BIOCHEMISTRY
including biophysical chemistry & molecular biology

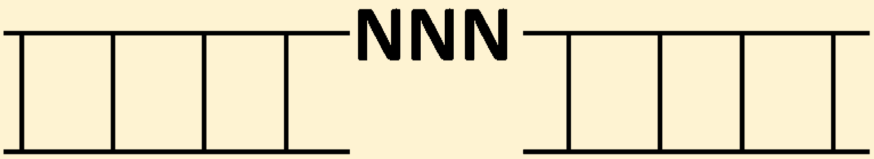
Article

pubs.acs.org/biochemistry

Improved Model to Predict the Free Energy Contribution of Trinucleotide Bulges to RNA Duplex Stability

Meghan H. Murray, Jessicah A. Hard, and Brent M. Znosko*

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



$$\Delta G^{\circ}_{37,\text{ trint bulge}} = (\Delta G^{\circ}_{\text{bulged nucleotides}}) + (\Delta G^{\circ}_{\text{GU or AU closing pair}})$$

$$\Delta G^{\circ}_{37,\text{trint bulge}} = (\Delta G^{\circ}_{3R} \text{ or } \Delta G^{\circ}_{3Y} \text{ or } \Delta G^{\circ}_{2R,1Y} \text{ or } \Delta G^{\circ}_{1R,2Y}) + (\Delta G^{\circ}_{GU} \text{ and/or } \Delta G^{\circ}_{AU})$$

ABSTRACT: Trinucleotide bulges in RNA commonly occur in nature. Yet, little data exists concerning the thermodynamic parameters of this motif. Algorithms that predict RNA secondary structure from sequence currently attribute a constant free energy value of 3.2 kcal/mol to all trinucleotide bulges, regardless of bulge sequence. To test the accuracy of this model, RNA duplexes that contain frequent naturally occurring trinucleotide bulges were optically melted, and their thermodynamic parameters—enthalpy, entropy, free energy, and melting temperature—were determined. The thermodynamic data were used to derive a new model to predict the free energy contribution of trinucleotide bulges to RNA duplex stability: $\Delta G^{\circ}_{37,\text{ trint bulge}} = \Delta G^{\circ}_{37,\text{ bulge}} + \Delta G^{\circ}_{37,\text{ AU}} + \Delta G^{\circ}_{37,\text{ GU}}$. The parameter $\Delta G^{\circ}_{37,\text{ bulge}}$ is variable depending upon the purine and pyrimidine composition of the bulge, $\Delta G^{\circ}_{37,\text{ AU}}$ is a 0.49 kcal/mol penalty for an A-U closing pair, and $\Delta G^{\circ}_{37,\text{ GU}}$ is a −0.56 kcal/mol bonus for a G-U closing pair. With both closing pair and bulge sequence taken into account, this new model predicts free energy values within 0.30 kcal/mol of the experimental value. The new model can be used by algorithms that predict RNA free energies as well as algorithms that use free energy minimization to predict RNA secondary structure from sequence.

■ INTRODUCTION

RNA has a more expansive functional role than merely serving as an intermediate between DNA and protein. For example, recent studies have investigated the catalytic role of ribozymes,[1] the role of snRNAs in mRNA splicing,[2] and the control of gene expression by riboswitches,[3] siRNAs,[4] and miRNAs.[5] As more functions of RNA are elucidated, greater interest develops in investigating RNA structure. Programs such as *RNAstructure*,[6] *mfold*,[7] and the Vienna package[8] predict RNA secondary structure from sequence. These predictive algorithms use a nearest neighbor model based on the thermodynamic parameters of various structural motifs. Current algorithms correctly predict ~73% of base pairs.[9] Any improvement to the nearest neighbor model could improve secondary structure prediction from sequence. Accurate RNA secondary structure prediction can provide clues about the tertiary structure of a given sequence.

One common naturally occurring secondary structure motif in RNA is a bulge. Bulges consist of one or more consecutive unpaired nucleotides on one side of an RNA duplex. Bulges greater than one nucleotide are assumed to prevent the base pairs adjacent to the bulge from stacking on one another.[9] Bulges serve in a variety of functions including viral replication,[10] ligand binding,[11] feedback regulation,[12] intron splicing,[13] gene expression,[14] as well as the formation of RNA tertiary structures.[15] Thermodynamic data have been collected on all combinations of single nucleotide bulges and adjacent base pairs, and algorithms predicting the free energy contribution of this motif have been developed from these data.[16−18] In contrast to the single nucleotide bulge motif, the thermodynamic parameters of the trinucleotide bulge motif have not been well characterized. This lack of data is unfortunate since trinucleotide bulges are prevalent in nature and have significant biological functions. Trinucleotide bulges have been discovered in all three domains of life. Most notably, two trinucleotide bulges form a bulge-helix-bulge motif that is conserved throughout all archaea and acts as the site of recognition and cleavage by splicing endonucleases.[19] Another notable trinucleotide bulge occurs in the HIV-1 TAR RNA hairpin fragment and is crucial for the interaction between the TAR RNA and Tat protein. Since this interaction is vital to mRNA synthesis, and ultimately, to the replication of the HIV-1 virus, this trinucleotide bulge has been widely studied as a potential drug target.[10]

Despite the natural prevalence of trinucleotide bulges, few studies have characterized the thermodynamic parameters of this motif.[20,21] In fact, the current predictive model is based on free energy values for only six trinucleotide bulges, and it

Received: February 17, 2014
Revised: May 5, 2014
Published: May 22, 2014

attributes a 3.2 kcal/mol free energy penalty to all trinucleotide bulges, regardless of sequence.[9] The aim of this study was to determine whether a sequence-independent model is the most accurate approach for predicting the free energy contribution of trinucleotide bulges. Reported here are the thermodynamic parameters for 19 trinucleotide bulges, most of which occur frequently in nature. These results can be incorporated into predictive models, allowing experimental rather than predicted values to be used for these naturally occurring bulges. For trinucleotide bulges whose thermodynamic parameters have not yet been characterized, an improved model for predicting their free energy contribution was derived.

■ MATERIALS AND METHODS

Compiling and Searching a Database for RNA Trinucleotide Bulges. A previously compiled database containing 1349 RNA secondary structures was used to identify common naturally occurring trinucleotide bulges and closing pair sequences. The database contained structures for 484 tRNAs, 309 5S rRNAs, 223 large subunit rRNAs, 123 small subunit rRNAs, 100 group I introns, 91 signal recognition particles, 16 large subunit rRNAs, and 3 group II introns.[22] This database was searched for trinucleotide bulges and their adjacent base pairs. In this work, G-U pairs were considered to be canonical base pairs.

Design of Sequences. RNA duplexes were composed of one strand that was 11 nucleotides long and a second strand that was 8 nucleotides long; the bulge and adjacent pairs were centered in the duplex with three additional Watson−Crick pairs on each side. In order to prevent end fraying during melting, the terminal pairs of each duplex were G-C pairs. The sequence of the trinucleotide bulge and nearest neighbor pairs was selected based on those that appeared most frequently in the secondary structure database described above. Prior to the experiments, all sequences were checked for possible competing unimolecular or alternative bimolecular folding. It was assumed that all of the duplexes studied here formed a trinucleotide bulge, as this was the fold with the lowest free energy fold. Any possible competing folds were at least ~2 kcal/mol less stable.

RNA Synthesis and Purification. The RNA oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA). The synthesis and purification procedures were standard and have been described previously.[23]

Optical Melting Experiments and Thermodynamics. Optical melting experiments were performed in 1 M NaCl, 20 mM sodium cacodylate, and 0.5 mM Na_2EDTA (pH 7.0). Optical melting curves were obtained, and the thermodynamic parameters for the entire duplex were determined as previously described.[24,25] Each duplex was melted at least nine times, using a different concentration each time, to ensure that the total oligonucleotide concentration range was at least 50-fold. The free energy contribution of the canonical pairs was determined using the nearest neighbor method (a 0.45 kcal/mol penalty was applied per each A-U pair adjacent to the bulge, as it was considered a terminal A-U pair).[26,27] This calculated free energy value was subtracted from the experimental free energy of the entire duplex, obtained from the $1/T_m$ vs ln C_T plots, in order to determine the free energy contribution of the trinucleotide bulge.[26] A reference duplex was not used because multiple different reference duplexes would have been needed. In addition, the nearest neighbor method is quite accurate for canonical pairs[26,27] and was used since the bulge parameters will ultimately be used in conjunction with nearest neighbor parameters as part of predictive models.

Linear Regression and Trinucleotide Bulge Thermodynamic Parameters. This study investigated the thermodynamic parameters of 16 trinucleotide bulge and closing pair combinations which occurred frequently in the secondary structure database. Data for 3 additional bulges[28] were added to these 16, and data for all 19 bulges were used during data analysis. These three additional bulges were studied previously by our laboratory,[28] but these particular bulge sequences were not found in the secondary structure database described above. Two previous studies did report thermodynamic parameters for some trinucleotide bulges;[20,21] however, due to non-two-state melts[20] and different buffer conditions,[21] these data were not included in this analysis. The experimental free energy contribution of the trinucleotide bulge was used as a constant when doing linear regression using the LINEST function in *Microsoft Excel*. Multiple parameters were tested, and the parameters that yielded the greatest predictive accuracy accounted for the closing base pair and the number of purine and pyrimidine nucleotides in the bulge.

■ RESULTS

Database Searching. In the database of 1349 RNA secondary structures, 410 trinucleotide bulges were found. The first data set in Table 1 presents the frequency and percent occurrence when bulge and closing pair sequence are specified. When the data is analyzed in this fashion, 122 unique bulge and closing pair combinations were identified. The top 20 most frequent bulge and corresponding closing pair sequences found in the database are shown in Table 1. These bulges and corresponding closing pair sequences account for ~65% of all trinucleotide bulges identified in the database search. Many of the bulges identified do not occur frequently, as 62% of the unique bulge and closing pair combinations found each account for <0.25% of the total number of trinucleotide bulges. The most frequent occurring bulge and corresponding closing pair sequence is $\begin{pmatrix} 5' & \mathrm{G} & \mathbf{GAC} & \mathrm{C} \\ 3' & \mathrm{U} & & \mathrm{G} \end{pmatrix}$, which accounts for ~12% of all of the bulges identified. All of the trinucleotide bulges with this bulge and closing pair sequence were discovered in 16S rRNA, suggesting that this sequence is conserved in this ribosomal subunit. However, the thermodynamic parameters of this bulge and closing pair combination were not characterized in this study due to possible competing conformations. Similarly, other bulge and closing pair combinations found frequently in the secondary structure database were also not studied here due to possible competing structures. For example, the most frequent bulge found in the database, 5′GAG3′, (Table 1) was not studied here. Based on the sequence of the bulge and nearest neighbors, it is possible that a different trinucleotide bulge could form, 5′GGA3′, resulting from the 3′ U in the bottom strand forming a G-U pair with the 3′-most G in the top strand. This ambiguity in the bulge sequence is akin to the ambiguity described for single nucleotide bulges.[16−18] Due to the ambiguity of the bulge sequence, this bulge was not studied; however, this study did investigate 7 of the top 10 most frequent bulge and corresponding closing pair sequences found in the database.

The second data set in Table 1 presents the frequency and percent occurrence when only the bulge sequence is specified. This analysis identified 49 unique trinucleotide bulge sequences in the database, out of a possible 64 bulge sequence

Table 1. Frequency of Occurrence of Trinucleotide Bulge Sequences in a Secondary Structure Database[a]

Dataset 1			Dataset 2			Dataset 3		
Bulge + NN[b]	Freq.[c]	%[d]	Bulge[e]	Freq.[c]	%[d]	NN[f]	Freq.[c]	%[d]
G GAG C U G	51	12.44	GAG	62	15.12	U C G G	102	24.88
A AAC G U U	27	6.59	GAU	51	12.44	G C U G	80	19.51
U CAU C G G	27	6.59	CAU	29	7.07	A G U U	35	8.54
U CUG U G G	23	5.61	AAC	27	6.59	U U G G	29	7.07
U GAU C G G	14	3.41	CAG	23	5.61	C C G G	26	6.34
U CAG C G G	13	3.17	CUG	23	5.61	U A G U	18	4.39
C GAU U G G	13	3.17	GUA	19	4.63	G G C C	17	4.15
C GUA C G G	13	3.17	GAC	16	3.90	C U G G	13	3.17
U CUA A G U	10	2.44	CUA	13	3.17	C U G A	11	2.68
U CCU C G G	9	2.20	CCU	11	2.68	U C A G	9	2.20
G GAU C U G	9	2.20	AAA	10	2.44	A C U G	8	1.95
G UAG C U G	9	2.20	GUU	10	2.44	G C C G	7	1.71
G AAA G C C	7	1.71	UAG	10	2.44	C G G C	7	1.71
U CAG U G G	7	1.71	CCG	9	2.20	U U G G	6	1.46
G GUU C U G	7	1.71	GAA	9	2.20	U U G A	4	0.98
U CCG U G G	6	1.46	GCU	8	1.95	C A G U	4	0.98
C GAC C G G	6	1.46	AAG	6	1.46	G A U U	4	0.98
U GAU C A G	6	1.46	UUA	6	1.46	U U A A	4	0.98
G GAG C C G	5	1.22	GGU	5	1.22	U G A C	4	0.98
C GAU U G A	5	1.22	AUA	4	0.98	A G U C	3	0.73

[a] Not all bulges found in the database are shown due to space limitations. [b] Trinucleotide bulge when bulge and nearest neighbor sequence is specified. [c] Frequency of occurrence in the database search. [d] Percent out of 410 trinucleotide bulges, the total number found in the database search. [e] Trinucleotide bulge when only the bulge sequence is specified. [f] Closing pairs of trinucleotide bulges are specified.

combinations. Table 1 shows the top 20 most frequent trinucleotide bulge sequences found, which account for ~86% of all of the trinucleotide bulges found in the database. The trinucleotide bulge sequences not included in Table 1 individually account for ≤4% of the total trinucleotide bulges found in the database search. The most frequent bulge sequence identified was 5′GAG3′ which accounts for 15% of all the trinucleotide bulges found in the database search. As mentioned earlier, the prevalence of the 5′GAG3′ bulge sequence in this database can be attributable to the fact that it appears to be conserved in 16S rRNA. This thermodynamic study investigates a total of 12 different bulge sequences (not considering the nearest neighbors), which represents 48% of the different types of trinucleotide bulges identified in the database search.

The third data set in Table 1 lists the frequency and percent occurrence when only the closing pair combination is specified. The database search yielded 31 of the 36 possible types of closing pair combinations, and 95% of all of the trinucleotide bulges are represented by the closing pair combinations in Table 1. The most frequent closing pair found in the database was $\left(\begin{smallmatrix}5' & \text{U} & \textbf{XXX} & \text{C}\\ 3' & \text{G} & & \text{G}\end{smallmatrix}\right)$, accounting for 25% of all closing pairs found. As was true with the predominant bulge sequences, this closing pair pattern appears to be conserved in 16S rRNA, but it also appears in other rRNA subunits as well. This thermodynamic study investigates a total of nine unique closing pair sequences, which account for 74% of closing pair sequences found in the database.

Thermodynamic Parameters. Table 2 shows the thermodynamic parameters for the formation of duplexes containing trinucleotide bulges. These thermodynamic parameters are derived from an analysis of individual melt curves and an analysis of the $1/T_M$ versus $\log(C_T)$ plots. The duplexes in the table are listed in order of decreasing frequency in the secondary structure database.

Contribution of the Trinucleotide Bulge to Duplex Thermodynamics. The contribution of the trinucleotide bulge to duplex thermodynamics (Table 3) was calculated as described in Materials and Methods. The free energy contribution of the measured trinucleotide bulges has a large variance, ranging from 2.3 to 5.6 kcal/mol. The most destabilizing trinucleotide bulge reported here is $\left(\begin{smallmatrix}5' & \text{A} & \textbf{UUU} & \text{G}\\ 3' & \text{U} & & \text{C}\end{smallmatrix}\right)$, while the least destabilizing trinucleotide bulge is $\left(\begin{smallmatrix}5' & \text{U} & \textbf{CAG} & \text{C}\\ 3' & \text{G} & & \text{G}\end{smallmatrix}\right)$.

Free Energy Parameters for Trinucleotide Bulges. Currently, a free energy penalty of 3.2 kcal/mol is attributed to all trinucleotide bulges regardless of bulge and closing pair sequence.[9] In order to improve prediction, multiple other models were tested, and a model that resulted in a low average deviation from the experimental bulge values with minimal parameters was derived. This model is

$$\Delta G°_{37,\text{trint bulge}} = \Delta G°_{37,\text{bulge}} + \Delta G°_{37,\text{AU}} + \Delta G°_{37,\text{GU}} \tag{1}$$

As shown in Table 4, $\Delta G°_{37,\text{ bulge}}$ is dependent upon the number of purines and pyrimidines in the bulge sequence. Thus, $\Delta G°_{37,\text{ bulge}}$ is a 4.19 kcal/mol penalty for three bulged purines, a 5.06 kcal/mol penalty for three bulged pyrimidines, a 3.36 kcal/mol penalty for two bulged purines and one bulged pyrimidine, and a 3.56 kcal/mol penalty for two bulged pyrimidines and one bulged purine. $\Delta G°_{37,\text{ AU}}$ is a 0.49 kcal/mol penalty for each A-U closing pair, and $\Delta G°_{37,\text{ GU}}$ is a −0.56 kcal/mol bonus for each G-U closing pair. When using this model, the average magnitude of difference between the predicted values and the experimental values was only 0.30 kcal/mol (Table 3).

■ DISCUSSION

Database Searching. The database search yielded 124 out of 2304 possible combinations of bulge and closing pair sequence combinations, 49 out of 64 possible bulge sequence

Table 2. Thermodynamic Parameters for the Formation of Duplexes Containing Trinucleotide Bulges[a]

		Analysis of T_M Dependence/ Errors (ln Plot)				Analysis of Melt Curve Fit/Errors			
Freq.[b]	Sequence[c]	ΔH° (kcal/mol)	ΔS° (cal/Kmol)	ΔG°$_{37}$ (kcal/mol)	T_m[d] (°C)	ΔH° (kcal/mol)	ΔS° (cal/Kmol)	ΔG°$_{37}$ (kcal/mol)	T_m[d] (°C)
27	GAC A**AAC**G CUG CUG U U GAC	-70.3 ± 8.7	-210.3 ± 28.5	-5.09 ± 0.31	30.8	-65.3 ± 9.8	-193.9 ± 32.5	-5.20 ± 0.38	30.8
27	GAC U**CAU**C CUG CUG G G GAC	-59.9 ± 3.8	-169.9 ± 12.2	-7.24 ± 0.08	40.7	-71.2 ± 8.2	-205.9 ± 26.3	-7.30 ± 0.18	40.4
14	GUC U**GAU**C CUG CAG G G GAC	-68.1 ± 8.6	-196.7 ± 27.9	-7.08 ± 0.34	39.5	-61.5 ± 9.3	-175.2 ± 30.0	-7.13 ± 0.19	40.0
13	GAC U**CAG**C CUG CUG G G GAC	-60.0 ± 10.6	-169.5 ± 33.4	-7.42 ± 0.76	41.6	-56.3 ± 13.2	-157.5 ± 42.2	-7.45 ± 0.52	42.2
13	GAC C**GAU**U CUG CUG G G GAC	-45.8 ± 3.9	-126.8 ± 12.7	-6.52 ± 0.14	36.9	-43.9 ± 16.0	-120.6 ± 51.4	-6.50 ± 0.26	36.8
13	GAC C**GUA**C CUG CUG G G GAC	-75.6 ± 11.1	-217.3 ± 35.2	-8.16 ± 0.41	43.8	-75.7 ± 15.8	-217.9 ± 50.6	-8.13 ± 0.30	43.7
10	GAC U**CUA**A CUG CUG G U GAC	-55.8 ± 3.7	-164.4 ± 12.3	-4.79 ± 0.17	27.6	-58.2 ± 7.4	-172.5 ± 24.6	-4.70 ± 0.36	27.6
7	GAC G**AAA**G CUG CUG C C GAC	-59.2 ± 7.1	-171.4 ± 23.4	-6.09 ± 0.28	34.7	-50.5 ± 4.7	-143.4 ± 16.0	-6.06 ± 0.73	34.2
5	GAC C**GAU**U CUG CUG G A GAC	-50.8 ± 6.0	-142.5 ± 19.6	-6.59 ± 0.26	37.3	-48.6 ± 6.0	-135.1 ± 19.2	-6.66 ± 0.16	37.9
4	GAC U**GAA**C CUG CUG G G GAC	-52.7 ± 5.6	-148.6 ± 18.1	-6.65 ± 0.19	37.7	-52.8 ± 8.8	-148.8 ± 27.9	-6.67 ± 0.24	37.8
3	GAC C**GAU**C CUG CUG G G GAC	-57.7 ±11.7	-161.0 ± 37.0	-7.72 ± 0.75	43.6	-60.4 ± 12.2	-169.5 ± 38.9	-7.83 ± 0.36	43.8
3	GAC U**GUA**C CUG CUG G G GAC	-50.9 ± 4.2	-141.7 ± 13.5	-6.99 ± 0.12	39.8	-52.8 ± 17.1	-147.5 ± 54.7	-6.99 ± 0.24	39.7
3	GAC A**UAC**G CUG CUG U U GAC	-59.2 ± 6.0	-174.5 ± 20.0	-5.08 ± 0.24	29.6	-60.0 ± 4.5	-177.0 ± 14.6	-5.05 ± 0.10	29.5
3	GAC G**UAU**C CUG CUG U G GAC	-50.4 ± 1.5	-142.7 ± 4.9	-6.14 ± 0.03	34.6	-57.0 ± 5.1	-164.3 ± 15.8	-6.07 ± 0.12	34.5
2	GAC U**AAG**C CUG CUG G G GAC	-38.1 ± 9.7	-104.6 ± 31.4	-5.70 ± 0.97	30.4	-42.8 ± 14.8	-119.5 ± 47.2	-5.76 ± 0.41	31.5
2	GAC U**CAA**G CUG CUG G C GAC	-56.4 ± 2.8	-162.0 ± 9.3	-6.12 ± 0.08	34.7	-56.1 ± 10.6	-161.1 ± 34.4	-6.16 ± 0.16	35.0
0	GCC A**UGU**G AGC[e] CGG U C UCG	-51.9 ± 5.2	-143.4 ± 16.7	-7.47 ± 0.18	42.7	-50.4 ± 4.7	-138.6 ± 14.5	-7.47 ± 0.22	42.9
0	GCC A**CUU**G AGC[e] CGG U C UCG	-61.3 ± 2.2	-176.2 ± 7.0	-6.61 ± 0.02	37.4	-59.8± 3.7	-171.4 ± 11.8	-6.62 ± 0.08	37.5
0	GCC A**UUU**G AGC[e] CGG U C UCG	-66.2 ± 2.1	-192.5 ± 6.8	-6.50 ± 0.02	36.8	-62.4 ± 3.1	-180.2 ± 10.1	-6.53 ± 0.09	37.0

[a] Measurements were made in 1.0 M NaCl, 10 mM sodium cacodylate, and 0.5 mM Na_2EDTA pH 7.0. [b] Frequency of occurrence obtained from database described in Materials and Methods. [c] The trinucleotide bulge is identified by bold letters. The nearest neighbors and bulge are set apart for easy identification. The top strand of each duplex is written 5′ to 3′, and each bottom strand is written 3′ to 5′. [d] All values are calculated at 10^{-4} M oligomer concentration. [e] Melt data from ref 28.

possibilities, and 33 out of a possible 36 closing pair combinations. It is likely that many of the possible sequence combinations that were not found in our database do exist in nature and would have been found in a larger secondary structure database. There may be a structural explanation for why certain sequence combinations occur more frequently than others, but such an explanation would require an extensive structural study.

The two most frequent bulge sequences in the database, 5′GAG3′ and 5′GAU3′, are similar and account for 15% and 12%, respectively, of the bulge sequences found in the database. While 5′GAA3′ and 5′GAC3′ were also found in the database search, it is interesting to note that these sequences, with the same first two nucleotides, only accounted for 4% and 2%, respectively, of bulges in the database. Similarly, the two most frequently occurring closing pair sequences for trinucleotide bulges in this database $\begin{pmatrix} 5' & \mathrm{U} & \mathbf{XXX} & \mathrm{C} \\ 3' & \mathrm{G} & & \mathrm{G} \end{pmatrix}$ and $\begin{pmatrix} 5' & \mathrm{G} & \mathbf{XXX} & \mathrm{C} \\ 3' & \mathrm{U} & & \mathrm{G} \end{pmatrix}$, are similar and account for 25% and 20%, respectively, of the trinucleotide bulges found in the database. However, the closing pair sequences that result from flipping the C-G pair, $\begin{pmatrix} 5' & \mathrm{U} & \mathbf{XXX} & \mathrm{G} \\ 3' & \mathrm{G} & & \mathrm{C} \end{pmatrix}$ and $\begin{pmatrix} 5' & \mathrm{G} & \mathbf{XXX} & \mathrm{G} \\ 3' & \mathrm{U} & & \mathrm{C} \end{pmatrix}$, occur far less frequently. The former sequence accounts for ~0.5% of all trinucleotide bulges found in the database, and the latter was not found in the database. There are no obvious explanations for these observed idiosyncracies.

Thermodynamic Contributions of a Trinucleotide Bulge to Duplex Thermodynamics. All of the trinucleotide

Table 3. Contribution of the Trinucleotide Bulge to Duplex Thermodynamics

		$\Delta G°_{37}$ (kcal/mol)				
			Sequence-Independent Model		Sequence-Dependent Model	
Freq.[a]	Sequence[b]	Measured[c]	Prediction[d]	Difference[e]	Prediction[f]	Difference[g]
27	A**AAC**G U U	3.41	3.2	-0.21	3.29	-0.12
27	U**CAU**C G G	2.48	3.2	0.72	3.00	0.52
14	U**GAU**C G G	2.64	3.2	0.56	2.80	0.16
13	U**CAG**C G G	2.30	3.2	0.90	2.80	0.50
13	C**GAU**U G G	3.23	3.2	-0.03	2.80	-0.43
13	C**GUA**C G G	3.05	3.2	0.15	3.36	0.31
10	U**CUA**A G U	3.46	3.2	-0.26	3.49	0.03
7	G**AAA**G C C	4.38	3.2	-1.18	4.19	-0.19
5	C**GAU**U G A	3.26	3.2	-0.06	3.85	0.59
4	U**GAA**C G G	3.07	3.2	0.13	3.63	0.56
3	C**GAU**C G G	3.49	3.2	-0.29	3.36	-0.13
3	U**GUA**C G G	2.73	3.2	0.47	2.80	0.07
3	A**UAC**G U U	3.42	3.2	-0.22	3.49	0.07
3	G**UAU**C U G	3.06	3.2	0.14	3.00	-0.06
2	U**AAG**C G G	4.02	3.2	-0.82	3.63	-0.39
2	U**CAA**G G C	3.76	3.2	-0.56	2.80	-0.96
0	A**UGU**G[h] U C	4.63	3.2	-1.43	4.05	-0.58
0	A**CUU**G[h] U C	5.49	3.2	-2.29	5.55	0.06
0	A**UUU**G[h] U C	5.60	3.2	-2.40	5.55	-0.05
Average[i]				0.67 ± 0.92		0.30 ± 0.41

[a] Frequency of occurrence obtained from database described in Materials and Methods. [b] The trinucleotide bulge is identified by bold letters. The top strand of each duplex is written 5′ to 3′, and each bottom strand is written 3′ to 5′. [c] The experimental free energy contribution of the bulge calculated as described in the text. [d] The free energy prediction made by the sequence-independent model (ref 9). [e] The difference between the free energy predicted by the sequence-independent model (ref 9) and the experimental free energy. [f] The free energy prediction made by the sequence dependent model proposed here. [g] The difference between the free energy predicted by the sequence-dependent model and the experimental free energy. [h] Data from ref 28. [i] The average (absolute value) deviation.

Table 4. Sequence-Dependent Model for Predicting the Free Energy Contribution of Trinucleotide Bulges

$\Delta G°_{37,\text{ trint bulge}}$ parameter	free energy contribution (kcal/mol)
$\Delta G°_{37,\text{ bulge}}$[a]	
3 purines	4.19 ± 0.30
3 pyrimidines	5.06 ± 0.45
2 purine, 1 pyrimidine	3.36 ± 0.26
1 purine, 2 pyrimidines	3.56 ± 0.37
$\Delta G°_{37,\text{ AU}}$[b]	0.49 ± 0.30
$\Delta G°_{37,\text{ GU}}$[c]	−0.56 ± 0.26

[a] Free energy contribution attributed to the three bulged nucleotides. One of the four values will be applied depending on the purine/pyrimidine composition of the bulge. [b] Free energy penalty penalty applied for each A-U closing pair of a trinucleotide bulge. [c] Free energy bonus applied for each G-U closing pair of a trinucleotide bulge.

bulges included in this study destabilize the duplex. This destabilization is expected, as it is presumed that every trinucleotide bulge disrupts what would otherwise be stabilizing stacking interactions between neighboring base pairs.[9] Also, the presence of the bulge may put strain on the adjacent base pairs, and the bend or kink at the site of the bulge may also destabilize the helix. Recall that the free energy contribution of the bulge ranged from 2.3 to 5.6 kcal/mol. This large variance suggests that the thermodynamic contribution attributed to the bulge may be dependent upon bulge and closing pair sequence, and thus, a sequence-dependent model would be preferred over a model that attributes a constant value to all trinucleotide bulges regardless of sequence.

At first glance, the thermodynamic data appear to suggest a relationship between frequency of occurrence in the database and stability of the bulge. Considering the 16 duplexes in this study that were found in the database, the average free energy contribution of the bulge for the three most frequent was 2.8 ± 0.5 kcal/mol, while the average free energy contribution of the bulge for the three least frequent was 3.5 ± 0.5 kcal/mol. The three additional sequences included in this study did not occur in the database search, and the average free energy of the bulge for these sequences was 5.2 ± 0.5 kcal/mol. From this analysis, less stable bulges appear to be occurring less frequently in the database.

However, a more detailed analysis suggests that no relationship between frequency and stability exists. Recall that the new model derived for predicting the free energy of formation of trinucleotide bulges has a parameter that was dependent upon the sequence of the bulge ($\Delta G°_{37,\text{ bulge}}$). The corresponding values for this parameter suggest that triple pyrimidine bulges and triple purine bulges are more destabilizing than bulges with a mix of purines and pyrimidines. If there is a relationship between bulge frequency and stability, then it would be expected that the most frequently occurring bulges would favor a mix of purines and pyrimidines. From analyzing the top 10 most frequent occurring trinucleotide bulge sequences from the database search (Table 1), bulges with a mix of purines and pyrimidines do not appear to be significantly favored over what would be statistically expected. This supports the conclusion that there is no relationship between frequency and stability. Conversely, there are 15 bulge sequences which did not appear in the database at all, making these sequences the least frequently occurring. If there was a relationship between bulge frequency and stability, then it would be expected that the least frequently occurring bulge sequences would favor all purine and all pyrimidine sequences. The 15 bulge sequences which did not appear in the database do not significantly favor all purine and all pyrimidine sequences over what would be statistically expected, giving further support to the conclusion that there is no relationship between frequency and stability.

Improving the Model Used to Predict Trinucleotide Bulge Thermodynamics. As shown in Table 4, every trinucleotide bulge receives a $\Delta G°_{37,\text{ bulge}}$ penalty depending upon the purine and pyrimidine composition of the bulge. All four possible $\Delta G°_{37,\text{ bulge}}$ penalties have a far greater magnitude than the G-U closing pair bonus. Thus, all trinucleotide bulges will be predicted to be destabilizing, regardless of bulge and closing pair sequence.

Several of the patterns observed here for trinucleotide bulges are consistent with what was reported previously for single nucleotide bulges. A study by Serra et al. (2007) on single nucleotide bulges concluded that the more stable a duplex is without the bulge, the greater the potential for a single nucleotide bulge to destabilize the duplex.[17] This trend was also evident in this study. The average free energy contribution for trinucleotide bulges located within the five most stable duplexes is 4.45 kcal/mol, while the average free energy contribution for trinucleotide bulges located within the five least stable duplexes is 3.32 kcal/mol. This suggests that the free energy contribution of bulges is somewhat dependent upon non-nearest neighbors.

In eq 1, all of the free energy parameters are penalties, except for the free energy contribution of a G-U closing pair, −0.56 kcal/mol. This bonus is very near −0.6 kcal/mol, the bonus attributed by Serra et al. (2007) to single nucleotide bulges with the closing pair sequence $\begin{pmatrix} 5' & \mathrm{U} & \mathbf{XXX} & \mathrm{X} \\ 3' & \mathrm{G} & & \mathrm{Y} \end{pmatrix}$.[17] Our study included a total of 10 trinucleotide bulges with one or more G-U closing pairs. Eight of these 10 bulges matched the closing pair orientation that was attributed the −0.6 kcal/mol bonus in the single nucleotide bulge study.[17] Thus, it is not surprising that the bonus discovered in this study is very near the previously discovered bonus for single nucleotide bulges.

The current model used by *RNAstructure* to predict the free energy contribution of trinucleotide bulges was derived by averaging the measured free energy contributions of six trinucleotide bulges.[6] Unlike that model, the model derived in this study is sequence-dependent and is based upon the measured free energy contribution of 19 trinucleotide bulges, a 3-fold increase in the sample size. On average, the sequence-independent model[9] predicts a free energy that is 0.67 kcal/mol different than the experimental value. On average, the sequence-dependent model proposed here predicts a free energy that is 0.30 kcal/mol different than the experimental value, a more than a 2-fold improvement over the sequence-independent model. Additionally, the standard deviation of the sequence-independent model is 0.92 kcal/mol, which is more than double the standard deviation of the sequence-dependent model, 0.41 kcal/mol (Table 3). Also, the sequence-dependent model works well on bulges that are predicted rather poorly by the sequence-independent model. For example, the prediction made by the sequence-independent model deviates from the measured value by more than 1 kcal/mol for four bulges, and three of these four bulges have measured free energy values which deviate by more than 2 kcal/mol from the predicted value. In contrast, the sequence-dependent model predicts the free energy of all the bulges studied within 1 kcal/mol of the measured free energy.

With the collection of thermodynamic data for 19 trinucleotide bulges, the model for predicting the free energy of formation of trinucleotide bulges was updated. While the current model used by *RNAstructure* attributes a constant free energy penalty to all trinucleotide bulges, the new model takes both bulge and closing pair sequence into consideration. This adjustment resulted in a sequence-dependent model with half the average error of the sequence-independent model.

AUTHOR INFORMATION

Corresponding Author

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

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number [R15GM085699]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Reymond, C., Beaudoin, J. D., and Perreault, J. P. (2009) Modulating RNA structure and catalysis: lessons from small cleaving ribozymes. *Cell. Mol. Life Sci. 66*, 3937−3950.

(2) Valadkhan, S. (2005) snRNAs as the catalysts of pre-mRNA splicing. *Curr. Opin. Chem. Biol. 9*, 603−608.

(3) Hollands, K., Proshkin, S., Sklyarova, S., Epshtein, V., Mironov, A., Nudler, E., and Groisman, E. A. (2012) Riboswitch control of Rho-dependent transcription termination. *Proc. Natl. Acad. Sci. U.S.A. 109*, 5376−5381.

(4) Gavrilov, K., and Saltzman, W. M. (2012) Therapeutic siRNA: Principles, challenges, and strategies. *Yale J. Biol. Med. 85*, 187−200.

(5) Lehmann, S. M., Kruger, C., Park, B., Derkow, K., Rosenberger, K., Baumgart, J., Trimbuch, T., Eom, G., Hinz, M., Kaul, D., Habbel, P., Kalin, R., Franzoni, E., Rybak, A., Nguyen, D., Veh, R., Ninnemann, O., Peters, O., Nitsch, R., Heppner, F. L., Golenbock, D., Schott, E., Ploegh, H. L., Wulczyn, F. G., and Lehnardt, S. (2012) An unconventional role for miRNA: Let-7 activates Toll-like receptor 7 and causes neurodegeneration. *Nat. Neurosci. 15*, 827−835.

(6) 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.

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

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

(9) 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.

(10) Davidson, A., Leeper, T. C., Athanassiou, Z., Patora-Komisarska, K., Karn, J., Robinson, J. A., and Varani, G. (2009) Simultaneous recognition of HIV-1 TAR RNA bulge and loop sequences by cyclic peptide mimics of Tat protein. *Proc. Natl. Acad. Sci. U.S.A. 106*, 11931−11936.

(11) Peattie, D. A., Douthwaite, S., Garrett, R. A., and Noller, H. F. (1981) A bulged double helix in a RNA-protein contact site. *Proc. Natl. Acad. Sci. U.S.A. 78*, 7331−7335.

(12) Climie, S. C., and Friesen, J. D. (1987) Feedback-regulation of the Rpljl-Rpobc ribosomal-protein operon of *Escherichia-coli* requires a region of messenger-RNA secondary structure. *J. Mol. Biol. 198*, 371−381.

(13) McManus, C. J., Schwartz, M. L., Butcher, S. E., and Brow, D. A. (2007) A dynamic bulge in the U6 RNA internal stem-loop functions in spliceosome assembly and activation. *RNA 13*, 2252−2265.

(14) Gerdeman, M. S., Henkin, T. M., and Hines, J. V. (2003) Solution structure of the Bacillus subtilis T-box antiterminator RNA: Seven nucleotide bulge characterized by stacking and flexibility. *J. Mol. Biol. 326*, 189−201.

(15) Woese, C. R., and Gutell, R. R. (1989) Evidence for several higher-order structural elements in ribosomal-RNA. *Proc. Natl. Acad. Sci. U.S.A. 86*, 3119−3122.

(16) Znosko, B. M., Silvestri, S. B., Volkman, H., Boswell, B., and Serra, M. J. (2002) Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges. *Biochemistry 41*, 10406−10417.

(17) Blose, J. M., Manni, M. L., Klapec, K. A., Stranger-Jones, Y., Zyra, A. C., Sim, V., Griffith, C. A., Long, J. D., and Serra, M. J. (2007) Non-nearest-neighbor dependence of the stability for RNA bulge loops based on the complete set of group I single-nucleotide bulge loops. *Biochemistry 46*, 15123−15135.

(18) McCann, M. D., Lim, G. F., Manni, M. L., Estes, J., Klapec, K. A., Frattini, G. D., Knarr, R. J., Gratton, J. L., and Serra, M. J. (2011) Non-nearest-neighbor dependence of the stability for RNA group II single-nucleotide bulge loops. *RNA 17*, 108−119.

(19) Hermann T, P. D. (2000) RNA bulges as architectural and recognition motifs. *Structure 8*, R47−54.

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

(21) Carter-O'Connell, I., Booth, D., Eason, B., and Grover, N. (2008) Thermodynamic examination of trinucleotide bulged RNA in the context of HIV-1 TAR RNA. *RNA 14*, 2550−2556.

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

(23) 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.

(24) Christiansen, M. E., and Znosko, B. M. (2008) Thermodynamic characterization of the complete set of sequence symmetric tandem mismatches in RNA and an improved model to predict the free energy contribution of sequence asymmetric tandem mismatches. *Biochemistry 47*, 4329−4336.

(25) Schroeder, S. J., and Turner, D. H. (2009) Optical melting measurements of nucleic acid thermodynamics. *Methods Enzymol. 468*, 371−387.

(26) 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.

(27) Chen, J. L., Dishler, A. L., Kennedy, S. D., Yildirim, I., Liu, B., Turner, D. H., and Serra, M. J. (2012) Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. *Biochemistry 51*, 3508−3522.

(28) Davis, A. R. Thermodynamic and structural properties of small RNA secondary structural motifs and their role in RNA-protein interactions, PhD Dissertation, Saint Louis University, St. Louis, MO, May 2010.