J Comput Aided Mol Des (2006) 20:629–645
DOI 10.1007/s10822-006-9083-x

ORIGINAL PAPER

Conformational specificity of non-canonical base pairs and higher order structures in nucleic acids: crystal structure database analysis

Shayantani Mukherjee · Manju Bansal · Dhananjay Bhattacharyya

Received: 29 June 2006 / Accepted: 26 September 2006 / Published online: 24 November 2006

Abstract Non-canonical base pairs contribute immensely to the structural and functional variability of RNA, which calls for a detailed characterization of their spatial conformation. Intra-base pair parameters, namely propeller, buckle, open-angle, stagger, shear and stretch describe structure of base pairs indicating planarity and proximity of association between the two bases. In order to study the conformational specificities of non-canonical base pairs occurring in RNA crystal structures, we have upgraded NUPARM software to calculate these intra-base pair parameters using a new base pairing edge specific axis system. Analysis of base pairs and base triples with the new edge specific axis system indicate the presence of specific structural signatures for different classes of non-canonical pairs and triples. Differentiating features could be identified for pairs in *cis* or *trans* orientation, as well as those involving sugar edges or C–H-mediated hydrogen bonds. It was seen that propeller for all types of base pairs in *cis* orientation are generally negative, while those for *trans* base pairs do not have any preference. Formation of a base triple is seen to reduce propeller of the associated base pair along with reduction of overall flexibility of the pairs. We noticed that base pairs involving sugar edge are generally more non-planar, with large propeller or buckle values, presumably to avoid steric clash between the bulky sugar moieties. These specific conformational signatures often provide an insight into their role in the structural and functional context of RNA.

Keywords Base pair conformation · RNA structure · Non-Watson-Crick · Intra-base pair parameters · Propeller twist

S. Mukherjee · D. Bhattacharyya (✉)
Biophysics Division, Saha Institute of Nuclear Physics,
1/AF Bidhannagar, Kolkata 700064, India
e-mail: dhananjay.bhattacharyya@saha.ac.in

M. Bansal
Molecular Biophysics Unit, Indian Institute of Science,
Bangalore 560012, India

D. Bhattacharyya
Center for Applied Mathematics and Computational
Science, Saha Institute of Nuclear Physics,
1/AF Bidhannagar, Kolkata 700064, India

Introduction

The diverse role of RNA in several cellular processes ranging from gene expression to enzymatic functions often helps us to get a glimpse of its structural and functional variability [1]. A thorough structural insight into the complicated regime of secondary and tertiary interactions occurring in RNA would facilitate a better understanding of its versatile biological functions. It has been known for long that, the regular stretches of A-form RNA are often defined by contiguous canonical Watson–Crick base pairs, formed through hydrogen bonds between complementary bases adenine–uracil and guanine–cytosine. Such regular A-form helices exhibit deep narrow major grooves and shallow wide minor grooves in contrast to a B-DNA structure consisting of wide major and narrow minor grooves. This characteristic architecture of RNA double helices results in the inaccessibility of discriminatory major groove edges and greater access of the comparatively uniform and shallow minor groove edges of canonical pairs, which fails to reveal the basis for structural

variability and molecular recognition in RNA. On the other hand, non-canonical base pairs occurring within secondary structural blocks, can impart variation in groove edges leading to potential interaction sites for specific RNA–RNA and RNA–protein recognition, and are thus of immense importance [2].

It has been reported earlier that apart from Watson–Crick edge, a base in RNA can also undergo pairing through Hoogsteen or the sugar edges [3–7]. Some of the bases can also be protonated due to local environmental stress and can form non-canonical pairs with other regular bases [8]. Several computational studies of non-canonical base pairs in RNA have led to the development of various types of characterization techniques and nomenclatures [7–17]. Primary goal of these methods is to detect all canonical and non-canonical base pairs occurring in nucleic acid three-dimensional structures considering three hydrogen-bonding edges of purines and pyrimidines and some protonated forms of bases. However, many of these methods detect base pairs barely stabilized by a single hydrogen bond or water/ion-mediated hydrogen bonds, whose strength of interaction remains questionable. Probably these are not very important in formation of RNA secondary structure, while it is expected that a good non-canonical base pair stabilized by at least two hydrogen bonds may play a significant role. It was suggested earlier by Saenger [18] that at least two hydrogen bonds are needed to hold a base pair as a planar entity. Recent study on quantum chemical geometry optimization of two bases also showed that in most cases these leads to base pairs with two hydrogen bonds [19–20]. Several methods have been developed to predict RNA secondary structure from nucleotide sequence using various experimental data. Mathews et al. had attempted to predict RNA secondary structure considering only the Watson–Crick base pairing criteria and number of hydrogen bonds in GC and AU base pairs, although accuracy of prediction is improved by incorporating other experimental data [21]. A recent study based on statistical frequencies of occurrences of different base pair stacks in RNA crystal structures could predict secondary structures with high accuracy [21]. However none of these studies considered the frequently occurring non-canonical base pairs and their stacking, although consideration of such base pairs can lead to significant improvement in RNA secondary structure prediction.

We have recently developed a new methodology (BPFIND) for detection of base pairs in RNA three-dimensional structures, which are stabilized by at least two direct hydrogen bonds of $N-H\cdots O/N$ or $C-H\cdots O/N$ types [17]. However, while there have been several attempts to detect the wide range of non-canonical base pairs, there has been no effort toward characterizing conformational features of these non-canonical base pairs in RNA. This lacuna is surprising, more so considering that, base pairs in RNA very often deviate from the ideal planar geometry of a canonical base pair and a qualitative description, only in terms of hydrogen-bonding edges or orientations does not describe the conformation in totality. Hence, a detailed structural analysis will help to reveal conformational specificities of several classes of non-canonical pairs, thereby highlighting their role in maintaining the complicated folded structure of RNA. Such a study would also help in understanding the basis for structural flexibility of RNA that often leads to functional diversity in terms of various receptor binding or ligand docking sites. There are previous reports on DNA base pair analysis, showing that changes in local geometry affect the stacking interaction thereby leading to changes in free energies of stacking [22]. Previous studies on both proteins and DNA have successfully derived knowledge-based energy functions from conformational flexibilities [23–25]. Thus, understanding base pair conformation and loss or gain of flexibility associated with different types of base pairs in RNA structures will facilitate in evaluating the entropy factor related to folding and formation of complicated motifs.

Following IUPAC-IUB convention the spatial arrangement of one base with respect to the other in a pair can be quantitatively defined with the help of three rotational and three translational intra-base pair parameters, namely; buckle, propeller, open-angle, shear, stagger and stretch [26]. Among these six parameters; shear, stretch and open-angle relate directly to the hydrogen-bonding pattern and proximity, while buckle, propeller and stagger describe the overall non-planarity of a base pair compared to the ideal coplanar geometry. By definition, buckle and shear values undergo sign reversal when they are calculated from the opposite sense, i.e., buckle and shear of an A:U base pair and the same pair converted to U:A have identical values but with opposite signs. This happens due to the presence of a pseudo-dyad symmetry axis along the short axis, perpendicular to the base pair normal and base pair long axis [18, 26]. In a similar manner, pseudo-dyad axis of a *trans* base pair lies along the base pair normal, making open and stagger as anti-symmetric parameters. These intra-base pair parameters have a direct resemblance to the three-dimensional conformation of a base pair, i.e., shear indicates sliding of one base with respect to the other in the base pair plane; stagger indicates out of plane motion of one base with respect to the other; stretch

indicates separation of the two bases relating to hydrogen-bonding distance; buckle indicates the amount of cusp formation; open indicates the angle between the two bases on the base pair plane and propeller is the twisting motion of the two bases about the base pair long axis [26]. Thus, implications about the extent of deformation compared to an ideal planar geometry of a base pair can be easily visualized from the parameter values. Conformational preference for negative propeller of base pairs in DNA structures is well established, which is the main feature of Calladine's rule dictating sequence directed DNA double helical structure [27]. Moreover larger propeller twist of A:T base pairs in DNA has been correlated to narrow minor groove width of A-tract DNA, which makes these regions suitable targets of most DNA binding antibiotics [28]. Such systematic study of base pair parameters of non-canonical base pairs in RNA three-dimensional structures, however, has not been reported. Here we have analyzed conformational features of the frequently occurring non-canonical base pairs in RNA using intra-base pair parameters. Our first attempt to determine their structural features by the widely used software, 3DNA [29], indicated its inadequacy for most base pair types, due to usage of Watson–Crick *cis* (canonical) base pair specific axis system. We have, therefore, defined a new base pair edge specific axis system and upgraded the NUPARM software to calculate the base pair parameters using this axis system. These follow the IUPAC-IUB guideline in all respects and the parameters are efficient in describing conformational characteristics of any base pair type, i.e., one can visualize the base pair geometry from these parameter values, along with the base pair type. Apart from containing various types of non-canonical pairs, three-dimensional structures of RNA are known to exhibit enormous intricacy with folds, loops, turns and knots interspersing stretches of regular double helical secondary structures. Higher order structures like base triples formed between a canonical or non-canonical base pair and a distant single stranded nucleotide often define and stabilize such complicated structural motifs in RNA [30–36]. We have also attempted to understand their structural signatures, which are often found to differ from the structures of the constituent base pairs.

Methods

Preparation of data set

The RNA structures have been selected from Protein Data Bank [37] solved by X-ray crystallography at 3.5 Å or better resolution, as available on April 2005 and described in our earlier paper on detection of canonical and non-canonical base pairs [17]. Only structures with more than 30 nucleotides (including RNA–protein complexes) are considered, as we are interested in studying the importance of naturally occurring non-Watson–Crick pairs in functional RNA only, which are generally large macromolecules. We have not considered the synthetic oligonucleotide duplex structures of RNA or RNA/DNA strands for this analysis. The data set of RNA structures is inclusive of all functional RNA molecules and an additional filter depending on sequence homology has been omitted. This has been purposefully done in order to include the maximum possible number of non-canonical base pair types in the respective data set. Moreover, in spite of sequence homology, changes in microenvironment often leads to discernable changes in local base pair geometry and omission of homology related sequences would not allow an in-depth study of such small conformational alterations. However, in some cases, consideration of all homology related RNA structures led to overemphasis of certain conformational features giving incorrect central tendencies. In those cases we have recalculated the parameters after omitting redundant structures.

All types of canonical or non-Watson–Crick type base pairs and base triples were identified using our recently developed BPFIND software [17]. Here we considered the three distinct hydrogen-bonding edges (viz. Watson–Crick, Hoogsteen and sugar) of the bases guanine, adenine, cytosine and uracil that can form base pairs in two types of orientation of the glycosidic bonds with respect to the pseudo-hydrogen bond axis, i.e., *cis* and *trans*. It should also be noted that BPFIND only detects pairs stabilized by at least two hydrogen bonds involving N, O or C heavy atoms of nucleotide base moieties and O2′ atoms in ribose sugars. We have selected base pairs by BPFIND, which have at least two direct hydrogen bonds between them with donor–acceptor distance less than 3.8 Å and pseudo-angles less than 120°. When two bases are found to be oriented in such a way that only protonated form of one of them can form two hydrogen bonds with the above criterion, we have considered the protonated pair. Examples of each type of non-canonical pair occurring in RNA three-dimensional structures along with their frequency of occurrence can be found at our website: http://www.saha.ac.in/biop/bioinformatics.html. Many RNA structures, especially tRNA, contain chemically modified bases such as 5MeC, PsU, DHU, etc., which are treated just like their standard counterparts while identifying base pairs formed by them [21]. For

example, DHU is considered as uracil while, 7MG is considered as guanine.

We have adopted a nomenclature to designate different types of base pair edges that is consistent with previous reports on various non-canonical base pair types [7, 17, 38]. Here a base can form a base pair involving 'W' edge (Watson–Crick edge), 'H' edge (Hoogsteen edge) or 'S' edge (sugar edge) of a normal base. Additionally, to distinguish protonated edges, we have marked them by '+' or 'z' for Watson–Crick or sugar edge, respectively. We have further distinguished base pairs stabilized by at least one $C-H\cdots N/O$ type hydrogen bond by marking them with lower case letters, 'w', 'h', and 's' for Watson–Crick, Hoogsteen and sugar edges, respectively. Thus, now the uppercase 'W', 'H' or 'S' letters indicate the respective base pair edges involving only $N-H\cdots N/O$ hydrogen bonds. To maintain concordance with earlier reports [17, 38], a specific base pair type is denoted by indicating the single letter codes for each of the two bases, followed by edges of the two bases involved in hydrogen bonding and finally stating the orientation of the pair, which can be either *cis* or *trans*. For example, A:U H:W *cis* denotes a base pair formed between Hoogsteen edge of Adenine and Watson–Crick edge of Uracil and involving $N-H\cdots N/O$ type of hydrogen bonds, while A:C w:s *cis* denotes a base pair formed between Watson–Crick edge of adenine and sugar edge of cytosine and involving one $C-H\cdots N/O$ type of hydrogen bond. Statistical analysis has been carried out only for the pairs with frequency of occurrence greater than 100 in the RNA crystal structure database.

We have also chosen very high-resolution (2.0 Å or better) double helical DNA structures, without any bound ligand or chemical modifications for calculating the intra-base pair parameters of regular Watson–Crick base pairs. The PDB-IDs of the selected DNA structures (30 A-DNA and 10 B-DNA), as obtained from NDB (April 2005) [39], are 118D, 126D, 137D, 138D, 160D, 1BNA, 1D13, 1D62, 1D78, 1D79, 1EHV, 221D, 240D, 243D, 251D, 260D, 272D, 295D, 2D94, 2D95, 307D, 317D, 348D, 349D, 368D, 369D, 371D, 395D, 396D, 399D, 414D, 440D, 441D, 7BNA, 9BNA, 9DNA, ADH010, ADH029, ADH034 and BDJ061. The double helical RNA and RNA/DNA hybrid structures, with 2.0 Å or better resolution, analyzed here are 157D, 161D, 165D, 168D, 1CSL, 1D4R, 1D88, 1D96, 1D9H, 1DQH, 1EVP, 1FUP, 1ICG, 1ID9, 1IDW, 1IHA, 1IK5, 1J9h, 1JZV, 1KD5, 1KFO, 1LNT, 1OSU, 1RXB, 1ZEV, 216D, 217D, 255D, 259D, 2AOP, 315D, 332D, 354D, 377D, 393D, 394D, 398D, 406D, 413D, 421D, 464D, 468D, 469D, 470D, 472D and 479D.

Results

Need for a new definition of intra-base pair parameters of non-canonical base pairs

Three rotational and three translational parameters are required for describing relative orientation of one rigid body with respect to another, as in the case of a base pair. Hence, relative orientations of two mutually orthogonal axis systems defined for two bases involved in pair formation, would provide all six intra-base pair parameters, viz., buckle, propeller, open-angle, shear, stagger and stretch. There are several excellent methodologies and softwares for calculation of intra-base pairs parameters, such as NEWHELIX, CURVES, NUPARM, 3DNA, etc. [28, 39–52]. Most of these were developed decades ago, time tested and widely used for calculation of intra- and inter-base pair parameters of double helical DNA. Among these, the 3DNA package distributed by NDB is the most recent and it can recognize non-canonical base pairs in RNA apart from the canonical pairs of DNA double helix [29]. Thus, we have attempted to use this software to quantitatively evaluate structural features of different types of non-canonical base pairs in RNA. The nomenclature that we have followed to designate different types of canonical and non-canonical base pairs, following earlier reports [7, 17], makes use of a single letter code for each of the two bases, followed by a description of the edges of the two bases involved in hydrogen bonding and finally stating the orientation of the pair (see Sect. 'Methods' for detail). One expects that these parameters will highlight two aspects of a base pair geometry: (1) quality of hydrogen bonds forming the base pairs in terms of its deviation from the ideal geometry and (2) relative orientation of the bases with respect to each other. Both these properties would essentially indicate the strength of association of the base pairs and hence will highlight the role of such pairs in RNA fold formation and recognition.

The parameters calculated by 3DNA software provide excellent informative values for canonical W:W *cis* base pairs (G:C W:W *cis* of Table 1). The IUPAC-IUB guidelines regarding sign convention is also followed exactly for W:W *cis* base pair parameters calculated by 3DNA. Furthermore, statistical analysis of base pair parameters of high-resolution DNA crystal structures calculated by 3DNA also highlights their sequence dependent conformational features (Table 2). In order to understand the efficiency of 3DNA in calculating conformational parameters of non-canonical base pairs, we have generated ideal base pairs of various kinds having near perfect

Table 1 Intra-base pair parameters of perfectly planar base pairs generated with optimum hydrogen bonds

Base pair type	Software used	Buckle	Open-angle	Propeller	Stagger	Shear	Stretch
G:C W:W *cis*	NUPARM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.8 (2.8)
	3DNA	0.0 (0.0)	–0.4 (–0.4)	0.0 (0.0)	0.0 (0.0)	–0.1 (0.1)	–0.2 (–0.2)
A:U H:W *cis*	NUPARM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.8 (2.8)
	3DNA	0.0 (0.0)	68.6 (–68.6)	0.0 (0.0)	0.0 (0.0)	0.7 (–0.7)	–3.7 (3.7)
A:U H:W *trans*	NUPARM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.8 (2.8)
	3DNA	0.0 (0.0)	–97.3 (–97.3)	0.0 (0.0)	0.0 (0.0)	–4.2 (4.21)	–2.1 (–2.1)
G:A S:H *trans*	NUPARM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	2.0 (2.0)	3.3 (3.3)
	3DNA	0.0 (0.0)	–17.4 (–17.4)	0.0 (0.0)	0.0 (0.0)	6.9 (–6.9)	–4.8 (–4.8)
G:G S:S *trans*	NUPARM	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	1.5 (1.5)	3.3 (3.3)
	3DNA	0.0 (0.0)	–180.0 (180.0)	0.0 (0.0)	0.0 (0.0)	–3.1 (3.1)	–7.9 (7.9)

A specific base pair type is denoted by indicating the single letter code for each of the two bases, followed by edges of the two bases involved in hydrogen bonding and finally stating the orientation of the pair. Base pairing through Watson–Crick, Hoogsteen and sugar edges are denoted by 'W', 'H' and 'S', respectively. Values in parenthesis are the parameters calculated for the same base pair in the reverse direction (see text for detail)

hydrogen bonds between the bases (Fig. 1) and calculated their intra-base pair parameters by 3DNA (Table 1). On alteration of the order of bases in these base pairs, 3DNA reported values with reversal of sign for most of the intra-base pair parameters, even for the A:U H:W *cis* base pair, and thus do not conform to the IUPAC-IUB nomenclature (i.e., only buckle and shear should change sign for *cis* pairs). Following the pseudo-dyad symmetry axis aligned perpendicular to the base pair planes for the *trans* base pairs, only the open and stagger values are expected to undergo sign reversal, but 3DNA values do not follow these criteria (Table 1). Possibly, usage of the W:W *cis* base edge specific axis system for all types of non-canonical pairs gives rise to such discrepancies. Moreover, the values of open-angle, shear and stretch as calculated by 3DNA for non-canonical pairs do not corroborate the conformation of the pairs. For example, the stretch values show high negative values for ideal planar base pairs with good hydrogen bonds, while they should be ~3.0, in order to depict the true separation of the bases. Similarly, open-angle values also show high negative or positive values and high shear values, which have no direct correlation with the actual sliding displacement of one base with respect to the other in the base pair plane. We have therefore adopted a base pair edge specific axis system for calculating equivalent parameters for these pairs.

Edge specific axis system for canonical/non-canonical base pairs

The IUPAC-IUB convention recommended the *X*-axis along the pseudo-dyad axis of the base pair, the *Y*-axis along the long axis of the base pair and the *Z*-axis along the base pair normal [26]. Most of the

Table 2 Average and standard deviation (shown in parenthesis) of intra-base pair parameters in DNA crystal structures

	A-DNA						B-DNA					
	G:C (226)			A:T (44)			G:C (61)			A:T (45)		
	Average (SD) by NUPARM	Average (SD) by 3DNA	ρ	Average (SD) by NUPARM	Average (SD) by 3DNA	ρ	Average (SD) by NUPARM	Average (SD) by 3DNA	ρ	Average (SD) by NUPARM	Average (SD) by 3DNA	ρ
Buckle (°)	–0.3 (9.6)	–0.3 (8.1)	0.97	0.0 (7.2)	0.0 (8.2)	0.95	–1.3 (7.6)	0.2 (8.1)	0.58	0.5 (3.9)	0.8 (3.8)	0.68
Open (°)	0.1 (2.7)	0.1 (2.3)	0.76	2.1 (3.6)	–0.3 (4.1)	0.94	–2.0 (2.8)	–2.1 (2.1)	0.57	4.7 (3.3)	2.1 (3.3)	0.77
Propeller (°)	–10.2 (5.3)	–11.1 (4.8)	0.94	–10.8 (4.9)	–10.6 (3.9)	0.95	–9.9 (6.5)	–9.8 (6.2)	0.65	–13.8 (4.5)	–13.7 (4.6)	0.96
Stagger (Å)	–0.1 (0.2)	0.0 (0.2)	0.90	–0.1 (0.2)	0.1 (0.2)	0.81	0.0 (0.2)	0.1 (0.2)	0.63	–0.1 (0.2)	0.0 (0.2)	0.94
Shear (Å)	0.0 (0.2)	0.0 (0.3)	0.91	0.0 (0.1)	0.0 (0.1)	0.87	0.0 (0.2)	0.0 (0.2)	0.79	0.1 (0.2)	0.1 (0.2)	0.89
Stretch (Å)	2.8 (0.1)	–0.2 (0.1)	0.91	2.8 (0.1)	–0.2 (0.1)	0.90	2.8 (0.2)	–0.2 (0.2)	0.91	2.7 (0.1)	–0.2 (0.1)	0.93

Intra-base pair parameters calculated by the proposed hydrogen-bonding edge specific axis system and 3DNA (23) for canonical A:T and G:C base pairs in the high-resolution X-ray crystal structures of DNA oligonucleotides are shown. The correlation coefficients (ρ) between the values calculated using the proposed edge specific reference frame and the standard Watson–Crick reference frame adopted in 3DNA are also listed. The frequencies of base pairs are given in parenthesis beside each type

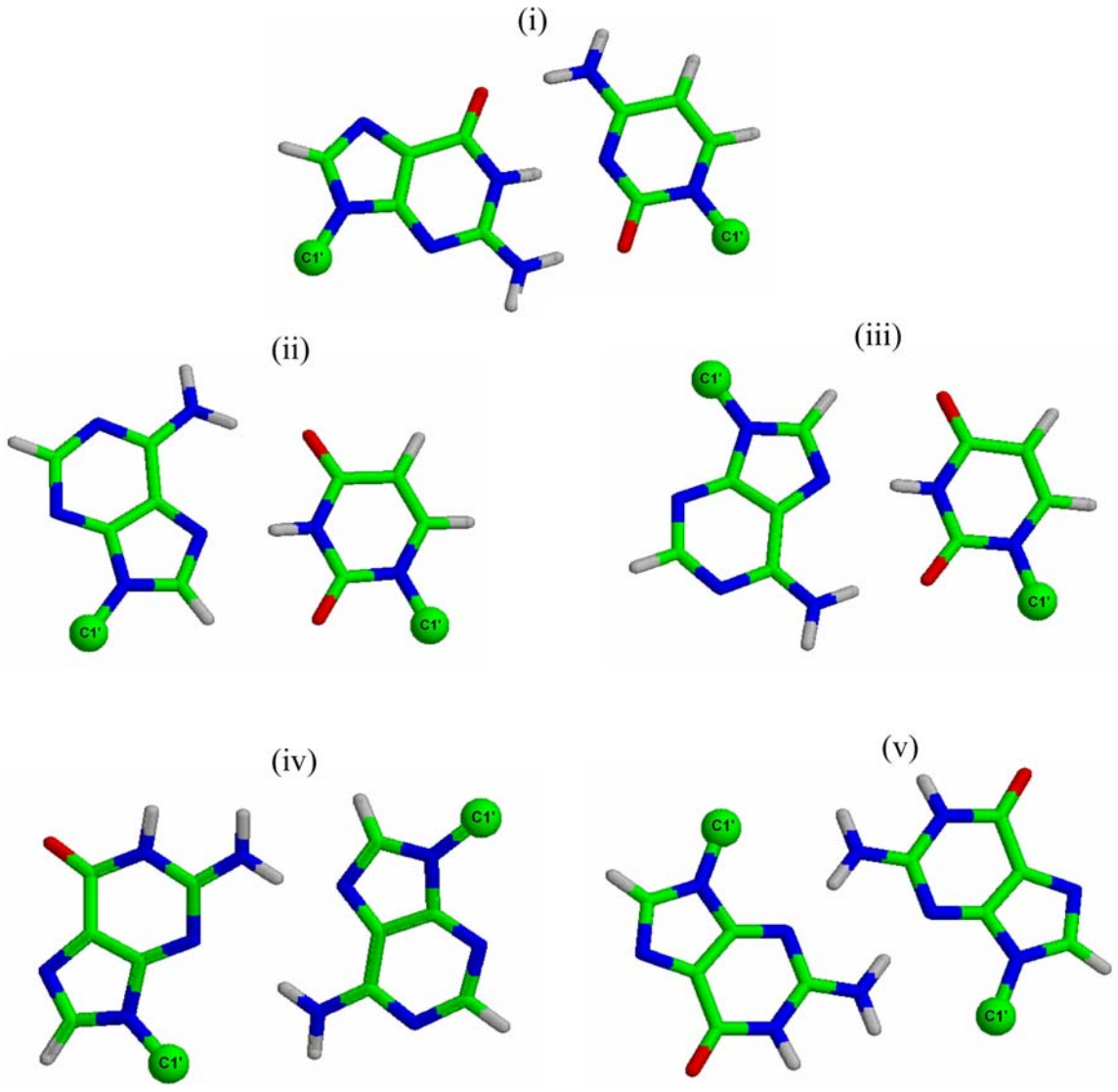


Fig. 1 Planar base pairs with optimum hydrogen bonds (i) G:C W:W *cis*, (ii) A:U H:W *cis*, (iii) A:U H:W *trans*, (iv) G:A S:H *trans* and (v) G:G S:S *trans*

base pair parameter determination softwares including previous version of NUPARM adopt this definition. Although such an axis system is quite powerful in describing inter-base pair or local doublet parameters (tilt, roll, twist, etc.), it is inadequate for calculating correct intra-base pair parameters. We have defined the axis systems for the two bases in a pair in accordance with the hydrogen-bonding edge of the bases involved. Primarily, the *X*-axis of the base is fixed perpendicular to the best mean plane through the base ring atoms, which can be either along or opposite to the $5' \rightarrow 3'$ strand direction. The *Y*-axis of each base in a pair is defined involving two hydrogen-bonding heavy atoms of the specific edge forming the hydrogen bonds. For example, *Y*-axis for W edge of purine points from N1 to N6/O6, that for H edge points from N7 to N6/O6 while the same for S edge points from C1′ to N3. Similarly, the *Y*-axis for W edge of pyrimidine is the vector from N3 to N4/O4, that for H edge points from C5 to N4/O4 and that for S edge points from C1′ to O2. We have chosen C1′ to define *Y*-axis for the S edge rather than the hydrogen-bonding heavy atom O2′, as C1′ lies in the base plane. Finally, the *Z*-axis of a base is fixed perpendicular to both *X*- and *Y*-axis following a right-handed axis system and it lies approximately parallel to the hydrogen bonds formed in the pair (Fig. 2). Following the above procedure, two sets of edge specific base axes are fixed for the bases involved

in a pair formation. It is thus ensured that, for base pairs in *cis* orientation, the *Y*-axis of each base is directed away from the C1′ atom (i.e., the *Y*-axis always points toward the major groove) and the *Z*-axes of both the bases (e.g., U1 and U2 in Fig. 3a) are directed toward the other base (U2 of Fig. 3a) of the pair. When the calculated *Z*-axis does not meet this criterion, directions of both *X*- and *Z*-axes are reversed, hence ambiguity in defining the direction of *X*-axis is corrected here. In the *trans* orientation, both *X*- and *Y*-axes of the first base are reversed again (Fig. 3b).

Calculation of intra-base pair parameters and regeneration of base pairs

The inter-base pair rotational and translational parameters (tilt, roll, twist, slide, shift and rise) for a base pair step are conventionally calculated in NUPARM using base pair specific axis systems for the two base pairs forming a doublet [49], while the intra-base pair parameters (propeller, buckle, open-angle, shear, stagger and stretch) are now calculated using base edge specific axis systems for the two bases forming a pair. The mathematical expression used in calculating buckle is analogous to that used for calculating tilt, while the expressions for open-angle, propeller, shear, stagger and stretch are identical to those used for calculating roll, twist, slide, shift and rise, respectively, but with the axes for the bases rather than the base pairs. Thus,

$$\text{buckle}\,(\kappa) = -2\sin^{-1}(\boldsymbol{Zm}\cdot\boldsymbol{Y}_1)$$

$$\text{open - angle}\,(\sigma) = -2\sin^{-1}(\boldsymbol{Zm}\cdot\boldsymbol{X}_1)$$

$$\text{propeller}\,(\pi) = \cos^{-1}((\boldsymbol{X}_1\times\boldsymbol{Zm})\cdot(\boldsymbol{X}_2\times\boldsymbol{Zm}))$$

$$\text{stagger}\,(S_x) = -\boldsymbol{Ym}\times\boldsymbol{M}$$

$$\text{shear}\,(S_y) = \boldsymbol{Xm}\cdot\boldsymbol{M}$$

$$\text{stretch}\,(S_z) = \boldsymbol{Zm}\cdot\boldsymbol{M}$$

where $\mathbf{X}_1$, $\mathbf{Y}_1$ and $\mathbf{Z}_1$ are unit vectors along the axes of the first base, $\mathbf{X}_2$, $\mathbf{Y}_2$ and $\mathbf{Z}_2$ are those for the second base. The components of the mean unit vector **Xm**, **Ym** and **Zm** are calculated as follows:

$$\boldsymbol{Xm} = (\boldsymbol{X}_1 + \boldsymbol{X}_2)/|(\boldsymbol{X}_1 + \boldsymbol{X}_2)|$$

$$\boldsymbol{Ym} = (\boldsymbol{Y}_1 + \boldsymbol{Y}_2)/|(\boldsymbol{Y}_1 + \boldsymbol{Y}_2)|$$

$$\boldsymbol{Zm} = \{(\boldsymbol{X}_1+\boldsymbol{X}_2)\times(\boldsymbol{Y}_1+\boldsymbol{Y}_2)\}/|(\boldsymbol{X}_1+\boldsymbol{X}_2)\times(\boldsymbol{Y}_1+\boldsymbol{Y}_2)|$$

The vector **M** is obtained by joining two base atoms, one from each base of the pair, chosen according to the hydrogen-bonding edge of that particular base,

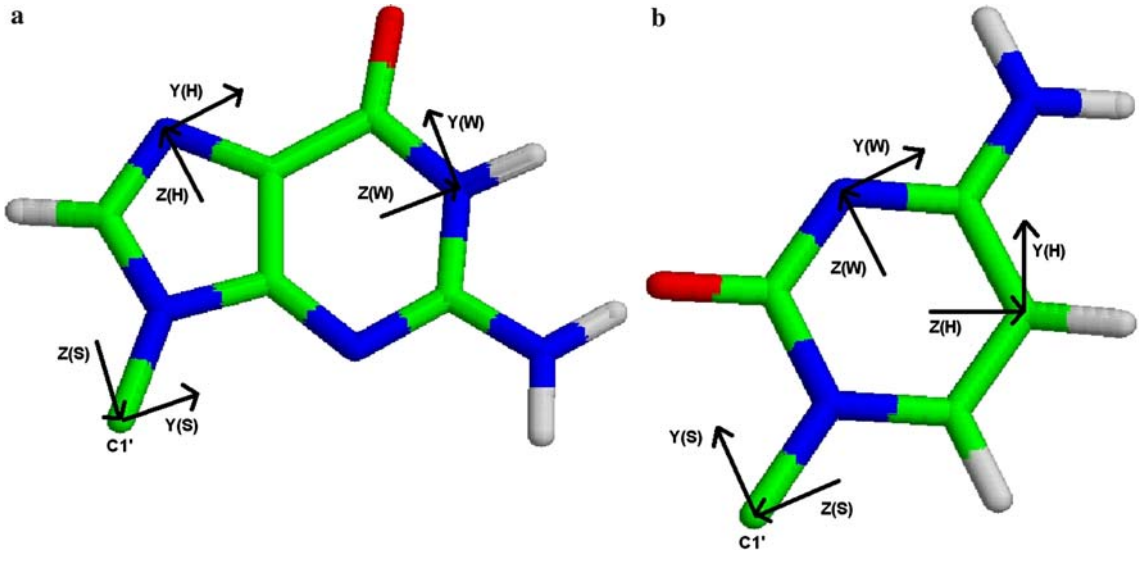


Fig. 2 Definitions of *Y*- and *Z*-axes for W, H and S edges of (a) a purine (represented by guanine) and (b) a pyrimidine base (represented by cytosine)

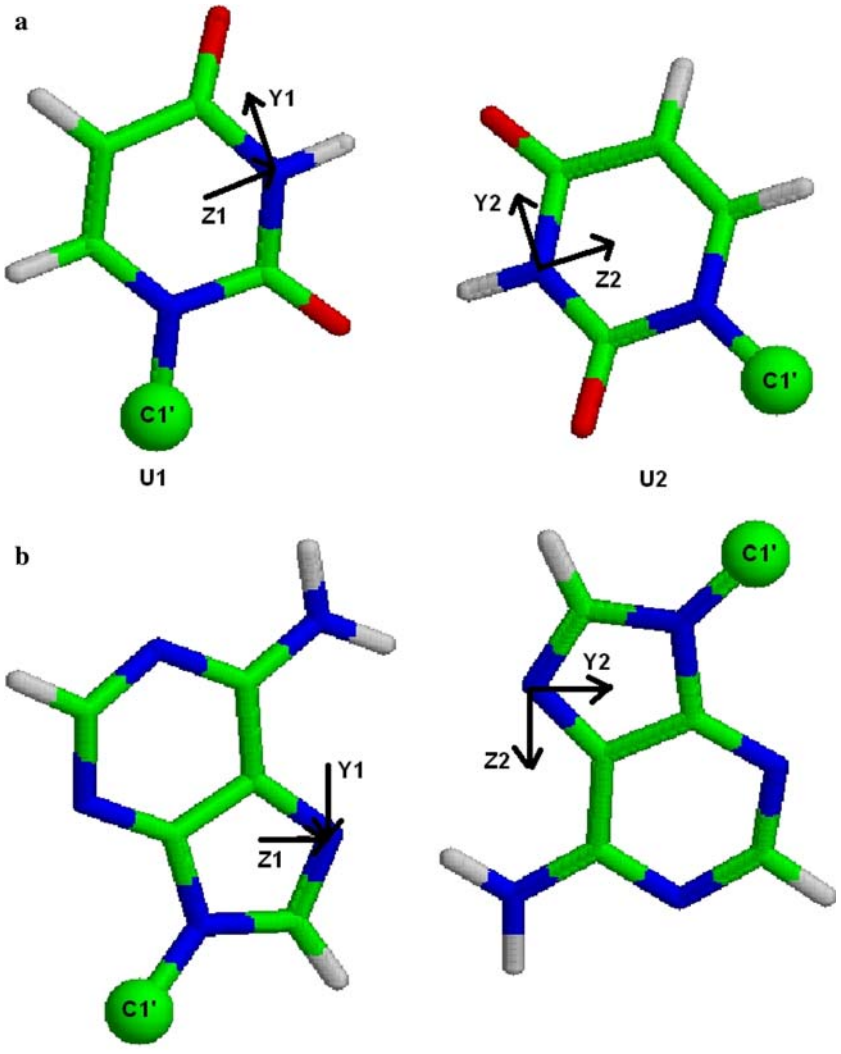


Fig. 3 Definitions of *Y*- and *Z*-axes for representative base pairs (a) U:U W:W *cis* pair observed between residue numbers 26 (chain id—0) and 517 (0) in 50S ribosomal unit (PDB ID—1FFK) and (b) A:A H:H *trans* pair observed between residue numbers 2691 (0) and 2703 (0) in 50S ribosomal unit (PDB ID—1QVG)

e.g., N1 of purine and N3 of pyrimidine for the W edge, N7 of purine or C5 of pyrimidine for the H edge and N3 of purine or C1′ of pyrimidine for the S edge.

Calculation of intra-base pair parameters using the above definitions has an extra advantage that values of rotational and translational parameters in local helical or cylindrically symmetric frame can be obtained from them following our previous derivations [50]. Thus the rotational parameters in local helix axis frame, viz. helical propeller (π_h), helical buckle (κ_h) and helical open (σ_h) can be deduced from values of buckle (κ), open-angle (σ) and propeller (π) using the following relations:

$$\pi_h = \sin^{-1}[\sin^2(\pi/2)\cos^2\{\sin^{-1}(R+T)^{1/2} + R + T\}]^{1/2}$$

$$\kappa_h = \sin^{-1}[(1/2T)\cot^2(\pi h/2)\{(1-R-T) - \{(1-R-T)^2 - 4RT\}^{1/2}\}]^{1/2}$$

$$\sigma_h = \sin^{-1}[(1/2T)\cot^2(\pi h/2)\{(1-R-T) - \{(1-R-T)^2 - 4RT\}^{1/2}\}]^{1/2}$$

where $R = \sin^2(\sigma/2)$ and $T = \sin^2(\kappa/2)$.

Similarly, the three local helical translational parameters (S_{hx}, S_{hy} and S_{hz}) can be deduced from values of stagger (S_x), shear (S_y), stretch (S_z) and the rotational parameters, viz. π_h, κ_h and σ_h using the following relations:

$$S_{hx} = 2B1\cos(\kappa_h)\sin(\pi_h) + 2B3\sin(\kappa_h)$$

$$S_{hy} = -2B2\ \cos(-\sigma_h)sin(\pi_h) + 2B3\sin(-\sigma_h)$$

$$\begin{aligned} S_{hz} &= -2B1\cos(-\sigma_h)\sin(\pi_h)\sin(\pi_h) \\ &\quad + 2B2\cos(\kappa_h)\sin(-\sigma_h)\sin(\pi_h) \\ &\quad + B3(1+\cos\pi_h)\cos(\kappa_h)\cos(-\sigma_h) \end{aligned}$$

where,

$$B1 = S_y[2\{1+\cos(\pi_h) + \{1-\cos(\pi_h)\sin2(-\sigma_h)\}\}]^{1/2}$$

$$B2 = S_x[2\{1+\cos(\pi_h) + \{1-\cos(\pi_h)\sin2(\kappa_h)\}\}]^{1/2}$$

$$\begin{aligned} B3 &= S_z[1+\cos(\pi_h) + \{1-\cos(\pi_h)\sin^2(-\sigma_h)\}] \\ &\quad [1+\cos(\pi_h) + \{1-\cos(\pi_h)\sin^2(\kappa_h)\}] \end{aligned}$$

Regeneration of a base pair with a given set of values of intra-base pair parameters can be done by: (1) rotation by κ_h about *Y*-axis and σ_h about *X*-axis performed on both the bases, (2) translation of S_{hx} along *X*-axis and S_{hy} along *Y*-axis performed on both the bases and (3) helical transformation, i.e., rotation by π_h and translation by S_{hz} along *Z*-axis, to one of them. Accuracy of this procedure for regeneration of base paired doublets has been discussed earlier [50] and a similar study for non-canonical base pairs is beyond the scope of the present paper and will be presented elsewhere.

In case of the canonical type base pairs with W edges of both bases in *cis* orientation, as expected the values of buckle and shear undergo a sign reversal when calculated from the opposite direction. Similar sign reversal is also seen for all types of base pairs involving sugar or Hoogsteen edges in *cis* orientation. However, in *trans* orientation the values of open-angle and stagger appear with sign reversal when calculated with reference to the second strand (Table 1). This difference is due to the change in axis of symmetry for *trans* orientation, i.e., instead of *Y* being along pseudo-dyad symmetry axis in the case of *cis* base pairs, *X* is the axis of the pseudo-dyad symmetry in case of *trans* base pairs.

Intra-base pair parameters for canonical Watson–Crick base pairs in DNA X-ray crystal structures

The base pair parameters for the representative planar base pairs, considering the new axis definition, are tabulated in Table 1. The parameters for the canonical G:C W:W *cis* represents the base pair conformation to a high accuracy, the only difference with 3DNA being the value of stretch, we calculate it as 2.8 Å rather than 0. In essence, our definition of stretch provides a direct quantitative measure of the separation of hydrogen-bonding base heavy atoms. In order to assess the efficiency of the edge specific axis system in defining canonical structures of A:T and G:C pairs, we have calculated the intra-base pair parameters for pairs in regular A-DNA and B-DNA crystal structures (Table 2). The average values of propeller for A:T and G:C base pairs differ significantly in B-DNA structures, while these are very similar in A-DNA structures. The large values of propeller for A:T base pairs in B-DNA is hypothesized to give rise to distinct sequence dependent effects, such as narrow minor groove in AT rich regions, which are absent in A-DNA. The open-angle values for A:T and G:C base pairs in B-DNA structures also differ widely, while these do not show such large variations in A-DNA crystal structures. Further, the values obtained by NUPARM (using edge specific axis system) have been compared with those obtained by 3DNA [29]. Results indicate that values of all six intra-base pair parameters calculated with NUPARM exhibit very high correlation coefficients with those calculated by 3DNA. Thus it can be deduced that the new edge specific axis system effectively quantifies the spatial arrangement of canonical base pairs in a manner similar to that of 3DNA. The intra-base pair parameters of canonical base pairs in RNA and RNA/DNA oligonucleotide structures (Table 3) are also quite similar to those in DNA.

Intra-base pair parameters of non-canonical pairs in RNA crystal structures

Values of the intra-base pair parameters of non-canonical pairs calculated by NUPARM with edge specific axis definition describe their true planar conformation (Fig. 1). All representative non-canonical pairs show 0 open-angle and about 2.8 Å stretch in accordance with their three-dimensional structure. It is evident from the figure of four non-canonical base pairs, that two base pairing arrangements (A:U H:W *cis* and *trans*) do not have any shear, while bases in other two arrangements are sheared with respect to one another (Fig. 1). The extent of shear is correctly assessed considering the new axis definition. Furthermore, the IUPAC-IUB suggestions of sign reversal are also diligently followed when parameters are calculated from the reverse directions. Thus, observing the performance of the edge specific axis system in describing non-canonical base pair structure, we have

Table 3 Mean and standard deviations (within parenthesis) of intra-base pair parameters of base pair types observed frequently in RNA crystal structures are tabulated for *cis* pairs and *trans* pairs

Base pair type	Frequency	Buckle (°) (SD)	Open (°) (SD)	Propeller (°) (SD)	Stagger (Å) (SD)	Shear (Å) (SD)	Stretch (Å) (SD)
cis pairs							
Canonical A:U W:W *cis* and G:C W:W *cis*	28,455	0.3 (11.6)	1.8 (4.5)	–8.0 (8.6)	–0.2 (0.4)	0.0 (0.3)	2.8 (0.1)
A:U W:W *cis*	6,861	–0.4 (10.0)	3.7 (5.1)	–8.4 (8.8)	–0.1 (0.4)	0.0 (0.3)	2.8 (0.1)
G:C W:W *cis*	21,614	0.6 (12.1)	1.3 (4.1)	–7.9 (8.6)	–0.2 (0.4)	0.0 (0.3)	2.9 (0.1)
A:U W:W (oligonucleotide)	75	0.0 (6.9)	3.5 (4.4)	–12.9 (5.4)	–0.0 (0.2)	0.0 (0.2)	2.8 (0.1)
G:C W:W (oligonucleotide)	252	0.1 (8.1)	0.0 (2.3)	–9. 8 (5.4)	–0.1 (0.2)	0.0 (0.2)	2.9 (0.1)
Non-Watson–Crick *cis* (involving only N – H···N/O hydrogen bonds)	4,536	0.9 (13.1)	–0.3 (8.4)	–9.0 (10.2)	–0.2 (0.5)	–1.6 (1.4)	2.8 (0.2)
A:G W:W	404	4.3 (18.8)	3.0 (9.1)	–9.4 (13.4)	–0.4 (0.5)	0.2 (0.5)	2.8 (0.2)
G:U W:W	2,769	0.1 (9.3)	0.5 (5.6)	–8.1 (7.6)	–0.2 (0.4)	–2.3 (0.4)	2.8 (0.2)
U:U W:W	360	0.5 (12.1)	–2.2 (6.5)	–14.8 (9.1)	–0.2 (0.6)	–2.3 (0.4)	2.9 (0.2)
A:C + :W	184	8.5 (15.7)	5.5 (9.5)	–12.5 (10.8)	–0.4 (0.4)	–2.2 (1.3)	2.8 (0.2)
A:U H:W	201	–3.0 (17.3)	2.7 (6.1)	–10.0 (11.0)	–0.1 (0.5)	0.0 (0.4)	2.8 (0.2)
G:G H:W	158	–5.3 (15.4)	–1.6 (6.4)	–6.5 (14.3)	0.0 (0.7)	3.0 (0.4)	2.9 (0.1)
Non-Watson–Crick *cis* (involving C – H···N/O hydrogen bonds)	586	–6.9 (26.4)	8.8 (33.2)	–7.0 (21.0)	–0.2 (0.8)	0.8 (1.5)	3.1 (0.8)
A:C w:s	143	–5.8 (21.4)	–31.9 (6.0)	–22.1 (13.9)	–0.7 (0.5)	0.1 (0.4)	4.2 (0.2)
A:G w:s	197	–21.7 (24.8)	49.0 (8.2)	–3.0 (24.6)	0.2 (0.9)	2.1 (0.4)	2.8 (0.2)
C:C + :s	113	4.0 (19.7)	–33.6 (5.2)	–28.6 (11.6)	–0.6 (0.5)	0.2 (0.4)	4.2 (0.3)
trans pairs							
Non-Watson–Crick *trans* (involving only N – H···N/O hydrogen bonds)	5,734	–1.5 (17.4)	0.2 (11.8)	–0.1 (15.2)	0.0 (0.5)	1.7 (1.1)	3.1 (0.3)
A:A W:W	266	10.2 (10.8)	2.7 (7.9)	3.9 (31.0)	–0.1 (0.7)	2.1 (0.6)	2.9 (0.2)
A:U W:W	210	0.5 (13.2)	–1.8 (7.9)	–0.4 (12.2)	0.0 (0.4)	–0.3 (0.4)	2.8 (0.2)
G:C W:W	132	–5.0 (18.1)	–1.1 (9.3)	–13.5 (12.4)	0.2 (0.4)	–2.2 (0.6)	2.9 (0.2)
A:A H:H	437	–13.5 (15.7)	1.0 (5.7)	5.2 (15.3)	0.0 (0.5)	2.4 (0.3)	2.8 (0.2)
A:G H:S	2,323	–2.3 (15.7)	1.3 (14.8)	0.8 (13.8)	0.0 (0.6)	2.9 (0.4)	3.3 (0.2)
A:A H:W	204	1.3 (13.2)	–1.3 (9.3)	–6.0 (25.0)	–0.2 (0.6)	2.4 (0.4)	2.9 (0.2)
A:C H:W	282	–2.1 (19.4)	–2.5 (7.5)	–9.2 (20.7)	0.0 (0.5)	2.4 (0.3)	3.0 (0.2)
A:U H:W	1,193	–5.8 (14.9)	–0.5 (7.2)	–2.7 (10.0)	0.0 (0.4)	0.1 (0.4)	2.8 (0.2)
G:G S:S	130	9.6 (23.0)	1.7 (7.2)	17.6 (12.6)	0.4 (0.7)	1.4 (0.3)	3.5 (0.2)
G:A S:W	350	13.5 (19.4)	2.0 (15.0)	1.3 (15.6)	0.1 (0.5)	1.8 (0.3)	3.4 (0.2)
Non-Watson–Crick *trans* (involving C – H···N/O hydrogen bonds)	2,336	–10.2 (26.2)	14.2 (27.8)	–8.9 (14.0)	0.0 (0.9)	1.5 (0.9)	3.1 (0.4)
A:A h:s	176	–6.0 (18.2)	–0.3 (7.7)	2.5 (8.6)	–0.1 (0.6)	2.4 (0.5)	2.7 (0.3)
A:C w:s	180	–0.5 (26.1)	–51.0 (9.5)	–16.7 (16.0)	–0.4 (0.8)	0.1 (0.3)	3.9 (0.2)
A:G s:s	1,517	–12.8 (27.2)	27.9 (6.6)	–8.6 (13.3)	0.1 (0.9)	1.7 (0.3)	3.1 (0.2)

Base pairs with frequency of occurrence greater than 100, in 145 RNA crystal structures are shown. A specific base pair type is denoted by indicating the single letter code for each of the two bases, followed by edges of the two bases involved in hydrogen bonding and finally stating the orientation of the pair. Watson–Crick, Hoogsteen and sugar edge involving N – H···N/O hydrogen bond donors/acceptors are denoted by 'W', 'H' and 'S', while those involving C – H···N/O hydrogen bond donors/acceptors are denoted by 'w', 'h' and 's'. Protonated Watson–Crick edge is denoted by '+'

attempted to characterize the conformational features of frequently non-canonical base pairs observed in RNA crystal structures. These intra-base pair parameters have been calculated for more than 42,000 pairs identified by the program BPFIND from 145 RNA crystal structures [17]. The average values of the six parameters, along with their standard deviations for different types of frequently occurring base pairs (frequency of occurrence above 100) are tabulated in Table 3. Detailed analysis of the parameters indicated characteristic distributions for different types of base pair geometries, as evident from their average and standard deviation values (Table 3). Propeller of A:U and G:C W:W *cis* (canonical) base pairs in RNA do not differ significantly (Table 2). As expected, G:U W:W *cis* base pairs have characteristic large shear values while the other parameters are similar to canonical base pairs. Most of the *cis* base pairs have negative propeller, small buckle and open-angles. These, however, do not follow any particular trend in *trans* base pairs. Using BPFIND, we have detected three different types of A:G base pairs, which occur frequently in the RNA crystal structures. Parameters of these three types (G:A W:W *cis*, G:A S:W *trans* and G:A S:H

trans) show values indicative of strong and stable base pairing. There are distinguishing features inherent to each base pair type, which are also evident from our parameter values. Similarly, different types of base pairs involving other bases, such as A:A, A:C, G:G and U:U, etc. were also identified, whose structural specifications could easily be highlighted through their parametric analyses. These structural variations, often specific to each class of base pair geometries can facilitate in understanding the structural asymmetries introduced in regular A-form RNA helices by non-canonical base pairs.

In spite of characteristic structural signatures of different classes of base pair types, standard deviation values for most of the pair types are within acceptable ranges, indicating reasonable conformational stabilities. The standard deviation value for propeller of A:A W:W *trans* pair is rather high compared to others. A detailed analysis of these pairs shows possibility of two $N-H\cdots N$ hydrogen bonds between them, although some of the pairs are highly non-planar, exhibiting large propeller values. These base pairs exhibiting unusual propeller values are all between residues 1746 and 1754 of 23S rRNA from *Haloarcula marismortui* and constitute a considerable part of the A:A W:W *trans* data set. They are also found to interact with a third base, thus forming a triplet, which further perturbs the geometry of the pair. After discarding such anomalous pairs, the average and standard deviation of propeller of this base pair type decreases to 10.2 (20.2).

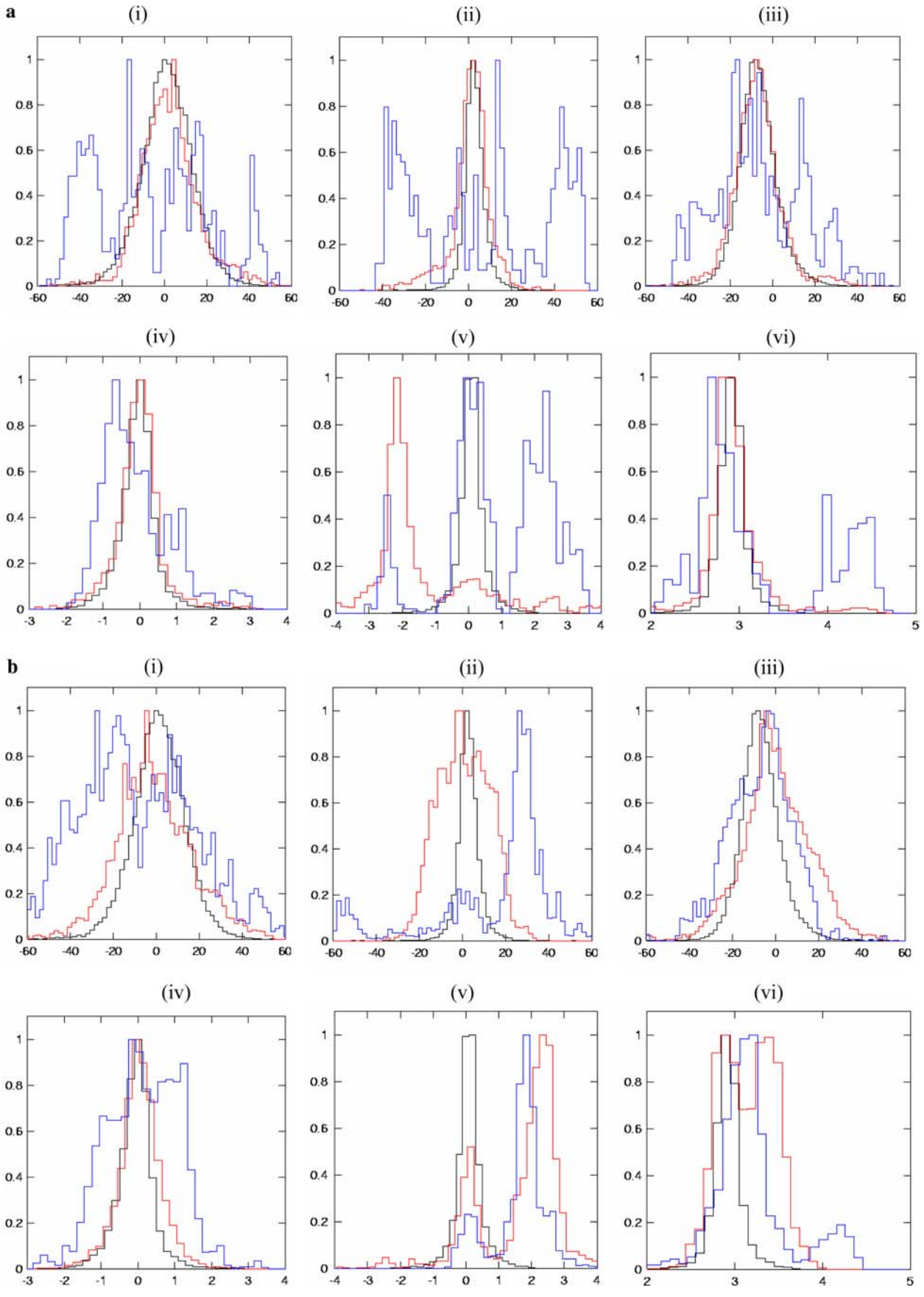
Structural specificities of non-canonical base pairs in comparison with the canonical pairs

To study the characteristic parametric distributions of *cis* and *trans* pairs in RNA and to understand the structural variations of pairs formed with $N-H\cdots N/O$ hydrogen bonds and those with $C-H\cdots N/O$ interactions, all base pairs detected in 145 RNA crystal structures have been classified into five subsets: (1) canonical set containing G:C and A:U W:W *cis* pairs, (2) non-canonical pairs in *cis* orientation involving $N-H\cdots N/O$ hydrogen bonds only, (3) non-canonical pairs in *cis* orientation involving $C-H\cdots N/O$ interactions, (4) non-canonical pairs in *trans* orientation involving $N-H\cdots N/O$ hydrogen bonds only and (5) non-canonical pairs in *trans* orientation involving $C-H\cdots N/O$ interactions. Comparison of the histograms of all six parameters (Fig. 4) indicates that the structural variability for the canonical set is always small when compared to others. Smaller values of standard deviations compared to others (Table 3) also indicate that they undergo relatively small structural variations while retaining their canonical mode of interaction. The distributions for non-canonical *cis* oriented pairs with $N-H\cdots N/O$ interactions overlap in most cases with that of the canonical set, except for shear. However, the spread of distributions are always wider for other intra-base pair parameters than that of canonical, indicating greater structural variability of these non-canonical pairs. Although, the frequency of occurrence is much higher for canonical base pairs as compared to a specific type of non-canonical base pair, each of the non-canonical base pair analyzed here has a population >100 and constitutes a statistically significant data set. Furthermore, the structural characteristics of non-canonical base pairs obtained from their statistical distributions are validated by visualization of several such base pair arrangements. Thus the conformational characteristics obtained from the statistical analyses presumably reflect the true preference of their three-dimensional geometry.

Analyses of base pair conformation indicate characteristic differences between distribution patterns for *trans* and *cis* orientations. The non-canonical *trans* distributions are most often wider than *cis* distributions. This indicates that *cis* pairs are geometrically more constrained than *trans* pairs, which may have occurred due to specific orientation of two bulky sugar moieties on the same side of the base pair hydrogen bonds. Differences are observed in case of propeller of *trans* pairs, whose distribution shifts toward more positive values with respect to the canonical pairs. Characteristic shifts toward negative axis are noted in distributions of buckle for *trans* pairs, while those of the canonical pairs are near the 0 mark. Moreover, stagger and stretch also show large positive values as compared to the canonical pairs. All these characteristic differences noted for *trans* pairs are also there when compared to the non-canonical *cis* pairs, pointing to the inherent structural variability of *trans* geometries.

It is further noted that parametric distributions for non-canonical *cis* pairs involving $C-H\cdots N/O$ interactions are much wider than canonical or non-canonical pairs with $N-H\cdots N/O$ interactions. Average values for buckle, open-angle and propeller are higher, with large standard deviation values. This indicates that pairs involving comparatively weaker $C-H\cdots N/O$ hydrogen bonds deviate considerably more from the ideal planar geometry and show more conformational flexibility. The same trend is also observed for pairs in *trans* orientation.

Apart from parameters like buckle, propeller, open-angle, stagger and stretch, distribution patterns for


a
(i)
(ii)
(iii)
(iv)
(v)
(vi)
b
(i)
(ii)
(iii)
(iv)
(v)
(vi)

◀ **Fig. 4** Histograms of (i) buckle (°), (ii) open-angle (°), (iii) propeller (°), (iv) stagger (Å), (v) shear (Å) and (vi) stretch (Å) of (a) canonical *cis* (*black*), non-canonical *cis* pairs involving N – H···N/O hydrogen bonds (*red*) and non-canonical *cis* pairs involving C – H···N/O hydrogen bonds (*blue*); (b) canonical *cis* (*black*), non-canonical *trans* involving N – H···N/O hydrogen bonds (*red*) and non-canonical *trans* involving C – H···N/O hydrogen bonds (*blue*). *Y-axis* in each of the figures represents the normalized frequency, while *X-axis* represents the parameter values

shear showed a huge variation for five subsets of base pairs. Distribution of the canonical set mostly follows a normal or Gaussian distribution pattern. The histogram exhibited one peak around 0 and values ranging on either side of the peak, which is the standard signature for regular DNA base pairs. However, distributions for non-canonical pairs exhibited multimodal nature, the peaks being intercepted by regions of very low level of occurrences. This is due to various types of hydrogen-bonding patterns exhibited by different types of non-canonical pairs, which are sometimes sheared with respect to the canonical mode. For example, the much studied G:U wobble pair designated as G:U W:W *cis* in our study exhibits a shear value of –2.3 Å (± 0.4), which is the expected structural signature of these base pair geometries. Thus, it should be noted that large values of shear for such base pair types are a characteristic feature of their geometry, rather than indicative of any kind of deformation.

To analyze base pairs formed through different hydrogen-bonding edges, we have compared their parameter distributions with those of canonical pairs formed through regular W edges in *cis* orientation. Characteristic distribution patterns are often noted for H or S edges compared to W, which can act as their structural signature. However, among several non-canonical base pair geometries, those involving S edge are found to exhibit maximum alterations compared to the canonical set. We observed that distributions of buckle, open-angle and propeller for pairs involving S edge in *trans* orientation are wider, while shear, stagger and stretch values are more positive compared to other non-canonical *trans* pairs. All these structural signatures point to the fact that pairs involving the sugar edge of guanines have to undergo more deformation in order to accommodate two closely placed bulky sugar moieties. Lower frequencies observed for base pairs involving S edge in *cis* orientation may also indicate the effect of the steric factor (placing of two sugar moieties on same side of the pseudo-hydrogen bond) in restricting their occurrences. Parameter values of the observed *cis* base pairs involving S edge also suggest large amount of deformation compared to other non-canonical *cis* pairs (Table 3). Further, it is noted that pairs, which have hydrogen bonds involving O2′ atom of sugar, have large values of open-angle and stretch although their visual examination shows quite linear and short hydrogen bonds. This occurs because the pairs are formed by involving the sugar oxygen atom O2′, while their base centered axis system is defined using the C1′ atom (which lies in the base plane) instead of O2′. This has been purposely done in order to retain the definition of base centered axis system that can truly highlight the extent of base pair planarity between two base rings in most of the cases. Furthermore, such an axis system also correctly calculates the values of buckle, propeller, stagger and shear for pairs involving O2′.

Special conformational characteristics of base triples in RNA

Along with double helical stem regions containing Watson–Crick, wobble and different types of non-canonical base pairs, there are several other structural motifs in RNA that involve interaction of distant bases with base pairs leading to the formation of higher order structures. Some such important structural motifs are the C-motif, A-minor and the kink-turn motif, where three bases interact to form base triples [54]. In order to understand the structural pattern of these important base triples, we have compared the structural features of the most frequently occurring triples along with those occurring as pairs (Table 4). Base triples occurring with regular G:C W:W *cis* and A:U W:W *cis* and some frequently observed non-canonical types are analyzed in detail. It is interesting to note that, presence of a third base along with a pair often alters the structure of the canonical/non-canonical base pair (Table 4). The propeller values, in particular, are seen to lose their general propensity of adopting negative values (as observed in all W:W *cis* pairs) in triples. Moreover, the presence of a third base restricts the structural variability of the base pair, as reflected by a decrease in the standard deviation values for most of the parameters, when compared to those in isolated base pairs. This general feature is clearly evident in the G:C W:W *cis*/G:G S:S *trans* triplet, where G undergoes pair formation with C involving W edge of both bases (G:C W:W *cis*) and also with another G, involving the S edge of both bases (G:G S:S *trans*). It is observed that due to hydrogen bonding of a third base G (S) with G:C W:W *cis*, parameters of the latter show some significant differences when compared to that of a regular G:C W:W *cis* base pair. Differences are observed for values of propeller, which assumes large

Table 4 Mean and standard deviations (within parenthesis) of intra-base pair parameters of base pairs involved in formation of base triples in RNA crystal structures

Type of pairs in triples	Frequency of occurrence as triples	Frequency of occurrence as pairs	Buckle (°)		Open (°)		Propeller (°)		Stagger (Å)		Shear (Å)		Stretch (Å)	
			In triples	In pairs	In triples	In pairs	In triples	In pairs	In triples	In pairs	In triples	In pairs	In triples	In pairs
Type of triple: A:U W:W *cis*/A:U H:W *trans*														
A:U W:W *cis*	43	6,613	14.1 (8.4)	–0.3 (10.0)	10.7 (3.6)	3.6 (5.1)	0.6 (13.2)	–8.4 (8.7)	0.2 (0.3)	–0.1 (0.4)	0.3 (0.2)	0.0 (0.3)	2.8 (0.1)	2.8 (0.1)
A:U H:W *trans*	43	1,002	17.9 (9.3)	6.9 (14.7)	2.8 (2.1)	0.2 (7.3)	9.6 (6.3)	2.6 (10.2)	0.4 (0.3)	0.0 (0.4)	0.2 (0.2)	0.1 (0.4)	2.8 (0.1)	2.8 (0.2)
Type of triple: A:U W:W *cis*/A:C H:W *trans*														
A:U W:W *cis*	116	6,613	1.4 (6.4)	–0.3 (10.0)	4.6 (2.9)	3.6 (5.1)	15.0 (5.6)	8.4 (8.7)	–0.3 (0.3)	–0.1 (0.4)	0.1 (0.2)	0.0 (0.3)	2.7 (0.1)	2.8 (0.1)
A:C H:W *trans*	116	158	8.1 (7.1)	8.0 (22.9)	6.4 (5.2)	0.8 (8.2)	9.6 (23.3)	7.2 (19.1)	0.1 (0.5)	–0.1 (0.5)	2.3 (0.4)	2.4 (0.4)	3.0 (0.1)	2.9 (0.2)
Type of triple: G:C W:W *cis*/G:G S:S *trans*														
G:C W:W *cis*	71	21,090	3.2 (8.7)	0.5 (12.1)	1.1 (4.7)	1.3 (4.1)	11.9 (6.0)	–8.0 (8.5)	0.0 (0.4)	–0.2 (0.4)	–0.2 (0.3)	–0.0 (0.3)	2.9 (0.1)	2.9 (0.1)
G:G S:S *trans*	71	62	31.1 (10.2)	1.2 (20.2)	1.2 (6.0)	0.6 (8.6)	5.4 (7.1)	23.6 (12.5)	0.8 (0.5)	0.0 (0.6)	2.3 (0.3	2.5 (0.1)	2.9 (0.2)	2.8 (0.1)
Type of triple: G:C W:W *cis*/G:A S:W *trans*														
G:C W:W *cis*	167	21,090	8.4 (7.2)	0.5 (12.1)	1.7 (3.7)	1.3 (4.1)	12.3 (8.7)	–8.0 (8.5)	0.0 (0.3)	–0.2 (0.4)	–0.1 (0.2)	–0.0 (0.3)	2.8 (0.1)	2.9 (0.1)
G:A S:W *trans*	167	140	9.1 (21.9)	18.1 (18.0)	1.4 (5.1)	4.2 (7.4)	1.1 (8.7)	5.2 (10.1)	0.0 (0.4)	0.1 (0.5)	2.3 (0.3)	2.2 (0.3)	3.0 (0.2)	3.0 (0.2)
Type of triple: A:A W:W *trans*/A:G H:S *trans*														
A:A W:W *trans*	61	143	5.8 (16.7)	10.5 (9.5)	–1.8 (6.1)	3.8 (5.9)	21.5 (8.4)	4.6 (22.8)	–0.0 (0.2)	0.3 (0.5)	1.9 (0.9)	2.3 (0.7)	2.9 (0.1)	2.9 (0.2)
A:G H:S *trans*	61	2,225	10.0 (16.6)	–2.6 (16.7)	–8.4 (3.3)	–0.1 (6.8)	11.7 (3.4)	0.8 (12.4)	0.5 (0.2)	0.0 (0.6)	2.6 (0.2)	2.8 (0.4)	3.1 (0.1)	2.9 (0.2)
Type of triple: A:A W:W *trans*/A:U H:W *trans*														
A:A W:W *trans*	86	143	12.0 (10.7)	10.5 (9.5)	–1.9 (10.9)	3.8 (5.9)	–0.2 (42.5)	4.6 (22.8)	0.5 (0.7)	0.3 (0.5)	1.9 (0.2)	2.3 (0.7)	2.8 (0.2)	2.9 (0.2)
A:U H:W *trans*	86	1,002	–1.8 (7.9)	6.9 (14.7)	–0.2 (7.3)	0.2 (7.3)	–0.5 (5.4)	2.6 (10.2)	–0.1 (0.2)	–0.0 (0.4)	–0.0 (0.2)	0.1 (0.4)	2.8 (0.1)	2.8 (0.2)

Values are separately tabulated for each type of pair when involved in triplet formation and also when involved only in pair formation in RNA. A base triple is denoted by indicating the two types of base pairs involved in triple formation separated by a slash

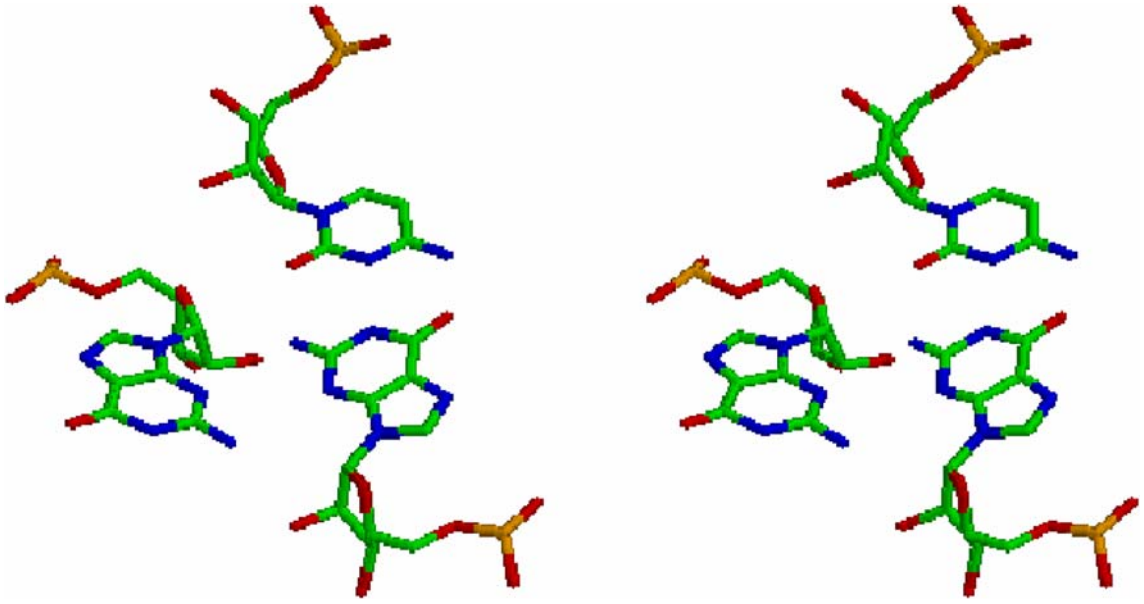


Fig. 5 Stereo view of a C(W):(W)G(S):(S)G base triplet, formed between residues 451:32:456 of chain '0' in 50S ribosomal unit (PDB ID—1FFK)

positive values when G:C W:W *cis* base pairs are involved in triplet formation, while regular G:C W:W *cis* show predominantly negative values with a distribution pattern ranging on both sides of the 0 value. Differences are also noted for buckle values of G:G S:S *trans*, which exhibits only high positive values when engaged in triple formation, whereas those of G:G S:S *trans* pairs show distribution ranging on both sides of 0°. Similarly, stagger of G:G S:S *trans* pair involved in triple formation shows large positive values only. These results indicate that formation of base triples can often lead to certain steric/repulsive constraints, which determine the structural specificities of the base triple (Fig. 5). This spatial conformation allows the base pairs in a triple to assume a limited range of parametric values (for example, for propeller, buckle or stagger) and bars other geometries that are possible for a G:C pair or a G:G pair not involved in triplet formation.

It is seen that the standard deviation value for propeller of A:A W:W *trans* pair involved in triplet formation is high (Table 4). As previously noted, the high value is due to inclusion of the same type of base pairs with deformed geometry, from similar 23S rRNA molecules from *H. marismortui*. After discarding such base pairs, the average and standard deviation of this base pair type decreases significantly. This survey, although not exhaustive, highlights the efficiency of the new edge specific axis system in understanding the specific structural characteristics of base triples that often differ from those of the base pairs forming the triples.

Conclusion

In this study, we have quantitatively described the geometric conformation of all types of canonical and non-canonical pairs occurring in RNA and stabilized by at least two hydrogen bonds formed between heavy atoms of the nucleotide base moieties and the sugar O2′ atom. This has been accomplished by calculating intra-base pair geometrical parameters using a new hydrogen-bonding edge specific axis definition that enables us to study the structural features of a variety of base–base arrangements, ranging from canonical and non-canonical base pairs to base triples and provides useful information regarding the conformational stability of such unusual pairs. Thus, one can visualize a base pair conformation from the set of six parameters and the base pair type.

The structural analysis reveals characteristic geometric features of non-canonical pairs in both *cis* and *trans* orientations in comparison to that of the canonical pairs. It is seen that *cis* base pairs, irrespective of their hydrogen-bonding edge, prefer to adopt negative propeller and small values (near 0) for the remaining five intra-base pair parameters. This trend is not observed in case of *trans* base pairs, which exhibit more conformational flexibility, possibly arising due to positioning of two bulky sugar moieties being far apart as compared to their *cis* counterparts. The conformational analysis reveals certain structural constraints faced by base pairs involving the sugar edge, for

example, placement of bulky sugar moieties in close vicinity also forces most of the S:S pairs to adopt non-planar geometries. The study also differentiates base pairs involving the same hydrogen-bonding edge of two bases, but exhibiting sheared geometries due to formation of different hydrogen-bonding patterns. Large shear values observed for all base pairs of a particular type are indicative of a characteristic feature of that base pair type and may not highlight structural deformation. The study further indicates that pairs involving weak $C-H\cdots N/O$ hydrogen bonds show greater conformational flexibility compared to those formed with $N-H\cdots N/O$ bonds. This perhaps indicates that base pairs involving stronger $N-H\cdots N/O$ hydrogen bonds play a more definitive role in three-dimensional folding of the RNA macromolecule due to their greater conformational stability. Although $C-H\cdots O$ type interactions appear quite frequently in nucleic acid [6, 7, 55] our recent study on energetics of cross-strand hydrogen bond indicated poor interaction energy for them [56]. On the other hand, pairs with weaker $C-H\cdots N/O$ hydrogen bonds may act as switches in RNA where it needs to open up quickly for performing some specific enzymatic function.

Our study indicates that structure of a base pair may get significantly altered when a third base interacts with it to form a base triple, which often play an important role in some functional motifs. The general effect of the third base on a base pair doublet is often manifested through reduction of propeller values and overall structural flexibility of the doublet. Hence, mere extrapolation of structural characteristics of base pair doublets does not help us to understand structural specificities of the triples. However, a clear understanding of the conformational specificities of higher order structures is essential to understand their role in maintaining the tertiary interactions present in a RNA three-dimensional structure. The analysis of non-canonical pairs and triples with the aid of edge specific axis system provides a physically meaningful and quantitative structural insight into the huge and diverse array of possible non-canonical base–base interactions. This quantitative information on geometries of base pair or base triple should further assist in determining sets of closely related isomorphic pairs/triples, which would have great utility in the areas of homology modeling and secondary structure prediction of RNA molecules. The study also aims to facilitate in understanding the nature of structural asymmetries introduced by non-canonical base pairs stacked within regular Watson–Crick helical stretches and thereby provide insight into the subtle structural variations leading to the formation of specific ligand binding sites.

Acknowledgments We are grateful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), India, for financial support. We are thankful to Ms Jhuma Das for technical support.

References

1. Gesteland RF, Cech TR, Atkins JF (1999) The RNA world, 2nd edn. Cold Spring Harbor Laboratory, New York
2. Hermann T, Westhof E (1999) Chem Biol 6:R335
3. Hoogsteen K (1963) Acta Crystallogr 16:907
4. Topal MD, Fresco JR (1976) Nature 263:289
5. Wahl MC, Rao ST, Sundaralingam M (1996) Nat Struct Biol 3:24
6. Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA (2001) Proc Natl Acad Sci USA 98:4899
7. Leontis NB, Westhof E (2001) RNA 7:499
8. Leontis NB, Stombaugh J, Westhof E (2002) Nucleic Acids Res 30:3497
9. Duarte CM, Pyle AM (1998) J Mol Biol 284:1465
10. Gendron P, Lemieux S, Major F (2001) J Mol Biol 308:919
11. Klosterman PS, Tamura M, Holbrook SR, Brenner SE (2002) Nucleic Acids Res 30:392
12. Lemieux S, Major F (2002) Nucleic Acids Res 30:4250
13. Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE (2002) Nucleic Acids Res 30:395
14. Walberer BJ, Cheng AC, Frankel AD (2003) J Mol Biol 327:767
15. Lee JC, Gutell RR (2004) J Mol Biol 344:1225
16. Sykes MT, Levitt M (2005) J Mol Biol 351:26
17. Das J, Mukherjee S, Mitra A, Bhattacharyya D (2006) J Biomol Struct Dyn 24:149
18. Saenger W (1984) Principles of nucleic acid structure. Springer, Berlin Heidelberg New York
19. Sponer JE, Spackova N, Kulhanek P, Leszczynski J, Sponer J (2005) J Phys Chem 109A:2292
20. Bhattacharyya D, Koripella SC, Mitra A, Rajendran VB, Sinha B (2006) Manuscript in preparation
21. Mathews DH, Sabina J, Zuker M, Turner DH (1999) J Mol Biol 288:911
22. Dima RI, Hyeon C, Thirumalai D (2005) J Mol Biol 347:53
23. Florian J, Sponer J, Warshel A (1999) J Phys Chem B 103:884
24. Go M, Go N (1976) Biopolymers 15:1119
25. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998) Proc Natl Acad Sci USA 95:11163
26. Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, Harvey SC, Heinmann U, Lu XJ, Neidle S, Shakked Z, Sklenar H, Suzuki M, Tung CS, Westhof E, Wolberger C, Berman HM (2001) J Mol Biol 313:229
27. Calladine CR (1982) J Mol Biol 161:343
28. Bhattacharyya D, Bansal M (1992) J Biomol Struct Dyn 10:213
29. Lu XJ, Olson WK (2003) Nucleic Acids Res 31:5108
30. Leontis NB, Westhof E (2003) Curr Opin Struct Biol 13:300
31. Stark A, Brennecke J, Russell RB, Cohen SM (2003) PloS Biol 1:397
32. Peyret N, Seneviratne A, Allawi HT, SantaLucia J Jr (1999) Biochemistry 38:3468
33. Michel F, Westhof E (1990) J Mol Biol 216:585
34. Gautheret D, Damberger SH, Gutell RR (1995) J Mol Biol 248:27
35. Gautheret D, Gutell RR (1997) Nucleic Acids Res 25:1559

36. Razga F, Koca J, Sponer J, Leontis NB (2005) Biophys J 88:3466
37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235
38. Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H, Westhof E (2003) Nucleic Acids Res 31:3450
39. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B (1992) Biophys J 63:751
40. El Hassan MA, Calladine CR (1995) J Mol Biol 251:648
41. Lu XJ, El Hassan MA, Hunter CA (1997) J Mol Biol 273:668
42. Gorin AA, Zhurkin VB, Olson WK (1995) J Mol Biol 247:34
43. Kosikov KM, Gorin AA, Zhurkin VB, Olson WK (1999) J Mol Biol 289:1301
44. Lavery R, Sklenar H (1988) J Biomol Struct Dyn 6:63
45. Lavery R, Sklenar H (1989) J Biomol Struct Dyn 6:655
46. Dickerson RE (1998) Nucleic Acids Res 26:1906
47. Soumpasis DM, Tung CS (1988) J Biomol Struct Dyn 6:397
48. Tung CS, Soumpasis DM, Hummer G (1994) J Biomol Struct Dyn 11:1327
49. Bansal M, Bhattacharyya D, Ravi B (1995) CABIOS 11:281
50. Bhattacharyya D, Bansal M (1989) J Biomol Struct Dyn 6:645
51. Babcock MS, Pednault EPD, Olson WK (1993) J Biomol Struct Dyn 11:597
52. Babcock MS, Pednault EPD, Olson WK (1994) J Mol Biol 237:125
53. Babcock MS, Olson WK (1994) J Mol Biol 237:98
54. Holbrook SR (2005) Curr Opin Struct Biol 15:302
55. Ghosh A, Bansal M (1999) J Mol Biol 294:1149
56. Bandyopadhyay D, Bhattacharyya D (2006) Bioploymers 83:313