

# Thermodynamic characterization of tandem mismatches found in naturally occurring RNA

Martha E. Christiansen and Brent M. Znosko\*

Department of Chemistry, Saint Louis University, St Louis, MO 63103, USA

Received March 6, 2009; Revised May 4, 2009; Accepted May 15, 2009

## ABSTRACT

Although all sequence symmetric tandem mismatches and some sequence asymmetric tandem mismatches have been thermodynamically characterized and a model has been proposed to predict the stability of previously unmeasured sequence asymmetric tandem mismatches [Christiansen, M.E. and Znosko, B.M. (2008) *Biochemistry*, 47, 4329–4336], experimental thermodynamic data for frequently occurring tandem mismatches is lacking. Since experimental data is preferred over a predictive model, the thermodynamic parameters for 25 frequently occurring tandem mismatches were determined. These new experimental values, on average, are 1.0 kcal/mol different from the values predicted for these mismatches using the previous model. The data for the sequence asymmetric tandem mismatches reported here were then combined with the data for 72 sequence asymmetric tandem mismatches that were published previously, and the parameters used to predict the thermodynamics of previously unmeasured sequence asymmetric tandem mismatches were updated. The average absolute difference between the measured values and the values predicted using these updated parameters is 0.5 kcal/mol. This updated model improves the prediction for tandem mismatches that were predicted rather poorly by the previous model. This new experimental data and updated predictive model allow for more accurate calculations of the free energy of RNA duplexes containing tandem mismatches, and, furthermore, should allow for improved prediction of secondary structure from sequence.

## INTRODUCTION

Non-canonical regions account for about half of the secondary structure of RNA (1). In particular, tandem

mismatches or  $2 \times 2$  nucleotide internal loops, are widespread and are found quite frequently. For example, tandem mismatches have been found to occur in the rRNA of *Escherichia coli* (2), *Haloflex volcanii* (3) and *Haloarcula marismortui* (4); the mitochondrial 16S-like rRNA of *Clamydomonas reinhardtii* (3); the genomic RNA of the dengue-3 virus (5); turnip crinkle viral RNA (6); and myotonic dystrophy type 2 RNA (7), to name a few. It is important to note, however, that tandem mismatches not only occur in a variety of different RNA and in different organisms, but they also serve functional roles within the RNA. For example, the aminoglycoside ampramycin binds to a tandem mismatch in bacteria and halts replication (8). Another functional role of tandem mismatches occurs in the J4/5 loop of the group I intron of *Candida albicans*. This tandem mismatch is the docking site for the first step of the self-splicing reaction required for processing of rRNA (9).

Due to the natural occurrence of tandem mismatches and their functional roles within certain RNAs, the ability to accurately predict tandem mismatches in secondary structures can lead to a better understanding of tertiary interactions and structure–function relationships and aid in the design of pharmaceuticals. As a result, thermodynamics of tandem mismatches have been studied quite extensively (10–19). Recently, we reported a complete periodic table of sequence symmetric tandem mismatches and proposed an updated model for predicting the free energy contribution of previously unmeasured sequence asymmetric tandem mismatches (19). However, despite these studies, only 8% of all possible tandem mismatch—nearest neighbor combinations have experimentally determined thermodynamic parameters. Due to a lack of data, current algorithms that predict RNA secondary structure from sequence (1,20–24) are based on several assumptions and may be oversimplified.

An obvious way to improve the current algorithms is to expand the experimental data set for tandem mismatches. Ideally, the algorithms should contain experimental parameters for each possible tandem mismatch (1095 when defined as the tandem mismatch and the adjacent nearest neighbors). Although this is unrealistic, what is

\*To whom correspondence should be addressed. Tel: +1 314 977 8567; Fax: +1 314 977 2521; Email: znoskob@slu.edu

possible is to measure the thermodynamics of tandem mismatches that occur frequently in nature. This would provide algorithms with experimental parameters for the tandem mismatches that users of these programs would most likely be interested in. Thus, the only tandem mismatches that would rely on a predictive model would be those that occur less frequently in nature. In addition, more experimental data would increase the amount of tandem mismatch data that is available from which the predictive model was derived, enhancing the accuracy of the model. With this in mind, the frequency of occurrence of tandem mismatches in a dataset of RNA secondary structures was determined. Twenty five of the most frequently occurring tandem mismatches were then optically melted in order to derive their corresponding thermodynamic parameters. In addition, this thermodynamic data was used to update the predictive model used for tandem mismatches that do not have experimental values. This additional experimental data and updated predictive model can now be incorporated into secondary structure algorithms.

## MATERIALS AND METHODS

### Compiling and searching a database for tandem mismatches

A database of 1899 RNA secondary structures containing 123 small subunit rRNAs (3), 223 large subunit rRNAs (25,26), 309 5S rRNAs (27), 484 tRNAs (28), 91 signal recognition particles (29), 16 RNase P RNAs (30), 100 group I introns (31,32), 3 group II introns (33) and 450 miRNAs (34–36). This database was searched for tandem mismatches, and the number of occurrences for each type of tandem mismatch was tabulated. In this work, G–U pairs are considered to be canonical base pairs.

### Design of sequences for optical melting studies

Sequences of tandem mismatches and nearest neighbors were designed to represent those found most frequently in the database described above. Each duplex chosen was 10 base pairs long, with the tandem mismatches occurring in the middle of the duplex. On either side of the tandem mismatch are the adjacent nearest neighbors and three additional canonical base pairs. The terminal base pairs are G–C base pairs in order to prevent end fraying of the sequence during melting. The duplexes were also designed to have a melting temperature between 35°C and 55°C and to have minimal formation of hairpin structures or mis-aligned duplexes.

### RNA synthesis and purification

Oligonucleotides were ordered from the Keck Lab at Yale University (New Haven, CT), Azco BioTech, Inc. (San Diego, CA) or Integrated DNA Technologies (Coralville, IA). The synthesis and purification of the oligonucleotides followed standard procedures that were described previously (19,37,38).

### Optical melting experiments and thermodynamics

Optical melting experiments were performed in 1 M NaCl, 20 mM sodium cacodylate and 0.5 mM Na<sub>2</sub>EDTA (pH 7.0). Melting curves (absorbance versus temperature) were obtained, and duplex thermodynamics were determined as described previously (19,37,38). The thermodynamic contributions of tandem mismatches to duplex thermodynamics ( $\Delta G^{\circ}_{37, \text{tandem mismatch}}$ ,  $\Delta H^{\circ}_{\text{tandem mismatch}}$  and  $\Delta S^{\circ}_{\text{tandem mismatch}}$ ) were determined by subtracting the Watson–Crick contribution (39) from the measured duplex thermodynamics. This type of calculation was described previously (19).

### Linear regression and tandem mismatch thermodynamic parameters

Data collected for the 25 sequence asymmetric tandem mismatches in this study were combined with previously published data for 72 sequence asymmetric tandem mismatches (12,15–18). Similar to what was done previously to derive parameters for sequence asymmetric tandem mismatches (19), six variables were used for linear regression: (i) tandem mismatches with a U•U pair adjacent to an R•R pair, tandem mismatches with one G•A or A•G pair adjacent to a Y•Y pair, or tandem mismatches with any combination of A•C, U•C, C•U, C•C, C•A or A•A pairs; (ii) tandem mismatches with any combination of adjacent G•A and A•G pairs or adjacent U•U pairs; (iii) tandem mismatches with a U•U pair adjacent to a Y•Y (not U•U), C•A, or A•C pair; (iv) tandem mismatches with a single G•G pair not adjacent to a U•U pair; (v) tandem mismatches with an A–U or U–A nearest neighbor; and (vi) tandem mismatches with a G–U or U–G nearest neighbor. The calculated experimental contribution of the tandem mismatch to duplex stability was used as a constant when doing linear regression. To simultaneously solve for each variable, the LINEST function of *Microsoft Excel* was used for linear regression. Based on this dataset, the fourth parameter did not contribute to tandem mismatch stability. Therefore, this parameter was removed and linear regression was re-run with the remaining parameters. Various other parameters and/or combinations of parameters were tried, but the parameters described here resulted in free energy contributions that agreed closely with the experimental data and error values that are comparable to those of the *RNAstructure* algorithm (1,20,21).

### Structural investigation

Six previously solved three-dimensional structures containing frequently occurring tandem mismatches were downloaded from the Protein Data Bank (PDB) (40). These structures were viewed using Insight II software (Accelrys Software Inc.), and base–base hydrogen bonding between the mismatch nucleotides was investigated. A hydrogen bond was considered to be present when the distance between the proton on the donor atom and the heavy atom acceptor was less than 2.5 Å and when the angle between the heavy atom donor, proton, and heavy atom acceptor was between 120° and 180°.

## RESULTS

### Database searching

The database described above was searched for tandem mismatches. In this database, 1092 tandem mismatches were found, averaging about one occurrence in every two sequences. Table 1 shows a summary of the database results obtained. The first set of data lists frequency and percent occurrence when the mismatch nucleotides and nearest neighbors are specified. Categorizing tandem mismatches in this fashion results in 350 types of mismatches in the database. The 30 tandem mismatch types listed in the first data set (Table 1) account for 51% of the total number of tandem mismatches found. The 320 types of tandem mismatches not shown account for the remaining 49%; however, each type represents <0.6% of the total number of tandem mismatches found. When categorized in this manner, previous studies account for only 21% of the total number of tandem mismatches found, but after adding the data reported here, this percentage increases to 59%. Similarly, previous studies thermodynamically characterized only seven types of tandem mismatches in the top 30, but after adding the data reported here, all of the tandem mismatches in the top 30 have been studied.

The second set of data (Table 1) lists frequency and percent occurrence when only the tandem mismatch sequence is specified (nearest neighbors are not considered). Categorizing tandem mismatches in this fashion results in 55 types of tandem mismatches in the database. The 30 tandem mismatches listed in the second data set (Table 1) account for 93% of the total number of tandem mismatches found. The 25 types of tandem mismatches not shown account for the remaining 7%, with each mismatch representing <0.5% of the total number of tandem mismatches found. When categorized in this manner, previous studies account for 93% of the total number of tandem mismatches found. Adding the data reported here did not change this percentage.

The third set of data (Table 1) lists frequency and percent occurrence of 5' and 3' nearest neighbor combinations. Categorizing tandem mismatches in this fashion results in 21 types of nearest neighbor combinations in the database, representing all possible types of nearest neighbor combinations. When categorized in this manner, previous studies account for 55% of all nearest neighbor combinations, but after adding the data reported here, this percentage increases to 92%.

The fourth set of data (Table 1) lists frequency and percent occurrence of the tandem mismatch nucleotides when A and G are categorized as purines (R) and C and U are categorized as pyrimidines (Y). Categorizing tandem mismatches in this fashion results in 10 types of tandem mismatches, representing all possible combinations. When categorized in this manner, previous studies account for 99.5% of all combinations. Adding the data reported here did not change this percentage. The only combination that has not been measured is 5'RR3'/3'YY5' (with 5'AA3'/3'CC5' being the only possible tandem mismatch sequence in this category).

### Thermodynamic parameters

Table 2 shows the thermodynamic parameters of duplex formation that were obtained from fitting each melting curve to the two-state model and from the van't Hoff plot of  $T_M^{-1}$  versus  $\log(C_T/4)$ . Data for the duplexes containing the 30 most frequently occurring tandem mismatches in the database are shown in order of decreasing frequency. However, data for 33 duplexes are shown because three tandem mismatches were melted with two different stem sequences.

### Contribution of tandem mismatches to duplex thermodynamics

The contributions of the 33 tandem mismatches to duplex stability are listed in Table 3. The examination of the thermodynamic contributions of tandem mismatches to duplex thermodynamics indicates a large variance in the obtained thermodynamic parameters. Tandem mismatch contributions to enthalpy, entropy and free energy changes range from  $-31.1$  to  $38.4$  kcal/mol,  $-101.7$  to  $117.1$  cal/(K mol) and  $-1.7$  to  $3.7$  kcal/mol, respectively.

### Updated model for predicting the thermodynamics of previously unmeasured sequence asymmetric tandem mismatches

An updated model to predict the thermodynamics of previously unmeasured sequence asymmetric tandem mismatches was derived by compiling data for 72 sequence asymmetric tandem mismatches from previous works (12,15–18) and the data for 25 sequence asymmetric tandem mismatches from this work. Linear regression was then used to derive nearest neighbor parameters for predicting the contribution of a tandem mismatch to duplex thermodynamics. These parameters are shown in Table 4. Table 4 also lists parameters for calculating enthalpy and entropy contributions of rare sequence asymmetric tandem mismatches.

We previously proposed a model for predicting the free energy contribution of previously unmeasured sequence asymmetric tandem mismatches (19). Since the database of sequence asymmetric tandem mismatches has increased from 72 to 97 tandem mismatches with the data reported here, we re-derived the parameters used in the previous model to determine if the additional data would change any of the parameters published previously (19). With the addition of the new data, there were some minor changes to the previous parameters (Table 4). Initially, as done previously, a parameter was derived for loops containing a G•G pair not adjacent to a U•U pair. It was found that these loops did not contribute to duplex stability. Therefore, this mismatch combination was removed as a parameter. Four of the remaining five parameters derived with the additional data are within experimental error of the previous data. The one value not within experimental error was the penalty added per A–U nearest neighbor; this penalty changed from 0.5 to 1.0 kcal/mol. The difference between the previous penalty and the new penalty is likely due to the small sample size (eight occurrences) used to generate this parameter previously; with the data

**Table 1.** Summary of database search results for tandem mismatches<sup>a</sup>

Dataset 1: Tandem mismatch with nearest neighbors				Dataset 2: Tandem mismatch				Dataset 3: 5' and 3' adjacent base pairs				Dataset 4: Tandem mismatch nucleotides classified as purine (R) or pyrimidine (Y)				
Tandem Mismatch <sup>b</sup>	Freq <sup>c</sup>	% <sup>d</sup>	Refs	Tandem Mismatch <sup>b</sup>	Freq <sup>c</sup>	% <sup>d</sup>	Refs	Closing bp	Freq <sup>c</sup>	% <sup>d</sup>	Refs	Tandem Mismatch <sup>b</sup>	Freq <sup>c</sup>	% <sup>d</sup>	Refs	
G GA G C AG U	57	5.2	This work	GA AG	217	19.9	This work (10,13,15,17)	C G	C G	134	12.3	This work (12,15,16)	RR RR	514	47.1	This work (12,16,19)
G AA G C AG C	43	3.9	(15)	AA AG	175	16.0	This work (15,17)	C G	A U	133	12.2	This work	YY YY	262	24.0	This work (11,15,18,19)
C GA G G AG C	41	3.8	(10)	UU UU	127	11.6	This work (11,14,15,19)	C G	G C	122	11.2	This work (10,11,12,18)	RY RY	142	13.0	This work (15,16,18)
A GA G U AG C	37	3.4	This work	AA GG	60	5.5	This work (15)	G C	G U	96	8.8	This work (17)	RR YR	63	5.8	This work (15,16)
C AA G G AG C	34	3.1	This work	UC UU	56	5.1	This work (15,18)	A U	G C	95	8.7	This work	RR RY	33	3.0	(15,16)
U UC U A UU A	30	2.7	This work	AC GU	36	3.3	This work (15)	A U	A U	90	8.2	This work	YR YY	30	2.7	(15,18)
U AA G A GG C	27	2.5	This work	AA AA	28	2.6	(10,12,14,15, 17,19)	G C	A U	78	7.1	This work	RY YY	28	2.6	This work (15,18)
C UU A G UU U	22	2.0	This work	AA CG	28	2.6	This work (15)	U A	A U	61	5.6	This work (14,16,19)	YR RY	8	0.7	(11,14,19)
C UU G G UU C	20	1.8	(11)	AC GA	26	2.4	This work (15)	C G	G U	57	5.2	This work	RY YR	6	0.5	(11,14,15,19)
U AC A A GU U	17	1.6	This work	CU CU	24	2.2	This work (15,18)	G C	C G	42	3.8	(13,14,19)	RR YY	6	0.5	
C AA G G CG U	16	1.5	This work	AU AU	19	1.7	(15,18)	A U	U A	40	3.7	This work (14,16,19)	Previously	1086	99.5	
C GA A G AG U	16	1.5	This work	AU CU	19	1.7	This work (15,18)	A U	G U	33	3.0	This work	New Total	1086	99.5	
C UU C G UU G	16	1.5	This work (15)	CA CG	18	1.6	This work (15)	A U	C G	32	2.9					
C AA A G AG U	14	1.3	This work	AU GU	18	1.6	This work (15)	C G	U G	22	2.0	(17)				
G CA A C CG U	14	1.3	This work	CU UU	18	1.6	(15)	U A	G U	16	1.5					
G AA G C AG U	13	1.2	This work	GU AU	15	1.4	(15,18)	A U	U G	14	1.3					
U AA G A AG C	13	1.2	This work	CU AU	13	1.2	(15,18)	G C	U G	13	1.2					
A AC U U GA A	13	1.2	This work	CG AG	12	1.1	(16)	G U	A U	9	0.8					
C GA G G AG U	13	1.2	This work	CC UU	10	0.9	(15)	G U	G U	2	0.2	(17)				
G UU A C UU U	12	1.1	This work	CA AG	12	1.1	(15)	U G	G U	2	0.2	(14,19)				
A GA G U AG U	11	1.0	This work	CC UU	11	1.0	(15)	G U	U G	1	0.1	(13,19)				
A UU G U UU C	11	1.0	This work	AA GA	10	0.9	(15,17)	Previously	600	54.9						
C AU C G GU G	10	0.9	This work (15)	CA UC	10	0.9		New Total	1008	92.3						
U AA C A GG G	10	0.9	This work	CA AC	8	0.7	(11,14,19)									
A AA G U AG U	9	0.8	This work	CC UC	8	0.7										
A AU A U CU U	9	0.8	This work	GA GG	8	0.7	(16)									
A CU A U CU U	9	0.8	This work	CG CG	7	0.6										
C GA C G AG G	9	0.8	(15)	CA UG	7	0.6										
U GA A A AG U	9	0.8	(15)	CU UC	7	0.6	(11,14,19)									
A AA G U AG C	7	0.6	This work	AC CA	6	0.5	(11,14,15,19)									
Previously	230	21.1		Previously	1017	93.1										
New Total	644	59.0		New Total	1017	93.1										

<sup>a</sup>Not all combinations in datasets 1 and 2 are shown because of space limitations. For each set of sequences, the top strand is written 5'–3', and the bottom strand is written 3'–5'.

<sup>b</sup>Duplexes are written in alphabetical order by the tandem mismatch nucleotides. If the tandem mismatch nucleotides are identical, then duplexes are written in alphabetical order by the nearest neighbors.

<sup>c</sup>Frequency of occurrence in the database.

<sup>d</sup>Percent out of 1092 tandem mismatches, the total number of tandem mismatches found in the database.

**Table 2.** Thermodynamic parameters for duplex formation<sup>a</sup>

Frequency <sup>b</sup>	Sequence <sup>c</sup>	Analysis of Curve Fit/Errors				Analysis of T <sub>m</sub> Dependence (ln plot)			
		−ΔG <sub>37</sub> <sup>o</sup> (kcal/mol)	−ΔH <sup>o</sup> (kcal/mol)	−ΔS <sup>o</sup> (eu)	T <sub>M</sub> (°C)	−ΔG <sub>37</sub> <sup>o</sup> (kcal/mol)	−ΔH <sup>o</sup> (kcal/mol)	−ΔS <sup>o</sup> (eu)	T <sub>M</sub> (°C)
57	GACG <sup>GA</sup> GCUG CUG <sub>AG</sub> UGAC	9.64 ± 0.29	77.5 ± 5.2	218.9 ± 16.0	50.0	9.66 ± 0.43	79.2 ± 10.5	224.1 ± 32.6	49.8
43	GAG <sup>AA</sup> GAG <sup>(15)</sup> CUC <sub>AG</sub> CUC	5.56 ± 0.2	54.0 ± 4.0	156.2 ± 13.4	31.5	5.42 ± 0.03	58.3 ± 1.4	170.6 ± 4.7	31.2
41	GGC <sup>GA</sup> GCC <sup>(10)</sup> CCG <sub>AG</sub> CGG	10.16	73.9	205.6	57.0	9.68	66.1	181.8	57.0
	GC <sup>GA</sup> GC <sup>(10)</sup> CG <sub>AG</sub> CG	4.17	28.6	78.7	21.5	4.01	30.4	85.0	20.1
37	GACA <sup>GA</sup> GCUG CUG <sub>AG</sub> CGAC	10.51 ± 0.51	84.3 ± 9.2	237.8 ± 28.0	52.4	10.45 ± 0.1	84.6 ± 3.2	239.0 ± 9.8	52.1
34	GACC <sup>AA</sup> GCUG CUG <sub>AG</sub> CGAC	12.55 ± 1.23	98.4 ± 18.3	276.8 ± 55.1	57.2	12.50 ± 0.49	99.1 ± 8.1	279.1 ± 24.6	56.9
30	GACU <sup>UC</sup> UCUG CUG <sub>AUU</sub> AGAC	7.68 ± 0.23	90.1 ± 8.4	265.6 ± 26.9	41.0	7.70 ± 0.22	91.8 ± 10.1	271.1 ± 32.2	41.0
27	GACU <sup>AA</sup> GCUG CUG <sub>AGG</sub> CGAC	7.92 ± 0.40	57.5 ± 8.1	159.7 ± 25.8	44.7	8.00 ± 0.67	66.1 ± 13.3	187.5 ± 41.7	44.0
21	GACC <sup>UU</sup> ACUG CUG <sub>GUU</sub> UGAC	8.54 ± 0.33	61.8 ± 10.9	171.6 ± 34.2	47.4	8.56 ± 0.11	68.4 ± 4.1	193.1 ± 12.9	46.5
19	CGC <sup>UU</sup> GCG <sup>(11)</sup> GCC <sub>UU</sub> CGC	7.26 ± 0.11	63.0 ± 3.5	179.7 ± 11.1	-	7.19 ± 0.09	61.3 ± 1.1	174.5 ± 3.5	44.8
17	GACU <sup>AC</sup> ACUG CUG <sub>AGU</sub> UGAC	6.52 ± 0.23	77.1 ± 11.9	227.5 ± 38.4	37.0	6.47 ± 0.10	71.9 ± 5.5	210.8 ± 17.7	36.7
16	GACC <sup>AA</sup> GCUG CUG <sub>CG</sub> UGAC	9.97 ± 0.22	92.2 ± 6.2	265.1 ± 19.3	49.0	9.94 ± 0.16	91.6 ± 4.6	263.3 ± 14.5	49.0
16	GACC <sup>GA</sup> ACUG CUG <sub>AG</sub> UGAC	8.08 ± 0.08	27.0 ± 4.7	61.0 ± 15.2	55.9	8.06 ± 0.03	27.5 ± 0.9	62.6 ± 2.8	55.2
14	GACC <sup>AA</sup> ACUG CUG <sub>AG</sub> UGAC	9.66 ± 0.16	80.7 ± 3.1	229.1 ± 9.7	49.5	9.74 ± 0.08	83.4 ± 2.5	237.4 ± 7.8	49.4
14	GACG <sup>CA</sup> ACUG CUG <sub>CG</sub> UGAC	6.83 ± 0.19	57.7 ± 5.1	164.1 ± 16.0	38.6	6.86 ± 0.03	69.4 ± 2.6	201.7 ± 8.4	38.5
13	GACC <sup>AA</sup> GCUG CUG <sub>AG</sub> UGAC	8.48 ± 0.10	61.0 ± 4.1	169.3 ± 12.9	47.2	8.47 ± 0.05	60.5 ± 1.9	167.7 ± 5.8	47.3
13	GACU <sup>AA</sup> GCUG CUG <sub>AG</sub> CGAC	9.48 ± 0.10	85.7 ± 3.1	245.7 ± 9.9	48.0	9.48 ± 0.07	86.6 ± 2.3	248.6 ± 7.3	47.9
13	GACA <sup>AC</sup> UCUG CUG <sub>GA</sub> AGAC	6.82 ± 0.16	82.9 ± 5.2	245.1 ± 16.8	38.1	6.83 ± 0.09	86.1 ± 5.9	255.4 ± 19.1	38.1
13	GACC <sup>GA</sup> GCUG CUG <sub>AG</sub> UGAC	11.33 ± 0.29	89.4 ± 5.0	251.6 ± 15.0	54.6	11.32 ± 0.2	89.5 ± 3.3	252.1 ± 10.0	54.5
11	GACA <sup>GA</sup> GCUG CUG <sub>AG</sub> UGAC	8.60 ± 0.10	82.4 ± 7.4	238.0 ± 23.5	45.0	8.64 ± 0.04	87.6 ± 1.9	254.4 ± 6.1	44.7
10	GACC <sup>AU</sup> CCUG CUG <sub>GU</sub> GGAC	9.87 ± 0.23	72.2 ± 7.4	201.0 ± 23.2	52.0	10.09 ± 0.14	80.0 ± 3.9	225.4 ± 12.2	51.4
	CUC <sup>AU</sup> CUC <sup>(15)</sup> GAG <sub>GU</sub> GAG	4.39 ± 0.2	48.3 ± 3.8	141.5 ± 12.8	23.9	4.34 ± 0.08	50.4 ± 1.4	148.4 ± 4.9	24.1
10	GACA <sup>UU</sup> GCUG CUG <sub>UU</sub> CGAC	9.59 ± 0.14	82.6 ± 7.9	235.4 ± 25.2	48.9	9.88 ± 0.09	94.3 ± 3.0	272.2 ± 9.5	48.4
10	GACC <sup>UU</sup> CCUG CUG <sub>GU</sub> GGAC	11.78 ± 1.00	91.0 ± 16.8	255.3 ± 50.9	56.0	12.03 ± 0.19	97.8 ± 3.6	276.5 ± 10.9	55.5

Continued

Table 2. Continued

Frequency <sup>b</sup>	Sequence <sup>c</sup>	Analysis of Curve Fit/Errors				Analysis of T <sub>m</sub> Dependence (ln plot)			
		−ΔG <sub>37</sub> <sup>o</sup> (kcal/mol)	−ΔH <sup>o</sup> (kcal/mol)	−ΔS <sup>o</sup> (eu)	T <sub>M</sub> (°C)	−ΔG <sub>37</sub> <sup>o</sup> (kcal/mol)	−ΔH <sup>o</sup> (kcal/mol)	−ΔS <sup>o</sup> (eu)	T <sub>M</sub> (°C)
	CUC <sup>UU</sup> CUC <sup>(15)</sup> GAG <sub>UU</sub> GAG	5.98 ± 0.10	67.6 ± 8.3	198.6 ± 26.8	34.5	5.88 ± 0.02	65.8 ± 1.2	193.2 ± 3.8	34.0
10	GACG <sup>UU</sup> ACUG CUGC <sub>UU</sub> UGAC	8.48 ± 0.15	75.8 ± 4.9	217.0 ± 15.8	45.2	8.83 ± 0.08	94.3 ± 3.5	275.4 ± 11.2	44.8
9	GACA <sup>AA</sup> GCGC CUGU <sub>AG</sub> UGAC	7.14 ± 0.06	72.4 ± 3.7	210.5 ± 12.1	39.6	7.12 ± 0.01	77.3 ± 1.6	226.3 ± 5.0	39.4
9	GACU <sup>AA</sup> CCUG CUGA <sub>GG</sub> GGAC	10.12 ± 0.39	83.7 ± 10.2	237.1 ± 31.6	50.9	10.34 ± 0.12	90.4 ± 3.2	258.2 ± 9.9	50.7
9	GACA <sup>AU</sup> ACUG CUGU <sub>CU</sub> UGAC	6.04 ± 0.16	65.4 ± 8.3	191.5 ± 26.8	34.7	5.96 ± 0.28	65.3 ± 8.2	191.4 ± 26.8	34.3
9	GACA <sup>CU</sup> ACUG CUGU <sub>CU</sub> UGAC	6.40 ± 0.05	77.6 ± 6.0	229.6 ± 19.2	36.5	6.37 ± 0.04	74.0 ± 3.2	218.1 ± 10.3	36.3
9	GAG <sup>GA</sup> GAG <sup>(15)</sup> CUC <sub>AG</sub> CUC	7.00 ± 0.2	62.0 ± 5.0	177.4 ± 16.6	39.4	6.95 ± 0.01	73.0 ± 1.9	213.1 ± 6.2	38.8
9	GCU <sup>GA</sup> AGC <sup>(13)</sup> CGA <sub>AG</sub> UCG	5.97 ± 0.17	56.2 ± 6	162.0 ± 19	38.6	5.88 ± 0.05	64.9 ± 3	190.2 ± 11	38.0
7	GACA <sup>AA</sup> GCGC CUGU <sub>AG</sub> CGAC	8.71 ± 0.27	80.1 ± 7.2	230.2 ± 22.4	45.7	8.70 ± 0.06	80.2 ± 2.4	230.5 ± 7.4	45.6

<sup>a</sup>Measurements were made in 1.0 M NaCl, 10 mM sodium cacodylate and 0.5 mM Na<sub>2</sub>EDTA at pH 7.0.

<sup>b</sup>Frequency of occurrence in the database described in Materials and Methods.

<sup>c</sup>The top strand of each duplex is written 5'–3', and each bottom strand is written 3'–5'.

reported here, the number of tandem mismatches with A–U nearest neighbors has quadrupled in size (31 occurrences). A small sample size should not be a problem with any other parameter, as all parameters now have at least 16 occurrences contributing to the derivation of that parameter. The penalty per A–U nearest neighbor is the same as the penalty per G–U nearest neighbor, 1.0 kcal/mol. Having an identical penalty for an A–U and G–U pair adjacent to a tandem mismatch is consistent with previous findings for tandem mismatches (1,20,21) and for 1 × 2 nucleotide internal loops (41). In our previous model (19), we recently proposed penalties that were slightly different, 0.5 kcal/mol per A–U nearest neighbor and 1.2 kcal/mol per G–U nearest neighbor. As stated previously, this difference was likely due to the small sample size of tandem mismatches with A–U nearest neighbors.

## DISCUSSION

### Database searching

Due to the size and diversity of the RNA secondary structure database that was searched, we have assumed that the number and type of tandem mismatches found in this database are representative of tandem mismatches found in naturally occurring RNA.

It is clear from the first set of data in Table 1 that the tandem mismatch—nearest neighbor combinations that have been studied previously are not representative of

those found in nature. When looking at the second set of data in Table 1, however, it appears as if the most frequent tandem mismatches (considering the mismatch nucleotides only) have already been studied. In 1997, Xia *et al.* (15) reported the thermodynamics of asymmetric tandem mismatches adjacent to G–C base pairs. As a result, most of the tandem mismatches (considering the mismatch nucleotides only) have already been studied. However, it has been shown that the stability of tandem mismatches depends not only on the identity of the nucleotides in the loop but also on the identity of the closing base pairs (1,10–21). Therefore, this work (i) focuses on frequently occurring tandem mismatches when considering both the nucleotides in the loop and the nearest neighbors, (ii) compliments the work done in 1997 by Xia *et al.* (15), and (iii) provides additional, useful information about the thermodynamic stability of tandem mismatches.

We previously compiled the thermodynamics of all 60 sequence symmetric tandem mismatches (19). It is interesting to note that only 3 of the 30 most frequently occurring tandem mismatches (loop nucleotides and nearest neighbors) are sequence symmetric (Table 1).

The database search to compile the values in Dataset 2 of Table 1 revealed 55 types of tandem mismatches in the database; representing all possible types of tandem mismatches.

Comparisons can be made between the most common nearest neighbor combinations listed here (Dataset 3 of

**Table 3.** Contributions of 33 tandem mismatches to duplex stability<sup>a</sup>

Frequency <sup>b</sup>	Sequence <sup>c</sup>	$\Delta G_{37}^{\circ}$ (kcal/mol)	$\Delta H^{\circ}$ (kcal/mol)	$\Delta S^{\circ}$ (eu)	Frequency <sup>b</sup>	Sequence <sup>c</sup>	$\Delta G_{37}^{\circ}$ (kcal/mol)	$\Delta H^{\circ}$ (kcal/mol)	$\Delta S^{\circ}$ (eu)
57	G <sup>GA</sup> G C <sub>AG</sub> U	-0.10	-14.8	-47.4	13	C <sup>GA</sup> G G <sub>AG</sub> U	-0.84	-22.4	-69.4
43	G <sup>AA</sup> G <sup>(15)</sup> C <sub>AG</sub> C	-0.18	-16.0	-51.1	11	A <sup>GA</sup> G U <sub>AG</sub> U	0.67	-23.4	-77.5
41	C <sup>GA</sup> G <sup>(10)</sup> G <sub>AG</sub> C	-0.69	-8.9	-26.5	10	C <sup>AU</sup> C G <sub>GU</sub> G	1.12	-12.1	-42.5
	C <sup>GA</sup> G <sup>(10)</sup> G <sub>AG</sub> C	-1.74	-10.5	-28.3		C <sup>AU</sup> C <sup>(15)</sup> G <sub>GU</sub> G	0.90	-8.1	-28.9
37	A <sup>GA</sup> G U <sub>AG</sub> C	-0.23	-18.1	-57.7	10	A <sup>UU</sup> G U <sub>UU</sub> C	0.34	-27.8	-90.9
34	C <sup>AA</sup> G G <sub>AG</sub> C	-1.13	-29.7	-92.0	10	C <sup>UU</sup> C G <sub>UU</sub> G	-0.82	-29.9	-93.6
30	U <sup>UC</sup> U A <sub>UU</sub> A	1.42	-27.7	-94.0		C <sup>UU</sup> C <sup>(15)</sup> G <sub>UU</sub> G	-0.64	-23.5	-73.7
27	U <sup>AA</sup> G A <sub>GG</sub> C	2.19	0.4	-6.0	10	G <sup>UU</sup> A C <sub>UU</sub> U	0.46	-31.1	-101.7
21	C <sup>UU</sup> A G <sub>UU</sub> U	1.63	-2.5	-13.4	9	A <sup>AA</sup> G U <sub>AG</sub> U	2.19	-13.1	-49.4
19	C <sup>UU</sup> G <sup>(11)</sup> G <sub>UU</sub> C	-0.44	-17.5	-55.0	9	U <sup>AA</sup> C A <sub>GG</sub> G	-0.31	-25.4	-80.9
17	U <sup>AC</sup> A A <sub>GU</sub> U	2.54	-8.9	-36.7	9	A <sup>AU</sup> A U <sub>CU</sub> U	3.08	-2.3	-17.5
16	C <sup>AA</sup> G G <sub>CG</sub> U	0.52	-24.5	-80.6	9	A <sup>CU</sup> A U <sub>CU</sub> U	2.67	-11.0	-44.2
16	C <sup>GA</sup> A G <sub>AG</sub> U	2.13	38.4	117.1	9	G <sup>GA</sup> G <sup>(15)</sup> C <sub>AG</sub> C	-1.71	-30.7	-93.6
14	C <sup>AA</sup> A G <sub>AG</sub> U	0.45	-17.5	-57.7	9	U <sup>GA</sup> A <sup>(13)</sup> A <sub>AG</sub> U	0.71	-13.5	-45.6
14	G <sup>CA</sup> A C <sub>CG</sub> U	2.43	-6.2	-28.0	7	A <sup>AA</sup> G U <sub>AG</sub> C	1.52	-13.7	-49.2
13	G <sup>AA</sup> G C <sub>AG</sub> U	1.09	3.9	9.0					
13	U <sup>AA</sup> G A <sub>AG</sub> C	0.71	-20.1	-67.1					
13	A <sup>AC</sup> U U <sub>GA</sub> A	2.32	-22.1	-78.5					

Continued

<sup>a</sup>Measurements were made in 1.0 M NaCl, 10 mM sodium cacodylate and 0.5 mM Na<sub>2</sub>EDTA at pH 7.0.<sup>b</sup>Frequency of occurrence in the database described in materials and methods<sup>c</sup>The top strand of each duplex is written 5' – 3', and each bottom strand is written 3' – 5'.

**Table 4.** Model for predicting the free energy contribution of previously unmeasured sequence asymmetric tandem mismatches at 37°C

Tandem mismatches with <sup>a</sup>	$\Delta G_{37}^{\circ}$ increments proposed previously (19) (kcal/mol)	$\Delta G_{37}^{\circ}$ increments (kcal/mol)	$\Delta H^{\circ}$ increments (kcal/mol)	$\Delta S^{\circ}$ increments (eu)
a U-U pair adjacent to an R-R pair or a G-A or A-G pair adjacent to a Y-Y pair or any combination of A-C, U-C, C-U, C-C, C-A, or A-A pairs	1.1 ± 0.1	1.0 ± 0.1	-6.2 ± 2.2	-23.5 ± 6.9
any combination of adjacent G-A and A-G pairs or two U-U pairs	-1.2 ± 0.3	-0.7 ± 0.2	-13.6 ± 3.3	-41.3 ± 10.4
a U-U pair adjacent to a Y-Y (not U-U), C-A, or A-C pair	0.8 ± 0.2	0.6 ± 0.2	-4.3 ± 3.0	-14.3 ± 9.5
a G-G pair not adjacent to a U-U pair	-0.3 ± 0.2	0	0	0
per A-U nearest neighbor	0.5 ± 0.2	1.0 ± 0.1	-5.3 ± 2.2	-20.6 ± 6.9
per G-U nearest neighbor	1.2 ± 0.1	1.0 ± 0.1	-5.0 ± 2.6	-19.5 ± 8.1

<sup>a</sup>Any other base pair combinations in a tandem mismatch do not contribute to duplex stability.

Table 1) and those found most frequently adjacent to single mismatches (38) and 1 × 2 nucleotide loops (41). For tandem mismatches, a 5'G-C and 3'C-G combination was the most frequent (see the first entry of Dataset 3 in Table 1). This combination was the sixth most common combination for single mismatches (38) and the most common combination for 1 × 2 nucleotide loops (41). The most common combination for single mismatches was a 5'G-C and 3'C-G combination (38) (which is the 10th most common for tandem mismatches). It is unclear why single mismatches prefer different nearest neighbor combinations than 1 × 2 loops and tandem mismatches.

When categorizing the loop nucleotides as purines and pyrimidines (Dataset 4 of Table 1), it is interesting to note that the type of mismatch with the most possible combinations, 5'RY3'/3'RY5' (with 16 possible sequences), is only the third most frequent. The top two types, all purine and all pyrimidine loops, each only have 10 possible sequences. Therefore, there appears to be a preference for all purine and all pyrimidine tandem mismatches, with all purine being the most preferred. In fact, all purine and all pyrimidine tandem mismatches account for 47% and 24%, respectively, of all the mismatches found in the database. Similar results were observed for 1 × 2 nucleotide internal loops; all purine (56%) and all pyrimidine (15%) internal loops accounted for ~75% of all the 1 × 2 nucleotide internal loops in the database (41).

#### Thermodynamic contributions of tandem mismatches to duplex thermodynamics

From the data in Table 3, it is evident that the stability of a tandem mismatch does not determine its frequency of occurrence. For example, the second most stable tandem mismatch (5'GGAG3'/3'CAGC5') is the 28th most common. Similarly, a tandem mismatch that contributes an unfavorable 3.1 kcal/mol toward duplex stability (5'AAUA3'/3'UCUU5') appears in the top 30. Interestingly, all four tandem mismatches that have favorable free energy contributions (5'GA3'/3'AG5', 5'AA3'/

3'AG5', 5'AA3'/3'GG5' and 5'UU3'/3'UU5') also have unfavorable free energy contributions when situated between different nearest neighbor combinations.

It is interesting to note that there are three tandem mismatches (loop nucleotides plus adjacent nearest neighbors) that were studied in two different stem sequences (Table 3). Two of the tandem mismatches, 5'CAUC3'/3'GGUG5' and 5'CUUC3'/3'GUUG5', differ by only 0.2 kcal/mol when placed in the two different stems. However, 5'CGAG3'/3'GAGC5' differed by 1.1 kcal/mol when placed in the two different stems. This difference is likely due to non-nearest neighbor effects. Similar non-nearest neighbor effects have been observed previously (37,38,41-44). It has recently been reported that the accuracy of RNA secondary structure prediction by free energy minimization is limited by non-nearest neighbor effects (45). Since these effects may be complicated to interpret, non-nearest neighbor effects were ignored here, and data were treated as if stability relied only upon immediate nearest neighbors and the identity of the loop nucleotides. The role of non-nearest neighbor effects, however, is currently being investigated.

#### Updated model for predicting thermodynamics of tandem mismatches

Because we have collected thermodynamic data for 25 tandem mismatches that previously did not have experimental values, when predicting the free energy contributions of these mismatches in an RNA duplex, the experimental values can be used. These new experimental values, on average, are 1.0 kcal/mol different from the values predicted for these mismatches using the previous model.

In order to test the accuracy of this new model, the free energies of all 97 sequence asymmetric tandem mismatches compiled here were predicted using the updated model (although this model would never be used for these loops since they have experimental parameters available) (see Supplementary Table S1). For the sequence



asymmetric tandem mismatches compiled here, the average absolute difference between the measured values and the values predicted using these updated parameters is  $0.53 \pm 0.41$  kcal/mol. Stated differently, 57% of the experimental free energies were predicted within 0.5 kcal/mol, 86% were predicted within 1.0 kcal/mol and 96% were predicted within 1.5 kcal/mol. As suggested by these numbers, there are some idiosyncrasies associated with the free energy contributions of tandem mismatches that we still do not understand. Using the model proposed previously (19) and the same dataset of asymmetric tandem mismatches, the average absolute difference between the measured values and the values predicted is  $0.58 \pm 0.51$  kcal/mol. Overall, there is not a significant difference between the two models. However, the new model does improve the prediction for tandem mismatches that were predicted rather poorly by the previous model. For example, for eight tandem mismatches that were predicted poorly (greater than 1 kcal/mol difference between the predicted value and the experimental value) by the previous model, the average difference between experimental and predicted values was 1.7 kcal/mol. For the same eight loops using the updated model, this average difference decreases to 0.9 kcal/mol.

### Hydrogen bonding patterns in tandem mismatches

In order to investigate both the influence of structure on the stability of tandem mismatches and if structural features are important for determining a tandem mismatch's frequency of occurrence, the hydrogen bonding patterns between tandem mismatch nucleotides in previously solved three-dimensional structures were investigated. Structures of four tandem mismatches in the top 30 were downloaded from the PDB. One of the tandem mismatches was the most frequently occurring tandem mismatch in the secondary structure database, 5'GGAG3'/3'CAGU5'. This tandem mismatch occurred 57 times in the secondary structure database and had a slightly favorable free energy contribution of  $-0.10$  kcal/mol. The PDB structure (1J5E) of the *Thermus thermophilus* 30S ribosomal subunit from the Ramakrishnan laboratory (46) revealed that both G•A pairs are forming sugar-edge/Hoogsteen pairs. One of the G•A pairs contained one base–base hydrogen bond from the G amino proton to the A N7. The other G•A pair contained two base–base hydrogen bonds, one between an A amino proton and G N3 and one between a G amino proton and A N7 (see Supplementary Figure S1). With the extensive hydrogen bonding within the tandem mismatch, it is unclear why this tandem mismatch is only slightly stabilizing.

The next tandem mismatch investigated structurally was the third most frequently occurring tandem mismatch in the secondary structure database, 5'CGAG3'/3'GAGC5'. This tandem mismatch occurred 41 times in the secondary structure database and had a favorable free energy contribution of  $-0.69$  or  $-1.74$  kcal/mol, depending on the identity of the stem duplex. Interestingly, this tandem mismatch also contains two adjacent G•A pairs, similar to the tandem mismatch described above. Hydrogen bonding of this tandem mismatch was

investigated in three different PDB structures, the PDB structure (1NKW) of the *Deinococcus radiodurans* large ribosomal subunit from the Yonath laboratory (47), the PDB structure (1S72) of the *H. marismortui* large ribosomal subunit from the Moore and Steitz laboratories (48) and the PDB structure (1YFV) of a synthetic RNA from the Turner laboratory (49). The tandem mismatch in each of these structures had slightly different features. In PDB ID 1NKW, both G•A pairs are in a sugar-edge/Hoogsteen conformation; however, there are no base–base hydrogen bonds (see Supplementary Figure S2). In both PDB ID 1S72 and 1YFV, the GA pairs are also in a sugar-edge/Hoogsteen conformation; however, both pairs in both structures have two base–base hydrogen bonds, a G amino proton to A N7 and an A amino proton to G N3 (see Supplementary Figure S3). It is difficult to interpret the influence of the structure on either the thermodynamics or frequency of occurrence for this tandem mismatch due to the difference in stability when this tandem mismatch is placed within different stems and due to the different hydrogen bonding patterns found for this tandem mismatch in various PDB structures.

A third tandem mismatch with two G•A pairs was investigated, 5'UGAG3'/3'AAGC5'. This tandem mismatch occurred 16 times in the secondary structure database and had an unfavorable free energy contribution of 2.13 kcal/mol. Structural features of this tandem mismatch were investigated in the PDB structure (1NYI) of the hammerhead ribozyme from the Scott laboratory (50). Once again, both G•A pairs are in a sugar-edge/Hoogsteen conformation with both pairs having an A amino proton to G N3 hydrogen bond and a G amino proton to A N7 hydrogen bond (see Supplementary Figure S3). After investigating all five structures containing tandem G•A pairs, it is not straightforward to identify the structural features that lead to the varying thermodynamic stabilities.

Perhaps major contributors to the thermodynamic stabilities are the identities of the closing base pairs. The last tandem mismatch discussed above, 5'UGAG3'/3'AAGC5', has one U–A and one G–C adjacent base pairs and was the most unfavorable thermodynamically of the three. The second tandem mismatch discussed above, 5'CGAG3'/3'GAGC5', has one C–G and one G–C adjacent base pairs and was the most favorable thermodynamically. The first tandem mismatch discussed above, 5'GGAG3'/3'CAGU5' has one G–C and one G–U adjacent base pair and was more stable than the third and less stable than the second tandem mismatch.

Another tandem mismatch that was investigated was 5'CUUG3'/3'GUUC5'. This tandem mismatch occurred 19 times in the secondary structure database and had a favorable free energy contribution of 0.44 kcal/mol. Structural features of this tandem mismatch were investigated in the PDB structure (280D) of an RNA dodecamer from the Kundrot laboratory (51). Both U•U pairs are in a Watson–Crick/Watson–Crick conformation with both pairs having two U imino proton to U carbonyl oxygen hydrogen bonds (see Supplementary Figure S4).

The base pairing between the tandem mismatch nucleotides in the PDB structures discussed here shed little

light on the relationship between structure and thermodynamics and the reason why nature prefers particular tandem mismatch sequences over others. In addition to base–base hydrogen bonds, perhaps base–backbone hydrogen bonds, the hydrogen bonding pattern of the adjacent base pairs, the amount of stacking between the loop nucleotides and between the loop nucleotides and the nearest neighbors, sugar puckers, dynamics, helix distortion, and available functional groups situated in the grooves would provide more insight to help answer these questions. Perhaps a more extensive PDB search and comparison that investigates a much larger number of tandem mismatches and more structural features would help in understanding structure–stability relationships and nature’s preference for certain loop tandem mismatch sequences.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

The authors would like to thank Mr Matthew Alexander and Miss Mary (Molly) Burke for checking all of our thermodynamic calculations.

## FUNDING

Saint Louis University College of Arts and Sciences; Department of Chemistry, Saint Louis University; a Saint Louis University Summer Research Award (to B.M.Z.); the Saint Louis University Beaumont Faculty Development Fund (to B.M.Z.); and two Sigma Xi Grants-in-Aid of Research (to M.E.C.). Award Number R15GM085699 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Institute of General Medical Sciences of the National Institutes of Health. Funding for open access charge: National Institute of General Medical Sciences.

*Conflict of interest statement.* None declared.

## REFERENCES

- 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.
- Gutell,R.R. and Woese,C.R. (1990) Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. *Proc. Natl Acad. Sci. USA*, **87**, 663–667.
- Gutell,R.R. (1994) Collection of small-subunit (16s- and 16s-like) ribosomal-RNA structures—1994. *Nucleic Acids Res.*, **22**, 3502–3507.
- 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 Å resolution. *Science*, **289**, 905–920.
- Shi,P.-Y., Brinton,M. A., Veal,J.M., Zhong,Y.Y. and Wilson,W.D. (1996) Evidence for the existence of a pseudoknot structure at the 3' terminus of the flavivirus genomic RNA. *Biochemistry*, **35**, 4222–4230.
- Zhang,J.C., Zhang,G.H., Guo,R., Shapiro,B.A. and Simon,A.E. (2006) A pseudoknot in a preactive form of a viral RNA is part of a structural switch activating minus-strand synthesis. *J. Virol.*, **80**, 9181–9191.
- Sobczak,K., de Mezer,M., Michlewski,G., Krol,J. and Krzyzosiak,W.J. (2003) RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Res.*, **31**, 5469–5482.
- DeNap,J.C.B., Thomas,J.R., Musk,D.J. and Hergenrother,P.J. (2004) Combating drug-resistant bacteria: small molecule mimics of plasmid incompatibility as antiplasmid compounds. *J. Am. Chem. Soc.*, **126**, 15402–15404.
- Disney,M.D., Haidaris,C.G. and Turner,D.H. (2001) Recognition elements for 5' exon substrate binding to the *Candida albicans* group I intron. *Biochemistry*, **40**, 6507–6519.
- SantaLucia,J. Jr., Kierzek,R. and Turner,D.H. (1990) Effects of GA mismatches on the structure and thermodynamics of RNA internal loops. *Biochemistry*, **29**, 8813–8819.
- 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.
- 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.
- 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.
- Wu,M., McDowell,J. A. and Turner,D.H. (1995) A periodic table of symmetric tandem mismatches in RNA. *Biochemistry*, **34**, 3204–3211.
- Xia,T.B., McDowell,J.A. and Turner,D.H. (1997) Thermodynamics of nonsymmetric tandem mismatches adjacent to G–C base pairs in RNA. *Biochemistry*, **36**, 12486–12497.
- Burkard,M.E., Xia,T. and Turner,D.H. (2001) Thermodynamics of RNA internal loops with a guanosine–guanosine pair adjacent to another noncanonical pair. *Biochemistry*, **40**, 2478–2483.
- Schroeder,S. J. and Turner,D.H. (2001) Thermodynamic stabilities of internal loops with GU closing pairs in RNA. *Biochemistry*, **40**, 11509–11517.
- Bourdelat-Parks,B.N. and Wartell,R.M. (2005) Thermodynamics of RNA duplexes with tandem mismatches containing a uracil–uracil pair flanked by C–G/G–C or G–C/A–U closing base pairs. *Biochemistry*, **44**, 16710–16717.
- 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.
- 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. USA*, **101**, 7287–7292.
- 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.
- Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
- Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
- Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
- Gutell,R.R., Gray,M.W. and Schnare,M.N. (1993) A compilation of large subunit (23s-like and 23s-like) ribosomal-RNA structures—1993. *Nucleic Acids Res.*, **21**, 3055–3074.
- Schnare,M.N., Damberger,S.H., Gray,M.W. and Gutell,R.R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23S-like) ribosomal RNA. *J. Mol. Biol.*, **256**, 701–719.

27. Szymanski, M., Specht, T., Barciszewska, M.Z., Barciszewski, J. and Erdmann, V.A. (1998) 5S rRNA data bank. *Nucleic Acids Res.*, **26**, 156–159.
28. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. and Steinberg, S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.
29. Larsen, N., Samuelsson, T. and Zwieb, C. (1998) The signal recognition particle database (SRPDB). *Nucleic Acids Res.*, **26**, 177–178.
30. Brown, J.W. (1998) The ribonuclease P database. *Nucleic Acids Res.*, **26**, 351–352.
31. 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.
32. Damberger, S.H. and Gutell, R.R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res.*, **22**, 3508–3510.
33. Michel, F., Umesono, K. and Ozeki, H. (1989) Comparative and functional-anatomy of group-I catalytic introns—a review. *Gene*, **82**, 5–30.
34. Griffiths-Jones, S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
35. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
36. Griffiths-Jones, S., Saini, H. K., van Dongen, S. and Enright, A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
37. 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.
38. Davis, A. R. and Znosko, B.M. (2007) Thermodynamic characterization of single mismatches found in naturally occurring RNA. *Biochemistry*, **46**, 13425–13436.
39. 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.
40. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
41. 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.
42. Kierzek, R., Burkard, M.E. and Turner, D.H. (1999) Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, **38**, 14214–14223.
43. Siegfried, N.A., Metzger, S.L. and Bevilacqua, P.C. (2007) Folding cooperativity in RNA and DNA is dependent on position in the helix. *Biochemistry*, **46**, 172–181.
44. Longfellow, C.E., Kierzek, R. and Turner, D.H. (1990) Thermodynamic and spectroscopic study of bulge loops in oligoribonucleotides. *Biochemistry*, **29**, 278–285.
45. Mathews, D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, **359**, 526–532.
46. Wimberly, B.T., Brodersen, D.E., Clemons, W.M., Morgan-Warren, R.J., Carter, A.P., Vornrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327–339.
47. Harms, J.M., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F. and Yonath, A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679–688.
48. Klein, D.J., Moore, P.B. and Steitz, T.A. (2004) The roles of ribosomal proteins in the structure, assembly and evolution of the large ribosomal subunit. *J. Mol. Biol.*, **340**, 141–177.
49. SantaLucia, J. Jr. and Turner, D.H. (1993) Structure of (rGGCGAG CC)<sub>2</sub> in solution from NMR and restrained molecular dynamics. *Biochemistry*, **32**, 12612–12623.
50. Dunham, C.M., Murray, J.B. and Scott, W.G. (2003) A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J. Mol. Biol.*, **332**, 327–336.
51. Lietzke, S.E., Barnes, C.L., Berglund, J.A. and Kundrot, C.E. (1996) The structure of an RNA dodecamer shows how tandem U–U base pairs increase the range of stable RNA structures and the diversity of recognition sites. *Structure*, **4**, 917–930.