BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

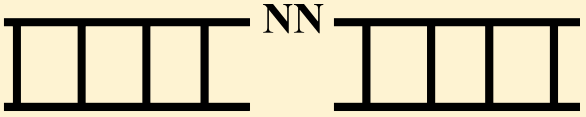
pubs.acs.org/biochemistry

Improved Model for Predicting the Free Energy Contribution of Dinucleotide Bulges to RNA Duplex Stability

Jeremy C. Tomcho, Magdalena R. Tillman, and Brent M. Znosko*

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

ABSTRACT: Predicting the secondary structure of RNA is an intermediate in predicting RNA three-dimensional structure. Commonly, determining RNA secondary structure from sequence uses free energy minimization and nearest neighbor parameters. Current algorithms utilize a sequence-independent model to predict free energy contributions of dinucleotide bulges. To determine if a sequence-dependent model would be more accurate, short RNA duplexes containing dinucleotide bulges with different sequences and nearest neighbor combinations were optically melted to derive thermodynamic parameters. These data suggested energy contributions of dinucleotide bulges were sequence-dependent, and a sequence-dependent model was derived. This model assigns free energy penalties based on the identity of nucleotides in the bulge (3.06 kcal/mol for two purines, 2.93 kcal/mol for two pyrimidines, 2.71 kcal/mol for 5′-purine-pyrimidine-3′, and 2.41 kcal/mol for 5′-pyrimidine-purine-3′). The predictive model also includes a 0.45 kcal/mol penalty for an A-U pair adjacent to the bulge and a −0.28 kcal/mol bonus for a G-U pair adjacent to the bulge. The new sequence-dependent model results in predicted values within, on average, 0.17 kcal/mol of experimental values, a significant improvement over the sequence-independent model. This model and new experimental values can be incorporated into algorithms that predict RNA stability and secondary structure from sequence.



$\Delta G^\circ_{37,\text{dint bulge}} = (\Delta G^\circ_{\text{bulge}}) + (\Delta G^\circ_{\text{GU or AU closing pair}})$

$\Delta G^\circ_{37,\text{dint bulge}} = (\Delta G^\circ_{RR} \text{ or } \Delta G^\circ_{YY} \text{ or } \Delta G^\circ_{RY} \text{ or } \Delta G^\circ_{YR}) + (\Delta G^\circ_{GU} \text{ and/or } \Delta G^\circ_{AU})$

RNA has more biological functions in nature than serving as an intermediate in protein synthesis. A few of the numerous additional roles are to catalyze reactions,[1] regulate function,[2] control gene expression through riboswitches,[3] and use snRNAs in mRNA splicing.[4] Many scientists are interested in predicting the free energy or secondary structure of a particular RNA sequence. *RNAstructure*,[5] *mfold*,[6] and the Vienna package[7] make these predictions using a nearest neighbor model based on thermodynamic parameters for all secondary structure motifs. The improvement in the free energy parameters for any particular motif could improve the secondary structure and free energy predictions made by these programs. Subsequently, the secondary structure can be utilized as an intermediary in tertiary structure prediction based on the sequence.[8]

Bulges are a common RNA secondary structure motif in nature. A bulge consists of one or more adjacent unpaired nucleotides in one strand of an RNA duplex and is assumed to prevent the adjacent base pairs from stacking with one another when the bulge consists of two or more nucleotides.[9] Bulges can perform a variety of functions such as serving roles in gene expression,[10] intron splicing,[11] ligand binding,[12] and the formation of tertiary structures.[13] Dinucleotide bulges occur naturally in several organisms, such as the telomerase holoenzyme of *Tetrahymena thermophila*,[14] the 5′-UTR of multiple enteroviruses such as *poliovirus* type 1, and rhinoviruses such as human rhinovirus-14. The dinucleotide bulge in these enteroviruses and rhinoviruses is part of a consensus sequence that has been shown to regulate translation as well as replication of these viruses.[15] One virus of specific interest is HIV-2 where dinucleotide bulges have previously been identified and determined to play a significant role in viral replication.[16]

Thermodynamic data have previously been collected for only six dinucleotide bulges. On the basis of this limited data set, a model that attributed a 2.8 kcal/mol free energy penalty to all dinucleotide bulges independent of the identity of the nucleotides in the bulge and the nearest neighbors adjacent to the bulge was derived.[17] However, the identities of these nucleotides have proven to be important in thermodynamic model development, as seen for single-nucleotide bulges[18] and trinucleotide bulges.[19] This study aims to determine whether the current sequence-independent model is the most accurate predictor of the free energy contribution of dinucleotide bulges. This study reports the thermodynamic parameters for 18 dinucleotide bulges (and includes an additional bulge from the literature), most of which frequently occur in nature. These experimental results can be incorporated into secondary structure prediction software. Also, a sequence-dependent model utilizing the identity of the nucleotides in the bulge and the nearest neighbors was derived. This improved model can also be incorporated into secondary structure prediction software to be used for predicting the free energy contributions of dinucleotide bulges for which we do not have experimental thermodynamic data.

MATERIALS AND METHODS

Compiling and Searching a Database for RNA Dinucleotide Bulges. A previously compiled database of

Received: April 29, 2015
Revised: July 10, 2015
Published: August 19, 2015

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

various RNA secondary structures[20] was searched for dinucleotide bulges. G-U pairs were considered to be canonical base pairs for this study. The resulting list of naturally occurring dinucleotide bulges was used to identify the dinucleotide bulges and the nearest neighbor combinations studied here.

Design of Sequences. On the basis of the list of dinucleotide bulges described above, short, synthetic RNA duplexes were designed. The RNA duplexes were composed of one strand with 10 nucleotides and another with 8 nucleotides. The dinucleotide bulge with its nearest neighbors was centered in the middle of the duplex with three additional Watson−Crick base pairs on either side. To prevent end fraying during the melting experiment, the terminal base pairs of the duplexes were G-C pairs. The designed sequences were then checked (by free energy calculations) for possible competing unimolecular or alternative bimolecular folding. A total of 18 duplexes with dinucleotide bulges were selected for this study.

RNA Synthesis and Purification. The RNA oligonucleotides were ordered from Integrated DNA Technologies (Coralville, IA). The synthesis and purification were standard, and the procedures have been described previously.[21]

Optical Melting Experiments and Thermodynamics. Optical melting experiments were performed in a buffer solution containing 1 M NaCl, 20 mM sodium cacodylate, and 0.5 mM EDTA (pH 7). The optical melting studies were performed utilizing standard procedures as described previously.[20,21] The resulting melt curves were analyzed with *Meltwin*.[22] Thermodynamic parameters were then obtained for each duplex as described previously.[23] Because multiple reference duplexes would have been necessary, the nearest neighbor model was utilized to account for the thermodynamic parameters of the stem.[24] In addition, the nearest neighbor model was utilized because nearest neighbor parameters will be used along with the newly derived bulge parameters to predict the free energy and secondary structure.

Linear Regression and Dinucleotide Bulge Thermodynamic Parameters. The data from the 18 dinucleotide bulge and closing pair combinations determined experimentally in this study were combined with data from one additional bulge previously studied.[17] Four other bulges were previously studied[17] but were not utilized during model development because of the possible competition from unwanted bimolecular associations of the two strands and non-two-state melting. For example, the authors of the previous study recognized these possible issues by stating that one of the sequences, $\binom{5'GCG\ AA\ GCG}{3'CGC\ \ \ CGC}$, had significant concentrations of homoduplex formation and that most strands did not melt in a two-state manner.[17] The experimental free energy contribution of the dinucleotide bulge was utilized as a constant when performing linear regression using the LINEST function in *Microsoft Excel*. Multiple combinations of parameters were tested, and those that yielded the greatest predictive accuracy accounted for the closing base pair as well as the identity (purine or pyrimidine) of the nucleotides that made up the dinucleotide bulge.

RESULTS

Database Searching. In the compiled database of RNA secondary structures, 1839 dinucleotide bulges were found. The first data set in Table 1 presents the frequency and percent occurrence when both the bulge and closing base pairs are specified. When the data were categorized in this manner, 220 unique bulge and closing pair combinations were identified. The top 20 most frequent bulge and closing pair sequences are listed in Table 1. These bulges and corresponding closing pair sequences account for 63% of all dinucleotide bulge and closing pair combinations found in the database. Many of the combinations found do not occur frequently. The most frequently occurring bulge and corresponding closing pair sequence is $\binom{5'G\ AA\ G}{3'C\ \ \ C}$, which accounts for 13% of all combinations found. The thermodynamic parameters for this bulge and closing pair sequence were previously studied, and those results are included in this study. Some combinations of bulge and corresponding closing pair sequences frequently found in the secondary structure database were not studied here because of possible competing structures; however, this study investigated 5 of the top 10 most frequent bulges and closing pair sequences in the database and incorporated the most frequent bulge from the literature.[17] These six bulges account for 27% of the combinations found in the database. Overall, 41% of the combinations found in the database were accounted for in this study.

The second data set in Table 1 displays the frequency and percent occurrence when only the bulge sequence is specified. There were 16 unique dinucleotide bulge sequences identified in the database, which includes all possible dinucleotide bulges. Table 1 includes all 16 of these unique sequences. The most frequent dinucleotide bulge is (5′ AA 3′), which accounts for 28% of all dinucleotide bulges in the database. This is the same dinucleotide bulge found within the most prevalent bulge and closing base pair combination discussed earlier. This study investigates the thermodynamics of seven different bulge sequences (not considering nearest neighbors), which represent 76% of the bulges present in the secondary structure database.

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

Contribution of the Dinucleotide Bulge to Duplex Thermodynamics. The dinucleotide bulge contribution to duplex thermodynamics is listed in Table 3. These values were calculated as described in Materials and Methods. The measured free energy contribution of the dinucleotide bulges ranged from 2.33 to 3.76 kcal/mol. The most destabilizing dinucleotide bulge is $\binom{5'U\ AC\ U}{3'A\ \ \ A}$, while the least destabilizing dinucleotide bulge is $\binom{5'G\ AU\ C}{3'C\ \ \ G}$.

Free Energy Parameters for Dinucleotide Bulges. Currently, the free energy penalty for all dinucleotide bulges regardless of bulge and closing pair sequence is 2.8 kcal/mol.[17] Multiple models were tested to improve upon the current method of prediction. The model that resulted in the lowest average deviation between the predicted and experimental values is

$$\Delta G^\circ_{37,\text{dint bulge}} = \Delta G^\circ_{37,\text{bulge}} + \Delta G^\circ_{37,\text{AU}} + \Delta G^\circ_{37,\text{GU}} \quad (1)$$

As shown in Table 4, $\Delta G^\circ_{37,\text{bulge}}$ is dependent upon the type of nucleotides in the bulge (purine or pyrimidine) and their order. Thus, $\Delta G^\circ_{37,\text{bulge}}$ is a 3.06 kcal/mol penalty for two bulged purines, a 2.94 kcal/mol penalty for two bulged pyrimidines, a 2.71 kcal/mol penalty for a 5′-purine-pyrimidine-3′ bulge, and a 2.41 kcal/mol penalty for a 5′-pyrimidine-purine-3′ bulge. $\Delta G^\circ_{37,\text{AU}}$ is a 0.45 kcal/mol penalty for each A-U pair adjacent to the bulge; $\Delta G^\circ_{37,\text{GU}}$ is a −0.28 bonus for each G-U base pair adjacent to the bulge. It is important to note the A-U adjacent pair penalty of 0.45 kcal/mol is in addition to the 0.45 kcal/mol

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

penalty for a terminal A-U pair that is used when calculating the free energy contribution of the stem. As an example, the stability of 5′GACUACCGUG3′/3′CUGAGCAC5′ is predicted to be −5.68 kcal/mol, as shown below:

$$\Delta G^\circ_{37}\begin{pmatrix}\text{GACUACCGUG}\\ \text{CUGA}\quad\text{GCAC}\end{pmatrix} = \Delta G^\circ_{37,i} + \Delta G^\circ_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} + \Delta G^\circ_{37,\text{terminal U-A}} + \Delta G^\circ_{37,\text{dint bulge}} + \Delta G^\circ_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{GU}\\ \text{CA}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix} \quad (2)$$

Substituting eq 1 into eq 2 yields

$$\Delta G^\circ_{37}\begin{pmatrix}\text{GACUACCGUG}\\ \text{CUGA}\quad\text{GCAC}\end{pmatrix} = \Delta G^\circ_{37,i} + \Delta G^\circ_{37}\begin{pmatrix}\text{GA}\\ \text{CU}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{AC}\\ \text{UG}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{CU}\\ \text{GA}\end{pmatrix} + \Delta G^\circ_{37,\text{terminal U-A}} + \Delta G^\circ_{37,\text{bulge}} + (\Delta G^\circ_{37,\text{AU}} + \Delta G^\circ_{37,\text{GU}}) + \Delta G^\circ_{37}\begin{pmatrix}\text{CG}\\ \text{GC}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{GU}\\ \text{CA}\end{pmatrix} + \Delta G^\circ_{37}\begin{pmatrix}\text{UG}\\ \text{AC}\end{pmatrix} \quad (3)$$

Replacing the terms with values yields

$$\begin{aligned}\Delta G^\circ_{37}&\begin{pmatrix}\text{GACUACCGUG}\\ \text{CUGA}\quad\text{GCAC}\end{pmatrix}\\ &= 4.09 + (-2.35) + (-2.24) + (-2.08) + 0.45\\ &\quad + (2.71 + 0.45 + 0) + (-2.36) + (-2.24)\\ &\quad + (-2.11)\\ &= -5.68\ \text{kcal/mol}\end{aligned} \quad (4)$$

which is close to the measured value of −5.86 kcal/mol. This new sequence-dependent model predicted the experimental values of the dinucleotide bulges with an average difference of 0.17 kcal/mol from the experimental value (Table 3).

DISCUSSION

Database Searching. While the database search yielded 220 possible combinations of bulge and closing pair sequence combinations, it is likely that some of the possible sequence combinations that were not found in the database do exist in nature and would have been found in a larger secondary structure database. There may also be a structural explanation for why certain sequence combinations occur more frequently than others in nature, though this would require extensive structural studies.

It is interesting to note that 8 of the top 10 most frequent bulge sequences in the database contain at least one A-U closing base pair and account for 28% of the bulge and corresponding closing base pair sequences in the database. Similarly, 3 of the top 10 most frequent sequences contain at least one G-U closing base pair and account for 11% of the total dinucleotide bulges and corresponding closing base pair sequences in the database. The remaining closing base pair, G-C, is present in 6 of the top 10 most frequent bulge sequences in the database and accounts for 31% of the total dinucleotide bulges and closing pair sequences in the database.

Thermodynamic Contributions of a Dinucleotide Bulge to Duplex Thermodynamics. All of the dinucleotide bulges included in this study destabilized the duplex. This destabilization is expected, and it is assumed that every dinucleotide bulge disrupts what otherwise would be stabilizing stacking interactions between neighboring nucleotides.[9] Also, the presence of a bulge likely places strain on the adjacent base pairs, and the resulting bend or kink at the bulge site may destabilize the helix.

The bulges that consist of only pyrimidines or only purines seem to impart a larger penalty than the bulges that consist of one pyrimidine and one purine. This seems to suggest that the identity of the bulge does play a role in the overall destabilization of the dinucleotide bulge. This conclusion favors a sequence-dependent model over a sequence-independent model that attributes a constant value for all dinucleotide bulges regardless of sequence. If nature selected bulges on the basis of stability, we would expect bulges of one purine and one pyrimidine to be the most common. However, the most common dinucleotide bulge (5′-AA-3′) contains two purines.

Petrov, Zirbel, and Leontis[25] have identified several structural motifs adopted by dinucleotide bulges (Figure 1). Some examples of these structural motifs include (a) a bulge with its nucleotides in an extensive stacking arrangement with the closing base pairs and the nucleotides in the bulge but no pairing by the bulge nucleotides, (b) a bulge with its nucleotides "flipped out" and not stacking on each other or the closing base pairs and with no pairing by the bulge nucleotides, (c) a bulge with one bulge nucleotide pairing with a closing nucleotide on the opposite strand creating a base triple, and (d) a bulge with one bulge nucleotide pairing with a closing nucleotide on the opposite strand creating a base triple and the other bulge nucleotide pairing with a closing nucleotide on the same strand creating a second base triple. Although we have observed some sequence stability patterns and Petrov, Zirbel, and Leontis identified some structural motifs, we were unable to correlate the thermodynamic contribution of the bulge to stacking, hydrogen bonding, or structural features in general. Our limited knowledge of structural motifs in dinucleotide bulges does not provide enough information about stacking of the bulge nucleotides with each other, stacking of the bulge nucleotides with the closing base pairs, stacking of the closing base pairs with each other, pairing of the bulge nucleotides with each other, or pairing of the bulge nucleotide with the closing base pairs to develop any meaningful relationship between structure and stability. A significant effort to identify sequence−structure patterns in dinucleotide bulges may shed some light on the relationship between thermodynamic stability and structure.

Improving the Model Used To Predict Dinucleotide Thermodynamics. The current model used by *RNAstructure* to predict the free energy contribution of dinucleotide bulges was derived by averaging the measured free energy of only six dinucleotide bulges. Four of these six sequences were not included in this study as they had issues with possible self-association and/or exhibited a non-two-state melt as discussed previously.[17] The two that were used had the same bulge and nearest neighbors. The large range of bulge contributions seen here, approximately 1.5 kcal/mol (Table 3), suggested that a sequence-dependent model may be better at predicting the free energy contribution of a dinucleotide bulge. Therefore, the data collected here (a 3-fold increase in sample size in comparison to

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

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

Dataset 1				Dataset 2			
Bulge and NN[b]	Frequency[c]	Percentage[d]	Reference[e]	Bulge[f]	Frequency[c]	Percentage[d]	Reference[f]
G AA G C C	240	13.05	g	**AA**	507	27.57	g,h
G CA U U A	85	4.62		**CA**	249	13.54	h
G UG A C U	81	4.40		**AC**	217	11.80	h
C CA G G C	76	4.13	h	**AG**	180	9.79	h
A AG G U C	75	4.08		**UG**	152	8.27	
U AA A G U	65	3.53		**GA**	144	7.83	h
U AC C A G	62	3.37	h	**UA**	132	7.18	
U AG U G A	52	2.83	h	**CG**	55	2.99	
U AC G A C	44	2.39	h	**UC**	40	2.18	h
U AC U A A	43	2.34	h	**AU**	39	2.12	h
G UA U C G	41	2.23		**GU**	32	1.74	
A AA C U G	39	2.12		**UU**	32	1.74	
G CA G C C	39	2.12	h	**CC**	18	0.98	
A AC G U C	38	2.07		**GG**	17	0.92	
U GA A G U	34	1.85	h	**GC**	14	0.76	
G AA U C A	33	1.79	h	**CU**	11	0.60	
C GA A G U	31	1.69					
U GA C A G	26	1.41					
U AA C A G	24	1.31					
C AG C G G	23	1.25					

[a]Not all bulges found in the database are shown because of space limitations. [b]Dinucleotide bulge and nearest neighbor sequence. [c]Frequency of occurrence in the database searched. [d]Percent of 1839 dinucleotide bulges, the total number found in the database search. [e]Reference in which data were reported. [f]Dinucleotide bulge sequence. [g]From ref 17. [h]From this work.

the previous dinucleotide bulge data set) were used to derive a sequence-dependent model (Table 4). On average, the sequence-independent model[17] predicts a free energy that is 0.42 kcal/mol different from the experimental value. The sequence-dependent model proposed here predicts an average free energy that is only 0.17 kcal/mol different from the experimental value, over a 2-fold improvement. Additionally, the standard deviation of the predictions using the sequence-independent model is 0.23 kcal/mol, which is almost double that of the sequence-dependent model (0.13 kcal/mol) (Table 3). It is important to note that bulges whose stabilities are poorly predicted by the sequence-independent model are more accurately predicted with the sequence-dependent model. For example, when using the sequence-independent model, there are seven bulges with predicted values varying from the experimental values by >0.5 kcal/mol, with the greatest difference being 0.96 kcal/mol. In contrast, the sequence-dependent model has only one bulge whose predicted value varies from the experimental value by >0.5 kcal/mol, more specifically, only 0.51 kcal/mol.

The sequence-dependent model predicts the experimental dinucleotide bulge contribution well and is consistent with the published model used to predict the free energy contribution of trinucleotide bulges.[19] Both models are sequence-dependent; they rely on the identity of the bulge nucleotides and the adjacent base pairs to predict the free energy contribution to the bulge. Also, for dinucleotide and trinucleotide bulges, bulges consisting of all pyrimidines or all purines are more destabilizing than bulges consisting of both pyrimidines and purines. Finally, both dinucleotide bulges and trinucleotide bulges utilize a penalty for an A-U adjacent pair (0.45 kcal/mol for dinucleotide bulges and 0.49 kcal/mol for trinucleotide bulges) and a bonus for a G-U

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

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

		Analysis of T_m Dependence/Errors				Analysis of Melt Curve Fits/Errors			
Frequency[b]	Sequence[c]	ΔH° (kcal/mol)	ΔS° (cal/Kmol)	ΔG°$_{37}$ (kcal/mol)	T_m (°C)[d]	ΔH° (kcal/mol)	ΔS° (cal/K·mol)	ΔG°$_{37}$ (kcal/mol)	T_m (°C)[d]
240	GC GAAG CGA[e] ACG C C GC	-54.4	-154	-6.7	38.2	-35.2	-91.6	-6.8	39.5
	GC GAAG UCA[e] ACG C C AG	-49.4	-138	-6.6	37.5	-40.9	-110	-6.8	39.0
76	GAG CCAG GUG CUC G C CAC	-70.5 ± 4.5	-198.4 ± 14.1	-8.92 ± 0.13	47.9	-66.4 ± 4.9	-185.6 ± 15.8	-8.85 ± 0.10	48.2
62	GAC UACC GUG CUG A G CAC	-61.6 ± 2.2	-179.7 ± 7.3	-5.86 ± 0.05	33.7	-62.3 ± 3.1	-182.1 ± 10.2	-5.86 ± 0.09	33.7
52	GAG UAGU GUC CUC G A CAG	-62.4 ± 6.3	-183.4 ± 20.8	-5.49 ± 0.22	31.9	-62.7 ± 4.9	-184.3 ± 15.9	-5.52 ± 0.13	32.1
44	GAC UACG CUG CUG A C GAC	-63.7 ± 3.7	-184.9 ± 12.1	-6.40 ± 0.07	36.4	-64.1 ± 6.8	-186.0 ± 21.8	-6.45 ± 0.17	36.6
43	GAC UACU CUG CUG A A GAC	-70.2 ± 10.3	-212.0 ± 34.2	-4.46 ± 0.48	28.1	-69.8 ± 15.3	-210.4 ± 50.9	-4.51 ± 0.53	28.3
39	GAG GCAG GUG CUC C C CAC	-89.9 ± 3.8	-262.4 ± 12.0	-8.55 ± 0.08	44.1	-85.5 ± 8.4	-248.3 ± 26.8	-8.51 ± 0.09	44.4
38	GAG AACG CUG CUC U C GAC	-61.7 ± 2.3	-178.3 ± 7.5	-6.43 ± 0.05	36.5	-60.2 ± 5.9	-173.1 ± 19.4	-6.50 ± 0.08	36.8
34	GAC UGAA CUG CUG G U GAC	-58.6 ± 3.0	-172.5 ± 10.1	-5.09 ± 0.11	29.5	-56.7 ± 3.8	-166.2 ± 12.5	-5.13 ± 0.17	29.5
33	GAG GAAU CUG CUC C A GAC	-65.0 ± 9.8	-188.9 ± 31.5	-6.45 ± 0.47	36.6	-63.0 ± 8.7	-182.1 ± 28.7	-6.52 ± 0.33	36.9
26	GAC UGAC CUG CUG A G GAC	-64.6 ± 8.8	-188.8 ± 28.6	-6.03 ± 0.36	34.6	-64.2 ± 10.9	-187.4 ± 35.6	-6.05 ± 0.32	34.7
24	GAC UAAC CUG CUG A G GAC	-68.4 ± 4.5	-200.5 ± 14.6	-6.20 ± 0.09	35.5	-68.1 ± 7.6	-199.5 ± 24.5	-6.23 ± 0.09	35.6
23	GAG CAGC GUG CUC G G CAC	-68.3 ± 3.8	-196.2 ± 12.1	-7.41 ± 0.06	41.0	-63.6 ± 3.4	-181.2 ± 10.6	-7.38 ± 0.10	41.2
21	GAC CAAC GUG CUG G G CAC	-72.2 ± 3.61	-209.0 ± 11.6	-7.35 ± 0.05	40.6	-67.0 ± 2.9	-192.3 ± 8.9	-7.34 ± 0.14	40.8
11	GAG GCAU CUG CUC C A GAC	-59.6 ± 2.51	-169.3 ± 8.1	-7.12 ± 0.03	40.1	-64.4 ± 11.7	-184.6 ± 37.7	-7.14 ± 0.18	40.0
11	GAC UAAC CUG CUG G G GAC	-53.4 ± 5.6	-149.3 ± 18.1	-7.08 ± 0.20	40.2	-67.2 ± 17.0	-194.0 ± 55.0	-7.01 ± 0.33	39.2
10	GAG GAUC GUG CUC C G CAC	-75.5 ± 2.9	-217.6 ± 9.2	-7.98 ± 0.05	43.1	-74.0 ± 8.0	-213.1 ± 25.4	-7.95 ± 0.17	43.1
7	GAC AUCG CUG CUG U U GAC	-61.3 ± 0.8	-181.0 ± 2.7	-5.17 ± 0.03	30.2	-62.9 ± 4.0	-186.2 ± 13.3	-5.10 ± 0.16	30.1
6	CUG GUCG CUC GAC C C GAG	-66.2 ± 0.9	-186.2 ± 2.8	-8.48 ± 0.02	46.4	-69.1 ± 3.3	-195.1 ± 10.4	-8.55 ± 0.10	46.4

[a]Measurements were taken in 1 M NaCl, 20 mM sodium cacodylate, and 0.5 M EDTA (pH 7). [b]Frequency of occurrence obtained from database described in Materials and Methods. [c]The dinucleotide bulge is identified in 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 an oligomer concentration of 10^{-4} M. [e]Melt data from ref 17.

adjacent pair (−0.28 kcal/mol for dinucleotide bulges and −0.56 kcal/mol for trinucleotide bulges).[19] This G-U bonus is also present in the predictive model for single-nucleotide bulges.[18]

With the collection of new thermodynamic data for 18 dinucleotide bulges (and one additional bulge from the literature), new experimental data for free energy and secondary

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

Table 3. Contribution of Dinucleotide Bulges to Duplex Thermodynamics

	ΔG°37 (kcal/mol)				
		Sequence Independent Model		Sequence Dependent Model	
Sequence[a]	Measured[b]	Prediction[c]	Difference[d]	Prediction[e]	Difference[f]
GAAG[g] C C	3.57	2.8	0.77	3.06	0.51
GAAG[g] C C	3.08	2.8	0.28	3.06	0.02
CCAG G C	2.45	2.8	0.35	2.41	0.04
UACC A G	2.98	2.8	0.18	3.16	0.18
UAGU G A	3.25	2.8	0.45	3.23	0.02
UACG A C	3.34	2.8	0.54	3.10	0.21
UACU A A	3.76	2.8	0.96	3.61	0.15
GCAG C C	2.66	2.8	0.14	2.41	0.25
AACG U C	3.42	2.8	0.62	3.16	0.26
UGAA G U	3.16	2.8	0.36	3.23	0.07
GAAU C A	3.24	2.8	0.44	3.51	0.27
UGAC A G	3.55	2.8	0.75	3.51	0.04
UAAC A G	3.38	2.8	0.58	3.51	0.13
CAGC G G	3.06	2.8	0.26	3.06	0.00
CAAC G G	3.12	2.8	0.32	3.06	0.06
GCAU C A	2.57	2.8	0.23	2.86	0.29
UAAC G G	2.64	2.8	0.16	2.78	0.14
GAUC C G	2.33	2.8	0.47	2.71	0.38
AUCG U U	3.33	2.8	0.53	3.10	0.23
GUCG C C	2.73	2.8	0.07	2.93	0.20
Average[h]			0.42 ± 0.23		0.17 ± 0.13

[a]The dinucleotide 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′. [b]Experimental free energy contribution of the bulge calculated as mentioned in the text. [c]Free energy prediction made by the sequence-independent model (ref 17). [d]Absolute difference between the free energy predicted by the sequence-independent model (ref 17) and the experimental free energy. [e]Free energy prediction made by the seqence-dependent model proposed here. [f]Absolute difference between the free energy predicted by the sequence-dependent model and the experimental free energy. [g]Data from ref 17. [h]Average (absolute value) deviation.

structure prediction software are available. The newly derived sequence-dependent predictive model can also be incorporated into prediction software. These should both improve prediction of RNA stability and secondary structure from sequence.

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

$\Delta G^\circ_{37,\text{dint bulge}}$ parameter	free energy contribution (kcal/mol)
$\Delta G^\circ_{37,\text{bulge}}$[a]	
two purines	3.06 ± 0.10
two pyrimidines	2.94 ± 0.20
one purine, one pyrimidine (5′RY3′)	2.71 ± 0.16
one pyrimidine, one purine (5′YR3′)	2.41 ± 0.15
$\Delta G^\circ_{37,\text{AU}}$[b] (kcal/mol)	0.45 ± 0.11
$\Delta G^\circ_{37,\text{GU}}$[c] (kcal/mol)	−0.28 ± 0.16

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

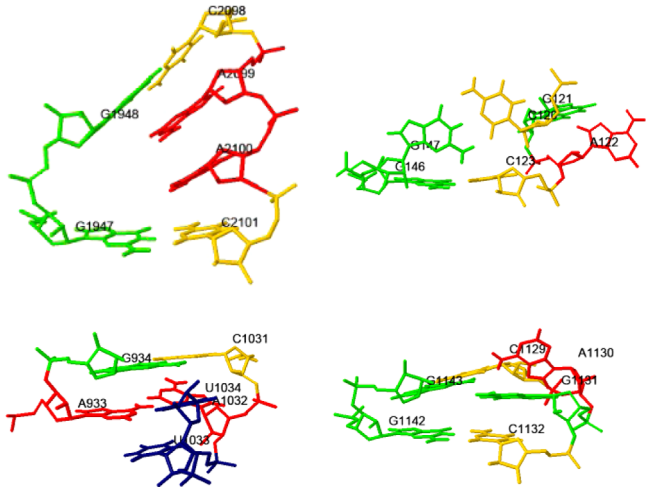


Figure 1. Dinucleotide bulge motifs identified from three-dimensional structures.[25] The nucleotides are colored by residue type, where A is red, C yellow, G green, and U blue, and are labeled by residue type and sequence number. Motif IL_04466.1 (top left) with sequence 5′CAAC3′/3′GG5′ [Protein Data Bank (PDB) entry 3U5H]. In this structure, the closing G-C pairs are canonical Watson−Crick pairs, and the bulge nucleotides are inserted into the helix and stack on each other and the C's of the closing G-C pair. The bulge nucleotides do not form any pairs. Motif IL_23448.1 (top right) with sequence 5′CGAC3′/3′GG5′ (PDB entry 4ERD). In this structure, the closing G-C pairs are canonical Watson−Crick pairs and the bulge nucleotides are "flipped out" and do not stack on each other or the closing base pairs. The bulge nucleotides do not form any pairs, and the bulge disrupts the closing base pairs from stacking with each other. Motif IL_37964.2 (bottom left) with sequence 5′CAUU3′/3′GA5′ (PDB entry 3J7A). In this structure, the closing A933-U1034 and G934-C1031 pairs are canonical Watson−Crick pairs and there is significant stacking between the residues. Bulge residue A1032 is pairing with A933 on the opposite strand, forming an A-A-U base triple. Motif IL_73000.4 (bottom right) with sequence 5′CAGC3′/3′GG5′ (PDB entry 2AW7). In this structure, the closing G-C pairs are canonical Watson−Crick pairs, with some stacking between the nucleotides. Bulge residue G1131 is pairing with G1143 on the opposite strand, forming a G-G-C base triple. In addition, bulge residue A1130 is pairing with the adjacent nucleotide C1129, forming a G-C-A base triple.

AUTHOR INFORMATION

Corresponding Author

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

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health via Grant 2R15GM085699-02.

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) Morris, K. V., and Mattick, J. S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet. 15*, 423−437.
(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) Valadkhan, S. (2005) snRNAs as the catalysts of pre-mRNA splicing. *Curr. Opin. Chem. Biol. 9*, 603−608.
(5) 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.
(6) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406−3415.
(7) Hofacker, I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res. 31*, 3429−3431.
(8) Li, X., Quon, G., Lipshitz, H. D., and Morris, Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA 16*, 1096−1107.
(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) 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*, 189−210.
(11) 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.
(12) 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.
(13) 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.
(14) O'Connor, C. M., and Collins, K. (2006) A novel RNA binding domain in *Tetrahymena* telomerase p65 initiates hierarchical assembly of telomerase holoenzyme. *Mol. Cell. Biol. 26*, 2029−2036.
(15) Du, Z., Yu, J., Ulyanov, N. B., Andino, R., and James, T. L. (2004) Solution structure of consensus stem-loop D RNA domain that plays important roles in translation and replication in enteroviruses and rhinoviruses. *Biochemistry 43*, 11959−11972.
(16) Rhim, H., and Rice, A. P. (1994) Functional significance of the dinucleotide bulge in stem-loop 1 and stem-loop 2 of HIV-2 TAR RNA. *Virology 202*, 202−211.
(17) Longfellow, C. E., Kierzek, R., and Turner, D. H. (1990) Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry 29*, 278−285.
(18) 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.
(19) Murray, M. H., Hard, J. A., and Znosko, B. M. (2014) Improved model to predict the free energy contribution of trinucleotide bulges to RNA duplex stability. *Biochemistry 53*, 3502−3508.
(20) 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 for predicting the free energy contribution of sequence asymmetric tandem mismatches. *Biochemistry 47*, 4329−4336.
(21) 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.
(22) McDowell, J. A. (1996) *Meltwin*, version 3.5.
(23) Vanegas, P. L., Horwitz, T. S., and Znosko, B. M. (2012) Effects of non-nearest neighbors on the thermodynamic stability of RNA GNRA hairpin tetraloops. *Biochemistry 51*, 2192−2198.
(24) 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.
(25) Petrov, A. I., Zirbel, C. L., and Leontis, N. B. (2013) Automated classification of RNA 3D motifs and the RNA 3D motif atlas. *RNA 19*, 1327−1340.

DOI: 10.1021/acs.biochem.5b00474
Biochemistry 2015, 54, 5290−5296