Canonical and Non-canonical Base Pairs in DNA or RNA: Structure, Function and Dynamics

Dhananjay Bhattacharyya[1] and Abhijit Mitra[2]

[1] Computational Science Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata 700064, INDIA
[2] Center for Computational Natural Sciences and Bioinformatics (CCNSB), International Institute of Information Technology (IIIT-H), Gachibowli, Hyderabad 500032, INDIA

ABSTRACT:

Non-canonical base pairs are planar hydrogen bonded pairs of nucleobases, having hydrogen bonding patterns which differ from the patterns observed in Watson-Crick base pairs, as in the classic double helical DNA. The structures of polynucleotide strands of both DNA and RNA molecules can be understood in terms of sugar-phosphate backbones consisting of phosphodiester-linked D-2'-deoxyribofuranose (D-ribofuranose in RNA) sugar moieties, with purine or pyrimidine nucleobases covalently linked to them. Here, the N9 atoms of the purines, guanine and adenine, and the N1 atoms of the pyrimidines, cytosine and thymine (uracil in RNA), respectively, form glycosidic linkages with the C1' atom of the sugars. These nucleobases can be schematically represented as triangles with one of their vertices linked to the sugar, and the three sides accounting for three edges through which they can form hydrogen bonds with other moieties, including with other nucleobases. As also explained in greater details later in this article, the side opposite to the sugar linked vertex is traditionally called the Watson-Crick edge, since they are involved in forming the Watson-Crick base pairs which constitute building blocks of double helical DNA. The two sides adjacent to the sugar-linked vertex are referred to, respectively, as the sugar and Hoogsteen (C-H for pyrimidines) edges.

Each of the four different nucleobases are characterized by distinct edge-specific distribution patterns of their respective hydrogen bond donor and acceptor atoms, complementarity with

which, in turn, define the hydrogen bonding patterns involved in base pairing. The double helical structures of DNA or RNA are generally known to have base pairs between complementary bases, Adenine:Thymine (Adenine:Uracil in RNA) or Guanine:Cytosine. They involve specific hydrogen bonding patterns corresponding to their respective Watson-Crick edges, and are considered as Canonical Base Pairs. At the same time, the helically twisted backbones in the double helical duplex DNA form two grooves, major and minor, through which the hydrogen bond donor and acceptor atoms corresponding respectively to the Hoogsteen and sugar edges are accessible for additional potential molecular recognition events.

Experimental evidences reveal that the nucleotide bases are also capable of forming a wide variety of pairing between bases in various geometries, having hydrogen bonding patterns different from those observed in Canonical Base Pairs [**Figure 1**]. These base pairs, which are generally referred to as Non-Canonical Base Pairs, are held together by multiple hydrogen bonds, and are mostly planar and stable. Most of these play very important roles in shaping the structure and function of different functional RNA molecules. In addition to their occurrences in several double stranded stem regions, most of the loops and bulges that appear in single-stranded RNA secondary structures form recurrent 3D motifs, where non-canonical base pairs play a central role. Non-canonical base pairs also play crucial roles in mediating the tertiary contacts in RNA 3D structures.

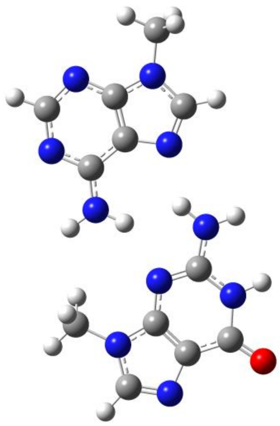
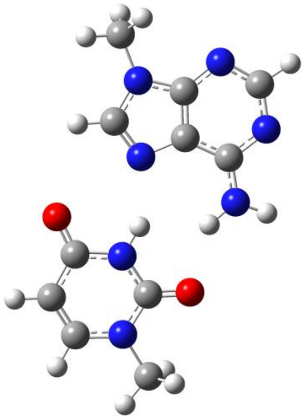
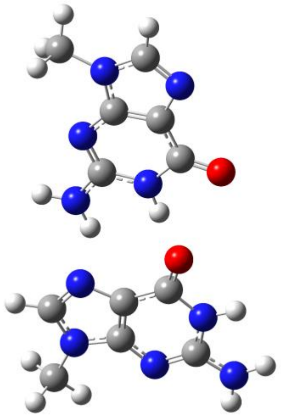
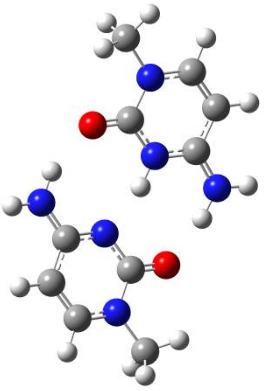
			
Adenine:Guanine *Trans* HS, found in many GNRA tetraloop	Adenine:Uracil *Trans* HW, found in many recurrent structural motifs	Guanine:Guanine *Cis* WH, found in G-Quadruplex of DNA or RNA and in many other motifs	Cytosine:Cytosine *Trans* WW, found in i-motif DNA and other forms. One of the Cyt base needs to be protonated to avoid electrostatic repulsion

Figure 1*: Examples of few frequently observed non-canonical base pairs, Adenine:Guanine trans Hoogsteen/Sugar-edge, Adenine:Uracil trans Hoogsteen/Watson-Crick, Guanine:Guanine cis Watson-Crick/Hoogsteen, Protonated Cytosine(+):Cytosine trans Watson-Crick/Watson-Crick*

Contents:

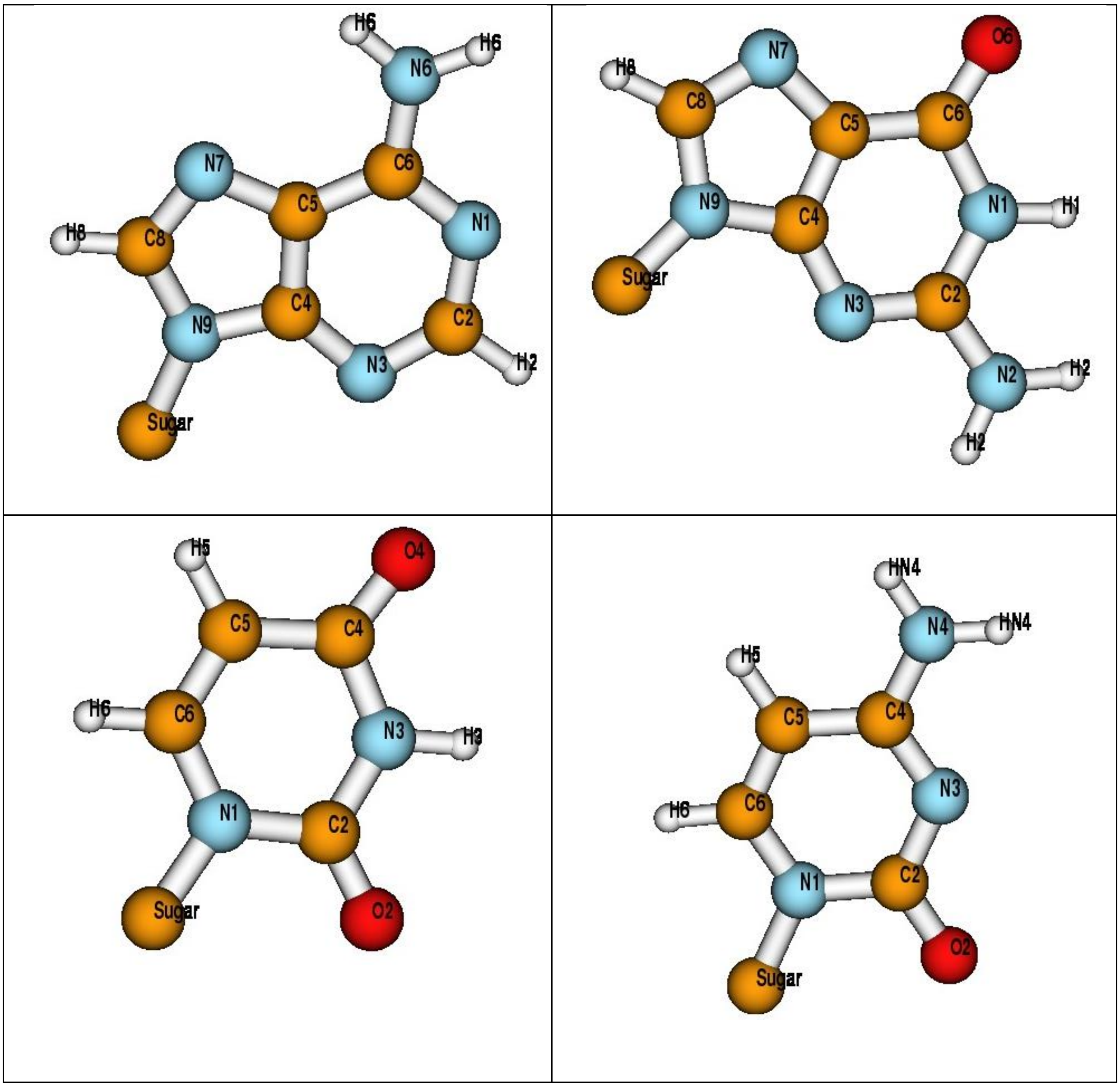


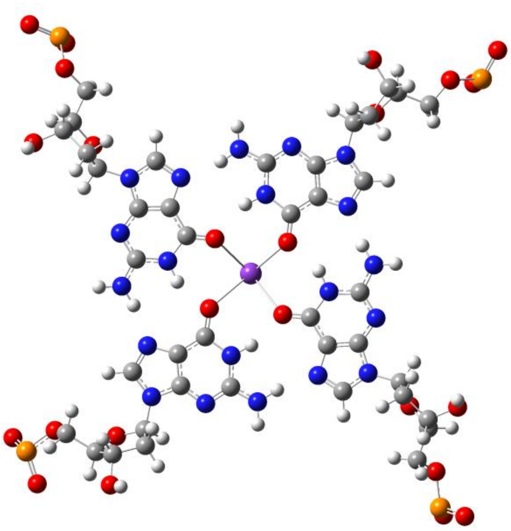
***Figure 2**: IUPAC-IUB recommended nomenclature of nucleotide base atoms of Adenine, Guanine, Uracil and Cytosine bases [created by MOLDEN]*

History

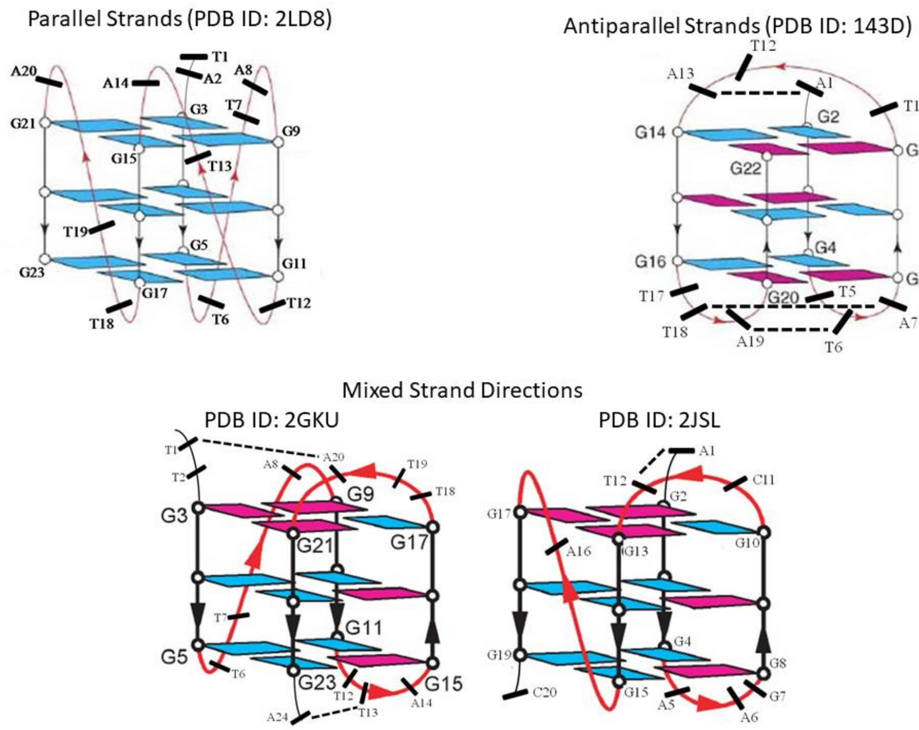
Double helical structures of DNA as well as in folded single stranded RNA are now known to be stabilized by Watson-Crick base pairing between the purines, Adenine and Guanine, with the pyrimidines, Thymine (or Uracil for RNA) and Cytosine. In this scheme, the N1 atoms of the purine residues form hydrogen bond with N3 atoms of the pyrimidine residues in A:T and G:C complementarity (see **Figure 2** for atom labeling scheme according to IUPAC-IUB convention). The second hydrogen bond in A:T base pairs involves the N6 amino group of Adenine and the O4 atom of Thymine (or Uracil in RNA). Similarly, the second hydrogen bond in G:C base pairs involves O6 atom and N4 amino group of Guanine and Cytosine, respectively. The G:C base pairs also have a third hydrogen bond involving the N2 amino group of Guanine and the O2 atom of Cytosine. However, even till about twenty years after this scheme was initially proposed by James D. Watson and Francis H.C. Crick[1], experimental evidences suggesting other forms of base-base interactions continued to draw the attention of researchers investigating the structure of DNA[2,3]. The first high resolution structure of a Adenine:Thymine base pair, as solved by Karst Hoogsteen by single crystal X-ray crystallography in 1959[4], revealed a structure with two hydrogen bonds involving N7 and N6 atoms of Adenine and N3 and O4 (or O2) atoms of Thymine, respectively [**Figure 1b and 2**], which was very different from what was proposed by Watson and Crick. In order to distinguish this alternate base pairing scheme from the Watson-Crick scheme, base pairs where a hydrogen bond involves the N7 atom of a purine residue have been referred to as Hoogsteen base pair, and later, the purine base edge which includes its N7 atom is referred to as its Hoogsteen edge. The first high resolution structure of Guanine:Cytosine pair, obtained by W. Guschelbauer also was similar to the Hoogsteen base pair, although this structure required an unusual protonation of N1 imino nitrogen of Cytosine, which is possible only at significantly lower pH[5]. Experimental evidences, including low resolution NMR studies[6] as well as high resolution X-ray crystallographic studies[7], supporting Watson-Crick base pairing were obtained as late as in the early 70's. Almost a decade later, with the advent of efficient DNA synthesis methods, Richard Dickerson[8] followed by several other groups, solved structures of the physiological double helical B-DNA of complete helical turn based on the crystals of synthetic DNA oligomers[9–12]. The pairing geometries of the A:T (A:U in RNA) and G:C pairs in these structures confirmed the common or canonical form of base pairing as proposed by Watson and Crick, while those with all other geometries, and compositions, are now referred to as non-canonical base pairs.

It was noticed that even in double stranded DNA, where canonical Watson Crick base pairs associate the two complementary antiparallel strands together, there were occasional occurrences of Hoogsteen and other non-Watson-Crick base pairs[13–17]. It was also proposed that Hoogsteen base pair formation could be a transient phenomenon within Watson-Crick base pair[18] dominated DNA double helices.

While canonical Watson-Crick base pairs are most prevalent and common in forming majority of chromosomal DNA or most functional RNAs, presence of stable non-canonical base pairs in DNA biology is also extremely important. An example of non-Watson-Crick, or non-canonical, base pairing can be found at the ends of chromosomal DNA. The 3'-ends of chromosomes contain single stranded overhangs with some conserved sequence motifs (such as TTAGGG in most vertebrates). The single stranded region adopts some definite three-dimensional structures, which has been solved by X-ray crystallography as well as by NMR spectroscopy[19–21]. The single strands containing the above sequence motifs are found to form interesting four stranded mini-helical structures stabilized by Hoogsteen base pairing between Guanine residues. In these structures, four Guanine residues form a near planar base quartet, referred to as G-quadruplex, where each Guanine participates in base pairing with its neighboring Guanine (**Figure 3**), involving their Watson-Crick and Hoogsteen edges in a cyclic manner. The four central carbonyl groups are often stabilized by potassium ions (K+). From the full genomic sequences of different organisms, it has been observed that telomere like sequences sometimes also interrupt double helical regions near transcription start site of some oncogenes, such as c-myc. It is possible that these sequence stretches form G-quadruplex like structures can suppress the expression of the related genes. The complementary Cytosine rich sequences, on the other strand, may adopt another similar four stranded structure, the i-motif, stabilized by Cytosine:Cytosine non-canonical base pairs.



(a)



(b)

Figure 3. *(a) Structure of a representative G-Quadruplex consisting of Hoogsteen base pairs between every neighboring Guanine residues (from PDB ID. 1KF1). (b) Three G-quadruplexes stack to form four stranded telomere with different topologies for d(GGGATTGGGATTGGGATTGGG) sequence.*

While non-canonical base pairs are still relatively rare in DNA, in RNA molecules, where generally a single polymeric strand folds onto itself to form various secondary and tertiary structures, the occurrence of non-Watson-Crick base pairs turns out to be far more prevalent. As early as in the 1970's, analysis of the crystal structure of Yeast tRNA^{Phe} showed that RNA structures possess significant non-canonical variations in base pairing schemes. Subsequently, the structures of Ribozymes, Ribosome, Riboswitches, etc. have highlighted their abundance, and hence the need for a comprehensive characterization of Non-Canonical Base Pairs. These three-dimensional RNA structures generally possess several secondary structural motifs, such as double helical stems, stems with hairpin loops, symmetric and asymmetric internal loops, kissing loops between two hairpin motifs, pseudoknots, continuous stacks between two segments of helices, multi helix junctions[22,23] etc. along with single stranded regions. These secondary structural motifs, except for the single stranded motifs, are stabilized by hydrogen bonded base pairs and several of these are non-canonical base pairs, including G:U Wobble base pairs.

It is notable in this context, that the Wobble hypothesis of Francis Crick predicted the possibility of G:U base pair, in place of the canonical G:C or A:U base pairs, also mediating the recognition between mRNA codons and tRNA anticodons, during protein synthesis. Today, as can be seen in the corresponding Wiki page, the G:U wobble base pair is the most numerously observed non-canonical base pair. While, because of its geometric similarity with the canonical base pairs, they frequently occur in the double helical stem regions of RNA structures, the geometric differences continue to draw the attention of nucleic acid researchers, providing new insights related to its structural significance. It may be noted that though, as in DNA, the base pairs in the folded RNA structures, give rise to double helical stems, its two cleft regions – the major groove and minor groove, differ in their respective dimensions from those in DNA double helices. Unlike for those in DNA, the sequence discriminating major grooves in RNA double helices are very narrow and deep. On the other hand the minor groove regions, though wide and shallow, do not carry much sequence specific information in terms of the hydrogen bonding donor-acceptor positioning of the corresponding base pair edges[24]. The G:U wobble base pairs, along with the various other non-canonical base pairs, introduce variations in the structures of RNA double helices, thus enhancing the accessibility of the discriminating major groove edges of associated base pairs. This has been seen to be very important for molecular recognition steps during tRNA aminoacylation as well as in ribosome functions[25].

Considering the immense importance of the non-canonical base pairs in RNA structure, folding and functions, researchers from multiple domains – biology, chemistry, physics, mathematics, computer science, etc., have joined in the effort to understand their structure, dynamics, function and their consequences. The complexities associated with experimental handling of RNA further underline the importance of diverse theoretical inputs towards addressing these issues.

Types of Non-canonical Base-pairs

Two bases may approach each other in various ways, eventually leading to specific molecular recognition mediated by, often non-canonical, base pairing interactions, in addition to strong stacking interactions. These are essential for the process of RNA single strands folding into three-dimensional structures. Early studies on such unusual base pairs by Jiri Sponer, Pavel Hobza and their group were somewhat disadvantaged due to the unavailability of suitable unambiguous systematic naming schemes[26]. While some of the observed base pair were assigned names following the Saenger nomenclature scheme[27], others were arbitrarily assigned names by different researchers. It may be mentioned that some attempts were also made by Michael Levitt and coworkers to classify base-base association in terms of adjacency of bases, through either pairing or stacking interactions[28]. There was clearly a need for a classification scheme for different types of non-canonical base pairs, which could comprehensively and unambiguously handle newer variants coming up due to the rapid increase in the sampling space. Different approaches which have evolved in response to this need are discussed below.

(i) Classification based on hydrogen bonding

The nucleotide bases are nearly planar heterocyclic moieties, with conjugated pi-electron cloud, and with several hydrogen bonding donors and accepters distributed around the edges, usually designated as W, H or S, based on whether the edges can respectively be involved in forming Watson-Crick base pair, Hoogsteen base pair, or, whether the edge is adjacent to the C2'-OH group of the ribose sugar. Eric Westhoff and Neocles Leontis[29] used these edge designations to propose a, currently widely accepted, nomenclature scheme for base pairs. The hydrogen bonding donor and acceptor atoms could thus be classified in terms of their positioning along their three edges, namely the Watson-Crick or W edge, the Hoogsteen or H edge, and the Sugar or S edge [**Figure 4**]. Since base pairs are mediated through hydrogen bonding interactions based on hydrogen bond donor-acceptor complementarity, this, in turn, provides a convenient bottoms-up approach towards classifying base pair geometries in terms of respective interacting edges of the participating bases. It may be noted that, unlike the Hoogsteen edge of purines, the corresponding edges of the pyrimidine bases do not have any polar hydrogen bond acceptor atom such as N7. However, these bases have C—H groups at their C6 and C5 atoms, which can act as weak hydrogen bond donors, as proposed by Gautam Desiraju[30]. The Hoogsteen edge, hence, is also called Hoogsteen/C-H edge in a unified scheme for designating equivalent positions of purines as well as pyrimidines. Thus, the total number of possible edge combinations involved in base pairing are 6, namely Watson-Crick/Watson-Crick (or W:W), Watson-Crick/Hoogsteen (or W:H), Watson-Crick/Sugar (or W:S), Hoogsteen/Hoogsteen (or H:H), Hoogsteen/Sugar (or H:S) and Sugar/Sugar (or S:S).

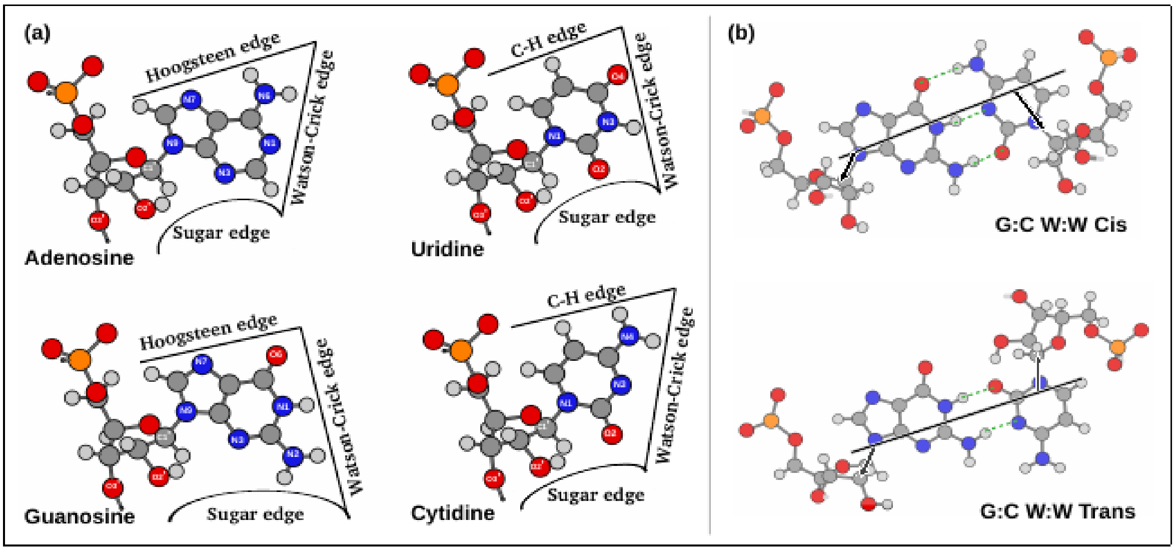


Figure 4*: (a) Three hydrogen bonding edges of the four nucleotides (Guanine), showing nomenclature of each edge and (b) Cis and Trans orientations of the sugar moieties of the two nucleotide residues glycosidic bonds of a base pair with respect to hydrogen bonding direction. The arrows in (b) indicate glycosidic bonds as vectors.*

In the canonical Watson-Crick base pairs, the glycosidic bonds attaching the N9 (of purine) and N1 (of pyrimidine) of the paired bases with their respective sugar moieties, are on the same side of the mean hydrogen bonding axis, and are hence called *Cis* Watson-Crick base pairs. However, the relative orientations of the two sugars may also be *Trans* with respect to the mean hydrogen bonding axis giving rise to a distinct *Trans* Watson-Crick geometric class, consisting of species which were earlier referred to as reverse Watson-Crick base pairs according to Saenger nomenclature[27]. The possibility of both *Cis* and *Trans* glycosidic bond orientation for each of the 6 possible edge combinations, gives rise to 12 geometric families of base pairs (**Table 1**).

According to the Leontis-Westhoff scheme[29], any base pair can be systematically and unambiguously named using the syntax <Base_1: Base_2><Edge_1: Edge_2><Glycosidic Bond Orientation> where Base_1 and Base_2 carry information on respective base identities and their nucleotide number. This nomenclature scheme also allows us to enumerate the total number of distinct possible base pair types. For a given glycosidic bond orientation, say *Cis*, the four naturally occurring bases each have three possible edges for formation of base pairs giving rise to 12 such possible base pairing edge identities, each of which can in principle form base pairing with any edge of another base, irrespective of complementarity. This gives rise to a 12x12 symmetric matrix displaying 144 pairwise permutations of base pairing edge identities, where, apart from the 12 diagonal entries, others include repeat combinations. Thus, there are 78 (= 12 + 132/2) unique entries corresponding to the *cis* glycosidic bond orientation. Considering both *cis* and *trans* glycosidic bond orientations, the number of base pair types amounts to 156.

Of course, this number 156 is only an indicator. It includes base-edge combinations where base pairs cannot be formed due to absence of hydrogen bond donor acceptor complementarities. For example, potential pairing between two Guanine residues utilizing their Watson-Crick edges in *cis* form (cWW) is not supported by hydrogen bonding donor-acceptor complementarity, and is never observed. This method of enumerating the possible number of distinct base pair types also does not consider possibilities of multimodality or bifurcated base pairs, or even instances of base pairs involving modified bases, protonated bases and water or ion mediation in hydrogen bond formation. Two Cytosine bases can form trans Watson-Crick/Watson-Crick (tWW) base pairing with their neutral as well as hemi protonated forms, possibly both, giving rise to the i-motif DNA. However, both C(+):C tWW and C:C tWW, are counted as one type among 156 possible types.

(ii) Classification based on isostericity

Although significant differences are there between structures of non-canonical base pairs belonging to different geometric families, some base pairs within the same geometric family have been found to substitute each other without disrupting the overall structure. These base pairs are called isosteric base pairs. Isosteric base pairs always belong to same geometric families, but all the base pairs in a particular geometric family are not always isosteric. Two base pairs are called isosteric if they meet the following three criteria: (i) The C1′–C1′ distances should be similar; (ii) the paired bases should be related by the similar rotation in 3D space; and (iii) H-bonds formation should occur between equivalent base positions[31,32]. A detailed approach towards quantifying isostericity, in terms of an IsoDiscrepancy Index (IDI), which can facilitate reliable prediction regarding which base pair substitutions can potentially occur in conserved motifs, was formulated by Neocles Leontis, Craig Zirbel and Eric Westhof[33]. Based on IDI values and available base pair structural data, the group maintains a curated online base pair catalogue and an updated set of Isostericity Matrices (IM) corresponding to each of the 12 geometric families. Using this resource, one can comprehensively classify different types of canonical and non-canonical base pairs in terms of their positions in the Isostericity Matrices. This approach, for example, indicates that the four base pair types: A:U cWW, U:A cWW, G:C cWW and C:G cWW are isosteric to each other. Thus, as also confirmed by detailed sequence comparisons, double mutations altering A:U cWW to U:A cWW or even to G:C cWW may not disturb the structure, and, unless stability issues are involved, the function of the related RNA. It was also found that G:U cWW is not really isosteric to U:G cWW, indicating that such double mutations may significantly affect the functioning of the corresponding RNA[25]. On the other hand, some of the base pairs which are stabilized involving Sugar edge of the bases are mutually isosteric.

(iii) Classification based on local strand direction

It may be noted here that because of the geometric relationship of the bases with the sugar phosphate backbone, these 12 geometric families of base pairs are associated with two possible local strand orientations, namely parallel and antiparallel. For the 6 families with edge combinations involving Watson-Crick and Sugar edges, W:W, W:S and S:S, *cis* and *trans* families are respectively associated with antiparallel and parallel local strand orientations [**Table 1**]. Introduction of the Hoogsteen edge, as one of the partners in the combination, causes an inversion in the relationship. Thus, for W:H and H:S, *cis* and *trans* respectively correspond to parallel and antiparallel local strand orientation. As expected, when both the edges are H, a double inversion is observed, and H:H *cis* and *trans* correspond respectively to antiparallel and parallel local strand orientations[29]. The annotation of local strand orientation in terms of parallel and antiparallel directions helps to understand which faces of the individual bases can be seen for a given base pair from the 5'- or the 3' sides [**Table 1**]. This annotation also helps in classifying the 12 geometries into two groups of 6 each, where the geometries can potentially interconvert within each group, by in-plane relative rotation of the bases. However, one should note that the above theory is applicable only when the glycosidic torsion angles of both the nucleotide residues are *anti*. Notably, crystallographic observations[34] and energetic[35] considerations indicate that *syn* glycosidic torsions are also quite possible. Hence the above classification of parallel or antiparallel nature of strand directions, by itself, does not always provide the correct understanding.

Table 1: *Different types of base pairing schemes and associated local strand orientations of their sugar-phosphate backbone*

Interacting edges	Glycosidic bond orientation	Nomenclature	Symbolic representation	Local Strand Direction
Watson-Crick/Watson-Crick	*Cis*	cWW or *cis* Watson-Crick/Watson-Crick	-●-	Antiparallel
Watson-Crick/Watson-Crick	*Trans*	tWW or *trans* Watson-Crick/Watson-Crick	-○-	Parallel

Watson-Crick/Hoogsteen	*Cis*	cWH or *cis* Watson-Crick/Hoogsteen	●-■	Parallel
Watson-Crick/Hoogsteen	*Trans*	tWH or *trans* Watson-Crick/Hoogsteen	○-□	Antiparallel
Watson-Crick/Sugar edge	*Cis*	cWS or *cis* Watson-Crick/Sugar edge	●▶	Antiparallel
Watson-Crick/Sugar edge	*Trans*	tWS or *trans*Watson-Crick/Sugar edge	○-▷	Parallel
Hoogsteen/Hoogsteen	*Cis*	cHH or *cis* Hoogsteen/Hoogsteen	-■-	Antiparallel
Hoogsteen/Hoogsteen	*Trans*	tHH or *trans* Hoogsteen/Hoogsteen	-□-	Parallel
Hoogsteen/Sugar edge	*Cis*	cHS or *cis* Hoogsteen/Sugar edge	■-▶	Parallel
Hoogsteen/Sugar edge	*Trans*	tHS or *trans* Hoogsteen/Sugar edge	□-▷	Antiparallel
Sugar edge/Sugar edge	*Cis*	cSS or *cis* Sugar-edge/Sugar-edge	-▶-	Antiparallel
Sugar edge/Sugar edge	*Trans*	tSS or *trans* Sugar-edge/Sugar-edge	-▷-	Parallel

Various functional RNA molecules are stabilized, in their specific folded pattern, by both canonical as well as non-canonical base pairs. Most tRNA molecules, for example, are known to have four short double helical segments, giving rise to a cloverleaf like two-dimensional structure. The three-dimensional structure of tRNA, however, takes an L-shape. As shown in **Figure5,** this is mediated by several non-canonical base pairs and base triplets. The D-loop and TψC loop are held together by several such base pairs. While it is not possible to include here the complete range of non-canonical base pair varieties, some of the frequently occurring representatives are shown in **Figure 1.** Interested readers are encouraged to browse through different websites such as NDB[9], RNABPDB[36], RNABP COGEST[37], etc., to get a better understanding.

It may be noted that the above scheme is valid for naturally occurring nucleotide bases.

However, there are plenty of examples of post-transcriptional chemical modifications of the bases, many of which are seen in tRNAs or ribosomes. It may be important to understand their structural features also[38,39].

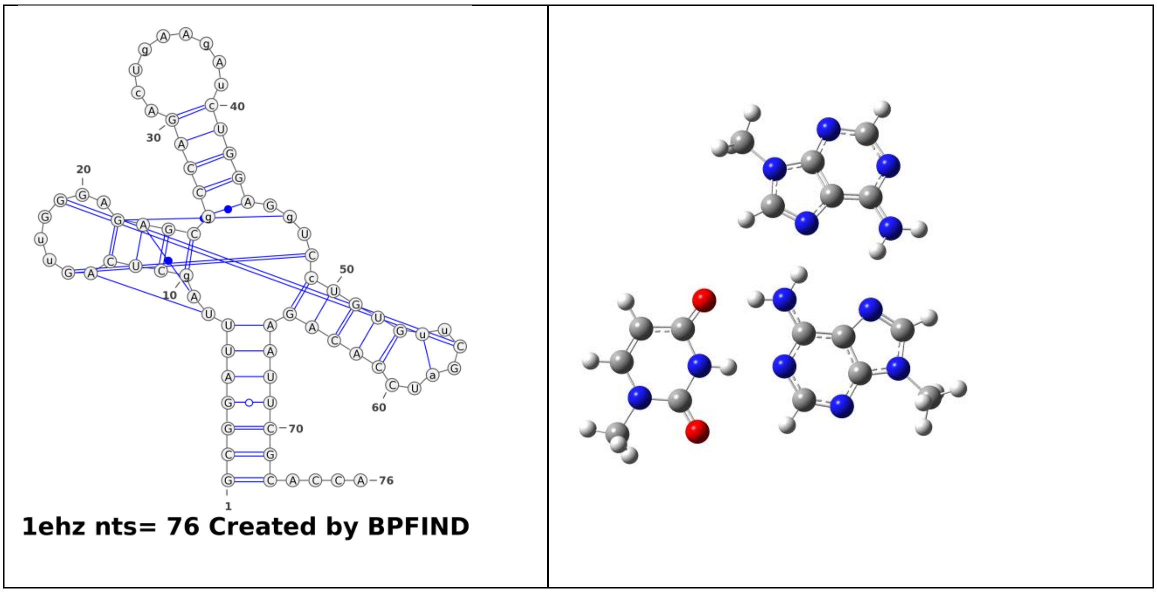


Figure 5*: (a) Cloverleaf model of tRNA^{Phe} [Picture created by VARNA[40] for PDB ID: 1EHZ] and (b) A typical base triplet involving residues 9, 12 and 23 of the same tRNA*

Identification of Non-canonical Base-pairs

In case of double helical DNA, identification of base pairs is quite trivial using molecular visualizers such as VMD, RasMol, etc. It is, however, not so simple for single stranded folded functional RNA molecules. Several algorithms have been implemented in software tools for the automated detection of base pairs in RNA structures solved by X-ray crystallography, NMR or other methods. Essentially the programs detect hydrogen bonds between two bases, and ensure their (near) planarity, before reporting that they constitute a base pair. Since most of the structures of RNA, available in public domain, are solved by X-ray crystallography, the positions of hydrogen atoms are rarely reported. Hence, detection of hydrogen bond becomes a non-trivial job.

The DSSR algorithm[41] by Lu and Wilma K. Olson considers two bases to be paired when they detect one or more hydrogen bond(/s) between the bases, by actually modeling the positions of the hydrogen atoms, and by ensuring the perpendiculars to the two bases being nearly parallel to each other. The positions of the hydrogen atoms can be deduced by converting Internal Coordinates (bond length, bond angle and torsion angle) along with positions of precursor atoms, such as amino group nitrogen atoms and those bonded to the nitrogen or Z-matrix to external Cartesian Coordinates. The base pairs identified by this method are listed in NDB[42] and FR3D[43] databases.

A unique way of identification of base pairs in RNA was incorporated in MC-Annotate[44] by Francois Major. In this algorithm they make use of the positions of the hydrogen atoms as well as lone-pair electrons using suitable molecular mechanics/dynamics force-fields[45] and derive hydrogen bond formation probabilities for them. The final identifications of base pairs are done based on these probabilities and approach of hydrogen atoms to lone-pairs electrons of nitrogen or oxygen. This method also attempted to classify the base pair nomenclature with additional information of each interacting edge, such as *Ws* indicating the sugar edge corner of the Watson-Crick edge, *Wh* representing the Hoogsteen edge corner of Watson-Crick edge, *Bw* indicating bifurcated three-center hydrogen bond involving both the hydrogen atoms of amino groups to form hydrogen bonds with a carbonyl oxygen involving both of its lone-pairs, etc. As claimed by the authors, this nomenclature scheme adds some additional features to the Leontis-Westhof (LW)[29] scheme and may be referred to as the LW+ scheme. A major advantage of this scheme lies in its ability to distinguish between alternative base pairing geometries, where multimodality is observed within an LW family. This method, however, does not consider the possible participation of the 2'-OH group of the ribose sugars in base pair formation.

Another algorithm, namely BPFIND by Bhattacharyya and coworkers[46], demands at least two hydrogen bonds using two distinct sets of donors and acceptors atoms between the bases. This hypothesis driven algorithm considers distances between two pairs of atoms (hydrogen bond donor (D1 and D2) and acceptor (A1 and A2) and four suitably chosen precursor atoms (PD1, PD2, PA1, PA2) corresponding to the D's and A's (as shown for a representative base pair in **Figure 6**). Small values of such distances in conjunction with large values of the angles defined by PD1—D1—A1, D1—A1—PA1, PD2—D2—A2, D2—A2—PA2 (close to 180^o or π^c) ensures two structural features which characterize well defined base pairs: i) the hydrogen bonds are strong and linear and ii) the two bases are co-planar. Notably, so long as one restricts the search to base pairs which are stabilized by at least two distinct hydrogen bonds, the above algorithms, by and large, yield the same set of base pairs in different RNA structures.

Sometimes in the crystal structures it is observed that two closely spaced bases are oriented in such a way that apart from the regular hydrogen bonds two additional electronegative hydrogen bond acceptor atoms are very close to each other, which may cause electrostatic repulsion. The concept of protonated base pairing, implicating a possible protonation of one of these electronegative, (potentially) hydrogen bond acceptor atoms thus converting it into a hydrogen bond donor, was introduced to explain stability of such geometries[31,32,46,47]. Some of the NMR derived structures also support the protonation hypothesis, but possibly more rigorous studies using neutron diffraction or other techniques would be able to confirm it. The quality of the crystal structures permitting, some algorithms also attempted to detect water or cation mediated base pair formation[31,32].

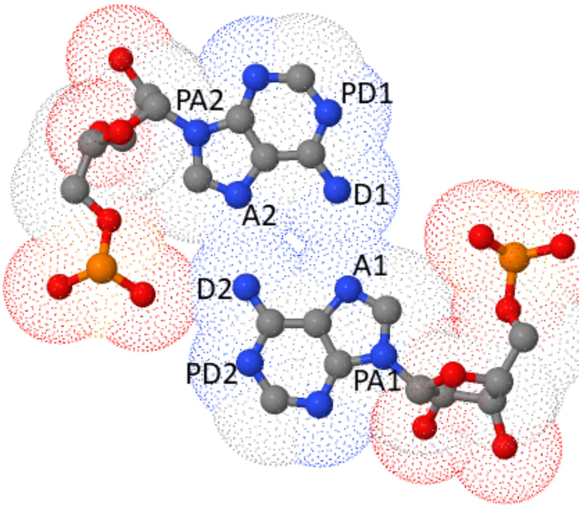


Figure 6*. Descriptions of the hydrogen bonding atoms, along with their precursors (as used by BPFIND algorithm*[46]*), for a typical non-canonical base pair.*

Strengths and stabilities of Non-canonical Base-pairs

The canonical Watson-Crick base pairs, G:C and A:T/U as well as most of the non-canonical ones are stabilized by two or more (e.g. 3 in the case of G:C) hydrogen bonds. Justifiably, a significant amount of research on non-canonical base pairs has been carried out towards benchmarking their strengths (interaction energies) and (geometric) stabilities against those of the canonical base pairs. It may be noted here that base pair geometries, as observed in the crystal structures, are often influenced by several interactions present in the crystal environment, thus perturbing their intrinsically stable geometries arising out of the hydrogen bonding and related interactions between the two bases. Therefore, in principle, it is possible that the observed geometries in some cases are intrinsically unstable, and that they are stabilized by other interactions provided by the environment. Several groups have attempted to determine the interaction energies in these non-canonical base pairs using different quantum chemistry based approaches, such as Density Functional Theory (DFT) or MP2 methods[48–56]. These methods were applied on suitably truncated, hydrogen-added, and geometry optimized models of the base (or nucleoside) pairs extracted from PDB structures.

Depending upon the optimization protocol, typically three types of interaction energies have been reported. In the first method, the base pair model geometries, isolated from their respective environments, are fully optimized without any constraints[48,50,52,55,56], thus providing the intrinsic geometries and interaction energies of the isolated models. This procedure, however, sometimes leads to optimized geometries of base pairs involving edges different from initial crystal geometry. Abhijit Mitra and collaborators also used an additional second protocol, where the heavy atom (non-hydrogen) coordinates are retained as in the crystal geometries, optimizing only the positions of the added hydrogen atoms[51,54,56]. In the third protocol, followed mostly by Jiri Sponer and his group[49], optimization was carried out with constraints on some angles and dihedrals. Given that the models are extracted from their respective crystal structures, and are isolated from their crystal environments, the second and the third protocols provide two different approaches towards approximating the environmental effects, without explicit considerations of any specific environmental interactions. This has further been addressed in some reports by considering specific environmental factors, such as coordination with Magnesium, or even some covalent modifications to the bases[50].

All the three protocols are useful in their respective contexts. Further, a comparison of the model geometries, obtained by the different protocols, provide an idea regarding both, the stabilities of the corresponding base pair geometries, as well as regarding the probable extent and nature of environmental influences. It was found that most non-canonical base pairs, having two or more hydrogen bonds, generally maintain the same hydrogen bonding pattern in the crystal and in fully optimized in isolation geometries, respectively, thus indicating their intrinsic geometric stabilities. Interaction energies calculated from these optimized models also indicated the energetic stabilities of the corresponding non-canonical base pairs. The previous notion that non-canonical base pairs are weaker than the Watson-Crick base pairs, was found to be incorrect. Interaction energies between the bases of Several base pairs, such as G:G tWW, G:G cWH, A:U cHW, G:A cWW, G:U cWW, etc., are found to be larger than that of canonical A:U cWW base pair[37].

Of course all non-canonical base pairs are not necessarily very strong or stable in terms of interaction energy. Several base pairs have been detected on the basis of weak hydrogen bonds involving C—H…O/N atoms, where interaction energies are rather small. Further, geometry optimizations of some of the observed base pairs, in particular, but not limited to those involving weak hydrogen bonds, or those stabilized by single hydrogen bonds, were found to adopt alternate geometries[51,52,57], thus indicating their intrinsic lack of geometric stability. These alteration of hydrogen bonding schemes, giving rise to changes in base pairing family upon free optimization, may have some functional implication in RNA, such as their action as conformational switch. Accordingly, as mentioned above in the Sponer's protocol, there have been some attempts to restrain the experimentally observed geometry while carrying out geometry optimization[49] for interaction energy calculations. Interestingly, in several cases, interaction energies calculated for these 'away from intrinsically stable' geometries also indicate good energetic stability.

Though the energetics and geometric stabilities of different non-canonical base pairs do not show any generalized correlations, analysis of several databases, such as RNABPDB[36] and RNABP COGEST[37], which catalogue structural and energetic features of some of the observed base pair and their stacks, reveal some interesting general trends. For example, geometry optimizations of several base pairs involving 2'-OH group of sugar residue resulted in significant alterations from their initial geometry. This is possibly due to flexibility of the sugar puckers and glycosidic torsions. The significantly high interaction energies of protonated base pairs, despite the high energy cost of base protonation, also deserve a special mention in this context. This can mostly be attributed to the additional charge-induced dipole interactions which are associated with protonated base pairs[47].

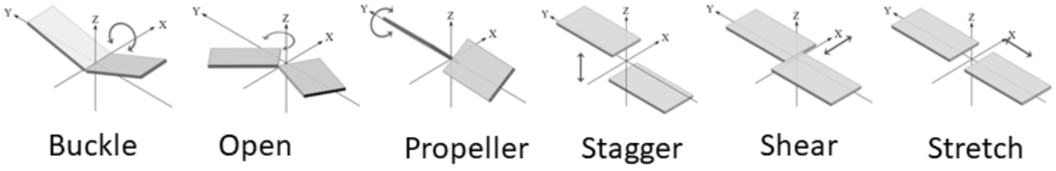


Figure 7. IUPAC recommended Intra base pair parameters used to describe geometry of canonical or non-canonical base pairs

Structural Features of Non-canonical Base-pairs

Structural features of a base-pair, formed by two planar rigid units, can be quantified, using six parameters – three translational and three rotational. IUPAC recommended parameters are Propeller, Buckle, Open Angle, Stagger, Shear and Stretch (**Figure 7**)[58]. Brief description of these in the context of DNA double helical structure can be found in Wiki. There are several publicly available software, such as Curves by Richard Lavery[59], 3DNAby Wilma Olson[60], NUPARM by Manju Bansal[61], etc., which may be used to calculate these parameters. While the first two calculate the parameters of canonical and non-canonical base-pairs relative to the standard canonical Watson-Crick base pairs geometry, the NUPARM algorithm calculates in absolute terms using base pairing edge specific axis system. Hence, for most non-canonical base-pairs, which involve non-Watson-Crick edges, some of the parameters (Open, Shear and Stretch) calculated by Curves or 3DNA are usually large even in their respective intrinsically most stable geometries. On the other hand, the values provided by NUPARM indicate the quality of hydrogen bonding and planarity of the two bases in a more realistic fashion. Thus, the NUPARM Stretch values, indicating separation of the two bases of a base pair, and which depend on optimal hydrogen bonding distances, are always around 3Ǻ. Some other general trends observed in the values of the above parameters may be of interest to note. Most of the *cis* base pairs are seen to have Propeller values around -10^o and small values of Buckle and Stagger. The Open and Shear values often depend on positions of the hydrogen bonding atoms. As for example, GU cWW wobble base pairs have Shear value around -2.2Ǻ while GC or AU cWW base pairs have Shear values around zero. The Open values for most base pairs are close to zero but the values are often rather large for those involving 2'-OH group of sugar in the NUPARM derived parameter set. The *trans* base pairs, however, do not show any systematic trend in their Propeller values.

Role of Non-canonical Base Pairs in RNA

The structural hierarchy in RNA is usually described in terms of a stem-loop 2D secondary structure, which further folds to form its 3D tertiary structure, stabilized by what are referred to as long range tertiary contacts. Most often the non-canonical base pairs are involved in those tertiary contacts or extra-stem base pairs. For example, some of the non-canonical base pairs in tRNA appear between the D-stem and TψC loops (Figure 5), which are close in the three-dimensional structure. Such base pairing interactions give stability to the L-shaped structure of tRNA. In this region, some base pairs are found to be additionally hydrogen bonded to a third base. Thus, as shown in **Figure 5**, the 23rd residue is simultaneously paired to 9th and 12th residues, together forming a base triple, the smallest member of the class of higher order multiplets.

Multiplets formed by Non-canonical Base-pairs

One base, in addition to forming proper planar base pairing with a second base, can often participate in base pair formation with a third base forming a base triple. One such classic example is in formation of DNA triple helix, where two bases of two antiparallel strands form consecutive Watson-Crick base pairs in a double helix and a base of a third strand form

Hoogsteen base pairing with the purine bases of the Watson-Crick base pairs. Many different types of base triples have been reported in the available RNA structures and have been elegantly classified in the literature[62]. Multiplets are however not limited to triplet formation. Four bases giving rise to a base quartet is now well documented in the structure of the G-quadruplex (**Figure 3**) characteristically found in the telomere. Here four Guanine residues pair up within themselves in a cyclic form involving Watson-Crick/Hoogsteen *cis* (cWH) base pairing scheme and each of the Guanine bases are found to be respectively interact with two other guanine bases. Three to four such base G-quadruplexes stack on top of the other to form a four stranded DNA structure. In addition to such a cyclic topology, several other topologies of base:base pairings are possible for higher order multiplets such as quartets, pentets etc.[63].

Non-canonical Base-pairs in Double Helical Regions

Non-canonical base pairs quite frequently appear within double helical regions of RNA. The G:U cWW non-canonical base pairs are seen very frequently within double helical regions as this base pair is nearly isosteric to the other canonical ones[31–33]. Due to complication of strand direction, as elaborated in the Classification section (Table 1), not all types of non-canonical base pairs can be accommodated within double helical regions with *anti* glycosidic torsion angles. However, many non-canonical base pairs, e.g. A:G tHS (*trans* Hoogsteen/Sugar edge) or A:U tHW (*trans* Hoogsteen/Watson-Crick), A:G cWW, etc., are often seen within double helical regions giving rise to symmetric internal loop like motifs. Attempts have been made recently to classify all such situations where two base pairs (canonical or non-canonical) stack in anti-parallel sense possibly giving rise to double helical regions in RNA structures[36]. These base pairs are quite stable, and they are able to maintain the helical property quite well. The backbone torsion angles around these residues are also generally within reasonable limits: C3'-endo sugar pucker with *anti* glycosidic torsion, α/γ around -60°/60°, β/ε around 180°.

Non canonical base pairs in recurrent structural motifs:

Non-canonical base pairs often appear in different structural motifs, including pseudoknots, with their special hydrogen bonding features. Structural features of these recurrent motifs have been archived in searchable databases, such as, FR3D[43] and RNA FRABASE[64]. Also, several of these motifs can be identified in a given query PDB file by the NASSAM[65] web-server. They are most frequently detected at the termini of double helical segment acting as capping residues, often preceding hairpin loops. The most frequently found non-canonical base pair, namely G:A tSH, is an integral part of GNRA tetraloops, where N can be any nucleotide residue and R is a purine residue. This motif shows some amount of flexibility and alterations of structural features depending on whether the Guanine and Adenine are paired or not. Several other types of tetraloops motifs, such as UNCG, YNMG, GNAC, CUYG, (where Y stands for pyrimidine and M is either Adenine or Cytosine) etc., have been found in available RNA structures. However, these do not generally show involvement of non-canonical base pairing. In addition to these common hairpin motifs, where the loop residues largely remain unpaired, there are also a few motifs where the loop residues make extensive interactions between themselves or with other residues external to the loop. A common example is the C-loop motif[66], where the bulging loop residues make non-canonical base pairing with the bases of double helical regions forming non-canonical base pairing (**Figure 8**). The extra base pairs in these cases give rise to additional stabilization to the composite double helix containing motif. Non-canonical base pairs are also involved in receptor-loop interaction, such as in T-loop motif[66] as shown in **Figure 9**.

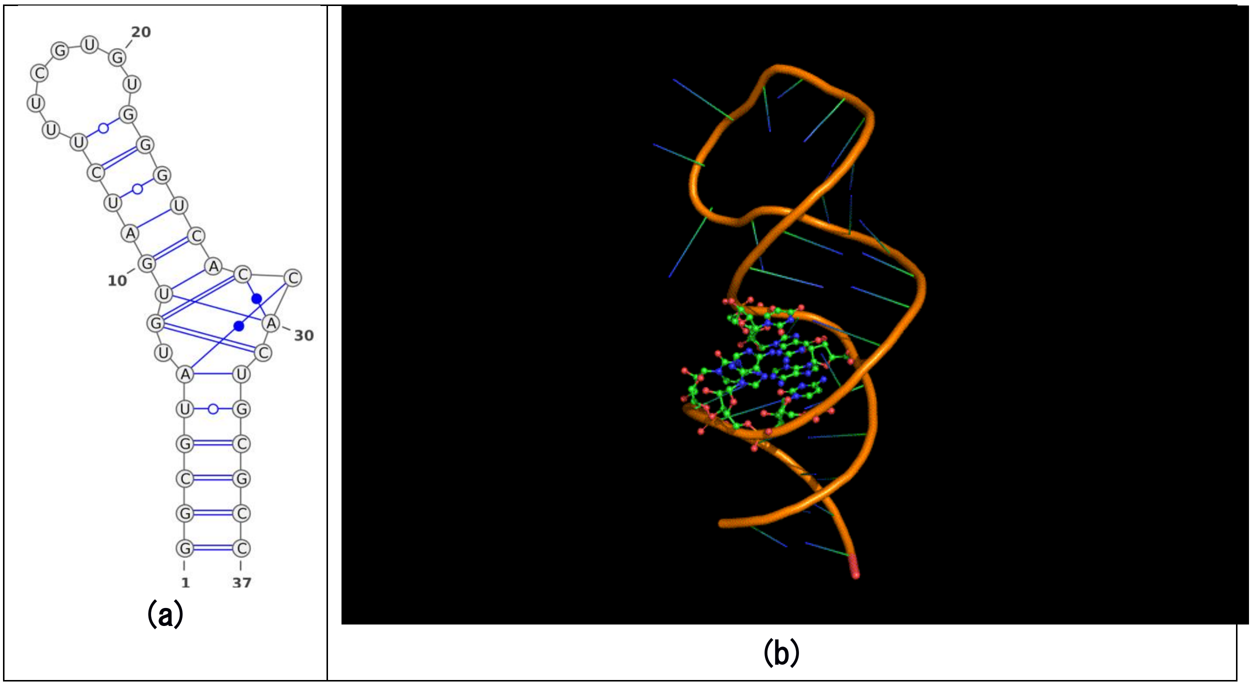


Figure 8. *An example of higher order structure (C-loop) in RNA by formation of base triples using non-canonical base pair from PDB ID 1KOG (a) by schematic representation and (b) by molecular visualizer.*

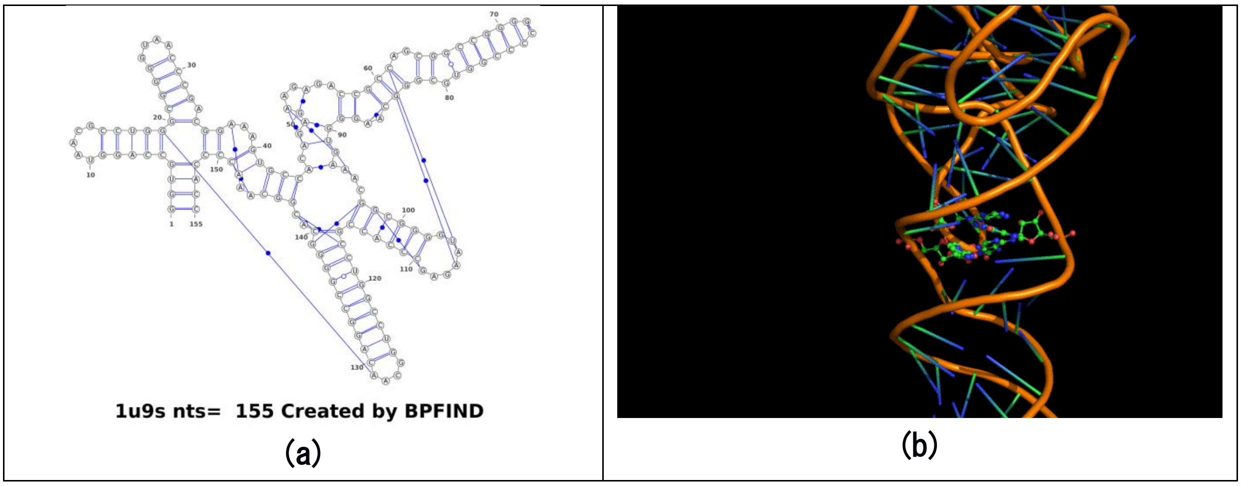


Figure 9. *An example of T-loop motif from an extended RNA 2D structural profile from PDB ID 1U9S by (a) schematic representation and (b) molecular visualizer. The loop residues 105 and 106 form base triples with the 61:84 Watson-Crick base pair.*

Another interesting example of the involvement of non-canonical base pairs in recurrent contexts was detected as the GAAA receptor motif, which consists of A:A cHS base pair followed by U:A tWH base pair stacked on both sides by G:C cWW base pairs. Here we have successive non-canonical base pairs within an antiparallel RNA double helical domain. Similarly there is an A:A cSH base pair involving two consecutive residues in this motif. Such pairing between consecutive residues, which is also termed as a dinucleotide platform motif, is quite commonly observed. They appear in many RNA structures and the pairing can also be between other bases, and can involve other base pairing edges. Such dinucleotide platform was reported in A:A, A:G, A:U, G:A, G:U base pairs belonging to the cSH class and also in A:A cHH base pairs. These motifs can alter the strand direction within a double helix by formation of kinks. Such dinucleotide platform along with triplet formation is also an integral component of the Sarcin-ricin motif[66].

Modeling of RNA structures containing Non-canonical base pairs:

Prediction of biomolecular structure from sequence alone is a long term goal of scientists working in the fields of bioinformatics, computational chemistry, statistical physics as well as in computer science. Prediction of protein structures from amino acid sequence by methods like homology modeling, comparative modeling, threading, etc were successful to some extent due to availability of about 1200 unique protein folds. Inspired by the protein experience, there are now several approaches towards predicting RNA structures, albeit with varying degrees of success. Any comprehensive discussion on RNA modeling is beyond the scope of this article, and one may browse the "List of RNA structure prediction software" for getting an idea about the growing interest in this area. Nevertheless, some general observations, as summarized below, may be useful in the current context.

It can be seen that most of the approaches are essentially limited to the prediction of RNA 2D stem-loop structure, also referred to as RNA secondary structure. For example, minimum computed free energy prediction of double helical regions of RNA sequences from the energy of base pairing and stacking interactions, essentially computationally derived from experimental thermodynamic data, was initially introduced by Ruth Nussinov and later by Michael Zuker. This, in turn, has inspired several related modified algorithms, including data on neighboring group interactions etc.[67] Most of these approaches, however, mainly consider data on canonical base pairing, with only a few which also consider thermodynamic data on Hoogsteen base pairs. Thus, in addition to the computational costs and complications associated with the identification of pseudoknots, all these methods also suffer from the drawback associated with the paucity of experimental data on non-canonical base pairs.

However, there are also several approaches which attempt at predicting the tertiary 3D structure corresponding to given predicted 2D structures. There are also a few involving 3D fragment based modeling[64], which are getting further facilitated with the increasing availability of motif wise curated RNA 3D structure data[68]. It is also, encouraging to note that there are now some software and servers, such as MC-Fold[69], RNAPDBee[70], RNAWolfe[71], etc. available for exploring non-canonical base pairing in RNA 3D structures. Some of these methods depend on structural database of RNA, such as FRABASE[64], to obtain 3D coordinates of motifs containing non-canonical base pairs and stitch the information with 3D structure of double helices containing canonical base pairs.

It may be relevant in this context, to mention about the approach towards 3D model building of double helical regions with both canonical and non-canonical base pairs used in 3DNA by Olson[41] or in RNAHelix by Bhattacharyya and Bansal[72]. These software suites use base pair parameters to generate 3D coordinates of individual dinucleotide steps, which can be extended to model double helices of arbitrary lengths with canonical or non-canonical base pairs.

The above mentioned methods attempt to model a single structure (2D or 3D) of a given

RNA sequence. However, growing evidences indicate that a given RNA sequence can adopt ensemble of structures and possibly interconvert between them[73]. This ensembles obviously adopt different base pairing patterns between different sets of residues[74]. Thus, there are enough pointers to suggest that the focus on modeling single structures appears to have been a bottleneck for accurate modeling of RNA structure.

The theoretical prediction of RNA 2D structure and consequently 3D structure can also be confirmed by different chemical probing methods. One of the latest such tools is SHAPE (Selective 2′-hydroxyl acylation analyzed by primer extension), and SHAPE-Directed RNA Secondary Structure Prediction[75] appears to be most promising. Coupled with mutational profiling, ensembles of RNA structures, which often include non-canonical base pairing, can be experimentally studied using the SHAPE-MaP approach[76]. One of the ways ahead today appears to be an integration of Zuker’s minimum free energy approach with experimentally derived SHAPE data, including simulated SHAPE data as outlined in[77],[78].

Conclusion

Hydrogen bond mediated interactions between nucleotide bases, leading to base-pair formation, constitute one of the most important class of attractive interactions which shape the structure, dynamics and function of nucleic acids. With the determination of the structure of double stranded DNA molecules fueling the development and phenomenal growth in the area of molecular biology, for a long time, nucleic acid research was focused primarily around the canonical G:C and A:T/U canonical base pairs. However, even in DNA, other types of base pairings, involving different geometries and base pairing partners, have been drawing attention in the context of structural and functional diversity. Occurrence of these non-canonical base pairs are far more abundant in RNA, where a single strand folds on to itself, often without the possibility of complementary canonical base pairs to stabilize the folds. The picture that emerges from ongoing research in the context of diverse structure, dynamics and function of RNA, is that the diversity may be rationalized in terms of the structure, dynamics and stabilities of over more than 100 types of base pairs, including non-canonical base pairs. The role of G:U W:W *cis* base pairs in the context of the Wobble hypothesis, or the Hoogsteen base pairing in the context of triple helices and G quartet formation were initial indicators. Most of the tertiary interactions shaping the complex folding and functions of 3D RNA are mediated through non-canonical base pairs. What is particularly notable is that non-canonical base pairs are capable of creating appropriate localized distortions to provide functionally important structural variations, not only in RNA, but even in double stranded DNA. This becomes even more significant in the context of non-canonical base pairs, occurring in the A-type double stranded regions of functional RNAs, which play an important role in molecular recognition of base sequence by locally distorting the otherwise inaccessible major groove. Thus, the field of non-canonical base pairing is still quite open for scientific contributions from different directions. In particular, a comprehensive characterization of non-canonical base pairs will have a far reaching impact on RNA biotechnology, both, in terms of prediction of structure as well as in terms of enriching our molecular level understanding of the functioning of non (protein) coding RNA.

References:

1 J. D. Watson and F. H. Crick, A structure for deoxyribose nucleic acid. 1953., *Nature*.

2 E. N. Nikolova, H. Zhou, F. L. Gottardo, H. S. Alvey, I. J. Kimsey and H. M. Al-Hashimi, *Biopolymers*, 2013.

3 E. Westhof and V. Fritsch, *Structure*, 2000, 8, R55–R65.

4 K. Hoogsteen, The structure of crystals containing a hydrogen-bonded complex of 1-methylthymine and 9-methyladenine, *Acta Crystallogr.*, 1959, **12**, 822–823.

5 Y. Courtois, P. Fromageot and W. Guschlbauer, Protonated Polynucleotide Structures: 3. An Optical

Rotatory Dispersion Study of the Protonation of DNA, *Eur. J. Biochem.*, 1968, **6**, 493–501.

6 D. J. Patel and A. E. Tonelli, Assignment of the proton nmr chemical shifts of the T□N3H and G□N1H proton resonances in isolated AT and GC Watson-Crick base pairs in double-stranded deoxy oligonucleotides in aqueous solution, *Biopolymers*, 1974, **13**, 1943–1964.

7 N. C. Seeman, J. M. Rosenberg, F. L. Suddath, J. J. Kim and A. Rich, RNA double-helical fragments at atomic resolution. I. The crystal and molecular structure of sodium adenylyl-3',5'-uridine hexahydrate., *J. Mol. Biol.*, 1976, **104**, 109–44.

8 H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura and R. E. Dickerson, Structure of a B-DNA dodecamer: conformation and dynamics., *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 2179–2183.

9 RNA Basepair Catalog, http://ndbserver.rutgers.edu/ndbmodule/services/BPCatalog/bpCatalog.html, (accessed 5 December 2019).

10 A. H. J. Wang, S. Fujii, J. H. Van Boom, G. A. Van Der Marel, S. A. A. Van Boeckel and A. Rich, Molecular structure of r(GCG)d(TATACGC): A DNA-RNA hybrid helix joined to double helical DNA, *Nature*, 1982, **299**, 601–604.

11 U. Heinemann and C. Alings, Crystallographic study of one turn of G/C-rich B-DNA, *J. Mol. Biol.*, 1989, **210**, 369–381.

12 A. C. Dock-Bregeon, B. Chevrier, A. Podjarny, J. Johnson, J. S. de Bear, G. R. Gough, P. T. Gilham and D. Moras, Crystallographic structure of an RNA helix: [U(UA)6A]2, *J. Mol. Biol.*, 1989, **209**, 459–474.

13 G. A. Patikoglou, J. L. Kim, L. Sun, S. H. Yang, T. Kodadek and S. K. Burley, TATA element recognition by the TATA box-binding protein has been conserved throughout evolution, *Genes Dev.*, 1999, **13**, 3217–3230.

14 J. Aishima, R. K. Gitti, J. E. Noah, H. H. Gan, T. Schlick and C. Wolberger, *Nucleic Acids Res.*, 2002, 30, 5244–5252.

15 D. T. Nair, R. E. Johnson, S. Prakash, L. Prakash and A. K. Aggarwal, Replication by human DNA polymerase-ι occurs by Hoogsteen base-pairing, *Nature*, 2004, **430**, 377–380.

16 M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig and Z. Shakked, Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs, *Nat. Struct. Mol. Biol.*, 2010, **17**, 423–429.

17 A. S. Ethayathulla, P. W. Tse, P. Monti, S. Nguyen, A. Inga, G. Fronza and H. Viadiu, Structure of p73 DNA-binding domain tetramer modulates p73 transactivation, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 6066–6071.

18 Y. Xu, J. McSally, I. Andricioaei and H. M. Al-Hashimi, Modulation of Hoogsteen dynamics on DNA recognition, *Nat. Commun.*, , DOI:10.1038/s41467-018-03516-1.

19 G. N. Parkinson, M. P. H. Lee and S. Neidle, Crystal structure of parallel quadruplexes from human telomeric DNA, *Nature*, 2002, **417**, 876–880.

20 K. N. Luu, A. T. Phan, V. Kuryavyi, L. Lacroix and D. J. Patel, Structure of the human telomere in K+ solution: An intramolecular (3 + 1) G-quadruplex scaffold, *J. Am. Chem. Soc.*, 2006, **128**, 9963–9970.

21 A. T. Phan, V. Kuryavyi, K. N. Luu and D. J. Patel, Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K+ solution, *Nucleic Acids Res.*, 2007, **35**, 6517–6525.

22 D. K. Hendrix, S. E. Brenner and S. R. Holbrook, *Q. Rev. Biophys.*, 2005, 38, 221–243.

23 C. Laing, S. Jung, A. Iqbal and T. Schlick, Tertiary Motifs Revealed in Analyses of Higher-Order RNA Junctions, *J. Mol. Biol.*, 2009, **393**, 67–82.

24 S. Halder and D. Bhattacharyya, *Prog. Biophys. Mol. Biol.*, 2013, 113, 264–283.

25 P. Ananth, G. Goldsmith and N. Yathindra, *RNA*, 2013, 19, 1038–1053.

26 J. Šponer, J. Leszczynski and P. Hobza, Structures and energies of hydrogen-bonded DNA base pairs. A nonempirical study with inclusion of electron correlation, *J. Phys. Chem.*, 1996, **100**, 1965–1974.

27 W. Saenger, *Principles of nucleic acid structure*, Springer-Verlag, 1984.

28 M. T. Sykes and M. Levitt, Describing RNA structure by libraries of clustered nucleotide doublets, *J. Mol. Biol.*, 2005, **351**, 26–38.

29 N. B. Leontis and E. Westhof, Geometric nomenclature and classification of RNA base pairs, *RNA*, 2001, **7**, 499–512.

30 G. R. (Gautam R. . Desiraju and T. Steiner, *The weak hydrogen bond : in structural chemistry and biology*, Oxford University Press, 2001.

31 N. B. Leontis, The non-Watson-Crick base pairs and their associated isostericity matrices, *Nucleic Acids Res.*, 2002, **30**, 3497–3531.

32 L. Nasalean, J. Stombaugh, C. L. Zirbel and N. B. Leontis, in *Non-Protein Coding RNAs*, Springer Berlin Heidelberg, 2008, pp. 1–26.

33 J. Stombaugh, C. L. Zirbel, E. Westhof and N. B. Leontis, Frequency and isostericity of RNA base

pairs, *Nucleic Acids Res.*, 2009, **37**, 2294–2312.

34 J. E. Sokoloski, S. A. Godfrey, S. E. Dombrowski and P. C. Bevilacqua, Prevalence of syn nucleobases in the active sites of functional RNAs, *RNA*, 2011, **17**, 1775–1787.

35 J. Reichert and J. Sühnel, The IMB Jena Image Library of Biological Macromolecules: 2002 update, *Nucleic Acids Res.*, 2002, **30**, 253–254.

36 RNA Base Pair Database(RNABPDB), http://hdrnas.saha.ac.in/rnabpdb/, (accessed 5 December 2019).

37 S. Bhattacharya, S. Mittal, S. Panigrahi, P. Sharma, P. S. P., R. Paul, S. Halder, A. Halder, D. Bhattacharyya and A. Mitra, RNABP COGEST: a resource for investigating functional RNAs, *Database*, , DOI:10.1093/database/bav011.

38 M. Chawla, R. Oliva, J. M. Bujnicki and L. Cavallo, An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies, *Nucleic Acids Res.*, 2015, **43**, 6714–6729.

39 P. P. Seelam, P. Sharma and A. Mitra, Structural landscape of base pairs containing post-transcriptional modifications in RNA, *RNA*, 2017, **23**, 847–859.

40 K. Darty, A. Denise and Y. Ponty, VARNA: Interactive drawing and editing of the RNA secondary structure, *Bioinformatics*, 2009, **25**, 1974–1975.

41 X. J. Lu and W. K. Olson, 3DNA: A software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Res.*, 2003, **31**, 5108–5121.

42 H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S. H. Hsieh, A. R. Srinivasan and B. Schneider, The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids, *Biophys. J.*, 1992, **63**, 751–759.

43 M. Sarver, C. L. Zirbel, J. Stombaugh, A. Mokdad and N. B. Leontis, FR3D: Finding local and composite recurrent structural motifs in RNA 3D structures, *J. Math. Biol.*, 2008, **56**, 215–252.

44 S. Lemieux, F. Major, N. Kim, C. Laing, S. Elmetwaly, S. Jung, J. Curuksu and T. Schlick, RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire\rGraph-based sampling for approximating global helical topologies of RNA, *Nucleic Acids Res*, 2002, **30**, 4250–4263.

45 C. I. Bayly, K. M. Merz, D. M. Ferguson, W. D. Cornell, T. Fox, J. W. Caldwell, P. A. Kollman, P. Cieplak, I. R. Gould and D. C. Spellmeyer, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.

46 J. Das, S. Mukherjee, A. Mitra and D. Bhattacharyya, Non-canonical base pairs and higher order

structures in nucleic acids: Crystal structure database analysis, *J. Biomol. Struct. Dyn.*, , DOI:10.1080/07391102.2006.10507108.

47 M. Chawla, P. Sharma, S. Halder, D. Bhattacharyya and A. Mitra, Protonation of base Pairs in RNA: Context analysis and quantum chemical investigations of their geometries and stabilities, *J. Phys. Chem. B*, , DOI:10.1021/jp106848h.

48 R. E. A. Kelly, Y. J. Lee and L. N. Kantorovich, Homopairing possibilities of the DNA base adenine, *J. Phys. Chem. B*, 2005, **109**, 11933–11939.

49 J. E. Sponer, J. Leszczynski, V. Sychrovský and J. Sponer, Sugar edge/sugar edge base pairs in RNA: stabilities and structures from quantum chemical calculations., *J. Phys. Chem. B*, 2005, **109**, 18680–9.

50 R. Oliva, L. Cavallo and A. Tramontano, Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions., *Nucleic Acids Res.*, 2006, **34**, 865–79.

51 D. Bhattacharyya, S. C. Koripella, A. Mitra, V. B. Rajendran and B. Sinha, in *Journal of Biosciences*, 2007, vol. 32, pp. 809–825.

52 A. Roy, S. Panigrahi, M. Bhattacharyya and D. Bhattacharyya, Structure, stability, and dynamics of canonical and noncanonical base pairs: Quantum chemical studies, *J. Phys. Chem. B*, , DOI:10.1021/jp076921e.

53 P. Sharma, A. Mitra, S. Sharma, H. Singh and D. Bhattacharyya, Quantum chemical studies of structures and binding in noncanonical rna base pairs: The trans watson-crick:watson-crick family, *J. Biomol. Struct. Dyn.*, , DOI:10.1080/07391102.2008.10507216.

54 P. Sharma, J. E. Šponer, J. Šponer, S. Sharma, D. Bhattacharyya and A. Mitra, On the role of the eis hoogsteen:Sugar-Edge family of base pairs in platforms and Triplets-Quantum chemical insights into RNA structural biology, *J. Phys. Chem. B*, 2010, **114**, 3307–3320.

55 O. O. Brovarets, Y. P. Yurenko and D. M. Hovorun, Intermolecular CH···O/N H-bonds in the biologically important pairs of natural nucleobases: A thorough quantum-chemical study, *J. Biomol. Struct. Dyn.*, 2014, **32**, 993–1022.

56 T. Marino, DFT investigation of the mismatched base pairs (T-Hg-T)3, (U-Hg-U)3, d(T-Hg-T)2, and d(U-Hg-U)2, *J. Mol. Model.*, , DOI:10.1007/s00894-014-2303-8.

57 A. Mládek, P. Sharma, A. Mitra, D. Bhattacharyya, J. Šponer and J. E. Šponer, Trans Hoogsteen/sugar edge base pairing in RNA. Structures, energies, and stabilities from quantum chemical calculations, *J. Phys. Chem. B*, , DOI:10.1021/jp808357m.

58 R. E. Dickerson, Definitions and nomenclature of nucleic acid structure components, *Nucleic Acids Res.*, 1989, **17**, 1797–1803.

59 C. Blanchet, M. Pasi, K. Zakrzewska and R. Lavery, CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures, *Nucleic Acids Res.*, 2011, **39**, W68–W73.

60 X. J. Lu and W. K. Olson, 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures, *Nat. Protoc.*, 2008, **3**, 1213–1227.

61 NUPARM-Plus-A Program for analyzing sequence dependent variations in nucleic acids, http://nucleix.mbu.iisc.ernet.in/nuparm/, (accessed 5 December 2019).

62 A. S. Abu Almakarem, A. I. Petrov, J. Stombaugh, C. L. Zirbel and N. B. Leontis, Comprehensive survey and geometric classification of base triples in RNA structures, *Nucleic Acids Res.*, 2012, **40**, 1407–1423.

63 S. Bhattacharya, A. Jhunjhunwala, A. Halder, D. Bhattacharyya and A. Mitra, Going beyond base-pairs: Topology-based characterization of base-multiplets in RNA, *RNA*, 2019, **25**, 573–589.

64 M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E. K. Burke, J. Blazewicz and R. W. Adamiak, RNA FRABASE 2.0: An advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures, *BMC Bioinformatics*, , DOI:10.1186/1471-2105-11-231.

65 H. Y. Hamdani, S. D. Appasamy, P. Willett, P. J. Artymiuk and M. Firdaus-Raih, NASSAM: A server to search for and annotate tertiary interactions and motifs in three-dimensional structures of complex RNA molecules, *Nucleic Acids Res.*, , DOI:10.1093/nar/gks513.

66 RNA 3D Motif Atlas, http://rna.bgsu.edu/rna3dhub/motifs, (accessed 5 December 2019).

67 Y. Tabei, K. Tsuda, T. Kin and K. Asai, SCARNA: fast and accurate structural alignment of RNA sequences by matching fixed-length stem fragments., *Bioinformatics*, 2006, **22**, 1723–9.

68 E. Baulin, V. Yacovlev, D. Khachko, S. Spirin and M. Roytberg, URS DataBase: universe of RNA structures and their motifs, *Database (Oxford).*, , DOI:10.1093/database/baw085.

69 M. Parisien and F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data, *Nature*, 2008, **452**, 51–55.

70 T. Zok, M. Antczak, M. Zurkowski, M. Popenda, J. Blazewicz, R. W. Adamiak and M. Szachniuk, RNApdbee 2.0: Multifunctional tool for RNA structure annotation, *Nucleic Acids Res.*, 2018, **46**, W30–W35.

71 A. Rybarczyk, N. Szostak, M. Antczak, T. Zok, M. Popenda, R. Adamiak, J. Blazewicz and M. Szachniuk, New in silico approach to assessing RNA secondary structures with non-canonical base pairs., *BMC Bioinformatics*, 2015, **16**, 276.

72 D. Bhattacharyya, S. Halder, S. Basu, D. Mukherjee, P. Kumar and M. Bansal, RNAHelix: computational modeling of nucleic acid structures with Watson–Crick and non-canonical base pairs, *J. Comput. Aided. Mol. Des.*, , DOI:10.1007/s10822-016-0007-0.

73 J. A. Cruz and E. Westhof, *Cell*, 2009, 136, 604–609.

74 P. S. Ray, J. Jia, P. Yao, M. Majumder, M. Hatzoglou and P. L. Fox, A stress-responsive RNA switch regulates VEGFA expression, *Nature*, 2009, **457**, 915–919.

75 J. T. Low and K. M. Weeks, *Methods*, 2010, 52, 150–158.

76 N. A. Siegfried, S. Busan, G. M. Rice, J. A. E. Nelson and K. M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP), *Nat. Methods*, 2014, **11**, 959–965.

77 S. Montaseri, M. Ganjtabesh and F. Zare-Mirakabad, Evolutionary Algorithm for RNA Secondary Structure Prediction Based on Simulated SHAPE Data, *PLoS One*, 2016, **11**, e0166965.

78 A. Spasic, S. M. Assmann, P. C. Bevilacqua and D. H. Mathews, Modeling RNA secondary structure folding ensembles using SHAPE mapping data., *Nucleic Acids Res.*, 2018, **46**, 314–323.