



Recognition of Nucleic Acid Bases and Base-pairs by Hydrogen Bonding to Amino Acid Side-chains

Alan C. Cheng^{1,2}, William W. Chen¹, Cynthia N. Fuhrmann¹ and Alan D. Frankel^{1*}

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

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

Sequence-specific protein–nucleic acid recognition is determined, in part, by hydrogen bonding interactions between amino acid side-chains and nucleotide bases. To examine the repertoire of possible interactions, we have calculated geometrically plausible arrangements in which amino acids hydrogen bond to unpaired bases, such as those found in RNA bulges and loops, or to the 53 possible RNA base-pairs. We find 32 possible interactions that involve two or more hydrogen bonds to the six unpaired bases (including protonated A and C), 17 of which have been observed. We find 186 “spanning” interactions to base-pairs in which the amino acid hydrogen bonds to both bases, in principle allowing particular base-pairs to be selectively targeted, and nine of these have been observed. Four calculated interactions span the Watson–Crick pairs and 15 span the G:U wobble pair, including two interesting arrangements with three hydrogen bonds to the Arg guanidinium group that have not yet been observed. The inherent donor–acceptor arrangements of the bases support many possible interactions to Asn (or Gln) and Ser (or Thr or Tyr), few interactions to Asp (or Glu) even though several already have been observed, and interactions to U (or T) only if the base is in an unpaired context, as also observed in several cases. This study highlights how complementary arrangements of donors and acceptors can contribute to base-specific recognition of RNA, predicts interactions not yet observed, and provides tools to analyze proposed contacts or design novel interactions.

© 2003 Elsevier Science Ltd. All rights reserved

Keywords: RNA structure; non-Watson–Crick base-pairs; DNA–protein interactions; RNA–protein interactions

*Corresponding author

Introduction

The ability of proteins to recognize specific RNA sites is important in many biological systems but the “rules” governing these interactions are not well understood. This is in part because the structural database of protein–RNA complexes is still relatively limited (despite the recent addition of the ribosomal subunit structures) and, perhaps more importantly, because RNA structures are so diverse. In addition to Watson–Crick helices, RNAs often contain non-Watson–Crick base-pairs, unpaired bases such as those in bulges and loops, and base-triples or other higher-order tertiary

interactions.^{1,2} Therefore many of the principles of protein–DNA recognition inferred from the large number of solved structures^{3,4} may not apply to RNA.

Despite the current gaps in knowledge, it is apparent that one important determinant of specificity in both DNA and RNA complexes is the complementary nature of hydrogen bonding interactions between polar groups on the protein side-chains and nucleic acid bases. A seminal study by Seeman *et al.* conducted before the structure of even a single protein–nucleic acid complex had been solved⁵ systematically examined the possible hydrogen bonding interactions between amino acid side-chains and groups along the edges of the Watson–Crick base-pairs. They concluded that interactions involving two hydrogen bonds would be required to uniquely distinguish each base-pair from the others, and inferred that two hydrogen bonds from a single functional group would

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

specify a site with higher precision than from two independent groups, analogous to the “chelate effect” in which formation of one bond favors formation of additional bonds by an increase in effective concentration.⁶ Based on their analysis, Seeman *et al.* predicted two interactions in which precisely positioned side-chains in the DNA major groove could discriminate amongst all the base-pairs: one in which the guanidinium group of Arg donates two hydrogen bonds to the O6 and N7 acceptor groups of guanine and a second in which the carboxamide group of Asn (or Gln) hydrogen bonds to the N7 acceptor and N6 donor groups of adenine. These interactions are indeed the most commonly observed in protein–DNA complexes,^{4,7–9} and the importance of such direct amino acid–base hydrogen bonds in determining sequence specificity has been confirmed by many structure–function studies.

Several detailed studies have analyzed the interactions observed in protein–DNA complexes, partly in efforts to determine whether a “recognition code” exists for DNA double helices.^{3,4,8–14} It seems clear that while no simple code exists, some common interaction patterns between amino acids and bases can be found, particularly within a given structural context such as the zinc finger or helix–turn–helix motif.^{3,11,15} Hydrogen bonding interactions to the bases comprise about two-thirds of all base-specific contacts,⁴ and interactions involving two hydrogen bonds from the side-chain are dominated by the Arg–G and Asn(Gln)–A interactions described above. The only other frequent bidentate interaction utilizes the amino group of Lys to hydrogen bond to the O6 and N7 acceptors of guanine, although other two hydrogen-bond interactions are found that utilize bifurcated bonds or donors and acceptors from the peptide backbone.⁴ A statistical survey of 28 protein–DNA complexes found that side-chains possessing both donor and acceptor atoms more frequently use the donor atom for hydrogen bonding.⁸ In addition to direct amino acid–base hydrogen bonds, it is clear that other types of interactions contribute to DNA recognition, including water-mediated hydrogen bonds, van der Waals contacts, and interactions to the sugar–phosphate backbone, and that the structural context in which the interactions are presented is an inherent part of the recognition process.^{3,4,10,13,14}

The analysis of protein–RNA interactions is at an earlier stage and the available structures represent only a small subset of possible RNA tertiary elements, but some characteristics are beginning to emerge.^{2,16–19} With respect to hydrogen bonding, perhaps the most obvious difference between RNA and DNA complexes is the use of the ribose 2'OH group in about a quarter of all hydrogen bonds.¹⁹ In addition, of all the observed base-specific interactions, hydrogen bonds appear less dominant than in DNA complexes,^{17–19} probably because a significant number of bases are not stacked within Watson–Crick duplexes

and consequently some bases are sequestered from solvent *via* van der Waals interactions with the protein. Nevertheless, the importance of hydrogen bonding for RNA-binding specificity is as apparent for RNA as it is for DNA. It is inherently more difficult to evaluate the possible amino acid hydrogen-bonding interactions to RNA than to DNA⁵ given the large number of possible RNA base configurations, and therefore a systematic computational approach is required. Here, we report the calculation of databases of hydrogen-bonding interactions between amino acids and bases or base-pairs that can occur in RNA structures. The databases include interactions between unpaired bases, such as those found in bulges or loops, and non-Watson–Crick base-pairs, some of which involve multiple hydrogen bonds that may be used to uniquely recognize bases in particular structural contexts. The databases may be useful not only for analyzing existing interactions but also for designing novel interactions in RNA-binding proteins or peptides, in combination with other experimental approaches.

Computational Approach

Database construction

The approach for generating hydrogen-bonding arrangements of amino acid side-chains with bases or base-pairs utilizes simple geometric and steric criteria and is illustrated in [Figure 1A](#). The program WASABI (what are the specific amino acid–base interactions) first forms a single linear hydrogen bond between a side-chain and base for every possible combination of donor and acceptor groups, and then samples the allowable three-dimensional conformations by rotating the side-chain around each of five angles (pivoting around the donor or acceptor heavy atoms) while still maintaining the initial bond. One rotation is symmetric along the axis of the hydrogen bond and would be redundant for the donor and acceptor sites. Each conformation is evaluated for the formation of additional hydrogen bonds (using the parameters shown in [Figure 1B](#)) and the absence of steric clashes. The “best” conformation, as judged by a scoring function that favors linear hydrogen bonds (see below), is identified and all unique hydrogen-bonding arrangements are stored, including those with single hydrogen bonds. This three-dimensional search algorithm is an adaptation of a two-dimensional version used to systematically generate base–base combinations.^{20,21} One limitation to the WASABI algorithm is that the length of the initial hydrogen bond remains fixed during the conformational search, but its length subsequently may be varied to generate multiple conformations of any hydrogen-bonding arrangement (see below).

Conformational searches were performed utilizing the nine hydrogen-bonding side-chain moieties

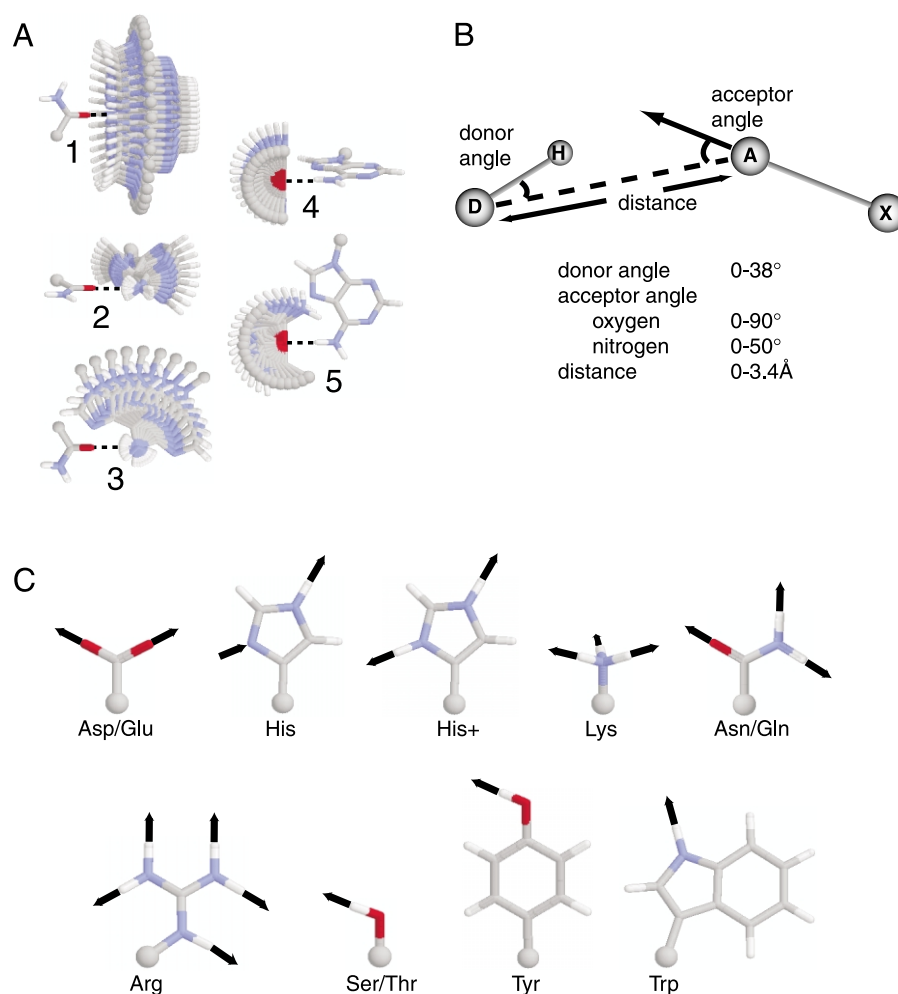


Figure 1. A, Schematic of the WASABI search method. Rotations 1–3 illustrate rotations about the donor atom and rotations 4 and 5 illustrate rotations about the acceptor atom, conceptually similar to first holding the amino acid fixed and rotating the base (1–3) and then holding the base fixed and rotating the amino acid (4, 5). B, The three parameters used to define a hydrogen bond: the acceptor angle, the donor angle, and distance between heavy atoms, with parameters listed for the different atom types. C, The nine hydrogen-bonding moieties of the amino acid side-chains, with arrows indicating donor and acceptor positions.

shown in Figure 1C, including unprotonated and protonated forms of histidine, and either with the six RNA bases (A, C, G, U, A⁺, C⁺) or 53 possible RNA base-pairs generated by Walberer *et al.*²¹ In addition, we constructed interactions with the additional DNA base, thymine, and the 17 possible base-pairs that utilize thymine.²⁰ These types of amino acid–DNA interactions may occur in the context of single-stranded sites or in helices with extruded bases.

For our steric parameters, van der Waals radii taken from AMBER²² were divided by $2^{1/6}$ to approximate hard sphere radii;²³ these radii were further reduced by a factor of 0.8 to include geometries slightly outside a reasonable steric range. Polar hydrogen atoms were assigned van der Waals radii of 0.2 Å and also were further reduced by the 0.8 steric parameter. Because the polar hydrogen atoms had such small radii, we implemented a filter that removed conformations in which two polar hydrogen atoms were closer

than 2.5 Å, approximately the distance between two oxygen atoms, thereby eliminating unfavorable arrangements with two nearby positive charges. Amino acid side-chains were constructed using the LEAP package with param96 residue definitions.²²

The parameters used to define a hydrogen bond (Figure 1B) were chosen based on the analysis of small molecule high-resolution crystal structures^{24,25} and on an empirical test of the donor angle parameter. A maximum distance of 3.4 Å was used for all hydrogen bonds. We estimated acceptor and donor parameters from the crystal structure analysis, which reported only acceptor–hydrogen–donor atom angles, by assuming a hydrogen–acceptor length of 2.0 Å and a donor–hydrogen bond length of 1.0 and using $\text{donor angle} = \sin^{-1}(D)$, where D is as described previously.²⁵ For the nitrogen acceptor angle, values of $0(\pm 22)^\circ$ for two-center hydrogen bonds and $0(\pm 45)^\circ$ for three-center hydrogen bonds were

Table 1. Numbers of computed amino acid–base and amino acid–base pair interactions

	Single bases	Base- pairs
WASABI output	470	7730
Remove backbone incompatible	470	7718
Remove U/T redundancies	426	5819
Remove Asn/Gln, Arg, His ⁺ redundancies	385	5076
Remove Tyr redundancies	344	4612
Remove bifurcated interactions	261	3519
Remove single H-bond interactions	36	457
Remove A ⁺ , C ⁺ redundancies	32	423
Remove non-spanning base-pair interactions	N/A	186

The raw output from WASABI was filtered to remove sterically restricted conformations, calculational and structural redundancies, and bifurcated and single hydrogen-bonded interactions, as described in the text. Bifurcated interactions refer only to those in which a donor or acceptor atom on the amino acid is simultaneously involved in two hydrogen bonds to a base and does not include those in which a bifurcated bond exists between two bases of a base-pair. The narrow donor angle parameter used to construct the base-pairs limits the number of bifurcated bonds.²¹

three standard deviations from the mean and therefore included virtually all observed hydrogen bonds. Thus, we used a nitrogen acceptor angle of $0(\pm 50)^\circ$ to include slightly unreasonable geometries and to allow us to evenly sample conformations using a 4° step size. We used an oxygen acceptor angle of $0(\pm 90)^\circ$ for similar reasons. To determine an appropriate cutoff for the donor angle parameter, we performed a set of WASABI calculations using angles from $0(\pm 30)^\circ$ to $0(\pm 40)^\circ$ and found that $0(\pm 36)^\circ$ generated all known interactions (see below) and that at least some of the additional conformations generated using a $0(\pm 38)^\circ$ angle appeared reasonable by inspection. A similar empirical approach was used to select the $0(\pm 18)^\circ$ donor angle parameter used to construct the base–base interaction database, which was substantially more restrictive due to the planar nature of the conformational search.²¹

As mentioned above, WASABI generates multiple conformations with the same hydrogen bonding arrangement and thus we devised an empirical scoring function in order to select a representative conformation with as planar an arrangement as possible. Scores (S) were calculated over all hydrogen bonds as follows:

$$S = \sum \{w_1(1 + [A/d]^4 - [B/d]^2) + w_2 \sin^2 \theta_d + w_3 \sin^2 \theta_a + w_4 \sin^2 \theta_p\},$$

where $w_1:w_2:w_3:w_4 = 100:30:1:1$ for oxygen acceptors and $100:30:10:0$ for nitrogen acceptors, θ_a is the acceptor angle, θ_d is the donor angle, θ_p is the angle between the plane of the side-chain and the plane of the base or base-pair, d is the distance between the heavy atom donor and acceptor, and

A and B are parameterized to mean hydrogen bond distances of 2.95 Å for N–O bonds, 2.73 Å for O–O bonds, and 2.90 Å for N–N bonds, which are average distances calculated from the database of known protein–nucleic acid complexes and similar to those previously reported.^{24–28}

Finally, we wished to ensure that each calculated arrangement could accommodate a nucleotide backbone and complete amino acid side-chain. We added C2'-endo or C3'-endo ribose sugars (generated using AMBER parameters) to each base in a combinatorial manner, rotating the sugars by 360° around the C1'–N1 bond in 2° increments. We similarly added all amino acid rotamers from Dunbrack & Cohen²⁹ (August 10, 1999 release) in a combinatorial manner and identified any model in which no set of sugar and rotamer conformations could be accommodated sterically. These models were analyzed further using DIVERSIGEN as described below. Although we used one hydrogen bond moiety to represent Asn(Gln), Asp(Glu), and Ser(Thr) side-chains (Figure 1C), rotamers of all represented amino acids also were added for these final steric tests. Interestingly, two interactions involving bifurcated hydrogen bonds with Ser and Thr were found to be sterically impossible but could occur with Tyr. Despite the larger size of the Tyr side-chain, the planarity of the aromatic ring makes the interaction more favorable than with the Ser or Thr side-chains (see Results).

Diversity generator

Each hydrogen-bonding arrangement is represented in our databases by a single conformation, however many three-dimensional conformations typically are possible for each arrangement. We constructed a diversity generating program, DIVERSIGEN, that begins with one conformation and creates a set of conformations chosen to represent the sterically accessible space for the particular hydrogen-bonding arrangement. For arrangements that could not accommodate the nucleotide sugars and side-chain rotamers (see above), we generated ten representative conformations and tested each for its ability to accommodate the sugars and rotamers. Those few arrangements that could not (see Table 1) were eliminated from the databases.

To generate conformational diversity, first the length of each hydrogen bond in a particular arrangement is set to three values, corresponding to short, median, and long distances that cover the experimentally observed range for each type of donor–acceptor pair. Next, the same five angles varied in the WASABI search again are incrementally varied, beginning with a large step size, and hydrogen bond distances and angles are monitored and steric tests performed using the parameters described above to retain plausible conformations with the appropriate hydrogen bonds. Conformations generated using each of the starting hydrogen bond lengths are retained, until a total of $\sim 10,000$ are generated (or 1000–4000 in a few

Table 2. Nucleic acid–protein complexes from the PDB utilized in this study

DNA complexes (NMR)	185d, 193d, 1a66, 1a6b, 1ahd, 1b69, 1bbx, 1bj6, 1c7u, 1c9g, 1dsc, 1dsd, 1e7j, 1f4s, 1f5e, 1fja, 1g4d, 1gcc, 1hry, 1hrz, 1ig4, 1iv6, 1j46, 1j47, 2j4w, 1j5k, 1kqq, 1l1m, 1l1v, 1lcc, 1lcd, 1mse, 1msf, 1nk2, 1nk3, 1rcs, 1tf3, 1tn9, 1yui, 1yuj, 2da8, 2ezd, 2eze, 2ezf, 2ezg, 2gat, 2hdc, 2lef, 2stt, 2stw, 3gat, 4gat, 5gat, 6gat, 7gat
DNA complexes (crystal)	1a02, 1a0a, 1a1f, 1a1g, 1a1h, 1a1i, 1a1j, 1a1k, 1a1l, 1a3q, 1a6y, 1a73, 1a74, 1aay, 1ais, 1akh, 1am9, 1an2, 1an4, 1aoi, 1apl, 1au7, 1awc, 1az0, 1azp, 1azq, 1b01, 1b3t, 1b72, 1b8l, 1b97, 1bc7, 1bc8, 1bdh, 1bdi, 1bdt, 1bdv, 1ber, 1bf4, 1bf5, 1bg1, 1bgb, 1bhm, 1bl0, 1bnk, 1bnz, 1bp7, 1bpx, 1bpy, 1bpz, 1bsu, 1bua, 1bvo, 1c0w, 1hlo, 1c7y, 1c8c, 1c9b, 1c9r, 1ca5, 1ca6, 1cbv, 1cdw, 1cey, 1cf6, 1cgp, 1cit, 1ckq, 1cl8, 1clq, 1cma, 1cqt, 1crx, 1cw0, 1cyq, 1cz0, 1d02, 1d0e, 1d1u, 1d2l, 1d3u, 1d5y, 1d66, 1db7, 1db8, 1db9, 1dbc, 1dc1, 1dct, 1ddn, 1dfm, 1dgc, 1dh3, 1diz, 1dnk, 1dp7, 1drg, 1dsz, 1du0, 1ea4, 1ecr, 1ej9, 1eqz, 1eri, 1evw, 1ewn, 1ewq, 1eyu, 1f0o, 1f3l, 1f44, 1f4k, 1f4r, 1f5t, 1f66, 1f6o, 1fiu, 1fjl, 1fjx, 1fn7, 1flo, 1fok, 1fos, 1fw6, 1g2d, 1g2f, 1g38, 1g9y, 1g9z, 1gdt, 1glu, 1gxp, 1h88, 1h89, 1h8a, 1h9d, 1hao, 1hap, 1hbx, 1hcq, 1hcr, 1hdd, 1hlv, 1huo, 1hut, 1huz, 1hw2, 1hwt, 1i6j, 1i8m, 1iaw, 1id3, 1if1, 1ig7, 1ig9, 1ign, 1ihf, 1ijs, 1imh, 1ipp, 1ijw, 1j59, 1j5o, 1jb7, 1jey, 1jfi, 1jfs, 1jft, 1jgg, 1jh9, 1jj6, 1jj8, 1jko, 1jqp, 1jqk, 1jkr, 1jmc, 1jto, 1k6o, 1k8g, 1kix, 1ksx, 1ksy, 1l1a, 1l2c, 1l2d, 1l3l, 1lat, 1lau, 1le8, 1lli, 1lmb, 1mdy, 1mey, 1mhd, 1mht, 1mj2, 1mjm, 1mjo, 1mjp, 1mqj, 1mmn, 1mvm, 1nfx, 1noy, 1oct, 1otc, 1par, 1pdn, 1per, 1pnr, 1pue, 1pvi, 1pyi, 1qai, 1qaj, 1qbj, 1qln, 1qp0, 1qp4, 1qp7, 1qp9, 1qpi, 1qps, 1qpz, 1qqa, 1qqb, 1qrh, 1qri, 1qrv, 1qsl, 1qum, 1ram, 1rbj, 1rcn, 1rep, 1rtd, 1run, 1ruo, 1rv5, 1rva, 1rvb, 1rvs, 1skn, 1srs, 1ssp, 1svc, 1t7p, 1tau, 1tc3, 1tf6, 1tgh, 1tro, 1trr, 1tsr, 1tup, 1uaa, 1ubd, 1vas, 1vkv, 1vol, 1vpw, 1wet, 1xbr, 1xrn, 1yfb, 1yft, 1zaa, 1zay, 1zme, 2bam, 2bpa, 2cgp, 2crx, 2dgc, 2drp, 2gli, 2hap, 2hdd, 2hmi, 2irf, 2kfn, 2kfz, 2kzm, 2kzz, 2nll, 2or1, 2pjr, 2pua, 2pub, 2puc, 2pud, 2pue, 2puf, 2pug, 2pvi, 2ram, 2rve, 2ssp, 2up1, 3bam, 3cro, 3crx, 3hdd, 3hts, 3mht, 3orc, 3pjr, 3pvi, 4crx, 4dpv, 4mht, 4rve, 4skn, 5crx, 5mht, 6cro, 6mht, 6pax, 7mht, 8mht, 9ant, 9mht
RNA complexes (NMR)	1alt, 1a4t, 1aju, 1akx, 1arj, 1aud, 1biv, 1ck5, 1ck8, 1cn8, 1cn9, 1d6k, 1dz5, 1ekz, 1etf, 1etg, 1exy, 1f6u, 1fje, 1g70, 1hji, 1i9f, 1k1g, 1l1c, 1koc, 1mnb, 1qfq, 1ull, 484d
RNA complexes (crystal)	1a34, 1a9n, 1aq3, 1aq4, 1asy, 1asz, 1av6, 1b23, 1b7f, 1bmj, 1c0a, 1c9s, 1cvj, 1cwp, 1cx0, 1d9f, 1dfu, 1di2, 1dk1, 1drz, 1dul, 1dzs, 1e6t, 1e7k, 1e7x, 1e8o, 1ec6, 1efw, 1eiy, 1ekc, 1euy, 1euy, 1exd, 1f7u, 1f7v, 1f7y, 1f8v, 1feu, 1ffy, 1ffj, 1fjg, 1g1x, 1g2e, 1g59, 1gax, 1gkv, 1gkw, 1gtf, 1gtn, 1gtr, 1gts, 1h4q, 1h4s, 1h8j, 1hc8, 1hdw, 1he0, 1he6, 1hp6, 1hq1, 1hys, 1i6h, 1i6u, 1il2, 1j5a, 1jbr, 1jbs, 1jbt, 1jid, 1jj2, 1jzx, 1jzy, 1k01, 1k8w, 1knz, 1kog, 1kq2, 1kuo, 1l9a, 1lng, 1lnr, 1mms, 1qa6, 1qf6, 1qrs, 1qrt, 1qru, 1qtq, 1qu2, 1qu3, 1ser, 1ttt, 1urn, 1zdh, 1zdi, 1zdi, 1zdk, 2a8v, 2bbv, 2fmt, 5msf, 6msf, 7msf

particularly sterically restricted cases). The step size used for each of the angles varied is adjusted iteratively to achieve the desired $\sim 10,000$ conformations. These conformations then are clustered into the desired number of representatives (10 to $\sim 10,000$), chosen to cover conformational space as completely as possible. It is particularly difficult to achieve a good representation when choosing a small number of conformations to represent a broad space. To assess whether a chosen set of conformations reasonably represents the space sampled, we define similarity of any two conformations as the Euclidean distance between the five parameters of the WASABI search. For conformations **a** and **b**, with parameter coordinates $\{a_1, a_2 \dots a_5\}$, $\{b_1, b_2 \dots b_5\}$, Euclidean distance (d_E) is defined by $d_E = \|\mathbf{a} - \mathbf{b}\|$. Thus, we are defining conformational similarity based on hydrogen bonding parameters and not on the r.m.s.d. of three-dimensional coordinates. The clustering routine produces conformations within each hydrogen bond class (short, median, long), with their number proportional to the number of conformations found for each in the WASABI search.

Database search of observed interactions

We identified all hydrogen bonding interactions between amino acids and bases or base-pairs in protein–DNA and protein–RNA complexes in the PDB (July 15, 2002 release; see Table 2). For this search, we slightly relaxed the hydrogen bonding

parameters, using a donor angle of $0(\pm 40)^\circ$ and a maximum distance of 3.5 Å to ensure that no plausible interactions would be missed. Of the 433 protein–DNA complexes examined, 378 were crystal structures, 22 were averaged NMR structures, and 33 were NMR ensembles. Of the 132 protein–RNA complexes examined, 103 were crystal structures, seven were averaged, NMR structures, and 22 were NMR ensembles. Only crystal structures with < 3.5 Å resolution were used. Hydrogen atoms were added using InsightII (Biosym) or AMBER PROTONATE, and the search program automatically calculated optimized placement of hydrogen atoms where rotatable hydroxyl and amine groups are involved. Optimized hydrogen positions were determined by rotating the hydrogen into the plane formed by points D, A, and X (Figure 1B) for each potential hydrogen bond. Polymerases and topoisomerases were considered non-specific binders and were not examined, and for crystal structures with multiple complexes in a unit cell, only one representative was included. For the 30 S and 50 S ribosomal structures, one representative structure for each was chosen (1fjf and 1jj2, respectively), and all others removed. We made no other attempts to remove other possible sources of redundancy, including similar structures solved by more than one group, similar structures reported at different levels of resolution or refinement, or mutant structures (one interaction was observed only in a mutant; see Results and Discussion). For the NMR

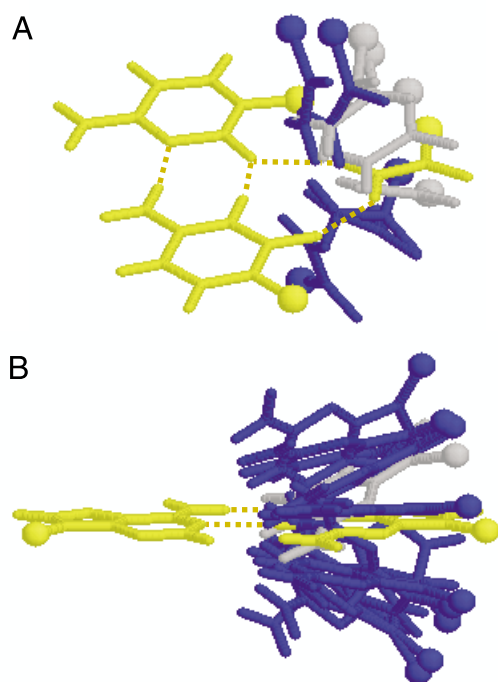


Figure 2. Generation of conformational diversity by DIVERSIGEN. A, Ten representative conformations of an Asn-C:C⁺ base-pair interaction. The starting model calculated by WASABI, shown in yellow, cannot accommodate the sugar backbone and full amino acid side-chain, the sterically allowed conformations are shown in blue, and clashing models are shown in gray. B, Ten conformations of the G:A base-pair found in the RRE,^{37,38} beginning with the initial planar conformation shown in yellow. The blue conformations represent those in which Asn can form spanning interactions, as described in the text, whereas the starting yellow conformation and gray conformations cannot.

structures, we scored the presence of an interaction if it was observed in the averaged structure or any member of an ensemble. Our goal for this study was to gather all the observed interactions rather than to compile precise statistics.

Results

Database construction

To better understand the ways in which RNA sites might be recognized by proteins in a base-specific manner, we calculated extensive databases of possible hydrogen bonding interactions between amino acid side-chains and either the six unpaired bases (A, C, G, U, A⁺, and C⁺) or the 53 possible RNA base-pairs in planar conformations.²¹ Although our focus is primarily on RNA, we also constructed databases that include thymine or the 17 possible thymine-containing base-pairs that might be found in single-stranded DNA structures. A simple geometric algorithm (WASABI; Figure 1A) was utilized in which a single hydrogen bond was

first formed between a hydrogen-bonding donor or acceptor of a side-chain moiety (Figure 1C) and a complementary group on a base, followed by a systematic conformational search that identified additional possible hydrogen bonds in sterically plausible configurations. Each of five hydrogen bond angles (Figure 1B) was varied in 4° steps such that no other donor or acceptor on any of the amino acid side-chains would move by more than 0.6 Å. The hydrogen bonding and steric parameters used were slightly beyond what would be considered energetically favorable to help generate thorough databases. A single conformation was chosen to represent each unique hydrogen-bonding arrangement using a scoring function that attempted to maintain relatively planar geometries when possible (see Computational Approach). The databases contain all possible amino acid–base and amino acid–base-pair arrangements with one or more hydrogen bond (Table 1), but we focus primarily on interactions containing two or more bonds that have defined orientations and may contribute to high binding specificity.

In addition to the simple steric criteria applied by WASABI to the side-chains and bases, we also wished to ensure that each hydrogen-bonding arrangement could accommodate the nucleic acid backbone and at least one reasonable conformation of a full amino acid side-chain. We added C2' and C3'-endo conformations of the ribose sugar and all amino acid rotamers²⁹ in a combinatorial manner to all arrangements and found that every interaction to the unpaired bases was acceptable, whereas 170 interactions to the base-pairs were sterically restricted. Given that the WASABI search utilized only planar conformations of the base-pairs²¹ and output only a single conformation for each hydrogen-bonding arrangement, we wished to generate a larger set of plausible conformations for any individual arrangement and to re-examine their abilities to accommodate the backbone moieties. We constructed the DIVERSIGEN algorithm, which generates a specified number of output conformations for a single hydrogen-bonding arrangement (see Computational Approach; Figure 2), and found that only 12 arrangements to the base-pairs remained sterically impossible when multiple conformations were tested (Table 1). DIVERSIGEN may be used to generate multiple conformations of amino acid interactions with bases or base-pairs (Figure 2A) or of interactions between bases (Figure 2B).

The modeled interactions were constructed using the hydrogen-bonding moieties shown in Figure 1C, assuming that interactions involving the carboxyl groups of Asp and Glu, the carboxamide groups of Asn and Gln, and the hydroxyl groups of Ser and Thr would be redundant. Indeed, we found that all arrangements could accommodate the extra lengths of the Glu, Gln, and Thr side-chains. Initially we considered the hydroxyl-containing moiety of Tyr as separate from Ser(Thr), but subsequently found that all

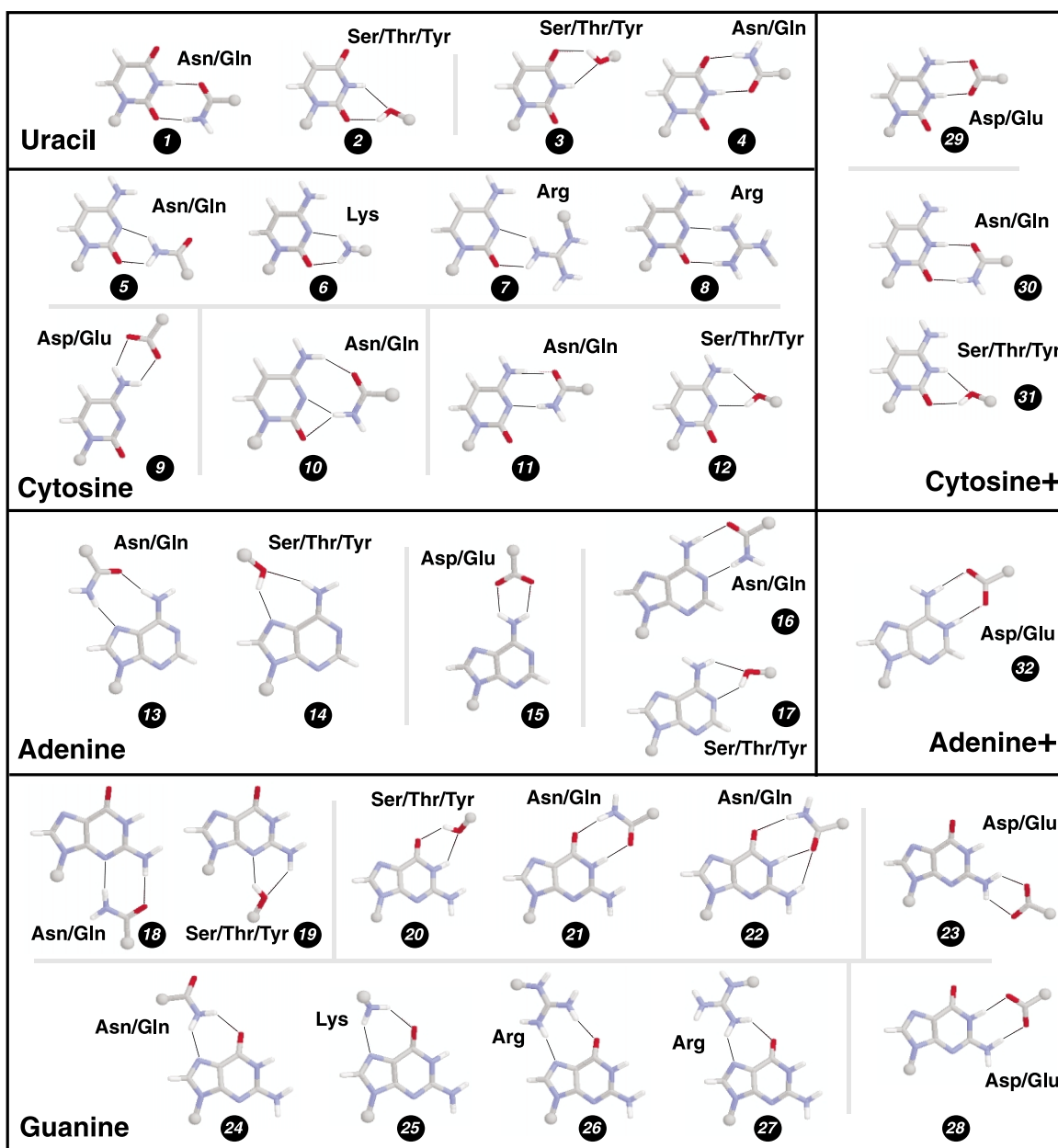


Figure 3. Amino acid–base interactions with two or more hydrogen bonds. Interactions are grouped by base type, and are subgrouped by the interacting face. Bifurcated versions of two interactions are shown (#10, #22) that probably are more stable than the related two hydrogen-bonded versions also present in the database. In these cases, the additional bifurcated interaction is absent only when the side-chain is placed considerably out of the plane of the base. Five interactions (#5, #7, #9, #15, #23) are observed only at the edge of our parameter range (requiring a donor angle of 38°) and may not be energetically favorable.

Ser(Thr) interactions could be represented by Tyr interactions despite the extra bulk of the Tyr ring. Interestingly, two Tyr base-pair arrangements cannot occur with Ser, both involving a bifurcated hydrogen bond to one base that is located close to the sugar of the second base, illustrating that steric clashes can occur with shorter side-chains that position peptide backbone atoms closer to the bases. All interactions with thymine were possible with uracil. Thus, the databases were appropriately filtered for all types of redundant interactions, including those with His⁺ and pseudo-

symmetric arrangements with Asn(Gln) and Arg moieties that produce structural redundancies (Table 1).

Amino acid–base interactions

There are 344 unique ways for the amino acid hydrogen-bonding moieties to interact with the bases, including A⁺ and C⁺ (Table 1). Of these, 225 utilize single hydrogen bonds, representing all possible acceptor–donor combinations. Some of the interactions, even with one hydrogen bond,

Table 3. Observed amino acid–base interactions

Interaction	Face	Number (DNA)	Number (RNA)	Observed interactions ^a
Ser/Thr/Tyr-U (#3)	WC	1	1	DNA: reverse transcriptase (1d0e) Ser67; RNA: Sxl (1b7f) Tyr164
Asn/Gln-U (#4)	WC	0	6	RNA: AspRS (1asy, 1asz) Gln138, (1c0a) Gln46, (1efw) Gln47, (1il2) Gln1046; GlnRS (1euq) Gln517
Lys-C (#6)	WC	1	1	DNA: nucleocapsid (1bj6) Lys34; RNA: nucleolin (1fje) Lys94
Arg-C (#8)	WC	1	4	DNA: oxoG glycosylase (1fn7) Arg 204 RNA: S15/S16/S18 (1ekc,1g1x) Arg74; AspRS (1il2) Arg225; GluRS (1g59) Arg358
Ser/Thr/Lys-C (#12)	WC	0	1	RNA: AspRS (1asz) Ser329
Asn/Gln-A (#13)	Major	83	1	RNA: 50S ribosome (1jj2) Asn44
Ser/Thr/Tyr-A (#14)	Major	11	7	RNA: U1A (1aud) Tyr12, (1dz5) Ser45, Thr88, Tyr12; MS2 coat (5msf, 6msf, 7msf) Thr45
Asn/Gln-A (#16)	WC	3	2	DNA: RNaseB (1rbj) Gln69, Asn71; methyltransferase (1g38) Asn105 RNA: U4 (1e7k) Ser96; Ribozyme (1hp6) Ser91
Ser/Thr/Tyr-A (#17)	Major	0	1	RNA: U2B''/A' (1a9n) Ser91
Asn/Gln-G (#18)	Minor	7	3	DNA: telomere BP (1otc) Gln135
Ser/Thr/Tyr-G (#19)	Minor	2	1	
Lys-G (#25)	Major	20	6	RNA: L30 (1ck8, 1cn9) Lys28; 50S (1jj2) Lys35; ProRS (1h4q, 1h4s) Lys369
Arg-G (#26)	Major	164	19	DNA: telomere BP (1otc) Arg274; RNA: AspRS (1c0a, 1il2) Arg222; Rev (1ull) Arg6, (484d) Arg41; Nucleolin (1fje) Arg49
Arg-G (#27)	Major	3	4	RNA: Nucleolin (1fje) Arg49
Asp/Glu-G (#28)	WC	10	15	DNA: telomere BP (1otc) Asp225, Glu45; (1jb7) Asp223, Asp225, Glu45, (1k8g, 1kix) Asp223, Asp25; UP1 (2up1) Asp42; RNA: TRAP (1c9s, 1gtf, 1gtm) Asp39, Glu36; AspRS (1il2) Glu93; ThrRS (1qf6, 1kog) Glu600; ProRS (1h4q, 1h4s) Asp354, (1h4s) Glu340; 50S (1jj2) Asp105, Glu71
Asp/Glu-C ⁺ (#29)	WC	3	0	DNA: HaeIII (1dct) Glu109; HhaI (1mht, 4mht) Glu119
Ser/Thr/Tyr-C ⁺ (#31)	WC	0	2	RNA: U1A (1aud, 1dz5) Tyr12

Numbers refer to the interactions shown in Figure 4. Face refers to the interacting surface of the base (Watson–Crick, major groove, minor groove) in a Watson–Crick helix. The observed cases in DNA and RNA are indicated, with details (protein name, pdb identifier, residue) provided for bases that are not in a Watson–Crick pair.






















































^a Interactions in the major or minor grooves are listed only if they are found in unpaired or non-Watson–Crick pairing contexts.

require that the amino acid be roughly perpendicular to the plane of the base to avoid steric clashes, particularly for His interactions with adenine N3. To simplify analysis of the database, we removed bifurcated hydrogen-bonding interactions and those involving the protonated bases (A⁺, C⁺) that do not form hydrogen bonds to the extra proton (Table 1). Thus, there are 32 unique interactions involving two or more hydrogen bonds between amino acid side-chains and the unpaired bases (shown in Figure 3). For two of the Asn(Gln) interactions, we present arrangements that include bifurcated bonds that probably are more stable than the non-bifurcated versions in the database. Of the 32 possible interactions, 12 involve Asn(Gln) and eight involve Ser(Thr/Tyr). Both types of side-chains show potential interactions to all bases except A⁺, and their dominance likely reflects the high frequency of adjacent acceptor and donor groups on the bases, as also described for base–base interactions.²¹ Asp(Glu), with two acceptors, shows five interactions, including one with A⁺ not possible with the unprotonated base, and none with U. Arg, with five hydrogen donors

on its guanidinium group, allows only four interactions, all with C and G.

To help evaluate the completeness of our database and to determine whether any rules might be inferred from known interactions, we identified amino acid–base hydrogen bonds in protein–nucleic acid complexes in the PDB (Table 2), using slightly relaxed hydrogen bond parameters (see Computational Approach) to help ensure that no plausible interactions would be missed. All observed interactions are found in our database, including 17 of the 32 possible two-hydrogen bonded arrangements (Table 3; Figure 3). There are 12 types of interactions in DNA complexes, including five in the major groove and two in the minor groove of Watson–Crick helices, with the Arg–G and Asn(Gln)–A interactions (#26 and #13, Figure 3) predicted by Seeman *et al.*⁵ dominating, as previously observed.^{4,7–9} Only six types of DNA interactions are found in which amino acids form two hydrogen bonds to a Watson–Crick face. A Ser–U interaction is observed in a reverse transcriptase complex (#3, Figure 3),³⁰ a Lys–C interaction (#6, Figure 3) is observed in a nucleocapsid-single-stranded DNA complex,³¹ an Arg–C

Table 4. Calculated amino acid–base-pair spanning interactions

Pur-Pur								
	Total	D/E	H	H+	K	N/Q	R	S/T/Y
AA 1 								
AA 15 	1					1		
AA 16 								
AA+ (1) 								
AA+ (16) 								
A+A+ (16) 								
GA 21 								
GA 22 	2					2		
GA 23 	2					2		
GA 24 								
GA+ (21) 								
GA+ (23) 								
GA+ 38 	1					1		
GA+ 39 								
GG 17 	2					1	1	
GG 18 	13		2			8	1	2
GG 19 								
GG 20 	10		1		1	5	2	1
Total	31	0	3	0	1	20	3	4
Pyr-Pyr								
	Total	D/E	H	H+	K	N/Q	R	S/T/Y
CC 25 	4					2		2
CC+ 40 	5	1			1	1	2	
CC+ 41 	2					2		
C+C+ 42 	5					2	1	2
[UC 29 	5				1	2	2	
[UC 30 	5				1	2	2	
[UC+ 43 	7				1	2	3	1
[UC+ 44 	7				1	2	3	1
[UU 26 	9				2	2	5	
[UU 27 	9				2	2	5	
[UU 28 	9				2	2	5	
Total	67	1	0	0	11	21	28	6
Pur-Pyr								
	Total	D/E	H	H+	K	N/Q	R	S/T/Y
AC 12 	2					1		1
AC 13 	5			1		1	2	1
AC+ 35 	1					1		
AC+ 37 	2					2		
A+C (12) 	2					2		
A+C 34 	3	1	1			1		
A+C 36 	2					2		
A+C+ (35) 								
[AU 1 	2					1	1	
[AU 3 	4					1	3	
[AU 2 	1					1		
[AU 4 	1					1		
[A+U (2) 								
[A+U (4) 								
GC 5 	2					2		
GC 6 	5	1			1	1	2	
GC 7 	4					3		1
GC+ 31 	9		1			5	1	2
GC+ 32 	8		1		1	4	2	
GC+ 33 	1					1		
[GU 8 	16			1	1	2	11	1
[GU 10 	16			1	1	2	11	1
[GU 9 	1					1		
[GU 11 	1					1		
Total	88	2	3	3	4	36	33	7

Base-pairs are listed according to the numbering used by Walberer *et al.*²¹ For cases in which the protonated base atom did not form an additional hydrogen bond in the pair, the number of the corresponding unprotonated pair is indicated in parentheses. The A⁺:C pair (12) is the only case with interactions not observed with the unprotonated partner. The arrangements marked with brackets indicate base arrangements in which a U is flipped, presenting essentially an identical donor and acceptor arrangement. The interactions observed with these related pairs are identical in all but one case, where steric restrictions differ.

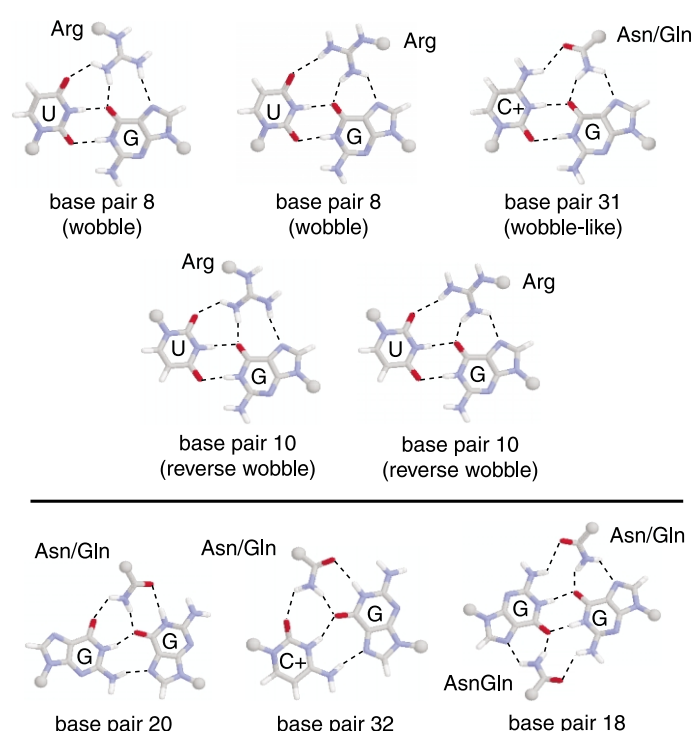


Figure 4. Spanning interactions utilizing three hydrogen bonds. The base-pair numbering is that used by Walberer *et al.*²¹ The top set shows interactions with the G:U wobble, related G:C⁺, and reverse wobble pairs, and the bottom set shows three interactions of Asn(Gln). Two symmetric Asn(Gln) interactions are shown to the symmetric G:G base-pair 18, which in principle might occur simultaneously.

interaction (#8, Figure 3) is observed in a base excision DNA repair complex,⁶⁷ Asn (or Gln)–A interactions (#16, Figure 3) are observed in RNaseB–DNA and methyltransferase–DNA complexes,^{32,68} Asp–G interactions (#28, Figure 3) are observed in two telomere-binding protein complexes,^{33,34} and Asp (or Glu)–C⁺ interactions (#29, Figure 3) are observed in *HhaI* and *HaeIII* methylase complexes in which cytosine bases are extruded from the DNA helix.^{35,36}

Despite the relatively small database of protein–RNA complexes, the diversity of amino acid–base interactions already seems apparent. There are 16 types of interactions with RNA bases, including eight in which amino acids form two hydrogen bonds to a Watson–Crick face (Table 3). Of these, Gln–U, Ser–C, and Ser–C⁺ interactions (#4, #12, #31; Figure 3) have been observed only in RNA complexes, whereas Ser–U, Lys–C, Arg–C, Asn (Gln)–A, and Asp–G interactions (#3, #6, #8, #16 and #28) have been observed in both DNA and RNA complexes. In addition to recognition of the Watson–Crick faces of the bases, some interactions to the major or minor groove faces are found in unpaired or non-Watson–Crick pairing contexts (Table 3), adding further to the diversity of interactions seen with RNAs. For recognition of RNA Watson–Crick pairs, the Arg–G interaction is the most common, as for DNA, and the Asn(Gln)–A interaction is observed rarely (only one), as noted previously.¹⁸

Amino acid–base-pair interactions

One potentially attractive strategy to uniquely recognize portions of an RNA involves simul-

taneous hydrogen bonding to both partners of a non-Watson–Crick base-pair.²¹ The Rev-RRE interaction appears to utilize such a strategy to recognize an unusual G:A base-pair.^{37,38} To systematically examine the possible amino acid interactions with base-pairs, we constructed a database using the 53 possible RNA base-pairs that are bridged by two or more hydrogen bonds (and 17 additional pairs that include thymine).²¹ After removing bifurcated and redundant interactions, as for the unpaired bases, we identified 186 “spanning” interactions in which two or more hydrogen bonds bridge across each pair (Table 1). Table 4 lists all interactions by the 53 RNA base-pairs defined by Walberer *et al.*²¹ As with the unpaired bases, the most common interactions utilize Asn(Gln) (77 arrangements), but in contrast, Arg interactions also are common (64 arrangements), whereas there are only three possible Asp(Glu) interactions, two to non-canonical G:C and C:A⁺ purine–pyrimidine pairs, one to a C:C⁺ pyrimidine–pyrimidine pair, and none to any purine–purine pair. Interestingly, very few Arg interactions are possible with the purine–purine base-pairs (just three arrangements) but are common to the purine–pyrimidine and pyrimidine–pyrimidine pairs.

Of the 186 possible spanning interactions, nine potentially form three hydrogen bonds, all using Asn(Gln) or Arg side-chains (Figure 4). The four interactions involving Arg are with G:U wobble or reverse wobble base-pairs (see below), whereas the five interactions involving Asn(Gln) are with four unusual base-pairs, two G:G and two G:C⁺ pairs. These six base-pairs are among the most commonly used for all spanning interactions

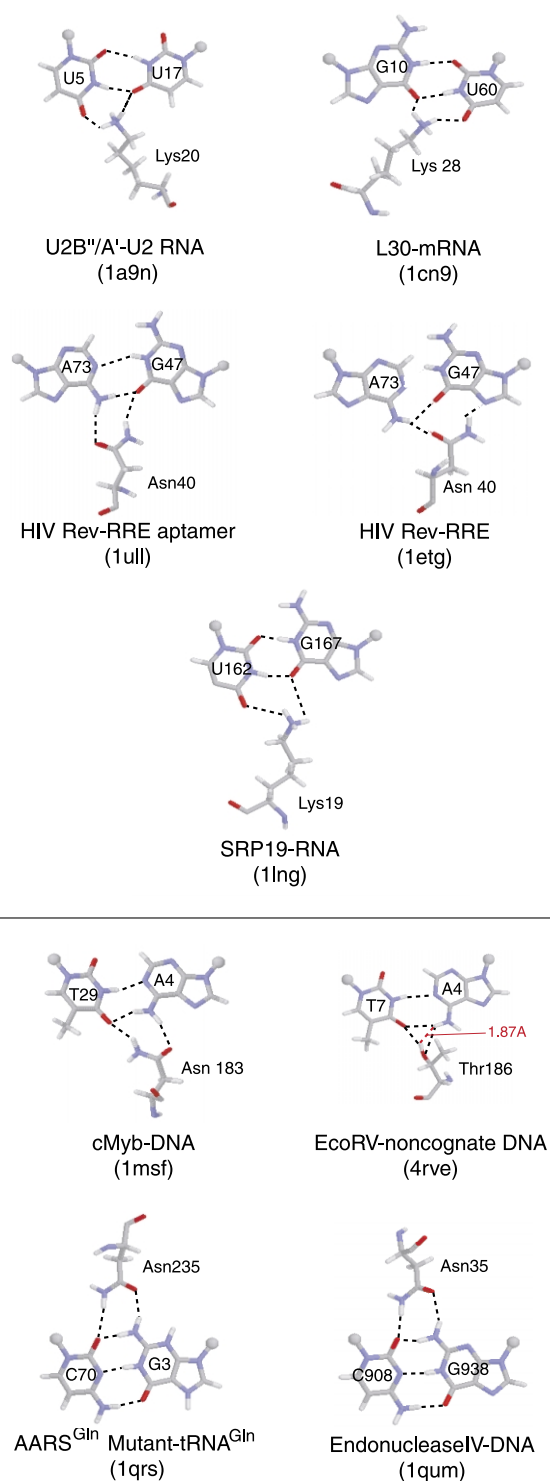


Figure 5. Observed spanning interactions. The top set shows interactions to non-Watson–Crick base-pairs in RNAs, and the bottom set shows interactions to Watson–Crick pairs both in DNA and RNA. PDB identifiers are shown in parentheses, and references are provided in the text. A possible polar hydrogen clash in the EcoRV complex is indicated in red (see text).

(base-pairs #8, 10, 18, 20, 31, 32; Table 4), reflecting the diversity of their donor and acceptor groups. Some of these pairs also are observed to form potential base-triple interactions in which a third

base, rather than an amino acid, is used to span the base-pair.²¹

Nine examples of spanning interactions have been observed (Figure 5), including four to Watson–Crick pairs and two to a G:U wobble pair (discussed below). Six of the nine are in RNA complexes and three of these involve a non-canonical, non-wobble base-pair. In the crystal structure of a spliceosomal U2B''–U2A' protein complex with a U2 snRNA hairpin, Lys20 makes a spanning interaction to a specificity-determining U:U base-pair located in the loop (Figure 5).³⁹ In NMR structures of an HIV Rev peptide bound to an RRE hairpin or to a related RNA aptamer,^{37,38} the carboxamide of Asn40 hydrogen bonds to both bases of an important G:A base-pair (Figure 5). The position of Asn40 in the two Rev peptide–RNA complexes is well-defined by the NMR data, but the Asn–G:A hydrogen bonding arrangements appear to differ (Figure 5). It is not yet clear whether the difference in these spanning interactions reflects the slightly different RNA contexts in which the G:A pair is presented or inaccuracies in the structures. A tight RRE-binding peptide identified from a combinatorial library probably utilizes a Gln side-chain, instead of Asn, in the context of a polyarginine framework to form a spanning interaction to the G:A pair.⁴⁰

Two of the nine observed spanning interactions were not found in our database. One of the reported Rev-RRE Asn–G:A interactions (Figure 5) was missing because the G:A pair in the complex is especially non-planar,³⁷ although it was readily identified when we first used DIVERSIGEN to generate ten conformations of the G:A pair (Figure 2B) and then used WASABI to generate all possible amino acid hydrogen bonding interactions. This case illustrates one limitation to approximating the base-pairs as nearly planar, as well as an approach to account for such interactions. The construction of subsequent databases may explicitly take into account the three-dimensional diversity of base-pairings. The second missing interaction, observed in the structure of an EcoRV–DNA complex,⁴¹ places two polar hydrogen atoms at a distance of 1.87 Å in a Thr–A:T spanning interaction (Figure 5) and was eliminated from the database because polar hydrogen atoms closer than 2.5 Å are considered to have clashing charges. Subsequent crystal structures of EcoRV bound to the same DNA site but with different flanking sequences suggest that Thr186 makes only one hydrogen bond to the O4 of T and that another side-chain (Asn185) may hydrogen bond to the N6 of the paired A,^{42–45} suggesting that the Thr–A:T interaction may not involve two hydrogen bonds.

Spanning interactions to Watson–Crick and wobble base-pairs

Because Watson–Crick base-pairs dominate in nucleic acid structures, followed by G:U wobble base-pairs in RNAs,^{46,47} we examined their possible

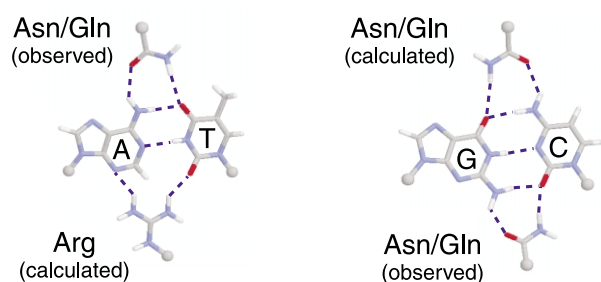


Figure 6. Possible and observed spanning interactions to the Watson–Crick base-pairs. The Asn–A:T major groove interaction has been observed in a c-Myb–DNA complex⁴⁸ and Asn–G:C minor groove interactions have been observed in EndoIV–DNA⁴⁹ and Gln tRNA synthetase–tRNA⁵⁰ complexes.

spanning interactions in more detail. We found four possible interactions with the two Watson–Crick base-pairs (Figure 6). Asn(Gln) can span either the major or minor groove of a G:C pair and the major groove of an A:U(T) pair, whereas Arg can span the minor groove of an A:U(T) pair. Two of these interactions have been observed: an Asn–A:T major groove interaction in a c-Myb–DNA complex and Asn–G:C minor groove interactions in both EndoIV–DNA and Gln tRNA synthetase–tRNA complexes (Figure 5).^{48–50} In the c-Myb complex,⁴⁸ Asn183 hydrogen bonds to both partners of an A:T pair in one of 25 members of an NMR ensemble. While seemingly not well populated, the interaction is within the constraints of the experimental data and mutation of Asn183 to Ala severely reduces binding activity.⁵¹ In the crystal structure of the EndoIV complex,⁴⁹ an Asn35 interaction to a G:C pair represents the only

direct side-chain–base hydrogen bonds in the complex. However, EndoIV is a DNA base excision repair endonuclease that recognizes abasic nucleotides within a protein pocket, so the role of a base-specific spanning interaction is unclear. In the crystal structure of a Gln tRNA synthetase mutant bound to its cognate tRNA,⁵⁰ the mutant Asn235 side-chain hydrogen bonds to both bases of the G3:C70 base-pair in the minor groove of the acceptor stem. Asn is able to make an additional hydrogen bond to the G:C base-pair compared to the wild-type Asp side-chain, consistent with the observed decrease in K_M corresponding to a gain in binding free energy of ~ 1.3 kcal/mol.

The wobble G:U base-pair is very common in RNA structures, and our database contains 16 possible spanning arrangements utilizing Arg, Lys, Asn(Gln), and Ser(Thr/Tyr) side-chains (Figure 7). The Arg and Lys interactions can occur only in the major groove, the Ser(Thr/Tyr) interaction only in the minor groove, and the Asn(Gln) interactions in both grooves. Seven of the 11 Arg–G:U hydrogen bonding arrangements require a non-planar orientation of the guanidinium group relative to the base pair (Figure 7B). One spanning interaction between Lys28 and a G:U wobble pair has been observed in the NMR structure of L30 bound to a hairpin site in its mRNA⁵² (Figure 5), with the position of Lys28 being well-defined by a large number of NOEs. The Lys28 side-chain also appears to make two additional hydrogen bonds to the surrounding RNA tertiary structure formed by this terminal G:U pair of a helix and an adjacent internal loop. Another spanning interaction between Lys19 and a G:U wobble pair has been observed in an SRP19–7SL RNA complex.⁵³

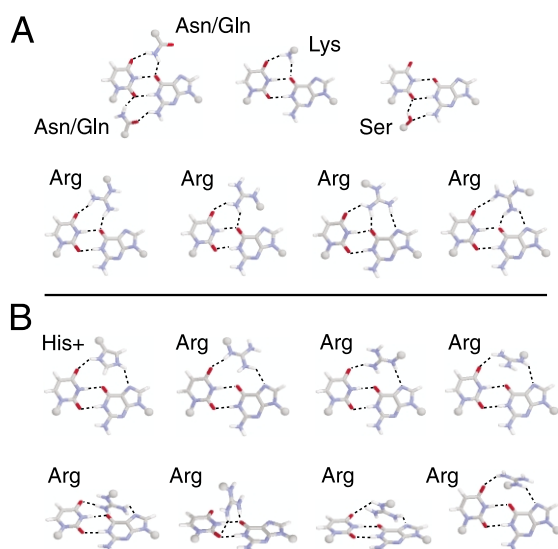


Figure 7. Possible spanning interactions to the G:U wobble pair. A, Interactions in which the side-chains are nearly coplanar with the base-pair and B, interactions that require non-planar orientations.

Discussion

The repertoire of interactions

An early study by Seeman *et al.* defined several ways in which base-pairs might be recognized in the context of a DNA double helix through hydrogen-bonding interactions to amino acid side-chains.⁵ In the case of RNA, bases can be presented in many more structural contexts, including unpaired configurations and a variety of non-Watson–Crick arrangements, and thus the number of possible recognition modes is expected to be quite large. As a first attempt to define the interactions possible within complex tertiary structures and perhaps identify some “rules” of recognition, we performed a systematic search for amino acid–base and amino acid–base-pair interaction patterns based on geometric and steric criteria. We identified ~ 5000 plausible interactions, including 32 with two hydrogen bonds to a single base and 186 with two or three hydrogen bonds that span a base-pair. Only 17 types of interactions to a single base and nine that span a base-pair have been

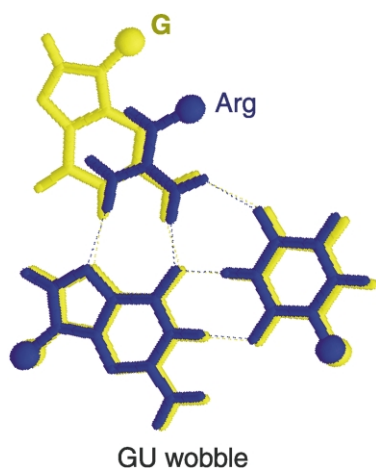


Figure 8. Similarity of a G-G:U base-triple and a modeled Arg-G:U wobble interaction. The base-triple has previously been observed in tRNA^{Asp}.⁶³

observed. Among those that recognize unpaired bases are interactions to the Watson–Crick face of cytosine bases extruded from DNA double helices in two methylase complexes^{35,36} and an Asp–G interaction in TRAP and threonyl-tRNA synthetase RNA complexes.^{54,55} The Asp–G interaction is essential for TRAP binding⁵⁶ and also is observed in the binding of GTP by G proteins, where binding specificity can be switched to xanthine (XTP) by a compensatory change to the donor and acceptor arrangement of Asn.⁵⁷ Our calculated interactions include other interesting arrangements in which amino acids recognize the Watson–Crick faces of unpaired bases or span the G:U wobble pair (see below), and some of these are likely to be observed as the database of RNA–protein complexes expands.

The calculated hydrogen bonding arrangements represent a potentially important class of base-specific interactions, although they do not include interactions with water molecules, backbone moieties, hydrophobic groups, or CH \cdots O bonds.^{17–19,58} To preliminarily assess whether the calculated interactions are energetically reasonable, we computed *in vacuo* interaction energies of the 28 arrangements involving the four unpaired, unprotonated bases (Figure 3) using quantum chemical methods (A.C.C. & A.D.F., unpublished results). Five interactions near the edge of our parameter range (38° donor angle) were unstable (see Figure 3 legend), and none of these has been observed. Of the remaining 23 arrangements, all but Lys–G (#25) appear to reside at stable energy minima, with good hydrogen bond geometries and favorable interaction energies. Thus, our hydrogen bonding criteria generally result in stable arrangements, although the energetic contribution of any individual interaction clearly will depend on its structural context and, in some cases, may be thermodynamically unfavorable while still contributing to binding specificity.⁵⁹

Importance of spanning interactions

One of the most interesting aspects of the databases is the identification of spanning interactions that, in principle, can provide unique ways to distinguish among the 53 possible base-pairs. While only nine spanning interactions have been observed so far, there is evidence that some are important in non-Watson–Crick base-pair recognition (Figure 5). In a U2B''/A'–U2 snRNA complex, Lys20 of U2B'' spans a U:U base-pair at the base of a loop, allowing discrimination between U2 and a related U1 hairpin recognized by U1A.^{39,60} In two Rev peptide–RRE complexes, Asn40 makes a spanning interaction to a G:A base-pair, and both the Asn and G:A pair are critical for recognition.^{37,38,61,62} In an L30–rRNA complex, Lys28 spans a highly conserved wobble G:U base-pair, and mutation of the Lys decreases binding affinity by 20–30-fold.⁵² Similarly, in an archaeal SRP19-7SL complex, Lys19 spans a G:U wobble pair at the base of a tetraloop and is one of only two base-specific interactions formed in the complex.⁵³

Our calculated spanning interactions suggest other strategies to recognize the G:U wobble pair (Figure 7A). Two arrangements in which different donor faces of the Arg guanidinium group are used to form three hydrogen bonds may be particularly favorable, and preliminary energy minimization and quantum chemical geometry calculations indicate that both are stable (data not shown). Interestingly, a model of the *Drosophila* ortholog of the U2B''/A'–snRNA complex described above positions Arg52 in the major groove of a G:U wobble pair where the Lys–U:U interaction is found,³⁹ perhaps replacing one spanning interaction with another. Indirect experimental evidence for an Arg–G:U interaction is provided by the existence of a G–G:U base-triple in tRNA^{Asp}.⁶³ In this triple, part of the Watson–Crick face of G forms three hydrogen bonds to a G:U wobble pair, presenting three donors in an arrangement virtually identical with that of the guanidinium group (Figure 8). Thus, the Arg–G:U and G–G:U interactions can be considered “pseudo-isomorphic”. Calculated databases of base-triples contain many types of spanning interactions where a third base hydrogen bonds to both partners of a base-pair,²¹ and two additional A⁺–G:U and C⁺–G:U arrangements are found that are pseudo-isomorphic to the proposed Arg–G:U interaction. The nearly equivalent arrangement of donors on the Arg guanidinium group and guanine base has been demonstrated by competition experiments that identified a guanosine-binding site in the *Tetrahymena* group I intron.^{64,65}

To date, two spanning interactions have been observed with Watson–Crick base-pairs but our studies suggest two other possible arrangements (Figure 6). A spanning interaction of Asn(Gln) with a G:C pair in the major groove seems especially plausible given that the arrangement of

donors and acceptors on a G:C pair are relatively symmetric in both the major and minor grooves (Figure 6), and given the precedent of the minor groove interaction. However, it is unclear how well such an interaction would discriminate between base-pairs because Asn(Gln) can similarly span the major groove of an A:U(T) pair (Figure 6). In contrast, the Asn(Gln) minor groove spanning interaction, observed in the Gln tRNA synthetase and EndoIV structures, can uniquely distinguish the donor/acceptor arrangements among all base-pairs using an appropriately positioned side-chain, as can a possible spanning interaction of Arg in the A:T minor groove (Figure 6).

In principle, several side-chains might be used to discriminate a G:U wobble pair from the Watson–Crick pairs. From inspection of Table 4, Lys, Ser, or Arg are able to form spanning interactions to the wobble pair but not to the Watson–Crick pairs, whereas Asn can span both types. Thus, if Lys, Ser, or Arg were positioned between the bases of a pair, accurate discrimination might be possible. We favor Arg for this purpose, given its potential to form the three hydrogen-bonded interaction described above.

Complementarity of donor–acceptor arrangements

In general, the bases and base-pairs display a high frequency of adjacent acceptor and donor groups²¹ and consequently, arrangements involving the carboxamide group of Asn(Gln) or the hydroxyl group of Ser(Thr/Tyr), which have complementary acceptor–donor pairs, are highly represented among the possible doubly hydrogen-bonded interactions (Figure 3, Table 4). Such interactions to the single bases are relatively commonly observed (Table 3). In contrast, there are few possible interactions to Asp(Glu), reflecting the limited number of adjacent donor group arrangements on the bases. For unpaired bases, six of the 32 possible arrangements involve Asp(Glu) (Figure 3), but only three have favorable hydrogen bond geometries. Of these, two interactions are to the protonated bases (A^+ and C^+) and one is to G. Interestingly, Asp- C^+ and Asp-G interactions already have been observed (Table 3) despite the involvement of the Watson–Crick face. A previous analysis of DNA–protein complexes revealed that interactions with Asp and Glu are rarely observed, and it was suggested that this probably reflects unfavorable electrostatic interactions between the negatively charged carboxyl group and DNA backbone.¹³ It seems that the arrangement of donors on the bases also inherently disfavors hydrogen-bonded Asp(Glu) interactions. The rarity of hydrogen bonding possibilities for Asp(Glu) may present a good strategy for base-specific recognition and, indeed, two out of the three types of interactions with unpaired bases already have been observed despite the relatively small size of the RNA struc-

tural database, with the Asp-G interaction appearing in five different RNA–protein complexes.

Because donors and acceptors become occupied in a base-pair, it is instructive to examine the doubly hydrogen-bonded interactions possible only in an unpaired context. Such interactions are candidates for recognizing bases in bulges or loops. Interestingly, every amino acid interaction to U (or T) can form two hydrogen bonds to a base only in the absence of any type of base-pairing (assuming two hydrogen bonds are required to form a base-pair). This is a consequence of the fact that U (or T) possesses a total of only three donor and acceptor groups and thus cannot simultaneously form two hydrogen bonds to both another base and to an amino acid, nor can bifurcated bonds be made to the middle N3 donor group. Thus, U bases in RNA bulges and loops in principle could be specified uniquely by two hydrogen bonds and, indeed, seven cases already have been observed (Table 3).

Utilization of the databases

The databases described may be useful for deducing specific amino acid–RNA contacts in conjunction with biochemical data, or by analyzing amino acid–base covariations or compensatory mutations, as attempted for an L11 ribosomal protein–rRNA complex.⁶⁶ In principle, the databases also may be used to engineer “isosteric” change-of-specificity variants, or to aid in designing novel sequence-specific binding proteins whose interactions are guided largely by hydrogen bonding interactions. The databases, named NAIL (nucleic acid interaction libraries), have been placed on a graphical web site† along with a set of filters that can be used to sort through the databases by criteria such as: number of hydrogen bonds, type of amino acid, and type of base or base-pair²¹.

Acknowledgements

We thank Peter Kollman, David Agard, and Wendell Lim, and Bernhard Walberer, Steve Landt, Aenoch Lynn and others members of the Frankel lab for helpful discussions, James Robertson for help with quantum chemical calculations, Wei Wang for advice on energetic calculations, David Konerding for computational advice, and Valerie Calabro, Chandreyee Das, Steve Landt, Robert Nakamura, and James Robertson for comments on the manuscript. We thank the Computer Graphics Laboratory (UCSF) for use of computing resources. This work was supported by NIH grants GM56531 and GM47478 (to A.D.F.) and by NIH training grants GM08284 and GM08388 (A.C.C.).

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

References

- Hermann, T. & Patel, D. J. (1999). Stitching together RNA tertiary architectures. *J. Mol. Biol.* **294**, 829–849.
- Draper, D. E. (1999). Themes in RNA–protein recognition. *J. Mol. Biol.* **293**, 255–270.
- Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
- Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucl. Acids Res.* **29**, 2860–2874.
- Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
- Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*, W. H. Freeman and Co, New York.
- Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61**, 1053–1095.
- Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA–complexes: in search of common principles. *J. Mol. Biol.* **253**, 370–382.
- Lustig, B. & Jernigan, R. L. (1995). Consistencies of individual DNA base–amino acid interactions in structures and sequences. *Nucl. Acids Res.* **23**, 4707–4711.
- Suzuki, M. (1994). A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
- Choo, Y. & Klug, A. (1997). Physical basis of a protein–DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
- Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucl. Acids Res.* **26**, 2306–2312.
- Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
- Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Struct. Funct. Genet.* **35**, 114–131.
- Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
- Steitz, T. A. (1999). *The RNA World* (Gesteland, R. F., Cech, T. R. & Atkins, J. F., eds), 2nd edit., pp. 427–450, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Jones, S., Daley, D. T. A., Luscombe, N. M., Berman, H. & Thornton, J. M. (2001). Protein–RNA interactions: a structural analysis. *Nucl. Acids Res.* **29**, 943–954.
- Allers, J. & Shamoo, Y. (2001). Structure-based analysis of protein–RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **311**, 75–86.
- Treger, M. & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA–protein interfaces. *J. Mol. Recog.* **14**, 199–214.
- Walberer, B. J. (2000). Construction and analysis of a complete database of hydrogen-bonded base combinations, PhD thesis, University of California, San Francisco.
- Walberer, B. J., Cheng, A. C., Frankel, A. D. (2003). Structural diversity and isomorphism of hydrogen-bonded base interactions in nucleic acids. *J. Mol. Biol.*
- 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.
- Israelachvili, J. N. (1989). *Intermolecular and Surface Forces*, Academic Press, New York.
- Taylor, R., Kennard, O. & Versichel, W. (1983). Geometry of the N–H–O=C hydrogen bond. *J. Am. Chem. Soc.* **105**, 5761–5766.
- Taylor, R. & Kennard, O. (1984). Hydrogen-bond geometry in organic crystals. *Accts. Chem. Res.* **17**, 320–326.
- Saenger, W. (1984). *Principles of Nucleic Acid Structure*, Springer, New York.
- Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Molecules*, Springer, Berlin.
- Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.
- Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein sidechain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
- Najmudin, S., Cote, M. L., Sun, D., Yohannan, S., Montano, S. P., Gu, J. & Georgiadis, M. M. (2000). Crystal structures of an N-terminal fragment from Moloney murine leukemia virus reverse transcriptase complexed with nucleic acid: functional implications for template–primer binding to the fingers domain. *J. Mol. Biol.* **296**, 613–632.
- Morellet, N., Demene, H., Teilleux, V., Huynh-Dinh, T., de Rocquigny, H., Fournie-Zaluski, M. C. & Rocques, B. P. (1998). Structure of the complex between the HIV-1 nucleocapsid protein NCp7 and the single-stranded pentanucleotide d(ACGCC). *J. Mol. Biol.* **283**, 419–434.
- Ko, T. P., Williams, R. & McPherson, A. (1996). Structure of a ribonuclease B + d(pA)₄ complex. *Acta Crystallog. sect. D*, **52**, 160–164.
- Horvath, M. P., Schweiker, V. L., Bevilacqua, J. M., Ruggles, J. A. & Schultz, S. C. (1998). Crystal structure of the Oxytricha nova telomere and binding protein complexed with single strand DNA. *Cell*, **95**, 963–974.
- Ding, J., Hayashi, M. K., Zhang, Y., Manche, L., Krainer, A. R. & Xu, R. M. (1999). Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev.* **13**, 1102–1115.
- Klimasauskas, S., Kumar, S., Roberts, R. J. & Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. *Cell*, **76**, 357–369.
- Reinisch, K. M., Chen, L., Verdine, G. L. & Lipscomb, W. N. (1995). The crystal structure of HaeIII methyltransferase covalently complexed to DNA: an extra-helical cytosine and rearranged base pairing. *Cell*, **82**, 143–153.
- Battiste, J. L., Mao, H., Rao, N. S., Tan, R., Muhandiram, D. R., Kay, L. E. *et al.* (1996). Alpha helix major groove recognition in an HIV-1 Rev peptide–RRE RNA complex. *Science*, **273**, 1547–1551.
- Ye, X., Gorin, A., Ellington, A. D. & Patel, D. J. (1996). Deep penetration of an alpha-helix into a widened

- RNA major groove in the HIV-1 rev peptide–RNA aptamer complex. *Nature Struct. Biol.* **3**, 1026–1033.
39. Price, S. R., Evans, P. R. & Nagai, K. (1998). Crystal structure of the spliceosomal U2B''–U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature*, **394**, 645–650.
 40. Tan, R. & Frankel, A. D. (1998). A novel glutamine–RNA interaction identified by screening libraries in mammalian cells. *Proc. Natl Acad. Sci. USA*, **95**, 4247–4252.
 41. Winkler, F., Banner, D., Oefner, C., Tsernoglou, D., Brown, R., Heathman, S. *et al.* (1993). The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* **12**, 1781–1795.
 42. Kostrewa, D. & Winkler, F. K. (1995). Mg²⁺ binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. *Biochemistry*, **34**, 683–696.
 43. Perona, J. & Martin, A. (1997). Conformational transitions and structural deformability of EcoRV endonuclease revealed by crystallographic analysis. *J. Mol. Biol.* **273**, 207–225.
 44. Horton, N. C. & Perona, J. J. (1998). Role of protein-induced bending in the specificity of DNA recognition: crystal structure of EcoRV endonuclease complexed with d(AAAGAT) + d(ATCTT). *J. Mol. Biol.* **277**, 779–787.
 45. Horton, N. C. & Perona, J. J. (1998). Recognition of flanking DNA sequences by EcoRV endonuclease involves alternative patterns of water-mediated contacts. *J. Biol. Chem.* **273**, 21721–21729.
 46. Masquida, B. & Westhof, E. (2000). On the wobble G:U and related pairs. *RNA*, **6**, 9–15.
 47. Varani, G. & McClain, W. H. (2000). The G:U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse systems. *EMBO Rep.*, **1**, 18–23.
 48. Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H. *et al.* (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell*, **79**, 639–648.
 49. Hosfield, D. J., Mol, C. D., Shen, B. & Tainer, J. A. (1998). Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. *Cell*, **95**, 135–146.
 50. Arnez, J. G. & Steitz, T. A. (1996). Crystal structures of three misacylating mutants of *Escherichia coli* glutamyl-tRNA synthetase complexed with tRNA(Gln) and ATP. *Biochemistry*, **35**, 14725–14733.
 51. Gabrielsen, O. S., Sentenac, A. & Fromageot, P. (1991). Specific DNA binding by c-Myb: evidence for a double helix-turn-helix-related motif. *Science*, **253**, 1140–1143.
 52. Mao, H., White, S. A. & Williamson, J. R. (1999). A novel loop-loop recognition motif in the yeast ribosomal protein L30 autoregulatory RNA complex. *Nature Struct. Biol.* **6**, 1139–1147.
 53. Hainzl, T., Huang, S. & Sauer-Eriksson, A. E. (2002). Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature*, **417**, 767–771.
 54. Antson, A. A., Dodson, E. J., Dodson, G., Greaves, R. B., Chen, X. & Gollnick, P. (1999). Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature*, **401**, 235–242.
 55. Sankaranarayanan, R., Dock-Bregeon, A.-C., Romby, P., Caillet, J., Springer, M., Rees, B. *et al.* (1999). The structure of threonyl-tRNA synthetase-tRNA(Thr) complex enlightens its repressor activity and reveals an essential zinc ion in the active site. *Cell*, **97**, 371–381.
 56. Elliott, M. B., Gottlieb, P. A. & Gollnick, P. (1999). Probing the TRAP–RNA interaction with nucleoside analogs. *RNA*, **5**, 1277–1289.
 57. Powers, T. & Walter, P. (1995). Reciprocal stimulation of GTP hydrolysis by two directly interacting GTPases. *Science*, **269**, 1422–1424.
 58. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L. & Zhurkin, V. B. (1998). A role for CH···O interactions in protein–DNA recognition. *J. Mol. Biol.* **277**, 1129–1140.
 59. Szwojdz, D. & Carey, J. (1997). Molecular and biological constraints on ligand-binding affinity and specificity. *Biopolymers*, **44**, 181–198.
 60. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. & Nagai, K. (1994). Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, **372**, 432–438.
 61. Iwai, S., Pritchard, C., Mann, D. A., Karn, J. & Gait, M. J. (1992). Recognition of the high affinity binding site in rev-response element RNA by the human immunodeficiency virus type-1 rev protein. *Nucl. Acids Res.* **20**, 6465–6472.
 62. Tan, R., Chen, L., Buettner, J. A., Hudson, D. & Frankel, A. D. (1993). RNA recognition by an isolated α helix. *Cell*, **73**, 1031–1040.
 63. Westhof, E., Dumas, P. & Moras, D. (1985). Crystallographic refinement of yeast aspartic acid transfer RNA. *J. Mol. Biol.* **184**, 119–145.
 64. Yarus, M. (1988). A specific amino acid binding site composed of RNA. *Science*, **240**, 1751–1758.
 65. Michel, F., Hanna, M., Green, R., Bartel, D. P. & Szostak, J. W. (1989). The guanosine binding site of the Tetrahymena ribozyme. *Nature*, **342**, 391–395.
 66. GuhaThakurta, D. & Draper, D. E. (1999). Protein–RNA sequence covariation in a ribosomal protein–rRNA complex. *Biochemistry*, **38**, 3633–3640.
 67. Norman, D. P., Bruner, S. D. & Verdine, G. L. (2001). Coupling of substrate recognition and catalysis by a human base-excision repair protein. *J. Am. Chem. Soc.* **123**, 359–360.
 68. Goedecke, K., Pignot, M., Goody, R. S., Scheidig, A. J. & Weinhold, E. (2001). Structure of the N6-adenine DNA methyltransferase M. TaqI in complex with DNA and a cofactor analog. *Nature Struct. Biol.* **8**, 121–125.

Edited by D. E. Draper

(Received 23 October 2002; accepted 23 December 2002)