BIOCHEMISTRY
including biophysical chemistry & molecular biology

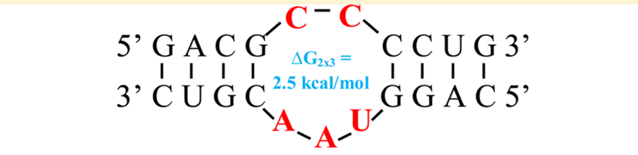
Article

pubs.acs.org/biochemistry

Thermodynamic Characterization of RNA 2 × 3 Nucleotide Internal Loops

Nina Z. Hausmann and Brent M. Znosko*

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



$$\Delta G^{\circ}_{37,2\times3} = \Delta G^{\circ}_{37,2\times3\text{ initiation}} + \Delta G^{\circ}_{37,\text{AU/GU}} + \Delta G^{\circ}_{37,5'\text{RG}/3'\text{YA}} + \Delta G^{\circ}_{37,5'\text{YG}/3'\text{RA}} + \Delta G^{\circ}_{37,\text{GG}} + G^{\circ}_{37,\text{UU}}$$

ABSTRACT: To better elucidate RNA structure−function relationships and to improve the design of pharmaceutical agents that target specific RNA motifs, an understanding of RNA primary, secondary, and tertiary structure is necessary. The prediction of RNA secondary structure from sequence is an intermediate step in predicting RNA three-dimensional structure. RNA secondary structure is typically predicted using a nearest neighbor model based on free energy parameters. The current free energy parameters for 2 × 3 nucleotide loops are based on a 23-member data set of 2 × 3 loops and internal loops of other sizes. A database of representative RNA secondary structures was searched to identify 2 × 3 nucleotide loops that occur in nature. Seventeen of the most frequent 2 × 3 nucleotide loops in this database were studied by optical melting experiments. Fifteen of these loops melted in a two-state manner, and the associated experimental $\Delta G^{\circ}_{37,2\times3}$ values are, on average, 0.6 and 0.7 kcal/mol different from the values predicted for these internal loops using the predictive models proposed by Lu, Turner, and Mathews [Lu, Z. J., Turner, D. H., and Mathews, D. H. (2006) *Nucleic Acids Res. 34*, 4912−4924] and Chen and Turner [Chen, G., and Turner, D. H. (2006) *Biochemistry 45*, 4025−4043], respectively. These new $\Delta G^{\circ}_{37,2\times3}$ values can be used to update the current algorithms that predict secondary structure from sequence. To improve free energy calculations for duplexes containing 2 × 3 nucleotide loops that still do not have experimentally determined free energy contributions, an updated predictive model was derived. This new model resulted from a linear regression analysis of the data reported here combined with 31 previously studied 2 × 3 nucleotide internal loops. Most of the values for the parameters in this new predictive model are within experimental error of those of the previous models, suggesting that approximations and assumptions associated with the derivation of the previous nearest neighbor parameters were valid. The updated predictive model predicts free energies of 2 × 3 nucleotide internal loops within 0.4 kcal/mol, on average, of the experimental free energy values. Both the experimental values and the updated predictive model can be used to improve secondary structure prediction from sequence.

While approximately half of RNA nucleotides are found in Watson−Crick regions, the other half are found in bulges, hairpin loops, internal loops, or multibranch loops.[1] One common non-Watson−Crick motif is a 2 × 3 nucleotide internal loop, which has been found to occur in the human immunodeficiency virus type 1 (HIV-1) Rev response element,[2,3] the large ribosomal subunit from *Haloarcula marismortui*,[4] the *Thermus thermophilus* 30S ribosomal subunit,[5] the P4−P6 domain of the *Tetrahymena thermophila*[6] and *Pneumocystis carinii*[7] group I introns, the purine-binding domain of the guanine riboswitch from the xpt-pbuX operon of *Bacillus subtilis*,[8] and in many other organisms (Figure 1). On the basis of the sequences, 2 × 3 nucleotide loops are predicted to form in the 23S ribosomal subunit of the fungal plant pathogen *Mucor racemosus*,[9,10] the 16S ribosomal subunit of the rodent parasite *Giardia muris*,[11] the group I intron of *Chlorella saccharophila*,[12,13] and the signal recognition particle from the flowering plant *Cineraria hybrida*,[14] to name a few.

Not only do 2 × 3 nucleotide loops occur in different types of RNA and in a wide range of organisms, but they also can be found in regions that serve important biological roles. One example is the 2 × 3 nucleotide loop found in the *Homo sapiens* 18S ribosomal decoding A site (Figure 1), which plays an important role in protein synthesis.[15] The decoding A site works to recognize the cognate interactions between the tRNA anticodon and the messenger codon. Specifically, when an activated tRNA with elongation factor EF-Tu and GTP is delivered to the A site, the adenines found within the 2 × 3 nucleotide loop change their conformation from an "off" state to an "on" state. Interactions between the adenosine residues (while in the "on" state) of the 2 × 3 nucleotide loop and the codon and anticodon contribute to the fidelity of the tRNA selection step.[15] Another example of a biological role played by 2 × 3 nucleotide loops is the Rev response element in mRNA from HIV-1 (Figure 1). The Rev response element contains a 2 × 3 nucleotide loop that interacts with Rev, a small regulatory protein of HIV-1 that controls viral replication.[16,17]

Because of the frequency of occurrence of 2 × 3 nucleotide loops in nature and their important biological roles, they are of interest to researchers in a wide variety of fields. Being able to accurately predict the thermodynamic contribution of 2 × 3 nucleotide loops to duplex stability is crucial in predicting RNA

Received: January 26, 2012
Revised: April 30, 2012
Published: June 21, 2012

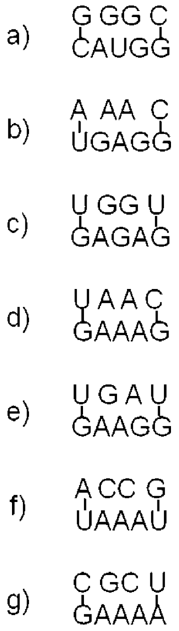


Figure 1. Secondary structures of naturally occurring 2 × 3 nucleotide internal loops. These loops are found in (a) the human immunodeficiency virus type 1 (HIV-1) Rev response element,[2,3] (b) the large ribosomal subunit from *H. marismortui*,[4] (c) the *T. thermophilus* 30S ribosomal subunit,[5] (d) the P4−P6 domain of the *T. thermophila* group I intron,[6] (e) the P4−P6 domain of the *P. carinii* group I intron,[7] (f) the purine-binding domain of the guanine riboswitch from the xpt-pbuX operon of *B. subtilis*,[8] and (g) the *H. sapiens* 18S ribosomal decoding A site.[15]

secondary structure from sequence. Accurate secondary structures can provide information about structure−function relationships, three-dimensional structures, and tertiary interactions, thereby aiding in the design of therapeutics. Because the thermodynamics of <3% of all possible 2 × 3 nucleotide loops have been experimentally measured, the thermodynamic contribution of most 2 × 3 nucleotide loops to duplex stability is predicted by an algorithm.[18,19] One algorithm[18] predicts the free energy contribution of 2 × 3 nucleotide loops to duplex stability using the following equation:

$$\begin{aligned}\Delta G^\circ_{37,2\times3} = {} & \Delta G^\circ_{37,\text{loop initiation}} + \Delta G^\circ_{37,\text{AU/GU}} \\ & + \Delta G^\circ_{37,\text{asym}} + \Delta G^\circ_{37,5'\text{YA}/3'\text{RG}} \\ & + \Delta G^\circ_{37,5'\text{RG}/3'\text{YA}} + \Delta G^\circ_{37,5'\text{YG}/3'\text{RA}} \\ & + \Delta G^\circ_{37,\text{GG}} + \Delta G^\circ_{37,\text{UU}} \qquad (1)\end{aligned}$$

where $\Delta G^\circ_{37,\text{loop initiation}}$ is an initiation term for an internal loop containing five nucleotides (2.0 kcal/mol); $\Delta G^\circ_{37,\text{AU/GU}}$ is a penalty applied per A-U, U-A, G-U, or U-G base pair adjacent to the loop (0.7 kcal/mol); $\Delta G^\circ_{37,\text{asym}}$ is a term for internal loops with unequal numbers of nucleotides on each side (0.6 kcal/mol); $\Delta G^\circ_{37,5'\text{YA}/3'\text{RG}}$, $\Delta G^\circ_{37,5'\text{RG}/3'\text{YA}}$, and $\Delta G^\circ_{37,5'\text{YG}/3'\text{RA}}$ are bonuses for the stack of a specific closing base pair with the first "pair" in the loop (−0.5, −1.2, and −1.1 kcal/mol, respectively); and $\Delta G^\circ_{37,\text{ GG}}$ and $\Delta G^\circ_{37,\text{ UU}}$ are bonuses for loops that contain a possible G·G or U·U pair as the first or last pair in the loop (−0.7 and −0.4 kcal/mol, respectively). While most of the parameters were derived from the published data set of 2 × 3 nucleotide loops,[7,20,21] $\Delta G^\circ_{37,\text{AU/GU}}$ and $\Delta G^\circ_{37,\text{asym}}$ were derived from data of loops of various sizes.[18]

A second, similar algorithm[19] predicts the free energy contribution of 2 × 3 nucleotide loops to duplex stability using the same equation (eq 1) with slightly different values. In this case, $\Delta G^\circ_{37,\text{loop initiation}}$ is 2.15 kcal/mol, $\Delta G^\circ_{37,\text{AU/GU}}$ is 0.73 kcal/mol, $\Delta G^\circ_{37,\text{asym}}$ is 0.45 kcal/mol, $\Delta G^\circ_{37,5'\text{YA}/3'\text{RG}}$ is −0.39 kcal/mol, $\Delta G^\circ_{37,5'\text{RG}/3'\text{YA}}$ is −1.06 kcal/mol, $\Delta G^\circ_{37,5'\text{YG}/3'\text{RA}}$ is −1.41 kcal/mol, $\Delta G^\circ_{37,\text{ GG}}$ is −0.74 kcal/mol, and $\Delta G^\circ_{37,\text{ UU}}$ is −0.34 kcal/mol. Like the first algorithm, most parameter values were derived from 2 × 3 experimental data, with the exception of $\Delta G^\circ_{37,\text{loop initiation}}$ and $\Delta G^\circ_{37,\text{asym}}$, which were derived from data of loops of various sizes.

The purpose of this study is twofold. First, by collecting thermodynamic data for previously unstudied 2 × 3 nucleotide loops that occur frequently in nature, we can directly incorporate these thermodynamic data into *RNAstructure*, an RNA secondary structure prediction program based on free energy minimization,[1,22] and other programs that predict RNA secondary structure.[23−26] This means that the thermodynamic contribution of these loops to RNA stability can be calculated from the experimental values instead of the predictive models currently used by *RNAstructure*. Second, by collecting thermodynamic data for additional loops, we can increase the size of the total database of available 2 × 3 thermodynamics. The larger database can then be used to derive an updated predictive model to be used to calculate the thermodynamic contribution of 2 × 3 loops that do not have experimental values.

To determine which 2 × 3 loops to study, we searched a database of 1349 RNA secondary structures for the most frequently occurring 2 × 3 loops. The results of this search guided our experiments, and the thermodynamics of 17 frequently occurring loops in the database are reported. The Znosko laboratory has previously published the thermodynamics of naturally occurring single mismatches (also termed 1 × 1 nucleotide internal loops),[27,28] tandem mismatches (also termed 2 × 2 nucleotide internal loops),[29] 1 × 2 nucleotide internal loops,[30] triloops (also termed hairpin loops of three nucleotides),[31] and tetraloops (also termed hairpin loops of four nucleotides).[32] This study is the next in a series of studies investigating the thermodynamics of small secondary structure motifs in RNA.

■ MATERIALS AND METHODS

Compiling and Searching a Database for 2 × 3 Nucleotide Loops. A database of 1349 RNA secondary structures containing various types of RNA was previously compiled.[31,32] This database was searched for 2 × 3 nucleotide loops, and the number of occurrences for each type of 2 × 3 nucleotide loop was tabulated (Table 1). For this work, G-U pairs were considered to be canonical base pairs.

Optical Melting Experiments. The general design of sequences to be used for optical melting studies has been previously described.[30] The sequences of loops and canonically base paired nearest neighbors were chosen on the basis of search results of the database described above. The focus of this study was the most frequently occurring 2 × 3 loop sequences in the database with their adjacent nearest neighbors. One frequently occurring loop [13 occurrences in the database (Table 1)], $\left(\begin{matrix}\text{CAAG}\\ \text{GGGGC}\end{matrix}\right)$, was not studied here because of the presence of multiple guanines in a row and the possible formation of unwanted structures and/or aggregation. In all duplex sequences, the loop nucleotides and their nearest

Table 1. Summary of a Database Search for 2 × 3 Nucleotide 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]
G **CC** C C **AAU** G	90	13.5	*e*	CC AAU	90	13.5	*e,g*	G C C G	105	15.8	*e*	RR RRR	127	19.1	*e-i*
A **AA** C U **CAC** G	39	5.9	*e*	AA CAC	40	6.0	*e*	G G C C	104	15.6	*e,g,i*	YY RRY	91	13.7	*e,g*
A **CA** G U **AUA** C	31	4.7	*e*	CA AUA	33	5.0	*e*	A G U C	60	9.0	*e,g*	RY RRY	89	13.4	*e-f*
G **AU** G C **CAU** C	18	2.7	*e*	AU AAU	25	3.8	*e-f*	A C U G	48	7.2	*e*	RR RYR	48	7.2	*e*
G **AU** G C **CGU** C	18	2.7	*e*	UA UAG	25	3.8		C G G C	43	6.5	*e,g,h*	RR YRY	42	6.3	*e*
U **AU** C G **GGU** G	17	2.6	*e*	GG AGA	22	3.3	*e,g*	U C G G	43	6.5	*e,f*	RY YRY	42	6.3	*e*
U **AA** A A **GAC** U	16	2.4	*e*	AU CAU	21	3.2	*e*	A A U U	28	4.2	*e*	YR YRR	41	6.2	
C **AA** G G **GGG** C	13	2.0		AU CGU	21	3.2	*e*	U G G U	28	4.2	*e*	YR RYR	35	5.3	*e*
U **GG** G G **GGA** U	13	2.0	*e*	AA GGG	19	2.9	*e*	U G G C	27	4.1	*e-f*	RR RRY	23	3.5	*e*
U **AU** G G **AAU** C	12	1.8	*f*	AU GGU	18	2.7	*e*	C U G A	24	3.6	*e*	YY YRR	15	2.3	*e*
G **GA** G C **AUA** C	11	1.7	*e*	AC GAU	17	2.6	*e*	C G G U	18	2.7		YY YRY	15	2.3	
G **GA** G C **AAA** C	10	1.5	*e*	GG GGA	17	2.6	*e*	U A A U	18	2.7	*e*	YY YYY	15	2.3	*g*
C **GC** U G **AAA** A	10	1.5	*e*	AA GAC	16	2.4	*e*	G A C U	15	2.3		RY RRR	11	1.7	*e,g*
U **AU** C G **AAU** G	9	1.4	*e*	AA AAG	12	1.8	*f,g*	G G C U	12	1.8		YR RRR	10	1.5	
A **AC** A U **GAU** U	7	1.1	*e*	CC UAA	12	1.8	*e*	U C A G	11	1.7		RR YYR	9	1.4	
A **CC** A U **UAA** U	6	0.9	*e*	GA AUA	12	1.8	*e*	A U U A	10	1.5		RR RYY	7	1.1	
A **CC** G U **UAA** C	6	0.9	*e*	GA AAA	11	1.7	*e-g*	C C G G	8	1.2	*g*	YR YRY	7	1.1	
U **GG** G G **AGA** U	6	0.9	*e*	GC AAA	11	1.7	*e,g*	U U A A	8	1.2	*g*	YR YYR	7	1.1	
U **GG** G G **AGA** C	5	0.8	*e*	GA AAG	9	1.4	f,g,h	U U G A	8	1.2		RR YRR	6	0.9	*g*
G **AA** G C **GAA** C	5	0.8		AC GAC	8	1.2		A G U U	6	0.9		RY RYY	5	0.8	
U **GG** U G **AGA** A	5	0.8		GA ACG	8	1.2		C A G U	6	0.9		RY RYR	4	0.6	*g*
G **GU** G C **AAU** C	5	0.8		AA GAA	7	1.1	*g*	U G A C	6	0.9		RY YYY	4	0.6	
C **UA** G G **UAG** U	5	0.8		CA AAG	7	1.1		U A G U	5	0.8	*g*	YY YYR	4	0.6	
G **UA** G C **UAG** U	5	0.8		GU AAU	7	1.1		G U C A	4	0.6		RR YYY	3	0.5	
C **AA** G G **GGG** U	4	0.6		AU AGU	6	0.9		U G A U	4	0.6		YY RYR	3	0.5	
Previously[j]	17	2.6		Previously[j]	198	29.7		Previously[j]	288	43.2		Previously[j]	343	51.5	
New Total[k]	329	49.4		New Total[k]	424	63.7		New Total[k]	544	81.7		New Total[k]	548	82.3	

[a] As described in the text, not all combinations are shown because of space limitations. [b] Frequency of occurrence in the database. [c] Percent of 666 loops, the total number of loops found in the database. [d] Reference where data are reported. [e] From this work. [f] From ref 38. [g] From ref 20. [h] From ref 7. [i] From ref 21. [j] The total number of 2 × 3 loops accounted for by previous studies and the corresponding percentage of the total number of loops found in the database. [k] The total number of 2 × 3 loops accounted for when the data from previous studies were combined with the data reported here and the corresponding percentage of the total number of loops found in the database.

neighbors were situated in the middle of a duplex, with three Watson–Crick base pairs on either side. Most of the loop–nearest neighbor combinations were inserted in the same stem, $\left(\begin{smallmatrix}\text{GACW}\mathbf{XX}\text{WCUG}\\ \text{CUGW}\mathbf{XXX}\text{WGAC}\end{smallmatrix}\right)$, where W is a nearest neighbor nucleotide and X is a nucleotide in the 2 × 3 nucleotide loop. Three loop–nearest neighbor combinations were inserted into a different stem because of the possible formation of competing structures involving this standard stem.

Oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA). The purification of the oligonucleotides followed standard procedures,[33,34] as described previously.[35] Calculations of concentrations and the formation of duplexes from single strands are standard and were also described previously.[30]

Newly formed duplexes were lyophilized and redissolved in 1 M NaCl, 20 mM sodium cacodylate, and 0.5 mM Na_2EDTA (pH 7.0). The melting scheme and details of the standard melting experiments were described previously.[30] *Meltwin*[36] was used to fit melting curves to a two-state model and to derive the duplex thermodynamics.[30]

Determination of the Contribution of 2 × 3 Nucleotide Loops to Duplex Thermodynamics. The measured free energy change for the duplex can be contributed to an initiation term, the free energy contribution of the 2 × 3 loop, and the sum of the nearest neighbor interactions of Watson–Crick base pairs in the stem.[37] For example

$$\begin{aligned}\Delta G^\circ{}_{37}\begin{pmatrix}\text{GACG}\mathbf{CC}\text{CCUG}\\ \text{CUGC}\mathbf{AAU}\text{GGAC}\end{pmatrix}&\\ = \Delta G^\circ{}_{37,\mathrm{i}} + \Delta G^\circ{}_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix}&\\ + \Delta G^\circ{}_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} + \Delta G^\circ{}_{37,2\times3} + \Delta G^\circ{}_{37}\begin{pmatrix}\text{CC}\\ \text{GG}\end{pmatrix}&\\ + \Delta G^\circ{}_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} + \Delta G^\circ{}_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix}& \qquad (2)\end{aligned}$$

where $\Delta G^\circ{}_{37,\mathrm{i}}$ is the free energy change for duplex initiation, 4.09 kcal/mol,[37] $\Delta G^\circ{}_{37,2\times3}$ is the free energy contribution from the 2 × 3 nucleotide loop, and the remainder of the terms are individual nearest neighbor values.[37] A rearrangement of the terms in this equation to solve for the stability of the 2 × 3 loop results in eq 3:

$$\begin{aligned}\Delta G^\circ{}_{37,2\times3} = \Delta G^\circ{}_{37}\begin{pmatrix}\text{GACG}\mathbf{CC}\text{CCUG}\\ \text{CUGC}\mathbf{AAU}\text{GGAC}\end{pmatrix}&\\ - \Delta G^\circ{}_{37,\mathrm{i}} - \Delta G^\circ{}_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} - \Delta G^\circ{}_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix}&\\ - \Delta G^\circ{}_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} - \Delta G^\circ{}_{37}\begin{pmatrix}\text{CC}\\ \text{GG}\end{pmatrix}&\\ - \Delta G^\circ{}_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} - \Delta G^\circ{}_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix}& \qquad (3)\end{aligned}$$

where $\Delta G^\circ{}_{37}\left(\begin{smallmatrix}\text{GACG}\mathbf{CC}\text{CCUG}\\ \text{CUGC}\mathbf{AAU}\text{GGAC}\end{smallmatrix}\right)$) is the value determined by optical melting experiments. More specifically

$$\begin{aligned}\Delta G^\circ{}_{37,2\times3} &= -7.17 - 4.09 - (-2.35) - (-2.24)\\ &\quad - (-2.36) - (-3.26) - (-2.08) - (-2.11)\\ &= 3.14\ \text{kcal/mol} \qquad (4)\end{aligned}$$

This method of calculating $\Delta G^\circ{}_{37,2\times3}$ values utilizes nearest neighbor parameters.[37] A different method of calculating $\Delta G^\circ{}_{37,2\times3}$ values utilizes reference duplexes. Reference duplexes were not used here because the stability of an additional duplex, the reference duplex, would need to be measured to calculate the thermodynamic contribution of each of the 2 × 3 nucleotide loops studied here. Therefore, all of the $\Delta G^\circ{}_{37,2\times3}$ values reported here were calculated by the nearest neighbor method. For loops previously studied,[7,20,21,38] regardless of how the loop data were calculated in the original publications, $\Delta G^\circ{}_{37,2\times3}$ values were also calculated by equations similar to eq 4, not from reference duplexes. Similar calculations were used for $\Delta H^\circ{}_{2\times3}$ and $\Delta S^\circ{}_{2\times3}$.

Chen and Turner thermodynamically characterized eight unique 2 × 3 nucleotide internal loops,[19] but all of these loops were inserted into a stem with a dangling purine nucleotide. Reference duplexes were used to calculate the contribution of the 2 × 3 nucleotide loops to duplex stability. Because the nearest neighbor method was used to calculate the contribution of the 2 × 3 nucleotide loops studied here to duplex stability (eqs 3 and 4), we also wanted to use that same method with the data of Chen and Turner. Unfortunately, there are no reliable nearest neighbor parameters for dangling purine nucleotides. As a result, these loops were omitted from the analysis described here.

One additional previously measured loop[20] was omitted from the analysis. On the basis of an analysis of the sequence and possible competing secondary structures, we were not confident that the duplex $\left(\begin{smallmatrix}\text{ACCU}\mathbf{GC}\text{UUGC}\\ \text{UGGA}\mathbf{ACA}\text{AACG}\end{smallmatrix}\right)$ was forming the desired 2 × 3 internal loop (other competing structures from the association of the two strands are possible); therefore, it was omitted from the analysis.

Linear Regression and Updated Predictive Model. The melting of two duplexes studied here, $\left(\begin{smallmatrix}\text{GACG}\mathbf{GA}\text{GCUG}\\ \text{CUGC}\mathbf{AUA}\text{CGAC}\end{smallmatrix}\right)$ and $\left(\begin{smallmatrix}\text{GACU}\mathbf{AA}\text{ACUG}\\ \text{CUGA}\mathbf{GAC}\text{UGAC}\end{smallmatrix}\right)$, was considered non-two-state. For $\left(\begin{smallmatrix}\text{GACG}\mathbf{GA}\text{GCUG}\\ \text{CUGC}\mathbf{AUA}\text{CGAC}\end{smallmatrix}\right)$, the data derived from the individual melting curves and from the T_M versus C_T plot resulted in large error values. For $\left(\begin{smallmatrix}\text{GACU}\mathbf{AA}\text{ACUG}\\ \text{CUGA}\mathbf{GAC}\text{UGAC}\end{smallmatrix}\right)$, the ΔH° values derived from the analysis of the melt curve fits and the analysis of the T_m dependence did not agree within 15%. As a result, these two duplexes were not included in the analysis described below. To create a larger data set of 2 × 3 nucleotide loop data, the data for the 15 sequences studied here were combined with the data for 31 previously published sequences,[7,20,21,38] resulting in a data set of 46 sequences containing 2 × 3 nucleotide internal loops. Data from these 46 sequences were subjected to linear regression using *Microsoft Excel*'s LINEST function. The loop thermodynamic values were used as constants, and many different combinations of parameters (using a variety of bonuses and penalties) were tried as variables. The combination of variables (which is similar to those in eq 1) that resulted in a model that best agreed with the experimental data included a term for 2 × 3 nucleotide loop initiation, an A-

Table 2. Thermodynamic Parameters for Duplex Formation[a]

		Analysis of Melt Curve Fit/Errors				Analysis of Tm Dependence/ Errors (ln Plot)			
Frequency[b]	Sequence	-ΔH° (kcal/mol)	-ΔS° (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)	-ΔH° (kcal/mol)	-ΔS° (cal/K·mol)	$-\Delta G°_{37}$ (kcal/mol)	T_m[c] (°C)
90	GACG**CC**CCUG CUGC**AAU**GGAC	69.4 ± 6.1	200.7 ± 19.7	7.21 ± 0.08	40.1	73.2 ± 4.1	213.0 ± 13.3	7.17 ± 0.06	39.7
39	GUCA**AA**CCAG CAGU**CAC**GGUC	77.1 ± 5.6	227.0 ± 18.0	6.72 ± 0.12	37.8	77.2 ± 7.2	227.3 ± 23.3	6.70 ± 0.15	37.7
31	GUCA**CA**GCAG CAGU**AUA**CGUC	58.0 ± 6.0	165.7 ± 19.2	6.63 ± 0.22	37.5	53.5 ± 8.2	151.0 ± 26.8	6.66 ± 0.42	37.8
18	GACG**AU**GCUG CUGC**CAU**CGAC	84.5 ± 6.9	243.3 ± 22.1	9.02 ± 0.19	46.4	87.5 ± 6.6	253.0 ± 20.9	9.06 ± 0.19	46.2
18	GACG**AU**GCAG CUGC**CGU**CGUC	79.2 ± 7.1	226.6 ± 22.3	8.90 ± 0.23	46.6	79.1 ± 3.1	226.6 ± 9.9	8.88 ± 0.08	46.5
17	GACU**AU**CCUG CUGG**GGU**GGAC	81.8 ± 8.3	238.6 ± 26.0	7.74 ± 0.37	41.7	86.8 ± 13.1	255.0 ± 41.9	7.75 ± 0.44	41.4
16	GACU**AA**ACUG CUGA**GAC**UGAC	(62.3)	(180.9)	(6.22)	(35.4)	(45.4)	(126.2)	(6.27)	(35.2)
13	GACU**GG**GCUG CUGG**GGA**UGAC	46.7 ± 12.5	129.9 ± 39.7	6.43 ± 0.30	36.3	42.9 ± 6.0	117.9 ± 19.5	6.36 ± 0.35	35.8
12	CUGU**AU**GACG[d] GACG**AAU**CUGC	69.4 ± 3.3	202.2 ± 10.9	6.6 ± 0.1	37.5	77.7 ± 2.2	229.3 ± 7.0	6.6 ± 0.1	37.2
11	GACG**GA**GCUG CUGC**AUA**CGAC	(24.5)	(53.5)	(7.92)	(55.7)	(23.5)	(50.7)	(7.82)	(54.9)
10	GACG**GA**GCUG CUGC**AAA**CGAC	82.1 ± 2.1	235.6 ± 6.6	9.06 ± 0.12	46.9	83.3 ± 4.5	239.2 ± 14.2	9.09 ± 0.12	46.8
10	GACC**GC**UCUG CUGG**AAA**AGAC	66.4 ± 4.3	190.4 ± 13.7	7.37 ± 0.12	41.0	67.7 ± 3.4	194.5 ± 10.8	7.36 ± 0.05	40.9
9	GACU**AU**CCUG CUGG**AAU**GGAC	80.2 ± 5.1	234.9 ± 16.3	7.36 ± 0.10	40.3	79.3 ± 2.1	232.0 ± 6.7	7.34 ± 0.02	40.2
7	GACA**AC**ACAG CUGU**GAU**UGUC	67.5 ± 4.2	201.0 ± 14.3	5.15 ± 0.27	30.8	67.1 ± 5.2	199.8 ± 17.2	5.14 ± 0.17	30.7
6	GACA**CC**ACUG CUGU**UAA**UGAC	75.4 ± 5.4	230.9 ± 18.0	3.81 ± 0.26	26.2	76.0 ± 3.4	232.6 ± 11.3	3.81 ± 0.14	26.3
6	GUCA**CC**GCUG CAGU**UAA**CGAC	62.4 ± 3.6	181.4 ± 12.0	6.11 ± 0.10	34.9	62.4 ± 1.2	181.4 ± 3.8	6.11 ± 0.02	34.9
6	GACU**GG**GCUG CUGG**AGA**UGAC	75.9 ± 8.2	228.3 ± 26.5	5.04 ± 0.20	31.0	74.7 ± 9.3	224.3 ± 30.8	5.10 ± 0.38	31.2
5	GACU**GG**GCUG CUGG**AGA**CGAC	58.5 ± 11.3	167.8 ± 36.6	6.47 ± 0.59	36.7	59.9 ± 10.4	172.1 ± 33.5	6.46 ± 0.58	36.6

[a] Measurements were taken in 1.0 M NaCl, 10 mM sodium cacodylate, and 0.5 mM Na_2EDTA (pH 7.0). Values in parentheses are approximate because of 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 an oligomer concentration of 10^{-4} M. [d] From ref 38.

U or G-U closure, a 5′RG3′/3′YA5′ stack, a 5′YG3′/3′RA5′ stack, a potential G·G pair, and a potential U·U pair. Internal loop thermodynamic parameters were derived for $\Delta G°_{37,2\times3}$, $\Delta H°_{2\times3}$, and $\Delta S°_{2\times3}$.

■ RESULTS

Database Searching. The database containing 1349 structures (described in Materials and Methods) was searched for 2 × 3 nucleotide loops. As a result, 666 2 × 3 nucleotide loops were found. On average, about one 2 × 3 nucleotide loop was found in every two structures. A summary of database results is shown in Table 1.

The first set of data in Table 1 represents the loops when the loop nucleotides and the nearest neighbor canonical base pairs are specified, resulting in a total of 252 different types of loops in the database. Many of the different types of loops do not occur frequently. In fact, 227 types of loops, accounting for 45% of the loops found in the database, individually account for ≤0.6% of the total number of loops found. Table 1 shows the frequency and percent occurrence of only the 25 most frequently occurring 2 × 3 nucleotide loops, which represent the other 55% of the 2 × 3 nucleotide loops found in the

Table 3. Contributions of Loops to Duplex Thermodynamics[a]

Frequency[b]	Sequence	ΔH°_{2x3} (kcal/mol)	ΔS°_{2x3} (cal/K·mol)	$\Delta G^\circ_{37,2x3}$ (kcal/mol)	Frequency[b]	Sequence	ΔH°_{2x3} (kcal/mol)	ΔS°_{2x3} (cal/K·mol)	$\Delta G^\circ_{37,2x3}$ (kcal/mol)
90	GACG**CC**CCUG CUGC**AAU**GGAC	-8.0 *-6.7*	-36.1 *-30.0*	3.14 *2.5*	0	UGAC**UU**CUCA[c] ACUG**UUU**GAGU	-17.7 *-26.1*	-59.9 *-87.0*	0.89 *0.7*
39	GUCA**AA**CCAG CAGU**CAC**GGUC	-12.2 *-0.4*	-50.2 *-12.6*	3.36 *3.4*	0	GAGC**AG**CGAC[c] CUCG**GAA**GCUG	-10.6 *-6.7*	-40.4 *-30.0*	1.91 *2.5*
31	GUCA**CA**GCAG CAGU**AUA**CGUC	13.0 *-0.4*	-30.3 *-12.6*	3.56 *3.4*	0	GAGC**GA**CGAC[c] CUCG**AAA**GCUG	-6.4 *-11.6*	-27.1 *-43.4*	1.94 *1.8*
18	GACG**AU**GCUG CUGC**CAU**CGAC	-20.8 *-16.4*	-71.9 *-58.5*	1.41 *1.6*	0	CCAC**GG**CUCC[c] GGUG**AAA**GAGG	-10.3 *-11.6*	-39.5 *-43.4*	1.99 *1.8*
18	GACG**AU**GCAG CUGC**CGU**CGUC	-12.4 *-16.4*	-45.5 *-58.5*	1.59 *1.6*	0	GAGC**AA**CGAC[c] CUCG**GAA**GCUG	-9.4 *-6.7*	-37.2 *-30.0*	2.10 *2.5*
17	GACU**AU**CCUG CUGG**GGU**GGAC	-20.2 *-10.1*	-72.6 *-41.1*	2.31 *2.5*	0	UGAC**UU**CUCA[c] ACUG**CUU**GAGU	-16.8 *-16.4*	-58.0 *-58.5*	1.26 *1.6*
13	GACU**GG**GCUG CUGG**GGA**UGAC	23.0 *10.5*	64.3 22.0	2.95 *3.5*	0	GAGC**AG**CGAC[c] CUCG**AAG**GCUG	-8.7 *-2.1*	-35.5 *-12.8*	2.21 *1.7*
12	CUGU**AU**GACG[c] GACG**AAU**CUGC	-13.3 *-10.1*	-52.6 *-41.1*	2.96 *2.5*	0	UCAC**UU**CUGA[c] AGUG**CUC**GACU	-12.8 *-6.7*	-50.5 *-30.0*	2.90 *2.5*
10	GACG**GA**GCUG CUGC**AAA**CGAC	-16.6 *-16.5*	-58.1 *-57.1*	1.38 *1.1*	0	UCAG**CC**GUGA[c] AGUC**AAU**CACU	1.0 *-6.7*	-6.8 *-30.0*	3.14 *2.5*
10	GACC**GC**UCUG CUGG**AAA**AGAC	-0.7 *-5.3*	-11.8 *-26.0*	2.94 *2.7*	0	UGAG**AA**GUCA[e] ACUC**AAA**CAGU	-3.9 *-6.7*	-17.8 *-30.0*	1.66 *2.5*
9	GACU**AU**CCUG CUGG**AAU**GGAC	-12.7 *-10.1*	-49.6 *-41.1*	2.72 *2.5*	0	UGAG**AA**GUCA[c] ACUC**CGA**CAGU	-4.2 *-6.7*	-18.9 *-30.0*	1.67 *2.5*
7	GACA**AC**ACAG CUGU**GAU**UGUC	-4.1 5.9	-25.9 *4.8*	3.90 *4.3*	0	GAGC**AG**CGAC[c] CUCG**AAA**GCUG	-8.8 *-6.7*	-36.8 -30.0	2.57 *2.5*
6	GACA**CC**ACUG CUGU**UAA**UGAC	-13.0 *5.9*	-58.7 *4.8*	5.23 *4.3*	0	GAGC**AA**CGAC[c] CUCG**AAA**GCUG	-2.1 *-6.7*	-15.8 *-30.0*	2.72 *2.5*
6	GUCA**CC**GCUG CAGU**UAA**CGAC	4.1 *-0.4*	-0.1 *-12.6*	4.11 *3.4*	0	UGAC**UU**CUCA[c] ACUG**CCU**GAGU	-17.1 *-16.4*	-61.4 *-58.5*	1.93 *1.6*
6	GACU**GG**GCUG CUGG**AGA**UGAC	-8.9 *1.0*	-42.1 *-8.6*	4.21 *3.6*	0	GAGU**AA**CGAC[f] CUCG**AAG**GCUG	-17.3 *-10.2*	-61.9 *-39.7*	1.80 *2.0*
5	GACU**GG**GCUG CUGG**AGA**CGAC	8.2 *-5.3*	14.5 *-26.0*	3.76 *2.7*	0	GAGU**GA**UGAC[f] CUCG**AAG**GCUG	-14.3 *-8.8*	-52.1 *-35.7*	1.85 *2.2*
3	CGAC**GA**GCAG[d] GCUG**AAG**CGUC	-26.9 *-16.5*	-89.2 *-56.8*	0.77 *1.1*	0	GAGC**AA**UGAC[f] CUCG**AAG**GCUG	-15.9 *-10.2*	-58.7 *-39.7*	2.26 *2.0*
2	CUGU**GG**ACGA[c] GACG**AGA**UGCU	1.2 *1.0*	-8.0 *-8.6*	3.66 *3.6*	0	GAGU**AA**CGAC[f] CUCG**AAA**GCUG	-4.7 *-0.4*	-25.6 *-12.6*	3.20 *3.4*
0	GAGC**GA**CGAC[c] CUCG**AAG**GCUG	-18.0 *-21.4*	-59.0 *-70.5*	0.20 *0.4*	0	GAGU**AA**UGAC[f] CUCG**AAG**GCUG	0.6 *-3.9*	-8.3 *-22.3*	3.25 *2.9*
0	CCAC**GG**CUCC[c] GGUG**AGA**GAGG	-14.0 *-11.6*	-49.5 *-43.4*	1.37 *1.8*	0	GAGU**GA**UGAC[f] CUCG**AAA**GCUG	11.4 *1.0*	25.7 *-8.6*	3.35 *3.6*
0	GAGC**AA**CGAC[c] CUCG**AAG**GCUG	-6.0 *-16.5*	-24.2 *-57.1*	1.48 1.1	0	GAGC**AA**UGAC[f] CUCG**AAA**GCUG	-0.5 *-0.4*	-13.1 *-12.6*	3.56 *3.4*
0	CGAC**GA**GCAG[c] GCUG**GAA**CGUC	-8.0 *-2.1*	-32.4 *-12.8*	2.04 *1.7*	0	GAGU**AG**UGAC[f] CUCG**AAA**GCUG	15.3 *5.9*	35.9 *4.8*	4.05 *4.3*
0	CCUCU**GC**GGUGA[c] GGAGA**AAA**CCGC	-5.0 *-5.3*	-26.4 *-26.0*	3.18 *2.7*	0	GAGU**AA**UGAC[f] CUCG**AAA**GCUG	17.9 *5.9*	44.0 *4.8*	4.15 *4.3*

[a]Values in italics are predicted values based on the model derived here. [b]Frequency of occurrence in the database described in Materials and Methods. [c]Derived from raw data available in ref 20. [d]Derived from raw data available in ref 7. [e]Derived from raw data available in ref 21. [f]Derived from raw data available in ref 38.

database. The most frequent loop is $\left(\begin{matrix}\text{GCCC}\\ \text{CAAUG}\end{matrix}\right)$, which accounts for 14% of all 2 × 3 nucleotide loops found in the database. Although it was pioneering work, the initial thermodynamic experiments on 2 × 3 nucleotide internal loops[7,20,21,38] only began to study the possible sequence combinations. When specifying the loop nucleotides and the nearest neighbor base pairs, previous studies have only characterized one type of loop in the top 25, representing only 3% of the total number of loops found, while the data collected here increase this to 18 loops characterized in the top 25, representing 49% of the total number of loops found.

The second set of data in Table 1 lists the frequency and percent occurrence when only the loop nucleotides are specified. When they are characterized in this fashion, a total of 135 different types of loops were found. In Table 1, only the 25 most frequently occurring 2 × 3 nucleotide loops, which account for 71% of all possible types of 2 × 3 nucleotide loops, are shown. The remaining 110 2 × 3 nucleotide loops represent the remaining 29% of the 2 × 3 nucleotide loops found in the database, individually representing <0.9% of the total number of loops found. The most frequent loop is $\left(\begin{matrix}\text{CC}\\ \text{AAU}\end{matrix}\right)$, accounting for 14% of the 2 × 3 nucleotide loops found in the database. Previous studies have characterized only eight different types of loops, representing only 30% of the 2 × 3 nucleotide loops, while the data collected here increase this to 19 different types of loops, representing 64% of the 2 × 3 nucleotide loops.

The third set of data in Table 1 lists the frequency and percent occurrence of 5′ and 3′ nearest neighbor combinations. This analysis of the data results in 33 of 36 possible types of nearest neighbor combinations. The three possible nearest neighbor combinations not found in the database all contain a G-U pair. The 25 nearest neighbor combinations listed in the third data set account for 97% of the total number of combinations found, while the remaining eight nearest neighbor combinations account for the remaining 3%. The most frequently occurring nearest neighbor combination is $\left(\begin{matrix}\text{GXXC}\\ \text{CXXXG}\end{matrix}\right)$, representing 16% of all nearest neighbor combinations found adjacent to 2 × 3 nucleotide loops. As shown in Table 1, previous studies account for only 43% of the possible nearest neighbor combinations. Once the data reported here were included, this percentage increased to 82%. Similarly, previous studies only thermodynamically characterized eight nearest neighbor combinations in the top 25. Once the data reported here were included, 14 nearest neighbor combinations in the top 25 were studied.

The fourth set of data in Table 1 gives the frequency and percent occurrence of the loop nucleotides when cytidine and uridine are classified as pyrimidines and adenosine and guanosine are classified as purines. This arrangement of data results in 25 different types of loop sequences that are found in the database, all of which are listed in Table 1. An all-purine loop is the most frequently occurring combination, representing 19% of all 2 × 3 nucleotide loops found in the database. Previous studies account for 52%, while the data reported here increase the percentage to 82%.

Thermodynamic Parameters. Thermodynamic parameters for duplex formation derived from an analysis of the individual melt curves and an analysis of the T_M dependence are listed in Table 2. The data listed in Table 2 correspond to 18 of the 19 most frequently occurring loops found in the database (Table 1, first set of data) and are listed in order of decreasing frequency. As previously stated, one sequence with three adjacent guanines was not studied because of possible competing structures.

Contribution of 2 × 3 Nucleotide Loops to Duplex Thermodynamics. The contribution of 2 × 3 nucleotide loops to duplex thermodynamics (Table 3) was calculated as described in Materials and Methods and as defined by eqs 3 and 4 for $\Delta G°_{37,2\times3}$. In addition to 16 duplexes studied here (and listed in Table 2), 30 additional duplexes that occur less frequently but were studied previously[7,20,21,38] are also included.

Free Energy Parameters for 2 × 3 Nucleotide Internal Loops. While analyzing the data listed in Table 3 by linear regression, we tested several models (a variety of penalty and bonus combinations) to predict the thermodynamics of 2 × 3 nucleotide loops. A combination of parameters and their corresponding values (Table 4) similar to those used in eq 1 produced an equation that resulted in predicted $\Delta G°_{37,2\times3}$ values that were very close to the experimental $\Delta G°_{37,2\times3}$ values reported here:

Table 4. 2 × 3 Nucleotide Loop Nearest Neighbor Parameters at 37 °C[a]

	$\Delta H°_{2x3}$ (kcal/mol)	$\Delta S°_{2x3}$ (cal/K·mol)	$\Delta G°_{37,2x3}$ (kcal/mol)
2x3 Loop Initiation[a]	-6.7 ± 2.1 (-3.6 ± 1.6)[b]	-30.0 ± 6.7	2.5 ± 0.1 (2.6 ± 0.1)[b] (2.6 ± 0.2)[c]
A-U or G-U Closure[d]	6.3 ± 1.5 (5.0 ± 0.7)[b]	17.4 ± 4.8	0.9 ± 0.1 (0.7 ± 0.1)[b] (0.7 ± 0.1)[c]
5′YA[e] 3′RG	n/a (-5.7 ± 3.8)[b]	n/a	n/a (-0.5 ± 0.2)[b] (-0.4 ± 0.2)[c]
5′RG[e] 3′YA	-9.8 ± 3.3 (-10.9 ± 2.7)[b]	-27.1 ± 10.7	-1.4 ± 0.2 (-1.2 ± 0.1)[b] (-1.1 ± 0.2)[c]
5′YG[e] 3′RA	-4.9 ± 2.5 (-8.6 ± 1.9)[b]	-13.4 ± 7.9	-0.7 ± 0.1 (-1.1 ± 0.1)[b] (-1.4 ± 0.1)[c]
G·G Mismatch[f]	-4.6 ± 4.9 (-9.0 ± 4.6)[b]	17.2 ± 15.7	-0.8 ± 0.3 (-0.7 ± 0.2)[b] (-0.7 ± 0.3)[c]
U·U Mismatch[f]	-9.7 ± 2.8 (-6.4 ± 2.5)[b]	-28.5 ± 9.1	-0.9 ± 0.2 (-0.4 ± 0.1)[b] (-0.3 ± 0.2)[c]

[a] These parameters apply to all 2 × 3 loops, regardless of sequence. [b] Numbers in parentheses are values proposed in ref 18. The initiation value listed here is the sum of the initiation value and asymmetry penalty from the original reference. [c] Numbers in parentheses are values proposed in ref 19. The initiation value listed here is the sum of the initiation value and asymmetry penalty from the original reference. [d] These parameters are applied per A-U or G-U closure. [e] These parameters are applied per each closing pair—first mismatch or closing pair—last mismatch combination. [f] These parameters are applied per each first—last mismatch.

$$\begin{aligned}\Delta G°_{37,2\times3} &= \Delta G°_{37,2\times3\ \text{initiation}} + \Delta G°_{37,\text{AU/GU}}\\ &\quad + \Delta G°_{37,5'\text{RG}/3'\text{YA}} + \Delta G°_{37,5'\text{YG}/3'\text{RA}}\\ &\quad + \Delta G°_{37,\text{GG}} + \Delta G°_{37,\text{UU}} \qquad (5)\end{aligned}$$

Note that the $\Delta G°_{37,2\times3\ \text{initiation}}$ and $\Delta G°_{37,\text{asym}}$ terms in eq 1 were combined into the $\Delta G°_{37,2\times3\ \text{initiation}}$ term in eq 5. Also, the $\Delta G°_{37,5'\text{YA}/3'\text{RG}}$ term in eq 1 was not included in this new model because it was not found to contribute significantly to loop stability. The values corresponding to each of these parameters can be found in Table 4. Most of the new $\Delta G^\circ_{37,2\times3}$ parameters are within experimental error of the previously proposed values.[18,19] The same set of parameters was used in a linear regression analysis of the enthalpy and entropy data. All of the values for the new $\Delta H^\circ_{2\times3}$ parameters (Table 4) are within experimental error of previous literature values.[18] All of the values for the $\Delta S^\circ_{2\times3}$ parameters are also listed in Table 4.

■ DISCUSSION

This study is one in a series of studies investigating the thermodynamics of small RNA secondary structure motifs that occur frequently in nature. This study focuses on 2 × 3 internal loops. A database of 1349 RNA secondary structures was used to determine the frequency of 2 × 3 nucleotide loops found in naturally occurring RNA (Table 1). The high frequency of occurrence demonstrates the interest in and importance of this particular secondary structure motif. Furthermore, this confirms the importance of generating an updated model for predicting the stability of 2 × 3 internal loops, which may lead to more accurate prediction of secondary structure from sequence. Although 2 × 3 nucleotide loops are common in nature, relatively few studies have investigated their thermodynamics,[7,19−21,38] and most of the loops investigated in these studies were not found in the database compiled (Table 1). As a result, stabilities of many commonly occurring loops are based on the predictive model derived from this data set of 23 2 × 3 loops that rarely occur in the database. In this study, 17 frequently occurring 2 × 3 nucleotide loops were thermodynamically characterized to provide experimental values for frequently occurring loops and to increase the size of the data set of experimental data from which updated predictive parameters can be derived for those 2 × 3 loops still without experimental values. It is our hope that the additional experimental values and the updated predictive model can be used to improve the prediction of secondary structure from sequence.

Database Searching. The database that was searched during this study contained 1349 secondary structures, representing eight different types of RNA. Although this database is not all-inclusive, it provides loop sequences that occur in nature.

Although it was pioneering work, the initial thermodynamic experiments on 2 × 3 nucleotide internal loops[7,20,21,38] only began to study the possible sequence combinations. After the addition of the data reported here, the percentage of loop sequences with nearest neighbors defined (Table 1, data set 1), loop sequences without the nearest neighbors defined (Table 1, data set 2), nearest neighbors without loop sequences (Table 1, data set 3), and loop sequences classified as purines or pyrimidines (Table 1, data set 4) now studied has increased significantly from the percentages previously studied. It is apparent from the first data set in Table 1 that few of the 2 × 3 loops found in the database had been thermodynamically characterized previously. In fact, only one of the top 25 most frequently occurring 2 × 3 loops had been studied. After the addition of the results reported here, 18 combinations in the top 25 have been thermodynamically characterized. It is interesting to note that the three most frequently occurring loop sequences, $\binom{\text{GCCC}}{\text{CAAUG}}$, $\binom{\text{AAAC}}{\text{UCACG}}$, and $\binom{\text{ACAG}}{\text{UAUAC}}$, do not contain mismatches that have previously been considered stabilizing (G·G, A·G, or U·U).[22,39] Similar to what was observed for 1 × 2 nucleotide internal loops,[30] the presence or absence of stabilizing mismatches does not determine the frequency of occurrence for 2 × 3 nucleotide loops.

The second set of data in Table 1 shows that 135 loop sequences were found in this database. It is not immediately obvious why these loop sequences are found in the database while the other 66% of the possible loop sequences are not. Perhaps there is a structural explanation for this trend, but an extensive structural study would need to be done. Or, these loops may be found in secondary structures that were not included in this database.

The third set of data in Table 1 shows that 33 of the 36 possible nearest neighbor combinations were found in the database. $\binom{\text{GXXC}}{\text{CXXXG}}$ is the most frequent nearest neighbor combination, representing ∼16% of the total number of loops, followed closely by $\binom{\text{GXXG}}{\text{CXXXC}}$, also representing ∼16% of the total number of loops. As shown here, when the 3′ adjacent pair is flipped from C-G to G-C, there is relatively no difference in the number of loops that have either combination. However, if the 5′ adjacent pair of $\binom{\text{GXXC}}{\text{CXXXG}}$ is flipped from G-C to C-G, the adjacent base pair combination goes from being the most frequent to the 17th most frequent. There are three possible nearest neighbor combinations that were not found in the database, $\binom{\text{GXXG}}{\text{UXXXU}}$, $\binom{\text{GXXU}}{\text{UXXXG}}$, and $\binom{\text{UXXU}}{\text{AXXXG}}$, all of which contain a G-U or U-G pair.

Thermodynamic Contribution of 2 × 3 Nucleotide Loops to Duplex Thermodynamics. The thermodynamic contribution of a 2 × 3 loop to duplex thermodynamics is varied. For example, contributions of loops to enthalpy, entropy, and free energy changes of the duplex range from −26.9 to 23.0 kcal/mol, −89.2 to 64.3 cal K^{-1} mol^{-1}, and 0.2 to 5.2 kcal/mol, respectively (Table 3). The frequency with which a loop was found in the database does not correlate with the stability of the loop. For example, two of the three most stable loops were not found in the database; the remaining was the 32nd most common in the database.

As expected, all of the internal loops studied here contribute unfavorably to the stability of the duplex (Table 3). The nine most stable loops have sequences that could possibly form G·A or U·U pairs, which have been previously shown to stabilize mismatches and loops.[30,39−44] However, some loops with possible G·A or U·U pairs are actually less stable than those with no possible G·A or U·U pairs, showing that free energy contributions are not completely influenced by the possible pairs that could form.

There does seem to be a correlation between the number of G-C or C-G pairs directly adjacent to the loop and the free energy contribution of the loop to duplex stability. For example, the 23 loops with two adjacent G-C or C-G pairs contribute an average of 2.2 kcal/mol to duplex stability. The 12 loops with only one G-C or C-G adjacent base pair and the 11 loops with no adjacent G-C or C-G base pairs contribute an average of 2.8 and 3.1 kcal/mol to duplex stability, respectively.

Derivation of an Updated Model for Predicting Loop Thermodynamics. *RNAstructure* currently uses a predictive model to approximate the free energy contribution of all 2 × 3 nucleotide loops (see eq 1). Because the size of the data set of available 2 × 3 nucleotide loop thermodynamics increased with the addition of the data reported here, this larger data set was used to derive an updated model to predict 2 × 3 thermodynamics. Several models were generated by using different combinations of parameters (data not shown). The model that agreed best with the experimental data, resulted in linearly independent parameters, and had small standard deviations was one using parameters similar to those of the current *RNAstructure* model (eq 1). In fact, some of the new parameters are within experimental error of those of the current *RNAstructure* model.

In the current *RNAstructure* model, $\Delta G^\circ_{37,\text{loop initiation}}$ is 2.0 kcal/mol and was derived from a data set of 23 2 × 3 nucleotide loops.[18,22] $\Delta G^\circ_{37,\text{asym}}$, a penalty of 0.6 kcal/mol for being an asymmetric internal loop, was derived from data of loops of various sizes.[18] Resulting from a linear regression analysis with the larger data set of 2 × 3 thermodynamics, a new parameter was derived, $\Delta G^\circ_{37,2\times3\text{ initiation}}$. This value, 2.5 ± 0.1 kcal/mol, is within experimental error of the sum of $\Delta G^\circ_{37,\text{loop initiation}}$ and $\Delta G^\circ_{37,\text{asym}}$ (Table 4), which are used by the current *RNAstructure* model.

In the current *RNAstructure* model, there is a penalty ($\Delta G^\circ_{37,\text{AU/GU}}$) for replacing a closing G-C or C-G pair with a closing A-U, U-A, G-U, or U-G pair. This value was derived from data of loops of various sizes.[18] From the new data set, linear regression was used to derive a $\Delta G^\circ_{37,\text{AU/GU}}$ term based only on the 2 × 3 nucleotide loop data. This value, 0.9 ± 0.1 kcal/mol, is within experimental error of *RNAstructure*'s current penalty, 0.7 ± 0.1 kcal/mol.[18]

One difference between the *RNAstructure* parameters and the parameters proposed here is the term for a 5′YA3′/3′RG5′ bonus for the first and last mismatch. *RNAstructure* assigns a bonus of −0.5 ± 0.2 kcal/mol to this stack. Similarly, the Chen and Turner model[19] assigns a bonus of −0.4 ± 0.2 kcal/mol to this stack. With the data set used here, this stack did not contribute to 2 × 3 nucleotide loop stability. The 5′RG3′/3′YA5′ bonus for the first−last mismatch derived here (−1.4 ± 0.2 kcal/mol) was within experimental error of the same parameter used by *RNAstructure* (−1.2 ± 0.1 kcal/mol). Lastly, the bonus for a 5′YG3′/3′RA5′ first−last mismatch derived here (−0.7 ± 0.1 kcal/mol) is smaller than the bonus currently used by *RNAstructure* (−1.1 ± 0.1 kcal/mol).

The final parameters used by the current model in *RNAstructure* are bonuses for stabilizing mismatches in the loop.[18] *RNAstructure* includes a −0.7 ± 0.2 kcal/mol bonus for potential G·G pairs and a −0.4 ± 0.1 kcal/mol bonus for potential U·U pairs. The model proposed here contains a G·G bonus of −0.8 ± 0.3 kcal/mol, which is within experimental error of the *RNAstructure* value, and a more favorable U·U bonus of −0.9 ± 0.2 kcal/mol.

The values for the enthalpy parameters derived here can also be compared to the enthalpy parameters published previously.[18] Except for the omission of the 5′YA3′/3′RG5′ parameter in the model proposed here, the values for all of the other parameters are within experimental error of the previous values (Table 4). Although the size of the 2 × 3 loop data set was increased, we were unable to derive a predictive model that was much different from the one currently used by *RNAstructure*. Although both the predictive model used currently by *RNAstructure* and the updated predictive model derived here work quite well for the data set of 2 × 3 internal loops, adding a look-up table with experimental values for 2 × 3 nucleotide loops may improve free energy calculations and the prediction of secondary structure from sequence.

AUTHOR INFORMATION

Corresponding Author

*Phone: (314) 977-8567. Fax: (314) 977-2521. E-mail: znoskob@slu.edu.

Funding

This work was supported by Grant R15GM08569 from the National Institute of General Medical Sciences.

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

R, purine nucleotides; Y, pyrimidine nucleotides.

REFERENCES

(1) 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.

(2) Ippolito, J. A., and Steitz, T. A. (2000) The structure of the HIV-1 RRE high affinity Rev binding site at 1.6 angstrom resolution. *J. Mol. Biol. 295*, 711−717.

(3) Battiste, J. L., Mao, H., Rao, N. S., Tan, R., Muhandiram, D. R., Kay, L. E., Frankel, A. D., and Williamson, J. R. (1996) α Helix-RNA major groove recognition in an HIV-1 Rev peptide RRE RNA complex. *Science 273*, 1547−1551.

(4) Ban, N., Nissen, P., Hansen, J., Moore, P. B., and Steitz, T. A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science 289*, 905−920.

(5) Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature 407*, 340−348.

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

(7) Schroeder, S. J., Fountain, M. A., Kennedy, S. D., Lukavsky, P. J., Puglisi, J. D., Krugh, T. R., and Turner, D. H. (2003) Thermodynamic stability and structural features of the J4/5 loop in a *Pneumocystis carinii* group I intron. *Biochemistry 42*, 14184−14196.

(8) Batey, R. T., Gilbert, S. D., and Montange, R. K. (2004) Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature 432*, 411−415.

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

(10) 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.

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

(12) 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.

(13) Damberger, S. H., and Gutell, R. R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res. 22*, 3508−3510.

(14) Larsen, N., Samuelsson, T., and Zwieb, C. (1998) The signal recognition particle database (SRPDB). *Nucleic Acids Res. 26*, 177−178.

(15) Kondo, J., Urzhumtsev, A., and Westhof, E. (2006) Two conformational states in the crystal structure of the *Homo sapiens* cytoplasmic ribosomal decoding A site. *Nucleic Acids Res. 34*, 676−685.

(16) Gait, M. J., and Karn, J. (1993) RNA recogntion by the human immuno-deficiency virus Tat and Rev proteins. *Trends Biochem. Sci. 18*, 255–259.

(17) Gait, M. J., and Karn, J. (1995) Progress in anti-HIV structure-based drug design. *Trends Biotechnol. 13*, 430–438.

(18) 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.

(19) Chen, G., and Turner, D. H. (2006) Consecutive GA pairs stabilize medium-size RNA internal loops. *Biochemistry 45*, 4025–4043.

(20) Schroeder, S. J., and Turner, D. H. (2000) Factors affecting the thermodynamic stability of small asymmetric internal loops in RNA. *Biochemistry 39*, 9257–9274.

(21) 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.

(22) 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.

(23) Hofacker, I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res. 31*, 3429–3431.

(24) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406–3415.

(25) Markham, N. R., and Zuker, M. (2008) UNAFold: Software for nucleic acid folding and hybridization. *Methods Mol. Biol. 453*, 3–31.

(26) Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., and Murphy, K. P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics 23*, i19–i28.

(27) Davis, A. R., and Znosko, B. M. (2007) Thermodynamic characterization of single mismatches found in naturally occurring RNA. *Biochemistry 46*, 13425–13436.

(28) Davis, A. R., and Znosko, B. M. (2008) Thermodynamic characterization of naturally occurring RNA single mismatches with G-U nearest neighbors. *Biochemistry 47*, 10178–10187.

(29) Christiansen, M. E., and Znosko, B. M. (2009) Thermodynamic characterization of tandem mismatches found in naturally occurring RNA. *Nucleic Acids Res. 37*, 4696–4706.

(30) Badhwar, J., Karri, S., Cass, C. K., Wunderlich, E. L., and Znosko, B. M. (2007) Thermodynamic characterization of RNA duplexes containing naturally occurring 1 × 2 nucleotide internal loops. *Biochemistry 46*, 14715–14724.

(31) Thulasi, P., Pandya, L. K., and Znosko, B. M. (2010) Thermodynamic characterization of RNA triloops. *Biochemistry 49*, 9058–9062.

(32) Sheehy, J. P., Davis, A. R., and Znosko, B. M. (2010) Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA 16*, 417–429.

(33) Stawinski, J., Stromberg, R., Thelin, M., and Westman, E. (1988) Evaluation of the use of the tert-butyldimethylsilyl group for 2′-protection in RNA-synthesis via the H-phosphonate approach. *Nucleosides Nucleotides 7*, 779–782.

(34) Chou, S. H., Flynn, P., and Reid, B. (1989) Solid-phase synthesis and high-resolution NMR-studies of two synthetic double-helical RNA dodecamers: r(CGCGAAUUCGCG) and r-(CGCGUAUACGCG). *Biochemistry 28*, 2422–2435.

(35) 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.

(36) 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.

(37) 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.

(38) Schroeder, S. J., and Turner, D. H. (2001) Thermodynamic stabilities of internal loops with GU closing pairs in RNA. *Biochemistry 40*, 11509–11517.

(39) 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.

(40) 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.

(41) 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.

(42) 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.

(43) SantaLucia, J., Jr., and Turner, D. H. (1993) Structure of $(rGGCGAGCC)_2$ in solution from NMR and restrained molecular dynamics. *Biochemistry 32*, 12612–12623.

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