A Comprehensive Survey on the Nature of Ring:Ring Nucleobase Stacking Interactions in RNA: Occurrence, Structural Variability and Classification of the Associated Contacts.

Ayush Jhunjhunwala,[†] Zakir Ali,[‡] Sohini Bhattacharya,[†] Antarip Halder,[†] Abhijit Mitra[†,*] and Purshotam Sharma[‡,*]

[†]Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology (IIIT-H), Gachibowli, Hyderabad, Telangana, 500032, India.

[‡]Computational Biochemistry Laboratory, Department of Chemistry and Centre for Advanced Studies in Chemistry, Panjab University, Chandigarh, 160014, India.

*email: abi_chem@iiit.ac.in (AM), psharma@pu.ac.in (PS)

ABSTRACT

The astonishing diversity in folding patterns of RNA 3D structures is crafted by myriads of noncovalent contacts, of which base pairing and stacking are the most prominent. Although the classification scheme proposed by Leontis and Westhof (RNA (2001), **7**, 499–512) has been widely accepted for annotation of RNA base pairs, the absence of an unambiguous classification system for base stacks appears to be a roadblock for exploring the stacking diversity in RNA. Here we provide a structurally-intuitive and unambiguous geometry cum topology based classification scheme for base stacking, where a base stack is classified in terms of interacting nucleobase faces and the relative orientation of the glycosidic bonds. This generates eight or six basic stacking geometrical families for heterodimeric or homodimeric stacks. Further classification in terms of the identity of the bases, and the nature of involvement of the purines (5-membered, 6-membered or both rings) leads to 384 distinct RNA base stacks. In addition, we develop an algorithm for automated identification and classification of stacks in RNA crystal structures, and investigate the stacking geometries within selected RNA structures. Overall, our comprehensive classification scheme of RNA base stacking is expected to facilitate structure-based RNA research.

INTRODUCTION

The exponential growth in the availability of atomic resolution RNA structures in recent times has not only highlighted the astonishing structural diversity of RNA, but has also created the basis for molecular level investigation of how it is related with the functional diversity, which includes regulatory and catalytic roles. As a result, it has become apparent that the earlier prevalent methodological apparatus used for investigating structural details of double stranded DNA need a major overhaul in order to deal with the complexity of RNA 3D structures.

The apparently complex RNA structures can be easily understood in terms of a hierarchy, by considering the 3D structures as a collection of 2D stem-loop motifs interconnected through noncovalent interactions between non-consecutive ribonucleotides in the form of stacking and base pairing (Figure 1). Although stacking interactions involve *nonspecific* face-to-face van der Waals contacts between nucleobases, base pairing consists of *specific* edge-to-edge hydrogen bonding between two nucleotides. Since both these attractive interactions give rise to diverse geometries, in terms of accessible nucleobase orientations, the development of a reliable method for automated identification and annotation of base pairs and base stacks is essential for investigating the structural diversity of RNA.

Though the two complementary strands of DNA double helices are held together by canonical A:T (A:U in RNA) and G:C base pairs, it has long since been recognised that there are hot spots involving noncanonical base pairs in DNA that involve different nucleobase edges and different nucleobases. Accordingly, there have been attempts, most notably by Saenger (1), at annotating the geometries of nucleic acid base pairs. Saenger's annotation or nomenclature scheme, however, had problems in dealing with RNA base pairs which involve a significant variety of noncanonical base pairs (2). In this context, shortly after the publication of the high resolution crystal structures of ribosome at the turn of century (3-6), Leontis and Westhof proposed an excellent classification scheme for the naming of all possible RNA base pairs (2). This scheme characterizes RNA base pairs in terms of three interaction edges (Watson–Crick (W) edge, Hoogsteen (H) edge, and Sugar (S) edge) and the mutual glycosidic orientation (*cis* and *trans*) of the interacting nucleotides, and classifies RNA base-pairs into 12 distinct geometric families (i.e. W:W, W:H, W:S, H:H, H:S or S:S families in *cis* or *trans* orientations). In addition to providing a framework for understanding the structural implications of sequence-based homology between geometrically-similar base pairs, this scheme facilitated the analysis of intrinsic structural characteristics and dynamical

stabilities of RNA base-pairs (see, for example, refs. 7-10). Subsequent studies, while extending this scheme to classify and annotate higher order structures such as base triplets and quartets in RNA (11,12) highlighted the need of supplementing the geometric framework with topological considerations.

Notwithstanding the availability of a classification scheme for base-pairing interactions, the annotation of attractive interactions stabilizing RNA 3D structures is incomplete, in the absence of a comprehensive classification scheme for RNA base stacks. In fact, very few studies have focussed on understanding the diversity of stacking interactions in RNA (13). Nevertheless, at least four nucleic acid annotation and visualization tools exist in literature that identify stacking interactions. These include MC-Annotate (14), ClaRNA (15), 3DNA (16) and FR3D (Find RNA 3D) (17). Specifically, MC-Annotate and ClaRNA use the stacking algorithm previously developed by Gabb et al. (13), which considers two bases to be stacked if the distance between their ring centres is less than 5.5 Å, the angle between the normals to the base planes is less than 30°, and the angle between the normal of one base plane and the vector connecting the centre of the two rings is less than 40°. However, ClaRNA introduces additional parameterization to measure the overlap area of bases (15), although the identity of the overlapping rings (5- or 6-membered) remains ambiguous in the output. Further, 3DNA program developed by Lu and co-workers classifies stacking interaction from planar projection of each ring as well as exocyclic atoms in consecutive bases or base pairs (16). To detect stacking interactions, this algorithm defines a middle reference frame between two potentially stacked bases or base pairs and quantifies the stacking interaction by evaluating the shared overlap areas of the base rings. Further, to account for stacking overlap involving exocyclic nucleobase substituents, an extended polygon including exocyclic atoms is used. While this approach may be useful for capturing N-H•••π type interactions involving exocyclic amino groups, in addition to the π – π stacking instances, the output of this program does not differentiate between stacks involving two different nucleobase interacting faces. Similar to what was noted while classifying higher order base multiplets (12), there is a need for introducing topological considerations.

In a way, the issue of the identity of interacting faces was addressed by Sarver *et al.* who classified the stacked bases according to the orientation of the interacting face that occurs in a regular helix within the FR3D database (17). Specifically, FR3D identifies the stacking interactions based on non-zero overlap of the convex hull of each base projected vertically onto the convex hull of the other. The side of the base that orients towards the 3′-

end in a regular helix is designated as the 3′ face, and the other face is called the 5′ face. The base stacks are accordingly classified using an alpha-numeric notation as X-Y s35/s55/s33, where X and Y represent the stacked nucleotides, the letter 's' denotes the stacking interaction, and the numbers 35, 55 or 33 denote the identity of the interacting faces (3′ or 5′) of the stacked bases. Although this scheme distinguishes between interstrand and intrastrand stacking, and identifies stacking interactions involving exocyclic base atoms, it considers the definition of stacking faces based on the strand direction, which, because of variations in the relative glycosidic orientation of the stacked bases, can often be nonintuitive in complex RNA structures. Further, the scheme does not consider the extent of overlap of the stacked nucleobase rings and thus fails to comprehensively annotate the inherently complex intermolecular interaction hyperspace of stacked bases characterized by multiple, closely-spaced minima that strongly depend on the relative orientations and extent of overlap of the interacting bases (18). Though the annotation of relative orientation of the stacked bases is a general concern for both purines and pyrimidines, the extent of overlap is specifically problematic in case of a purine, which can either stack with another purine/pyrimidine through its 5-membered ring, 6-membered ring or both rings.

To overcome these shortcomings and to facilitate comprehensive characterization and annotation of RNA base stacks, here we provide an unambiguous and structurally-simple geometry cum topology-based classification scheme for RNA base stacks. We classify each RNA base stack in terms of the nucleobase face (α or β) participating in stacking and the mutual orientation of the glycosidic bonds (*cis* or *trans*), which gives rise to eight basic stacking geometrical families. We further classify each stacking family in terms of the identity of the purine ring (5-membered, 6-membered or both) participating in stacking, which leads to a number of distinct stacking topologies that depend on the identity of the stacked nucleobases. In addition, we identify examples of different stacking geometries in RNA crystal structures to verify their occurrence in the macromolecular context. Overall, the major goals of this article include: (i) introduction of a geometrical cum topological classification system for base stacking interactions in RNA, (ii) identification of examples of different base stacks, and (iii) to investigate the occurrence context of selected base stacks in RNA 3D structures.

MATERIALS AND METHODS

A dataset containing all RNA crystal structures deposited in the Protein Data Bank (PDB) till 21st March 2019 with the resolution of up to 3.5 Å was used to identify the occurrences of

base stacking geometries considered in the present work (Table S1). To detect stacking interactions, the purine five-membered ring consisting of C4, C5, N7, C8 and N9 atoms and the purine or pyrimidine six-membered ring consisting of N1, C2, N3, C4, C5 and C6 atoms were considered (Figure S1). Each PDB file was parsed using a python script to extract the Cartesian coordinates of atoms constituting these rings, as well as the C1′ of the attached sugar. As part of exception handling, nucleotides with incomplete information (generally missing ring atoms), or cases where two nucleotides belonging to different chains possess same atomic coordinates (e.g. in PDB code: 488d), were ignored.

To calculate ring-ring stacking, two types of rings were considered- the six-membered heterocyclic ring of each canonical (purine or pyrimidine) nucleobase and the five-membered heterocyclic ring of each purine (adenine or guanine). As a first step, the coordinates of centroid of each ring (*i.e.* x_c, y_c, z_c) were determined using the cartesian coordinates (*i.e.* (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3),……(x_n, y_n, z_n)) of 1, 2, ….n (n = number of atoms that constitute the ring (5 or 6)) atoms that constitute the ring, using the following relations:

$$x_c = \frac{x_1 + x_2 + \ldots + x_n}{n}, y_c = \frac{y_1 + y_2 + \ldots + y_n}{n}, z_c = \frac{z_1 + z_2 + \ldots + z_n}{n}$$

Subsequently, the position vector $\vec{r}_i$ of each ith ring atom was determined by using the cartesian coordinates (x_i, y_i, z_i) of the atom i, and the coordinates of the ring centroid. Specifically,

$$\vec{r}_i = (x_i - x_c)\hat{\imath} + (y_i - y_c)\hat{\jmath} + (z_i - z_c)\hat{k}$$

Further, the distance vector ($\vec{d}_{\mathrm{ab}}$, Figure 2) that connects the centroids of two stacked rings *a* and *b* was determined using the coordinates of their centroids (x_{ac}, y_{ac}, z_{ac} and x_{bc}, y_{bc}, z_{bc} respectively).

$$\vec{d}_{\mathrm{ab}} = (x_{bc} - x_{ac})\hat{\imath} + (y_{bc} - y_{ac})\hat{\jmath} + (z_{bc} - z_{ac})\hat{k}$$

To calculate ring-ring stacking, a mean plane for each ring was defined using the two orthogonal vectors $\vec{R}_1$ and $\vec{R}_2$ with the help of the method suggested by Cremer and Pople (28). Specifically,

$$\vec{R}_1 = \sum_{i=1}^{n} \vec{r}_i \, cos[2\pi(i-1)/n], \vec{R}_2 = \sum_{i=1}^{n} \vec{r}_i \, sin[2\pi(i-1)/n]$$

Further, $\vec{N}$, the vector normal to the mean plane vectors $\vec{R}_1$ and $\vec{R}_2$ of each ring (Figure 2), was determined as their cross product.

$$\vec{N} = \frac{\vec{R}_1 \times \vec{R}_2}{|\vec{R}_1 \times \vec{R}_2|}$$

$\vec{N}$ was then used to calculate θ_{ab} (Figure 2), the angle between the normals ($\vec{N}_a$ and $\vec{N}_b$) of the stacked rings *a* and *b*, as previously suggested by Gaab *et al* (22). Specifically,

$$\theta_{ab} = cos^{-1}(\vec{N}_a \cdot \vec{N}_b)$$

Further, the angles of $\vec{d}_{ab}$ with $\vec{N}_a$and $\vec{N}_b$ (*i.e.* τ_a and τ_b respectively, Figure 2) for the stacked rings were calculated as:

$$\tau_a = cos^{-1}\left[N_a \cdot \frac{\vec{d}ab}{|\vec{d}ab|}\right] \text{ and } \tau_b = cos^{-1}\left[N_b \cdot \frac{\vec{d}ab}{|\vec{d}ab|}\right]$$

θ_{ab} describes the tilt of one ring with respect to the other, whereas $\tau_{a/}\tau_b$, in conjunction with $|\vec{d}_{ab}|$, determines the horizontal shift of one monomer with respect to the other, which in turn quantifies the extent of overlap of one base/ring with respect to the other. For locating all the stacked pairs in RNA crystal structures, the following criteria established in a previous study (13), were used was used: $\vec{d}_{ab}$ < 5.5 Å, θ_{ab} < 30°, τ_a or τ_b < 40°.

Each nucleobase stack, thus identified, was further annotated as *consecutive* or *non-consecutive*, depending on the absolute value of the difference between the nucleotide residue numbers in the PDB file. Here, if the absolute value of the difference between the nucleotide numbers equals 1, then the bases are annotated as *consecutive* stacks, else they are annotated as *non-consecutive* stacks (Figure 3A). Further, the mutual orientation of glycosidic bonds of the stacked nucleotides a and b was determined by calculating the torsion angle (σ_{ab}), defined by $(C1')_a$–O_a–O_b– $(C1')_b$, where O_a and O_b are the coordinates of the centroids of the respective base rings connected to sugar through glycosidic bond (five membered ring in case of purines and six membered ring in case of pyrimidines). The orientation of the stacked bases is annotated as *cis* if $90^\circ \geq \sigma_{ab} \geq 0^\circ$ and *trans* if $180° \geq \sigma_{ab} \geq 90°$ (Figure S2). Analysis of the distribution of σ torsion angle reveals that although majority of consecutive stacks exhibit *cis* orientation, the non-consecutive stacks span a wide variation in σ, and are thereby distributed among *cis* and *trans* orientations (Figure 3B). This, in turn, suggests that the σ parameter is indeed required to unambiguously characterize base stacks in RNA structures.

In addition, the identity of the interaction face of the nucleobases (i.e. α or β), Figure S3)) were defined, without any reference to the strand direction, by using the nomenclature scheme previously proposed in the context of polycyclic systems (20). Thus, the interacting purine face is designated as α if the standard atom numbering of the 6-membered purine ring increases in the clockwise direction when viewed from the space between the stacked bases, and the other face is called β (Figure S3). Similar protocol is followed while assigning faces to pyrimidines (Figure S3). Accordingly, the faces were designated computationally, using

the direction of normal to the mean plane vectors defined for each stacked ring. Since the mean plane vectors are calculated from atom positions around the ring (N1 to N6), the direction of the normal vectors ($\vec{N}_a$ or $\vec{N}_b$) is such that it always extends from the β face. Thus, if the angle between the distance vector $\vec{d}$ and $\vec{N}_a$ or $\vec{N}_b$ is greater than 90° the face is α, else it is assigned as β.

Using the above protocol, we developed an automated method for classification of the base stacks present in RNA structures. The classification has been executed on the entire dataset of structures downloaded from PDB (*vide supra*). Further, this automated method for classification is made accessible through the web server (http://stackdetect.iiit.ac.in), where a user can generate an output file containing information on each base stack present in a given RNA PDB file, along with the values of the relevant classification parameters (i.e. the extent of overlap, mutual glycosidic orientation, distance between nucleotides in the sequence space as well as the geometrical paramters ($\vec{d}_{ab}$, θ_{ab}, τ_a, τ_b and σ_{ab})). The consolidated output files that characterize each stack identified in PDB dataset in terms of its type and the geometrical parameters ($\vec{d}_{ab}$, θ_{ab}, τ_a, τ_b and σ_{ab}) are also provided on the web portal (http://stackdetect.iiit.ac.in/tnd/).

RESULTS AND DISCUSSION

Classification of RNA base stacks: basic base stacking families

As explained above, RNA nucleobases possess two stacking faces α and β, (Figure S3) which can be designated in absolute terms, i.e. without any reference to strand directions. This leads to four combinations of face orientations for each sequence of stacked bases: α||α, α||β, β||α and β||β (Figure 4). However, due to intrinsic self-symmetry, α||β and β||α arrangements are identical for homodimer stacks (i.e. A||A, C||C, G||G and U||U), which leads to only three distinct face orientations (α||α, α||β and β||β) for such stacks.

Each of the three (homodimeric) or four (heterodimeric) stacking-face combinations can be further classified in terms of the (*cis* and *trans*) glycosidic orientations (Figure 4). In analogy with the base pairing nomenclature of Leontis and Westhof (2), if the glycosidic bonds of the stacked nucleotides point in the same direction, the orientation is designated as *cis*, whereas the other orientation is designated as *trans* (Figure S2). Although it is theoretically possible to use the mutual glycosidic orientation of the stacked bases as a parameter for understanding the similarity and interchangeability of stacking contacts in RNA, unlike base pairing (2), the proof of evolutionary covariation in RNA base stacks is largely nonexistent in literature. This probably stems from the fact that unlike specific and

directional nature of RNA base pairing contacts (2), stacking contacts are nonspecific and dispersive in nature. Nevertheless, a detailed future comparative analysis of specific structural and functional contexts of base stacks in homologous 3D RNA structures may probably help evolve a specific geometrical parameter for understanding covariation patterns (if any) in RNA base stacks.

Overall, the consideration of mutual orientation of the glycosidic bonds, in conjunction with the stacking-face combinations lead to 8 theoretically possible stacking geometries for heterodimer stacks and 6 geometries for homodimer stacks.

Topological variants of RNA base stacks

Purine||Purine stacks: Depending on whether the five-membered, six-membered or both rings participate in stacking, a heteropurine stack (i.e. A||G) may involve 15 distinct topologies. (Figure S4A). In contrast, each homopurine stack (A||A or G||G) may involve 11 topologies if both the purines interact with the same face (i.e. α||α or β||β, Figure S4B). However, if the interacting faces are different (i.e. α||β or β||α), the homopurine stacks may involve 15 topologies (Figure S4A). The combination of 15 A||G topologies with four face orientations (α||α, α||β, β||α and β||β) and two glycosidic orientations (*cis* and *trans*) leads to 120 distinct stacking arrangements. In contrast, the combination of 11 topologies each for α||α and β||β, and of 15 topologies for α||β face, with two glycosidic orientations, leads to 74 types of stacks for A||A and G||G respectively. Overall, this leads to (120 + 74*2) 268 distinct purine||purine stacking arrangements.

Among the purine||purine stacks, the three most dominant topologies involve the cis orientation (Table S2). The most prominent topology among these is provided by the 5||5, 6||56 α||β cis stacking arrangement, since it maximizes the stacking overlap, and can occur between two consecutive stacked bases (Table S2). However, 10 cis stacking arrangements and 10 *trans* arrangements are extremely uncommon (Table S3), since they can occur only at uncommon σ values (close to 90°). Consequently, one example each of only two of these topologies were found in the crystal structures at σ of 83.5° and 85.3° (Table S3). In contrast, 20 additional stacking arrangements possible at reasonable σ values were not found in the analyzed structures. These arrangements can, however, possibly occur in RNA crystal structures available in future (Table S4). Overall, examples of 230 out of 268 purine||purine stacking arrangements were found in the dataset (Tables S5 – S124).

Purine||Pyrimidine and Pyrimidine||Pyrimidine stacks: Since pyrimidines contain only one (i.e. six-membered) ring, the stacking of pyrimidines with five membered and/or six-

membered ring of purines leads to 3 distinct topologies (Figure S4C). The combination of these 3 topologies with 4 face orientations (Figure S3) and 2 glycosidic orientations (Figure 4) leads to 24 each of A||C, A||U, C||G and G||U stacks. Overall, this leads to 96 distinct purine||pyrimidine stacking arrangements. Of these, the stacking arrangement involving the combination of the maximum-overlap topology (5||6, 6||6) and α||α *cis* orientation is the most dominant one, and occurs in more than a third of purine||pyrimidine stacks detected (Table S125). However, unlike as observed in purine||purine stacks, examples of all 96 possible stacking arrangements were encountered in the analysed dataset (Tables S126 – S149).

In contrast, pyrimidine||pyrimidine stacks involve only one topology (Figure S4D). This, in conjunction with 4 face orientations and 2 glycosidic orientations leads to 8 distinct C||U stacking geometries. Similarly, the combination of 3 face orientations and 2 glycosidic orientations leads to 6 C||C and 6 U||U stacks. This leads to 20 distinct pyrimidine||pyrimidine stacking interactions. Among these, the α||β (or β||α) *cis* is the most prominent stacking arrangement, which occurs in majority (~93%) of the cases (Table S150). As in the case of purine||pyrimidine stacks, examples of all 20 stacking arrangements were also found in the RNA structures (Tables S151 – S158).

Overall 268 purine||purine, 96 purine||pyrimidine and 20 pyrimidine||pyrimidine stacks give rise to 384 distinct topological variants of RNA base stacks.

Examples of base stacking geometries in RNA crystal structures

Each stacking arrangement can be designated by stating the base sequence, the interacting faces, respectively, of each of the two bases (α or β face), the relative glycosidic bond orientation (*cis* or *trans*) and the stacking topology. Figure 4 explains the protocol for identification of interacting faces of a base stack using the U513||C512 α||β *cis* stack found in glutamyl-tRNA (PDB: 1g59, (21)). When viewed from the direction of the interacting face in the stack, the numbering of atoms in C512 increments in the counter clockwise direction. Hence, the interacting face is designated as β (Figure 4B). On the other hand, when similarly viewed, the numbering of atoms in U513 increments in clockwise progression (Figure 4C) and the face is assigned as α. This, in combination with the *cis* glycosidic orientation (Figure 4D) leads to annotation of this particular stack as U513||C512 α||β *cis*.

Although similar naming convention can be used in case of purine||purine and purine||pyrimidine stacks, additional information regarding the stacking topology also needs to be specified in these cases. For example, the designation of A2841||G2842 α||α 5||56 *cis* stack found in PDB 1kc8, indicates the stacking of A and G through their respective α faces,

where the five membered ring of A stacks with both the five membered and six membered rings of G in a *cis* glycosidic orientation. Similarly, G1163||U1164 α||β 5||6, 6||6 *trans* stack found in PDB 3ccu, implies the stacking of α face of G with the β face of U in a *trans* glycosidic orientation, where both five- and six-membered rings of G stack with the six-membered ring of U.

Each of the 384 theoretically-possible stacking arrangements can occur between nucleotides that occur either in tandem position in sequence space or are separated by other nucleotides within the sequence space. For example, Figure 5 shows the PDB examples of six possible stacking arrangements of a U||U stack. Although the first three arrangements illustrate the stacking of two Us that are *consecutive* in sequence space, the remaining three arrangements are taken from structures where the stacked Us are *non-consecutive* in sequence space. Examples of *consecutive* and *non-consecutive* stacking corresponding to all base stacks identified from analysis of the complete RNA dataset are provided in the Supplementary Information (Tables S5 – S124, S126 – S149 and S151 – S158). Further, a comparison of nomenclature of stacks identified by our method with the available methods is provided in Table S159.

In addition to base stacking contacts, T-shaped contacts between nucleobases were searched by locating the near-orthogonal orientation of the interacting bases. This was geometrically executed by fixing the θ_{ab} parameter in the range of 60° to 90°. Though the proportion of T-shaped contacts (~17%) is substantially smaller than stacking contacts (~83%) within the total number of contacts (Table S160), they do have a significant occurrence in absolute terms. However, due to their substantial difference from sandwich-type stacking contacts considered in the present work, any detailed analysis of T-shaped contacts ought to be a part of a separate study.

Annotation of stacking contacts in RNA 2D representations.

As previously suggested for base pairs (2), it is desirable to annotate the stacking arrangements in a RNA molecule on a standard 2D secondary structure representation. This would help to visually recognize the essential features of stacking interactions in an RNA motif. In this context, we propose that to designate base stacking in general, one could use the symbol "||". Further, we suggest filled or hollow arrow symbol to unambiguously specify the stacking face. Thus, when the head of the arrow symbol interacts with the other arrow, it represents the α interacting face of the base. Alternatively, if the tail of the arrow symbol interacts with the other arrow, it represents the β interacting face. Further, two filled arrows

denote *cis* orientation, whereas hollow arrows represent *trans* stacking orientation. For example, the symbol A⬅➡A and A⇦⇨A would represent β||β *cis* and β||β *trans* arrangement respectively. Symbols for all the variants of distinct mutual stacking arrangements can be seen in Table 1.

Stacking contacts in representative 2D RNA structures

To further illustrate the symbol convention for denoting stacking variants, we present examples of 2D representations of three well-known RNA motifs used previously to represent distinct base pairing geometries (2). These include the loop E of bacterial 5S rRNA (PDB code: 364d (22)), the sarcin/ricin motif from the large bacterial ribosomal subunit (PDB code: 430d, (23)) and the Domain IV of Signal Recognition Particle (SRP) 4.5 S RNA (PDB codes: 1dul and 1duh).

Loop E of bacterial 5S rRNA: Panel A in Figure 6 represents the secondary structure of the loop E of bacterial 5S rRNA (22). The highly conserved loop E is the binding site for the ribosomal protein L24 in the *E. coli* ribosome (24). All bases of this symmetric "internal loop" are paired. Further, all the base pairs are stacked on top of each other to form a chain like structure (Figure 6A). Although base pairing interactions within the motif are well known and have been previously annotated by Leontis and Westhof (see green symbols in Figure 6A, (2)), here we provide complete description of complex stacking interactions within the motif using our classification scheme.

As shown in Figure 6A, the A104||G105 *consecutive* stack, which involves a α||β 5||5 *cis* interaction, is designated using filled arrows (indicating the *cis* geometry), with the arrow placed above A104 has its head (representing its α face) pointing towards G105, and the other arrow placed below G105 with its tail (representing its β face) facing A104 (Figure 6A). However, A104||A73 6||5 and A99||A78 5||56, 6||56 are β||β *trans* cross-strand stacking interactions, which are indicated by hollow arrow symbols sharing their tails respectively. Figure 6A also shows how the numerical topological descriptors can be shown to indicate that different set of rings are involved in the interactions.

On similar lines, remaining base stacks are annotated in Figure 6A. However, although FR3D algorithm detects an interstrand N-H•••π contact between G75 and G112, such a contact is not identified by our algorithm, since we focus on ring stacking in this work (Table S161). Nevertheless, the classification of such contacts can easily be done along similar lines. Further, considering only base-pairing interactions, Leontis and Westhof observed that the bacterial loop E motif comprises of two isosteric submotifs oriented in opposite (palindromic)

directions (2). While this may apparently imply the presence of a C2 symmetry, this is clearly ruled out when stacking interactions within the motif are taken into consideration. Thus, these relationships can easily be depicted, in RNA motifs, with the help of the symbols we have proposed.

Sarcin/ricin motif of the large ribosomal subunit: The highly conserved sarcin/ricin motif (24) in Domain E of eukaryotic 5S rRNA structure comprises of a GAGA hairpin loop (not shown in figure) and an asymmetric internal loop (Figure 6B). Here the base G10 lies in the same plane as the base pair U11||A20 and thus represent a "bulge". Although the base pairing geometries of the five base pairs (i.e. A12:G19, G10:U11, U11:A20, A9:A21 and U7:C33) were previously described by Leontis and Westhof (see green symbols in Figure 6 (2)), analysis of the intra-motif stacking relationships between bases reveals six distinct base||base stacks. Two of these are, the A12||A20 β||β 56||5 *trans* cross-strand stacking interaction, and G19 (base pairing with A12) stacked upon G10 in α||α 56||5 *cis* geometry. Further, the stacking interaction between U11 and A9 reported earlier by Leontis and Westhof (2) is topologically annotated as α||β 6||5 according to our classification scheme. In addition, A9||C8, A20||A21 and A21||C22 stacks may be annotated, respectively, as α||β 5||6 *cis*, α||β 5||5,6||56 *cis* and α||α 6||6 *cis.*

Domain IV of SRP 4.5 S rRNA: This symmetric internal loop motif is very similar to the submotifs of the bacterial loop E motif (Figure 6A). Our analysis reveals that this motif comprises of a G165||A164 β||α 6||6,5||56 *cis* stacking along with A164||C163 β||β 5||6 *cis* stacking and a A164||A148 β||β 6||56 *trans* cross-strand stacking interaction (Figure 6C). These interactions correspond to G105||A104, A104||U103 and A104||A73 stacking interactions respectively in the E submotif, albeit with different set of rings stacked (Figure 6A). Interestingly, in SRP, C147 and G162 show cross-strand stacking interaction in β||α 5||56 *trans* geometry, which does not occur in the loop E motif (Figure 6A).

Analysis of the occurrence context of select stacking topologies in 3D RNA structures

To illustrate how the proposed classification scheme can enhance our understanding of stacking interactions in RNA 3D structures, we annotate the stacking interactions in certain specific RNA motifs. First, we annotate the stacking interactions around the G:C linchpin pair (i.e. G2:C74 pair (blue) in Figure S5) in the tRNA-like structure from the 3′-end of the turnip yellow mosaic virus (PDB id: 4p5j (25)). This pair has been previously compared with the U64:A85 base pair, observed in SAM-I riboswitch from *Thermoanaerobacter tengcongensis* (PDB:2gis (27)), and similarities in the base-stacking interactions around the

two pairs have been noted (26). However, detailed comparison of the two structures reveals geometrical differences in the stacking patterns around the G:C linchpin pair and the U:A pair (Figure S5 (26)).

We further compare the stacking interactions in three specific diloop motifs- the CUAG (C55 to G58) loop of the CRISPR Cas9-sgRNA-DNA ternary complex (chains B and C, PDB id: 4oo8 (28)), the UUGA diloop (26), which is recognized by the sequence specific RNA binding sterile alpha motif domain of yeast post *trans*criptional regulator Vts1p (PDB id: 2f8k (29)) and the CUUG diloop (26), which was previously characterized as a tetraloop (PDB code: 1rng (30)). A previous comparison had revealed that although all three loops belong to the 'diloop' category, UUGA and CUUG possess similar structural features that markedly differ from those of CUUG (26). This is further substantiated by our analysis, which reveals significantly smaller stacking overlap in CUUP loop compared to the other two loops (Figure S7). In addition, our analysis highlights subtle differences in the extent of stacking overlap within UUGA and CUUG loops. Overall, both these examples illustrate how annotation of stacking contacts in RNA 3D structures can aid in visual understanding of the subtle structural differences in apparently similar motifs.

CONCLUSIONS

In the present work, we propose a classification scheme for annotation and characterization of base stacks in RNA structures. We further develop an automated method for classification of base stacks. In addition, we have proposed symbolic representations to annotate stacking interactions in secondary structures as well as in 3D structural diagrams of RNA. The utility of our classification scheme is further demonstrated by analysing specific RNA structural and functional contexts that involve the role of base stacks. Overall, our proposed scheme for classification of RNA base stacks is expected to highlight the rich diversity in RNA base stacking, and further appreciate the role of stacking interactions in biology.

Tables and Figures

Table 1. Symbols for representations of eight stacking geometrical families.

No.	Glycosidic bond orientation	Interacting faces	Symbols
1	*cis*	α‖α	⬍
2	*trans*	α‖α	⇳
3	*cis*	α‖β	⬍
4	*trans*	α‖β	⇳
5	*cis*	β‖α	⬍
6	*trans*	β‖α	⇳
7	*cis*	β‖β	⬍
8	*trans*	β‖β	⇳

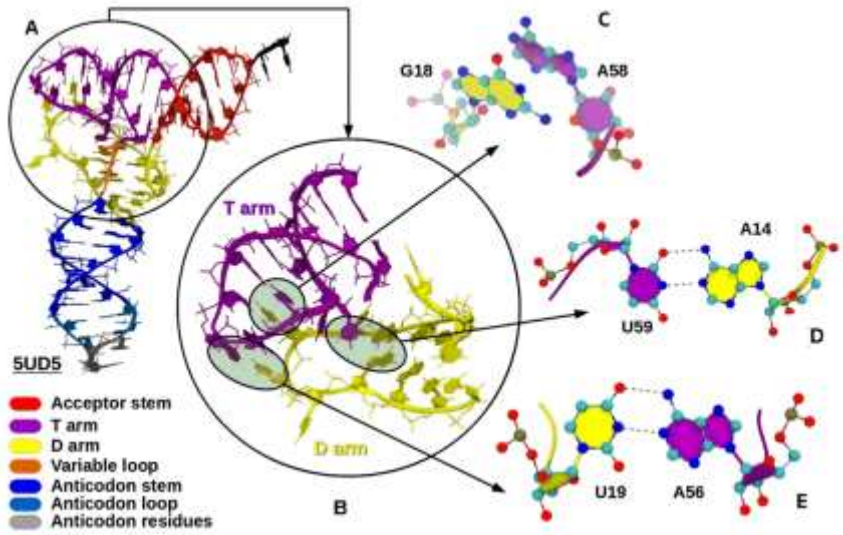

Figure 1. (A) Representation of a tRNA structure (PDB code: 5UD5) illustrating the noncovalent interactions between (B) T arm and D arm of tRNA. (C and D) Examples of Base pairing (C) and stacking interactions ((D) between nucleobases of D arm and T arm.

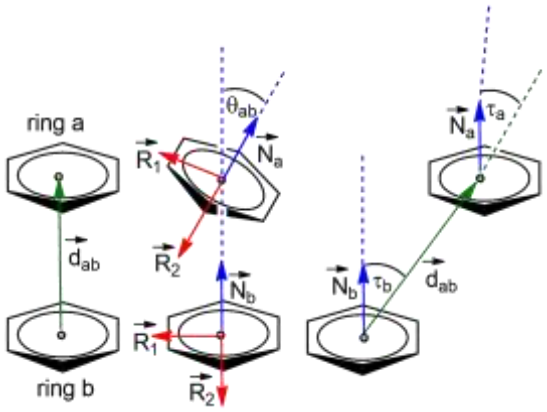

Figure 2. Representation of geometrical parameters ($\vec{d}_{ab}$, θ_{ab}, τ_a and τ_b) used to locate base||base stacking in the crystal structures of RNA.

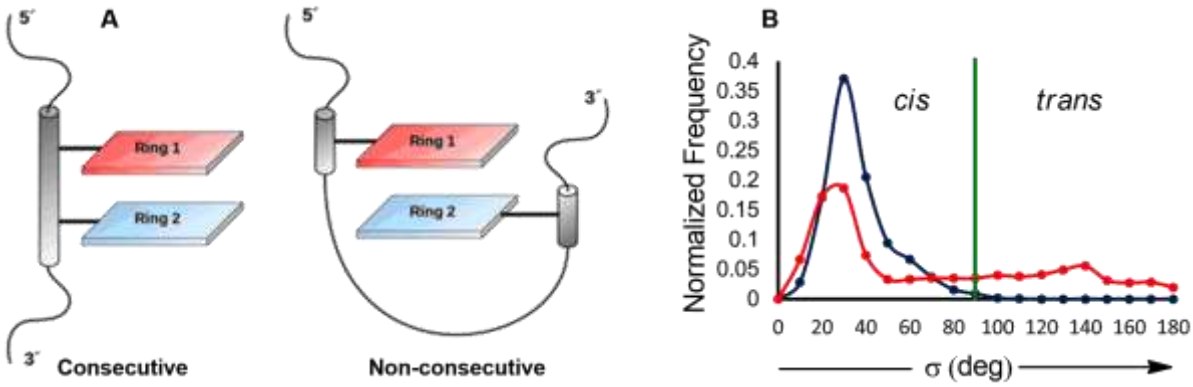

Figure 3. (A) Cartoon representation of *consecutive* and *non-consecutive* stacks. (B) Distribution of σ torsion angle in consecutive (blue) and non-consecutive (red) stacks in the analyzed structures.

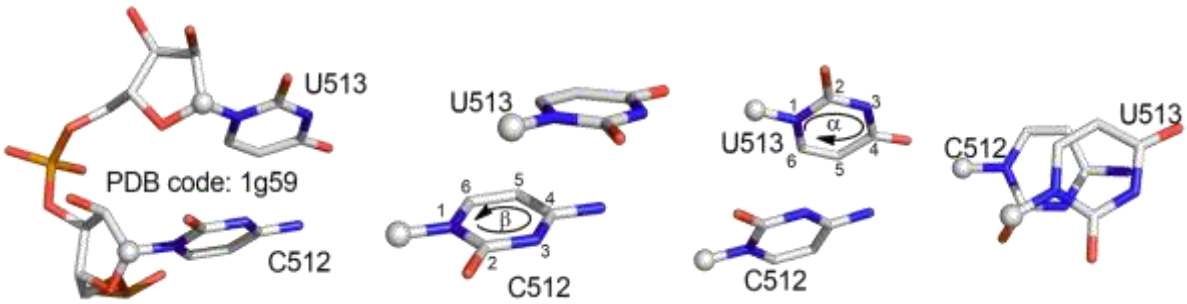

Figure 4. Illustration of the assignment of α and β face orientations with the help of a U||U α||β *cis* stack.

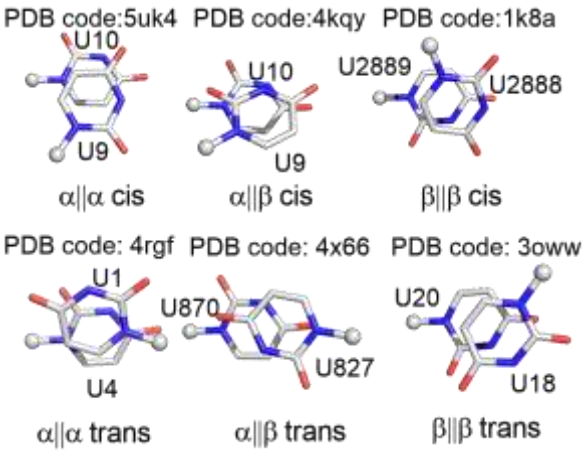

Figure 5. Six possible stacking arrangements of U||U stacks. Top three arrangements are shown using *consecutive* stacks, whereas bottom three arrangements illustrate non-consecutive stacks occurring in RNA crystal structures.

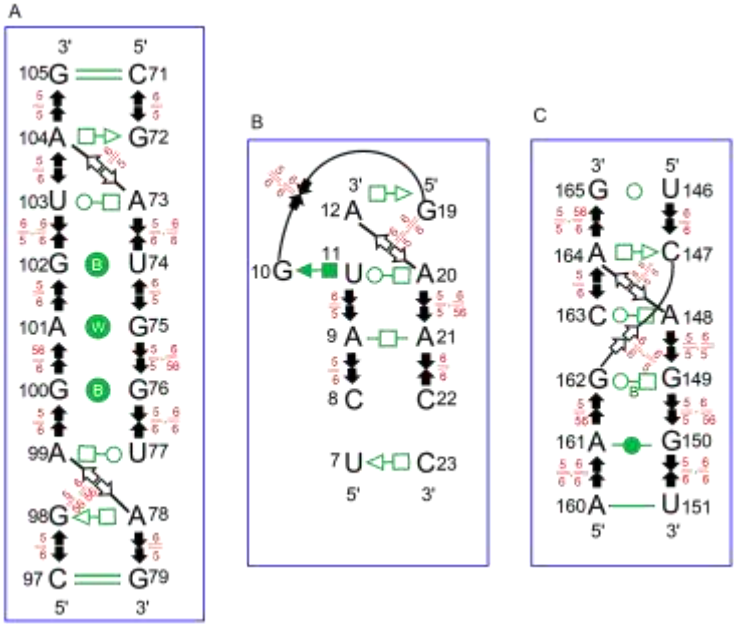

Figure 6. Annotation of stacking interactions using arrow representations in the (A) loop E motif of bacterial 5S rRNA (PDB code: 364D), (B) sarcin/ricin motif of the large ribosomal subunit (PDB code: 430D) and (C) domain IV of SRP 4.5 S rRNA (PDB code: 1DUL and 1DUH). Green symbols represent base pairing interactions previously described by Leontis and Westhof (2).

References.

1. Saenger, W. (1984) Principles of nucleic acid structure. Springer-Verlag New York 1984.
2. Leontis, N.B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA*, **7**, 499-512.
3. Cate, J.H., Yusupov, M.M., Yusupova, G.Z., Earnest, T.N. and Noller, H.F. (1999) X-ray crystal structures of 70S ribosome functional complexes. *Science*, **285**, 2095-2104.
4. Nissen, P., Hansen, J., Ban, N., Moore, P.B. and Steitz, T.A. (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920-930.
5. Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D., Bashan, A., Bartels, H., Agmon, I. and Franceschi, F. (2000) Structure of functionally activated small ribosomal subunit at 3.3 Å resolution. *Cell*, **102**, 615-623.
6. Wimberly, B.T., Brodersen, D.E., Clemons Jr, W.M., Morgan-Warren, R.J., Carter, A.P., Vonrhein, C., Hartsch, T. and Ramakrishnan, V. (2000) Structure of the 30S ribosomal subunit. *Nature*, **407**, 327.
7. Halder, S. and Bhattacharyya, D. (2013) RNA structure and dynamics: a base pairing perspective. *Prog. Biophys. Mol. Biol.*, **113**, 264-283.
8. Havrila, M., Réblová, K., Zirbel, C.L., Leontis, N.B. and Šponer, J. (2013) Isosteric and nonisosteric base pairs in RNA motifs: Molecular dynamics and bioinformatics study of the sarcin–ricin internal loop. *J. Phys. Chem. B*, **117**, 14302-14319.
9. Seelam, P.P., Sharma, P. and Mitra, A. (2017) Structural landscape of base pairs containing post-transcriptional modifications in RNA. *RNA*, **23**, 847-859.
10. Chawla, M., Oliva, R., Bujnicki, J.M. and Cavallo, L. (2015) An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Res.*, **43**, 6714-6729.
11. Bhattacharya, S., Jhunjhunwala, A., Halder, A., Bhattacharyya, D. and Mitra, A. (2019) Going beyond base-pairs: topology-based characterization of base-multiplets in RNA. *RNA*, **25**, 573-589.
12. Abu Almakarem, A.S., Petrov, A.I., Stombaugh, J., Zirbel, C.L. and Leontis, N.B. (2011) Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Res.*, **40**, 1407-1423.
13. Gabb, H., Sanghani, S., Robert, C. and Prevost, C. (1996) Finding and visualizing nucleic acid base stacking. *J. Mol. Graph.*, **14**, 6-11.
14. Gendron, P., Lemieux, S. and Major, F. (2001) Quantitative analysis of nucleic acid three-dimensional structures. *J. Mol. Biol.*, **308**, 919-936.
15. Waleń, T., Chojnowski, G., Gierski, P. and Bujnicki, J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151-e151.
16. Lu, X.-J. and Olson, W.K. (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols*, **3**, 1213.
17. Sarver, M., Zirbel, C.L., Stombaugh, J., Mokdad, A. and Leontis, N.B. (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *J. Math. Biol.*, **56**, 215-252.
18. Šponer, J., Riley, K.E. and Hobza, P. (2008) Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.*, **10**, 2595-2610.
19. Cremer, D.t. and Pople, J. (1975) General definition of ring puckering coordinates. *J. Am. Chem. Soc.*, **97**, 1354-1358.

20. Rose, I.A., Hanson, K.R., Wilkinson, K.D. and Wimmer, M.J. (1980) A suggestion for naming faces of ring compounds. *Proc. Natl. Acad. Sci. U. S. A.*, **77**, 2439-2441.
21. Sekine, S.-i., Nureki, O., Shimada, A., Vassylyev, D.G. and Yokoyama, S. (2001) Structural basis for anticodon recognition by discriminating glutamyl-tRNA synthetase. *Nature Struct. Mol. Biol.*, **8**, 203.
22. Correll, C.C., Freeborn, B., Moore, P.B. and Steitz, T.A. (1997) Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain. *Cell*, **91**, 705-712.
23. Correll, C.C., Munishkin, A., Chan, Y.-L., Ren, Z., Wool, I.G. and Steitz, T.A. (1998) Crystal structure of the ribosomal RNA domain essential for binding elongation factors. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 13436-13441.
24. Leontis, N.B. and Westhof, E. (1998) A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs. *J. Mol. Biol.*, **283**, 571-583.
25. Colussi, T.M., Costantino, D.A., Hammond, J.A., Ruehle, G.M., Nix, J.C. and Kieft, J.S. (2014) The structural basis of transfer RNA mimicry and conformational plasticity by a viral RNA. *Nature*, **511**, 366.
26. Lu, X.-J., Bussemaker, H.J. and Olson, W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142-e142.
27. Montange, R.K. and Batey, R.T. (2006) Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, **441**, 1172.
28. Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F. and Nureki, O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935-949.
29. Aviv, T., Lin, Z., Ben-Ari, G., Smibert, C.A. and Sicheri, F. (2006) Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. *Nature Struct. Mol. Biol.*, **13**, 168.
30. Jucker, F.M. and Pardi, A. (1995) Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. *Biochemistry*, **34**, 14416-14427.