, W. H. (1950). Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **18**, 1600-1606.

Jaeger, J. A., Turner, D. H. & Zuker, M. (1989). Improved predictions of secondary structures for RNA. *Proc. Natl Acad. Sci. USA,* **86**, 7706-7710.

Jaeger, L., Westhof, E. & Michel, F. (1993). Monitoring of cooperative unfolding of the sunY group I intron of bacteriophage T4. *J. Mol. Biol.* **234**, 331-346.

Jaeger, L., Michel, F. & Westhof, E. (1994). Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J. Mol. Biol.* **236**, 1271-1276.

James, B. D., Olsen, G. J. & Pace, N. R. (1989). Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol.* **180**, 227-239.

Jucker, F. M. & Pardi, A. (1995). Solution structure of the CUUG hairpin loop - a novel RNA tetraloop motif. *Biochemistry,* **34**, 14416-14427.

Kean, J. M. & Draper, D. E. (1985). Secondary structure of a 345-base RNA fragment covering the S8/S15 protein binding domain of Escherichia coli 16 S ribosomal RNA. *Biochemistry,* **24**, 5052-5061.

Kim, J., Walter, A. E. & Turner, D. H. (1996). Thermodynamics of coaxially stacked helices with GA and CC mismatches. *Biochemistry,* **35**, 13753-13761.

Knapp, G. (1989). Enzymatic approaches to probing RNA secondary and tertiary structure. *Methods Enzymol.* **180**, 192-212.

Laing, L. G. & Draper, D. E. (1994). Thermodynamics of RNA folding in a conserved ribosomal RNA domain. *J. Mol. Biol.* **237**, 560-576.

Larsen, N., Samuelsson, T. & Zwieb, C. (1998). The signal recognition particle database (SRPDB). *Nucl. Acids Res.* **26**, 177-178.

Lehnert, V., Jaeger, L., Michel, F. & Westhof, E. (1996). New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the *Tetrahymena thermophila* ribozyme. *Chem. Biol.* **3**, 993-1009.

Longfellow, C. E., Kierzek, R. & Turner, D. H. (1990). Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry,* **29**, 278-285.

Lück, R., Steger, G. & Riesner, D. (1996). Thermodynamic prediction of conserved secondary structure: application to the RRE element of HIV, the tRNA-like element of CMV and the mRNA of prion protein. *J. Mol. Biol.* **258**, 813-826.

Massire, C., Jaeger, L. & Westhof, E. (1998). Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J. Mol. Biol.* **279**, 773-793.

Mathews, D. H., Banerjee, A. R., Luan, D. D., Eickbush, T. H. & Turner, D. H. (1997). Secondary structure model of the RNA recognized by the reverse transcriptase from the R2 retrotransposable element. *RNA,* **3**, 1-16.

Mathews, D. H., Andre, T. C., Kim, J., Turner, D. H. & Zuker, M. (1998). An updated recursive algorithm for RNA secondary structure prediction with improved thermodynamic parameters. In *Molecular Modeling of Nucleic Acids* (Leontis, N. B. & SantaLucia, J., Jr, eds), pp. 246-257, American Chemical Society, New York.

McCaskill, J. S. (1990). The equilibrium partition function and base-pair probabilities for RNA secondary structure. *Biopolymers,* **29**, 1105-1119.

McDowell, J. A. & Turner, D. H. (1996). Investigation of the structural basis for thermodynamic stabilities of tandem GU mismatches: solution structure of $(\mathrm{rGAG\underline{GU}CUC})_2$ by two-dimensional NMR and simulated annealing. *Biochemistry,* **35**, 14077-14089.

McDowell, J. A., He, L., Chen, X. & Turner, D. H. (1997). Investigation of the structural basis for thermodynamic stabilities of tandem GU wobble pairs: NMR structures of $(\mathrm{rGGA\underline{GU}UCC})_2$ and $(\mathrm{rGGA\underline{UG}UCC})_2$. *Biochemistry,* **36**, 8030-8038.

Merouch, M. & Chow, C. S. (1999). Thermodynamics of RNA hairpins containing single internal mismatches. *Nucl. Acids Res.* **27**, 1118-1125.

Michel, F. & Westhof, E. (1990). Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**, 585-610.

Michel, F., Umesono, K. & Ozeki, H. (1989). Comparative and functional anatomy of group II catalytic introns - a review. *Gene,* **82**, 5-30.

Morse, S. E. & Draper, D. E. (1995). Purine-purine mismatches in RNA helixes: evidence for protonated G·A pairs and next-nearest neighbor effects. *Nucl. Acids Res.* **23**, 302-306.

Murphy, F. L. & Cech, T. R. (1994). GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J. Mol. Biol.* **236**, 49-63.

O'Donnell-Maloney, M. J., Smith, C. L. & Cantor, C. R. (1996). The development of microfabricated arrays for DNA sequencing and analysis. *Trends Biotechnol.* **14**, 401-407.

Pace, N. R., Thomas, B. C. & Woese, C. R. (1999). Probing RNA structure, function, and history by comparative analysis. In *The RNA World* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds), 2nd edit., pp. 113-141, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Peritz, A. E., Kierzek, R., Sugimoto, N. & Turner, D. H. (1991). Thermodynamic study of internal loops in oligoribonucleotides: Symmetric loops are more stable than asymmetric loops. *Biochemistry,* **30**, 6428-6436.

Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature,* **372**, 111-114.

SantaLucia, J., Jr, Kierzek, R. & Turner, D. H. (1991a). Functional group substitutions as probes of hydrogen bonding between GA mismatches in RNA internal loops. *J. Am. Chem. Soc.* **113**, 4313-4322.

SantaLucia, J., Jr, Kierzek, R. & Turner, D. H. (1991b). Stabilities of consecutive A·C, C·C, G·G, U·C, and U·U mismatches in RNA internal loops: evidence for stable hydrogen-bonded U·U and C·C+ pairs. *Biochemistry,* **30**, 8242-8251.

Schnare, M. N., Damberger, S. H., Gray, M. W. & Gutell, R. R. (1996). Comprehensive comparison of structural characteristics in Eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J. Mol. Biol.* **256**, 701-719.

Schroeder, S., Kim, J. & Turner, D. H. (1996). G·A and U·U mismatches can stabilize RNA internal loops of three nucleotides. *Biochemistry,* **35**, 16105-16109.

Serra, M. J. & Turner, D. H. (1995). Predicting thermodynamic properties of RNA. *Methods Enzymol.* **259**, 242-261.

Serra, M. J., Lyttle, M. H., Axenson, T. J., Schadt, C. A. & Turner, D. H. (1993). RNA hairpin loop stability depends on closing pair. *Nucl. Acids Res.* **21**, 3845-3849.

Serra, M. J., Axenson, T. J. & Turner, D. H. (1994). A model for the stabilities of RNA hairpins based on a study of the sequence dependence of stability for hairpins of six nucleotides. *Biochemistry,* **33**, 14289-14296.

Serra, M. J., Barnes, T. W., Betschart, K., Gutierrez, M. J., Sprouse, K. J., Riley, C. K., Stewart, L. & Temel, R. E. (1997). Improved parameters for the prediction of RNA hairpin stability. *Biochemistry,* **36**, 4844-4851.

Speek, M. & Lind, A. (1982). Structural analyses of *E. coli* 5S RNA fragments, their associates and complexes with proteins L18 and L25. *Nucl. Acids Res.* **10**, 947-965.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.* **26**, 148-153.

Sugimoto, N., Kierzek, R., Freier, S. M. & Turner, D. H. (1986). Energetics of internal GU mismatches in ribooligonucleotide helixes. *Biochemistry,* **25**, 5755-5759.

Szymanski, M., Specht, T., Barciszewska, M. Z., Barciszewski, J. & Erdmann, V. A. (1998). 5S rRNA data bank. *Nucl. Acids Res.* **26**, 156-159.

Tranguch, A. J. & Engelke, D. R. (1993). Comparative structural-analysis of nuclear RNase-P RNAs from yeast. *J. Biol. Chem.* **268**, 14045-14055.

Tranguch, A. J., Kinderberger, D. W., Rohlman, C. E., Lee, J. & Engelke, D. R. (1994). Structure-sensitive RNA footprinting of yeast nuclear ribonuclease P. *Biochemistry,* **33**, 1778-1787.

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

Van Batenburg, F. H. D., Gultyaev, A. P. & Pleij, C. W. A. (1995). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **174**, 269-280.

Varani, G., Cheong, C. & Tinoco, I., Jr (1991). Structure of an unusually stable RNA hairpin. *Biochemistry,* **30**, 3280-3289.

Walter, A. E., Turner, D. H., Kim, J., Lyttle, M. H., Müller, P., Mathews, D. H. & Zuker, M. (1994a). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA,* **91**, 9218-9222.

Walter, A. E. & Turner, D. H. (1994b). Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. *Biochemistry,* **33**, 12715-12719.

Walter, A. E., Wu, M. & Turner, D. H. (1994c). The stability and structure of tandem GA mismatches in RNA depend on closing base-pairs. *Biochemistry,* **33**, 11349-11354.

Waring, R. B. & Davies, R. W. (1984). Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing - a review. *Gene.* **28**, 277-291.

Weeks, K. M. & Crothers, D. M. (1993). Major groove accessibility of RNA. *Science,* **261**, 1574-1577.

Williams, K. P. & Bartel, D. P. (1996). Phylogenetic analysis of tmRNA secondary structure. *RNA,* **2**, 1306-1310.

Woese, C. R., Gutell, R. R., Gupta, R. & Noller, H. F. (1983). Detailed analysis of the higher order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.* **47**, 621-669.

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

Wu, M., McDowell, J. A. & Turner, D. H. (1995). A periodic table of symmetric tandem mismatches in RNA. *Biochemistry,* **34**, 3204-3211.

Xia, T., McDowell, J. A. & Turner, D. H. (1997). Thermodynamics of nonsymmetric tandem mismatches adjacent to G·C base pairs in RNA. *Biochemistry,* **36**, 12486-12487.

Xia, T., SantaLucia, J., Jr, Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C. & Turner, D. H. (1998). Parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry,* **37**, 14719-14735.

Zhu, J. & Wartell, R. M. (1997). The relative stabilities of base-pair stacking interactions and single mismatches in long RNA measured by temperature gradient gel electrophoresis. *Biochemistry,* **36**, 15326-15335.

Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science,* **244**, 48-52.

Zuker, M. & Jacobson, A. B. (1995). ''Well-determined'' regions in RNA secondary structure predictions. Applications to small and large subunit rRNA. *Nucl. Acids Res.* **23**, 2791-2798.

Zuker, M. & Jacobson, A. B. (1998). Using reliability information to annotate RNA secondary structures. *RNA,* **4**, 669-679.

Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9**, 133-148.

Edited by I. Tinoco

(Received 29 October 1998; received in revised form 9 March 1999; accepted 9 March 1999)

JMB Online

http://www.academicpress.com/jmb

Supplementary material comprising eight Tables is available from JMB Online.