doi:10.1016/S0022-2836(03)00090-1 J. Mol. Biol. (2003) 327, 767–780

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Structural Diversity and Isomorphism of Hydrogen-bonded Base Interactions in Nucleic Acids

Bernhard J. Walberer[1,2], Alan C. Cheng[1,2] and Alan D. Frankel[1]*

[1]*Department of Biochemistry and Biophysics University of California 513 Parnassus Avenue San Francisco CA 94143-0448, USA*

[2]*Graduate Group in Biophysics University of California San Francisco CA 94143-0448, USA*

**Corresponding author*

The wide structural diversity of RNA results in part from the diversity of non-Watson–Crick interactions between bases. To examine the repertoire of possible hydrogen bond interactions among bases, we computed databases of base-pairs and base-triples by systematically matching all possible hydrogen-bond donors and acceptors between bases and evaluating the geometries of each planar configuration. For base-pairs, we find 53 arrangements having at least two hydrogen bonds, including 23 pairs with protonated bases that have not previously been modeled. A comparison with experimentally observed base-pairs reveals an unexpected G:U pair recently observed in the ribosome. For base-triples, we find 840 arrangements in which the three bases are constrained by a total of at least three hydrogen bonds. Base-triples in particular exhibit a wide range of structural diversity, suggesting how compact or elongated nucleic acid structures may be constructed using different hydrogen-bonding patterns. Base-pair and base-triple conformations were systematically compared to identify structurally isomorphic combinations, and the experimentally observed arrangements within double and triple helices are among the most isomorphic. Unexpectedly, however, other combinations in the database are even more isomorphic, including several in which all-purine arrangements overlap with all-pyrimidine arrangements. These studies highlight some of the combinatoric and geometric versatility of base interactions and help provide a framework for analyzing and modeling isomorphic interactions and potentially for designing novel nucleic acid structures.

Keywords: base-pairs; base-triples; computer modeling

Introduction

Hydrogen bonding interactions often make important contributions to the assembly of macromolecular structures and to the specificity of macromolecular interactions, due in part to their geometric constraints.[1–4] In nucleic acids, hydrogen bonds between bases help establish the overall organization of a polynucleotide structure by participating in the formation of helical secondary structures and stabilization of specific three-dimensional shapes. Although Watson–Crick base-pairing is the most common type of interaction, it is widely appreciated that non-Watson–Crick base-pairs, triples, and quadruples are critical for forming non-helical structures, such as those found in RNAs.[5,6] Base-triples, for example, can provide key interactions in assembling a tertiary structure by docking a base-pair in a helical region to a single-stranded nucleotide distant in the polynucleotide chain. These types of interactions have provided key constraints for modeling nucleic acid structure.[7–10] Although hydrogen bonds between bases clearly are not the sole determinant of nucleic acid structure or thermodynamic stability, base stacking, van der Waals interactions, electrostatic interactions, and interactions involving the sugar, phosphate backbone all play major roles, they are often an important component of structural specificity.

Early modeling studies of base-pairing interactions identified 28 arrangements in which two bases can adopt fixed relative geometries, being tethered together by two or three hydrogen bonds.[3,11,12] Later modeling studies identified one additional G:C arrangement.[13,14] Most of these

Present address: A. C. Cheng, Pfizer Discovery Technology Center, 620 Memorial Drive, Cambridge, MA 02139, USA.

E-mail address of the corresponding author: frankel@cgl.ucsf.edu

predicted pairings have been observed experimentally either in RNA structures or in DNA helices containing mismatched bases.[5] The potential diversity of base-triples is less well understood and relatively few have been observed. Base-triples initially were inferred from the observation that particular homopolymers could form three-stranded RNA structures under some conditions.[3] Three major classes of DNA triple helices now are well defined and involve the docking of a third strand into the major groove of a Watson–Crick helix *via* hydrogen bonding to the Hoogsteen face of the base-paired purines.[15] Another type of triplex structure recently has been observed in an RNA pseudoknot in which the third strand docks into the RNA minor groove.[16] Other types of base-triples, some completely lacking Watson–Crick pairing, have been observed in tRNAs, the ribosome, and other RNA structures.[17–21]

Given the importance of base–base interactions in defining nucleic acid structure and function, we undertook a computational approach to systematically explore how bases can recognize each other through specific sets of hydrogen bonds. We calculated databases of base-pairs and base-triples, including interactions with protonated adenine and cytosine bases, which are known to occur in some structures even near neutral pH[3] and which provide additional types of hydrogen-bonding schemes. Analysis of the interactions has identified some unexpected arrangements and structural similarities between combinations. The databases and derived isomorphic relationships may be used together with biochemical and genetic covariation data to model plausible base arrangements within a structure, which rarely are defined unambiguously from the data alone, especially for base triples.[17,22–24] In addition, some multiply hydrogen-bonded arrangements or those with highly isomorphic partners may be particularly interesting candidates for designing novel nucleic acid structures.

Computational Approach

Modeling approach

The overall scheme for systematically calculating arrangements of planar, hydrogen-bonded base-pairs and base-triples is based on simple geometric and steric considerations and is shown in Figure 1(a). First, a single hydrogen bond is formed between two bases in a planar configuration for every possible combination of donor and acceptor groups, with a total of two bonds formed for a base-triple, one between each pairwise partner. Each single hydrogen bond is formed with the two possible planar orientations of the bases, generated by 180° rotation about the bond. Base nitrogen atoms with covalently bonded hydrogen atoms are treated as donors, and nitrogen and oxygen atoms with free electron pairs, other than purine N9 or pyrimdine N1 atoms, are treated as acceptors. Next, the allowed geometric space of each preformed hydrogen bond is systematically sampled by rotating one base around two angles with respect to the other base (see Figure 1(b)), and each conformation is evaluated for the formation of additional hydrogen bonds and the absence of steric clashes. Finally, the best conformation, as judged by a simple scoring function that favors linear hydrogen bonds (see below), is identified and all unique hydrogen-bonding arrangements are stored in a database, including those with single hydrogen bonds between pairs of bases.

Adenine (A), guanine (G), cytidine (C), and uracil (U) bases having covalent bond distances and angles within one standard deviation of average values[3] were chosen arbitrarily from one tRNA crystal structure (pdb identifier 4tra; residues 66, 42, 27, and 52). Hydrogen atoms were added using Insight II (MSI). Protonated A and C were generated by adding single hydrogen atoms to the N1 or N3 positions, respectively, leaving the heterocyclic rings unchanged. Possible interactions of thymine (T) were considered to be a subset of those formed with U, reasoning that the C5 methyl group of T cannot form hydrogen bonds and only would add an additional steric constraint.

Three parameters were used to define acceptable hydrogen bond geometries: the distance between the donor and acceptor heavy atoms and two angles centered at their atomic positions (Figure 1(b)). The initial hydrogen bond was formed at the observed mean distance for each donor–acceptor pair[3] and was not subsequently varied. The maximum length for any additional hydrogen bond was 3.2 Å and the minimum length was determined by the donor and acceptor van der Waals radii. Allowable ranges for the donor angle ($0 \pm 18°$) and acceptor angle ($0 \pm 51°$ for a nitrogen atom and $0 \pm 90°$ for a carbonyl oxygen atom) were chosen (along with a steric parameter) to include slightly unreasonable geometries and are integral multiples of the angular step size used to generate the various conformations. The donor and acceptor angles were varied in step sizes of 3°, which generates a large number of conformations (455 for each N–H···N bond and 793 for each N–H···O bond). For base-triples, two hydrogen bonds were formed initially between the two pairs of bases, resulting in $\sim 10^5$–10^6 planar conformations as each hydrogen bond was varied independently. A steric test was performed for each conformation by calculating all pairwise atomic distances prior to determining whether additional hydrogen bonds could be present. A steric parameter (0.8) was used to reduce the hard sphere radii of all atoms[25] and was chosen empirically to ensure that all known base-pair arrangements would be identified. An additional parameter was added to remove conformations containing close contacts between similarly charged groups. For this purpose, atoms that can serve as either a

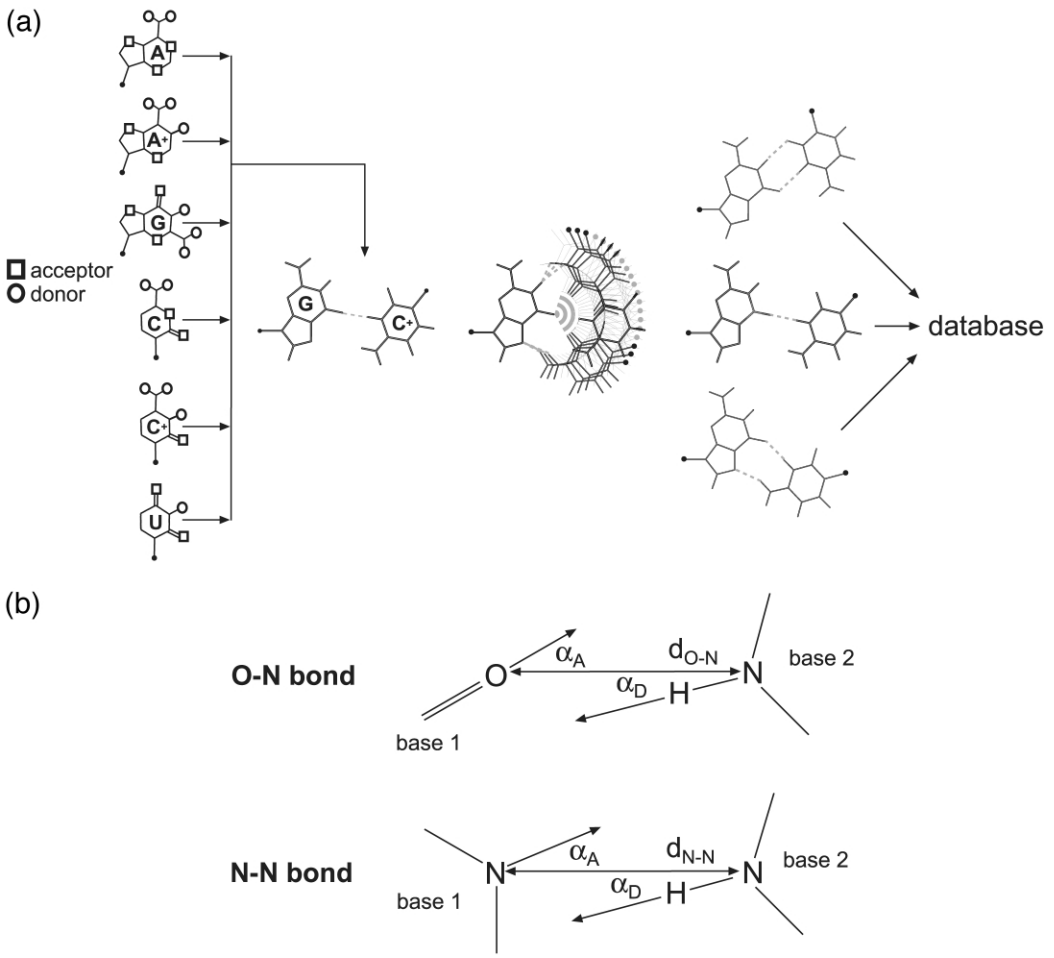

Figure 1. (a) The key steps of the modeling approach, illustrated for a base-pair. The six individual bases are shown on the left, with hydrogen bond donors (open circles) and acceptors (open squares) and the sugar C1′ atom (black dot) indicated. A single hydrogen bond is formed between two bases for every combination of donor and acceptor and for each of the two possible planar orientations of the two bases (only one orientation is shown). The bases are then rotated systematically around the acceptor and donor positions in the plane of the preformed hydrogen bond (middle). For clarity, only a few rotations around the acceptor (the O6 carbonyl oxygen atom of G in this example) are shown; the independent rotations around the donor (the N3 imino nitrogen atom of C^+) are not shown. Each conformation is tested for steric clashes and for formation of additional hydrogen bonds (dotted lines), and one conformation of every unique hydrogen bonded arrangement is retained. (b) The three parameters used to define a hydrogen bond; the acceptor angle (α_A), the donor angle (α_D), and distance between heavy atoms (d_{N-N} for amino-imino hydrogen bonds and d_{O-N} for carbonyl-amino hydrogen bonds). Although the acceptor angle for a carbonyl oxygen atom is defined in a linear manner with respect to the oxygen, we utilize a wider parameter than for a nitrogen acceptor when defining a hydrogen bond to account for its wider electron density distribution (see Computational Approach).

hydrogen bond donor or acceptor were considered negatively charged and polar hydrogen atoms were considered positively charged. A minimum separation of 2.9 Å was chosen, which still retained all known pairing arrangements.

Because the modeling procedure generates many possible conformations for each hydrogen-bonding arrangement, a simple scoring function was used to retain the most geometrically ideal conformation as a representative. A value was calculated for each hydrogen bond, $F(\alpha, d) = \sum\{C_\alpha(1 - \cos^2\alpha) + C_d|d^2_{mean} - d^2_{calc}|\}$, where α is the donor angle, d_{mean} is the mean observed donor–acceptor distance (from Saenger[3], d_{calc} is the calculated distance, and $C\alpha$ and C_d are constants ($C\alpha = 100$ and $C_d = 2000$) chosen empirically to give approximately equal weights to deviations in the distance and angle. This value is zero when $\alpha = 0°$ and $d_{calc} = d_{mean}$. A score for each conformation was assigned by summing all hydrogen bonds in a given arrangement, and the conformation with the lowest value was retained in the database. Due to the weak dependence of hydrogen bond strength on the acceptor angle[1–4] the donor angle alone was used in scoring. Redundant arrangements resulting from the combinatorial docking strategy were removed. Calculations were performed on two 300 MHz Pentium II processors using the Linux operating system. The base-pair database was calculated in about one minute and the base-triple database in about

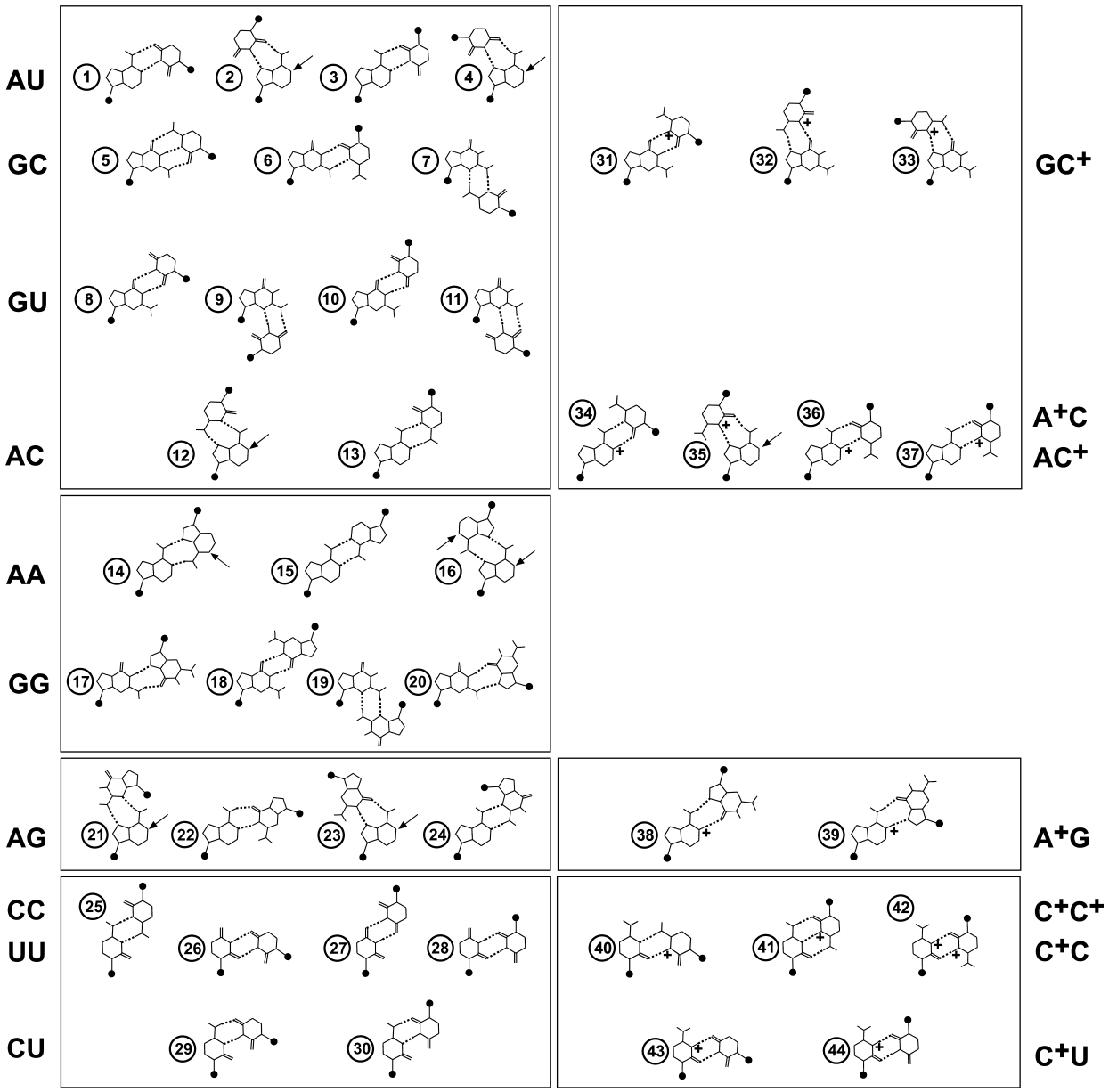

Figure 2. The 44 structurally distinct base-pairs with at least two hydrogen bonds. The 30 unprotonated base-pairs are shown on the left and the 14 unique protonated arrangements on the right. Arrows indicate positions that can be protonated but do not result in additional hydrogen bonds; thus we consider only 44 of the 53 possible base-pairs structurally distinct. Nine of the 44 base-pairs (#11, 25, 32, 35–37, 42–44) have not been observed.

three weeks. Both the base-pair and triple databases included all unprotonated and protonated combinations.

Arrangements in which two bases are tethered by a single hydrogen bond were classified as flexible and those in which each base is involved in at least two hydrogen bonds as fixed. Base-triples in which all three possible pairs were tethered by a single bond also were considered fixed. Some conformations contained bifurcated hydrogen bonds, but to simplify the database, these arrangements were removed because their individual, non-bifurcated arrangements present in the database had very similar conformations, with only small shifts in the relative orientations of the bases. Any arrangement was considered structurally redundant if its pattern of hydrogen bonding was a subset of another and, for fixed arrangements, if the two fell within a 0.5 Å all-atom r.m.s. limit.

Two additional filters based on solvation and backbone conformational criteria were used to remove improbable conformations of the fixed base arrangements. First, some arrangements contained charged groups not involved in a hydrogen bond and without sufficient space for a water molecule, an energetically unfavorable situation. Such arrangements were identified by bonding a 2.8 Å diameter sphere approximating a water molecule to the unoccupied acceptor or donor atom(s), rotating the sphere in 3° increments ($\pm 90°$),

and measuring steric accessibility. A single, unprotonated C:C base-pair was eliminated by this filter, which is related to the C:C$^+$ pair observed experimentally only in its protonated form (see Figure 2, #41). Nearly 50% of the base-triple database was eliminated by this filter. Second, we wished to ensure that all arrangements could sterically accomodate the sugar–phosphate backbone. The backbone was represented by a set of ribose sugar conformations with an enlarged C5′ atom (1.9 Å instead of 1.7 Å) to approximate the rest of the backbone. We calculated 141 conformations of the ribose ring in all possible sugar puckers[26] and placed them into four sets, the first containing the two most common C2′-*endo* and C3′-*endo* conformations, the second containing nine additional commonly observed conformations, the third containing 33 still within the range of observed conformations,[3] and the fourth containing the remaining 97, many of which are likely to be energetically unfavorable but are sterically possible. The first set of sugars was added to all fixed base arrangements combinatorially, varying the glycosidic bond angle in 10° increments, and tested for steric problems. Sets 2–4 were used sequentially only for those arrangements that failed the steric test with the common sugar conformations. All base-pairs could accommodate the C2′-*endo* and C3′-*endo* rings, and only 11 base-triples could not. Four base-triples could not accommodate any backbone conformation and were removed from the database.

Energy minimization

The computed databases were calculated by a simple geometric search and evaluated only planar arrangements. To test whether these conformations are energetically reasonable, we performed energy minimization of each arrangement using the program X-PLOR[27] and examined changes in structure. Only unprotonated bases were examined using partial charges provided by Dr Peter Kollman (personal communication); partial charges for protonated bases were not available. All other parameters were from Cornell *et al.*[25] Each base was treated as a rigid body and hydrogen bonds were maintained by constraining the coordinates of the donor and acceptor atoms with a harmonic force field (force constant = 20 kcal mol^{-1} Å^{-1}). Energy minimization was terminated after 10,000 steps or if the norm of the energy gradient was less than 0.001. Individual energies were estimated for the starting and energy-minimized structures as the sum of the van der Waals and electrostatic interaction energies between the individual bases, and r.m.s.d. values were calculated between the initial and minimized structures. The few unfavorable conformations observed prior to minimization almost exclusively reflected steric clashes; this was expected because the databases were calculated with reduced steric parameters compared to those used in the energy calculations.

Base interactions in known structures

To further evaluate the quality of the calculated databases, we compared them to compilations of computed hydrogen-bonded base arrangements in polynucleotide-containing structures, including nucleic acids complexed to proteins and the 30 S and 50 S ribosomal subunits. We generated one compilation utilizing hydrogen bonding citeria described below, and also performed comparisons to databases constructed by independent methods[28,29]† to reduce possible biases of hydrogen bond identification.

The hydrogen bond parameters (0–3.4 Å distance and 0 ± 35° donor angle) used for our search of observed interactions were less stringent than those used to compute the databases to allow for some experimental imprecision. Interactions were classified solely based on their hydrogen bonding patterns, allowing comparisons between the calculated planar conformations and observed non-planar conformations. The N1 position of A and N3 position of C were treated as acceptors and, if not involved in a hydrogen bond, also as donors to check for possible hydrogen bonds to protonated bases. Potential donors and acceptors on modified bases also were considered, defining these atoms based on their covalently bonded partners. Base-pairs or triples that are part of larger hydrogen-bonded arrangements were output in all possible arrangements. Observed arrangements were matched to the corresponding arrangements in the model databases. Many observed base arrangements contained bifurcated hydrogen bonds, particularly given the wider parameters used, and these were output as the two singly hydrogen-bonded forms, as described above. Due to the relaxed hydrogen bonding parameters used in this search, we identified ten fixed base-pairs (of 4741), 52 fixed triples (of 1955), 82 flexible pairs (of 2049), and 333 flexible triples (of 1304) that had no matching counterparts in our databases. All of these had either significant steric clashes, hydrogen bond values near the limits of the relaxed parameter ranges, or exhibited highly non-planar geometries. The absence of these arrangements points to two limitations of our approach: (1) the length of our initially formed hydrogen bond remains fixed, and (2) highly non-planar geometries are not well modeled. Nevertheless, all of the absent arrangements represent extreme cases and probably are not stabilized principally by hydrogen bonding. In searching for interactions within the ribosomal subunits, we used AMBER PROTONATE[25] to add hydrogen atoms to 30 S (1HRO) and 50 S (1FFK) subunit structures. The G:U base-pair discussed in the text also was observed in other reported 30 S structures (1FJF and 1FJG).[18,19,30–36]

† www.lbit.iro.umontreal.ca.prion.bchs.uh.edu/bp_type

Table 1. Numbers of computed base-pairs and base-triples

	Base-pairs		Base-triples	
	Unprotonated	Protonated	Unprotonated	Protonated
Fixed	30	23	307	533
Flexible	140	180	9,475	21,057

Calculation of isomorphic combinations

All pairwise combinations of base-pairs and all pairwise combinations of base-triples were overlapped to determine their degree of structural similarity, or isomorphism. The overlap was performed for each of the two possible pairwise arrangements of bases for base-pairs and for each of the six possible pairwise arrangements of bases for base-triples such that the positions of the glycosidic bonds were optimally aligned. The quality of each overlapped combination was defined by an I (Isomorphism) value; $I = \sqrt{\{(\Delta\alpha)^2 + \omega(\mathrm{rms})^2\}}$, where rms is the r.m.s.d. of all overlapped glycosidic bonds, $\Delta\alpha$ is the average angle between the overlapped glycosidic bonds, and ω is the slope of the curve that bisects the distribution of overlaps of either base-pairs or base-triples in the (rms, $\Delta\alpha$) plane (see Results and Figure 6). A low I value indicates a high degree of structural similarity. Both rms and bond angle parameters are required to completely specify the degree of overlap.[17]

Results

Database construction

We computed databases of hydrogen-bonded base-pairs and base-triples utilizing a geometric approach in which a single hydrogen bond was formed between two bases in a planar configuration, followed by a conformational search to identify additional hydrogen bonds (Figure 1(a)). The hydrogen bonding and steric parameters of the search were set to help generate thorough databases. All arrangements identified can sterically accommodate the sugar–phosphate backbone. The computed databases contain all combinations in which two bases are bridged by at least one hydrogen bond. In all, there are 373 base-pairs and 31,372 base-triples (Table 1), of which 53 pairs and 840 triples have "fixed" spatial conformations constrained by at least two hydrogen bonds (for pairs) or three hydrogen bonds (for triples; see the hydrogen bonding arrangements in Table 2).

In addition to the steric criteria used to construct the databases, we applied three additional tests to ensure that the calculated conformations were structurally and energetically reasonable. First, we eliminated arrangements in which two charged groups not involved in a hydrogen bond were too close to be solvated. Second, we estimated interaction energies of each arrangement, performed energy minimization, and examined changes in structure. Fewer than 1% of the base-pairs or triples showed unfavorable energies prior to minimization, estimated as the sum of the van der Waals and electrostatic interaction energies between the individual bases, and none was unfavorable after minimization. r.m.s.d. values between the initial and minimized structures showed relatively small deviations. Third, we assessed the quality of the calculated databases by comparisons to observed base interactions; virtually all observed interactions were present, except for a small number (see Computational Approach) that are significantly non-planar or have poor hydrogen bonding geometries and probably are not reasonable.

The calculated databases include the protonated bases, A^+ and C^+, with 14 pairs and 319 triples having fixed conformations that utilize the protonated positions for hydrogen bonding and thereby generate structurally distinct interactions. Thus, protonation can significantly expand the diversity of base interactions. In reality, the diversity of interactions is even greater than our calculated databases in that we do not explicitly consider water-mediated hydrogen bonds, hydrogen bonds to 2′ OH or phosphate backbone groups, or CH···O or CH···N bonds.[16,21,37,38] Nevertheless, arrangements involving these interactions usually include at least one direct hydrogen bond between base moieties, which are included in the databases, and thus are indirectly represented in our databases. By the same token, the databases indirectly

Table 2. Distribution of fixed base-pairs within base-triples, sorted by purine (R) and pyrimidine (Y) composition

Composition	0 fixed pairs	1 fixed pair	2 fixed pairs	Total
RRR	40	21	84	145
RRY	125	88	138	351
RYY	92	121	58	271
YYY	23	50	0	73
Total	280	280	280	840

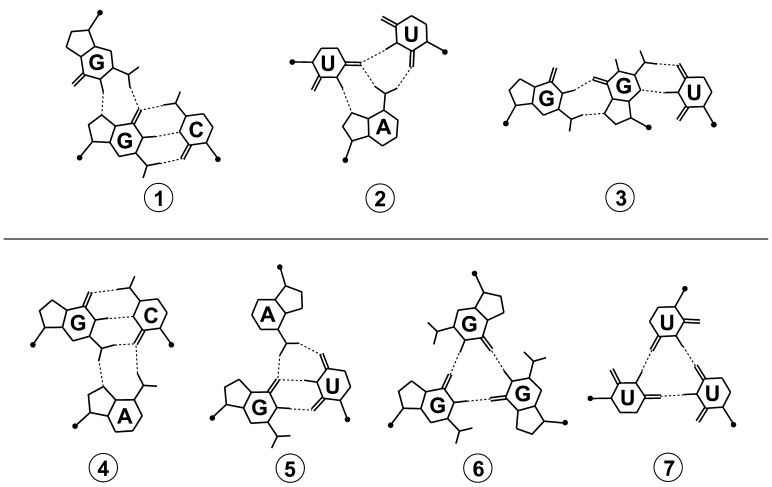

Figure 3. Examples that highlight the structural diversity of base-triples. Triples in the upper row (#1–3) have been observed whereas triples in the lower row (#4–7) have not. The characteristics and shapes of various triples are described in the text.

encompass arrangements having bifurcated hydrogen bonds, which are represented by the corresponding pair of singly bonded interactions. For the purposes of this work, we discuss only fixed arrangements of base-pairs and triples tethered *via* base–base hydrogen bonds.

Base-pairs

For base-pairs, our modeling yielded 44 distinct fixed arrangements (with two or three hydrogen bonds), including both unprotonated and protonated bases (Figure 2), and 35 of these have been observed. (Note that the remaining nine of the total of 53 base-pairs mentioned above are variants in which bases are protonated but no additional hydrogen bonds are formed and thus have redundant hydrogen bonding patterns; see Figure 2). Of the 30 unprotonated arrangements, two unexpected G:U pairs were found (#9, #11 in Figure 2) that were not identified in previous modeling studies,[13,14] probably due to the close proximity of the carbonyl groups of U and the C1′ atom of G, and have not been included in recent base-pair compilations.[39] In both G:U pairs, the pyrimidine base hydrogen bonds to the N3 imino and N2 amino groups of G, similar to a G:C arrangement (#7) that was previously modeled but not included in earlier base-pair compilations.[40] One of the G:U pairs (#9) now has been observed, first in NMR-derived models of a thrombin-binding DNA aptamer[41–43] and more recently in crystal structures of both ribosomal subunits (G1389:U1435 in the large subunit[18] and G362:U49 in the small subunit.[19]) Most base-pair arrangements not yet observed involve protonated bases, although some protonated arrangements with structurally similar neutral forms (for example, see the reverse Hoogsteen A:C pair #12 and the protonated A:C$^+$ pair #35 in Figure 2) may have been assigned as the neutral form during structure determination. Nonetheless it seems clear that

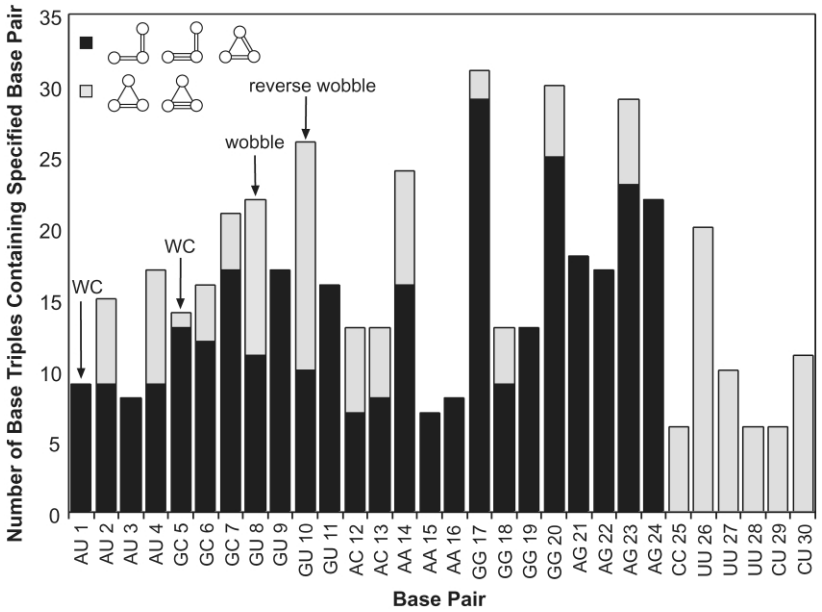

Figure 4. Recognition of fixed base-pairs by a third base. The number of base-triple combinations involving each of the 30 unprotonated base-pairs is plotted (numbers refer to the arrangements shown in Figure 2). For each base-pair, the number of triples in which the incoming base forms a single hydrogen to each base-pair partner (see top left corner) is indicated by shaded bars. These are the only possible arrangements for the six pyrimidine–pyrimidine base-pairs (#25–30).

protonation adds considerably to the diversity of possible base-pairs.

Base-triples

The number of possible base-triple combinations is very large, and obtaining a view of their structural repertiore requires a systematic approach such as that described here. Our modeling yielded 840 structurally distinct base-triples having fixed arrangements, including 307 unprotonated and 533 protonated triples, with only a small fraction of the total observed so far. The 50 S ribosomal subunit contains 27 triples, including nine without any Watson–Crick base-pairing and three involving a protonated base; the 30 S subunit contains ten triples, including four without Watson–Crick pairing. In our database, fixed arrangements containing every possible base composition are observed, and more than two-thirds of the triples contain single hydrogen bonds between at least two bases. The maximum number of hydrogen bonds within any triple is five, and there are 28 such arrangements. Selected examples of base-triples within the database (Figure 3) illustrate some of their wide structural diversity.

Base-triples can be thought of as being composed of either two or three base-pairs, but the formation of a fixed triple does not require that any pair itself be fixed by two hydrogen bonds (see Figure 3). Indeed, one-third of the triples are composed of three pairs, each bridged by a single hydrogen bond (Table 2). The remaining two-thirds of the triples contain one or two fixed base-pairs (Table 2). The number of arrangements in which a third base can be used to recognize each of the 30 unprotonated fixed base-pairs (see Figure 2) is shown in Figure 4. It is evident that six of the base-pairs only can be bridged by single hydrogen bonds to each base.

Isomorphic combinations

Base combinations that overlap closely in the positions of their glycosidic bonds can often functionally substitute for one another in the context of a given structure.[17,23,44,45] Given our extensive databases of base-pairs and triples, we wished to determine which combinations were most isomorphic. We systematically overlapped all pairwise combinations in each database, optimizing the overlap of all glycosidic bonds in all different

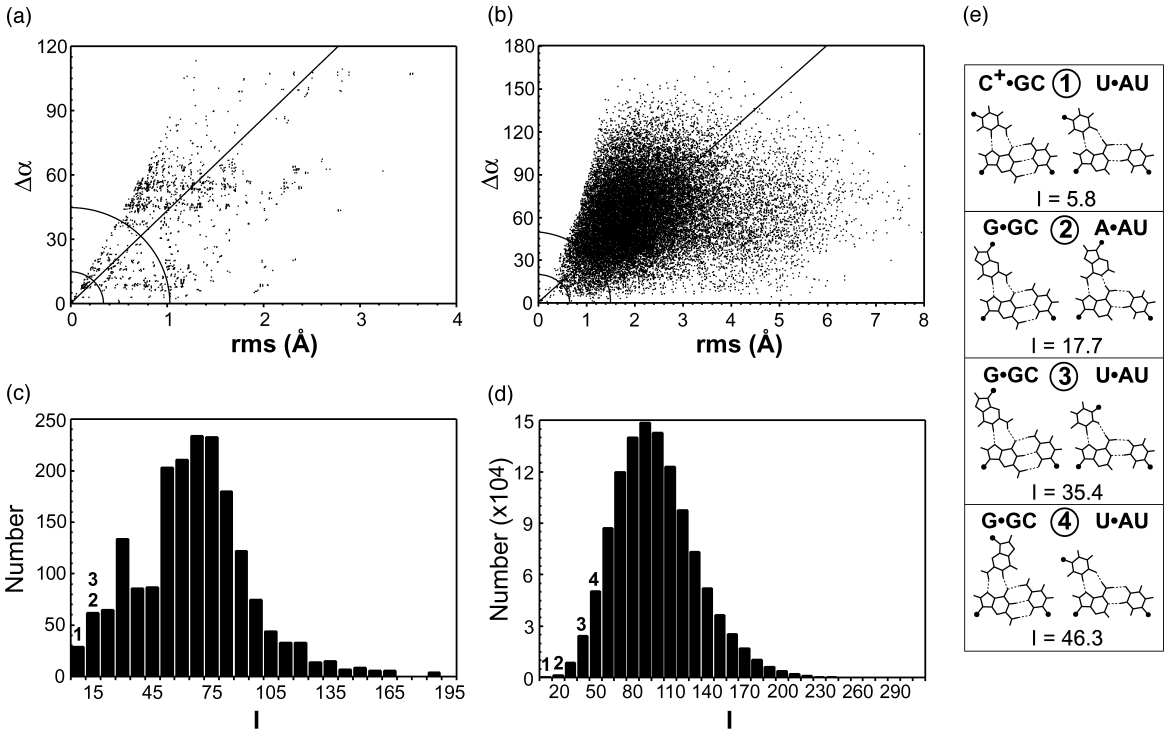

Figure 5. Isomorphic relationships between base-pairs and base-triples. The glycosidic bonds for all pairwise combinations of base-pairs or base-triples were overlapped, and the r.m.s.d. values of the glycosidic atoms and the average difference in angle of the overlapped bonds were plotted. The entire distribution for base-pairs (a) and a randomly chosen 2.5% of the base triples (b) are shown. Isomorphic (*I*) values were calculated (see Computational Approach) and their distributions for base-pairs (c) and base triples (d) were plotted. In (a) and (b), the curves that bisect the distributions of overlaps and whose slopes were used to calculate *I* values (see Computational Approach) are shown, as well as curves that represent constant *I* values ($I = 15$ and $I = 45$ for base-pairs, and $I = 20$ and $I = 40$ for base-triples). In (c), the overlapped Watson–Crick base-pairs (1) and each Watson–Crick base-pair overlapped with the GU wobble pair (2,3) are indicated by the numbers. In (d), the overlapped base-triples corresponding to those observed in triple helices (e) also are indicated by the corresponding numbers (1–4).

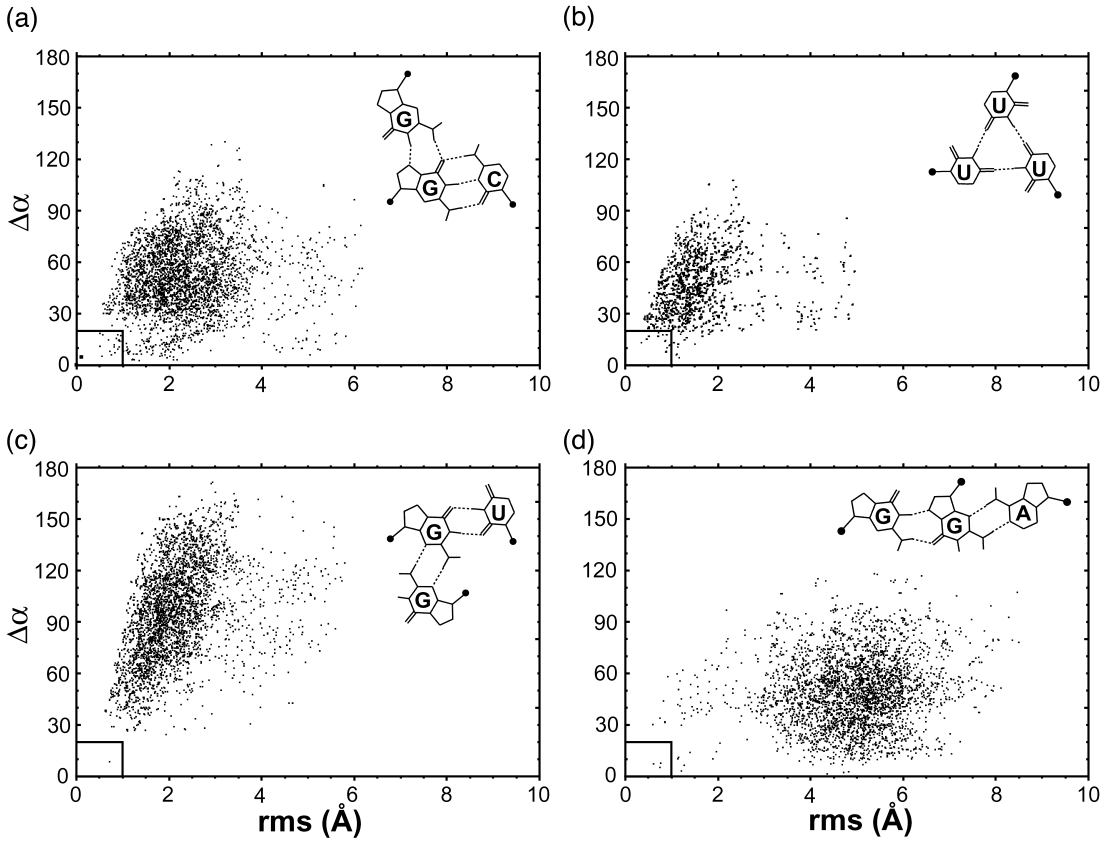

Figure 6. Comparisons of individual base-triples with the database. For a given base-triple, there typically are only few very close overlaps, indicated by the arbitrary box drawn near the origin. Compact base-triples tend to show the largest number of close overlaps (a), (b), independent of their base composition. Triples with unusual glycosidic bond arrangements (c) or elongated triples (d) tend to have few closely overlapping partners.

alignments (two for pairs and six for triples), and derived a value, *I*, that approximates the quality of each overlap. An *I* value of zero represents a perfect overlap of glycosidic bonds. The *I* values do not specify whether the *syn*/*anti*-orientations are the same, which is necessary if two base combinations are to be interchanged in any structural context.[46] No particular *I* value is used to establish that two combinations are isomorphic, unlike a previous analysis,[17] but rather the *I* values specify the relative degree of isomorphism.

To visualize "isomorphism" within the entire databases, we plotted the r.m.s.d. values of all overlapped glycosidic atoms and the average deviation of their bond angles (Δα) for all pairwise combinations of base-pairs and triples (Figure 5(a) and (b)), and plotted histograms for the corresponding (*I*) values derived from these two parameters (Figure 5(c) and (d)). Points closest to the origin in Figure 5(a) and (b) represent the most similar overlaps and have the lowest *I* values. For base-pairs, the most isomorphic combination is between the Watson–Crick pairs, as expected, and each Watson–Crick pair also is highly isomorphic to the G:U wobble pair (Figure 5(c)). For base-triples, the three known arrangements found in mixed-sequence triple helices (Figure 5(e), #1–3) are among the most isomorphic combinations in the database (Figure 5(d)), and a fourth (#4) that does not form mixed triple helices[15] has a slightly higher *I* value. A relatively large number of triple combinations overlap extremely closely (Figure 5(b); rms <0.1 Å and $\Delta\alpha < 2°$), and these represent cases in which U is substituted by C^+, using a nearly identical ad face (see Discussion and Figure 7) for hydrogen bonding. The equivalence of U and C^+ also is seen with base-pairs.

In addition to plotting the isomorphic relationships for the entire database, it can be instructive to plot the distribution of all overlaps for a chosen base-triple. This might be used, for example, to predict an isomorphic substitution for a known or inferred base-triple. Four examples are shown in Figure 6, utilizing base-triples with differing characteristics. From these plots, it is apparent that for any given triple, very few others closely overlap and might be considered isomorphic (arbitrary boxes indicating rms <1 Å and $\Delta\alpha < 20°$ are shown). Isomorphic arrangements are especially rare for those involving the N3–N2 hydrogen bonding face of G (Figure 6(c)) or for extended base-triples (Figure 6(d)), which comprise only 35 of the 840 fixed base-triples.

Discussion

Protonated base-pairs and unanticipated arrangements

We computed 44 unique arrangements of base-pairs containing at least two hydrogen bonds, including 14 arrangements with protonated bases (Figure 2). Unexpectedly, we found two G:U pairs (#9, #11 in Figure 2) not identified by previous modeling studies[13,14] and not generally included in base-pair compilations.[39] Interestingly, one of these (#9) has been observed in both ribosomal subunits and in a DNA aptamer.[18,19,41–43] Among the protonated arrangements, the C:C$^+$ base-pair (#41) is the only pair, aside from the Watson–Crick G:C, to contain three hydrogen bonds and has been observed in intercalated helical structures.[47] The unprotonated form of this pair has been included in previous base-pair compilations[39] but probably is energetically unfavorable and has not been observed. In general, protonated arrangements are expected to be relatively less frequent than neutral arrangements and, indeed, most arrangements not yet observed involve a protonated base. To our knowledge, this study provides the first systematic modeling of base-pairs involving protonated bases and extends the repertoire of possible pairing arrangements.

Diversity of base-triples

From inspection of the database, it is evident that the structures of base-triples can be very diverse and can be used to build distinct RNA architectural elements. A few selected examples, including observed (#1–3) and unobserved (#4–7) cases (Figure 3), illustrate several aspects of this diversity: (1) some triples involve Watson–Crick base-pairs (#1, #4) and are the most commonly observed given that a third base need only dock against an already paired Watson–Crick helix. Nevertheless, most of the computed base-triples do not utilize Watson–Crick pairing and some of these have been observed (#2, #3). (2) A third base can span across a base-pair, forming hydrogen bonds to both partners (#4, #5). In contrast, the regular triple helical structures observed to date have involved hydrogen bonding of a third strand to only one strand of a Watson–Crick helix.[15] (3) The shapes of some triples are relatively triangulated and compact, particularly when a third base spans both partners of a base-pair (#2, #4–7), whereas others are elongated and extended (#3). (4) Similar shapes can be formed by triples with rather different base compositions, even involving all-purine and all-pyrimidine arrangements (#6, #7). (5) Fixed triple arrangements can be formed with just three hydrogen bonds when they are used to close a ring (#6, #7). (6) As with base-pairs, protonation substantially widens the diversity of possible base-triples, representing nearly two-thirds of the calculated triples. The diversity of base-triples provides opportunities to test predicted interactions in a variety of structural contexts, as described below, although it clearly must be recognized that the energetic components of nucleic acid structure are far more complex than the simple geometric and hydrogen bonding criteria applied here.

Consequences of the donor and acceptor arrangements of the bases

The ability of bases to form fixed base-pair and triple arrangements can be understood by considering the number and spatial arrangement of hydrogen bond donor and acceptor groups on each base (Figure 7). To form a fixed base-pair, two adjacent acceptor (a) or donor (d) groups must be complementary to those of its pairing partner. The six bases (including A$^+$ and C$^+$) contain a total of nine ad, three dd, and two aa faces and thus those displaying the most ad faces should produce the most abundant pairing arrangements. G, A, and U each contain two ad faces and indeed generate the greatest diversity (Figure 2). Interestingly, protonation of A decreases its diversity, converting an ad face into a dd face. Also, the pyrimidine U is as diverse as the purine A with respect to its base-pairing potential, despite having fewer total donor and acceptor groups. Rather than considering the explicit arrangements of donors and acceptors, it also is possible to classify base interactions according to the three "edges" common to all bases: a Watson–Crick edge, a Hoogsteen edge, and a sugar edge (which also includes the 2′OH group).[48]

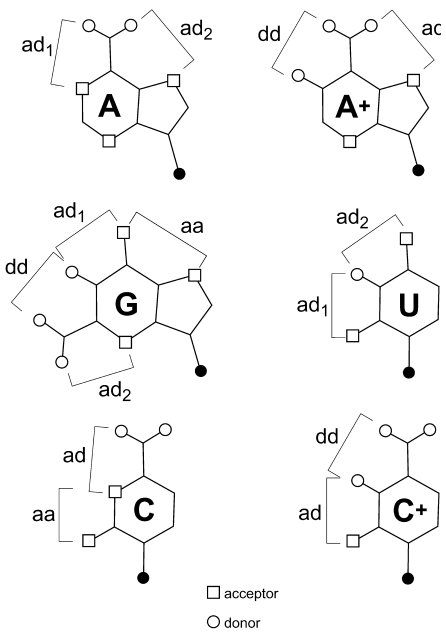

Figure 7. Distribution of functional groups around each base. All pairwise arrangements of hydrogen bond acceptors (a) and donors (d) that can simultaneously form two hydrogen bonds to another base are indicated.

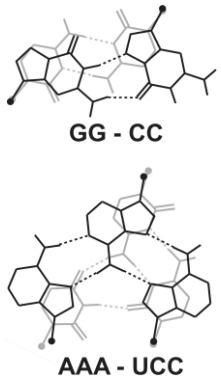

Figure 8. Examples of all-purine (black) and all-pyrimidine (grey) arrangements whose glycosidic bonds can be closely overlapped. Despite the large differences in ring size, several such closely overlapping arrangements can be found.

Once in a base-pair, the donor and acceptor arrangements present new surfaces for recognition by a third base. In some cases, one face is still available to form two hydrogen bonds with the incoming base, thereby forming two independently fixed base-pairs. However, in most cases (two-thirds), the incoming base spans both partners of a base-pair with just one hydrogen bond to each (Table 2; Figure 3; #6, #7), and in fact several base-pairs can only be bridged in this manner (Figure 4). Sorting the database by purine/pyrimidine content (Table 2) shows that no all-pyrimidine arrangement can be composed of two independently fixed pairs, as pyrimidines do not possess a sufficient number of donors and acceptors. In addition, the formation of elongated or extended triples (see Figure 3; #3) requires at least one purine, and it must be located in the middle of the structure to provide two binding faces that each form two hydrogen bonds to another base. Among observed bases-triples, those composed of two independently fixed pairs are more typical (see #1, #3 in Figure 3) although singly hydrogen bonded spanning interactions are found (see #2 in Figure 3).

Isomorphic relationships

Perhaps one of the most interesting aspects of the calculated databases is the ability to systematically examine isomorphic relationships amongst base-pairs and especially amongst the large database of triples. For base-pairs, the Watson–Crick and G:U wobble pairs found in double helices are among the most isomorphic combinations (Figure 5(C)) but, interestingly, some other combinations are even more isomorphic, including an all-purine G:G pair and an all-pyrimidine C:C pair (Figure 8). Similarly, the four known base-triples found in triple helices are among the most isomorphic

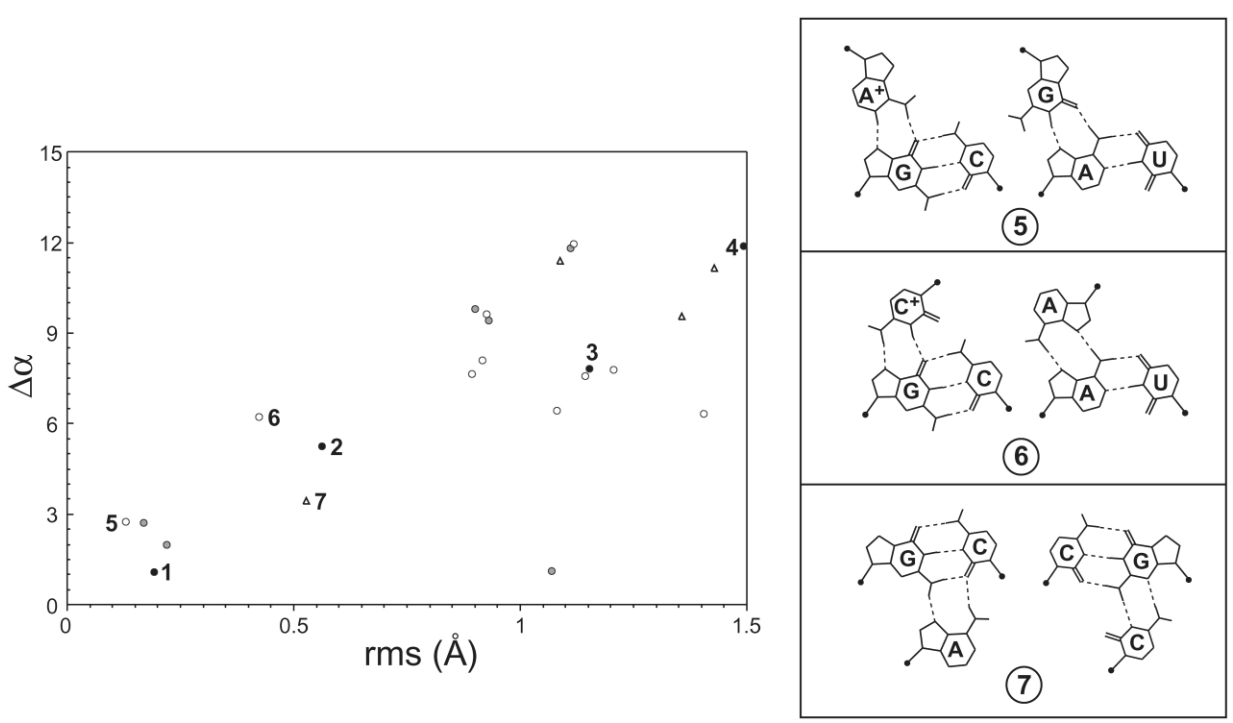

Figure 9. Isomorphic relationships of base-triples containing Watson–Crick base-pairs. The plot shows only those arrangements in which the Watson–Crick pairs are correctly aligned. Alignments of A:U and G:C pairs are indicated by circles, and alignments of G:C and C:G pairs by triangles. Gray circles indicate overlaps in which the third base is either the same or is protonated and thus would not be expected to bind in a sequence-specific manner. No overlapped arrangements of A:U and U:A pairs, or of A:U and C:G pairs, were observed within the range of parameters shown and thus none would be considered closely isomorphic. The four sets of overlapped triples in observed triple helices are numbered 1–4 (as in Figure 5), and three potentially novel arrangements, discussed in the text, are numbered 5–7.

but many other combinations are at least as isomorphic, including all-purine and all-pyrimidine arrangements (Figure 8). Given that arrangements with such different base compositions can overlap so closely, it seems unlikely that simple rules will be found to identify isomorphic replacements, as might be desired to test a predicted interaction. Rather, a systematic computational approach may be more effective, particularly for triples, where the database may be sorted according to similarity to a chosen arrangement (Figure 6).

The existing helical structures are based on isomorphic combinations but do not necessarily represent the complete repertoire of stable structures. Again, a computational approach may help identify other interesting structures, illustrated for triple helices as follows: in all four known types of triple helices, a third strand is used to recognize a purine-rich strand of a Watson–Crick helix, and these structures form in part because the triples are reasonably isomorphic.[15] To identify other possible arrangements that might form triple helices, we examined the overlap of all triples that contain a Watson–Crick pair (Figure 9) and found five that are at least as isomorphic as the two most isomorphic known triples (#1 and #2; see Figure 5(E)). Two of the five use the same third base to recognize the Watson–Crick pairs but three others (#5–7; Figure 9) use different bases and thus are candidates for forming novel, sequence-specific triple helices. Only #5 maintains the same *syn/anti* orientation for both base-triples, which probably is important for triple helix stability,[15] and likely is the best candidate. Arrangements #5 and #6 use a third base to recognize a purine in the major groove whereas #7 recognizes G:C or C:G pairs in the minor groove and in principle uses an A to span across the base-pair. It will be especially interesting to determine if regular helices can be constructed from entirely non-Watson–Crick arrangements, using the most isomorphic combinations found in the database, in part testing the role of isomorphism in determining helix stability.

Conclusions and utilization of the databases

The databases described contain all geometrically and energetically plausible planar base-pairs and triple combinations, including flexible arrangements tethered by single hydrogen bonds, fixed arrangements with multiple hydrogen bonds, and arrangements utilizing protonated bases. The databases represent all generally recognized planar and non-planar arrangements found in known structures, and it is reasonable to expect that they encompass the vast majority of, if not all, hydrogen-bonded base combinations that may be found. While interactions other than hydrogen bonds were not included in the modeling and will contribute to the structure and stability of any given base interaction, it is clear that many of the described hydrogen bonding arrangements make important contributions in a variety of structural contexts. Given the vast number of sequence and hydrogen bonding possibilities, especially for triples, it seems that systematic approaches taking into account all possible base arrangements will be required to successfully model structures in many unknown cases. Even at the base-pair level, we found two unexpected G:U base-pairs, one now observed in the ribosome, highlighting current limitations to our knowledge of nucleic acid structure and emphasizing the need for such approaches.

To help analyze the large number of possible base–base interactions, particularly for base-triples, it is desirable to sort interactions by their structural characteristics and to visualize individual structures. We have placed the databases, named NAIL (nucleic acid interaction libraries), on a graphical web site†, have devised a set of filters that can be used to identify desired interactions, and have set up a rudimentary user interface. The following filters can be applied: (1) base-pairs or triples, (2) fixed or flexible, (3) base composition, (4) unprotonated or protonated, (5) number of hydrogen bonds, (6) distribution of hydrogen bonds in base-triples, (7) shape of base-triples, (8) specify fixed base-pairs within triples, (9) identify best overlapped combinations, and (10) overlap two specific combinations.

Acknowledgements

We thank Peter Kollman, Robert Fletterick, Aenoch Lynn, and William Chen for helpful discussions and advice, and Richard Shafer, Elizabeth Blackburn, and Valerie Calabro for comments on the manuscript. This work was supported by NIH grants GM56531 and GM47478 (to A.D.F.) and by NIH training grants GM08284 and GM08388 (to A.C.C.).

References

1. Taylor, R., Kennard, O. & Versichel, W. (1983). Geometry of the N–H–O═C hydrogen bond. *J. Am. Chem Soc.* **105**, 5761–5766.
2. Taylor, R. & Kennard, O. (1984). Hydrogen-bond geometry in organic crystals. *Accts. Chem Res.* **17**, 320–326.
3. Saenger, W. (1984). *Principles of Nucleic Acid Structure*, Springer, New York.
4. Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Molecules*, Springer, Berlin.
5. Burkard, M. E., Turner, D. H. & Tinoco, I., Jr (1999). *The RNA World* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds), 2nd edit., pp. 233–264, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
6. Hermann, T. & Patel, D. J. (1999). Stitching together RNA tertiary architectures. *J. Mol. Biol.* **294**, 829–849.

† //www.ucsf.edu/frankel/frankel_homepage.html

7. Levitt, M. (1969). Detailed molecular model for transfer ribonucleic acid. *Nature*, **224**, 759–763.
8. Michel, F. & Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**, 585–610.
9. Gautheret, D., Major, F. & Cedergren, R. (1993). Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol.* **229**, 1049–1064.
10. Major, F., Gautheret, D. & Cedergren, R. (1993). Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc. Natl Acad. Sci. USA*, **90**, 9408–9412.
11. Donohue, J. & Trueblood, K. N. (1960). Base pairing in DNA. *J. Mol. Biol.* **2**, 363–371.
12. Donohue, J. (1956). Hydrogen-bonded helical configurations of polynucleotides. *Proc. Natl Acad. Sci. USA*, **42**, 60–65.
13. Hobza, P. & Sandorfy, C. (1987). Nonempirical calculations on all the 29 possible DNA base-pairs. *J. Am. Chem. Soc.* **109**, 1302–1307.
14. Poltev, V. I. & Shulyupina, N. V. (1986). Simulation of interactions between nucleic acid bases by refined atom–atom potential functions. *J. Biomol. Struct. Dynam.* **3**, 739–765.
15. Sun, J., Garestier, T. & Helene, C. (1996). Oligonucleotide directed triple helix formation. *Curr. Opin. Struct. Biol.* **6**, 327–333.
16. Su, L., Chen, L., Egli, M., Berger, J. M. & Rich, A. (1999). Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature Struct. Biol.* **6**, 285–292.
17. Gautheret, D. & Gutell, R. R. (1997). Inferring the conformation of RNA base-pairs and triples from patterns of sequence variation. *Nucl. Acids Res.* **25**, 1559–1564.
18. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
19. Carter, A. P., Clemons, W. M., Jr, Brodersen, D. E., Morgan-Warren, R. J., Hartsch, T., Wimberly, B. T. & Ramakrishnan, V. (2001). Crystal structure of an initiation factor bound to the 30 S ribosomal subunit. *Science*, **291**, 498–501.
20. Baugh, C., Grate, D. & Wilson, C. (2000). 2.8 Å crystal structure of the malachite green aptamer. *J. Mol. Biol.* **301**, 117–128.
21. Sussman, D. & Wilson, C. (2000). A water channel on the core of the vitamin B(12) RNA aptamer. *Struct. Fold. Des.* **8**, 719–727.
22. Winker, S., Overbeek, R., Woese, C. R., Olsen, G. J. & Pfluger, N. (1990). Structure detection through automated covariance serach. *Comput. Appl. Biosci.* **6**, 365–371.
23. Gautheret, D., Damberger, S. H. & Gutell, R. R. (1995). Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol.* **248**, 27–43.
24. Tanner, M. A., Anderson, E. M., Gutell, R. R. & Cech, T. R. (1997). Mutagenesis and comparative sequence analysis of a base-triple joining the two domains of group I ribozymes. *RNA*, **3**, 1037–1051.
25. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M. *et al.* (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
26. Walberer, B. J. (2000). Construction and analysis of a complete database of hydrogen-bonded base combinations, PhD thesis, University of California, San Francisco.
27. Brunger, A. T. (1992). *X-PLOR, Version 3.1. A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT.
28. Lemieux, S. & Major, F. (2002). RNA canonical and non-canonical base-pairing types: a recognition method and complete repertoire. *Nucl. Acids Res.* **30**, 4250–4263.
29. Nagaswamy, U., Larios-Sanz, M., Hury, J., Collins, S., Zhang, Z., Zhao, Q. & Fox, G. E. (2002). NCIR: a database of non-canonical interactions in known RNA structures. *Nucl. Acids Res.* **30**, 395–397.
30. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science*, **289**, 920–930.
31. Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Jr., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C. *et al.* (2000). Structure of the 30 S ribosomal subunit. *Nature*, **407**, 327–339.
32. Carter, A. P., Clemons, W. M., Jr, Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2000). Functional insights from the structure of the 30 S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 306–307.
33. Brodersen, D. E., Clemons, W. M., Jr, Carter, A. P., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2000). The structural basis for the action of the antibiotics tetracycline, pactamycin, and hygromycin B on the 30 S ribosomal subunit. *Cell*, **103**, 1143–1154.
34. Ramakrishnan, V. & Moore, P. B. (2001). Atomic structures at last: the ribosome in 2000. *Curr. Opin. Struct. Biol.* **11**, 144–154.
35. Ogle, J. M., Brodersen, D. E., Clemons, W. M., Jr, Tarry, M. J., Carter, A. P. & Ramakrishnan, V. (2001). Recognition of cognate transfer RNA by the 30 S ribosomal subunit. *Science*, **292**, 897–902.
36. Schluenzen, F., Tocilj, A., Zarivach, R., Harms, J., Gluehmann, M., Janell, D. *et al.* (2000). Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution. *Cell*, **102**, 615–623.
37. Baeyens, K. J., De Bondt, H. L. & Holbrook, S. R. (1995). Structure of an RNA double helix including uracil–uracil base-pairs in an internal loop. *Nature Struct. Biol.* **2**, 56–62.
38. Brandl, M., Meyer, M. & Suhnel, J. (2001). Quantum-chemical analysis of C–H···O and C–H···N interactions in RNA base-pairs—H-bond *versus* anti-H-bond pattern. *J. Biomol. Struct. Dynam.* **18**, 545–555.
39. Burkard, M. E., Turner, D. H. & Tinoco, I., Jr (1999). *The RNA World* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds), 2nd edit., pp. 675–680, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
40. Tinoco, I., Jr (1993). *The RNA World* (Gesteland, R. F. & Atkins, J. F., eds), pp. 603–607, Cold Spring Harbor Press, Cold Spring Harbor, NY.
41. Schultze, P., Macaya, R. F. & Feigon, J. (1994). Three-dimensional solution structure of the thrombin-binding DNA aptamer d(GGTTGGTGTGGTTGG). *J. Mol. Biol.* **235**, 1532–1547.
42. Macaya, R. F., Schultze, P., Smith, F. W., Roe, J. A. & Feigon, J. (1993). Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution. *Proc. Natl Acad. Sci. USA*, **90**, 3745–3749.

43. Wang, K. Y., Krawczyk, S. H., Bischofberger, N., Swaminathan, S. & Bolton, P. H. (1993). The tertiary structure of a DNA aptamer which binds to and inhibits thrombin determines activity. *Biochemistry*, **32**, 11285–11292.
44. Sampson, J. R., Behlen, L. S., DiRenzo, A. B. & Uhlenbeck, O. C. (1992). Recognition of yeast $tRNA^{Phe}$ by its cognate yeast phenylalanyl-tRNA synthetase: an analysis of specificity. *Biochemistry*, **31**, 4161–4167.
45. Tao, J., Chen, L. & Frankel, A. D. (1997). Dissection of the proposed base-triple in human immunodeficiency virus TAR RNA indicates the importance of the Hoogsteen interaction. *Biochemistry*, **36**, 3491–3495.
46. Heus, H. A., Wijmenga, S. S., Hoppe, H. & Hilbers, C. W. (1997). The detailed structure of tandem G·A mismatched base-pair motifs in RNA duplexes is context dependent. *J. Mol. Biol.* **271**, 147–158.
47. Gehring, K., Leroy, J. L. & Gueron, M. (1993). A tetrameric DNA structure with protonated cytosine·cytosine base-pairs. *Nature*, **363**, 561–565.
48. Leontis, N. B. & Westhof, E. (2001). Geometric nomenclature and classification of RNA base-pairs. *RNA*, **7**, 499–512.

Edited by D. E. Draper

(Received 23 October 2002; accepted 23 December 2002)