6-1-2003

Tools for the Automatic Identification and Classification of RNA Base Pairs

Neocles B. Leontis
Bowling Green State University - Main Campus, leontis@bgsu.edu

Huanwang Yang

Fabrice Jossinet

Li Chen

John Westbrook

See next page for additional authors

Follow this and additional works at: https://scholarworks.bgsu.edu/chem_pub

Part of the Chemistry Commons

Repository Citation
Leontis, Neocles B.; Yang, Huanwang; Jossinet, Fabrice; Chen, Li; Westbrook, John; Berman, Helen; and Westhof, Eric, "Tools for the Automatic Identification and Classification of RNA Base Pairs" (2003). Chemistry Faculty Publications. 3.
https://scholarworks.bgsu.edu/chem_pub/3

This Article is brought to you for free and open access by the Chemistry at ScholarWorks@BGSU. It has been accepted for inclusion in Chemistry Faculty Publications by an authorized administrator of ScholarWorks@BGSU.
Tools for the automatic identification and classification of RNA base pairs

Huanwang Yang, Fabrice Jossinet, Neocles Leontis, Li Chen, John Westbrook, Helen Berman and Eric Westhof

Department of Chemistry and Chemical Biology, Rutgers University, NJ 08854-8087, USA, 1Institut de Biologie Moléculaire et Cellulaire du CNRS, Université Louis Pasteur, 15 Rue René Descartes, F-67084 Strasbourg Cedex, France and 2Chemistry Department, Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, OH 43403, USA

Received February 6, 2003; Revised and Accepted March 20, 2003

ABSTRACT

Three programs have been developed to aid in the classification and visualization of RNA structure. BPViewer provides a web interface for displaying three-dimensional (3D) coordinates of individual base pairs or base pair collections. A web server, RNAView, automatically identifies and classifies the types of base pairs that are formed in nucleic acid structures by various combinations of the three edges, Watson–Crick, Hoogsteen and the Sugar edge. RNAView produces two-dimensional (2D) diagrams of secondary and tertiary structure in either Postscript, VRML or RNAML formats. The application RNAMLview can be used to rearrange various parts of the RNAView 2D diagram to generate a standard representation (like the cloverleaf structure of tRNAs) or any layout desired by the user. A 2D diagram can be rapidly reformatted using RNAMLview since all the parts of RNA (like helices and single strands) are dynamically linked while moving the selected parts. With the base pair annotation and the 2D graphic display, RNA motifs are rapidly identified and classified. A survey has been carried out for 41 unique structures selected from the NDB database. The statistics for the occurrence of each edge and of each of the 12 bp families are given for the combinations of the four bases: A, G, U and C. The program also allows for visualization of the base pair interactions by using a symbolic convention previously proposed for base pairs. The web servers for BPViewer and RNAView are available at http://ndbserver.rutgers.edu/services/. The application RNAMLview can also be downloaded from this site. The 2D diagrams produced by RNAView are available for RNA structures in the Nucleic Acid Database (NDB) at http://ndbserver.rutgers.edu/atlas/.

INTRODUCTION

Nucleic acid structures are being published at a great pace following enhanced X-ray, NMR and chemical synthesis technologies. The complex and intricate RNA structures solved to date reveal a great diversity of base to base interactions (reviewed in 1). Various approaches exist for describing nucleic acid base pairs (2). Recently, Leontis and Westhof (LW) (3) proposed a base pair classification for nucleic acids based on the observation that the planar edge-to-edge hydrogen bonding interactions between two bases involve one of three distinct edges: the Watson–Crick edge, the Hoogsteen edge and the Sugar edge. The classification applies to those base pairs with at least two H-bonds. To define the relative orientation of the two bases, a line is drawn parallel to and between the two connecting H-bonds. The interaction of the bases is called trans if the glycosidic bonds of the interacting nucleotides lie on opposite sides of the line. It is called cis if the glycosidic bonds of the nucleotides are on the same side of the line. This classification gives rise to 12 basic geometric families of pairing arrangements between nucleic acid bases. For each geometric type, the relative orientations of the strands can be easily deduced. The base pair interaction diagram is given in Figure 1. The 12 different families of base pairs together with the local strand orientations are given in Table 1. A graphical convention for displaying non-Watson–Crick interactions on a secondary structure diagram has also been proposed (3). According to the proposed convention, a circle represents the Watson–Crick edge of a base, a square represents its Hoogsteen edge and a triangle represents the Sugar edge. Cis interactions are designated by filled symbols and trans base pairs by open symbols. The symbols corresponding to each base pair family are also given in Table 1.
The analysis of the detailed three-dimensional (3D) interactions for each type of base pair interaction has been greatly facilitated by the program, BPViewer. The geometrical classification along with the symbolic diagrams provides a set of simple mnemonics for each type of base pair interaction. The full diagram convention simplifies and clarifies the annotation, description and comparison of secondary structure, RNA motifs and tertiary interactions present in a folded RNA. We have developed a program, RNAview, to quickly perform LW base pair classifications and RNA motif searches. The program generates the secondary structure with base pair symbols and tertiary interactions. RNA View also exports the VRML (Virtual Reality Markup Language) file so that the user can dynamically visualize the 3D structure. The detailed secondary structure annotations are also exported in the RNAML format (4). A companion program, RNAMLview, accepts RNAML input and allows the dynamic editing of the two-dimensional (2D) projections produced by RNAview. Web servers for BPViewer and RNA View are provided by the Nucleic Acid Database (NDB) (5). The present set of programs is complementary to others (9,10) and the results of either can be used or compared using the RNAML files.

**METHODS AND ALGORITHMS**

**Finding the base pairs**

*Choice of coordinate frame.* A standard reference frame is used for base pair determination (11). Models of the seven bases (A, U, G, C, T, I and P) were generated in a survey of high resolution crystal structures of nucleic acid analogs stored in the recent version of the Cambridge Structure Database (12). The coordinate frames are chosen so that the complementary bases form an ideal, planar Watson–Crick base pair in the undistorted reference state. The hydrogen bond donor–acceptor distance (N, O)...(N, O), the C1′...C1′ distance, the angles N9-C10...C10 and N1-C10...C10 between the base paired purine and pyrimidine should be consistent with values observed in the relevant small molecule in the Cambridge Structure Database. A right-handed coordinate frame is generated and attached to each base. The x-axis points in the direction of the major groove, along the perpendicular bisector of the C1′...C1′ vector. The origins are between the Watson–Crick edges of the standard base pair (Fig. 2) and are at the intersection point between the x-axis and a line connecting the pyrimidine Y(C6) and purine R(C8) atoms. The y-axis points in the direction of the sugar–phosphate backbone (of the sequence strand), parallel to the C1′...C1′ vector. The z-axis is determined by the right-handed rule, i.e. \( z = x \times y \). For right-handed A- and B-DNA the z-axis accordingly points in the 5′- to 3′-direction of the sequence strand. For a standard base pair, the origins of the two coordinate systems, one for the pyrimidine and one for the purine, overlap. Here, the two coordinate axes are separated slightly for easy visualization. The x-directions for the two local coordinates are the same while the y and z directions are opposite (Fig. 2). The coordinates satisfying the above criteria can be found in Table 1 of (11).

**Determination of base pairs.** For determining base pairs in a real structure, a least-square fit is carried out for each base. Each standard base, as described above, is fitted to the

---

**Table 1.** The local strand orientations of the 12 families of base pairs

<table>
<thead>
<tr>
<th>No.</th>
<th>Glycosidic Bond Orientation</th>
<th>Interacting Edges</th>
<th>Symbol</th>
<th>Default Local Strand Orientation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cis</td>
<td>Watson-Crick / Watson-Crick</td>
<td>–</td>
<td>Anti-parallel</td>
</tr>
<tr>
<td>2</td>
<td>Trans</td>
<td>Watson-Crick / Watson-Crick</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>3</td>
<td>Cis</td>
<td>Watson-Crick / Hoogsteen</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>4</td>
<td>Trans</td>
<td>Watson-Crick / Hoogsteen</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>5</td>
<td>Cis</td>
<td>Watson-Crick / Sugar Edge</td>
<td>–</td>
<td>Anti-parallel</td>
</tr>
<tr>
<td>6</td>
<td>Trans</td>
<td>Watson-Crick / Sugar Edge</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>7</td>
<td>Cis</td>
<td>Hoogsteen / Hoogsteen</td>
<td>–</td>
<td>Anti-parallel</td>
</tr>
<tr>
<td>8</td>
<td>Trans</td>
<td>Hoogsteen / Hoogsteen</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>9</td>
<td>Cis</td>
<td>Hoogsteen / Sugar Edge</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>10</td>
<td>Trans</td>
<td>Hoogsteen / Sugar Edge</td>
<td>–</td>
<td>Parallel</td>
</tr>
<tr>
<td>11</td>
<td>Cis</td>
<td>Sugar Edge / Sugar Edge</td>
<td>–</td>
<td>Anti-parallel</td>
</tr>
<tr>
<td>12</td>
<td>Trans</td>
<td>Sugar Edge / Sugar Edge</td>
<td>–</td>
<td>Parallel</td>
</tr>
</tbody>
</table>

The 12 families of edge to edge base pairs formed by nucleic acid bases as defined by the relative orientations of the glycosidic bonds of the interaction bases (column 2) and the edges used in the interaction (column 3). The symbolic representation is given in column 4. The local strand orientation is given in column 5. Table reproduced from (3) with permission from Oxford University Press.
corresponding base in the real structure. However, some residues in the structure may be chemically modified. In such a situation, the program will use the standard base which best matches the modified base by least square fitting. For example, the modified residue +C is fitted to the corresponding standard coordinates for C. When fitting to each real base, the standard coordinate frame is rigidly rotated and translated to bring the standard base into coincidence with the real base. In carrying out the least square fittings, several useful parameters, including shear, stretch, stagger, buckle, propeller and opening, can be easily calculated. There are two important parameters for base pair determination. One is the angle between the planes of the two bases and the other is the vertical distance between these planes. The angle of the two bases in the real structure can be calculated by the equation: \((180°/\pi) \times \cos(\vec{z}_i \cdot \vec{z}_j)\), where \(\vec{z}_i \cdot \vec{z}_j\) is the dot product of unit vectors \(\vec{z}_i\) and \(\vec{z}_j\), which are the normal vectors to the planes of base \(i\) and base \(j\), respectively. The vertical distance, \(d_v\), of two bases is calculated as the norm of the vector dot product, i.e., \(|\vec{n} \cdot \vec{d}|\), where \(\vec{n}\) is the unit vector averaged over vectors \(\vec{z}_i\) and \(\vec{z}_j\) and \(\vec{d}\) is a vector connecting the two origins.

To determine the glycosidic bond configuration (cis or trans) for each basepair, we generated \(\vec{V}_{12}\), the vector connecting the geometric centers of the ring systems of the interacting bases. \(\vec{V}_{12}\) is directed from the geometric center of base 1 to the center of base 2. \(\vec{V}_{1C}\) and \(\vec{V}_{2C}\) are vectors coinciding with the glycosidic bonds of bases 1 and 2. They are directed from Sugar C1’ to the corresponding glycosidic nitrogen (N1 for pyrimidines or N9 for purines) of the base. When the dot product \((\vec{V}_{12} \times \vec{V}_{1C} \cdot (\vec{V}_{12} \times \vec{V}_{2C}) < 0\), the base pair is in the trans orientation. Otherwise, it is in the cis configuration.

**The Watson–Crick base pairs.** The criteria for determining the canonical Watson–Crick base pairs are the following:

- The base pair must be either AU, AT, GC or IC.
- The angle between the two local \(x\)-axes must lie in the range \(0°–17°\); the angle between the two local \(y\)-axes must in the range \(157°–180°\); and the angle between the two local \(z\)-axes must be in the range \(0°–30°\). This defines the standard Watson–Crick geometry in which the glycosidic bond configuration is automatically cis.
- The distance separating the two base pair origins must be \(<2.5\,\text{Å}\).
- The vertical distance between the two base planes must be \(<1.5\,\text{Å}\).

**The non-Watson–Crick base pairs.** If the base pair does not belong to the canonical Watson–Crick base pair as defined above, it is examined to see whether it belongs to one of the other 12 non-Watson–Crick base pair families. The criteria for determining the type of base pair family are the following:

- The angle \(\varphi\) between two base planes (or equivalently, between the two local \(z\)-axes) must be \(<65°\).
- The vertical distance \(d_v\) between the two bases planes must be \(<2.5\,\text{Å}\).
- At least two hydrogen bonds must exist, one of which should be a H-bond that occurs between donor and acceptor atoms, both of which belong to a base, and with the distance between the two atoms (N or O) \(<3.4\,\text{Å}\). The second H-bond is determined with extended distances. The distance between donor and acceptor for base (N, O) base (N, O) \(<3.75\,\text{Å}\), for base (N, O) base (C–H) \(<3.9\,\text{Å}\), for base (N, O) ribose (O2’) \(<3.75\,\text{Å}\).

If the pair has only two H-bonds and they are bifurcated with two acceptor atoms and one donor atom, the pair will be further examined. Only those with better geometry (angle \(\varphi\) between two base planes \(<50°\), and vertical distance \(d_v\), between the two base planes \(<2.1\,\text{Å}\) will be assigned to one of the 12 families. Another special case for rejecting a base pair is when \(\varphi < 10°\) and \(d_v > 2.2\,\text{Å}\).

All the pairs which do not belong to one of the 12 families will be classified as a tertiary interaction if there exists at least one H-bond that meets the following criteria: the distance between donor and acceptor for base (N, O) base (N, O) \(<3.4\,\text{Å}\); for base (N, O) base (O2’) \(<3.4\,\text{Å}\); for base...
(N, O) ... base (O1P/O2P) < 3.2 Å; and for all others < 3.1 Å. The tertiary interactions are represented in our graphic display by red dashed lines.

The geometrical edge for each base of the 12 families is determined by counting the contact distances between the two bases. In order to correctly identify the base edges, all the atoms on the edge have to be used and the distance between heavy atoms (the contact distance) is set to 4.0 Å. However, for the bifurcated pairs with only two H-bonds, the distance is set to 4.3 Å, since in this type of pair the bases are normally further from each other.

A stacked arrangement of bases does not belong to the 12 families. However, since stacking is believed to play a major role in nucleic acid structure folding, we also listed the stacked bases. The criteria for stacking are the following: the distance between the two centers of the ring is < 5.7 Å, the vertical distance between the two base planes is < 2.7 Å, and the base angles are < 40°.

Graphic representations of secondary structure with tertiary interactions

In establishing the above base pairing criteria it was necessary to visualize a large number of these interactions. To simplify this process, a web visualization tool, BPViewer, was developed. This tool provides the means to navigate through the list of base pairs, selecting and viewing the 3D coordinates for individual or sets of base pairs of interest.

In order to convey the 3D contacts present in a folded RNA, it is usually necessary to represent in a planar drawing the secondary and tertiary base pairs and interactions. The secondary base pairs are the cis Watson–Crick pairs together with the wobble pairs and those are normally represented in standard 2D diagrams. The tertiary interactions can be decomposed into base–base or base–backbone interactions. Most generally, the base–base interactions belong to one of the 12 families of the pairs discussed above. We wanted to produce a program able to deduce such a diagram starting from a set of 3D coordinates.

We have developed a program called RNAview that provides a detailed annotation for the secondary and tertiary structures of a nucleic acid with the diagram convention for the 12 families of pairs. Figure 3A gives the flow chart of the program. A minimum of two consecutive Watson–Crick base pairs is needed to define a helical region. If several adjacent helices with their axes are roughly in the same direction, a longer pseudo-helix axis in this direction will be determined. All the helix axes are projected onto the 2D plane which is a least square plane formed by the atoms from all the helices. The 2D picture is automatically rotated with the longest pseudo-helix axis along the vertical direction (y coordinate for a postscript file). The 2D coordinates (x, y) of each base is generated perpendicular to each axis. All the loops are represented by circles. The standard Watson–Crick base pairs as defined above are presented by two short parallel lines for G–C pairs and a single short line for A–U pairs. The G–U wobble pairs are represented by small circles. The other 12 non-Watson–Crick base pair interactions are fully implemented with diagram conventions described in Table 1. The tertiary interactions as defined above are represented by single red dashed lines. All the standard nucleic acids are displayed by the single letter codes (i.e. A, U, G and C). If the base is modified, it will be displayed as a lower case letter code (i.e. a, u, g and c). For each nucleic acid, if the base is in the syn conformation with respect to the Sugar (note the difference with cis as defined above), the nucleic acid name is displayed as a red letter. The graphic secondary structure is exported as a postscript file. All the calculated structure annotations (such as the helices, single strands, bulges, loops, 2D coordinate of each base, the LW type of base pair, base modification carried out from the PDB file, etc.) are stored in the RNAML file, which uses the XML (eXtensible Markup Language) format as data exchange. Another output of the program is the VRML format. Detailed base pair interactions are displayed the same as the above conventions, but the squares are substituted by cubes, triangles by cones, circles by spheres and lines by sticks. The VRML file gives 3D dynamic structures that can be displayed from a web navigator with the VRML plug in.

The secondary structure with tertiary interactions from most of the NDB can be obtained without further modification. Some examples are given in Figures 4–6. For large complicated structures, some elements of the 2D picture may overlap. Therefore, it may be necessary to modify the automatically generated projections.

Another graphics application, RNAMLview, is able to manipulate the 2D view of RNA secondary structure stored in a RNAML file. RNAMLview is a versatile graphical editor that permits the user to adjust the layout and presentation of intramolecular interactions. The interactions are divided into secondary and tertiary interactions. Base pairing interactions are displayed by the same diagrams as those of the RNAview program. For each base in the secondary representation, the user is presented options for displaying the tertiary interactions.

An RNAML file can store different secondary representations for a single molecule. Each representation is defined in the RNAML file as a separate model. RNAMLview will generate distinct graphical representations for each model and
each representation will be displayed in a separate panel of the application and can be manipulated independently. The RNAMLview program uses the Java language (e.g., http://www.java.sun.com) to provide platform and operating system independence. The open source JDOM library (e.g., http://www.jdom.org) is used to parse the RNAML file. The program constructs two graphical views of an RNAML file: (i) a tree representation with all the elements stored; and (ii) a 2D picture for each secondary representation stored. The program creates an object model of the secondary representation (Fig. 7). This model includes all components of RNA secondary and tertiary structure such as the representations of helix, single strand, terminal loop and terminal single strand. All these components are modeled in Java as `Structure` objects in an inheritance relationship described in Figure 3.

The programs RNAview and RNAMLview are fully integrated using the data exchange protocol of RNAML. Figure 3B shows a typical scenario in which either a PDB, mmCIF or RNAML file is read by RNAview. RNAview calculates all intra-molecular interactions and stores this information in RNAML. RNAMLview reads the RNAML file produced by RNAview and, for the convenience of the user, the RNAview is wrapped by RNAMLview so that a user can directly import PDB, mmCIF or RNAML into the graphical interface. Using the RNAMLview graphic interface, a user can save the new representation in RNAML, SVG or PS formats.

RNAMLview can be used in a variety of ways. Figure 8 shows the procedure for generating a conventional secondary diagram. An NDB file is selected (in this example we use TR0001 (13)) and the program calls the RNAview program to generate a projection. Using RNAMLview, this projection can be manipulated to produce the familiar Holley diagram. The program can also be used to edit a complex secondary structure projection. A structure of P4–P6 domain (14) (Fig. 9A) is given for demonstration. Figure 9B was directly obtained using RNAview. Although technically correct, many hydrogen bonds are overlapped. RNAMLview was used to modify the projection diagram to produce Figure 9C. In this figure the overlap of hydrogen bonds is minimized.

**RESULTS**

**Graphic representation of RNA**

We have provided two graphic programs (RNAview and RNAMLview) that allow rapid visualization of RNA secondary structure. We calculated all the selected 41 RNA structures in ~1 min by using a PC (Dell 340, Pentium 4, with CPU 1.8G) with Linux operating system. This can greatly expedite looking for RNA patterns [or motifs (15)] from the fully annotated RNA structures. RNAMLview can be used to manipulate the 2D structure so that each element does not overlap for some complicated structures. The advantage of this program is that all the base pair interaction patterns are dynamically linked, while moving individual elements.

Some examples of the results of the programs are given in Figures 4–6 and 9. Figure 4A shows the 3D structure of the sarcin/ricin domain from *E. coli* 23S rRNA [NDBID UR0007 (16)]. Each block represents the base of nucleic acid: A, red; C, yellow; G, green; U, cyan. The secondary structure (B) was obtained directly from the RNAview program without further modification. The red single dashed lines mean that the base pairs only involve one H-bond.

**Figure 4.** The crystal structure of the sarcin/ricin domain from *E. coli* 23S rRNA [NDBID UR0007 (16)]. (A) The 3D structure. Each block represents the base of nucleic acid: A, red; C, yellow; G, green; U, cyan. The secondary structure (B) was obtained directly from the RNAview program without further modification. The red single dashed lines mean that the base pairs only involve one H-bond.

**Figure 5.** Crystal structure of hammerhead ribozyme [NDBID UHX026 (17)]. One biological unit is used for demonstration. (A) The 3D structure. Color code is the same as in Figure 4. The secondary structure (B) was obtained directly from the RNAview program without further modification. The red single dashed lines mean that the base pair only involve one H-bond.

Figure 5A shows the 3D crystal structure of hammerhead ribozyme (17). The structure consists of two chains with 47 nucleic acids forming 10 bp. The two helices are roughly in the same direction. Therefore, the RNAview program gives one longer pseudo-helix on the y axis.

Figure 5B was directly obtained from RNAview. The structure consists of a single chain with 27 nucleic acids forming 10 bp. The two helices are roughly in the same direction. Therefore, the RNAview program gives one longer pseudo-helix on the y axis.

**Figure 5.** Crystal structure of hammerhead ribozyme [NDBID UHX026 (17)]. One biological unit is used for demonstration. (A) The 3D structure. Color code is the same as in Figure 4. The secondary structure (B) was obtained directly from the RNAview program without further modification. The red single dashed lines mean that the base pair only involve one H-bond.
Figure 6. Crystal structure of tRNA [NDBID TR0001 (13)]. (A) The 3D structure. Color code is the same as in Figure 4. The secondary structure (B) was obtained directly from the RNAview program without further modification. The red single dashed lines mean that the base pair involve only one H-bond. The modified bases are given the corresponding lower case letter. The letter P means pseudouracil. The four red letters (a, A, G, g) indicate that the nucleic acids have syn conformations. Modified bases are given as lower case.

Figure 7. The correspondence of the RNAML tree structure with the RNA object model generated and used by RNAMLview.

Figure 6 shows the transfer RNA crystal structure (13). The structure is a single strand and forms four double helices [the acceptor, anticodon, D and T arms as annotated in (18)]. There are 30 bp from the total 76 bases. Figure 6A shows the 3D structure. Figure 6B shows the secondary structure projection that was directly obtained from RNAview. The four helices form two pseudo-helices which are perpendicular to each other. The view shows an L-shaped fold on the least square plane. Bases at the junctions (Terminal T and D loops) involve many tertiary and trans interactions. The 2D structure
is comparable to the edited picture by Ferré-D’Amare and Doudna (18); it took \(~0.1\) s (Del 340, Pentium 4, CPU 1.8G, with Linux system) to get the fully annotated structure by the RNAview program.

Figure 9 shows the crystal structure of a group I ribozyme P4–P6 domain (14). The P4–P6 domain can assemble with the remainder of the introns in \(trans\) to reconstitute catalytic activity. The structure has one chain for each monomer, consisting of nine anti-parallel helices. There are 67 bp from the total 157 bases. Figure 9A shows the 3D structure. Figure 9B shows the secondary structure that was directly obtained from RNAview. The nine anti-parallel helices make four long pseudo-helices. The two pseudo-helix axes [also annotated as P5a and P5b by Ferré-D’Amare and Doudna (18)] on the left side of Figure 9B make an angle of \(~155^\circ\). The two pseudo-helix axes on the right make an angle of \(~160^\circ\). Since this is a complicated structure, some elements of the secondary structure from the RNAview are overlapped. Therefore, the graphic editor (RNAMLview) is required to rearrange the overlapped elements. Figure 9C shows the secondary structure that was modified by RNAMLview program. It is seen that the secondary structure gives detailed annotation of the base pair interactions. The CPU time to generate this picture by RNAview program only takes \(0.3\) s (Del 340, Pentium 4, CPU 1.8G, with Linux system).

**Base pair survey and statistics**

Various examples of the 12 families of base pairs have been given (1). Here we report some statistics for these families of base pairs identified with the criteria described above. We selected 41 unique, well-refined X-ray crystal structures with resolutions of at least 3.0 A˚ from the current NDB database. If the structures were from the same source and similar, only one well-refined structure or the one with the highest resolution was selected. Nucleotides, dinucleotides and standard double-stranded helices were excluded. The selection of structures is representative of the different types of available crystal structures in which RNA is found alone or complexed to other molecules.

Calculations have been performed for four groups of structures (Table 2). The first group contains the structures of mostly single-stranded RNAs. The second group includes the protein–RNA complexes. The third group contains only the 30S and 50S ribosomal subunits. The fourth group contains protein–tRNA complexes.

The edge occurrences for pyrimidines (Y) and purines (R) for the ensemble of structures (group I) are shown in Figure 10. It is seen that the Watson–Crick edge dominates all other edges, since it includes both the standard and non-standard pairs with Watson–Crick geometry. While, for purine bases,
interactions at the Hoogsteen and Sugar edges represent each \( \sim 15\% \) of base pairs, for pyrimidine bases the same edges represent only \( \sim 3\% \). These numbers are the highest in the highly structured 50S and 30S rRNAs. Clearly, in purines, the N3 and N7 nitrogen atoms can make extra H bonds compared to pyrimidines. Interestingly also, the adenine bases present a clear preference for the Hoogsteen edge over the Sugar edge, while the opposite is true for the guanine bases. This can be understood from the edges of the R bases. For adenine, the amino group which is active as a donor in H bonds is on the Hoogsteen side, while for guanosine this group is on the Sugar side. For the pyrimidine bases, the Sugar edge is invariant.

The occurrences of each type of base pair in each of the 12 families are given in Table 3 according to the family and the base pair type. As expected, the \( \text{cis} \) Watson–Crick/Watson–Crick pairs dominate the statistics and the Hoogsteen–Hoogsteen pairs in \( \text{cis} \) are not represented. This indicates that when the two glycosidic bonds are on the same side of the line connecting to the center of the two bases, the two bases are very favorable to form the Watson–Crick geometry but not at all to form Hoogsteen geometry. Among the non-canonical pairs, the \( \text{cis} \) families represent \( <10\% \) and the \( \text{trans} \) families \( \sim 20\% \). The \( \text{cis} \) Sugar/Sugar family is dominated by the isosteric A·G and G·A pairs [some pairs present in the tables of (1) would not appear in the numbers calculated here because of the condition imposed about the presence of at least one H-bond between heavy atoms]. In the \( \text{cis} \) Watson–Crick/Hoogsteen family, one finds mainly U·A and G·G pairs, while in the \( \text{cis} \) Watson–Crick/Sugar family, the isosteric A·A and A·C pairs are the most frequently observed (interestingly, the C·A pairs are much less frequent than the A·C pairs). The \( \text{cis} \) Hoogsteen/Sugar family, the platform family (19), contains essentially the A·A and U·G pairs with about twice as many U·G as A·A pairs.
In the trans families, sharp differences appear. For example the most frequent trans Watson–Crick/Watson–Crick pairs are the A/C and G/C and then A/C U pairs. The trans Watson–Crick/Hoogsteen family is dominated by A/C U and U/C A pairs. The typical hydrogen bond pattern is N3-H...N7 and N6-H...O2, where N6 and N7 are for adenine.

In the family with a highest population, the trans Hoogsteen/Sugar, the most significant pair is again the A-G pair with the A sharing its Hoogsteen edge with the Sugar edge of the G exclusively.

The LW classification is based on geometry as defined by the types of edges with H-bonding interactions. In order to understand an RNA fold and recognize its variations, it is more important to visualize and remember which edges interact rather than the types of atoms in close contacts (which can, in principle, be deduced from the knowledge of the interacting edges). The proposed classification is clearly a simplification but it still allows, at the same time, constraints on the topology of the interacting strands and variability in the precise types and numbers of H-bonds being formed. Variability in number and types of H-bonds may have several origins, like molecular dynamics, local environments, availability of water and ionic contacts, or experimental, refinement or human errors in the crystallographic work. Here, we do not tabulate single and bifurcated H-bonds.

### Table 2. Representative RNA structures from the NDB grouped according to the structure type

<table>
<thead>
<tr>
<th>NDB ID</th>
<th>Structure description</th>
<th>Res. (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td>Sarcin/Ricin RNA domain</td>
<td>1.1</td>
<td>(16)</td>
</tr>
<tr>
<td>UR0007</td>
<td>Vitamin B12 binding RNA, cobalamin (vitamin B12)</td>
<td>3.0</td>
<td>(20)</td>
</tr>
<tr>
<td>UR0008</td>
<td>Malachite green RNA aptamer, tetramethyl-rosamine</td>
<td>2.8</td>
<td>(21)</td>
</tr>
<tr>
<td>UR0011</td>
<td>HIV-1(Lai) genomic RNA DIS.</td>
<td>2.6</td>
<td>(22)</td>
</tr>
<tr>
<td>UR0015</td>
<td>J1labc</td>
<td>2.9</td>
<td>(23)</td>
</tr>
<tr>
<td>UR0018</td>
<td>A5g group II self-splicing intron</td>
<td>3.0</td>
<td>(24)</td>
</tr>
<tr>
<td>UR0020</td>
<td>RNA pseudoknot</td>
<td>1.3</td>
<td>(25)</td>
</tr>
<tr>
<td>URT068</td>
<td>Tetraloop, hairpin</td>
<td>3.0</td>
<td>(26)</td>
</tr>
<tr>
<td>TR0001</td>
<td>transfer RNA (Phc)</td>
<td>1.9</td>
<td>(13)</td>
</tr>
<tr>
<td>UR0012</td>
<td>P4–P6 mutant RNA group I domain</td>
<td>2.3</td>
<td>(14)</td>
</tr>
<tr>
<td>DR0004</td>
<td>RNA aptamer/cobalamin</td>
<td>2.3</td>
<td>(27)</td>
</tr>
<tr>
<td>DR0005</td>
<td>Biotin-binding RNA pseudoknot, biotin</td>
<td>1.3</td>
<td>(28)</td>
</tr>
<tr>
<td>UHX026</td>
<td>RNA/DNA hammerhead ribozyme</td>
<td>2.6</td>
<td>(17)</td>
</tr>
<tr>
<td>UR0009</td>
<td>4.5S RNA domain IV</td>
<td>2.7</td>
<td>(29)</td>
</tr>
</tbody>
</table>

Protein RNA complex

<table>
<thead>
<tr>
<th>NDB ID</th>
<th>Structure description</th>
<th>Res. (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR0037</td>
<td>Signal recognition protein/RNA complex</td>
<td>1.5</td>
<td>(30)</td>
</tr>
<tr>
<td>PR0054</td>
<td>Signal recognition particle 19kDa protein/RNA complex</td>
<td>1.8</td>
<td>(31)</td>
</tr>
<tr>
<td>PRV002</td>
<td>U1a mutant/RNA complex + glycerol</td>
<td>1.9</td>
<td>(32)</td>
</tr>
<tr>
<td>PRV020</td>
<td>Satellite tobacco mosaic virus/RNA complex</td>
<td>1.8</td>
<td>(33)</td>
</tr>
<tr>
<td>PR0018</td>
<td>Ribosomal protein L25/S rRNA complex</td>
<td>1.8</td>
<td>(34)</td>
</tr>
<tr>
<td>PR0055</td>
<td>15S RNA binding protein/RNA complex</td>
<td>2.9</td>
<td>(35)</td>
</tr>
<tr>
<td>PR0074</td>
<td>U1 ribonucleoprotein/RNA Hairpin ribozyme complex</td>
<td>2.4</td>
<td>(36)</td>
</tr>
<tr>
<td>PR0052</td>
<td>Restrictocin/29-mer SRD RNA complex</td>
<td>2.0</td>
<td>(37)</td>
</tr>
<tr>
<td>PTO016</td>
<td>Spliceosomal U2B/U2A protein complex bound to a fragment of U2 small nuclear RNA</td>
<td>2.4</td>
<td>(38)</td>
</tr>
</tbody>
</table>

Ribosomal subunits

<table>
<thead>
<tr>
<th>NDB ID</th>
<th>Structure description</th>
<th>Res. (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR0033</td>
<td>Refined large ribosomal subunit (50 S)</td>
<td>2.4</td>
<td>(45)</td>
</tr>
<tr>
<td>RR0056</td>
<td>30 S ribosomal subunit</td>
<td>3.0</td>
<td>(46)</td>
</tr>
</tbody>
</table>

Protein tRNA complexes

<table>
<thead>
<tr>
<th>NDB ID</th>
<th>Structure description</th>
<th>Res. (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR0006</td>
<td>Theonyl-tRNA synthetase/tRNA(Thr) complex</td>
<td>2.9</td>
<td>(47)</td>
</tr>
<tr>
<td>PR0014</td>
<td>Isoeucyl-tRNA synthetase</td>
<td>2.2</td>
<td>(48)</td>
</tr>
<tr>
<td>PR0019</td>
<td>Aspartyl-tRNA synthetase complex</td>
<td>2.4</td>
<td>(49)</td>
</tr>
<tr>
<td>PR0029</td>
<td>Valyl-tRNA synthetase complex</td>
<td>2.9</td>
<td>(50)</td>
</tr>
<tr>
<td>PR0030</td>
<td>Arginyl-tRNA synthetase complex</td>
<td>2.2</td>
<td>(51)</td>
</tr>
<tr>
<td>PR0057</td>
<td>Prolyl-tRNA synthetase complex</td>
<td>2.9</td>
<td>(52)</td>
</tr>
<tr>
<td>PR0059</td>
<td>Methionyl-tRNA synthetase complex</td>
<td>2.8</td>
<td>(53)</td>
</tr>
<tr>
<td>PTE003</td>
<td>Glutamyl-tRNA synthetase/tRNA complex</td>
<td>2.3</td>
<td>(54)</td>
</tr>
<tr>
<td>PR0004</td>
<td>Elongation factor Tu/Cysteinyl tRNA complex</td>
<td>2.6</td>
<td>(55)</td>
</tr>
<tr>
<td>PR0060</td>
<td>tRNA pseudouridine synthase B (e.c. 4.2.1.70)/tRNA complex</td>
<td>1.9</td>
<td>(56)</td>
</tr>
</tbody>
</table>
The first column gives the LW type of base pairs. The three letters in the first column are as follows: W, Watson–Crick edge; H, Hoogsteen edge; S, Sugar edge; e, the glycosidic bond orientation is cis; and t, the orientation is trans. SUM gives the summation of the columns and rows. Examples of the combination of the letters HS means that the base on the right is Hoogsteen edge and on the left is Sugar edge and the two glycosidic bond orientation are cis. The integer numbers indicate the occurrence of LW pair type and the base combinations. The last row and last column give the percentages (%) with respect to the total number. For simplicity, equivalent base pairs were not merged in this table (e.g. WWc A U and WWc U A or WS t A G and SW t G A).

### CONCLUSIONS

In order to identify and classify automatically RNA base pairs, three programs have been developed. BPView allows dynamic navigation through the base pairing of the 3D RNA structure. A web interface to this tool is available from http://ndbserver.rutgers.edu/services/BPviewer/. RNAView and RNAMLView produce projections of the structures following the nomenclature previously proposed (3) and produce visual representations as well as outputs in the accepted RNAML syntax for exchange of data (4). The base pairing interactions are annotated on secondary structure diagrams which can be easily manipulated with a symbolic representation which allows for a rapid identification of RNA motifs (15). RNAView is available as a web service at http://ndbserver.rutgers.edu/services/rna-viewer/. The application RNAMLView may be downloaded from http://ndbserver.rutgers.edu/services/ and a new version, as a web service, will be available soon. In addition RNAView 2D projections have been calculated for all structures in the NDB and these are available on the NDB Atlas pages (http://ndbserver.rutgers.edu/atlases/).

### ACKNOWLEDGEMENTS

We would like to thank Dr Xiang-Jun Lu and Dr Zukang Feng for helpful discussions. This work was funded by grants from the NSF and DOE supporting the NDB project and by the Institut Universitaire de France. N.L. is supported by NIH 2R15 GM55898.

### REFERENCES

Centre: computer-based search, retrieval, analysis and display of information.


alanine-tRNA at 1.93 Å resolution: a classic structure revisited. RNA 6, 1091–1105.

basis of the enhanced stability of a mutant ribosome domain and a detailed
view of RNA solvent interactions. Structure, 9, 221–231.

Escherichia coli 23 S rRNA sarcin/ricin domain: the structure at 1.11 Å
resolution. EMBO J., 18, 1091–1097.


Crystal structures of coaxially-stacked kissing complexes of the HIV-1


Metal ions and


30. Batey,R.T., Sagar,M.B. and Doudna,J.A. (2001) Structural and energetic
analysis of RNA recognition by a universally conserved protein from the
motor protein complex bound to a fragment of U1 spliceosomal RNA.

structure at 1.92 Å resolution of the RNA-binding domain of the U1A
432–438.

structure of satellite tobacco mosaic virus at 1.8 Å resolution. J. Mol.

Transition state stabilization by a catalytic RNA. Science, 298,
1421–1424.

spliceosomal U2B*-U2A* protein complex bound to a fragment of U2

40. Tishchenko,S., Nikulin,A., Fomenkova,N., Nevskaya,N., Nikonov,O.,
Dumas,P., Moine,H., Ehresmann,B., Ehresmann,C., Piendl,W. et al. (2001)
Detailed analysis of RNA-protein interactions within the ribosomal protein
S8-rRNA complex from the Archaea Methanococcus jannaschii. J. Mol.
Biol., 311, 311–324.

42. Wimberly,B.T., Guymon,R., McCutcheon,J.P., White,S.W. and
Ramakrishnan,V. (1999) Detailed view of a ribosomal active site: the

44. Fedorov,V.R., Meshcheryakov,V.A., Gongadze,G.M., Fomenkova,N.P.,
Nevskaya,N.A., Selmer,M., Lauberg,M., Kristensen,O., Al-Karadaghi,S.,
Liljas,A. et al. (2001) Structure of ribosomal protein TLS complexed with
RNA provides new insights into the CTC family of stress proteins. Acta

The Kink-Turn: a new RNA secondary structure motif. EMBO J., 20,
4214–4221.

Selection of tRNA by the ribosome requires a transition from an open to a

The structure of threonyl-tRNA synthetase-rRNA(Thr) complex enlightens
its repressor activity and reveals an essential zinc ion in the active site.
Cell, 97, 371–381.

Ile-tRNA synthetase structure with tRNA(Ile) and Mucipropacin. Science,
285, 1074–1077.

Synthesis of aspartyl-tRNA(Asp) in Escherichia coli—a snapshot of the
second step. EMBO J., 18, 6532–6541.

56. Fukai,S., Nureki,O., Sekine,S., Shimada,A., Tao,J., Vassylevy,D.G. and
Yokoyama,S. (2000) Structural basis for double-sieve discrimination of
arginyl-tRNA synthetase: induced conformations during substrates
binding. EMBO J., 19, 5599–5610.

substrate induced conformational changes ensures the amino acid
specificty of Thermus thermophilus prolyl-tRNA synthetase: comparison

structure of methionyl-tRNA(34

How glutaminyl-tRNA synthetase selects glutamine. Structure, 6,
439–449.

structure of Cys-tRNAHyx- EF-Tu-GDPNP reveals general and specific
features in the ternary complex and in tRNA. Structure, 7, 143–156.

Psx55 pseudouridine synthase: nucleotide flipping by an RNA-modifying
catalyst. Enzyme, 107, 929–939.