Published online 28 September 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 3* ***1081–1094***
doi:10.1093/nar/gkq793

Structural characterization of naturally occurring RNA single mismatches

Amber R. Davis, Charles C. Kirkpatrick and Brent M. Znosko*

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

Received May 27, 2010; Revised August 16, 2010; Accepted August 21, 2010

ABSTRACT

RNA is known to be involved in several cellular processes; however, it is only active when it is folded into its correct 3D conformation. The folding, bending and twisting of an RNA molecule is dependent upon the multitude of canonical and non-canonical secondary structure motifs. These motifs contribute to the structural complexity of RNA but also serve important integral biological functions, such as serving as recognition and binding sites for other biomolecules or small ligands. One of the most prevalent types of RNA secondary structure motifs are single mismatches, which occur when two canonical pairs are separated by a single non-canonical pair. To determine sequence–structure relationships and to identify structural patterns, we have systematically located, annotated and compared all available occurrences of the 30 most frequently occurring single mismatch-nearest neighbor sequence combinations found in experimentally determined 3D structures of RNA-containing molecules deposited into the Protein Data Bank. Hydrogen bonding, stacking and interaction of nucleotide edges for the mismatched and nearest neighbor base pairs are described and compared, allowing for the identification of several structural patterns. Such a database and comparison will allow researchers to gain insight into the structural features of unstudied sequences and to quickly look-up studied sequences.

INTRODUCTION

RNA is known to perform a variety of biological functions and to be involved in several cellular processes; however, it is only active when in its correct 3D conformation. The structural complexity and wide repertoire of structural components of RNA allows this biomolecule to effectively carry out a multitude of key functions. RNA consists of canonical double helical regions, along with non-canonical regions, such as internal loops, bulges, hairpins and multi-branch loops, which have implications for folding and stability of the correct tertiary and quaternary structures. Often times, these motifs are important for a variety of biological functions, such as serving as binding sites for proteins (1–10), metals (11–13), small molecules (14–19), or other nucleic acids (20). The scaffold of RNA tertiary structure is a result of the secondary structural components, which introduce kinks and turns in the RNA structure while providing available hydrogen bond donor and/or acceptor sites allowing for intermolecular interactions. Therefore, an understanding of the 3D conformation of RNA secondary structure motifs will give insight into RNA function.

An understanding of the structural propensities of common RNA secondary structure motifs should improve the prediction of RNA structure, function and recognition (21). Much work has been done to improve the prediction of RNA secondary structure from sequence (22–31), and methods are being developed to predict RNA tertiary structure (32–39). While the methods of NMR, crystallography and cryo-electron microscopy provide definitive tertiary structure information, they are not capable of keeping pace with the discovery of new and interesting RNA sequences. However, these tools have revealed a wide range of base pairing geometries commonly found in RNA (40,41). These different geometries have been shown to contribute to the complexity of RNA tertiary structure (42,43). Therefore, an understanding of these base–base conformations may allow for further understanding and accuracy in the prediction of RNA secondary and tertiary structure. One possible approach to begin developing a method to predict tertiary structure of RNA is to identify structural patterns for a given motif by structurally characterizing each occurrence of that motif in available 3D structures. Such structures have been deposited into the Protein Data Bank (PDB) (44–48), a world-wide archive of structural data of biomolecules,

*To whom correspondence should be addressed. Tel: +1 314 977 8567; Fax: +1 314 977 2521; Email: znoskob@slu.edu which includes all RNA structures solved by NMR, crystallography and cryo-electron microscopy. Currently, there are over 1600 structures containing RNA in the PDB (44–48) (accessed on 12 August 2009).

The structural characterization and comparison of all structures containing a particular secondary structure motif is not a trivial task; however, several laboratories have made significant contributions to analyzing RNA motifs found in the structures deposited in the PDB (44–48). The Fox laboratory has developed an internet-based, interactive database of non-canonical base pairs found in known RNA structures (NCIR). It contains over 2000 non-canonical base pairs with descriptions of the associated structural properties, such as sequence context, sugar pucker and glycosidic bond orientation (49,50). The Olson laboratory has also developed a user friendly internet-based database [the RNA base-pair structure (BPS) database] of canonical and non-canonical base pairs found in determined RNA structures. It contains over 91 000 bp and approximately 4000 higher-order base interactions. The database provides representative figures of the observed spatial patterns and the annotation of the structural and chemical features for each base pair (51). The Gutell laboratory has contributed a significant amount of data by investigating the occurrence and diversity of various motifs (52–54). The laboratories of Leontis and Westhof have provided a standardized method for the naming and classification of the various orientations of RNA base pairs to allow for unambiguous communication (55–62). The Brenner and Holbrook laboratories have developed the Structural Classification of RNA (SCOR) database, which provides details about the 3D structure, function, tertiary interactions and phylogentic relationships of RNA secondary structure motifs (63–65). The Major laboratory has developed computational tools which are compliant with the RNA ontology (66) and are incorporated into the computer program, *MC-Annotate*, which is capable of interpreting and labeling RNA base pairs and base stacking interactions of a given 3D structure (67–69). The Major laboratory has also developed the computer program *MC-Search*, which determines the locations of user-defined structural motifs in RNA (69–71). These efforts have advanced the understanding of the structural details of RNA and have provided tools to analyze RNA tertiary structure. However, with the exception of the recent structural characterization of hairpin triloops (69), no effort has been put forth to systematically locate, annotate and compare occurrences of a particular RNA secondary structure motif.

This work is focused on systematically locating, annotating and comparing the most frequently occurring RNA single mismatches in nature. Single mismatches are known to be the most frequently occurring secondary structure motif in ribosomal RNA (72) and often times serve integral structural and/or functional roles (73–83). Using the computer search algorithm *MC-Search*, single mismatches have been located in the deposited structures found in the PDB. The structural characteristics of each occurrence were then objectively annotated using *MC-Annotate*. The resulting data for each located and annotated single mismatch were exported into *Microsoft Excel* to allow for the extraction of the most frequently occurring single mismatch-nearest neighbor sequence combinations (84). Hydrogen bonding, stacking and interaction of nucleotide edges for the mismatched and nearest neighbor base pairs are described and compared, allowing for the identification of several structural patterns. Such a database and comparison will allow researchers to gain insight into the structural features of unstudied sequences and quickly look-up studied sequences. It is important to distinguish this work from previous databases, such as the NCIR and BPS databases. Both the NCIR and BPS databases contain structure information about non-canonical pairs in all secondary structure motifs. This work focuses on non-canonical pairs in single mismatches exclusively, allowing for the identification of structural patterns specific to isolated non-canonical pairs.

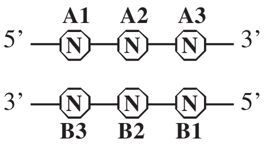
MATERIALS AND METHODS

Creation of a 3D RNA structure database

To create a database of previously solved RNA 3D structures, the PDB was searched for molecules containing RNA using the Molecule/Chain Type (since changed to Macromolecule Type) query in the Advanced Search menu on the PDB website (44–48) and selecting the molecules to contain RNA. All query results were selected and downloaded as uncompressed, .pdb formatted files. This search was conducted on 12 August 2009 and, therefore, includes all RNA-containing structures deposited into the PDB up to this date. The search was not limited by experimental method or resolution, but the resulting data is limited by the quality of the data deposited into the PDB.

Single mismatch database

The programs *MC-Search* (69–71) and *MC-Annotate* (67–69) were utilized to create the single mismatch database, and it is important to note they were not modified from the version provided by the authors. *MC-Search* (version 0.5) (69–71) was used to locate all single mismatches in the 3D structure database. In order to search 3D structures to locate a secondary structure motif, *MC-Search* requires an input descriptor (Figure 1). In simple terms, the input descriptor defines the size and type of the secondary structure motifs of interest. In order to define a single mismatch, 6 nt are involved, the 2 nt in the mismatch and the 2 nt in each of the 2 bp on either side of the mismatch. The type of interaction between the 2 nt in each pair was defined in the input descriptor, thereby limiting the nearest neighbor pairs to canonical pairs and the mismatch pair to a non-canonical pair. The pairing relations for the *MC-Search* input descriptor are defined by Roman (85–87) and Arabic (88,89) numerals, which indicate the presence of two or three hydrogen bonds and bifurcated or single hydrogen bonds, respectively. For example, Roman numeral XX (85–87) represents an A-U base pair with two hydrogen bonds (from A-N6 to U-NH3



```
sequence(RNA A1 NNN)
sequence(RNA B1 NNN)
relation(
  A1 B3 {XX or XXI or XXIII….88 or 83 or 89}
  A3 B1 {XX or XXI or XXIII….88 or 83 or 89}
  A2 B2 none or {! XX and !XXI and !XXIII….!88 and !83 and !89} )
```

Figure 1. Single mismatch graph (top) and *MC-Search* input descriptor (bottom). The nucleotides are numbered A1 to A3 and B1 to B3 in the 5′ to 3′ direction. The 'A' and 'B' letter designations specify opposing RNA strands. The letter 'N' represents any nucleotide. The input descriptor identifies the canonical nearest neighbors by limiting the allowed pairing interactions to the canonical pairs defined by the Roman (85–87) and Arabic (88,89) numerals. Not all possible numerals for A–U, U–A, G–C, C–G, G–U and U–G pairs are shown here due to space limitations. The input descriptor identifies the mismatched nucleotides by allowing an interaction defined by no hydrogen bonds, while also prohibiting the canonical pairing interactions defined by the Roman and Arabic numerals.

and A-NH6 to U-O4) between the Watson–Crick face of each base with a *cis*-glycosidic bond orientation. Other Roman numerals represent other pairs in a similar fashion (85–87). Arabic numeral 51 (88,89) represents an A-U base pair with one hydrogen bond (A-NH6 to U-O4) between the Watson–Crick face of each base with a *trans*-glycosidic bond orientation. Other Arabic numerals represent other pairs in a similar fashion (88,89).

For the nearest neighbor pair, any pair described by the Roman or Arabic numeral naming system of base pairs was allowed, thereby allowing most conformations of G-C, C-G, A-U, U-A, G-U and U-G pairs. Conversely, the mismatch nucleotides were defined as any pair not described by the Roman and Arabic numeral naming system of base pairs, thereby disallowing the pairs previously listed. Once the input descriptor contained this information, *MC-Search* was able to locate all of the single mismatches in the three dimensional RNA structural database. For each single mismatch located in this manner, the nucleotides involved in the single mismatch-nearest neighbor sequence combination were 'clipped' (i.e. all nucleotides not involved in the single mismatch or nearest neighbor were removed) and saved as a .pdb file to allow for quick annotation and a simple 3D graphic to be produced.

Once the results from the *MC-Search* and *MC-Annotate* scripts were tabulated, the results were searched for false-positives. A false-positive results, for example, when *MC-Annotate* does not annotate a G-C pair with a Roman or Arabic numeral. As a result, this G-C pair is considered a single mismatch. All G-C, C-G, A-U, U-A, G-U and U-G identified by the scripts as single mismatches were considered false positives and were removed from the database of true single mismatches.

Single mismatch annotation

The located single mismatches were structurally characterized by the program *MC-Annotate* (version 1.6.2) (67–69), which analyzes the atomic coordinates to determine the nucleotide interactions and classifies the type of base pairing. *MC-Annotate* utilizes four characterization parameters which include: (i) residue conformation, (ii) adjacent stackings, (iii) non-adjacent stackings and (iv) base-pairs. The residue conformation defines the sugar pucker as *endo* or *exo* and the glycosidic bond orientation as *syn* or *anti*. The adjacent and non-adjacent stackings define the relative orientation of each base, which are identified by *MC-Annotate* utilizing the method proposed by Gabb *et al.* (90). The nomenclature used to describe these orientations was proposed by Major and Thibault (91), which includes four base-stacking types: upward, downward, outward and inward. The nomenclature incorporated to illustrate the base pairing annotations is based on the Leontis and Westhof (56,57) classification scheme, which describes the interacting edges [i.e. the Watson–Crick (W), Hoogsteen (H) and Sugar (S) edges] of the two bases. This scheme has been further defined and described previously by Lemieux and Major (68). The resulting data for each located and annotated single mismatch were exported into *Microsoft Excel*.

Analysis of data and identification of structural patterns

Due to the excessive amount of data generated from the search and annotation (4899 single mismatches identified), the analysis of the data and the identification of structural patterns focused on the 30 most frequently occurring single mismatches in nature (84). To allow for the extraction of the most frequently occurring single mismatch-nearest neighbor sequence combinations (84) and further allow for the identification of structural patterns, the Leontis and Westhof (56,57) naming scheme was utilized when determining *general* structural trends and patterns because annotation is subject to interpretation and small geometrical variations (32), which could arise due to experimental conditions.

It is important to note some single mismatches have been excluded from the following analysis. In order to prevent over-counting and to simplify the analysis, ensembles of structures determined by NMR were excluded from the analysis. PDB structures consisting of a single averaged NMR structure, however, were included. Several clipped PDB files were not included in the analysis for various reasons (i.e. 13 single mismatch containing PDB files were not in the correct .pdb format, which prevented nucleotide annotation by *MC-Annotate*). These PDB files are denoted in Supplementary Table S1. Lastly, it is important to note the structural trends and patterns may be skewed due to repetitive representation of a molecule in the PDB. For example, the crystal structure of the large ribosomal subunit of *Haloarcula marismortui* has been solved unbound (PDB I.D. 1ffk) and bound (PDB I.D. 1n8r) to antibiotics.

RESULTS

3D RNA structure database

The PDB (44–48) search returned 1666 RNA-containing structures which were then used to create the 3D RNA structure database. A complete listing of the obtained structures can be found in the Supplementary Data (Supplementary Table S1).

Single mismatch structural database

Incorporation of a single mismatch-specific input descriptor into the *MC-Search* (69–71) program followed by a search of the structures contained in the 3D RNA structure database returned an extremely large dataset. Each of these 4899 identified single mismatches were structurally characterized using *MC-Annotate*. Of the 30 most frequently occurring single mismatches in a secondary structure database (84), 21 were located in the 3D structure database (Table 1 and Supplementary Table S2) and are the focus of the rest of this study. The nine frequently occurring single mismatch-nearest neighbor sequences (84) not found in the structural database were: $\begin{bmatrix} {}^{5'}\mathrm{A\underline{U}C}^{3'} \\ {}^{3'}\mathrm{U\underline{U}G}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{A\underline{U}A}^{3'} \\ {}^{3'}\mathrm{U\underline{U}U}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{A\underline{A}G}^{3'} \\ {}^{3'}\mathrm{U\underline{C}C}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{G\underline{U}G}^{3'} \\ {}^{3'}\mathrm{C\underline{U}U}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}G}^{3'} \\ {}^{3'}\mathrm{G\underline{G}C}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{C}U}^{3'} \\ {}^{3'}\mathrm{A\underline{U}A}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}A}^{3'} \\ {}^{3'}\mathrm{A\underline{A}U}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{A\underline{A}A}^{3'} \\ {}^{3'}\mathrm{U\underline{C}U}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{A\underline{C}U}^{3'} \\ {}^{3'}\mathrm{U\underline{U}A}^{5'} \end{bmatrix}$, with frequencies of 94, 62, 54, 43, 38, 38, 34, 34 and 34, respectively (84). For each of the remaining single mismatch-nearest neighbor combinations found in the top 30 (84), a wide variance in the number of times they were found in the structural database resulted (Table 1). Single mismatches were found in a wide repertoire of RNAs, including ribosomal RNAs (free and bound to antibiotics and proteins), riboswitches, tRNAs and viral RNAs.

Due to the immense amount of data collected, a table summarizing the common structural characteristics for each single mismatch-nearest neighbor sequence combination in the top 30 (84) is provided in Table 1 and Supplementary Table S2. To determine structural classes, or specimens (69), among each sequence combination, four parameters were considered: interacting edges for both the single mismatch nucleotides and the nearest neighbor base pairs and hydrogen bond patterns for both the single mismatch nucleotides and the nearest neighbor base pairs. Interactions involving a mismatched nucleotide and a nearest neighbor nucleotide were only considered when occurring in >5% of the total population for each single mismatch-nearest neighbor sequence combination.

DISCUSSION

A·G single mismatches

A·G single mismatches are the most frequently occurring single mismatch type found in the secondary structure database (84) when categorized by only the mismatched nucleotides. There are 10 A·G mismatch-nearest neighbor sequence combinations found in the 30 most frequently occurring single mismatches (84), and nine are represented in the RNA single mismatch structural database (Table 1 and Supplementary Table S2), with a total of 1462 occurrences. These nine can be divided into three groups based upon the geometric configuration of the mismatch nucleotides. The first group consists of the most common geometric orientation of the mismatched nucleotides, $^{5'}$(A)H/$^{3'}$(G)S pairing, antiparallel, *trans* glycosidic bond conformation, with 83% of the total occurrences found with these characteristics (Figure 2). $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{A\underline{G}G}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}G}^{3'} \\ {}^{3'}\mathrm{A\underline{G}C}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}U}^{3'} \\ {}^{3'}\mathrm{A\underline{G}A}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}A}^{3'} \\ {}^{3'}\mathrm{A\underline{G}U}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{A\underline{A}C}^{3'} \\ {}^{3'}\mathrm{U\underline{G}G}^{5'} \end{bmatrix}$ are the five sequence combinations with these geometric features, and, interestingly, they each contain a U-A or A-U base pair on the 5′ side of the A·G mismatch. Considering these five single mismatch-nearest neighbor sequence combinations, the most common base-pair orientation and hydrogen bonding pattern of the 5′ and 3′ nearest neighbors are $^{5'}$(U)W/$^{3'}$(A)H pairing, antiparallel, *trans* XXIV and $^{5'}$W/$^{3'}$W pairing, antiparallel, *cis* XIX, respectively. Although the orientation of the 5′ nearest neighbors are reversed for $\begin{bmatrix} {}^{5'}\mathrm{A\underline{A}C}^{3'} \\ {}^{3'}\mathrm{U\underline{G}G}^{5'} \end{bmatrix}$ (A–U instead of U–A), the A–U pair still exhibits a $^{5'}$(U)W/$^{3'}$(A)H pair. It is interesting to note the 5′ A–U or U–A nearest neighbor does not have the expected $^{5'}$W/$^{3'}$W pairing. Perhaps this is due to the structural perturbation resulting from the accommodation of the A·G mismatch, a purine–purine mismatch. The helical geometry may be disrupted to accommodate this type of noncanonical base pair. However, it is unclear why the 3′ nearest neighbor is not similarly disrupted.

The second group of A·G mismatches consist of mismatch nucleotides with $^{5'}$(A)W/$^{3'}$(G)W pairing, antiparallel, *cis* orientation forming two hydrogen bonds in the VIII pattern. $\begin{bmatrix} {}^{5'}\mathrm{C\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{G}G}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{G}G}^{5'} \end{bmatrix}$ are the two sequence combinations with these geometric features. They have similar nearest neighbors, with $^{5'}$Y/$^{3'}$G (where Y is a pyrimidine) and $^{5'}$C/$^{3'}$G on the 5′ and 3′ side of the A·G single mismatch, respectively. The 5′ and 3′ nearest neighbors are both characterized as $^{5'}$W/$^{3'}$W pairing, antiparallel, *cis* XIX.

The third group of A·G mismatches consists of mismatch nucleotides which are annotated not to form any interactions with each other. $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}C}^{3'} \\ {}^{3'}\mathrm{C\underline{G}G}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{A\underline{A}U}^{3'} \\ {}^{3'}\mathrm{U\underline{G}G}^{5'} \end{bmatrix}$ are the two sequence combinations with these geometric features. No interactions are found

Table 1. Summary of the structural orientation and interaction of the 30 frequently occurring single mismatches[a]

Single mismatch sequence[b]	Relative natural frequency[c]	PDB occurrences[d]	Number of similar PDB occurrences[e]	Structural orientation and hydrogen bonding pattern[f]		
				Single mismatch	5′ nearest neighbors	3′ nearest neighbors
AG						
UAC AGG	157	356				
			163	A-G Hh/Ss Bh/O2′ pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			90	"	U-A Ww/Hh pairing antiparallel *trans* XXIV	"
			49	A-G Hh/Ss pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			19	"	U-A Ww/Hh pairing antiparallel *trans* XXIV	"
UAG AGC	97	511				
			211	A-G Hh/Ss Bh/O2′ pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	G-C Ww/Ww pairing antiparallel *cis* XIX
			178	"	U-A Ww/Hh pairing antiparallel *trans* XXIV	"
			26	A-G Hh/Ss pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	G-C Ww/Ww pairing antiparallel *cis* XIX
			25	A-G O2′/Bs Ww/O2′ pairing	U-A Ws/Hh pairing antiparallel *trans* XXIV	G-C Ww/Ww pairing antiparallel *cis* XIX
AAU UGG	89	23				
			16	No interaction	A-U Hh/Ws pairing antiparallel *trans* XXIV	U-G Ws/Ww pairing antiparallel *cis* XXVIII
			7	"	A-U Hh/Ww pairing antiparallel *trans* XXIV	"
CAC GGG	53	17				
			17	A-G Ww/Ww pairing antiparallel *cis* VIII	C-G Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
UAU AGA	53	53				
			27	A-G Hh/Ss Bh/O2′ pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	U-A Ww/Ww pairing antiparallel *cis* XX
			23	"	U-A Ww/Hh pairing antiparallel *trans* XXIV	"
UAC GGG	40	79				
			53	A-G Ww/Ww pairing antiparallel *cis* VIII	U-G Wh/Ws pairing antiparallel *cis* one_hbond 94	C-G Ww/Ww pairing antiparallel *cis* XIX
			23	"	U-G Ws/Ww pairing antiparallel *cis* XXVIII	"
GAC CGG	35	4				
			3	No Interaction	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
			1	"	G-C Ww/Ws pairing antiparallel *cis* one_hbond 130	"
UAA AGU	35	216				
			81	A-G Hh/Ss Bh/O2′ pairing antiparallel *trans* XI	U-A Ww/Hh pairing antiparallel *trans* XXIV	A-U Ww/Ww pairing antiparallel *cis* XX
			61	"	U-A Ws/Hh pairing antiparallel *trans* XXIV	"
			35	A-G Hh/Ss pairing antiparallel *trans* XI	U-A Ws/Hh pairing antiparallel *trans* XXIV	A-U Ww/Ww pairing antiparallel *cis* XX
			17	"	U-A Ww/Hh pairing antiparallel *trans* XXIV	"
AAC UGG	34	203				
			62	A-G Hh/Ss Bh/O2′ pairing antiparallel *trans* XI	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			10	"	A-U Hh/Ww pairing antiparallel *trans* XXIV	"
			58	A-G Hh/Ss pairing antiparallel *trans* XI	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			10	"	A-U Hh/Ww pairing antiparallel *trans* XXIV	"
			32	No interaction	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			25	"	A-U Hh/Ww pairing antiparallel *trans* XXIV	"

(continued)

Table 1. Continued

Single mismatch sequence[b]	Relative natural frequency[c]	PDB occurrences[d]	Number of similar PDB occurrences[e]	Structural orientation and hydrogen bonding pattern[f]		
				Single mismatch	5′ nearest neighbors	3′ nearest neighbors
UU						
GUC CUG	183	231				
			118	U-U Ws/Ww pairing antiparallel *cis* XVI	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
			70	U-U Wh/Ww pairing antiparallel *cis* one_hbond	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
			36	No Interaction	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
CUG GUU	104	92				
			43 36	U-U Ws/Ww pairing antiparallel *cis* XVI "	C-G Ww/Ww pairing antiparallel *cis* XIX "	G-U Ww/Ws pairing antiparallel *cis* XXVIII G-U Ww/Bs Bs/O2′ pairing antiparallel *cis* 84
			5	U-U Ws/Ww pairing antiparallel *cis* one_hbond	C-G Ww/Ww pairing antiparallel *cis* XIX	G-U Ww/Ws pairing antiparallel *cis* XXVIII
CUC GUG	41	50				
			49	U-U Ww/Ws pairing antiparallel *cis* XVI	C-G Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
AUG UUC	36	30				
			15	No interaction	A-U Ww/Ww pairing antiparallel *cis* XX	G-C Ww/Ww pairing antiparallel *cis* XIX
			8	U-U Ws/Ww pairing antiparallel *cis* XVI	A-U Ww/Ww pairing antiparallel *cis* XX	G-C Ww/Ww pairing antiparallel *cis* XIX
			3	U-U Wh/Ww pairing antiparallel *cis* one_hbond	A-U Ww/Ww pairing antiparallel *cis* XX	G-C Ww/Ww pairing antiparallel *cis* XIX
			2	U-U Ww/Hw pairing parallel *cis* one_hbond 82	A-U Ww/Ww pairing antiparallel *cis* XX	G-C Ww/Ww pairing antiparallel *cis* XIX
AC						
AAC UCG	69	73				
			22	A-C Hh/Ww pairing antiparallel *trans* XXV	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			16	A-C Hh/Wh pairing antiparallel *trans* one_hbond	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
			10 5	No interaction "	A-U Hh/Ws pairing antiparallel *trans* XXIV A-U Hw/Ss Bh/O2′ pairing antiparallel *trans* one_hbond 45	C-G Ww/Ww pairing antiparallel *cis* XIX C-G Ww/Ww pairing antiparallel *cis* XIX
			8	A-C Hh/Ww pairing antiparallel *trans* one_hbond	A-U Hh/Ws pairing antiparallel *trans* XXIV	C-G Ww/Ww pairing antiparallel *cis* XIX
CAG GCC	60	9				
			9	A-C Wh/Ww pairing antiparallel *cis* one_hbond 75	C-G Ww/Ww pairing antiparallel *cis* XIX	G-C Ww/Ww pairing antiparallel *cis* XIX
CAC GCG	47	1				
			1	A-C Wh/Ww pairing antiparallel *cis* one_hbond 75	C-G Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
GAU CCA	45	2				
			1 1	A-C Ww/Hw pairing antiparallel *cis* one_hbond "	G-C Ww/Ww pairing antiparallel *cis* XIX "	U-A Ws/Bh pairing antiparallel *trans* one_hbond 46 U-A Ws/Wh pairing antiparallel *trans* one_hbond 46
GAG CCC	42	1				
			1	No interaction	G-C Ww/Ww pairing antiparallel *cis* XIX	G-C Ww/Ww pairing antiparallel *cis* XIX
GAC CCG	36	3				
			3	No interaction	Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX

(continued)

Table 1. Continued

Single mismatch sequence[b]	Relative natural frequency[c]	PDB occurrences[d]	Number of similar PDB occurrences[e]	Structural orientation and hydrogen bonding pattern[f]		
				Single mismatch	5′ nearest neighbors	3′ nearest neighbors
CU						
GCC CUG	48	76				
			38	C-U Wh/Wh pairing antiparallel *cis* one_hbond	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
			35	No interaction	G-C Ww/Ww pairing antiparallel *cis* XIX	C-G Ww/Ww pairing antiparallel *cis* XIX
GG						
AGG UGC	50	24				
			10 1	No interaction "	A-U Hh/Ws pairing antiparallel *trans* XXIV A-U Ww/Ww pairing antiparallel *cis* XX	G-C Ww/Ww pairing antiparallel *cis* XIX G-C Sw/Ww pairing antiparallel *cis* one_hbond 125
			8	G-G Hh/Bs pairing antiparallel *trans* 34	A-U Hh/Ws pairing antiparallel *trans* XXIV	G-C Ww/Ww pairing antiparallel *cis* XIX
			5	G-G Hh/Bs pairing antiparallel *trans* one_hbond 112	A-U Hh/Ws pairing antiparallel *trans* XXIV	G-C Ww/Ww pairing antiparallel *cis* XIX

[a] All possible orientations and hydrogen bonding patterns are not shown for each single mismatch-nearest neighbor combination. Only those representing at least 5% of total occurrences are included.

[b] For each sequence, the top strand is written 5′–3′, and the bottom strand is written 3′–5′. Duplexes are written in alphabetical order by the loop nucleotide (A over G, not G over A). If the loop nucleotides are identical, then duplexes are written in alphabetical order by the nearest neighbors (CUG over GUU, not GUU over CUG).

[c] Frequency of occurrence in the database (84).

[d] Number of times each single mismatch-nearest neighbor sequence combination was located in the three dimensional RNA structure database compiled from structures deposited into the PDB.

[e] Number of occurrences in each subclass, which is determined among each sequence combination, considering four parameters: interacting edges for the single mismatch nucleotides and the nearest neighbor base pairs and hydrogen bond patterns for the single mismatch nucleotides and the nearest neighbor base pairs.

[f] Annotated orientations and hydrogen bonding patterns of the single mismatch and 5′- and 3′-nearest neighbor nucleotides, which is described in 'Materials and Methods' section.

between the A·G mismatch nucleotides in $\begin{bmatrix} {}^{5'}G\underline{A}C^{3'} \\ {}^{3'}C\underline{G}G^{5'} \end{bmatrix}$ because the A is flipped out from the center of the helix and is interacting with the surrounding solvent. The nucleotides of the base pairing nearest neighbors for $\begin{bmatrix} {}^{5'}G\underline{A}C^{3'} \\ {}^{3'}C\underline{G}G^{5'} \end{bmatrix}$ were most commonly annotated to both be in the ${}^{5'}$W/${}^{3'}$W pairing, antiparallel, *cis* orientation forming three hydrogen bonds in the XIX pattern (one of the four examples was annotated to form only one hydrogen bond in the 130 base-pairing pattern). Although $\begin{bmatrix} {}^{5'}G\underline{A}C^{3'} \\ {}^{3'}C\underline{G}G^{5'} \end{bmatrix}$ contains similar nearest neighbor sequence combinations and geometries as $\begin{bmatrix} {}^{5'}C\underline{A}C^{3'} \\ {}^{3'}G\underline{G}G^{5'} \end{bmatrix}$ (discussed above in the second group), the geometry of the single mismatch is different. $\begin{bmatrix} {}^{5'}A\underline{A}U^{3'} \\ {}^{3'}U\underline{G}G^{5'} \end{bmatrix}$ also is annotated not to have any interactions between the mismatched nucleotides; however, the geometries of the 5′ and 3′ nearest neighbors are the same as those in the first group discussed above, ${}^{5'}$(U)W/${}^{3'}$(A)H pairing, antiparallel, *trans* XXIV and ${}^{5'}$(U)W/${}^{3'}$(G)W pairing, antiparallel, *cis* XIX, respectively.

Inter- and intra-strand interactions involving a mismatched nucleotide and a nearest neighbor nucleotide were found to occur prevalently in eight of the nine A·G mismatch-nearest neighbor sequence combinations (data not shown). The sequence without these types of interactions is $\begin{bmatrix} {}^{5'}C\underline{A}C^{3'} \\ {}^{3'}G\underline{G}G^{5'} \end{bmatrix}$, and it is unclear why this A·G mismatch does not engage in these types of interactions. Characterizing the single mismatch-nearest neighbor sequences as $\begin{bmatrix} {}^{5'}A\underline{B}C^{3'} \\ {}^{3'}F\underline{E}D^{5'} \end{bmatrix}$, all eight involved an inter-strand interaction between nucleotides A and E. The sequence combinations of $\begin{bmatrix} {}^{5'}U\underline{A}C^{3'} \\ {}^{3'}A\underline{G}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}A\underline{A}C^{3'} \\ {}^{3'}U\underline{G}G^{5'} \end{bmatrix}$ also formed an intra-strand interaction between nucleotides B and C through the O2P/Bh (i.e. one of the free oxygen atoms at the phosphorous between nucleotides B and C is the hydrogen bond acceptor which forms a bifurcated hydrogen bond with the two amino hydrogen atoms found on the Hoogsteen edge of the C) adjacent pairing with upward stacking. It is interesting to note, these two sequences only differ by the orientation of their 5′ nearest neighbor. The sequences $\begin{bmatrix} {}^{5'}A\underline{A}U^{3'} \\ {}^{3'}U\underline{G}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}A\underline{A}C^{3'} \\ {}^{3'}U\underline{G}G^{5'} \end{bmatrix}$ formed an intra-strand interaction between nucleotides F and E, and $\begin{bmatrix} {}^{5'}A\underline{A}U^{3'} \\ {}^{3'}U\underline{G}G^{5'} \end{bmatrix}$ has an additional intra-strand interaction between nucleotides E and D. These types of interactions may contribute to single mismatch stability and are, therefore, important to understand and further study their effects.

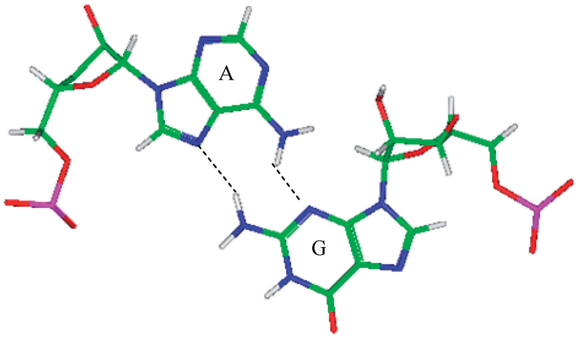


Figure 2. Representation of an A·G mismatch in the $^{5'}$(A)H/$^{3'}$(G)S pairing, antiparallel, *trans* orientation with XI hydrogen bonding pattern (PDB ID 1C04), which is the most common orientation and interaction determined for the most frequently occurring A·G mismatch-nearest neighbor combinations (84) that were also represented in the PDB.

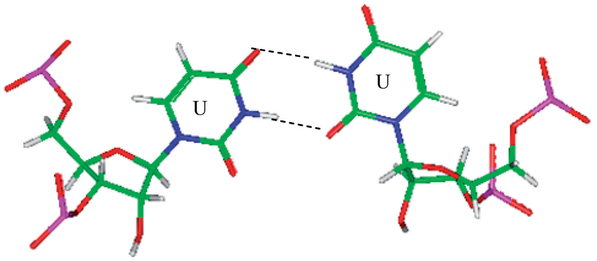


Figure 3. Representation of a U·U mismatch in the $^{5'}$(U)W/$^{3'}$(U)W pairing, antiparallel, *cis* orientation with XVI hydrogen bonding pattern (PDB ID 1FJG), which is the most common orientation and interaction determined for the most frequently occurring U·U mismatch-nearest neighbor combinations (84) that were also represented in the PDB.

An interesting structural and thermodynamic comparison is found for the two mismatch-nearest neighbor sequence combinations of $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{A\underline{G}G}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{G}G}^{5'} \end{bmatrix}$, which only differ by the identity of the 5′ nearest-neighbor, U-A versus U-G, respectively; however, they have experimental free energy values of −0.6 and 1.2 kcal/mol (84). There are 356 examples of $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{A\underline{G}G}^{5'} \end{bmatrix}$ found in the structural database, and the 5′ nearest neighbor, A·G mismatch and the 3′ nearest neighbor nucleotides are annotated to have the following characteristics in 90% of these occurrences: $^{5'}$(U)W/$^{3'}$(A)H pairing antiparallel *trans* XXIV (two hydrogen bonds), $^{5'}$(A)H/$^{3'}$(G)S pairing antiparallel *trans* XI (two hydrogen bonds) and $^{5'}$(C)W/$^{3'}$(G)W pairing antiparallel *cis* XIX (three hydrogen bonds), respectively. Additionally, this mismatch-nearest neighbor sequence generally forms intra- and inter-strand interactions, which are described above. There are 79 examples of $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{G}G}^{5'} \end{bmatrix}$ found in the structural database, and the 5′ nearest neighbor, A·G mismatch and the 3′ nearest neighbor nucleotides are annotated to have the following characteristics in 67% of these occurrences: $^{5'}$(U)W/$^{3'}$(G)W pairing antiparallel *cis* one_hbond (one hydrogen bond), $^{5'}$(A)W/$^{3'}$(G)W pairing antiparallel *cis* VII (two hydrogen bonds), and $^{5'}$(C)W/$^{3'}$(G)W pairing antiparallel *cis* XIX (three hydrogen bonds), respectively. It is important to note another 29% of the occurrences of $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{G}G}^{5'} \end{bmatrix}$ have similar structural characteristics and only differ by the hydrogen bonding pattern of the 5′ nearest neighbor, which is annotated to be XXVIII (two hydrogen bonds). However, this mismatch-nearest neighbor sequence is not annotated to engage in intra- and inter-strand interactions. Comparing the structural and interaction differences between these two mismatch-nearest neighbor sequences to the difference in free energy contribution of the respective single mismatches to duplex stability, it is unclear what the major contributing factor is that is resulting in such a large difference in thermodynamic stability. However, the additional stability of $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}C}^{3'} \\ {}^{3'}\mathrm{A\underline{G}G}^{5'} \end{bmatrix}$ may partially be a result of the additional intra- and inter-strand hydrogen bonding.

U·U single mismatches

There are seven U·U RNA single mismatch-nearest neighbor combinations found in the top 30 naturally occurring single mismatches (84), and four of these combinations, which include $\begin{bmatrix} {}^{5'}\mathrm{G\underline{U}C}^{3'} \\ {}^{3'}\mathrm{C\underline{U}G}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{C\underline{U}G}^{3'} \\ {}^{3'}\mathrm{G\underline{U}U}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{C\underline{U}C}^{3'} \\ {}^{3'}\mathrm{G\underline{U}G}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{A\underline{U}G}^{3'} \\ {}^{3'}\mathrm{U\underline{U}C}^{5'} \end{bmatrix}$, are represented in the RNA single mismatch structural database with a total of 403 occurrences (Table 1). Comparing these sequence combinations, the most common orientation of mismatch and nearest neighbor nucleotides for each are similar. Most commonly, the U·U mismatch nucleotides adopt the $^{5'}$W/$^{3'}$W pairing, antiparallel, *cis* conformation in 344 (85%) of the occurrences. When the U·U mismatches are found in this orientation, XVI and one_hbond (note this hydrogen bonding pattern has not been defined by an Arabic numeral in the literature) are the two hydrogen bonding patterns observed for 257 (75%) (Figure 3) and 87 (25%) of these occurrences, respectively. Also, when only considering this U·U conformation, 343 (∼100%) and 302 (88%) of the 5′ and 3′ nearest neighbor base pairs, respectively, are interacting in the $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX orientation. Interestingly, the 5′ nearest neighbors vary in sequence identity, including G-C, C-G and A-U, but they are all observed with the same type of orientation and interaction. The 3′ nearest neighbors also vary in sequence identity, including G-C, C-G and G-U; however,

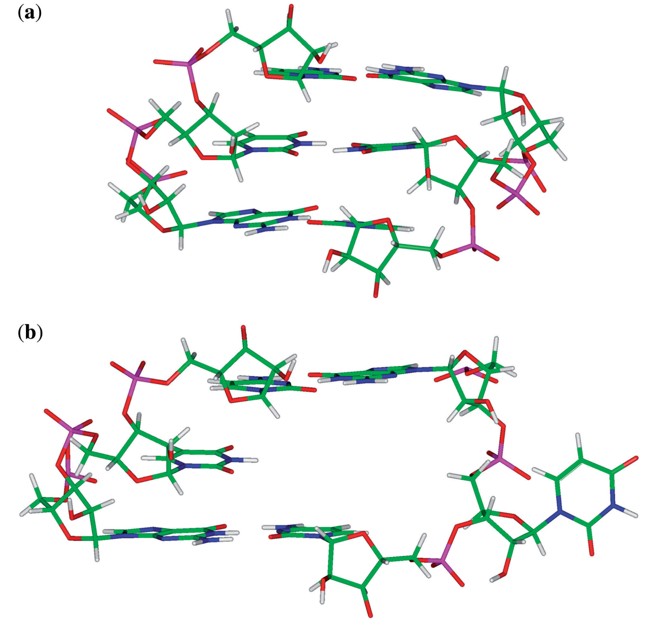


Figure 4. Representation of $\begin{bmatrix} {}^{5'}G\underline{U}C^{3'} \\ {}^{3'}C\underline{U}G^{5'} \end{bmatrix}$in the hydrogen bonded, stacked orientation (PDB ID 1O9M) (**a**) and in the non-hydrogen bonded, unstacked orientation (PDB ID 1O9M) (**b**).

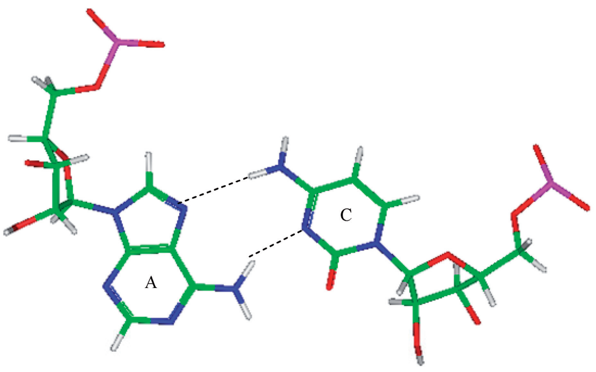


Figure 5. Representation of an A·C mismatch in the ${}^{5'}$(A)H/${}^{3'}$(C)W pairing, antiparallel, *trans* orientation with XXV hydrogen bonding pattern (PDB ID 1FJG)), which is the most common orientation and interaction determined for the A·C mismatch-nearest neighbor combination of $\begin{bmatrix} {}^{5'}A\underline{A}C^{3'} \\ {}^{3'}U\underline{C}G^{5'} \end{bmatrix}$. This mismatch-nearest neighbor sequence combination is found in the 30 most frequently occurring single mismatches (84) and accounts for 80% of the total A·C mismatches found in this study.

the 3′ nearest neighbor of the sequence combination $\begin{bmatrix} {}^{5'}C\underline{U}G^{3'} \\ {}^{3'}G\underline{U}U^{5'} \end{bmatrix}$ is observed to always have the same orientation but with the two different hydrogen bonding patterns of XIX (forming three hydrogen bonds) and XXVIII (forming two hydrogen bonds) for 40 (44%) and 50 (56%) of the occurrences, respectively.

It is interesting to note for these four U·U single mismatch-nearest neighbor sequence combinations, $\begin{bmatrix} {}^{5'}G\underline{U}C^{3'} \\ {}^{3'}C\underline{U}G^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}C\underline{U}G^{3'} \\ {}^{3'}G\underline{U}U^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}C\underline{U}C^{3'} \\ {}^{3'}G\underline{U}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}A\underline{U}G^{3'} \\ {}^{3'}U\underline{U}C^{5'} \end{bmatrix}$, there is at least one occurrence found for each where the U·U mismatch nucleotides are found to have no interaction with each other and are observed to be flipped-out from the center of the helix or to be positioned in such a way where hydrogen bonding is not possible through the ${}^{5'}$W/${}^{3'}$W paring type (data not shown). Furthermore, U·U mismatch nucleotides involved in the $\begin{bmatrix} {}^{5'}G\underline{U}C^{3'} \\ {}^{3'}C\underline{U}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}A\underline{U}G^{3'} \\ {}^{3'}U\underline{U}C^{5'} \end{bmatrix}$ sequence combinations are annotated to have no interaction for 16 and 50% of the total hits of each, respectively. This may suggest U·U mismatches are dynamic and interact with the surrounding environment under certain conditions, such as what is observed for the $\begin{bmatrix} {}^{5'}G\underline{U}C^{3'} \\ {}^{3'}C\underline{U}G^{5'} \end{bmatrix}$ sequence combination, which is annotated and observed to be in a hydrogen bonded (one or two bonds formed), stacked conformation (Figure 4a) and a non-hydrogen bonded, unstacked conformation, where one of the U nucleotides involved in the single mismatch is flipped-out from the center of the helix and is interacting with surrounding solvent (Figure 4b) in 84 and 16% of the occurrences, respectively. However, it is further interesting to note both of these geometric orientations were annotated to have the same ${}^{5'}$W/${}^{3'}$W nearest neighbors; therefore, it appears the difference in spatial arrangement of the mismatched nucleotides does not affect that of the adjacent base pairs. This loop sequence was thermodynamically measured to contribute favorably to duplex stability (92), which may result from the ability of one of the loop nucleotides to rotate between two positions without distorting the geometrical orientation of the nearest neighbors.

A·C single mismatches

Six of the eight A·C RNA single mismatch-nearest neighbor sequence combinations of the 30 most frequently occurring single mismatches in nature (84) are found in the RNA single mismatch structural database compiled here (Table 1). Considering these six combinations, a total of 89 A·C RNA single mismatch occurrences are found in the database; however, $\begin{bmatrix} {}^{5'}A\underline{A}C^{3'} \\ {}^{3'}U\underline{C}G^{5'} \end{bmatrix}$ accounts for 73 (82%) of these hits, with all other combinations accounting for only 4% each, on average. The mismatched nucleotides of $\begin{bmatrix} {}^{5'}A\underline{A}C^{3'} \\ {}^{3'}U\underline{C}G^{5'} \end{bmatrix}$ are most commonly observed in the ${}^{5'}$(A) H/${}^{3'}$(C) W pairing, antiparallel, *trans* orientation with the XXV (forming two hydrogen bonds) (Figure 5) or one_hbond hydrogen bonding pattern (each occurring ~50% of the time). When A·C mismatches are found with this type of orientation and these interactions, the 5′ and 3′ nearest neighbors are always found in the

$^{5'}$(A)Hh/$^{3'}$(U)Ws pairing, antiparallel, *trans* XXIV and $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX orientation and interaction, respectively. Similar to A·G single mismatches, the 5′ nearest neighbor does not have the expected $^{5'}$W/$^{3'}$W pairing. Contrary to A·G mismatches, A·C mismatches are not expected to disrupt the neighboring base pairs because this type of mismatch is comprised of one purine and pyrimidine base; therefore, it is similar in size to a canonical pair. This mismatch-nearest neighbor sequence combination was also found to engage in intra- and inter-strand interactions similar to what is observed for A·G mismatches. If the mismatch-nearest neighbor sequence is simply characterized as above, then inter- and intra-strand interactions are observed to form between nucleotides A and E and nucleotides B and C, respectively.

The remaining five A·C mismatch-nearest neighbor sequence combinations include $\begin{bmatrix} {}^{5'}\mathrm{C\underline{A}G}^{3'} \\ {}^{3'}\mathrm{G\underline{C}C}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{C\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{C}G}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}U}^{3'} \\ {}^{3'}\mathrm{C\underline{C}A}^{5'} \end{bmatrix}$, $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}G}^{3'} \\ {}^{3'}\mathrm{C\underline{C}C}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}C}^{3'} \\ {}^{3'}\mathrm{C\underline{C}G}^{5'} \end{bmatrix}$. These five can be divided into three groups based upon the geometric configuration of the mismatch nucleotides. The first group consists of the sequences $\begin{bmatrix} {}^{5'}\mathrm{C\underline{A}G}^{3'} \\ {}^{3'}\mathrm{G\underline{C}C}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{C\underline{A}C}^{3'} \\ {}^{3'}\mathrm{G\underline{C}G}^{5'} \end{bmatrix}$ and the mismatched nucleotides are annotated with $^{5'}$(A)Wh/$^{3'}$(C)Ww pairing, antiparallel, *cis* 75 (one hydrogen bond) geometric features. The second group consists of the sequences $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}G}^{3'} \\ {}^{3'}\mathrm{C\underline{C}C}^{5'} \end{bmatrix}$ and $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}C}^{3'} \\ {}^{3'}\mathrm{C\underline{C}G}^{5'} \end{bmatrix}$ and are annotated to have no interaction. Interestingly, the first and second groups exhibit the same 5′ and 3′ nearest neighbor orientations and interactions. These nearest neighbors are annotated to both be in the $^{5'}$Ww/$^{3'}$Ww pairing antiparallel *cis* orientation forming the canonical three hydrogen bonds in the XIX pattern. All four of these sequence combinations have G–C or C–G nearest neighbor base pairs at both the 5′ and 3′ side of the mismatch. Based upon the similarities in the type and orientation of the adjacent base pairs in these two groups, it is unclear why the A·C mismatched nucleotides are adopting different conformations.

The third group only consists of the $\begin{bmatrix} {}^{5'}\mathrm{G\underline{A}U}^{3'} \\ {}^{3'}\mathrm{C\underline{C}A}^{5'} \end{bmatrix}$ sequence combination, and the mismatched nucleotides are annotated to be in the $^{5'}$(A)Ww/$^{3'}$(C)Hw pairing antiparallel *cis*, one_hbond orientation. The 5′ nearest neighbor of this mismatch-nearest neighbor sequence exhibits the same geometric orientation and hydrogen bonding pattern as the first and second group of A·C mismatches. However, the 3′ nearest neighbor is unique in identity and orientation when compared to these groups. The U-A base pair at this position is either annotated to be in the $^{5'}$(A)W/$^{3'}$(C)Bh or $^{5'}$(A)W/$^{3'}$(C)W pairing, antiparallel, *trans* orientation with the 46 (one hydrogen bond) hydrogen bonding pattern.

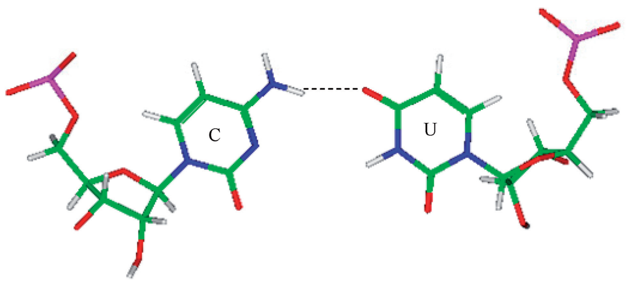


Figure 6. Representation of a C·U mismatch in the $^{5'}$(C)W/$^{3'}$(U)W pairing, antiparallel, *cis* orientation with one_hbond hydrogen bonding pattern (PDB ID 1FJG), which is the most common orientation and interaction determined for the most frequently occurring C·U mismatch-nearest neighbor combinations (84) that were also represented in the PDB.

C·U single mismatches

C·U RNA single mismatches are the fourth most frequently occurring mismatch type, with three C·U mismatch-nearest neighbor sequences found in the 30 most frequently occurring single mismatches (84). Only one of these combinations is represented in the RNA single mismatch structure database presented here. There are 76 occurrences of $\begin{bmatrix} {}^{5'}\mathrm{G\underline{C}C}^{3'} \\ {}^{3'}\mathrm{C\underline{U}G}^{5'} \end{bmatrix}$, and the C·U mismatch nucleotides are either in the $^{5'}$(C)W/$^{3'}$(U)W pairing, antiparallel, *cis* one_hbond conformation (Figure 6) or the nucleotides are annotated to have no interaction. However, it is important to note the C·U mismatches annotated to have no interaction are also observed in the 5′(C)W/3′(U)W orientation. The 5′ and 3′ nearest neighbor base pairs are both in the $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX orientation.

A·A single mismatches

A·A RNA single mismatches are the fifth most frequently occurring mismatch type (84). Additionally, there is only one A·A mismatch-nearest neighbor sequence combination, $\begin{bmatrix} {}^{5'}\mathrm{U\underline{A}A}^{3'} \\ {}^{3'}\mathrm{A\underline{A}U}^{5'} \end{bmatrix}$, found in the top 30, and it is not represented in the RNA 3D structure database. Therefore, this work does not contain structural information for this type of mismatch, but we are currently working to locate and annotate other A·A mismatch-nearest neighbor sequence combinations.

G·G single mismatches

G·G RNA single mismatches are the sixth most frequently occurring type of mismatch in nature (84). There is only one example of this mismatch type in the top 30 single

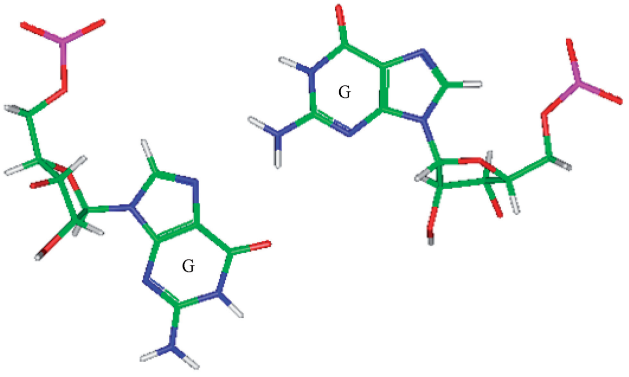


Figure 7. Representation of a G·G mismatch annotated as having no interaction (PDB ID 2QAM), which is the most common orientation and interaction determined for the most frequently occurring G·G mismatch-nearest neighbor combination, $\begin{bmatrix} ^{5'}A\underline{G}G^{3'} \\ ^{3'}U\underline{G}C^{5'} \end{bmatrix}$(84) that was also represented in the PDB.

mismatches, $\begin{bmatrix} ^{5'}A\underline{G}G^{3'} \\ ^{3'}U\underline{G}C^{5'} \end{bmatrix}$, and it is represented in the database presented here with 24 occurrences. The G·G mismatch nucleotides are either annotated to have no interaction (Figure 7) or in the $^{5'}$H/$^{3'}$Bs pairing, antiparallel, *trans* conformation. When the nucleotides are interacting, the two hydrogen bond patterns annotated are 34 (bifurcated hydrogen bond) or 112 (one hydrogen bond). However, the G·G mismatches annotated to have no interaction are also observed in the $^{5'}$H/$^{3'}$S orientation. Interestingly, regardless of the orientation and interaction of the mismatched nuceotides, the 5′ and 3′ nearest neighbor base pairs are always found in the $^{5'}$Hh/Ws$^{3'}$ pairing, antiparallel, *trans* XXIV and $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX conformations, respectively. Once again, it is interesting to note the 5′ nearest neighbor does not form the canonical $^{5'}$W/$^{3'}$W pairing type.

C·C single mismatches

C·C RNA single mismatches are the least frequently occurring mismatch type, and there are no C·C mismatch-nearest neighbor combinations found in the top 30 frequently occurring singe mismatches (84). Therefore, this work does not contain structural information for this type of mismatch, but we are currently working to locate and annotate C·C mismatch-nearest neighbor sequence combinations.

$\begin{bmatrix} ^{5'}\mathbf{G\underline{X}C}^{3'} \\ ^{3'}\mathbf{C\underline{X}G}^{5'} \end{bmatrix}$ Nearest neighbor comparison

There are four examples in the top 30 of the nearest neighbor combination $\begin{bmatrix} ^{5'}G\underline{X}C^{3'} \\ ^{3'}C\underline{X}G^{5'} \end{bmatrix}$, where X is any nucleotide, and all are represented here, which include $\begin{bmatrix} ^{5'}G\underline{A}C^{3'} \\ ^{3'}C\underline{G}G^{5'} \end{bmatrix}$, $\begin{bmatrix} ^{5'}G\underline{U}C^{3'} \\ ^{3'}C\underline{U}G^{5'} \end{bmatrix}$, $\begin{bmatrix} ^{5'}G\underline{A}C^{3'} \\ ^{3'}C\underline{C}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} ^{5'}G\underline{C}C^{3'} \\ ^{3'}C\underline{U}G^{5'} \end{bmatrix}$. It is important to note all three possible types of mismatches are present in this group: R·Y, R·R and Y·Y, when A and G are categorized as purines (R) and C and U are categorized as pyrimidines (Y). R·Y mismatches are similar in size to a canonical base pair since they are comprised of one purine and one pyrimidine; therefore, R·Y single mismatches are not likely disrupting the duplex backbone. R·R and Y·Y single mismatches are likely to disrupt the duplex backbone by causing the backbone to bulge-out or –in, respectively, to accommodate the mismatched nucleotides. Conversely, regardless of the mismatch type for these four sequence combinations, the 5′ and 3′ nearest neighbors are both in the $^{5'}$W/$^{3'}$W pairing, antiparallel, *cis* XIX conformation in ~99% of the occurrences.

$\begin{bmatrix} ^{5'}\mathbf{A\underline{X}C}^{3'} \\ ^{3'}\mathbf{U\underline{X}G}^{5'} \end{bmatrix}$ Nearest neighbor comparison

There are three examples in the top 30 of the nearest neighbor combination $\begin{bmatrix} ^{5'}A\underline{X}C^{3'} \\ ^{3'}U\underline{X}G^{5'} \end{bmatrix}$, but only two are represented in the RNA structural database, $\begin{bmatrix} ^{5'}A\underline{A}C^{3'} \\ ^{3'}U\underline{C}G^{5'} \end{bmatrix}$and $\begin{bmatrix} ^{5'}A\underline{A}C^{3'} \\ ^{3'}U\underline{G}G^{5'} \end{bmatrix}$. It is important to note the difference of mismatch type, R·Y versus R·R, for reasons stated in the previous section in regards to the size of the nucleotides comprising the mismatched base pair and the hypothesized effect on the backbone. Interestingly, the 5′ and 3′ nearest neighbors are most commonly found in the $^{5'}$H/$^{3'}$W pairing, antiparallel, *trans* XXIV and $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX conformations, respectively.

$\begin{bmatrix} ^{5'}\mathbf{C\underline{X}C}^{3'} \\ ^{3'}\mathbf{G\underline{X}G}^{5'} \end{bmatrix}$ Nearest neighbor comparison

There are three examples in the top 30 of the nearest neighbor combination $\begin{bmatrix} ^{5'}C\underline{X}C^{3'} \\ ^{3'}G\underline{X}G^{5'} \end{bmatrix}$, which are all represented in the structural database and include $\begin{bmatrix} ^{5'}C\underline{A}C^{3'} \\ ^{3'}G\underline{G}G^{5'} \end{bmatrix}$, $\begin{bmatrix} ^{5'}C\underline{U}C^{3'} \\ ^{3'}G\underline{U}G^{5'} \end{bmatrix}$ and $\begin{bmatrix} ^{5'}C\underline{A}C^{3'} \\ ^{3'}G\underline{C}G^{5'} \end{bmatrix}$. Similar to the previous nearest neighbor sequence combinations, both the 5′ and 3′ nearest neighbors are found in the $^{5'}$Ww/$^{3'}$Ww pairing, antiparallel, *cis* XIX conformation, in ~100% of the occurrences. It is interesting to note the 3′ nearest neighbor for $\begin{bmatrix} ^{5'}G\underline{X}C^{3'} \\ ^{3'}C\underline{X}G^{5'} \end{bmatrix}$, $\begin{bmatrix} ^{5'}A\underline{X}C^{3'} \\ ^{3'}U\underline{X}G^{5'} \end{bmatrix}$, and $\begin{bmatrix} ^{5'}C\underline{X}C^{3'} \\ ^{3'}G\underline{X}G^{5'} \end{bmatrix}$ is C-G, and the orientation and interaction of this base pair is found to be the same for each, regardless of the identities of 5′ nearest neighbor base pair and the mismatch nucleotides.

$\begin{bmatrix} ^{5'}\mathbf{A\underline{X}G}^{3'} \\ ^{3'}\mathbf{U\underline{X}C}^{5'} \end{bmatrix}$ Nearest neighbor comparison

There are three examples in the top 30 of the nearest neighbor combination $\begin{bmatrix} ^{5'}A\underline{X}G^{3'} \\ ^{3'}U\underline{X}C^{5'} \end{bmatrix}$, but only two are found in the structural database, $\begin{bmatrix} ^{5'}A\underline{U}G^{3'} \\ ^{3'}U\underline{U}C^{5'} \end{bmatrix}$ and

$\begin{bmatrix} {}^{5'}\mathrm{A\underline{G}G}^{\,3'} \\ {}^{3'}\mathrm{U\underline{G}C}^{\,5'} \end{bmatrix}$. The 5′ nearest neighbor conformation is different for each sequence combination. However, the 3′ nearest neighbor is identical in 98% of the total occurrences and is found to be ${}^{5'}$Ww/Ww${}^{3'}$ pairing, antiparallel, *cis* XIX, which is the same orientation and hydrogen bond pattern found in the above nearest neighbor comparisons.

In conclusion, the PDB is a rich source of structural information, and this work has undertaken the task of systematically locating, annotating and comparing the most frequently occurring RNA single mismatches in nature. The 2046 single mismatches presented here (Table 1 and Supplementary Table S2) account for only 42% of the total number of single mismatches found in the available PDB structures. Therefore, this study only begins to investigate the available data, and we are currently looking at and comparing the remaining single mismatches to identify more structural patterns.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Francois Major for providing them with executable versions of *MC-Search* and *MC-Annotate* and for providing assistance with the software in the context of the work described here. The authors would also like to thank Pamela Vanegas for assistance in refining the search script.

FUNDING

National Institute of General Medical Sciences (R15GM085699 to B.M.Z.). Monsanto Scholars Graduate Fellowship and the Saint Louis University Graduate School Dissertation Fellowship to A.R.D. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Mao,H., White,S.A. and Williamson,J.R. (1999) A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex. *Nat. Struct. Biol.*, **6**, 1139–1147.
2. Lee,J.-H., Culver,G., Carpenter,S. and Dobbs,D. (2008) Analysis of the EIAV rev-responsive element (RRE) reveals a conserved RNA motif required for high affinity rev binding in bond HIV-1 and EIAV. *PLoS ONE*, **3**, e2272.
3. Jones,S., Daley,D.T.A., Luscombe,N.M., Berman,H.M. and Thornton,J.M. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.
4. Beuth,B., García-Mayoral,M.F., Taylor,I.A. and Ramos,A. (2007) Scaffold-independent analysis of RNA-protein interactions: the nova-1 KH3-RNA complex. *J. Am. Chem. Soc.*, **129**, 10205–10210.
5. Messias,A.C. and Sattler,M. (2004) Structural basis of single-stranded RNA recognition. *Acc. Chem. Res.*, **37**, 279–287.
6. Hall,K.B. (2002) RNA-protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 283–288.
7. Hori,T., Taguchi,Y., Uesugi,S. and Kurihara,Y. (2005) The RNA ligands for mouse proline-rich RNA-binding protein (mouse Prrp) contain two consensus sequences in separate loop structure. *Nucleic Acids Res.*, **33**, 190–200.
8. Dubey,A.K., Baker,C.S., Romeo,T. and Babitzke,P. (2005) RNA sequence and secondary structure participate in high-affinity CsrA-RNA interaction. *RNA*, **11**, 1579–1587.
9. Nagai,K. (1996) RNA-protein complexes. *Curr. Opin. Struct. Biol.*, **6**, 53–61.
10. Steitz,T.A. (1999) *RNA Recognition by Proteins*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
11. Huppler,A., Nikstad,L.J., Allmann,A.M., Brow,D.A. and Butcher,S.E. (2002) Metal binding and base ionization in the U6 RNA intramolecular step-loop structure. *Nat. Struct. Biol.*, **9**, 431–435.
12. Grilley,D., Misra,V., Caliskan,G. and Draper,D.E. (2007) Importance of partially unfolded conformations for Mg^{2+}-induced folding of RNA tertiary structure: structural models and free energies of Mg^{2+} interactions. *Biochemistry*, **46**, 10266–10278.
13. Casiano-Negroni,A., Sun,X. and Al-Hashimi,H.M. (2007) Probing Na+-induced changes in the HIV-1 TAR conformational dynamics using NMR residual dipolar couplings: new insights into the role of counterions and electrostatic interactions in adaptive recognition. *Biochemistry*, **46**, 6525–6535.
14. Donarski,J., Shammas,C., Banks,R. and Ramesh,V. (2006) NMR and molecular modelling studies of the binding of amicetin antibiotic to conserved secondar structural motifs of 23S ribosomal RNA. *J. Antibiot.*, **59**, 177–183.
15. Liu,X., Thomas,J.R. and Hergenrother,P.J. (2004) Deoxystreptamine dimers bind to RNA hairpin loops. *J. Am. Chem. Soc.*, **126**, 9196–9197.
16. Chushak,Y. and Stone,M.O. (2009) In *silico* selection of RNA aptamers. *Nucleic Acids Res.*, **37**, e87.
17. Meyer,S.T. and Hergenrother,P.J. (2009) Small molecular ligands for bulged RNA secondary structures. *Org. Lett.*, **11**, 4052–4055.
18. Childs-Disney,J.L., Wu,M., Pushechnikov,A., Aminova,O. and Disney,M.D. (2007) A small molecule microarray platform to select RNA internal loop-ligand interactions. *ACS Chem. Biol.*, **2**, 745–754.
19. Gallego,J. and Varani,G. (2001) Targeting RNA with small-molecule drugs: Therapeutic promise and chemical challenges. *Accounts Chem. Res.*, **34**, 836–843.
20. Chang,K.-Y. and Tinoco,I. Jr (1997) The structure of an RNA "kissing" hairpin complex of the HIV TAG hairpin loop and its complement. *J. Mol. Biol.*, **269**, 52–66.
21. Shankar,N., Kennedy,S.D., Chen,G., Krugh,T.R. and Turner,D.H. (2006) The NMR structure of an internal loop from 23S ribosomal RNA differs from its structure in crystals of 50S ribosomal subunits. *Biochemistry*, **45**, 11776–11789.
22. 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.
23. 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.
24. 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.
25. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
26. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
27. Lu,Z.J., Gloor,J.W. and Mathews,D.H. (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA*, **15**, 1805–1813.
28. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2007) Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, **23**, i19–i28.
29. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.

30. Hamada,M., Kiryu,H., Sato,K., Mituyama,T. and Asai,K. (2009) Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**, 465–473.
31. Dowell,R.D. and Eddy,S.R. (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 71–84.
32. Parisien,M., Cruz,J.A., Westhof,É. and Major,F. (2009) New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, **15**, 1875–1885.
33. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci.*, **104**, 114664–114669.
34. Ding,F., Sharma,S., Chalasani,P., Demidov,V.V., Broude,N.E. and Dokholyan,N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164–1173.
35. Jonikas,M.A., Radmer,R.J., Laederach,A., Das,R., Pearlman,S., Herschlag,D. and Altman,R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, **15**, 189–199.
36. Martinez,H.M., Maizel,J.V. and Shapiro,B.A. (2008) RNA2D3D: a program for generating, viewing, and comparing three-dimensional models of RNA. *J. Biomol. Struct. Dyn.*, **25**, 669–683.
37. Massire,C. and Westhof,E. (1998) MANIP: an interactive tool for modelling RNA. *J. Mol. Graphics Modell*, **16**, 197–205, 255–257.
38. Michel,F. and Westhof,E. (1990) Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
39. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
40. Batey,R.T., Rambo,R.P., Lucast,L., Rha,B. and Doudna,J.A. (1999) Tertiary motifs in RNA structure and folding. *Angew. Chem., Int. Ed.*, **38**, 2326–2343.
41. Westhof,E. and Fritsch,V. (2000) RNA folding: beyond Watson-Crick pairs. *Structure with Folding & Design*, **8**, R55–R65.
42. Ferré-D'Amare,A.R. and a,D.J.A. (1999) RNA folds: insights from recent crystal structures. *Annu. Rev. Biophys. Biophys. Chem.*, **28**, 57–73.
43. Hermann,T.a.P.D.J. (1999) Stitching together RNA tertiary architectures. *J. Mol. Biol.*, **294**, 829–849.
44. 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.
45. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
46. Westbrook,J., Feng,Z., Chen,L., Huanwang,Y. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
47. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al.* (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
48. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
49. Nagaswamy,U., Voss,N., Zhang,Z.D. and Fox,G.E. (2000) Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res.*, **28**, 375–376.
50. Nagaswamy,U., Larios-Sanz,M., Hury,J., Collins,S., Zhang,Z.D., Zhao,Q. and Fox,G.E. (2002) NCIR: A database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.*, **30**, 395–397.
51. Xin,Y. and Olson,W.K. (2009) BPS: a database of RNA base-pair structures. *Nucleic Acids Res.*, **37**, D38–D88.
52. Schnare,M.N., Damberger,S.H., Gray,M.W. and Gutell,R.R. (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. *J. Mol. Biol.*, **256**, 701–719.
53. Gautheret,D., Konings,D. and Gutell,R.R. (1994) A major family of motifs involving G.A mismatches in ribosomal RNA. *J. Mol. Biol.*, **242**, 1–8.
54. Gautheret,D., Konings,D. and Gutell,R.R. (1995) GU base-pairing motifs in ribosomal-RNA. *RNA*, **1**, 807–814.
55. Leontis,N.B. and Westhof,E. (1998) Conserved geometrical base-pairing patterns in RNA. *Q. Rev. Biophys.*, **31**, 399–455.
56. Leontis,N.B. and Westhof,E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499–512.
57. Leontis,N.B. and Westhof,E. (2002) Survey and summary: the non-Watson-Crick pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
58. Lescoute,A. and Westhof,E. (2006) The interaction networks of structured RNAs. *Nucleic Acids Res.*, **34**, 6587–6604.
59. Leontis,N.B., Lescoute,A. and Westhof,E. (2006) The building blocks and motifs of RNA architecture. *Curr. Opin. Struct. Biol.*, **16**, 279–287.
60. Lescoute,A., Leonteis,N.B., Massire,C. and Westhof,E. (2005) Recurrent structural RNA motifs. isostericity matrices and sequence alignments. *Nucleic Acids Res.*, **33**, 2395–2409.
61. Leontis,N.B. and Westhof,E. (2003) Analysis of RNA motifs. *Curr. Opin. Struct. Biol.*, **13**, 300–308.
62. Leontis,N.B. and Westhof,E. (2002) The annotation of RNA motifs. *Comparative Funct Genomics*, **3**, 518–524.
63. Klosterman,P.S., Tamura,M., Holbrook,S.R. and Brenner,S.E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.*, **30**, 392–394.
64. Klosterman,P.S., Hendrix,D.K., Tamura,M., Holbrook,S.R. and Brenner,S.E. (2004) Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucleic Acids Res.*, **32**, 2342–2352.
65. Tamura,M., Hendrix,D.K., Klosterman,P.S., Schimmelman,N.R.B., Brenner,S.E. and Holbrook,S.R. (2004) SCOR: structrual classification of RNA, version 2.0. *Nucleic Acids Res.*, **32**, D182–D184.
66. Leontis,N.B., Altman,R.B., Berman,H.M., Brenner,S.E., Brown,J.W., Engelke,D.R., Harvey,S.C., Holbrook,S.R., Jossinet,F., Lewis,S.E. *et al.* (2006) The RNA Ontology Consortium: an open invitation to the RNA community. *RNA*, **12**, 533–541.
67. Gendron,P., Lemieux,S. and Major,F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919–936.
68. Lemieux,S. and Major,F. (2002) RNA canonical and non-canonical base-pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.*, **30**, 4250–4263.
69. Lisi,V. and Major,F. (2007) A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence-structure relationships. *RNA*, **13**, 1537–1545.
70. Hoffmann,B., Mitchell,G.T., Gendron,P., Major,F., Anderson,A.A., Collins,R.A. and Legault,P. (2003) NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. *Proc. Natl Acad. Sci. USA*, **100**, 7003–7008.
71. Olivier,C., Poirier,G., Gendron,P., Boisgontier,A., Major,F. and Chartrand,P. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol. Cell. Biol.*, **25**, 4752–4766.
72. 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.
73. Calin-Jageman,I. and Nicholson,A.W. (2003) Mutational analysis of an RNA internal loop as a reactivity epitope for *Escherichia coli* ribonuclease III substrates. *Biochemistry*, **42**, 5025–5034.
74. Saito,H. and Richardson,C.C. (1981) Processing of mRNA by ribonuclease III regulates expression of gene 1.2 of bacteriophage T7. *Cell*, **27**, 533–542.
75. Du,T. and Zamore,P.D. (2005) MicroPrimer: the biogenesis and function of microRNA. *Development*, **132**, 4645–4652.
76. Bae,S.H., Cheong,H.K., Lee,J.H., Cheong,C., Kainosho,M. and Choi,B.S. (2001) Structural features of an influenza virus promoter and their implications for viral RNA synthesis. *Proc. Natl Acad. Sci. USA*, **98**, 10602–10607.

77. Huthoff,H. and Berkhout,B. (2002) Multiple secondary structure rearrangements during HIV-1 RNA dimerization. *Biochemistry*, **41**, 10439–10445.
78. Schüler,M., Connell,S.R., Lescoute,A., Giesebrecht,J., Dabrowski,M., Schroeer,B., Mielke,T., Penczek,P.A., Westhof,E. and Spahn,C.M.T. (2006) Structure of the ribosome-bound cricket paralysis virus IRES RNA. *Nat. Struct. Mol. Biol.*, **13**, 1092–1096.
79. Wientges,J., Putz,J., Giege,R., Florentz,C. and Schwienhorst,A. (2000) Selection of viral RNA-derived tRNA-like structures with improved valylation activities. *Biochemistry*, **39**, 6207–6218.
80. Thunder,C., Witwer,C., Hofacker,I.L. and Stadler,P.F. (2004) Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.*, **85**, 1113–1124.
81. 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.
82. Everett,C.M. and Wood,N.W. (2004) Trinucleotide repeats and neurodegenerative disease. *Brain*, **127**, 2385–2405.
83. Ranum,L.P.W. and Day,J.W. (2004) Myotonic dystrophy: RNA pathogenesis comes into focus. *Amer. J. Hum. Gen.*, **74**, 793–804.
84. Davis,A.R. and Znosko,B.M. (2007) Thermodynamic characterization of single mismatches found in naturally occurring RNA. *Biochemistry*, **46**, 13425–13436.
85. Donohue,J. and Trueblood,K.N. (1960) Base-pairing in DNA. *J. Mol. Biol.*, **2**, 363–371.
86. Donohue,J. (1956) Hydrogen-bonded helical configurations of polynucleotides. *Proc. Natl Acad. Sci. USA*, **42**, 60–65.
87. Saenger,W. (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag New York, Inc., NY.
88. Gautheret,D. and Gutell,R.R. (1997) Inferring the conformation of RNA base pairs and triples from patterns of sequence variation. *Nucleic Acids Res.*, **25**, 1559–1564.
89. Lemieux,S., Chartrand,P., Cedergren,R. and Major,F. (1998) Modeling active RNA structures using the intersection of conformational space: application to the lead-activated ribozyme. *RNA*, **4**, 739–749.
90. Gabb,H.A., Sanghani,S.R., Rober,C.H. and Prevost,C. (1996) Finding and visualizing nucleic acid base stacking. *J. Mol. Graphics Modell.*, **14**, 23–24.
91. Major,F. and Thibault,P. (2007) RNA tertiary structure prediction. In Lengauer,T. (ed.), *Bioinformatics: From Genentics to Therapies*. Wiley-VCH, Weinheim, Germany, pp. 491–539.
92. Kierzek,R., Burkard,M.E. and Turner,D.H. (1999) Thermodynamics of single mismatches in RNA duplexes. *Biochemistry*, **38**, 14214–14223.