1728–1739 *Nucleic Acids Research, 1999, Vol. 27, No. 7*

Crystal structure of a 14 bp RNA duplex with non-symmetrical tandem G·U wobble base pairs

Jaishree Trikha, David J. Filman and James M. Hogle*

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

Received October 29, 1998; Revised and Accepted February 1, 1999

ABSTRACT

Adjacent G·U wobble base pairs are frequently found in rRNA. Atomic structures of small RNA motifs help to provide a better understanding of the effects of various tandem mismatches on duplex structure and stability, thereby providing better rules for RNA structure prediction and validation. The crystal structure of an RNA duplex containing the sequence r(GGUAUUGC-GGUACC)$_2$ has been solved at 2.1 Å resolution using experimental phases. Novel refinement strategies were needed for building the correct solvent model. At present, this is the only short RNA duplex structure containing 5′-U-U-3′/3′-G-G-5′ non-symmetric tandem G·U wobble base pairs. In the 14mer duplex, the six central base pairs are all displaced away from the helix axis, yielding significant changes in local backbone conformation, helix parameters and charge distribution that may provide specific recognition sites for biologically relevant ligand binding. The greatest deviations from A-form helix occur where the guanine of a wobble base pair stacks over a purine from the opposite strand. In this vicinity, the intra-strand phosphate distances increase significantly, and the major groove width increases up to 3 Å. Structural comparisons with other short duplexes containing symmetrical tandem G·U or G·T wobble base pairs show that nearest-neighbor sequence dependencies govern helical twist and the occurrence of cross-strand purine stacks.

INTRODUCTION

G·U wobble base pairs are the most common non-Watson–Crick base pairs in tRNAs and rRNAs. Indeed, ~15% of the base pairs in *Escherichia coli* 16S RNA are G·U mismatches (1). The existence of G·U pairs was initially used to account for genetic code degeneracy (2). However, more recent studies have found that G·U mismatches are often conserved, suggesting that they may have specific functional and/or structural roles (3,4). G·U wobble pairing has been shown to play a central role in the autocatalytic splicing reactions of group I introns. The guanine from the exon at the 5′ splice site is paired with a uridine from the internal guide sequence of the intron (5–7). In group I introns, G·U base pairing also forms a metal binding site that may be required for correct folding (8). Specific protein interactions with G·U wobble base pairs have been observed in tRNA synthetases (9,10). In addition, G·U mismatches play a significant role in mRNA codon recognition by the tRNA anticodon during protein synthesis and, in particular, are a specific feature of alanyl-tRNA required for mRNA recognition (9,10). Single G·U base pairs are commonly observed in both tRNAs and rRNAs, and adjacent G·U base pairs are very frequently found in rRNAs (3,11,12).

Comparisons of rRNA sequences have shown that the symmetrical tandem mismatches with the sequence 5′-U-G-3′/3′-G-U-5′ are the most common (11,12). This arrangement has been designated motif I (3). Motif I has been shown to be more stable than other mismatches, independent of its Watson–Crick sequence context (13). The symmetrical 5′-G-U-3′/3′-U-G-5′ mismatches (motif II) represent a smaller, though still appreciable fraction of the observed tandem mismatches (3). The presence of motif II may be either stabilizing or destabilizing, depending on the identities and orientations of neighboring Watson–Crick base pairs (14). The stabilizing influence of the symmetric G·U tandem pairs may be related to their ability to form cross-strand purine–purine stacks (15–17). In this distinctive stacking pattern, the six-membered ring of guanine from the G·U lies directly over the six-membered ring of a purine from the opposite strand, rather than the pyrimidine from its own strand. In some known structural examples of motif I: U·G (16,18) and T·G (19), cross-strand stacking involves the two purines of the tandem mismatch stacking with one another. In motif II, cross-strand stacking may involve the purine from a flanking Watson–Crick base pair (15,20). However, cross-strand stacking is clearly not obligatory in motif II, as observed in case of a short DNA sequence with symmetrical tandem G·T repeats (21), where the purine·pyrimidine orientation of the flanking Watson–Crick pair precludes cross-strand stacking.

Non-symmetrical tandem mismatches, such as 5′-U-U-3′/3′-G-G-5′ (motif III) are less well studied than symmetrical ones, partly because longer self-complementary sequences or heterologous duplexes are needed. Non-symmetrical tandem mismatch pairs are less stable, and less common in rRNA than motif I, but slightly more stable and more common than motif II tandem pairs (3,11,12,22). They are also essential components of the metal binding sites of self-splicing group I introns that are required for folding (8,23,24).

There is considerable interest in understanding how RNA sequence specifies structure (25,26). Mismatched base pairs are frequently interspersed among the Watson–Crick base pairs that tend to stabilize regular A-form duplex structures. These mismatches may provide sequence-specific punctuation marks in the RNA structure. Perforce, the interactions that stabilize complicated tertiary folding patterns may involve portions of the structure which do not form stable duplexes. The rules governing duplex stability are not entirely straightforward because the stabilizing or destabilizing effect of the base pair depends critically on the sequence context in which the mismatch appears.

*To whom correspondence should be addressed. Tel: +1 617 432 3918; Fax: +1 617 432 4360; Email: hogle@hogles.med.harvard.edu

Possible relevant factors include the identity and orientation of its neighboring Watson–Crick base pairs, and whether the mismatched pair appears alone or in tandem with other mismatched pairs. For tandem mismatched pairs, the thermodynamic results (13,14,22) suggest that the three base steps formed by two tandem mismatches and their two flanking Watson–Crick pairs make independent contributions to helix stability unless size discrepancies cause extreme distortion of the main chain. Thus, a consideration of nearest-neighbor energies is often sufficient to predict the effect of the mismatch on duplex stability for tandem mismatches with similar C1′–C1′ dimensions (including tandem G·U pairs). Understanding the factors that affect duplex stability of various sequences is relevant to the success of structure prediction. Sequence-structure correlates for RNA will prove increasingly valuable as computational techniques for structure prediction and validation improve. To obtain this information, simple RNA structures are currently being studied by crystallography (15,16,27–36) and by NMR (24,37), and the results are being correlated with assessments of folding stability (13,14,18,22).

We report here the 2.1 Å resolution crystal structure of the RNA duplex, r(GGUAUUGCGGUACC)$_2$, containing two sets of adjacent G·U base pairs in the sequence 5′-U-U-3′/3′-G-G-5′. The G·U pairs are located close to the middle of the duplex, and are separated by a central pair of G·C base pairs (Fig. 1a). The 14 nt sequence was originally synthesized as a part of an ongoing project to study the 5′ non-coding region of poliovirus RNA and its complex with the viral protease/polymerase precursor 3CD as this sequence contains the 3CD binding site. Though apparently monomeric in dilute solution, the 14mer formed a duplex at the concentrations used for crystallization. Along with (23), the 14mer duplex is one of only two crystallized examples of a non-symmetrical G·U tandem mismatch. Furthermore, because metal amines were not required for crystallization, it should also be useful in identifying which aspects of the 5′-U-U-3′/3′-G-G-5′ structure are independent of ion binding. The RNA duplex is also expected to be of interest in addressing fundamental issues of RNA sequence/structure correlates. The structure provides two crystallographically independent examples of motif III, which provide an excellent opportunity to test the applicability of the nearest neighbor model (38) to RNA duplexes containing non-symmetrical G·U tandem pairs.

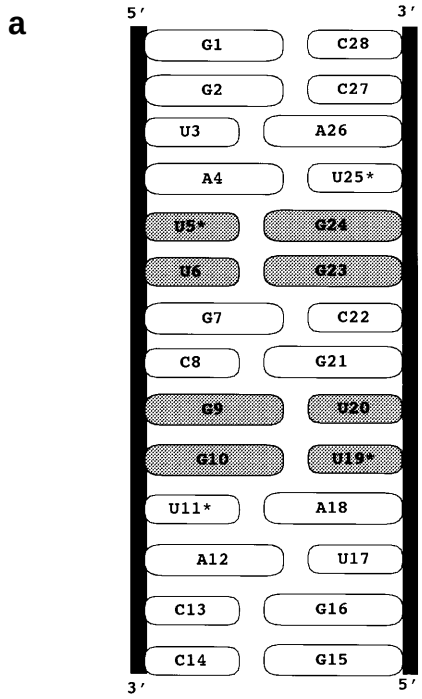


Figure 1. (Above and opposite) Schematic representations of the 14mer duplex. (**a**) Base-pairing pattern. G·U base pairs are shaded. Uridines with * were replaced by bromo-uridines for MIR phasing. (**b**) Experimentally phased electron density map at 2.85 Å resolution. A portion of the map is shown in stereo with the final model superimposed. (**c**) The overall shape of the 14mer, shown in stereo. The two strands and their backbone ribbon representations are colored yellow and purple. The wobble base pairs are indicated in blue and red. The three axes shown in green were each calculated separately from the coordinates of G1·C28–G7·C22, C8·G21–A11·U18 and U12·A17–C14·G15, using the program CURVES (54). (c) was made with InsightII (MSI).

MATERIALS AND METHODS

RNA synthesis and purification

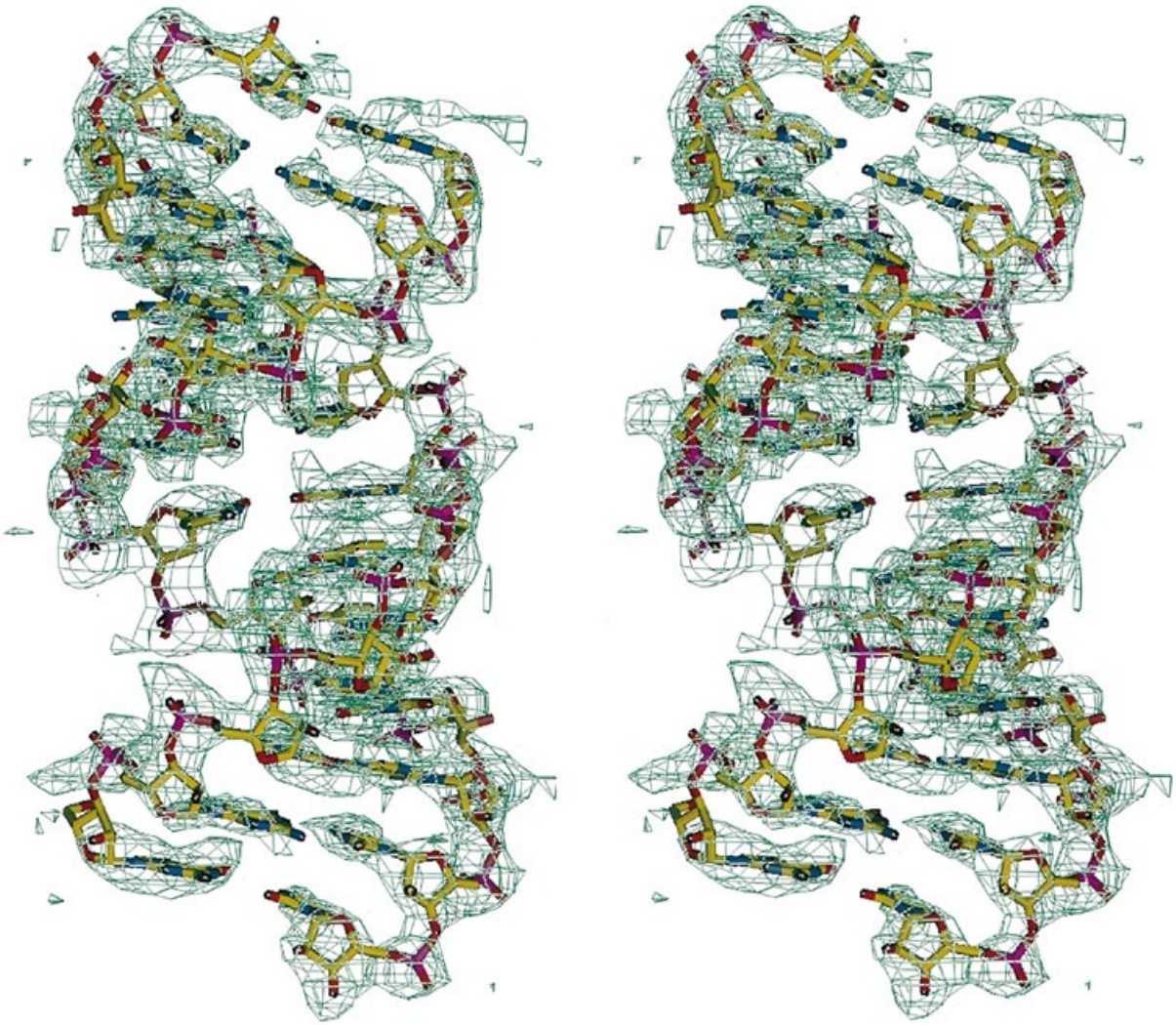
The oligonucleotides were synthesized using standard solid phase phosphoramidite chemistry on an automated nucleic acid synthesizer (Applied Biosystems DNA/RNA synthesizer, Model 392) using standard protocols with a 12 min coupling step (39). β-cyanoethyl, 2′ tertbutyldimethylsilyl ribose phosphoramidites with standard base protecting groups were used (40). The oligonucleotides were cleaved from the solid support and deprotected using 3:1 (v/v) ammonia/ethanol at 55°C overnight. The mixture was concentrated and dried in a lyophilizer. The 2′ hydroxyls were deprotected by overnight incubation with 1 M tetrabutylammonium fluoride for 48 h followed by ethanol precipitation. The sample was then dissolved in water and purified by anion exchange HPLC using a gradient of 0–1 M NaCl. Pooled fractions with high OD_{260} were precipitated with acetate and ethanol, and the precipitate was resuspended in water.

Three of the RNA sequences that were synthesized yielded useful crystals: one using standard amidites, and two with bromo substitutions at the C5 positions of uridine for use in heavy atom phase determination. One of the substituted sequences had bromo-uridine at position 11, and the other had bromo-uridine at positions 5 and 11 (Fig. 1a).

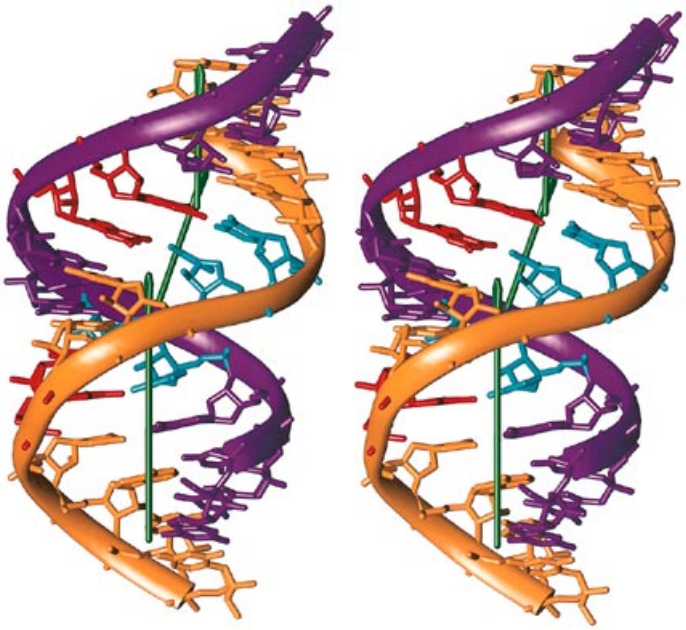
Crystallization and data collection

RNA was dissolved at a concentration of 8 mg/ml in 50 mM sodium cacodylate at pH 6.5 containing 15% 2-methyl-2,4 pentanediol (MPD) and was then crystallized in hanging drops by vapor diffusion against 100 mM sodium cacodylate and 30% MPD at room temperature. Crystals normally grew overnight and were typically $0.3 \times 0.3 \times 0.7$ mm in size. Native and brominated crystals were isomorphous to 3 Å resolution and belonged to space group $P6_322$ with unit cell dimensions a = b = 45.4 Å, c = 147.9 Å , $\alpha = \beta = 90°$ and $\gamma = 120°$. Crystals used for data collection were stabilized in 35% MPD in water, mounted in a rayon loop, frozen in a stream of nitrogen at –160°C, and maintained at this temperature throughout data collection. X-ray diffraction data to 2.4 Å resolution were collected on MAR detector using a Siemens rotating anode X-ray generator operated at 50 kV and 100 mA. Higher resolution data sets (2.1 Å) for the native crystal and data sets containing Bijvoet pairs for the singly brominated derivative were collected at CHESS on beam line F1 using a 1K CCD detector (Area Detector System Inc.). Data were integrated using DENZO and merged using SCALEPACK (41). Data collection statistics for the native and derivative data sets are

b



c



summarized in Table 1a. Due to base pair stacking and the alignment of helices parallel to the unit cell axis, **c**, extremely strong meridional reflections were present in the data between 3.0 and 2.6 Å resolution. In longer exposures that were calibrated for collecting the weaker data correctly, these strong reflections exhibited long visible tails that made it difficult to integrate the data. To properly collect data in this resolution range an additional native data set was collected with shorter exposure times.

Structure determination

The initial phase determination regarded the non-brominated (rotating anode) data as native, and obtained isomorphous phasing contributions from the 11-Br and 5,11-di-Br data sets, with no anomalous phasing contributions. Bromine positions were identified using SHELX-90 (42) in Patterson search mode and confirmed by difference Patterson maps. Coordinates for the two 11-Br sites were identical in both derivatives. Heavy atom positions were refined and phases calculated to 3.0 Å with MLPHARE (43). In the resulting electron density map, most of the backbone atoms and some of the bases were apparent. An A-form RNA double helix was built into the map, using the known bromine positions as sequence guides, and optimized using X-PLOR (44). Phases from this preliminary native model were applied to heavy atom difference Fourier amplitudes, and the resulting difference maps clearly showed the heavy atom positions.

Later on, the data were reprocessed, with the extremely strong meridional reflections integrated separately. Using this improved data set, the experimental phases were recalculated with MLPHARE, this time regarding the 11-Br derivative as native. Differences involving the doubly brominated RNA were represented by two 5-Br sites with positive occupancies; the non-brominated RNA crystals were assumed to have two 11-Br sites with negative occupancies; and anomalous contributions from the 11-Br synchrotron data were included. All occupancies refined to nearly equal absolute values. Phasing statistics for this calculation were acceptable (Table 1b). In the resulting experimentally phased electron density map the ribose phosphate backbone of the molecule was clearly visible and continuous for all residues. Electron density was well-defined for all of the nucleotide bases, and each base shows obvious connectivity with the corresponding ribose except for G1, G7, C8, G15, G21 and G28 (Fig. 1b).

The atomic model was built using O (45) and FRODO (46), and refined using X-PLOR 3.851. The RNA–DNA topology and parameter set from the nucleic acid database was used for refinement (47). Early in the refinement process, at 2.89 Å resolution and below, MIR phases were included as targets, and rigid body, atomic positional and simulated annealing refinements were carried out. Subsequently, the phase restraints were omitted and all of the native synchrotron data between 20.0 and 2.1 Å were included.

In the latter stages of the refinement, resolution-dependent bin scaling was incorporated via an X-PLOR macro. The final model includes a set of 47 resolution-dependent bin scales, each based on a window of five adjacent shells to ensure that each scale is based on a sufficient number of reflections. Each time the macro was invoked, honest R-factors were calculated relative to the unscaled values of F_{obs}. Then a scaled version of the F_{obs} set was supplied to the gradient-calculating routines for evaluating parameter shifts. Despite the fact that bin scaling is not intrinsic in X-PLOR, this approach allowed all of the available low resolution data to be included, while preventing scaling errors from causing incorrect parameter shifts.

Modeling fixed solvent

The most unusual aspect of the refinement was that, until the very final stages of the structure determination, the correct fixed solvent positions did not correspond to the largest chemically plausible features in the map. Indeed, the largest positive peaks in the early difference maps fell directly on the phosphorus atoms of the model. The problem was first evident when candidate solvents identified by the usual criteria failed to improve R or R_{free} significantly. Seen in retrospect, 75% of the solvent sites built conventionally turned out to be wrong. Subtle but concerted shifts of the RNA coordinates during refinement—particularly the electron-rich phosphate groups—may have compensated for the missing scatterers and flattened the difference map so ‘successfully’ that the missing solvent positions and noise peaks became indistinguishable. (Analogous problems occur in locating the hydrogen atoms in small molecule crystal structures containing heavy scatterers.)

An automatic process, mostly dependent on R_{free}, was used to distinguish authentic solvent sites from noise. By definition, the ‘test set’ reflections had not participated in previous refinements, so their ΔF values were still effective for identifying the missing scatterers. In each round of model-building, a long list of potential solvent sites was created (typically 25–50), accepting low contour peaks provided that their shapes and potential for hydrogen bonding were appropriate. In the X-PLOR macro, candidate solvents were repeatedly toggled between ‘occupied’ and ‘unoccupied’ states, to see what combination of sites had the most favorable statistical impact. Old waters were always introduced in the ‘on’ state, and new waters were introduced in the ‘off’ state, assigned a nominal 1% occupancy to avoid biasing the phases while permitting positional improvement. The selection process was alternated with positional and temperature factor refinement, based on working set reflections only, until convergence, a point where the set of provisionally accepted waters stopped changing. Water molecules with 1% occupancy were then eliminated, the model was re-refined, a new map was calculated, and the map was searched once again for additional sites. After this process was repeated several times, the solvent composition stabilized and the final solvent model was obtained.

At the conclusion of this process, the atomic positions for the RNA did not appear to have changed significantly. Nevertheless, the large peaks at the phosphorus positions had disappeared from the difference maps. The number of solvent molecules is only slightly larger than before (72), and their positions have improved. Thus, fixed solvents with acceptable hydrogen bonding patterns now correspond to the largest peaks in the solvent region of the map, as they should, and both the working R and free R have improved by about three percentage points.

The partly automated approach that was used for selecting solvent sites relied primarily on the test set reflections (R_{free}) to decide which solvents to retain, and on the working set reflections to decide their precise positions. The potential drawback of this approach is that the R_{free} statistic is no longer an unbiased indicator of the validity of the model. In the future, a possible alternative strategy would be to use only a subset of the ‘test’ reflections in the occupancy-toggling calculation. It would also be useful to have a more computationally sophisticated solution to the combinatorial problem of identifying the correct subset of the water list.

RESULTS

Structure solution

The 14mer RNA duplex crystallized in space group $P6_322$, with one duplex per asymmetric unit (Table 1). Initial phases to 2.85 Å resolution were obtained from isomorphous bromine substitutions with an anomalous contribution (Table 1b). In the experimentally phased electron density map (Fig. 1b), the ribose phosphate backbone of the molecule was clearly visible and continuous for all residues. Electron density was well-defined for all of the nucleotide bases, and each base showed obvious connectivity with the corresponding ribose except for G1, G7, C8, G15, G21 and C28. An atomic model was constructed, using the known bromine positions as guides, and refined to 2.1 Å resolution using XPLOR (44). Fixed solvent positions were identified using an automated trials procedure (see Materials and Methods) after conventional model-building failed.

Table 1. (**a**) Data collection

Data set	Native (CHESS)	Native (CuKα)	5,11-di-Br-derivative (CuKα)	11-Br derivative (CHESS)	11-Br derivative (CuKα)
No. of observations	18 099	55 193	40 317	17 663	29 097
No. of unique reflections	4969	3629	3456	2679	2272
Resolution (Å)	2.1	2.45	2.45	2.45	2.85
Completeness (%)	85.6	97.3	95.5	72.2	95.3
R_{sym} (%)	2.8	5.9	11.0	4.6	9.8

Where $R_{sym} = \Sigma_{i,(h,k,l)} | I_{i(h,k,l)} - I_{(h,k,l)} | / \Sigma_{i,(hkl)} | I_{(hkl)} |$. $I_{(hkl)}$ is the statistically weighted average intensity of symmetry equivalent reflections.

(**b**) Phasing: using the 11-Br dataset as native

Data set	Resolution (Å)	R_{cullis} acentric	centric	Phasing Power acentric	centric
non-brominated	2.85	0.58	0.51	2.36	1.77
5,11-di-Br derivative	2.85	0.54	0.48	2.38	1.95
11-Br anomalous signal	2.85	0.94	0.89	0.61	0.46
	Resolution (Å)	Figure of merit acentric	centric		
	10–2.85	0.64	0.85		
	3.15–2.85	0.57	0.78		

$R_{cullis} = \Sigma_{(hkl)} | | \mathbf{F}_{PH} | - | \mathbf{F}_P + \mathbf{F}_H | | / \Sigma_{(hkl)} | \mathbf{F}_{PH} - \mathbf{F}_P |$.
Phasing Power $= \Sigma_{(hkl)} | \mathbf{F}_H | / \Sigma_{(hkl)} | | \mathbf{F}_{PH} | - | \mathbf{F}_P + \mathbf{F}_H | |$.
F_P, F_{PH} and F_H represent structure factors for native and derivative data and for the heavy atom model, respectively.

(**c**) Statistics after model refinement

Unit Cell (Å)	a = b = 45.4, c = 147.9		
Space Group	P 6_322		
Resolution Shell Limits (Å)	Completeness (%)		
20.0 – 2.33	92.5		
2.33 – 2.24	72		
2.24 – 2.17	58		
2.17 – 2.10	44		
20.0 – 2.10	84.3		
Number of unique	Centric	Non-centric	Total
reflections	1319	3540	4859
R_{cryst} (%)	24.4	21.8	22.5
R_{free} (%)	34.8	25.2	27.7
r.m.s. disagreement with idealized standards			
Bond Lengths (Å)	0.010		
Bond Angles (°)	1.22		
Dihedral angles (°)	1.54		
Improper dihedrals (°)	7.55		
Non-crystallographic symmetry (approximate): rotation matrix to superimpose nucleotides 1–14 onto 15–28, and *vice versa*. (This is applied at the center of mass of the duplex.)			
	–0.6003	0.7997	–0.0130
	0.7997	0.5999	–0.0259
	–0.0130	–0.0259	–0.9996

This matrix specifies a 180° rotation about the eigenvector: (0.44705, 0.89439, –0.01449).

The final model includes 592 non-hydrogen atoms of the RNA duplex and 72 fixed water molecules, individual isotropic temperature factors, and 47 bin scales. All of the solvents share a single collective occupancy value of q = 0.5 because the addition of that one extra parameter yields a 1–2% improvement in R_{free}, relative to refinements run with q = 1.0. The geometry of the final model is acceptable (Table 1c). When Luzzati's statistical treatment (48,49) is used, the centric and non-centric R_{free} values (Table 1c) suggest an r.m.s. coordinate error of ~0.2 Å. Note that the data set contains an unusually large percentage of centric reflections (27%), where the centric R values are expected to be higher for any given level of coordinate uncertainty.

Overall conformation

The 14mer RNA sequence crystallizes as an irregular RNA duplex which is bent at two points. To a good approximation, the helix can be divided into three straight segments arranged to form a broad S (Fig. 1c), with the two outermost segments lying nearly parallel, at angles of 107° and 104° relative to the middle segment. The two strands of the RNA have been assigned residue numbers 1–14 and 15–28. The three straight segments consist of base pairs 1–7:28–22, 8–11:21–18 and 12–14:17–15.

The 14mer duplex can be superimposed on a canonical A-form RNA duplex with an r.m.s.d. of 1.6 Å for all atoms. A similar value is calculated when the nucleotides involved in G·U wobble pairing are omitted. In contrast, when the nucleotides involved in Watson–Crick base pairing are divided into three duplex segments, the r.m.s.d. for each of the segments is <1.0 Å. This shows that more than half of the mean square discrepancy is caused by the departure of the segments from co-linearity.

Although the sequence of the 14mer duplex is perfectly symmetrical (Fig. 1a), its structure in the crystal is slightly asymmetrical, presumably due to crystal packing forces. If the two ends of the duplex are exchanged, a least squares superposition of the rotated and unrotated copies (Table 1c) yields an r.m.s. difference between equivalent atoms of 1.3 Å. The largest discrepancy occurs at residue 1, the site of an intermolecular packing contact, where the r.m.s. difference between chains is 1.9 Å. The next largest differences, ranging from 1.4 to 1.5 Å, involve residues 5, 6, 11 and 12. These values represent identical chemical species in different crystal packing environments.

Crystal packing

The RNA duplexes are packed end-to-end in the crystals. Each duplex stacks with a 2-fold symmetry-related copy of itself to form a pseudo-continuous column. The local helix axis of each 14mer lies parallel to the crystallographic **c** axis, and translated laterally by ~9 Å to lie alongside of it. To maintain an unbroken succession of base-stacking contacts between neighbors, each duplex must be under-wound slightly at its ends. The extent of under-winding can be assessed at the pseudo-CpG step between neighboring duplexes, where the helical twist angle is –34.9°, and the helical rise is 3.86 Å. These values are very similar to those for an authentic base pair step in A-form RNA, and demonstrate the continuity of the stacking pattern.

A useful way to conceptualize the crystal packing is to view the translationally repeating unit as a super-helical coil 12 molecules long that spirals around the **c** axis completely once in every three unit cells (444 Å). Three 3-fold-related copies of the super-helix intertwine around the **c** axis to generate an infinite fiber. Full-cell translations in the **a**–**b** plane yield a hexagonally packed array of these fibers spaced 45.4 Å apart, with each fiber having six neighbors.

Superficially, packing of the individual duplexes in the **a**–**b** plane also appears to be hexagonal, forming a continuous sheet of molecules between z = 0 and z = 1/4. Each hexagonal sheet is incomplete, however, because there is a very large volume of solvent around one of the crystallographic 3-fold axes. This gives each duplex only four near neighbors, instead of six. The large solvent volume appears to be wide enough to accommodate an additional disordered duplex, though there is no clear evidence that any such duplex is actually present. The side-by-side packing of helices is held together primarily by backbone-to-backbone contacts through the O2′ of ribose moieties. Water molecules also bridge the backbone atoms of symmetry related molecules. Atoms of the bases are involved in helix packing contacts on only two occasions: these involve interactions in the minor groove of N2 from residues G23 and G21 with the O2′ of symmetry related duplexes. Side-to-side packing contacts via the O2′ groups of RNA, which are frequently seen in RNA crystal structures, are likely to be characteristic of RNA's biologically relevant interactions with its protein or nucleic acid ligands, and to stabilize tertiary folding (reviewed in 50).

Helix parameters

The conformation of the 14mer duplex is generally A-form. Many of the standard helical parameters are A-like throughout the molecule; and all of the furanose rings have c3′-endo pucker, as indicated by their δ angles. Some of the most important differences from A-form helix are global, and affect all of the nucleotides in the duplex. In particular, the tilt angles are uniformly smaller (0.3 ± 4.5°), the roll angles (8.4 ± 2.0°) and the rise per base step (3.3 ± 0.3 Å) are always larger, and the slide (–2.1 ± 0.3Å) and propeller twist (–10.0° ± 5.1°) are consistently more negative. Non-localized changes such as these show how the stresses of wobble pairing can be distributed over the length of the molecule. Notably, all of these parameters involve the positions and orientations of the bases, and indicate systematic changes in the extent of base stacking contacts.

Global effects on the molecular shape are also evident in the width and depth of the major and minor grooves. The major groove width increases monotonically from 4.5 Å at U5.G24 to ~7.3 Å at G10.U19, reflecting the asymmetry of the molecule. The minor groove width ranges from 9.3 to 10.8 Å near the Watson–Crick base pairs, but drops significantly around the G·U wobble pairs, becoming as low as 8.3 Å at G9·U20. Thus, the major groove is consistently wider than the 4.1 Å expected for A-form RNA, and the minor groove is consistently narrower than 11.3 Å. In the extreme, these differ from A-form values by as much as 3 Å.

Most of the other deviations from A-form values are localized, and primarily affect helical parameters in the immediate vicinity of the wobble pairs or at the helix ends. For example, exceptional values of the glycosyl torsion angles (χ) occur at C8 (–141°), G9 (178°), G23 (180°), and G24 (–171°) and also, trivially, at one end of the duplex where two symmetry-related copies of the base G15 (–176°) participate in a crystal packing contact.

Reorientation of G·U pairs about the helix axis, relative to their Watson–Crick paired neighbors, is stabilized, perforce, by the G·U wobble hydrogen bonding pattern. One parameter that reflects this difference in orientation is lambda (λ), the angle between the glycosidic bond and the C1′–C1′ vector. In the Watson–Crick base pairs of A-form helices, G and U have nearly equal λ values: 54.4° for G and 57.4° for U (a similarity corresponding to the concept of a pseudo-dyad). In contrast, the wobble-paired G and U bases of the 14mer have dissimilar λ values: 43.7° ± 1.4 for G, and 69.6° ± 1.5 for U.

Reorientation of the G·U base pairs profoundly affects helical twist angles at the base steps between the tandem wobble pairs and their Watson–Crick paired neighbors. Significant under-winding of the base steps U6·G23/G7·C22, and C8·G21/G9·U20 is reflected by helical twist angles of 21.5 and 18.0°. Under-twisting of the helix corresponds to the cross-strand stacking of G9 over G21, rather than over C8, and the stacking of G7 over G23, rather than over U6 (Fig. 2). A compensating over-winding of the base steps U5·G24/A4·U25 and G10·U19/U11·A18 is indicated by helical twist angles of 37.9 and 36.2°. The overall S-shape of the duplex reflects the fact that the compensation is only partial. The helical twist angles between non-symmetric tandem G·U pairs (but not symmetric ones) have A-like values because the orientations of both G·U pairs are twisted in the same direction.

Exceptional values of the main chain torsions occur around residues G9 and G23, and correspond to the rotation of the guanines into the minor groove. Compensating changes in the torsions α (167.3 and 147.8°) and γ (168.9 and 179.1°) reflect a locally extended conformation of the backbone, with the distance between adjacent phosphorus atoms increased from 5.9 to 7.0 Å on average. Correlated changes in α and γ are common in A-type RNA duplexes (51), but are not obligatory results of G·U pairing.

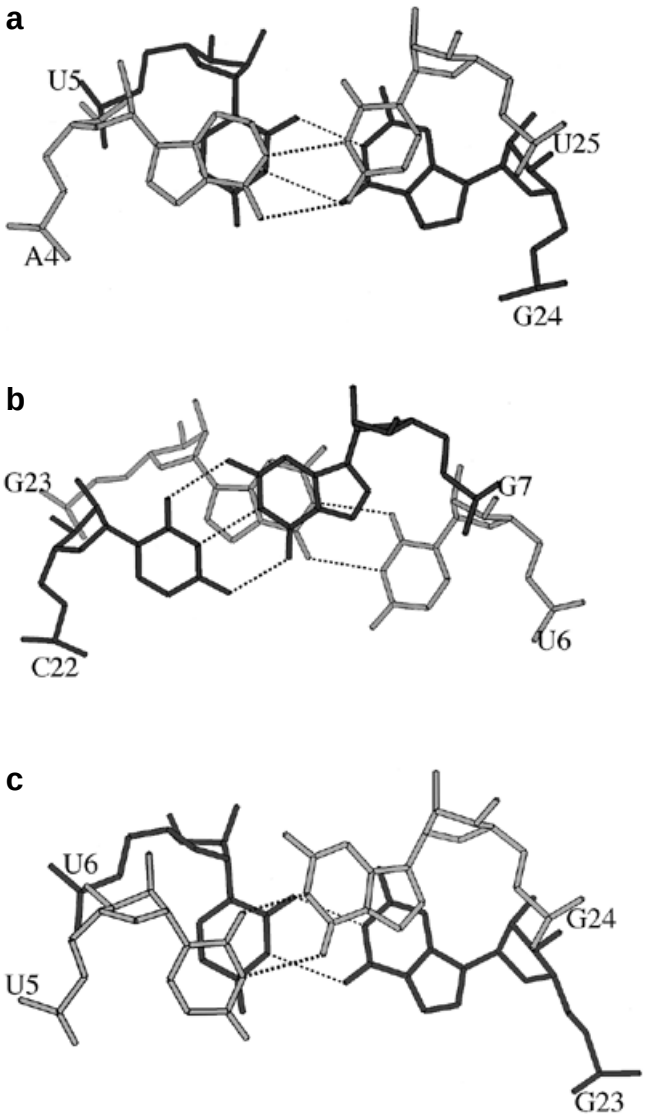


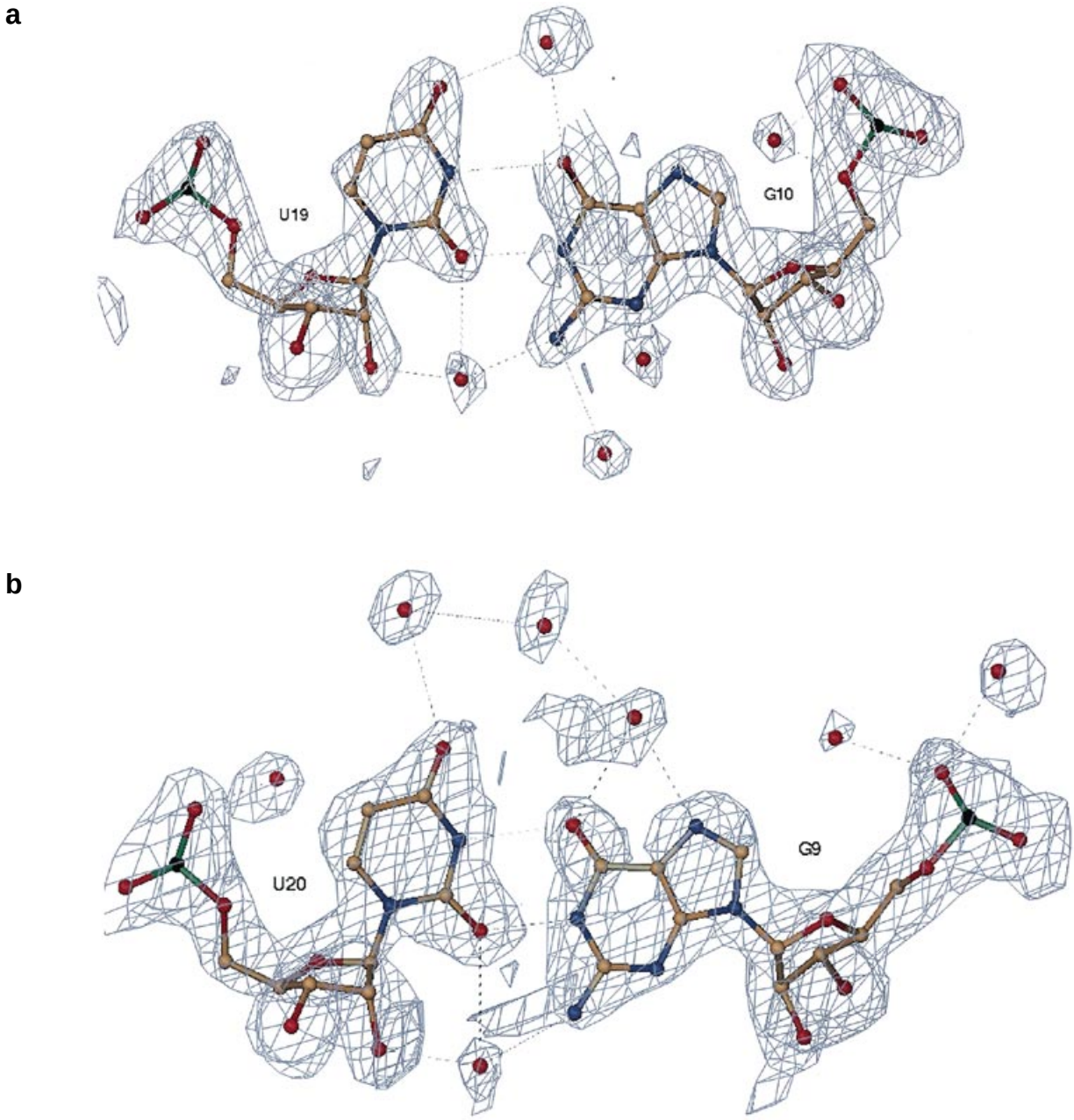
Figure 2. The stacking of wobble base pairs with flanking base pairs. (**a**) The stacking of A4·U25 (light gray) over U5·G24 (dark gray) illustrates the effect of over-twisting the helix. (**b**) The cross-strand stacking of guanine residues is due to under-twisting of the helix at the base step between the G7·C22 (dark gray) and U6·G23 (light gray) base pairs. (**c**) At the base step between the wobble pairs U5·G24 (light gray) and U6·G23 (dark gray) the bases are partially overlapped.

Influence of G·U pairs on hydration

The 14mer structure provides an opportunity to study the effects of non-symmetric tandem G·U base pairs on the hydration of an RNA duplex, absent the influence of the polyamines and complex metal amines that are frequently components of RNA crystallization experiments. At 'low resolution', the overall distribution of solvent shows considerable order: in both the major and minor grooves, a majority of the solvent molecules are concentrated in the middle part of the duplex, near the wobble base pairs, rather than at the ends of the duplex. The tendency of the tandem wobble pairs to order the solvent around them more strongly than the Watson–Crick pairs do is consistent with their known role in forming ion-binding sites (8,23,24). When the solvation model is analyzed in greater detail, only a few of the fixed solvent molecules fit obvious patterns. Specifically, in the minor groove, each of the four G·U wobble pairs has an 'invariant' solvent that bridges between N2 of guanine and O2 and O2′ of uridine (see below). For the other fixed waters, the RNA sequence failed to specify the solvent structure uniquely. Thus, when the approximate non-crystallographic symmetry operator (Table 1c) was used to rotate the duplex onto itself, the rotated solvent model showed little correspondence with the unrotated model.

Locally, solvation of the G·U pairs (Fig. 3a–d) is dominated by the wobble hydrogen bonding pattern. In each G·U pair, the ring nitrogens N1 of G and N3 of U donate hydrogen bonds to the keto oxygens O2 of U and O6 of G, respectively. The hydrogen bonds have normal geometries, and are not bifurcated, though this point was not resolved until late in refinement. In all four G·U pairs, the N2 amino group of G, an unpaired hydrogen bond donor, protrudes into the minor groove, and the O4 keto group of U, an unpaired acceptor, is presented in the major groove. Tandem G·U pairing causes large concentrations of charge on the molecular surface (Fig. 4), and thereby produces specific recognition sites.

The network of water molecules around each G.U base pair (Fig. 3a–d) compensates for the unpaired donor and acceptor. The only solvent position common to all four G·U pairs bridges N2 of guanine and O2 and O2′ of uridine. This 'invariant' minor groove water is frequently flanked by ribose or phosphate of neighboring



bases that help to define the boundaries of its binding site. Analogous solvent positions are commonly reported for other G·U and G·T mismatches (15,16,21,29,31).

The remaining waters near the G·U wobble pairs are not well conserved, in either the minor or major groove, possibly because each G·U pair presents more than one potential hydrogen bonding group along each solvent-exposed edge. For example, in the major groove, the O4 of uridine and the O6 of guanine form one possible binding site, while the O6 and N7 of guanine form a second possible site. In the U19·G10 wobble pair, the first site is found (Fig. 3a); both such sites are occupied in U20·G9 (Fig. 3b) and U6·G23 (Fig. 3c); and neither shows significant electron density in U5.G24 (Fig. 3d).

In the minor groove, the 'invariant' water is sometimes accompanied by a second fixed water. This second water could be located adjacent to the amino group (N2) or ring nitrogen (N3) of guanine, or shared by both. Guanines G9, G10 and G23 all are solvated at N2, while G24 is not (Fig. 3). The solvation pattern of the G·U pairs in the minor groove is further complicated by the proximity of symmetry-related molecules. In the U5·G24 base pair, the guanine ring nitrogen (N3) is located too close to a symmetry-related phosphate to permit hydration (4.4 Å). In U20·G9, the water at N2 hydrogen bonds to a symmetry-related phosphate. In U6·G23 and U19·G10, the 'invariant' waters bridge contacts with symmetry-related ribose and phosphate groups, respectively. The protrusion of the guanine N2 into the minor

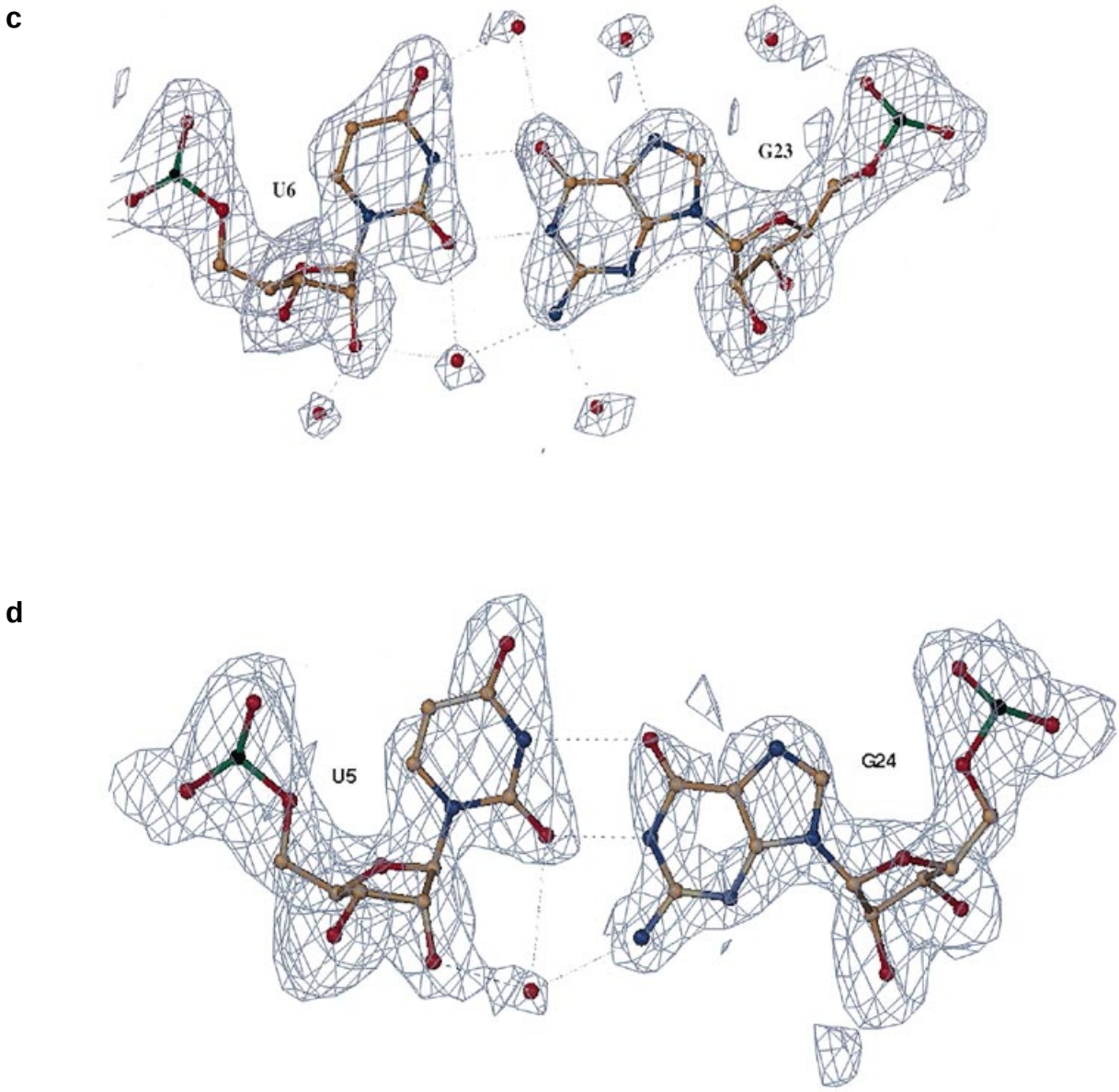


Figure 3. (Above and opposite). The four G·U base pairs and the solvent molecules located in their immediate vicinity. The refined model is shown superimposed on portions of the model-phased F_o map. Hydrogen bonds are indicated by broken lines. (**a**) G10·U19 (**b**) G9·U20 (**c**) U6·G23 (**d**) U5·G24. Electron density figures were made with O (45).

groove helps to explain the coincidence that all four G.U pairs participate in side-to-side crystal packing contacts.

DISCUSSION

Tandem mismatches and stacking

Normal base stacking patterns for A-form RNA duplexes have been described by Arnott *et al.* (51,52). Three common stacking motifs are as follows: (i) for 5′-pyrimidine-purine-3′ base steps, the six-membered rings of pyrimidine and purine from the same strand commonly overlap; (ii) for 5′-purine-pyrimidine-3′ base steps, some limited cross-strand overlap is seen between the distal ends of the purines; (iii) for 5′-purine-purine-3′ base steps, the purines are typically oriented similarly, but offset so that the five-membered ring of the second purine overlaps the six-membered ring of the first purine. In each case, the base step exhibits an average helical twist of ~33°.

In other short nucleic acid duplexes that include tandem wobble G·U or G·T base pairs (15–17,21,29) some of the same characteristic base stacking patterns are seen, but their sequence dependencies are radically altered. The six-membered rings of pyrimidines and purines tend to stack directly above one another, forming striking discrete clusters of up to 5 bases (Fig. 5). Although six-membered ring stacks of this kind, albeit shorter ones, are indeed common in A-form duplexes, the presence of the G·U mismatches changes the sequence dependence of stack formation from pyrimidine·purine to purine·pyrimidine. In each of the long clusters seen in the 14mer, the nucleus of the cluster is formed by the guanine of a wobble pair in a cross-strand

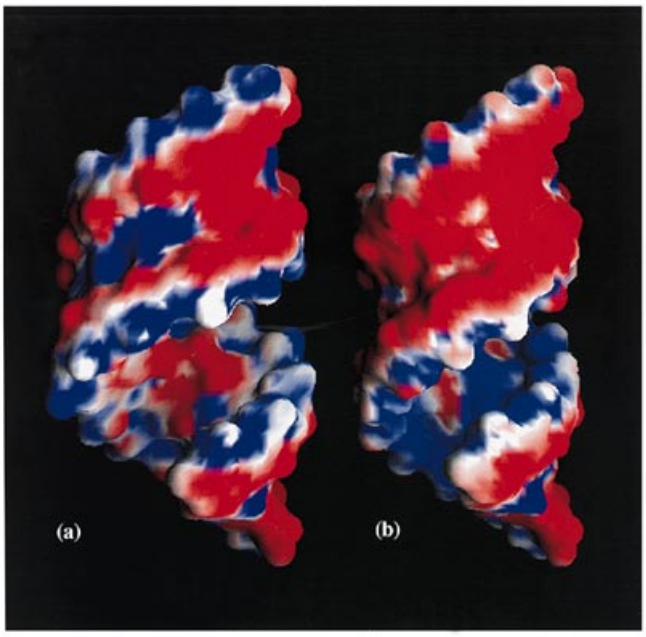


Figure 4. The distribution of partial charges on the surface of the 14mer is significantly altered by the presence of tandem 5′-U-U-3′/3′-G-G-5′ wobble base pairs. (**a**) The 14mer duplex. (**b**) A helix of the same sequence with Watson–Crick base pairs. In the 14mer, the minor groove (top) is narrower and shows an unusual cluster of positive charge due to solvent-exposed nitrogens. The major groove (bottom) is relatively wider, and exposes an additional patch of negative charge due to keto oxygens from the bases. The electrostatic surface was calculated using GRASP (55).

purine·purine stack. Such stacking requires abnormally large or small helical twist angles, and corresponds to a rotation of the base pair about an axis parallel to the helix axis, a rotation that the 'wobble' hydrogen bonding pattern stabilizes. Cross-strand stacking produced a distinctive arrangement in which the six-membered ring of the purine was sandwiched almost perfectly between the six-membered rings of its flanking bases: the purine from the opposite strand on one face, and the pyrimidine from its own strand on the other face.

An analysis of short RNA and DNA duplexes with tandem G·U and G·T mismatches (Fig. 5) shows that the pattern of extended base stacking is directly related to the sequence of the tandem mismatch and the purine·pyrimidine orientation of its flanking Watson–Crick base pairs. Thus, duplexes with symmetrical tandem 5′-U-G-3′/3′-G-U-5′ and 5′-T-G-3′/3′-G-T-5′ sequences (motif I, the most common in rRNA and the most stable) consistently showed cross-strand purine·purine stacking between the two wobble pairs, and same-strand stacking with flanking Watson–Crick base pairs. The base step between the two wobble pairs was invariably under-twisted (with a typical helical twist angle of ≤20°), and base steps between wobble pairs and flanking Watson–Crick base pairs were over-twisted (~40° or more). In contrast duplexes with symmetrical tandem 5′-G-U-3′/3′-U-G-5′ and 5′-G-T-3′/3′-T-G-5′ sequences (motif II, somewhat less common and less stable) never showed cross-strand stacking between the wobble pairs. Due to the purine·pyrimidine orientation of the flanking Watson–Crick base pair, cross-strand stacking with the flanking base pair was possible in the motif II RNA example, but not the DNA example. In motif II, the base step between the wobble pairs was always over-twisted (~50° on average), and the flanking steps were under-twisted (~20°).

It is interesting that the pattern of over- and under-twisted base steps in the 14mer duplex (motif III, Fig. 5e) can be predicted quite accurately by referring to sequentially analogous base steps in the motif I and II structures. The observed pattern of cross-strand and same-strand stacking follows as a consequence of the predicted over- and under-twisting. Thus in the 14mer, over-twisting (~40°) occurs between wobble pair G10·U19 and U11·A18, its flanking Watson–Crick base pair, and between G24·U5 and U25·A4, the symmetry-equivalent base pair step. Base pair steps with analogous sequences (i.e., 5′-G-pyrimidine-3′/3′-U-purine-5′ and 5′-G-pyrimidine-3′/3′-T-purine-5′) are also found in the duplexes containing motif I (15,19,29). At these base steps, the helices are similarly over-twisted (~40°) and same-strand stacking of six-membered rings, with nearly complete overlap, is seen on both strands. Though this stacking mode is common in A-form helices, it is normally associated with 5′-pyrimidine-purine-3′/3′-purine-pyrimidine-5′ base steps instead.

Under-twisting in the 14mer (~20°) occurs between U20·G9 and G21·C8, its flanking Watson–Crick base pair, and between U6·G23 and G7·C22, the symmetry-equivalent base step. Convincing cross-strand purine-purine stacks are seen at both of these base steps (Fig. 5). An analogous base-pair sequence (5′-U-purine-3′/3′-G-pyrimidine-5′) occurs in the motif II RNA example (16), where under-twisting (~20°) also leads to cross-strand purine stacking. Observe that a similar under-twisting (~20°) in the motif II DNA example (21) fails to extend the central base-stacking cluster any further because the flanking Watson–Crick pair has the same pyrimidine·purine orientation as the wobble pair. In that case, cross-strand stacking was precluded by sequence, and same-strand stacking was inconsistent with under-twisting.

The most striking outcome of the comparison is that the configuration of each base step depends almost exclusively on the two base pairs that form the base step, and is independent of the remainder of the sequence. This observation provides a structural explanation for the success of the nearest-neighbor model in predicting the thermodynamic stability of such duplexes (53). The analysis also shows that duplexes containing tandem G·U pairs share certain structural characteristics that distinguish them from A-form helices. Notably, the cross-strand purine stacks, and under- and over-twisted base steps stabilize the formation of discrete base-stacking clusters, each containing two to five bases whose six-membered rings are stacked almost directly above one another (Fig. 5).

Biological significance

The presence of adjacent G·U base pairs affects the structure of the duplex in several ways that might be exploited in specific RNA recognition. The recognition of nucleic acid duplexes could involve direct read-out of the bases, changes in the conformation of the backbone, or the recognition of solvent that is ordered in a sequence-specific way. For example, as noted above, the width of the major groove in the 14mer is increased near the G·U base pairs, as compared to an ideal A-form helix. The proximity of wobble base pairs also affects the helix conformation in nearby Watson–Crick base pairs. Such structural variations due to the presence of wobble base pairs could be features essential for protein-nucleic acid recognition.

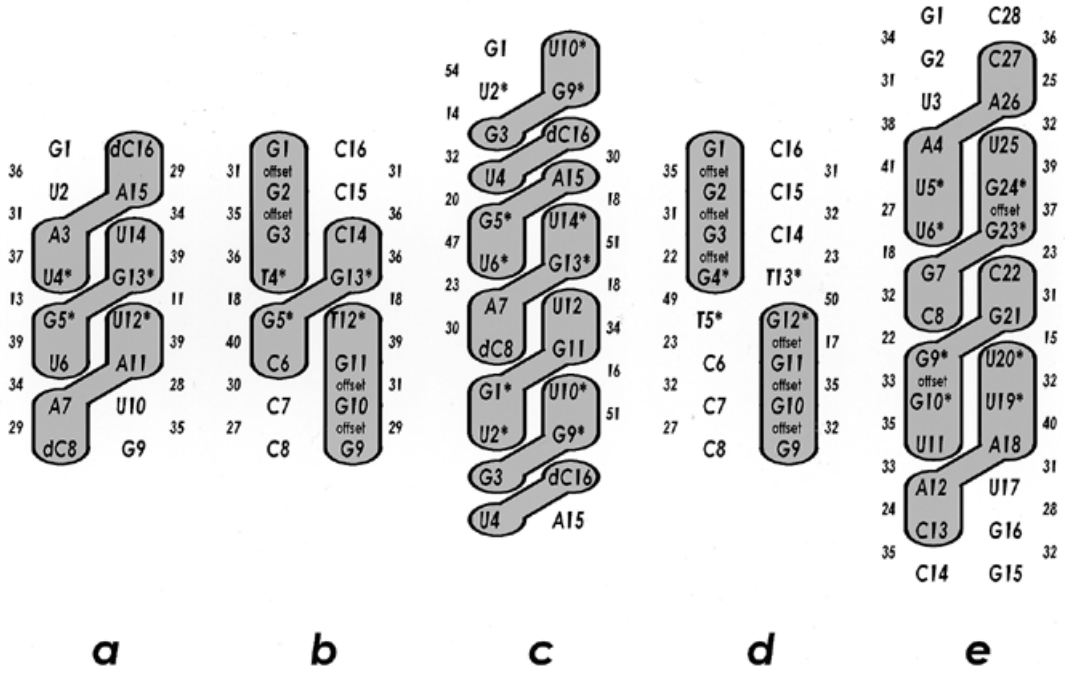


Figure 5. Base stacking patterns in nucleic acid duplexes containing tandem U·G or T·G base pairs. Bases enclosed in a closed curve belong to the same discrete stacking cluster. In these clusters, six-membered rings of purines are sandwiched directly between the six-membered rings of flanking base pairs. In base steps labeled OFFSET, successive purines from the same strand are only partly overlapped, as per the usual purine stacking arrangement in A-form helices. Numerical values in italics are local helical twist angles, as calculated by CURVES 5.1 (54). (**a**) Motif I RNA: Biswas *et al*. (16); (**b**) Motif I DNA: Rabinovich *et al*. (19); (**c**) Motif II RNA: Biswas and Sundaralingam (15); (**d**) Motif II DNA: Kneale *et al*. (21); (**e**) Motif III RNA: 14mer. Note that the helical twist values reported by CURVES consistently differ from published values for the same structures, which were calculated by the methods of (56) (d) and (57) (a and c). At base steps with wobble pairs, atypically large angles in one calculation correspond to atypically small values in the other, and produce opposite assessments of which base steps are under- and over-twisted. It is worth emphasizing that the implication of this figure—the conformational similarity of sequentially analogous base steps—remains valid but is less dramatic when the alternative calculation is used.

Local sequence-specific changes in shape or charge could also be the basis for sequence-specific recognition. Notably, G·U wobble pairing causes a number of ring substituents to project significantly into the major and minor grooves (Figs 2 and 3). The hydrogen bond donors and acceptors presented on the molecular surface are consequently different, and as a result, there is an unusual concentration of negative charge in the major groove, and of positive charge in the minor groove (Fig. 4). These pronounced concentrations of charge are unique to and characteristic of the 5′-U-U-3′/3′-G-G-5′ tandem pairs reported here.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the generosity of M. Zhang, A. Anderson and C. Frederick for their instructions and help in RNA synthesis. This work was supported by NIH grants AI20566 and AI32480 (to J.M.H.) and by the Harvard Center for Structural Biology and Giovani Armenise-Harvard Foundation for Advanced Scientific Research. Coordinates and structure factors for the 14mer duplex have been submitted to the nucleic acid database (NDB accession no. AR0008). J.M.H. is affiliated to the Committee on Higher Degrees in Biophysics, Harvard University.

REFERENCES

1 Stern,S., Weiser,B. and Noller,H.F. (1988) *J. Mol. Biol.*, **204**, 447–481.
2 Crick,F.H.C. (1966) *J. Mol. Biol.*, **19**, 548–555.
3 Gautheret,D., Konigs,D. and Gutell,R.R. (1995) *RNA*, **1**, 807–814.
4 van Knippenberg,P.H., Formenoy,L.J. and Heus,H.A. (1990) *Biochim. Biophys. Acta*, **1050**, 14–17.
5 Golden,B.L., Gooding,A.R., Podell,E.R. and Cech,T.R. (1998) *Science*, **282**, 259–264.
6 Cech,T.R. (1987) *Science*, **236**, 1532–1539.
7 Strobel,A.A. and Cech,T.R. (1995) *Science*, **267**, 675–679.
8 Cate,J.H. and Doudna,J.A. (1996) *Structure*, **4**, 1221–1229.
9 Hou,Y.M. and Schimmel,P. (1988) *Nature*, **333**, 140–145.
10 McClain,W.H. and Floss,K. (1988) *Science*, **240**, 793–796.
11 Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) *Nucleic Acids Res.*, **20**, 5785–5795.
12 Gutell,R.R. (1994) *Nucleic Acids Res.*, **22**, 3502–3507.
13 He,L., Kierzek,R., SantaLucia,J.,Jr, Walter,A.E. and Turner,D.H. (1991) *Biochemistry*, **30**, 11124–11132.
14 Wu,M., McDowell,J.A. and Turner,D.H. (1995) *Biochemistry*, **34**, 3204–3211.
15 Biswas,R. and Sundaralingam,M. (1997) *J. Mol. Biol.*, **270**, 511–519.
16 Biswas,R., Wahl,M.C., Ban,C. and Sundaralingam,M. (1997) *J. Mol. Biol.*, **267**, 1149–1156.
17 Dallas,A. and Moore,P.B. (1997) *Structure*, **5**, 1639–1653.
18 McDowell,J.A., He,L., Chen,X. and Turner,D.H. (1997) *Biochemistry*, **36**, 8030–8038.
19 Rabinovich,D., Haran,T., Eisenstein,M. and Shakkad,Z. (1988) *J. Mol. Biol.*, **200**, 151–162.
20 McDowell,J.A. and Turner,D.H. (1996) *Biochemistry*, **35**, 14077–14089.
21 Kneale,G., Brown,T. and Kennard,O. (1985) *J. Mol. Biol.*, **186**, 805–814.
22 Xia,T., McDowell,J.A. and Turner,D.H. (1997) *Biochemistry*, **36**, 12486–12497.
23 Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996) *Science*, 273, 1678–1685.
24 Kieft,J.S. and Tinoco,I.,Jr (1997) *Structure*, **5**, 713–721.
25 Pyle,A.M. and Green,J.B. (1995) *Curr. Opin. Struct. Biol.*, **5**, 303–310.

26 Zuker,M. (1994) *Methods Mol. Biol*., **25**, 267–294.
27 Baeyens,K.J., Hendrik,L.B. and Holbrook,S.R. (1995) *Nature Struct. Biol*., **2**, 56–62.
28 Baeyens,K.J., Hendrik,L.B., Pardi,A. and Holbrook,S.R. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 12851–12855.
29 Correll,C.C., Freeborn,B., Moore,P.B. and Steitz,T.A. (1997) *Cell*, **91**, 705–712.
30 Dock-Bregeon,A.C., Chevrier,B., Podjarny,A., Johnson,J., de Bear,J.S., Gough,G.R., Gilham,P.T. and Moras,D. (1989) *J. Mol. Biol*., **209**, 459–474.
31 Holbrook,S.R., Cheong,C., Tinoco,I.,Jr and Kim,S.-H. (1991) *Nature*, **353**, 579–582.
32 Leonard,G.A., McAuley-Hecht,K.E., Ebel,S., Lough,D.M., Brown,T. and Hunter,W.N. (1994) *Structure*, **2**, 483–494.
33 Lietzke,S.E., Barnes,C.L., Berglund,A.J. and Kundrot,C.E. (1996) *Structure*, **4**, 917–930.
34 Portmann,S., Usman,N. and Egli,M. (1995) *Biochemistry*, **34**, 7569–7575.
35 Schindelin,H., Zhang,M., Bald,R., Furste,J.-P., Erdmann,V.A. and Heinemann,U. (1995) *J. Mol. Biol*., **249**, 595–603.
36 Wahl,M.C., Rao,S.T. and Sundaralingam,M. (1996) *Nat. Struct. Biol*., **3**, 24–31.
37 Luebke,K.J., Landry,S.M. and Tinoco,I.,Jr (1997) *Biochemistry*, **36**, 10246–10255.
38 Jaeger,J.A., SantaLucia,J.,Jr and Tinoco,I.,Jr (1993) *Annu. Rev. Biochem*., **62**, 255–287.
39 Scaringe,S., Francklyn,C. and Usman,N. (1990) *Nucleic Acids Res*., **18**, 5433–5441.
40 Usman,N., Ogilvie,K.K., Jiang,M.-V. and Cendergren,R. (1987) *J. Am. Chem. Soc*., **109**, 7845–7854.
41 Otwinowski,Z. and Minor,W. (1997) *Methods Enzymol*., **276**, 307–326.
42 Sheldrick,G.M. (1990) *Acta Crystallogr. A*, **46**, 467–473.
43 CCP4. (1994) *Acta Crystallogr. D*, **50**, 760–763.
44 Brünger,A.T. (1992) *X-PLOR v.3.1. A System for X-ray Crystallography and NMR*. Yale University Press, New Haven, CT.
45 Jones,T.A., Zou,J.Y., Cowan,S.W. and Kjeldgaard,M. (1991) *Acta Crystallogr. A*, **47**, 110–119.
46 Jones,T.A. (1985) *Methods Enzymol*., **115**, 157–171.
47 Parkinson,G., Vojtechovsky,J., Clowney,L., Brünger A.T. and Berman,H.M. (1997) *Acta Crystallogr. D*, **52**, 57–64.
48 Adams,P.D., Pannu,N.S., Read,R. and Brünger,A.T. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.
49 Luzatti,P.V. (1952) *Acta Crystallogr*., **5**, 802–810.
50 Doudna,J.A. and Cate,J.H. (1997) *Curr. Opin. Struct. Biol*., **7**, 310–316.
51 Saenger,W. (1984) *Principles of Nucleic Acid Structure*. Springer-Verlag New York Inc., NY.
52 Arnott,S., Chandrasekaran,R., Hukins,D.W.L., Smith,P.J.C. and Watts,L. (1974) *J. Mol. Biol*., **88**, 523–533.
53 Turner,D.H., Freier,S.M. and Sugimoto,N. (1988) *Annu. Rev. Biophys. Chem*., **17**, 167–192.
54 Lavery,R. and Sklenar,H. (1988) *J. Biomol. Struct. Dynam*., **6**, 63–91.
55 Nicholas,A. (1992) *GRASP: Graphical Representation and Analysis of Surface Properties*. Columbia University, NY.
56 Dickerson,R.E. and Drew,H.R. (1981) *J. Mol. Biol*., **149**, 761–786.
57 Bhattacharyya,D. and Bansal,M. (1990) *J. Biomol. Struct. Dynam*., **8**, 539–572.