Thermodynamic Characterization of RNA Duplexes Containing Naturally Occurring 1 × 2 Nucleotide Internal Loops†

Jaya Badhwar, Saradasri Karri, Cody K. Cass, Erica L. Wunderlich, and Brent M. Znosko*

Department of Chemistry, Saint Louis University, Saint Louis, Missouri 63103

Received May 25, 2007; Revised Manuscript Received August 7, 2007

ABSTRACT: Thermodynamic data for RNA 1 × 2 nucleotide internal loops are lacking. Thermodynamic data that are available for 1 × 2 loops, however, are for loops that rarely occur in nature. In order to identify the most frequently occurring 1 × 2 nucleotide internal loops, a database of 955 RNA secondary structures was compiled and searched. Twenty-four RNA duplexes containing the most common 1 × 2 nucleotide loops were optically melted, and the thermodynamic parameters $\Delta H°$, $\Delta S°$, $\Delta G°_{37}$, and T_M for each duplex were determined. This data set more than doubles the number of 1 × 2 nucleotide loops previously studied. A table of experimental free energy contributions for frequently occurring 1 × 2 nucleotide loops (as opposed to a predictive model) is likely to result in better prediction of RNA secondary structure from sequence. In order to improve free energy calculations for duplexes containing 1 × 2 nucleotide loops that do not have experimental free energy contributions, the data collected here were combined with data from 21 previously studied 1 × 2 loops. Using linear regression, the entire dataset was used to derive nearest neighbor parameters that can be used to predict the thermodynamics of previously unmeasured 1 × 2 nucleotide loops. The $\Delta G°_{37,loop}$ and $\Delta H°_{loop}$ nearest neighbor parameters derived here were compared to values that were published previously for 1 × 2 nucleotide loops but were derived from either a significantly smaller dataset of 1 × 2 nucleotide loops or from internal loops of various sizes [Lu, Z. J., Turner, D. H., and Mathews, D. H. (2006) *Nucleic Acids Res. 34*, 4912−4924]. Most of these values were found to be within experimental error, suggesting that previous approximations and assumptions associated with the derivation of those nearest neighbor parameters were valid. $\Delta S°_{loop}$ nearest neighbor parameters are also reported for 1 × 2 nucleotide loops. Both the experimental thermodynamics and the nearest neighbor parameters reported here can be used to improve secondary structure prediction from sequence.

Sequencing projects, such as the Human Genome Project, are generating sequence information at a rate greater than a million nucleotides per day. While sequences of many important RNAs have been determined, little is known about structure−function relationships of RNA, primarily due to the lack of definitive secondary and tertiary structural information about RNA. X-ray crystallography and NMR methods are providing an increasing number of three-dimensional RNA structures, but it is unlikely that these methods will keep pace with the rate at which interesting sequences are being discovered. Thus, there is a need for reliable, rapid methods to predict secondary and tertiary structures of RNA. Being able to predict secondary and tertiary structures of RNA provides a foundation for determining structure−function relationships of RNA and for targeting RNA with therapeutics.

† Partial funding for this project was provided by the St. Louis University College of Arts and Sciences, St. Louis University Department of Chemistry, a St. Louis University Summer Research Award (B.M.Z.), the St. Louis University Faculty Development Fund (B.M.Z.), two Sigma Xi Grants-in-Aide of Research (J.B. and S.K.), and the Students and Teachers as Research Scientists (STARS) Program (E.L.W.).

* To whom correspondence should be addressed. Phone: (314) 977-8567. Fax: (314) 977-2521. E-mail: znoskob@slu.edu.

The cornerstone for modeling RNA structure is an accurate secondary structure model. The most popular method for modeling secondary structure is free energy minimization (*1*−*6*). In this method, each possible secondary structure motif is given a free energy parameter, and these values are added to predict the total free energy of forming a particular secondary structure. The secondary structure with the lowest free energy is predicted to be the predominant species in solution. An improvement of the thermodynamic parameters for a particular motif may result in an improvement in the accuracy of the predicted secondary structure.

The main structural motif of nucleic acids is the double helix with Watson−Crick base pairs. Thus, an understanding of principles governing double helix formation is essential for understanding and predicting properties of nucleic acids. Thermodynamics of Watson−Crick pairs have been extensively studied and are well characterized (*7*). However, since noncanonical regions (bulges, internal loops, hairpins, and multibranch loops) are also prevalent in RNA secondary structures, an understanding of the properties of these motifs is also necessary. Presently, the lack of experimental parameters for non-Watson−Crick regions is a major limitation of the current algorithms used to predict RNA secondary structure from sequence (*1*−*4*).

10.1021/bi701024w CCC: $37.00
Published on Web 11/20/2007

One common non-Watson−Crick motif is a 1 × 2 nucleotide internal loop. A 1 × 2 internal loop forms when one nucleotide in one strand opposes two nucleotides in a second strand, where the one nucleotide cannot form a canonical pair with either of the two opposing nucleotides. Not only do 1 × 2 loops occur frequently in nature (discussed below), but the loops contain nucleotides with functional groups available to serve as binding sites for proteins or drugs (e.g., aminoglycosides (*8*)) or for tertiary interactions, demonstrating a structural and functional role in biology.

Differences in internal loop stability arise from differences in the nucleotide sequence of the loop and differences in the base pairs adjacent to the loop. Considering only the three nucleotides in the loop, there are 26 possible 1 × 2 loops. Previous studies (*9*−*11*) have investigated 16 of these loops. Considering just the combinations of canonical base pairs closing the loop, there are 36 nearest neighbor combinations. Previous studies (*9*−*11*) have investigated only two (6%) of these combinations. Considering both nucleotides in the loop and nearest neighbors, there are 936 possible 1 × 2 loops, of which 21 (or 2%) have been previously studied (*9*−*11*). Evidently, there is a lack of experimental data for 1 × 2 loops.

RNAstructure (*1*−*3*) and *mfold* (*4*) are computer programs used to predict RNA secondary structure from sequence. *RNAstructure*, for example, uses two methods to calculate the free energy contribution of a 1 × 2 loop. If a loop has been measured thermodynamically, then the measured value is used. Because only 2% of all possible 1 × 2 loops have been previously measured, this method is rarely used. If a loop has not been measured thermodynamically, then *RNAstructure* approximates the free energy contribution by (*1*)

$$\Delta G^\circ_{37,\text{loop}} = \Delta G^\circ_{37,\text{loop initiation}} + \Delta G^\circ_{37,\text{AU/GU}} + \Delta G^\circ_{37,\text{asym}} + \Delta G^\circ_{37,\text{first noncanonical pairs}} \quad (1)$$

Here, $\Delta G^\circ_{37,\text{loop initiation}}$ is the free energy of initiation for a 1 × 2 loop (1.7 kcal/mol), $\Delta G^\circ_{37,\text{AU/GU}}$ is a penalty for replacing a closing G-C pair with an A-U or G-U pair (0.7 kcal/mol), $\Delta G^\circ_{37,\text{asym}}$ is a penalty for having an unequal number of nucleotides on each side of the loop (0.6 kcal/mol), and $\Delta G^\circ_{37,\text{first noncanonical pairs}}$ is a bonus for stabilizing mismatches in the loop (−1.2 kcal/mol for all G·G and most G·A pairs and −0.8 kcal/mol for U·U pairs). $\Delta G^\circ_{37,\text{loop initiation}}$ and $\Delta G^\circ_{37,\text{first noncanonical pairs}}$ were derived from the published dataset of 1 × 2 loops measured previously (*1*, *9*−*11*), while $\Delta G^\circ_{37,\text{AU/GU}}$ and $\Delta G^\circ_{37,\text{asym}}$ were derived from data of loops of various sizes (*1*). We have studied an additional 24 duplexes with 1 × 2 internal loops to (1) provide experimental free energy contributions for more types of 1 × 2 loops, especially those that occur frequently in nature, that can be utilized by algorithms such as *RNAstructure*; (2) determine if there is a better model than eq 1 to approximate the stability of previously unmeasured 1 × 2 loops; and (3) if eq 1 is the best model, refine values for $\Delta G^\circ_{37,\text{loop initiation}}$ and $\Delta G^\circ_{37,\text{first noncanonical pairs}}$ and derive 1 × 2 loop-specific $\Delta G^\circ_{37,\text{AU/GU}}$ and $\Delta G^\circ_{37,\text{asym}}$ values.

Although it is unrealistic to perform optical melting experiments with all 936 types of 1 × 2 loops, it is not unrealistic to optically melt the most frequently occurring 1 × 2 loops in known secondary structures. Therefore, a database of 955 secondary structures was compiled, and the type and frequency of 1 × 2 loops were tabulated. Here, thermodynamics for 24 frequently occurring 1 × 2 loops are reported.

MATERIALS AND METHODS

Compiling and Searching a Database for 1 × 2 Loops. A database of secondary structures that includes 151,503 nucleotides and 43,519 base pairs consisting of 22 small subunit rRNAs (*12*), 5 large subunit rRNAs (*13*, *14*), 309 5S rRNAs (*15*), 484 tRNAs (*16*), 91 signal recognition particles (*17*), 16 RNase P RNAs (*18*), 25 group I introns (*19*, *20*), and 3 group II introns (*21*), whose structures were determined by sequence comparison, was assembled. The database was searched for 1 × 2 loops, and the number of occurrence for each type of loop was tabulated. In this study, G-U pairs were considered canonical pairs. For example, $\binom{\text{GG C}}{\text{CUAG}}$ is considered a single adenosine bulge with a 3′ U-G nearest neighbor (not a 1 × 2 loop). Similarly, $\binom{\text{G G A C}}{\text{C U AA G}}$ is considered an all adenosine 1 × 2 loop with a 5′ G-U nearest neighbor (not a 2 × 3 internal loop).

Design of Sequences for Optical Studies. Duplexes containing 1 × 2 loops were designed to have melting temperatures between 35 and 55 °C and to have minimal formation of hairpin structures or misaligned duplexes. Terminal G-C pairs were chosen to prevent end fraying during melting experiments. The sequences of loops and nearest neighbors were chosen to reflect the most frequently occurring loops in the database described above. Loops were placed directly in the center of the duplex.

RNA Synthesis and Purification. Some oligonucleotides were synthesized on CPG support with an Applied Biosystems 392 DNA/RNA synthesizer at the University of Rochester (Rochester, NY). Some oligonucleotides were ordered from the Keck Lab at Yale University (New Haven, CT) or from Azco BioTech, Inc. (San Diego, CA). The synthesis and purification of the oligonucleotides followed standard procedures and were described previously (*22*).

Concentration Calculations and Duplex Formation. Using Beer's Law, the total concentration of each single strand was calculated from the high temperature absorbance at 280 nm and the extinction coefficient. Samples were diluted so that the absorbance was between 0.2 and 2.0. The absorbance was measured at 80 °C to disrupt any single-strand folding. Extinction coefficients of single strands were calculated using *RNACalc* (*23*). Individual single strand concentrations were used to mix equal molar amounts of non-self-complementary strands to form a duplex containing a 1 × 2 internal loop.

Optical Melting Experiments. Newly formed duplexes were lyophilized and redissolved in 1 M NaCl, 20 mM sodium cacodylate, and 0.5 mM Na_2EDTA, pH = 7.0. A melt scheme that consisted of a three dilution series was designed, resulting in nine samples to allow for a concentration range typically >50-fold. Using a heating rate of 1 °C/min on a Beckman-Coulter DU800 spectrophotometer with a Beckman-Coulter high performance temperature controller, curves of absorbance at 280 nm versus temperature were obtained.

Determination of Thermodynamic Parameters for Duplexes. Meltwin (*24*) was used to fit melting curves to a two-state model, assuming linear sloping baselines and temperature-independent ΔH° and ΔS° values (*25*, *26*). Additionally, T_M

values at different concentrations were used to calculate thermodynamic parameters according to Borer et al. (*27*):

$$T_M^{-1} = (2.303R/\Delta H^\circ)\log(C_T/4) + (\Delta S^\circ/\Delta H^\circ) \quad (2)$$

where C_T is total strand concentration.[1] For transitions that conform to the two-state model, ΔH° values from the two methods generally agree within 15%. The Gibb's free energy change at 37 °C was calculated as

$$\Delta G^\circ_{37} = \Delta H^\circ - (310.15\ \text{K})\Delta S^\circ \quad (3)$$

Determination of the Contribution of 1 × 2 Loops to Duplex Thermodynamics. The total free energy change for duplex formation can be approximated by a nearest neighbor model (*7*) that is the sum of energy increments for helix initiation, nearest neighbor interactions between base pairs, and the loop contribution. For example:

$$\Delta G^\circ_{37}\begin{pmatrix}\text{GACG A ACUG}\\ \text{CUGCCAUGAC}\end{pmatrix} = \Delta G^\circ_{37,i} + \Delta G^\circ_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} + \Delta G^\circ_{37,\text{loop}} + \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix} \quad (4)$$

where $\Delta G^\circ_{37,i}$ is the free energy change for duplex initiation, 4.09 kcal/mol (*7*), $\Delta G^\circ_{37,\text{loop}}$ is the free energy contribution from the 1 × 2 loop, and the remainder of the terms are individual nearest neighbor values (*7*). Therefore, rearranging eq 4 can solve for the contribution of the loop to duplex stability:

$$\Delta G^\circ_{37,\text{loop}} = \Delta G^\circ_{37}\begin{pmatrix}\text{GACG A ACUG}\\ \text{CUGCCAUGAC}\end{pmatrix} - \Delta G^\circ_{37,i} - \Delta G^\circ_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} - \Delta G^\circ_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix} \quad (5)$$

Here, $\Delta G^\circ_{37}\left(\begin{smallmatrix}\text{GACG A ACUG}\\ \text{CUGCCAUGAC}\end{smallmatrix}\right)$ is the value determined by optical melting experiments. More specifically:

$$\Delta G^\circ_{37,\text{loop}} = -6.83 - 4.09 - (-2.35) - (-2.24) - (-2.36) - (-2.24) - (-2.08) - (-2.11) = 2.46\ \text{kcal/mol} \quad (6)$$

Although previous studies (*9*−*11*) have used reference duplexes to calculate $\Delta G^\circ_{37,\text{loop}}$, all $\Delta G^\circ_{37,\text{loop}}$ values for loops studied here have been calculated as described above. Since most of the sequences studied here contain different nearest neighbor combinations, a reference duplex would have been required for each loop studied. Therefore, for simplicity, nearest neighbor parameters were used instead of reference duplexes. Similar calculations were done for $\Delta H^\circ_{\text{loop}}$ and $\Delta S^\circ_{\text{loop}}$.

Linear Regression and Internal Loop Thermodynamic Parameters. Data collected here were combined with previously published data on 1 × 2 loops (*9*−*11*). The melting of two duplexes, $\left(\begin{smallmatrix}\text{GACU A GCUG}\\ \text{CUGGGGUGAC}\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}\text{GACC A ACUG}\\ \text{CUGGAGUGAC}\end{smallmatrix}\right)$, was considered non-two-state. The data derived from the individual melting curves and from the T_M versus C_T plot did not agree, and/or the errors associated with the data were large. As a result, the data from these two duplexes were omitted from linear regression. Data from another sequence, $\left(\begin{smallmatrix}\text{GGCC C CGCC}\\ \text{CCGGAUGCGG}\end{smallmatrix}\right)$, was also omitted from linear regression. This loop was 1.7 kcal/mol less stable than any other loop and was 2.8 kcal/mol less stable than the same loop located within a different stem sequence. Two other sequences were omitted from linear regression, $\left(\begin{smallmatrix}\text{GGCG A CGCC}\\ \text{CCGUGCGCGG}\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}\text{GACG A CCUG}\\ \text{CUGCCGGGAC}\end{smallmatrix}\right)$. The thermodynamic contributions calculated for these loops were significantly different from what was expected. After further investigation, it appears as if a bimolecular association of one of the strands with itself may be competing with the bimolecular association of the two separate strands. Since the identity of the duplex(es) in solution could not be confirmed with certainty, the data for these two sequences were omitted from the analysis by linear regression. Data from all five of these sequences were omitted from the analysis by linear regression as well as any trends or averages discussed in the remainder of this work. As a result, data from 40 sequences (21 studied previously and 19 studied here) were used for linear regression. A term for 1 × 2 loop initiation, an A-U or G-U closure, a G·A (except for 5′RA/3′YG) or G·G mismatch, and a U·U mismatch were used as variables. Other combinations of bonuses and penalties were tried, but the combinations of variables listed here (and the same as those in eq 1) resulted in a model that agreed best with the experimental data. The loop thermodynamic values were used as the constants. Excel's LINEST function used linear regression to solve for each variable. Internal loop thermodynamic parameters were derived for $\Delta G^\circ_{37,\text{loop}}$, $\Delta H^\circ_{\text{loop}}$, and $\Delta S^\circ_{\text{loop}}$.

RESULTS

Database Searching. The database containing 955 structures and 151,503 nucleotides (described in Materials and Methods) was searched for 1 × 2 loops. In this database, 817 1 × 2 loops are found, averaging 0.86 1 × 2 loops per structure and one 1 × 2 loop per 185 nucleotides. However, no 1 × 2 loops are found in published tRNA secondary structures, therefore, these averages are higher for non-tRNA sequences. Non-tRNA sequences average 1.7 1 × 2 loops per structure.

A summary of database results is shown in Table 1. The first set of data lists frequency and percent occurrence when the loop nucleotides and nearest neighbors are specified. Categorizing 1 × 2 loops in this fashion results in 189 types of loops in the database. The 25 loop types listed in the first data set (Table 1) account for 61% of the total number of loops found. The 164 types of loops not shown account for the remaining 39%, however, each type represents <0.8% of the total number of loops found. When categorized in this manner, previous studies account for only 20% of the total number of loops found, but after adding the data reported here, this percentage increases to 64%. Similarly, previous studies thermodynamically characterized only six types of loops in the top 25, but after adding the data reported here, all of the loops in the top 25 have been studied.

The second set of data (Table 1) lists frequency and percent occurrence when only the loop sequence is specified.

[1] Abbreviations: C_T, total strand concentration; R, purine nucleotides (adenosine and guanosine); Y, pyrimidine nucleotides (uridine and cytidine).

Table 1: Summary of Database Search for 1 × 2 Loops[a]

loop with nearest neighbors						loop				5′ and 3′ adjacent base pairs				loop nucleotides classified as purine or pyrimidine			
	loop		freq[b]	%[c]	ref[d]	loop	freq[b]	%[c]	ref[d]	closing bp	freq[b]	%[c]	ref[d]	loop	freq[b]	%[c]	ref[d]
C G	A AA	C G	50	6.1	*e*, *f*	A AA	140	17.1	*e*, *f*, *h*	C C G G	181	22.2	*e*−*h*	R RR	455	55.7	*e*−*h*
U G	A GG	G U	38	4.7	*h*	A AG	112	13.7	*e*, *g*, *h*	C G G U	74	9.1	*h*	Y YY	125	15.3	*e*, *g*, *h*
G C	A CA	A U	32	3.9	*h*	A GG	91	11.1	*e*, *h*	C A G U	64	7.8	*h*	R YR	74	9.1	*h*
G C	U UC	U A	30	3.7	*h*	C AU	60	7.3	*e*, *h*	U G G U	47	5.8	*h*	Y RY	64	7.8	*e*, *h*
G U	A GC	C G	29	3.5	*h*	G AA	58	7.1	*e*, *h*	G U C A	41	5.0	*h*	R RY	50	6.1	*h*
C G	G AA	C G	29	3.5	*e*, *h*	A CA	54	6.6	*h*	G A C U	40	4.9	*h*	Y YR	21	2.6	
C G	A AA	A U	28	3.4	*h*	U UC	40	4.9	*e*, *h*	A G U U	38	4.7	*h*	Y RR	15	1.8	*e*
A U	C AU	G C	28	3.4	*h*	A GC	36	4.4	*h*	A C U G	37	4.5	*h*	R YY	13	1.6	*e*
C G	A AA	G U	27	3.3	*h*	U UU	36	4.4	*e*, *g*, *h*	G C U G	37	4.5	*h*	previously new total	672 796	82.3 97.4	
A U	A AG	G U	27	3.3	*h*	G GA	24	2.9	*e*, *h*	A G U C	32	3.9	*h*				
C G	C AU	C G	24	2.9	*e*, *h*	C UU	22	2.7		A A U U	29	3.5					
C G	A GG	G U	21	2.6	*h*	A CG	20	2.4	*h*	C G G C	29	3.5	*h*				
A U	G AA	C G	21	2.6	*h*	C AA	15	1.8	*e*	G C C G	22	2.7	*h*				
C G	G GA	C G	20	2.4	*e*, *h*	A AC	14	1.7		U G A C	19	2.3	*h*				
U A	A AG	G C	15	1.8	*h*	A CC	13	1.6	*e*	U G G C	16	2.0					
C G	A AG	A U	10	1.2	*h*	C UA	12	1.5		U C A G	15	1.8					
C G	U UU	C G	10	1.2	*e*, *g*, *h*	A GA	11	1.3	*e*	U A A U	14	1.7	*h*				
C G	A AG	C G	9	1.1	*e*, *h*	G AG	11	1.3	*e*	U U A A	13	1.6	*h*				
C G	A AG	G C	9	1.1	*h*	C CA	9	1.1		G G C C	12	1.5	*e*−*g*				
C G	A AG	G U	8	1.0	*h*	G GG	8	1.0		G A U U	11	1.3					
U A	A AG	U A	8	1.0	*h*	U CU	8	1.0	*e*, *g*	C U G A	10	1.2					
C G	U UU	G U	8	1.0	*h*	C CC	6	0.7	*g*	G G U C	6	0.7					
U A	A AA	A U	7	0.9	*h*	C CU	6	0.7	*g*	U A G U	6	0.7					
G C	A CG	C G	7	0.9	*h*	U CC	5	0.6	*g*	G G C U	5	0.6					
C G	A GG	G C	7	0.9	*h*	C AC	4	0.5		G U U A	4	0.5					
previously[i] new total[j]			163 523	20.0 64.0		previously[i] new total[j]	636 746	77.8 91.3		previously[i] new total[j]	193 700	23.6 85.7					

[a] As described in the text, not all combinations are shown due to space limitations. [b] Frequency of occurrence in database. [c] Percent out of 817 loops, the total number of loops found in the database. [d] Reference where data are reported. [e] Reference (*9*). [f] Reference (*11*). [g] Reference (*10*). [h] This work. [i] The total number of 1 × 2 loops accounted for and the corresponding percentage of the total number of loops found in the database by previous studies. [j] The total number of 1 × 2 loops accounted for and the corresponding percentage of the total number of loops found in the database when the data from previous studies were combined with the data reported here.

Categorizing 1 × 2 loops in this fashion results in 26 types of loops in the database, representing all possible types of 1 × 2 loops. The 25 loop types listed in the second data set account for 99.8% of the total number of loops found. The one type of loop that is not shown, $\binom{C}{UC}$, accounts for the remaining 0.2%. When categorized in this manner, previous studies account for 78% of the total number of loops found, and after adding the data reported here, this percentage increases to 91%. Similarly, previous studies thermodynamically characterized 16 types of loops in the top 25, and after

adding the data reported here, 19 types of loops in the top 25 have been studied.

The third set of data (Table 1) lists frequency and percent occurrence of 5′ and 3′ nearest neighbor combinations. Categorizing 1 × 2 loops in this fashion results in 30 types of nearest neighbor combinations in the database. It is interesting to note that six possible nearest neighbor combinations were not represented in this database, all of which involve a G-U pair. The 25 nearest neighbor combinations listed in the third data set account for 98% of the total number of combinations found. The five nearest neighbor combinations not shown account for the remaining 2%. When categorized in this manner, previous studies account for only 24% of the total number of loops found, but after adding the data reported here, this percent increases to 86%. Similarly, previous studies thermodynamically characterized only two nearest neighbor combinations in the top 25, but after adding the data reported here, 16 nearest neighbor combinations in the top 25 have been studied.

The last set of data (Table 1) lists frequency and percent occurrence of the loop sequence when adenosine and guanosine are classified as purines (R) and uridine and cytidine are classified as pyrimidines (Y). Categorizing 1 × 2 loops in this fashion results in eight types of loops in the database, representing all possible 1 × 2 combinations. When categorized in this manner, previous studies account for 82% of the total number of loops found, and after adding the data reported here, this percent increases to 97%. Similarly, previous studies thermodynamically characterized five of the eight types of loops, and after adding the data reported here, seven of the eight types of loops have been studied.

Thermodynamic Parameters. Thermodynamic parameters of duplex formation derived from T_M^{-1} vs log C_T and from fitting the shape of each melting curve to the two-state model are listed in Table 2. The data shown in Table 2 correspond with the 25 most frequent loops found in the database (Table 1, first set of data) and are listed in order of decreasing frequency. Although Table 2 corresponds to the 25 most frequent loops, data for 32 duplexes are shown because six loops were melted with different stems.

Contribution of 1 × 2 Loops to Duplex Thermodynamics. The contribution of 1 × 2 loops to duplex thermodynamics is described in Materials and Methods and is defined by eqs 5 and 6 for $\Delta G^\circ_{37,loop}$. These contributions are listed in Table 3. In addition to the 32 duplexes listed in Table 2, 13 additional duplexes that occur less frequently but have been studied previously (*9*−*11*) are also listed.

1 × 2 Loop Free Energy Parameters. Using linear regression with the data listed in Table 3, several models to calculate the thermodynamics of 1 × 2 loops were tested. The model described in eq 1 resulted in the $\Delta G^\circ_{37,loop}$ parameters (Table 4) that were closest to the experimental data reported here. All of the new $\Delta G^\circ_{37,loop}$ parameters are within experimental error of the *RNAstructure* values (*1*). Similar calculations and a similar comparison were done for ΔH°_{loop} parameters (Table 4). All of the new parameters except for the A-U or G-U closure parameter are within experimental error of literature values (*1*). Parameters were also derived for ΔS°_{loop} (Table 4).

DISCUSSION

RNA sequencing projects are generating an abundance of sequence information. In order to learn more about structure−function relationships of RNA and RNA−RNA, RNA−DNA, RNA−protein, and RNA−drug interactions and to improve rational drug design, many scientists are interested in secondary and tertiary structures of RNA. Computer algorithms use thermodynamic parameters to predict secondary structure from sequence (*1*−*4*). Because these algorithms rely on thermodynamic parameters for every type of motif (Watson−Crick pairs, bulges, internal loops, hairpins, etc.), the accuracy of predicted secondary structures depends on the accuracy of the thermodynamic parameters for each motif. Although 1 × 2 nucleotide loops are common in nature, relatively few studies have investigated their thermodynamics (*9*−*11*), resulting in thermodynamic parameters that rely on several approximations and assumptions. Additionally, the 1 × 2 loops studied previously (*9*−*11*) rarely occur in known secondary structures (Table 1). As a result, stabilities of many commonly occurring loops are based on assumptions and approximations.

In this study, the 25 most frequently occurring 1 × 2 loops have been thermodynamically characterized in an attempt to improve the current model used to predict the stability of RNA duplexes containing 1 × 2 loops, and therefore, improve secondary structure prediction from sequence.

Database Searching. The database compiled for this study contains 955 structures from eight different kinds of RNAs. We have assumed that the number and variety of structures in this database approximates the number and types of 1 × 2 loops found in nature. It is important to note, however, that the results found from searching this database may be slightly skewed. For example, the most prevalent type of RNA found in this database was tRNA, accounting for roughly half of the structures in the database. However, no 1 × 2 loops were found in any of the tRNA secondary structures. The second largest type of RNA in the database is 5S rRNA, with over two hundred more secondary structures than the next most prevalent type of RNA. Therefore, 1 × 2 loops prevalent in 5S rRNA may be over-represented, and 1 × 2 loops rarely found in 5S rRNA may be under-represented. However, the database is large enough and diverse enough to contain 196 loop and nearest neighbor combinations and all 26 possible loop combinations when considering the three nucleotides in the loop.

It is clear from the first set of data in Table 1 that the pioneering experiments on 1 × 2 loops (*9*−*11*) provided results for only 6 of the 25 1 × 2 loop sequences found most commonly in the database. The results reported here expand the database to include all combinations in the top 25. The first set of data in Table 1 provides some interesting results. The most frequently occurring loop sequence, $\binom{C\ A\ C}{G\ A\ A\ G}$, and four others in the top ten do not contain mismatches that have been considered stabilizing (G·G, A·G, or U·U) (*3*, *9*). Clearly, the presence or absence of stabilizing mismatches does not determine the frequency of occurrence for 1 × 2 loops.

The second set of data in Table 1 shows that all 26 possible loop sequences (25 shown and 1 not shown) were found in this database. Although this database only contains 955 secondary structures and may be biased toward tRNA and

Table 2: Thermodynamic Parameters for Duplex Formation[a]

		analysis of melt curve fit/errors				analysis of T_m dependence/errors (ln plot)			
freq[b]	sequence	$-\Delta H°$ (kcal/mol)	$-\Delta S°$ (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)	$-\Delta H°$ (kcal/mol)	$-\Delta S°$ (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)
50	UGAC **A** CUCA[d] ACUG**AA**GAGU	57.3 ± 4.9	162.5 ± 15.9	6.86 ± 0.08	38.8	62.3 ± 0.6	178.8 ± 2.0	6.80 ± 0.01	38.4
	UGAC **A** CUCA[e] ACUG**AA**GAGU	46.1	127.3	6.61	37.6	58.0	166.2	6.42	36.4
38	GACU **A** GCUG CUGG**GG**UGAC	(58.4)	(168.0)	(6.32)	(35.9)	(72.4)	(214.1)	(6.03)	(34.9)
32	GACG **A** ACUG CUGC**CA**UGAC	76.4 ± 12.7	224.3 ± 41.4	6.85 ± 0.25	38.3	77.4 ± 10.9	227.7 ± 35.3	6.81 ± 0.32	38.1
30	GACG **U** UCUG CUGC**UC**AGAC	64.3 ± 8.7	185.3 ± 27.3	6.79 ± 0.26	38.3	69.3 ± 12.4	201.6 ± 39.7	6.75 ± 0.53	38.0
29	GGCG **A** CGCC[f] CCGU**GC**GCGG	79.0 ± 12.0	209.1 ± 36.8	14.12 ± 0.67	70.0	76.9 ± 19.7	202.8 ± 57.6	13.99 ± 2.108	70.3
29	GACC **G** CCUG CUGG**AA**GGAC	54.5 ± 5.6	147.5 ± 17.5	8.72 ± 0.18	50.0	55.4 ± 11.0	150.4 ± 34.0	8.75 ± 0.68	49.9
	UGAC **G** CUCA[d] ACUG**AA**GAGU	63.8 ± 0.1	182.2 ± 0.2	7.34 ± 0.01	41.0	64.3 ± 0.3	183.6 ± 0.9	7.35 ± 0.01	41.0
28	GACC **A** ACUG CUGG**AA**UGAC	65.5 ± 3.8	187.5 ± 12.0	7.38 ± 0.24	41.1	78.0 ± 6.2	227.8 ± 19.9	7.38 ± 0.11	40.4
28	GACA **C** GCUG CUGU**AU**CGAC	66.9 ± 3.5	194.3 ± 11.4	6.64 ± 0.14	37.5	73.9 ± 4.8	217.2 ±± 15.6	6.59 ± 0.07	37.3
27	GACC **A** GCUG CUGU**AG**UGAC	45.7 ± 2.7	122.9 ± 8.4	7.59 ± 0.22	44.3	47.2 ± 7.2	127.8 ± 22.7	7.57 ± 0.40	44.0
27	GACA **A** GCUG CUGU**AG**UGAC	80.7 ± 9.6	239.5 ± 30.7	6.42 ± 0.14	36.6	78.9 ± 8.3	233.8 ± 27.0	6.40 ± 0.17	36.5
24	UGAC **C** CUCA[d] ACUG**AU**GAGU	57.9 ± 5.7	165.3 ± 18.3	6.68 ± 0.06	37.8	56.3 ± 0.9	160.1 ± 3.0	6.61 ± 0.01	37.4
	GGCC **C** CGCC CCGG**AU**GCGG	53.9 ± 4.8	142.9 ± 14.7	9.61 ± 0.30	55.8	54.9 ± 3.0	145.9 ± 9.2	9.60 ± 0.15	55.4
21	GACC **A** GCUG CUGG**GG**UGAC	55.8 ± 4.6	152.4 ± 14.5	8.52 ± 0.21	48.5	61.0 ± 12.8	169.1 ± 40.8	8.50 ± 0.70	47.3
21	GACA **G** CCUG CUGU**AA**GGAC	77.3 ± 7.5	223.6 ± 23.9	7.97 ± 0.14	42.9	73.80 ± 3.9	212.6 ± 12.4	7.92 ± 0.06	42.9
20	GACC **G** CCUG CUGG**GA**GGAC	72.8 ± 9.1	201.5 ± 28.0	10.30 ± 0.48	53.9	75.5 ± 6.2	209.9 ± 19.1	10.37 ± 0.30	53.6
	UGAC **G** CUCA[d] ACUG**GA**GAGU	65.9 ± 3.9	185.5 ± 12.1	8.34 ± 0.11	45.8	66.8 ± 0.7	188.6 ± 2.4	8.32 ± 0.02	45.5
15	GACU **A** GCUG CUGA**AG**CGAC	75.8 ± 6.6	217.1 ± 8.4	8.44 ± 0.33	45.0	77.2 ± 12.1	221.6 ± 38.4	8.43 ± 0.45	44.8
10	GACC **A** ACUG CUGG**AG**UGAC	(32.3)	(78.7)	(7.83)	(50.1)	(26.0)	(59.1)	(7.64)	(50.9)
10	GACC **U** CCUG CUGG**UU**GGAC	86.2 ± 13.5	241.1 ± 42.0	11.37 ± 0.49	55.5	85.3 ± 10.5	238.5 ± 32.2	11.34 ± 0.57	55.5
	UGAC **U** CUCA[d] ACUG**UU**GAGU	63.6 ± 2.0	179.8 ± 6.7	7.83 ± 0.06	43.5	69.4 ± 1.0	198.3 ± 3.2	7.90 ± 0.02	43.3
	AGGC **U** CGGA[g] UCCG**UU**GCCU	87.4 ± 4.3	246.7 ± 12.7	10.87 ± 0.31	53.2	78.0 ± 1.9	217.7 ± 5.7	10.43 ± 0.08	53.3
9	GACC **A** CCUG CUGG**AG**GGAC	76.2 ± 14.9	212.3 ± 45.1	10.37 ± 0.94	53.5	76.6 ± 12.9	213.6 ± 39.6	10.33 ± 0.78	53.2
	UGAC **A** CUCA[d] ACUG**AG**GAGU	65.4 ± 0.9	183.9 ± 2.7	8.32 ± 0.08	45.8	65.0 ± 1.8	182.8 ± 5.7	8.32 ± 0.05	45.8
9	GACC **A** GCUG CUGG**AG**CGAC	77.2 ± 8.5	216.2 ± 26.1	10.15 ± 0.36	52.2	79.4 ± 4.3	223.2 ± 13.2	10.17 ± 0.17	51.9
8	GACC **A** GCUG CUGG**AG**UGAC	(83.8)	(238.6)	(9.80)	(49.6)	(80.3)	(227.3)	(9.81)	(50.2)
8	GACU **A** UCUG CUGA**AG**AGAC	69.8 ± 7.7	204.1 ± 25.0	6.54 ± 0.13	37.0	67.8 ± 2.2	197.4 ± 7.1	6.52 ± 0.03	37.0
8	GACC **U** GCUG CUGG**UU**UGAC	88.3 ± 10.8	256.8 ± 35.1	8.68 ± 0.38	44.7	88.7 ± 15.6	258.0 ± 49.2	8.65 ± 0.53	44.6
7	GACU **A** ACUG CUGA**AA**UGAC	68.2 ± 10.8	202.0 ± 35.1	5.55 ± 0.22	32.6	61.2 ± 2.8	179.4 ± 9.3	5.57 ± 0.07	32.2

Table 2 (Continued)

		analysis of melt curve fit/errors				analysis of T_m dependence/errors (ln plot)			
freq[b]	sequence	$-\Delta H°$ (kcal/mol)	$-\Delta S°$ (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)	$-\Delta H°$ (kcal/mol)	$-\Delta S°$ (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)
7	GACG **A** CCUG[f] CUGC**CG**GGAC	74.7 ± 8.2	204.9 ± 24.6	11.13 ± 0.61	57.3	71.4 ± 15.4	194.9 ± 46.8	10.93 ± 1.15	57.4
7	GACC **A** GCUG CUGG**GG**CGAC	76.1 ± 17.6	214.2 ± 56.1	9.70 ± 0.33	50.5	78.7 ± 9.2	222.0 ± 28.7	9.80 ± 0.42	50.5

[a] Measurements were made in 1.0 M NaCl, 10 mM sodium cacodylate, and 0.5 mM Na_2EDTA, pH 7.0. Values in parentheses are approximate due to non-two-state melts and/or large errors associated with the derived thermodynamic parameters. [b] Frequency of occurrence in the database described in Materials and Methods. [c] Calculated at 10^{-4} M oligomer concentration. [d] Reference (*9*). [e] Reference (*11*). [f] It appears as if a bimolecular association of one of the strands with itself may be competing with the bimolecular association of the two separate strands. The formation of the 1 × 2 loop listed could not be confirmed with certainty. These values have been omitted from linear regression, trends, and averages. [g] Reference (*10*).

5S rRNA motifs, this database is large enough and diverse enough to account for all possible 1 × 2 loop sequences.

The third set of data in Table 1 shows the nearest neighbor combinations of 1 × 2 loops. The most frequent nearest neighbor combination is $\binom{C\ X\ C}{G\ X\ X\ G}$, representing 22% of the total number of loops. This is the only combination in the top ten, however, that contains two G-C pairs. It is interesting to note that four combinations in the top ten contain G-U pairs. Although G-U pairs occur much less frequently than G-C and A-U pairs (*28*), several of the most common 1 × 2 nearest neighbor combinations contain G-U pairs. This is consistent with previous research which shows that 42% of G-U pairs in small and large subunit rRNA occur at loop−helix junctions (*28*). It is also interesting to note that six possible nearest neighbor combinations are not found in the database. All six of these combinations contain a U-G pair, with five containing a U-G pair 3′ of the loop. Since G-U pairs are not as common as G-C and A-U pairs, it is possible that U-G pairs 3′ of the loop would occur in a more extensive secondary structure database. It is also possible that there is an unfavorable interaction that prevents nature from containing many loops with a 3′ U-G pair. In addition, it is obvious from the third set of data in Table 1 that previous studies have focused on 1 × 2 loops closed by C-G pairs. This study provides data for loops closed by A-U and G-U pairs as well.

It is interesting to note the pattern observed in the last set of data in Table 1 when loop nucleotides are classified as purines (R) or pyrimidines (Y). For simplicity, nucleotide positions in the loop will be referred to as $\binom{a}{bc}$. The two most frequent types of loops contain all purines (a = b = c = R) or all pyrimidines (a = b = c = Y), with all purine and all pyrimidine loops accounting for 56 and 15% of all loops found, respectively. The second most frequent group in this category are loops in which a = c ≠ b. For example, $\binom{R}{YR}$ and $\binom{Y}{RY}$ make up 9 and 8% of all loops found, respectively. The third most frequent group in this category are loops in which a = b ≠ c. For example, $\binom{R}{RY}$ and $\binom{Y}{YR}$ make up 6 and 3% of all loops found, respectively. The fourth and least frequent group in this set of data are loops in which a ≠ b = c. For example, $\binom{Y}{RR}$ and $\binom{R}{YY}$ each only make up 2% of all loops found. In each group of data, the loop with more purines occurs more frequently than the loop with more pyrimidines. When considering the frequencies, however, one must consider the number of possibilities for each combination. For example, when a = b = c, there are eight different 1 × 2 loop sequences that are possible. When a = c ≠ b and when a = b ≠ c, there are two different 1 × 2 loop sequences that are possible. Finally, when a ≠ b = c, there is only one 1 × 2 loop sequence that is possible. The number of 1 × 2 loop sequences that are possible in each group, however, does not fully explain the trends observed here, such as why the purine-rich members of each group occur more frequently than the pyrimidine-rich. These data suggest that placement of purines and pyrimidines in 1 × 2 loops are important when nature selects the sequence of 1 × 2 loops.

Thermodynamic Contributions of a 1 × 2 Loop to Duplex Thermodynamics. There is a wide range of contributions of a 1 × 2 loop to duplex thermodynamics. Loop contributions to enthalpy, entropy, and free energy changes range from −21.6 to 22.1 kcal/mol, −75.3 to 54.9 cal/(K·mol), and −0.95 to 5.3 kcal/mol, respectively (Table 3). There does not appear to be a correlation between the thermodynamic contribution of a loop and its frequency of occurrence in the database. For example, the three most stable loops are the 12th, 14th, and 15th most common loops found in the database, and the three least stable loops are the sixth, seventh, and eighth most common loops found in the database.

Except for 5′CUC3′/3′GUUG5′, all of the internal loops destabilize the duplex (Table 3). The 15 most stable sequences have possible G·A or U·U pairings. Possible G·A or U·U pairings have been previously shown to stabilize mismatches and loops (*9*, *29*−*33*). However, not all duplexes with possible G·A or U·U pairings are more stable than those loops with no possible G·A or U·U pairings. Thus, there are some idiosyncrasies associated with the free energy contribution of loops with possible G·A or U·U pairings.

There does seem to be a correlation between the number of G-C pairs directly adjacent to the loop and the free energy contribution to duplex stability. For example, the 27 loops with two adjacent G-C pairs contribute an average of 1.6 kcal/mol to duplex stability. The ten loops with one G-C adjacent base pair and the three loops with no adjacent G-C base pairs contribute an average of 2.2 and 3.0 kcal/mol to duplex stability, respectively.

It is interesting to note that there are five loops (loop nucleotides plus adjacent nearest neighbors) that were studied in at least two different stem sequences of the same length (Table 3). The differences in contributions of the loops to

Table 3: Loop Contribution to Duplex Thermodynamics[a]

freq[b]	sequence	$\Delta H°_{loop}$ (kcal/mol)	$\Delta S°_{loop}$ (cal/K·mol)	$\Delta G°_{37,loop}$ (kcal/mol)	freq[b]	sequence	$\Delta H°_{loop}$ (kcal/mol)	$\Delta S°_{loop}$ (cal/K·mol)	$\Delta G°_{37,loop}$ (kcal/mol)
50	UGAC **A** CUCA[c]	0.4	−5.9	2.28	9	GACC **A** CCUG	−8.7	−30.7	0.88
	ACUG**AA**GAGU	*3.2*	*3.1*	*2.2*		CUGG**AG**GGAC	*−3.1*	*−13.1*	*1.1*
	UGAC **A** CUCA[d]	4.7	6.7	2.66		UGAC **A** CUCA[c]	−2.3	−9.9	0.76
	ACUG**AA**GAGU	*3.2*	*3.1*	*2.2*		ACUG**AG**GAGU	*−3.1*	*−13.1*	*1.1*
38	GACU **A** GCUG	(−6.6)	(−31.9)	(3.28)	9	GACC **A** GCUG	−10.0	−36.1	1.20
	CUGG**GG**UGAC	*−10.1*	*−40.3*	*2.5*		CUGG**AG**CGAC	*−3.1*	*−13.1*	*1.1*
32	GACG **A** ACUG	−14.2	−54.0	2.48	8	GACC **A** GCUG	(−13.2)	(−44.6)	(0.65)
	CUGC**CA**UGAC	*−0.3*	*−10.5*	*2.9*		CUGG**AG**UGAC	*−6.6*	*−26.7*	*1.8*
30	GACG **U** UCUG	−5.1	−24.9	2.65	8	GACU **A** UCUG	−3.7	−20.3	2.6
	CUGC**UC**AGAC	*−12.2*	*−46.0*	1.9		CUGA**AG**AGAC	*−10.1*	*−40.3*	*2.5*
29	GGCG **A** CGCC[e]	−7.7	−21.9	−0.95	8	GACC **U** GCUG	−21.6	−75.3	1.81
	CCGU**GC**GCGG	*−6.6*	*−26.7*	*1.8*		CUGG**UU**UGAC	*−12.2*	*−46.0*	*1.9*
29	GACC **G** CCUG	12.5	32.5	2.46	7	GACU **A** ACUG	1.8	−5.3	3.44
	CUGG**AA**GGAC	*−3.1*	*−13.1*	*1.1*		CUGA**AA**UGAC	*−3.8*	*−24.1*	*3.6*
	UGAC **G** CUCA[c]	−1.6	−10.8	1.73	7	GACG **A** CCUG[e]	−6.2	−18.0	−0.62
	ACUG**AA**GAGU	*−3.1*	*−13.1*	*1.1*		CUGC**CG**GGAC	*−3.1*	*−13.1*	*1.1*
28	GACC **A** ACUG	−12.1	−48.1	2.81	7	GACC **A** GCUG	−9.3	−34.9	1.57
	CUGG**AA**UGAC	*−0.3*	*−10.5*	*2.9*		CUGG**GG**CGAC	*−3.1*	*−13.1*	*1.1*
28	GACA **C** GCUG	−7.4	−35.9	3.63	3	UGAC **A** CUCA[c]	−2.5	−13.5	1.66
	CUGU**AU**CGAC	*−0.3*	*−10.5*	*2.9*		ACUG**CC**GAGU	*3.2*	*3.1*	*2.2*
27	GACC **A** GCUG	19.9	54.9	2.89	3	UGAC **C** CUCA[c]	7.8	17.1	2.46
	CUGG**AA**UGAC	*−0.3*	*−10.5*	*2.9*		ACUG**CC**GAGU	*3.2*	*3.1*	*2.2*
27	GACA **A** GCUG	−14.7	−56.9	2.91	3	UGAC **C** CUCA[c]	1.3	−2.1	1.92
	CUGU**AG**UGAC	*−10.1*	*−40.3*	*2.5*		ACUG**CU**GAGU	*3.2*	*3.1*	*2.2*
24	UGAC **C** CUCA[c]	6.4	12.8	2.47	2	UGAC **A** CUCA[c]	2.5	4.7	1.07
	ACUG**AU**GAGU	*3.2*	*3.1*	*2.2*		ACUG**GA**GAGU	*−3.1*	*−13.1*	*1.1*
	GGCC **C** CGCC	22.1	54.2	5.29	1	UGAC **C** CUCA[c]	1.6	−2.3	2.3
	CCGG**AU**GCGG	*3.2*	*3.1*	*2.2*		ACUG**AA**GAGU	*3.2*	*3.1*	*2.2*
21	GACC **A** GCUG	6.1	13.6	1.96	1	UGAC **U** CUCA[c]	−5.8	−23.6	1.54
	CUGG**GG**UGAC	*−6.6*	*−26.7*	*1.8*		ACUG**UC**GAGU	*−8.7*	*−32.4*	*1.2*
21	GACA **G** CCUG	−8.8	−35.5	2.14	1	UCAG **U** GUGA[f]	−4.7	−20.4	1.59
	CUGU**AA**GGAC	*−6.6*	*−26.7*	*1.8*		AGUC**CU**CACU	*−8.7*	*−32.4*	*1.9*
20	GACC **G** CCUG	−7.6	−27.0	0.84	0	UCAG **A** GUGA[f]	−5.0	−19.8	1.15
	CUGG**GA**GGAC	*−3.1*	*−13.1*	*1.1*		AGUC**AG**CACU	*−3.1*	*−13.1*	*1.1*
	UGAC **G** CUCA[c]	−4.1	−15.8	0.76	0	UGAC **G** CUCA[c]	−3.1	−13.9	1.16
	ACUG**GA**GAGU	*−3.1*	*−13.1*	*1.1*		ACUG**AG**GAGU	*−3.1*	*−13.1*	*1.1*
15	GACU **A** GCUG	−10.7	−40.1	1.76	0	UGAG **A** GUCA[c]	2.4	1.1	2.11
	CUGA**AG**CGAC	*−6.6*	*−26.7*	*1.8*		ACUC**GA**CAGU	*−3.1*	*−13.1*	*2.2*
10	GACC **A** ACUG	(39.9)	(120.6)	(2.55)	0	UGAC **U** CUCA[c]	5.8	11.6	2.22
	CUGG**AG**UGAC	*−6.6*	*−26.7*	*1.8*		ACUG**CC**GAGU	*3.2*	*3.1*	*2.2*
10	GACC **U** CCUG	−17.4	−55.6	−0.13	0	UGAG **A** GUCA[c]	5.8	11.6	2.22
	CUGG**UU**GGAC	*−8.7*	*−32.4*	*1.2*		ACUC**AA**CAGU	*3.2*	*3.1*	*2.2*
	UGAC **U** CUCA[c]	−6.7	−25.5	1.18	0	UGAG **C** GUCA[f]	7.8	17.1	2.46
	ACUG**UU**GAGU	*−8.7*	*−32.4*	*1.2*		ACUC**CC**CAGU	*3.2*	*3.1*	*2.2*
	AGGC **U** CGGA[f]	−7.3	−28.5	1.57					
	UCCG**UU**GCCU	*−8.7*	*−32.4*	*1.2*					

[a] Values in italics are predicted values based on the model derived here. Values in parentheses are approximate due to non-two-state melts and/or large errors associated with the derived thermodynamic parameters. [b] Frequency of occurrence in the database described in Materials and Methods. [c] Derived from raw data availabe in ref (*9*). [d] Derived from raw data available in ref (*11*). [e] It appears as if a bimolecular association of one of the strands with itself may be competing with the bimolecular association of the two separate strands. The formation of the 1 × 2 loop listed could not be confirmed with certainty. These values have been omitted from linear regression, trends, and averages. [f] Derived from raw data available in ref (*10*).

duplex stabilities vary with the stem sequence. For example, the free energy contribution of 5′CGC3′/3′GGAG5′ only differed by 0.1 kcal/mol when it was placed into two different stems, however, the free energy contribution of 5′CUC3′/3′GUUG5′ differed by 1.7 kcal/mol when it was placed into two different stems. Thus, changing base pairs not adjacent to the loop can make a substantial difference in the free energy contribution of the loop. Similar non-nearest neighbor effects have been observed for bulges (*34*) and single mismatches (*35*). For example, bulges in the 5′GCGA$_n$GCGA3′/3′CGCCGCU5′ series consistently destabilize the duplex at least 2 kcal/mol more than those in the 5′GCGA$_n$GUCA3′/3′CGCCAGU5′ series, although these bulges have the same bulge nucleotides and same nearest neighbor nucleotides. It has recently been reported that the accuracy of RNA secondary structure prediction by free energy minimization is limited by non-nearest neighbor effects (*36*). Since non-nearest neighbor effects may be complicated to interpret and to include in algorithms such as *RNAstructure* and *mfold*, non-nearest neighbor effects were ignored here, and data was treated as if stability relied only upon immediate nearest neighbors.

Table 4: 1 × 2 Loop Nearest Neighbor Parameters at 37 °C[a]

	ΔH° (kcal/mol)	ΔS° (cal/K·mol)	ΔG°_{37} (kcal/mol)
1 × 2 loop initiation[b]	3.2 ± 1.9 (3.6 ± 1.4)	3.1 ± 6.2	2.2 ± 0.1 (2.3 ± 0.1)
A-U or G-U closure[c]	−3.5 ± 1.9 (5.0 ± 0.7)	−13.6 ± 5.6	0.7 ± 0.1 (0.7 ± 0.1)
A·G or G·G mismatch[d]	−6.3 ± 2.5 (−5.8 ± 1.5)	−16.2 ± 7.6	−1.1 ± 0.2 (−1.2 ± 0.2)
U·U mismatch[d]	−11.9 ± 3.3 (−10.1 ± 1.7)	−35.5 ± 10.2	−1.0 ± 0.2 (−0.8 ± 0.2)

[a] Numbers in parentheses are values currently used by *RNAstructure* and/or were published previously (*1*). [b] These parameters apply to all 1 × 2 loops, regardless of sequence. [c] These parameters are applied per A-U or G-U closure. [d] These parameters should only be applied once per loop for the type of possible mismatch listed and are not applied for 5′RA/3′YG loops.

Improving/Testing the Model Currently Used To Calculate Loop Thermodynamics. RNAstructure uses two methods to calculate the free energy contribution of a 1 × 2 loop (*1*−*3*). If a loop has been measured thermodynamically, *RNAstructure* uses the measured value or average of measured values. Previous to these experiments, only six of the 25 most frequent 1 × 2 loops found in the database had been measured thermodynamically (Table 1, first set of data). Now, the top 25 most frequently occurring loops in the database have experimental data. Since most scientists are likely interested in loops that do occur in nature, theses loops have experimental data, and scientists no longer have to rely on a model based on several approximations and assumptions for these loops.

For those loops that do not have experimental numbers, *RNAstructure* approximates the free energy contribution as described in the introductory comments and shown by eq 1. Since the number of 1 × 2 loops with experimental values has been doubled with this work, the data from this larger dataset was used to derive new 1 × 2 thermodynamic parameters in order to evaluate the accuracy of the current model used to calculate the thermodynamic contribution of 1 × 2 loops. Several models were generated by using different combinations of parameters (data not shown). The model that most agreed with the experimental data, resulted in linearly independent parameters, and had small standard deviations was one using parameters identical to the current *RNAstructure* model (*1*−*3*) (eq 1). In fact, most of the new parameters are within experimental error of the current *RNAstructure* model.

In *RNAstructure*, $\Delta G^\circ_{37,\text{loop initiation}}$ (1.7 kcal/mol) was derived from a small dataset of 1 × 2 loops (*1*). $\Delta G^\circ_{37,\text{asym}}$ is a penalty (0.7 kcal/mol) for having an unequal number of nucleotides on each side of the loop. This value was derived from data of loops of various sizes (*1*). Here, linear regression was used to derive a new parameter, $\Delta G^\circ_{37,1\times2\text{ initiation}}$, based on the new dataset of 1 × 2 loops. This value, 2.2 ± 0.1 kcal/mol, is within experimental error of the sum of $\Delta G^\circ_{37,\text{loop initiation}}$ and $\Delta G^\circ_{37,\text{asym}}$ (2.3 ± 0.1 kcal/mol) (Table 4).

RNAstructure also uses a penalty for replacing a closing G-C pair with an A-U or G-U pair, $\Delta G^\circ_{37,\text{AU/GU}}$. This value was derived from data of loops of various sizes (*1*). From the new dataset, linear regression was used to derive a $\Delta G^\circ_{37,\text{AU/GU}}$ term based only on the 1 × 2 loop data. This value, 0.7 ± 0.1 kcal/mol, is identical to *RNAstructure's* current penalty (0.7 kcal/mol) (*1*).

The final parameter used by the current model in *RNAstructure* is $\Delta G^\circ_{37,\text{first noncanonical pairs}}$, a bonus for stabilizing mismatches in the loop (*1*). This parameter consists of two values: (1) a bonus for all G·G mismatches and G·A mismatches, excluding 5′RA/3′YG loops (−1.2 ± 0.2 kcal/mol), and (2) a bonus for U·U mismatches (−0.8 ± 0.2 kcal/mol). These values were derived from the limited database of 1 × 2 loops. Linear regression with the larger dataset was used to derive new values for these parameters. For an A·G (excluding 5′RA/3′YG) or G·G mismatch, a bonus of −1.2 ± 0.2 kcal/mol was derived and is within experimental error of the current value. An all guanosine 1 × 2 loop is the only 1 × 2 loop that can have a G·G mismatch without also having a G·A mismatch. An all guanosine 1 × 2 loop has not been studied experimentally (due to aggregation issues). However, since G·G mismatches appear to stabilize loops of other sizes (*10*), G·A and G·G mismatches have been assigned the same bonus. The new bonus for U·U mismatches, −1.0 ± 0.2 kcal/mol, is also within experimental error of the bonus used in the current model, −0.8 ± 0.2 kcal/mol.

A similar comparison can be made between the enthalpy parameters published previously (*1*) and those derived here. All of the parameters are within experimental error except for the A-U or G-U closure parameter (Table 4). Previously, this parameter was determined to be slightly favorable (−3.5 ± 1.9 kcal/mol) (*1*). Here, this parameter was derived to be slightly unfavorable (5.0 ± 0.7 kcal/mol). It is unclear why this single parameter varies between what has been determined previously and what has been derived here. The agreement between the previously published free energy and enthalpy parameters and those derived here is somewhat encouraging. The previous parameters either were derived from a very small dataset or were derived from a dataset of internal loops of various sizes. Since the parameters published previously are in agreement with those published here, it suggests that (1) the small dataset of previously studied 1 × 2 loops was representative of 1 × 2 loops in general and (2) deriving 1 × 2 loop parameters from loops of other sizes can be a valid approximation.

ACKNOWLEDGMENT

The authors would like to thank Doug Turner for insightful discussions and the Turner lab for use of their RNA synthesizer. The authors would also like to thank Dave Mathews for his suggestions with linear regression.

REFERENCES

1. Lu, Z. J., Turner, D. H., and Mathews, D. H. (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation, *Nucleic Acids Res. 34*, 4912−4924.
2. Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, *J. Mol. Biol. 288*, 911−940.
3. Mathews, D. H., Disney, M. D., Childs, J. C., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure, *Proc. Natl. Acad. Sci. U.S.A. 101*, 7287−7292.

4. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res. 31*, 3406−3415.
5. Hofacker, I. L. (2003) Vienna RNA secondary structure server, *Nucleic Acids Res. 31*, 3429−3431.
6. Mathews, D. H., Schroeder, S. J., Turner, D. H., and Zuker, M. (2005). Predicting RNA Secondary Structure, in *The RNA World*, 3rd ed. (Gesteland, R. F., Cech, T. R., and Atkins, J. F., Eds.) pp 631−657, Cold Spring Harbor Laboratory Press, Cold Spring Harbor.
7. Xia, T., SantaLucia, J., Jr., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs, *Biochemistry 37*, 14719−14735.
8. Recht, M. I., Fourmy, D., Blanchard, S. C., Dahlquist, K. D., and Puglisi, J. D. (1996) RNA sequence determinants for aminoglycoside binding to an A-site rRNA model oligonucleotide, *J. Mol. Biol. 262*, 421−436.
9. Schroeder, S., Kim, J., and Turner, D. H. (1996) G·A and U·U mismatches can stabilize RNA internal loops of three nucleotides, *Biochemistry 35*, 16105−16109.
10. Schroeder, S. J., and Turner, D. H. (2000) Factors affecting the thermodynamic stability of small asymmetric internal loops in RNA, *Biochemistry 39*, 9257−9274.
11. Peritz, A. E., Kierzek, R., Sugimoto, N., and Turner, D. H. (1991) Thermodynamic study of internal loops in oligoribonucleotides: Symmetric loops are more stable than asymmetric loops, *Biochemistry 30*, 6428−6436.
12. Gutell, R. R. (1994) Collection of small-subunit (16s- and 16s-like) ribosomal-RNA structures - 1994, *Nucleic Acids Res. 22*, 3502−3507.
13. Gutell, R. R., Gray, M. W., and Schnare, M. N. (1993) A compilation of large subunit (23s-like and 23s-like) ribosomal-RNA structures - 1993, *Nucleic Acids Res. 21*, 3055−3074.
14. Schnare, M. N., Damberger, S. H., Gray, M. W., and Gutell, R. R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA, *J. Mol. Biol. 256*, 701−719.
15. Szymanski, M., Specht, T., Barciszewska, M. Z., Barciszewski, J., and Erdmann, V. A. (1998) 5S rRNA data bank, *Nucleic Acids Res. 26*, 156−159.
16. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes, *Nucleic Acids Res. 26*, 148−153.
17. Larsen, N., Samuelsson, T., and Zwieb, C. (1998) The signal recognition particle database (SRPDB), *Nucleic Acids Res. 26*, 177−178.
18. Brown, J. W. (1998) The ribonuclease P database, *Nucleic Acids Res. 26*, 351−352.
19. Damberger, S. H., and Gutell, R. R. (1994) A comparative database of group I intron structures, *Nucleic Acids Res. 22*, 3508−3510.
20. Waring, R. B., and Davies, R. W. (1984) Assessment of a model for intron RNA secondary structure relevant to RNA self-splicing - A review, *Gene 28*, 277−291.
21. Michel, F., Umesono, K., and Ozeki, H. (1989) Comparative and functional-anatomy of group-Ii catalytic introns - A review, *Gene 82*, 5−30.
22. Wright, D. J., Rice, J. L., Yanker, D. M., and Znosko, B. M. (2007) Nearest neighbor parameters for inosine-uridine pairs in RNA duplexes, *Biochemistry 46*, 4625−4634.
23. McDowell, J. A. (1995) RNA Calculations v. 1.1.
24. McDowell, J. A. (1996) MeltWin v. 3.5: Melt Curve Processing Program.
25. Petersheim, M., and Turner, D. H. (1983) Base-stacking and base-pairing contributions to helix stability: Thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp, *Biochemistry 22*, 256−263.
26. McDowell, J. A., and Turner, D. H. (1996) Investigation of the structural basis for thermodynamic stabilities of tandem GU mismatches: solution structure of $(rGAGGUCUC)_2$ by two-dimensional NMR and simulated annealing, *Biochemistry 35*, 14077−14089.
27. Borer, P. N., Dengler, B., Tinoco, I., and Uhlenbeck, O. (1974) Stability of ribonucleic-acid double-stranded helices, *J. Mol. Biol. 86*, 843−853.
28. Gautheret, D., Konings, D., and Gutell, R. R. (1995) GU base-pairing motifs in ribosomal-RNA, *RNA 1*, 807−814.
29. Walter, A. E., Wu, M., and Turner, D. H. (1994) The stability and structure of tandem GA mismatches in RNA depend on closing base pairs, *Biochemistry 33*, 11349−11354.
30. SantaLucia, J., Jr., Kierzek, R., and Turner, D. H. (1991) 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.
31. SantaLucia, J., Kierzek, R., and Turner, D. H. (1991) Functional-group substitutions as probes of hydrogen-bonding between GA mismatches in RNA internal loops, *J. Am. Chem. Soc. 113*, 4313−4322.
32. SantaLucia, J., Jr., and Turner, D. H. (1993) Structure of $(rGGCGAGCC)_2$ in solution from NMR and restrained molecular dynamics, *Biochemistry 32*, 12612−23.
33. Wu, M., McDowell, J. A., and Turner, D. H. (1995) A periodic table of symmetric tandem mismatches in RNA, *Biochemistry 34*, 3204−3211.
34. Longfellow, C. E., Kierzek, R., and Turner, D. H. (1990) Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides, *Biochemistry 29*, 278−285.
35. Kierzek, R., Burkard, M. E., and Turner, D. H. (1999) Thermodynamics of single mismatches in RNA duplexes, *Biochemistry 38*, 14214−14223.
36. Mathews, D. H. (2006) Revolutions in RNA secondary structure prediction, *J. Mol. Biol. 359*, 526−532.

BI701024W