*RNA* (2001), *7*:499–512. Cambridge University Press. Printed in the USA.

NOMENCLATURE PROPOSAL

# Geometric nomenclature and classification of RNA base pairs

**NEOCLES B. LEONTIS[1] and ERIC WESTHOF[2]**

[1] Chemistry Department and Center for Biomolecular Sciences, Bowling Green State University,
Bowling Green, Ohio 43403, USA
[2] Institut de Biologie Moléculaire et Cellulaire du Centre National de la Recherche Scientifique,
Modélisation et Simulations des Acides Nucléiques, Unité Propre de Recherche 9002,
F-67084 Strasbourg Cedex, France

## ABSTRACT

**Non-Watson–Crick base pairs mediate specific interactions responsible for RNA–RNA self-assembly and RNA–protein recognition. An unambiguous and descriptive nomenclature with well-defined and nonoverlapping parameters is needed to communicate concisely structural information about RNA base pairs. The definitions should reflect underlying molecular structures and interactions and, thus, facilitate automated annotation, classification, and comparison of new RNA structures. We propose a classification based on the observation that the planar edge-to-edge, hydrogen-bonding interactions between RNA bases involve one of three distinct edges: the Watson–Crick edge, the Hoogsteen edge, and the Sugar edge (which includes the 2′-OH and which has also been referred to as the Shallow-groove edge). Bases can interact in either of two orientations with respect to the glycosidic bonds, *cis* or *trans* relative to the hydrogen bonds. This gives rise to 12 basic geometric types with at least two H bonds connecting the bases. For each geometric type, the relative orientations of the strands can be easily deduced. High-resolution examples of 11 of the 12 geometries are presently available. Bifurcated pairs, in which a single exocyclic carbonyl or amino group of one base directly contacts the edge of a second base, and water-inserted pairs, in which single functional groups on each base interact directly, are intermediate between two of the standard geometries. The nomenclature facilitates the recognition of isosteric relationships among base pairs within each geometry, and thus facilitates the recognition of recurrent three-dimensional motifs from comparison of homologous sequences. Graphical conventions are proposed for displaying non-Watson–Crick interactions on a secondary structure diagram. The utility of the classification in homology modeling of RNA tertiary motifs is illustrated.**

**Keywords: bifurcated; Hoogsteen edge; isostericity; nomenclature; non-Watson–Crick base pairing; Shallow-groove; Sugar-edge; water-inserted; Watson–Crick edge**

## INTRODUCTION

Nucleic acid bases interact by stacking or by abutting edge-to-edge. Whereas stacking interactions provide most of the driving force for folding, the edge-to-edge interactions, mediated by hydrogen bonding between complementary arrays of electrically polarized atoms, provide directionality and specificity. The standard or canonical Watson–Crick pairs are characterized by their remarkable isostericity, which gives rise to the regular A-form double helix, and allows each of the four combinations to substitute for any of the others without distorting the three-dimensional helical structure. The canonical Watson–Crick pairs, however, represent only one of many possible edge-to-edge interactions (Leontis & Westhof, 1998c). The rapid progress of RNA crystallography has revealed a rich variety of base-pairing geometries (Batey et al., 1999; Westhof & Fritsch, 2000). This variety gives rise, in turn, to a multitude of complex tertiary structural motifs, as revealed by recent progress in RNA structural biology (Ferré-D'Amaré & Doudna, 1999; Hermann & Patel, 1999).

We feel that the growing literature on RNA structural biology is hampered by the lack of a systematic nomenclature for base pairing interactions. Historical, ambiguous, and, sometimes, confusing terms are used (e.g., “reverse” and “flipped”). Frequently, recourse is

Reprint requests to: Neocles B. Leontis, Chemistry Department and Center for Biomolecular Sciences, Overman Hall, Bowling Green State University, Bowling Green, Ohio 43403, USA; e-mail: Leontis@bgnet.bgsu.edu, or to Eric Westhof, e-mail: E.Westhof@ibmc.u-strasbg.fr.

made to stating the functional groups involved in the H-bonding interactions, which impedes ready visualization of the interactions. Further, the relationships with the relative strand orientations are obscured. We suggest that the utility of the recently compiled and exhaustive database of noncanonical base pairs observed in X-ray and NMR structures, http://prion.bchs.uh.edu/bp_type/, could be significantly enhanced by organizing the base pairs along geometric principles (Nagaswamy et al., 2000). Indeed, pairwise analysis of hydrogen-bonded, edge-to-edge interactions reveals recurrent geometric patterns that provide a natural (i.e., structural) means of classification. Such a classification can serve to organize the observed pairs into isosteric families (Leontis & Westhof, 1998c, 1999) and thus provides for systematic and descriptive nomenclature that facilitates prediction of isosteric pairs, necessary steps in motif recognition in RNA sequences.

The classification that we propose is based on the observation that, while only about 60% of bases in structured RNAs participate in canonical Watson–Crick base pairs, the great majority of the remainder participate in some other kind of edge-to-edge interactions with one or more other bases. This is borne out in the atomic-resolution structures of the large and small ribosomal subunits, the solution of which has expanded our database of RNA structure several-fold (Ban et al., 2000; Schluenzen et al., 2000; Wimberly et al., 2000). The non-Watson–Crick pairs define, in large part, the tertiary structure of an RNA. Thus, the tertiary structure can be decomposed into a collection of three-dimensional motifs held together by pairwise interactions that can be specified simply by indicating the interacting edges and the relative orientations of the glycosidic bonds of the two bases.

First it will be shown that there are 12 basic families of base pairs. Examples from each family will be presented and the correspondences between the proposed nomenclature and some current usage will be presented. Next the default strand orientations for each base pair type will be presented with simple rules for their visualization, extending previous work (Lavery et al., 1992; Westhof, 1992). Finally, the utility of the nomenclature in summarizing RNA tertiary structure in a two-dimensional format will be illustrated.

Because a nomenclature is fundamentally a working and networking tool, the adoption of a nomenclature, regardless of its merits, must be the result, in the end, of a consensus agreement between the members of a given scientific community. Therefore, we wish to arouse discussions and not controversies. Informal groups, working on the establishment of conventions useful for RNA research, gather regularly at the annual RNA Society meetings with the coordination of Russ B. Altman (russ.altman@stanford.edu). The proposed nomenclature has been presented and discussed at the RNA 2000 meeting.

## RESULTS AND DISCUSSION

### Twelve basic geometric families

RNA purine and pyrimidine bases present three edges for H-bonding interactions, as shown in Figure 1 (left panel). These are the *Watson–Crick edge*, the *Hoogsteen edge* (for purines) or the equivalent *“CH” edge* (for pyrimidines), and the *Sugar edge*, so-named because it includes the 2′-hydroxyl group. Although “Hoogsteen edge” applies only to purines, it will be used to refer also to the CH edge of pyrimidines, as the atoms involved are normally found in the *Deep groove* of the A-type helix, which corresponds to the *Major groove* of B-DNA. In previous works, the third edge was named “Shallow-groove edge” (Leontis & Westhof, 1998c), because the bases interacting with that edge are located in the RNA helix Shallow groove, which is equivalent to the B-DNA *Minor groove*. We thought it was important to emphasize the distinct and characteristic geometrical differences between the two major helices, the B-DNA type and the A-RNA type. However, with time, the word “minor” as applied to nucleic acid helices has been decoupled from its geometrical meaning. Although names should help memory, they should not convey mistaken meanings (the A-RNA shallow groove is anything but “minor,” either regarding function or shape). The designation “sugar-edge” has the advantage that it may be applied to B-DNA as well as to A-RNA.

A given edge of one base can potentially interact in a plane with any one of the three edges of a second base, and can do so in either the *cis* or *trans* orientation of the glycosidic bonds (this nomenclature was used before, see, e.g., Sundaralingam, 1977). The *cis* and *trans* orientations, which follow the usual stereochemical meanings, are illustrated in the right panel of Figure 1 for two bases interacting with their Watson–Crick edges. Thus, 12 distinct edge-to-edge interactions are possible. Each pairing geometry is designated by stating the interacting edges of each of the two bases (Watson–Crick, Hoogsteen, or Sugar edge) and the relative glycosidic bond orientation, *cis* or *trans*. The order in which the base pairs are listed in Table 1 is determined by a historically based priority rule: Watson–Crick edge > Hoogsteen edge > Sugar edge. The 12 base pair geometries are listed in Table 1, with the local strand orientations in the default *anti* configurations of the bases with respect to the sugars. Examples taken from high-resolution X-ray structures of 11 of the 12 basic types are shown in Figures 2 and 3.

When one of the interacting bases occurs in the rare *syn* configuration of the glycosidic bond, the local strand orientations given in Table 1 are reversed. Thus, in Z-DNA, the G=C Watson–Crick pair with the G in *syn* presents a locally *parallel* orientation of the strands (the O4′-atoms of the sugars of the paired bases point in the same directions), despite the globally *antiparallel*

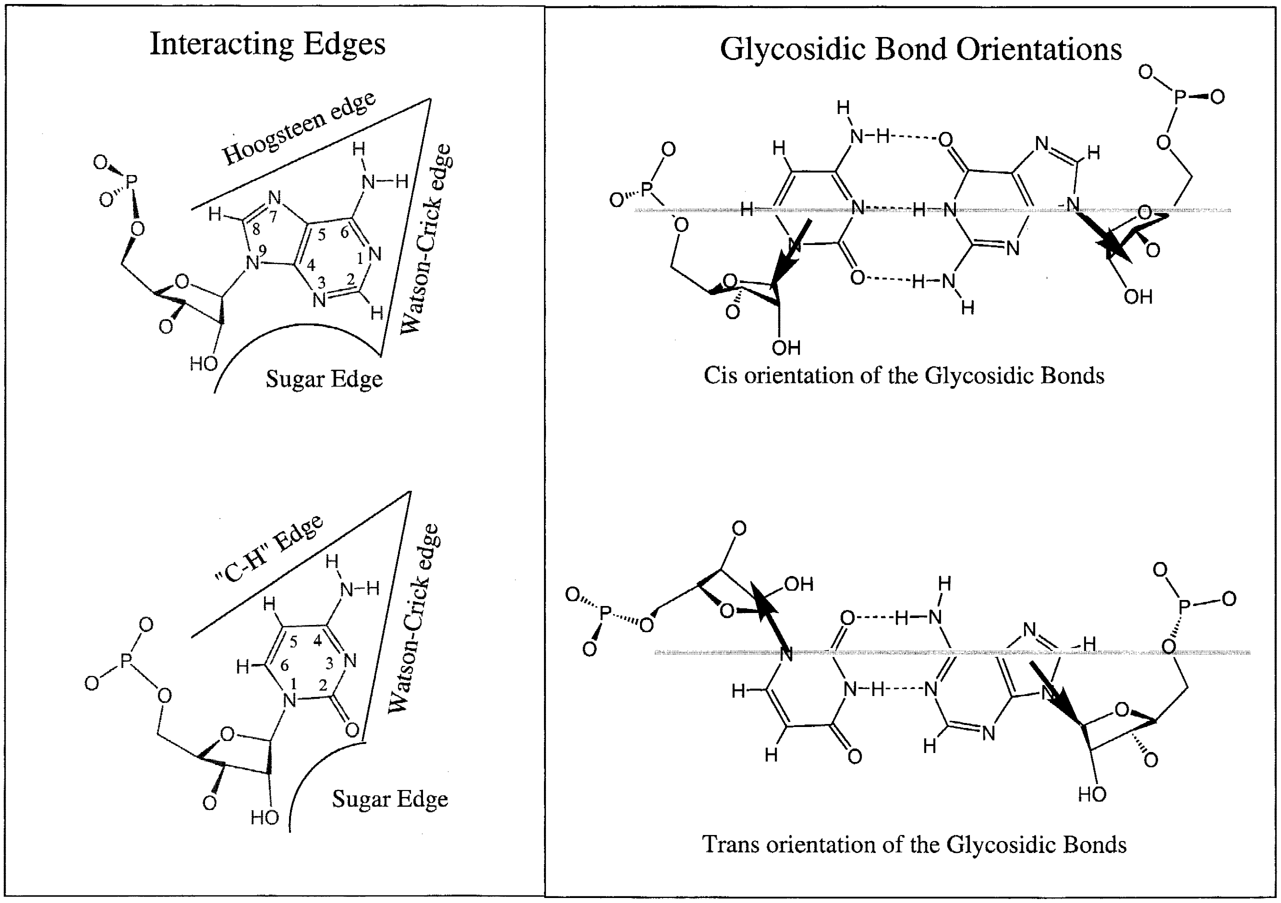

Interacting Edges
Hoogsteen edge
Watson-Crick edge
Sugar Edge
"C-H" Edge
Watson-Crick edge
Sugar Edge
Glycosidic Bond Orientations
Cis orientation of the Glycosidic Bonds
Trans orientation of the Glycosidic Bonds

**FIGURE 1.** Left panel: Purine (A or G, indicated by "R") and pyrimidine (C or U, indicated by "Y") bases provide three edges for interaction, as shown for adenosine and cytosine. The Watson–Crick edge comprises A(N6)/G(O6), R(N1), A(C2)/G(N2), U(O4)/C(N4), Y(N3), and Y(O2). The Hoogsteen edge comprises A(N6)/G(O6), R(N7), U(O4)/C(N4), and Y(C5). The Sugar-edge comprises A(C2)/G(N2), R(N3), Y(O2), and the ribose hydroxyl group, O2′. Right panel: The *cis* and *trans* orientations are defined relative to a line drawn parallel to and between the *base-to-base* hydrogen bonds in the case of two hydrogen bonds or, in the case of three hydrogen bonds, along the middle hydrogen bond.

orientation of the strands. In the very rare case that both bases are *syn*, the strand orientations revert to those given in Table 1. Thus, the proposed system eliminates the need to speak of "flipped" bases, "reverse" orientations, or to explicitly state the donor and acceptor atoms. With a mental image of the edges that each base of an RNA nucleotide presents for interaction, one can easily visualize and memorize the essential geometry of each interaction. To facilitate the adoption of the proposed nomenclature, we present in Table 2 the correspondence between our nomenclature and the base-pair designations given in the web-accessible database cited above, grouped according to geometric type.

The canonical A-U and G=C pairs belong to the *cis* Watson–Crick/Watson–Crick (W.C./W.C.) geometry. The so-called wobble pairs also belong to this group. Originally, the term "wobble" designated the pairing between the noncomplementary bases G and U and pairs

**TABLE 1.** The 12 main families of base pairs between nucleic acid bases together with the local strand orientation (which assumes that all bases are in the default *anti* conformation; a *syn* orientation would imply a reversal of orientation; for the global orientation, the stereochemistry at the phosphate groups has to be considered).

| No. | Glycosidic bond orientation | Interacting edges | Local strand orientation |
|---|---|---|---|
| 1 | *Cis* | Watson–Crick/Watson–Crick | Antiparallel |
| 2 | *Trans* | Watson–Crick/Watson–Crick | Parallel |
| 3 | *Cis* | Watson–Crick/Hoogsteen | Parallel |
| 4 | *Trans* | Watson–Crick/Hoogsteen | Antiparallel |
| 5 | *Cis* | Watson–Crick/Sugar Edge | Antiparallel |
| 6 | *Trans* | Watson–Crick/Sugar Edge | Parallel |
| 7 | *Cis* | Hoogsteen/Hoogsteen | Antiparallel |
| 8 | *Trans* | Hoogsteen/Hoogsteen | Parallel |
| 9 | *Cis* | Hoogsteen/Sugar Edge | Parallel |
| 10 | *Trans* | Hoogsteen/Sugar Edge | Antiparallel |
| 11 | *Cis* | Sugar Edge/Sugar Edge | Antiparallel |
| 12 | *Trans* | Sugar Edge/Sugar Edge | Parallel |

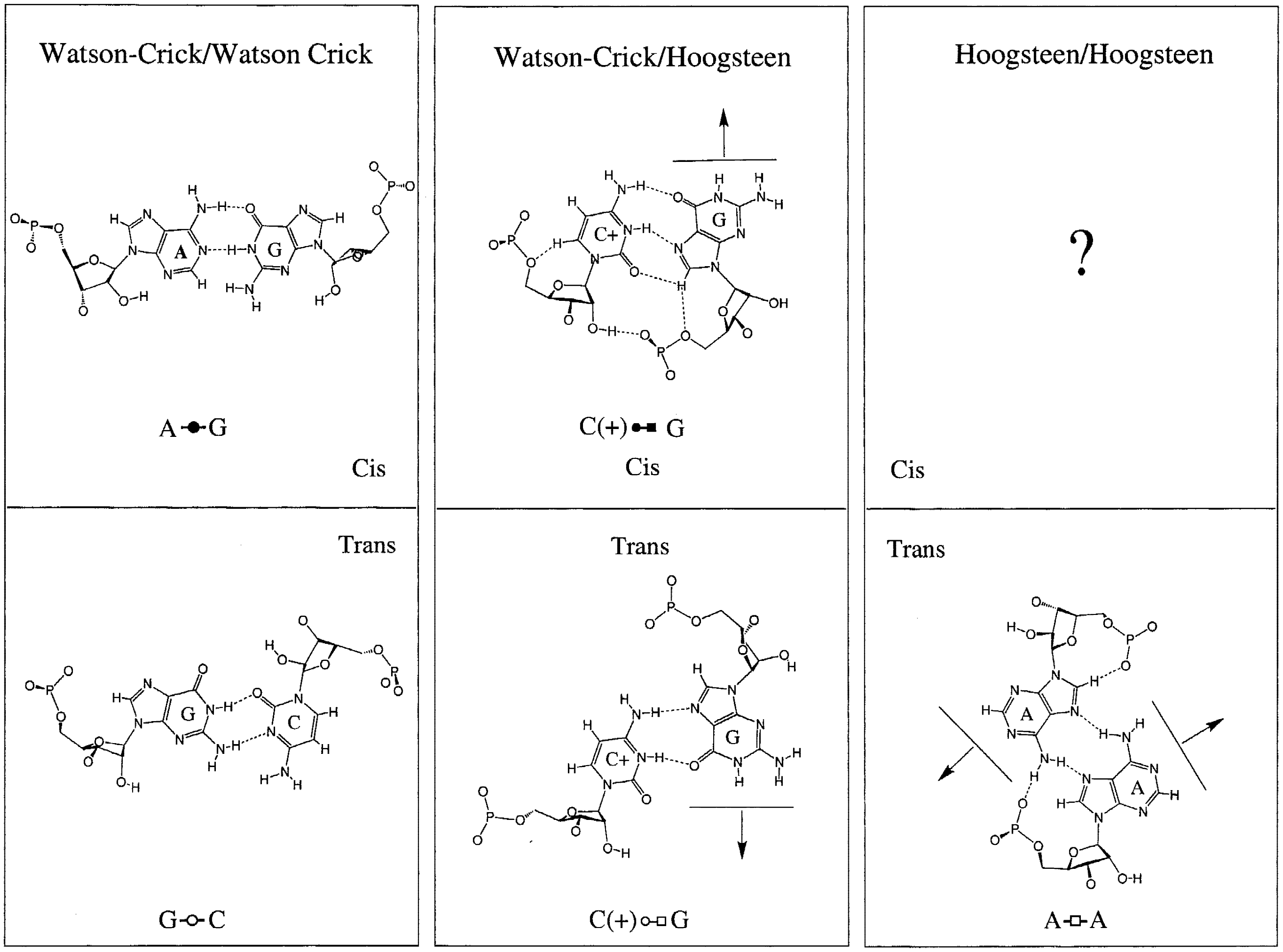


**FIGURE 2.** Six possible edge-to-edge base pairing geometries involving Watson–Crick and Hoogsteen edges in all combinations. Upper left: *Cis* Watson–Crick/Watson–Crick A•G NDB file URX053 (Cate et al., 1996). Lower left: *Trans* Watson–Crick/Watson–Crick G•C (Westhof et al., 1988). Upper center: *Cis* Watson–Crick/Hoogsteen C(+)•G, NDB file URX053 (Cate et al., 1996). Lower center: *Trans* Watson–Crick/Hoogsteen C(+)•G, UR0004 (Su et al., 1999). Lower left: *Trans* Hoogsteen/Hoogsteen A•A, TRNA09 (Westhof et al., 1988). No high resolution example of cis Hoogsteen/Hoogsteen was identified (upper right panel). Arrows designate Watson–Crick edges available for further interactions with other RNA units, proteins, or small molecules. The designation of each base pair using the symbols proposed in Figure 6 is also shown.

involving the modified residue inosine (Crick, 1966). Such pairs are characterized geometrically by a shift of one base relative to the other. We feel the term "wobble" should be restricted to those pairs in *cis* W.C./W.C. with a shift of the pyrimidine base and should not be extended to *trans* W.C./W.C. pairs even in those cases where a shift occurs. Although wobble pairs often can substitute for canonical pairs or constitute intermediates between them, they are not strictly isosteric with them (nor do they share the property of being self-isosteric). That is, a wobble GoU is not isosteric to its switched occurrence, UoG. Likewise, although N1-protonated adenosine forms a pair with cytosine that is isosteric to wobble GoU, the wobble type A(+)oC pair is not isosteric to CoA(+) nor is it isosteric to UoG. Recent reviews of wobble pairs are available (Masquida & Westhof, 2000; Varani & McClain, 2000).

## Strand orientations

The understanding of RNA folding and architecture, as well as interactive three-dimensional modeling, requires keeping track of the relative orientations of the strands to which the interacting bases belong. In Figures 4 and 5, each base-pairing geometry is displayed schematically using two right triangles abutting edge-to-edge. In each triangle, the sides adjacent to the right angle represent the Watson–Crick and Sugar edges of each base. The hypotenuse of the triangle represents the Hoogsteen edge. A cross or circle in the corner where the Hoogsteen and Sugar edges meet indicates the orientation of the sugar-phosphate backbone relative to the plane of the page (5′ to 3′ or 3′ to 5′). The six *cis* and the six *trans* edge-to-edge pairing geometries are displayed in separate, symmetric 3 × 3 ma-

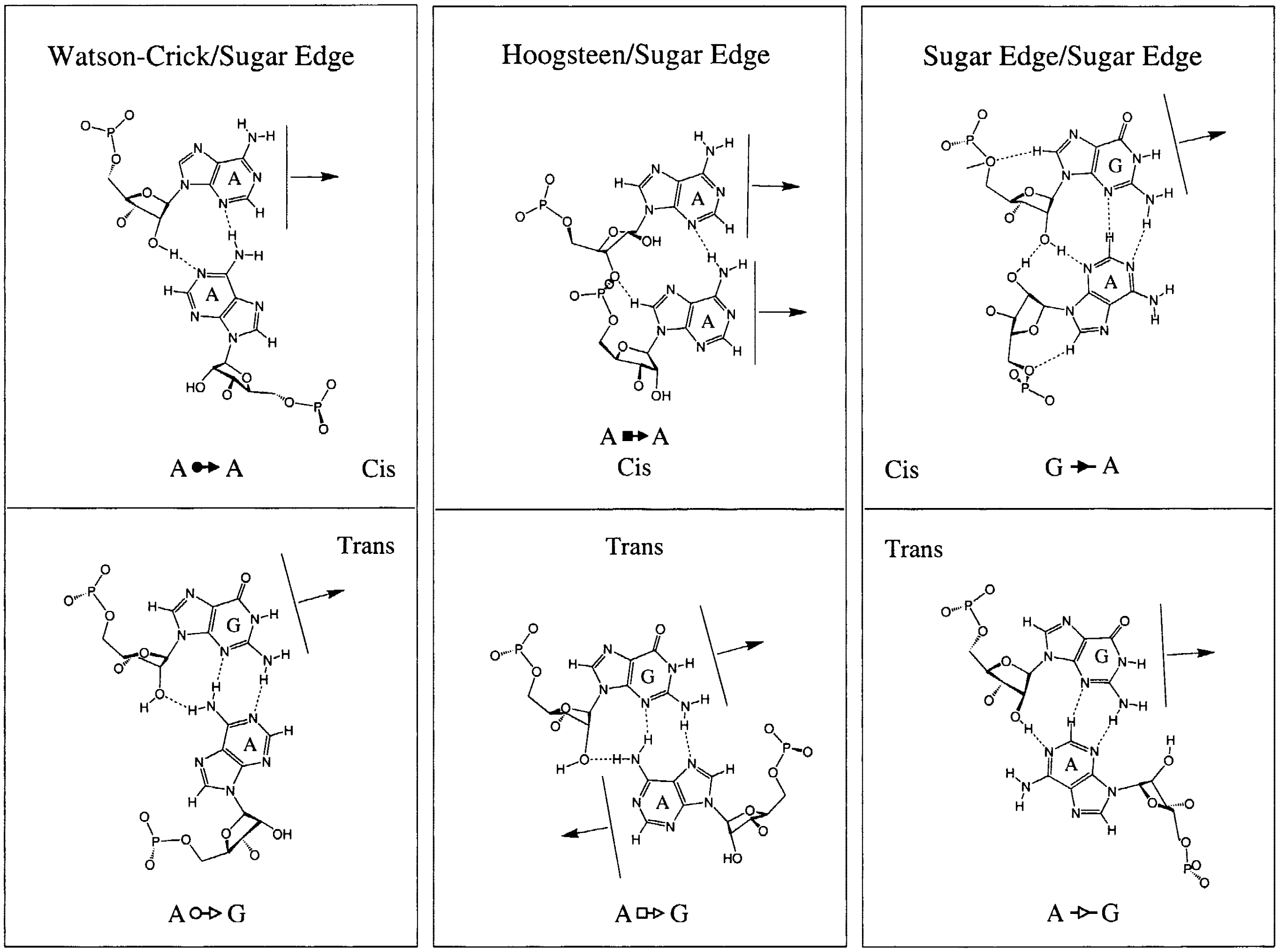


**FIGURE 3.** Six possible edge-to-edge base pairing geometries involving the Sugar edge in all combinations. Upper left: *Cis* Watson–Crick/Sugar Edge A•A, NDB file TRN007 (Westhof et al., 1988). Lower left: *Trans* Watson–Crick/Sugar Edge A•G, NDB file UR0004 (Su et al., 1999). Upper center: *Cis* Hoogsteen/Sugar Edge A•A, URX053 (Cate et al., 1996). Lower center: *Trans* Hoogsteen/Sugar Edge A•G, URL064 (Correll et al., 1997). Upper right: *Cis* Sugar Edge/Sugar Edge A•G,URX053 (Cate et al., 1996). Lower right: *Trans* Sugar Edge/Sugar Edge A•G UR0004 (Su et al., 1999). As in Figure 2, arrows designate Watson–Crick edges available for further interactions with other RNA units, proteins, or small molecules and the symbolic designation of each base pair according to Figure 6 is also shown.

trices. The elements of each matrix are arranged in the order Watson–Crick, Hoogsteen, and Sugar edge. Thus, the W.C./W.C. pair is placed in the first row, first column and the Hoogsteen/Sugar-edge pair in the second row, third column. In these diagrams, the position and strand orientation of the base on the left is fixed in space.

When arranged in this manner, the base pairs on the main diagonal of each matrix have the same strand orientation, antiparallel for *cis* pairs and parallel for *trans* pairs. Those in the first diagonal next to the main diagonal have opposite strand orientations, parallel for *cis* pairs and antiparallel for *trans* pairs. The strand orientation of the corner element (first row, third column) reverts to that of the main diagonal. Thus, any purely horizontal or vertical move in the table, corresponding to the change of one edge while retaining the *cis* or *trans* geometry, changes the strand orientation, whereas any diagonal move retains the strand orientation.

## Annotation of two-dimensional diagrams

It is desirable to present the non-Watson–Crick pairs of an RNA molecule on a standard two-dimensional drawing. This helps to recognize and to communicate succinctly in a visually accessible manner the essential features of a motif. This, in turn, facilitates recognition of shared three-dimensional tertiary motifs and foldings. Such diagrams should show, in addition to the classical secondary structure (contiguous canonical pairs forming A-form double-stranded helices maintained by Watson–Crick and wobble pairs), all non-Watson–Crick pairs, all points in the covalent chain at which the strand polarity reverses direction, and key base-stacking interactions, to the degree possible without overly cluttering the picture. As is usually done, nucleotides should be numbered sequentially (5′ to 3′) to aid in tracing the covalent chain. Nucleotides are

**TABLE 2.** Correspondence of proposed names to the numbering of Saenger (1984) and the nomenclature used in a recent compilation (Nagaswamy et al., 2000).

| | Proposed nomenclature | Saenger | Recent designation |
|---|---|---|---|
| 1. *Cis* Watson–Crick/Watson–Crick | G•A *cis* W.C./W.C. | VIII | GA Imino |
| | C•C *cis* W.C./W.C. (wobble) | | CC N3(+)-carbonyl, amino-N3 |
| | G•U *cis* W.C./W.C. (wobble) | XXVIII | |
| | U•C *cis* W.C./W.C. | XVIII | UC 4-carbonyl-amino |
| | U•U *cis* (wobble) W.C./W.C. | XVI | UU imino-carbonyl |
| 2. *Trans* Watson–Crick/Watson–Crick | A•U *trans* W.C./W.C. | XXI | AU Reverse Watson–Crick |
| | A•A *trans* W.C./W.C. | I | AA N1-amino, symmetric |
| | G•G *trans* W.C./W.C. | III | GG N1-carbonyl, symmetric |
| | G•C *trans* W.C./W.C. | XXII | GC Reverse Watson–Crick |
| | A•C *trans* W.C./W.C. | XXVI | AC Reverse Wobble |
| | G•U *trans* W.C./W.C. | XXVII | GU Reverse Wobble |
| | U•C *trans* W.C./W.C. | XVII | |
| | C•C *trans* W.C./W.C. | XIV, XV | |
| | U•U *trans* W.C./W.C. | XII, XIII | UU 4(2)-carbonyl-imino, symmetric |
| 3. *Cis* Watson–Crick/Hoogsteen | G•G *cis* W.C./Hoogsteen | VI | GG N1-carbonyl, N7-amino |
| | U•A *cis* W.C./Hoogsteen | XXIII | AU Hoogsteen |
| | G•A *cis* W.C./Hoogsteen | IX | GA N1-N7, carbonyl-amino |
| | A+•G *cis* W.C./Hoogsteen | | GA+ carbonyl-amino, N7-N1 |
| 4. *Trans* Watson–Crick/Hoogsteen | A•A *trans* W.C./Hoogsteen | V | AA N7-amino |
| | G•G *trans* W.C./Hoogsteen | VII | GG N7-imino |
| | U•A *trans* W.C./Hoogsteen | XXIV | AU Reverse Hoogsteen |
| | C•A *trans* W.C./Hoogsteen | XXV | AC Reverse Hoogsteen |
| 5. *Cis* Watson–Crick/Sugar-edge | A•G *cis* W.C./Sugar-edge | | GA N3-amino (1 bond) |
| | A•U *cis* W.C./Sugar-edge | | AU amino-2-carbonyl |
| 6. *Trans* Watson–Crick/Sugar-edge | A•G *trans* W.C./Sugar-edge | X | GA N3-amino, amino-N1 |
| | C•G *trans* W.C./Sugar-edge | | GC N3-amino, amino-N3 |
| 7. *Cis* Hoogsteen/Hoogsteen | | | |
| 8. *Trans* Hoogsteen/Hoogsteen | A•A *trans* Hoogsteen/Hoogsteen | II | AA N7-amino, symmetric |
| 10. *Trans* Hoogsteen/Sugar-edge | A•G *trans* Hoogsteen/Sugar-edge | XI | GA Sheared |
| | A•A *trans* Hoogsteen/Sugar-edge | | AA N3-amino |
| | C•U *trans* Hoogsteen/Sugar-edge | | UC 2-carbonyl-amino (1 bond) |
| 12. *Trans* Sugar-edge/Sugar-edge | G•G *trans* Sugar-edge/Sugar-edge | IV | GG N3-amino, symmetric |

indicated by single, black, capital letters (A, G, C, or U) as usual, except when the base adopts a *syn* conformation about the glycosidic bond, in which case the letter could be printed either bold or colored red. A red or dotted arrow may be drawn to indicate that a change in strand polarity occurs between two nucleotides. To designate canonical Watson–Crick and wobble pairs, one could use the symbols "–" for *both* AU and GC pairs and "●" for the wobble GU pair (Damberger & Gutell, 1994), but the convention "–" for AU pairs, "=" for GC pairs, and "○" for GU wobble pairs is more explicit (Michel et al., 1982) and allows the use of "●" as a generic designation for non-Watson–Crick pairs in text. Both conventions are noted in Figure 6.

Finally, we suggest a set of black-and-white symbols to accurately specify each kind of non-Watson–Crick edge-to-edge pairing interaction on a secondary structure diagram. We propose three symbols: *circles* for Watson–Crick edges, *squares* for Hoogsteen edges, and *triangles* for Sugar edges. The *cis* and *trans* orientations can be distinguished by *filled* and *open* symbols, respectively. When the same edge is used by the two bases, only one symbol is necessary (bp 1, 2, 7, 8, 11, and 12 in Fig. 6).

When an interaction involves two different edges, it is necessary to designate which edge corresponds to which base. For example, "AG *cis* Watson–Crick/Hoogsteen" designates a pair in which the Watson–Crick edge of the A interacts with the Hoogsteen edge of the G. To distinguish the XY and YX pairs in such cases in two-dimensional diagrams, we suggest using a horizontal line connecting the two symbols corresponding to the two interacting edges, as shown in Figure 6, for bp 3, 4, 5, 6, 9, and 10. In some situations it may be desirable to use a more compact symbol to designate an interaction. Thus, for each non-Watson–Crick pair we also propose compact symbols consisting of the symbol for one edge inside of the symbol for the other. The inner symbol is filled or open to designate *cis* and *trans*. A vertical line may be placed adjacent to the base interacting with the higher priority edge, following the convention discussed above.

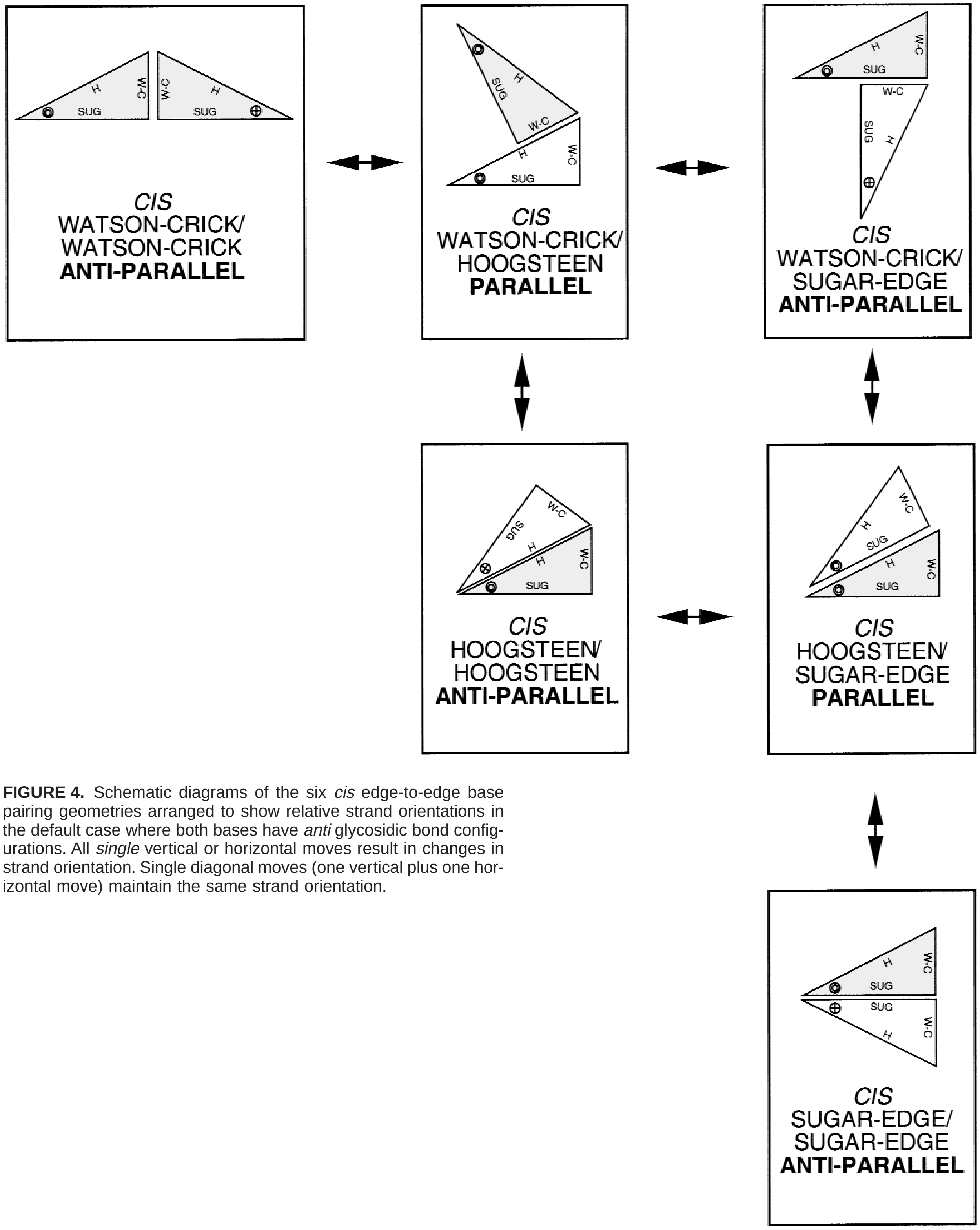


**FIGURE 4.** Schematic diagrams of the six *cis* edge-to-edge base pairing geometries arranged to show relative strand orientations in the default case where both bases have *anti* glycosidic bond configurations. All *single* vertical or horizontal moves result in changes in strand orientation. Single diagonal moves (one vertical plus one horizontal move) maintain the same strand orientation.

### *Asymmetry of the cis* Sugar-edge/ Sugar-edge pairs

The *cis* and *trans* W.C./W.C., the *trans* Hoogsteen/ Hoogsteen, and the *trans* Sugar-edge/Sugar-edge geometries (bp 1, 2, 8, and 12 in Fig. 6) are symmetric, with the interacting bases related by a twofold rotation about an axis passing either vertically or horizontally through the center of the base pair. The *cis* Sugar-edge/Sugar-edge geometry (bp 11 in Fig. 6), however,

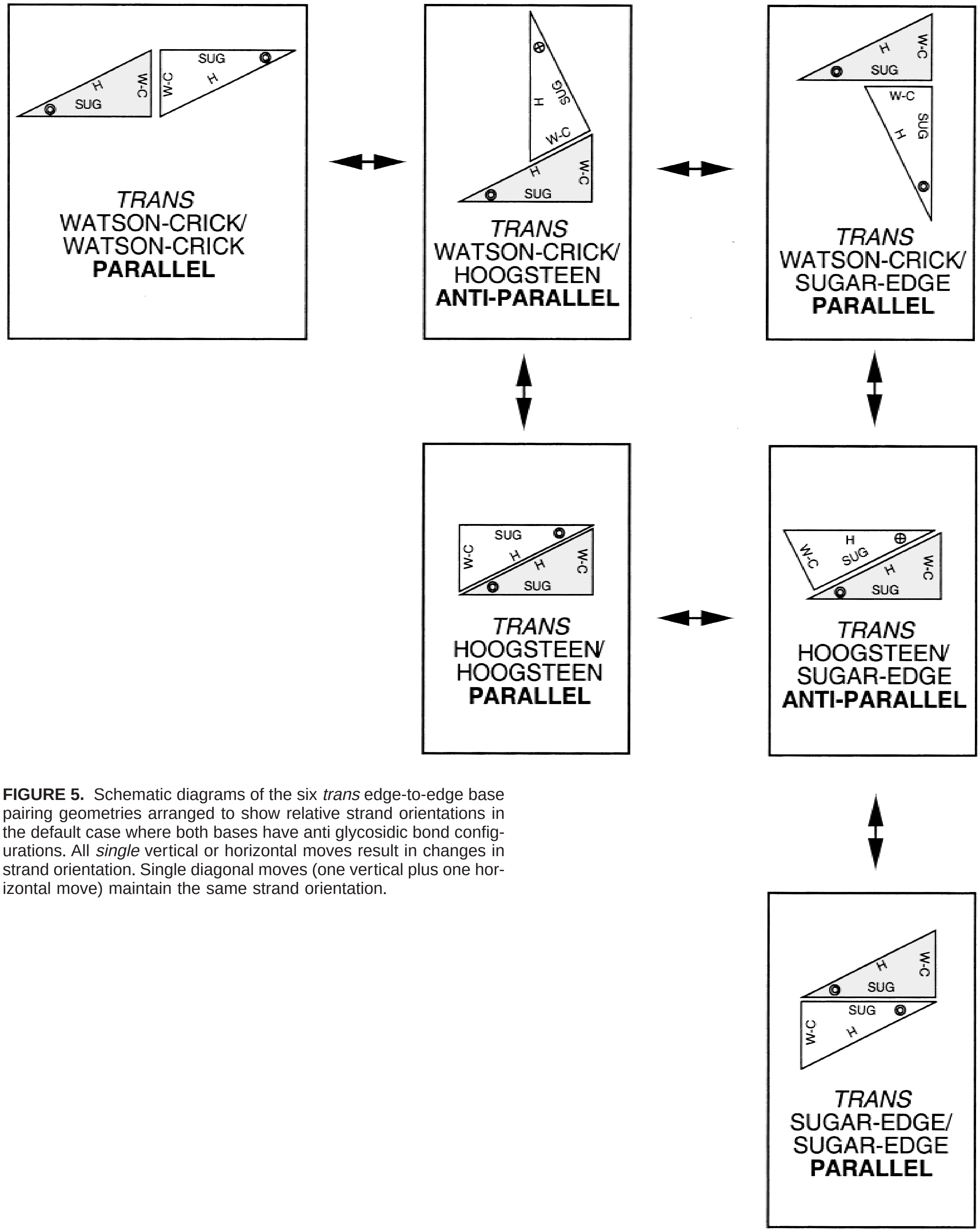


**FIGURE 5.** Schematic diagrams of the six *trans* edge-to-edge base pairing geometries arranged to show relative strand orientations in the default case where both bases have anti glycosidic bond configurations. All *single* vertical or horizontal moves result in changes in strand orientation. Single diagonal moves (one vertical plus one horizontal move) maintain the same strand orientation.

is not symmetric. To illustrate this point, two different A•G *cis* Sugar-edge/Sugar-edge pairs are shown in Figure 7. In these pairs, the 2′-OH of one of the nucleotides H bonds with both the 2′-OH and the base of the other nucleotide. The 2′-OH of the other nucleotide only H bonds with the 2′-OH of the first nucleotide. Thus, in the pair shown on the left in Figure 7, the 2′-OH of the adenosine H bonds to both the base and the 2′-OH of the guanosine, whereas in the pair shown on the right, the roles of the bases are reversed. For

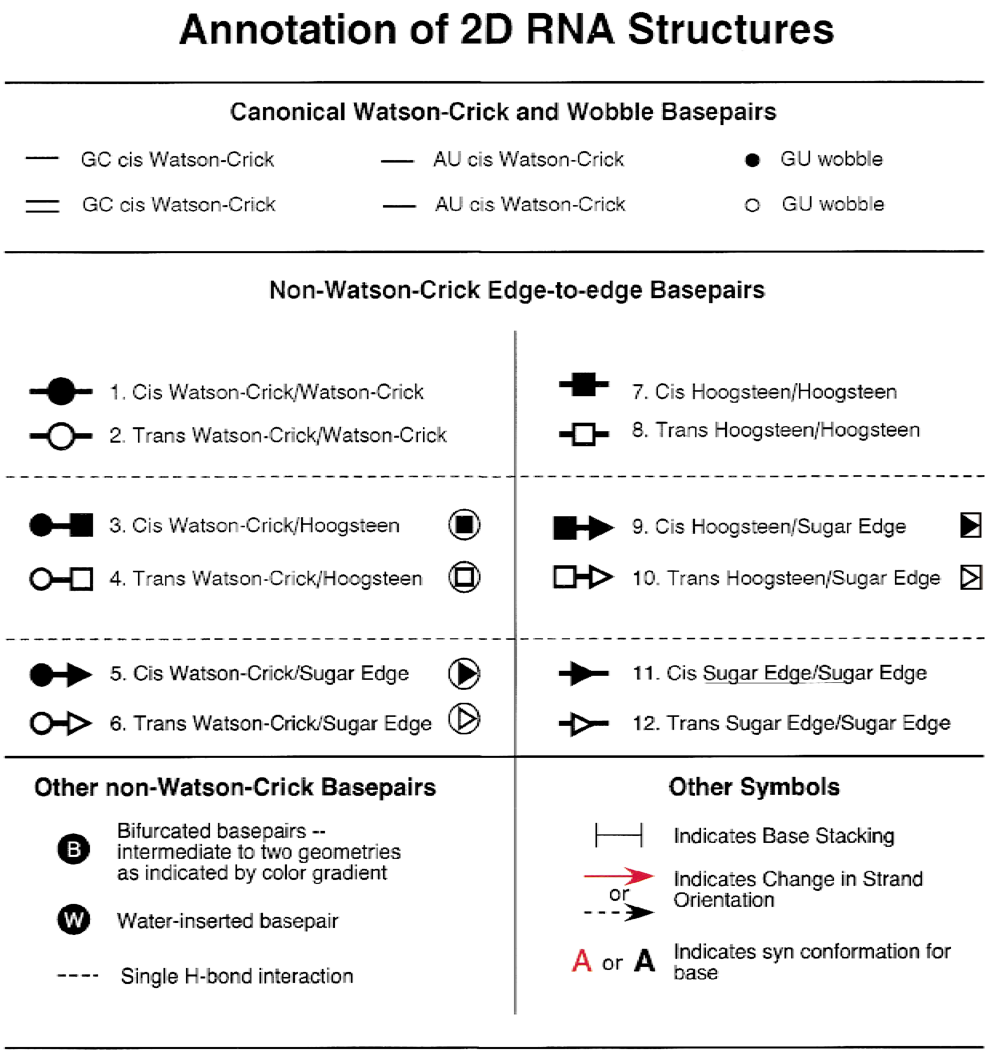

Annotation of 2D RNA Structures
Canonical Watson-Crick and Wobble Basepairs
GC cis Watson-Crick
AU cis Watson-Crick
GU wobble
GC cis Watson-Crick
AU cis Watson-Crick
GU wobble
Non-Watson-Crick Edge-to-edge Basepairs
1. Cis Watson-Crick/Watson-Crick
2. Trans Watson-Crick/Watson-Crick
7. Cis Hoogsteen/Hoogsteen
8. Trans Hoogsteen/Hoogsteen
3. Cis Watson-Crick/Hoogsteen
4. Trans Watson-Crick/Hoogsteen
9. Cis Hoogsteen/Sugar Edge
10. Trans Hoogsteen/Sugar Edge
5. Cis Watson-Crick/Sugar Edge
6. Trans Watson-Crick/Sugar Edge
11. Cis Sugar Edge/Sugar Edge
12. Trans Sugar Edge/Sugar Edge
Other non-Watson-Crick Basepairs
B
Bifurcated basepairs -- intermediate to two geometries as indicated by color gradient
W
Water-inserted basepair
Single H-bond interaction
Other Symbols
Indicates Base Stacking
or
Indicates Change in Strand Orientation
A or A
Indicates syn conformation for base

**FIGURE 6.** Suggested symbols for indicating tertiary interactions and other three-dimensional structural features in two-dimensional representations of RNA structures.

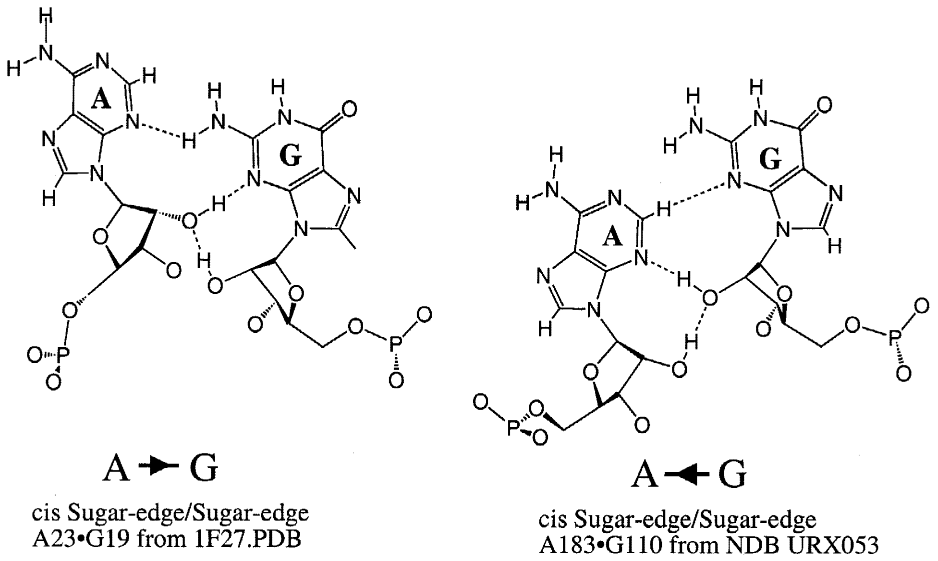
the pair on the left in Figure 7, the triangle is oriented to point to the G and vice versa for the pair on the right. Thus, the filled triangle, representing the *cis* Sugar-edge/Sugar-edge interaction, points away from the nucleotide that uses its 2′-hydroxyl to H bond to both the base and 2′-hydroxyl of the other nucleotide.

### Bifurcated and water-inserted base pairs

Most base–base interactions observed in high-resolution structures fit neatly into this classification framework. Pairs that feature bifurcated hydrogen bonds, however, are intermediate between two edge-to-edge geometries. The bifurcated pairs involve formally chelated (or three-centered) H bonds in which two H atoms point to a single acceptor atom; thus, they have been observed between the Watson–Crick edge of one base and one functional group of the second base.

Examples of bifurcated pairs that are intermediate to the canonical *cis* Watson–Crick/Watson–Crick and the *trans* Watson–Crick/Hoogsteen geometries are shown in Figure 8A. These are isosteric G•U and G•G pairs in which the exocyclic carbonyl oxygen atoms, UO4 or GO6, interact with the Watson–Crick edge of G (N1 and N2). They occur in loop E of bacterial 5S rRNA (Correll et al., 1997) and are isosteric to A•C and A•A, which covary with G•U and G•G in 5S sequences (Leontis & Westhof, 1998a). These pairs can be indicated in two-dimensional representations by a circle with the letter B inscribed, with white letters on black background as they are derived from the *cis* W.C./W.C. geometry (see Fig. 6).

A G•G pair having the bifurcated Hoogsteen geometry occurs in the 4.5 S RNA of the signal recognition particle RNA and is shown in Figure 8B (Batey et al., 2000; Jovine et al., 2000). In this pair, the N2 amino group of one G hydrogen bonds to the N7 and O6 acceptors of the other G. In this configuration, to ascertain that we are indeed dealing with a bifurcated H-bonded system would require high-resolution data;

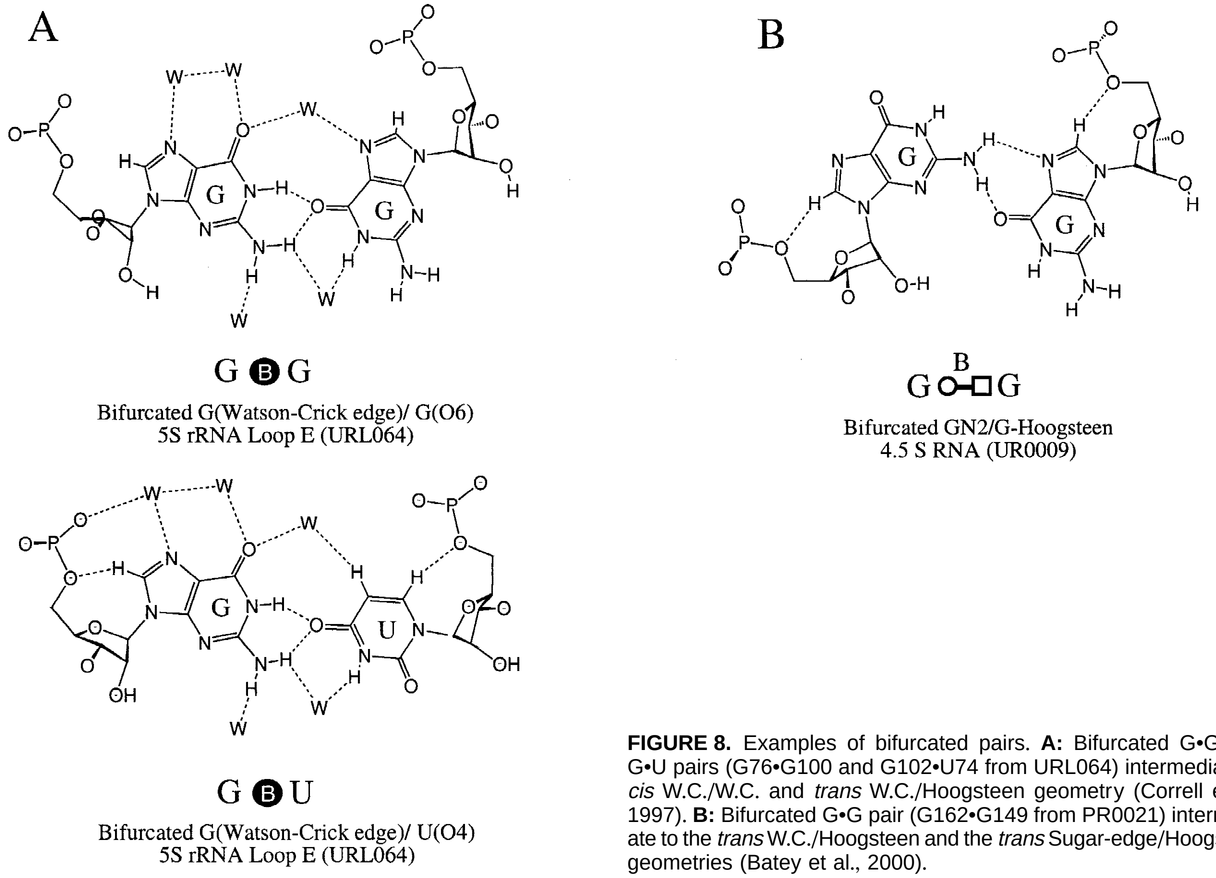


A➤G
cis Sugar-edge/Sugar-edge
A23•G19 from 1F27.PDB

A◀G
cis Sugar-edge/Sugar-edge
A183•G110 from NDB URX053

**FIGURE 7.** Two different A•G *cis* Sugar-Edge/Sugar-edge pairs. The triangle points from the nucleotide having the 2′-OH that H bonds to both the base and 2′-OH of the other nucleotide. This nucleotide is A23 in the pair from 1F27.PDB (left panel) and G110 in the pair from URX053 (right panel).

therefore, by analogy and to indicate that the pair involves unusual geometries we suggest extending the use of “bifurcated.” This pair is intermediate between the *trans* W.C./Hoogsteen and the *trans* Sugar-edge/Hoogsteen geometries and is therefore designated by adding a B to the symbol for *trans* W.C./Hoogsteen.



G Ⓑ G
Bifurcated G(Watson-Crick edge)/ G(O6)
5S rRNA Loop E (URL064)

G Ⓑ U
Bifurcated G(Watson-Crick edge)/ U(O4)
5S rRNA Loop E (URL064)

B
G○-□G
Bifurcated GN2/G-Hoogsteen
4.5 S RNA (UR0009)

**FIGURE 8.** Examples of bifurcated pairs. **A:** Bifurcated G•G and G•U pairs (G76•G100 and G102•U74 from URL064) intermediate to *cis* W.C./W.C. and *trans* W.C./Hoogsteen geometry (Correll et al., 1997). **B:** Bifurcated G•G pair (G162•G149 from PR0021) intermediate to the *trans* W.C./Hoogsteen and the *trans* Sugar-edge/Hoogsteen geometries (Batey et al., 2000).

Water-inserted pairs have been observed in several high-resolution structures, as recently reviewed (Leontis & Westhof, 1998c). They often result from an opening of a regular type geometry by a rotation of one base with respect to the other and insertion of one water molecule (see, e.g., Fig. 3 of Westhof & Fritsch, 2000). We propose that these be designated using the letter W inscribed white on black or black on white depending on whether the interaction is *cis* or *trans* (see Fig. 6). Pairs in which the inserted water molecule replaces a hydrogen bond in a *cis* pair are designated *cis* and likewise for *trans*.

## Examples of two-dimensional representations of RNA tertiary structure

To illustrate these conventions, we present in Figure 9 examples of two-dimensional representations of RNA motifs with tertiary interactions added. The left panel shows the loop E of bacterial 5S rRNA from NDB file URL064 (Correll et al., 1997). All bases of this symmetric "internal loop," in fact, are paired. A104•G72 comprise a *trans* Hoogsteen/Sugar-edge pair. This is designated using an open symbol (indicating the *trans* geometry) comprising a square, placed next to A104 (for the Hoogsteen edge), connected to a triangle, placed next to G72 (for the Sugar-edge). The same interaction occurs between A78 and G98, but the orientation is reversed, with the Hoogsteen base, A78, on the right. The symbols we propose make these relationships immediately clear. U103•A73 and U77•A99 are *trans* W.C./Hoogsteen pairs, and are indicated by open symbols comprising circles (placed next to the Us) connected to squares (placed next to the As). In the U103•A73 pair, the Watson–Crick base (U103) occurs on the left, whereas the situation is reversed for the U77•A99 pair. G102•U74 and G76•G100 are isosteric *cis* bifurcated pairs, intermediate between the *cis* W.C./W.C. and the *trans* W.C./Hoogsteen geometry. These interactions are indicated by black circles with white B inscribed. A101•G75 is a water-inserted *cis* W.C./W.C. pair. Thus, it is designated by a black circle with a W superimposed. This representation reveals that the bacterial loop E motif in fact comprises two isosteric submotifs oriented in opposite (palindromic) directions.

### *Sarcin/ricin motif from large ribosomal subunit*

The next example (middle panel, Fig. 9) is the highly conserved sarcin/ricin motif (Leontis & Westhof, 1998b). This motif also occurs in loop E of eukaryal 5S rRNA and should not be confused with bacterial loop E. The sequence shown is that of rat 28S rRNA, NDB file UR0002 (Correll et al., 1998). The structure comprises a GAGA hairpin loop (not shown) and an asymmetric "internal loop." The dotted arrows between C8 and A9 and between A9 and G10 indicate the local strand reversal that occurs at A9. The positioning of A9 beneath U11 indicates the stacking between these two residues. The "bulged" base, G10, is actually hydrogen bonded to U11 and lies in the same plane as the U11•A20 *trans* W.C./Hoogsteen pair. This is indicated by placing all three bases at the same horizontal level on the page. The G10•U11 pair is *cis* Sugar-edge/Hoogsteen whereas the G19•A12 and U7•C23 pairs are *trans* Sugar-edge/Hoogsteen.

### *Domain IV of SRP 4.5S RNA*

The SRP motif has been observed as the RNA alone (Jovine et al., 2000) and in complex to SRP protein 54

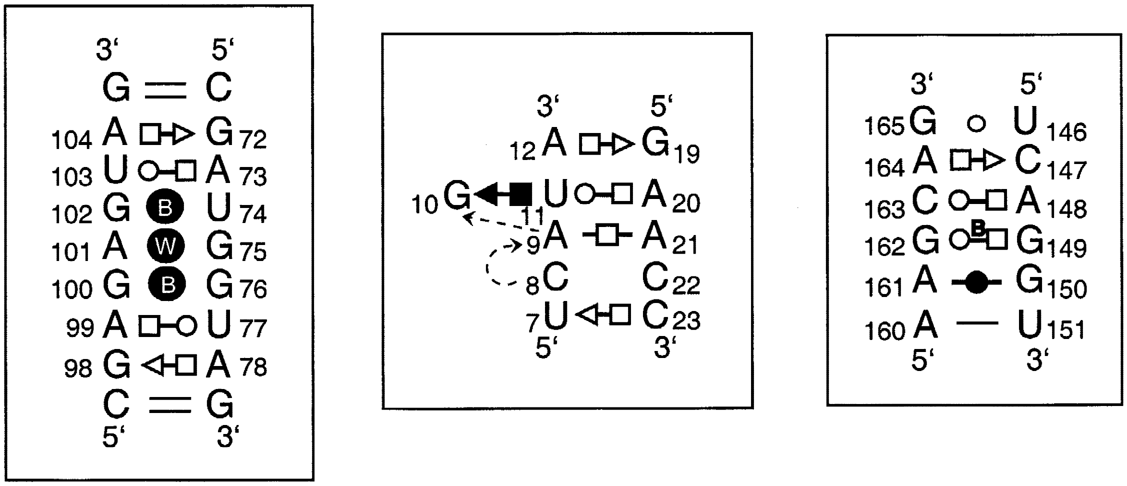


**FIGURE 9.** Left panel: two-dimensional representation of the tertiary structure of loop E of bacterial 5S rRNA (NDB file URL064). Center: two-dimensional representation of the tertiary structure of the sarcin/ricin (S-turn) motif of bacterial 23S rRNA (NDB file UR0002). Right panel: two-dimensional representation of the tertiary structure of the internal loop of Domain IV of the SRP 4.5 S RNA (NDB files PR0021 and UR0009).

(Batey et al., 2000). As shown in the right panel of Figure 9, this symmetric internal loop is very similar to the submotifs of the bacterial loop E motif (Fig. 9, left panel). The SRP motif comprises a *trans* Hoogsteen/Sugar-edge A•C pair adjoining a *trans* W.C./Hoogsteen C•A pair followed by a *trans* bifurcated G•G pair and a *cis* W.C./W.C. A•G pair. The trans Hoogsteen/Sugar-edge A•C pair corresponds to the A•G pair in the loop E submotif and is isosteric to it (Fig. 10, left). The *trans* W.C./Hoogsteen C•A pair corresponds to the U•A pair in loop E (Fig. 10, right). The *cis* W.C./W.C. A•G corresponds to the water-inserted A•G in the loop E motif, which is also *cis* W.C./W.C., with an H-bond between AN6 and GO6 and the water molecule bridging the imino nitrogens. The bifurcated G•G in the SRP differs slightly from the pair in loop E, as shown above in Figure 8. The loop E submotif occurs also in helix 20 of 16S rRNA (Wimberly et al., 2000), as was predicted (Leontis & Westhof, 1998a). Interestingly, the G•G bifurcated pair in 16S rRNA is identical to the pair in the SRP loop (*trans* bifurcated as in Fig. 8B).

## Recognition of motif similarity in annotated three-dimensional structures

Because the classification facilitates the comparison between different three-dimensional structures to identify common three-dimensional motifs, it further aids in predicting families of isosteric pairings that can substitute for each other in homologous RNA molecules. Since three-dimensional structures of homologous RNA molecules are more strongly conserved than their individual sequences, covariation data can be used to identify bases involved in tertiary interactions and even indicate the most likely pairing geometry. This approach was successfully applied for predicting potential sarcin-ricin motifs (also frequently referred to as "S-turn" or "eukaryal 5S loop E" motifs) and bacterial loop E motifs in 16S and 23S rRNAs (Leontis & Westhof, 1998a, 1998b). All these motifs, except for one, were later identified in crystal structures of the ribosome 70S and its subunits (Cate et al., 1999; Nissen et al., 2000; Schluenzen et al., 2000; Wimberly et al., 2000). In ad-

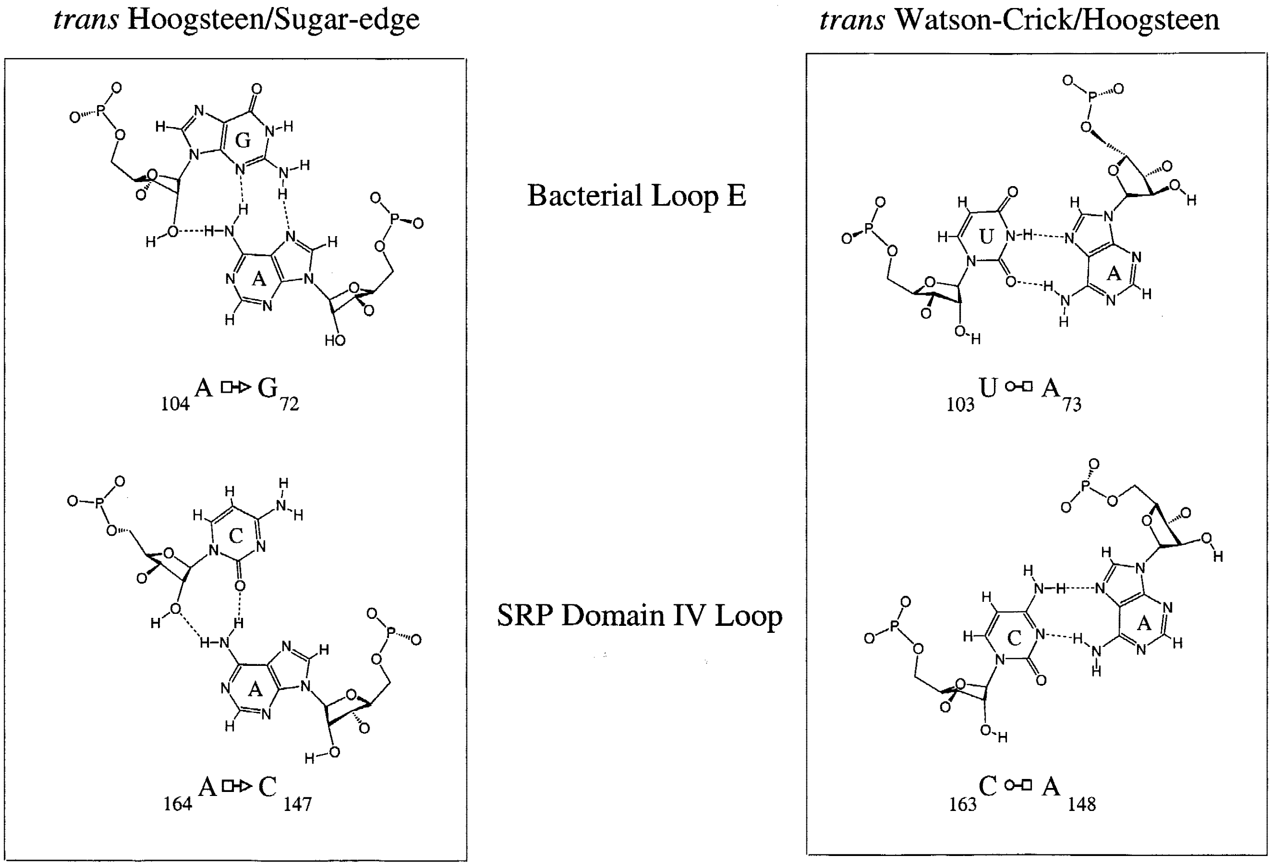


**FIGURE 10.** Comparison of isosteric base pairs in bacterial loop E (URL064) and the internal loop of Domain IV SRP 4.5S RNA (Correll et al., 1997; Batey et al., 2000). Left panel: *trans* Hoogsteen/Sugar edge pairs A104•G72 from loop E and A164•C147 from 4.5 S RNA. Right panel: *trans* Watson–Crick/Hoogsteen pairs U103•A73 from loop E and C163•A148 from 4.5 S RNA.

dition, a bacterial loop E motif, predicted to occur in domain IV of the 4.5S RNA in the signal recognition particle (Leontis & Westhof, 1998a), was later observed by X-ray crystallography (Batey et al., 2000; Jovine et al., 2000). Further, used in conjunction with experimental evidence, motif prediction, based on phylogeny and sequence-specific criteria, can be applied to structure prediction of RNA domains. Recently, such a method combining motif recognition with the NMR signature attached to the three-dimensional structure, led to the rapid identification of a sarcin/ricin (i.e., *eukaryal* 5S loop E) motif in a domain of the IRES element in the hepatitic C virus (Klinck et al., 2000).

## CONCLUSIONS

The proposed nomenclature and classification provides a succinct and coherent way to communicate RNA structural information in oral and written presentations. Moreover, it facilitates the two-dimensional representation of complex three-dimensional structures. Thus, we also propose conventions that present the essential three-dimensional features of RNA structures in a visually accessible and appealing two-dimensional format, including: (1) all canonical and non-Watson–Crick pairs, (2) changes in strand polarity in the folding of the RNA, (3) the occurrence of *syn* bases, and (4) essential stacking interactions. The added information incorporated in two-dimensional representations of RNA molecules helps in recognizing and memorizing similarities between motifs.

## MATERIALS AND METHODS

This work relied on visual examination of high-resolution X-ray crystal structures to determine hydrogen-bonding patterns. Structures were obtained from the Nucleic Acid Database, http://ndbserver.rutgers.edu/NDB, and the Protein Data Bank, http://www.rcsb.org/pdb/, and were manipulated with the Swiss PDB Viewer program, available from http://www.expasy.ch/spdbv/ (Guex & Peitsch, 1997). Hydrogen-bonding diagrams were prepared using the Chem3D and ChemDraw Pro programs (CambridgeSoft Corporation). Diagrams were prepared using Appelworks and Canvas.

## ACKNOWLEDGMENTS

We are indebted to Pascal Auffinger, Luc Jaeger, and François Michel for numerous discussions, and Carl Correll, François Major, Steve Harvey, Wilma Olson, Muttaya Sundaralingam, and Brian Wimberly for suggestions regarding motif and base-pair nomenclature. We thank Jesse Stombaugh for assistance with figure preparation (JS was supported by National Science Foundation-Research Experiences for Undergraduates grant CHE-9732463). This work was supported by National Institutes of Health Grant 2R15 GM55898 to NBL. EW wishes to thank the Institut Universitaire de France for support.

*Received December 13, 2000; returned for revision December 28, 2000; revised manuscript received January 24, 2001*

## REFERENCES

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution [see comments]. *Science 289*:905–920.

Batey RT, Rambo RP, Doudna JA. 1999. Tertiary motifs in RNA structure and folding. *Angew Chem Int Ed Engl 38*:2326–2343.

Batey RT, Rambo RP, Lucast L, Rha B, Doudna JA. 2000. Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science 287*:1232–1239.

Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science 273*:1678–1684.

Cate JH, Yusupov MM, Yusupova GZ, Earnest TN, Noller HF. 1999. X-ray crystal structures of 70S ribosome functional complexes [see comments]. *Science 285*:2095–2104.

Correll CC, Freeborn B, Moore PB, Steitz TA. 1997. Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell 91*:705–712.

Correll CC, Munishkin A, Chan YL, Ren Z, Wool IG, Steitz TA. 1998. Crystal structure of the ribosomal RNA domain essential for binding elongation factors. *Proc Natl Acad Sci USA 95*:13436–13441.

Crick FH. 1966. Codon–anticodon pairing: The wobble hypothesis. *J Mol Biol 19*:548–555.

Damberger SH, Gutell RR. 1994. A comparative database of group I intron structures. *Nucleic Acids Res 22*:3508–3510.

Ferré-D'Amaré AR, Doudna JA. 1999. RNA folds: Insights from recent crystal structures. *Annu Rev Biophys Biomol Struct 28*:57–73.

Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis 18*:2714–2723.

Hermann T, Patel DJ. 1999. Stitching together RNA tertiary architectures. *J Mol Biol 294*:829–849.

Jovine L, Hainzl T, Oubridge C, Scott WG, Li J, Sixma TK, Wonacott A, Skarzynski T, Nagai K. 2000. Crystal structure of the Ffh and EF-G binding sites in the conserved domain IV of *Escherichia coli* 4.5S RNA. *Struct Fold Des 8*:527–540.

Klinck R, Westhof E, Walker S, Afshar M, Collier A, Aboul-Ela F. 2000. A potential RNA drug target in the hepatitis C virus internal ribosomal entry site. *RNA 6*:1423–1431.

Lavery R, Zakrzewska K, Sun JS, Harvey SC. 1992. A comprehensive classification of nucleic acid structural families based on strand direction and base pairing. *Nucleic Acids Res 20*:5011–5016.

Leontis NB, Westhof E. 1998a. The 5S rRNA loop E: Chemical probing and phylogenetic data versus crystal structure. *RNA 4*:1134–1153.

Leontis NB, Westhof E. 1998b. A common motif organizes the structure of multi-helix loops in 16S and 23S ribosomal RNAs. *J Mol Biol 283*:571–583.

Leontis NB, Westhof E. 1998c. Conserved geometrical base-pairing patterns in RNA. *Quart Rev Biophysics 31*:399–455.

Leontis NB, Westhof E. 1999. Recurrent RNA motifs: Analysis at the base pair level. In: Barciszewki J, Clark BFC, eds. *RNA biochemistry and biotechnology*. Boston: Kluwer Academic Publishers. pp 45–61.

Masquida B, Westhof E. 2000. On the wobble GoU and related pairs. *RNA 6*:9–15.

Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie 64*:867–881.

Nagaswamy U, Voss N, Zhang Z, Fox GE. 2000. Database of non-canonical base pairs found in known RNA structures. *Nucleic Acids Res 28*:375–376.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosome activity in peptide bond synthesis [see comments]. *Science 289*:920–930.

Saenger W. 1984. *Principles of nucleic acid structure.* New York: Springer Verlag.

Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A. 2000. Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell 102*:615–623.

Su L, Chen L, Egli M, Berger JM, Rich A. 1999. Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nat Struct Biol 6*:285–292.

Sundaralingam M. 1977. Non-Watson-Crick base pairs in ribonucleic acids. *Int J Quant Chem: Quant Biol Symp 4*:11–23.

Varani G, McClain WH. 2000. The G•U wobble pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports 1*:18–23.

Westhof E. 1992. Westhof's rule [letter]. *Nature 358*:459–460.

Westhof E, Dumas P, Moras D. 1988. Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr A 44*:112–123.

Westhof E, Fritsch V. 2000. RNA folding: Beyond Watson–Crick pairs. *Struct Fold Des 8*:R55–R65.

Wimberly BT, Brodersen DE, Clemons WM Jr, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit [In Process Citation]. *Nature 407*:327–339.