The packaging signal of HIV-1 RNA contains a stem-loop structure, SL1, which serves as the dimerization initiation site for two identical copies of the genome and is important for packaging of the RNA genome into the budding virion and for overall infectivity. SL1 spontaneously dimerizes via a palindromic hexanucleotide sequence in its apical loop, forming a metastable kissing dimer form. Incubation with nucleocapsid protein causes this form to refold to a thermodynamically stable mature linear dimer. Here, we present an NMR structure of the latter form of the full-length SL1 sequence of the Lai HIV-1 isolate. The structure was refined using nuclear Overhauser effect and residual dipolar coupling data. The structure presents a symmetric homodimer of two RNA strands of 35 nucleotides each; it includes five stems separated by four internal loops. The central palindromic stem is surrounded by two symmetric adenine-rich 1–2 internal loops, A-bulges. All three adenines in each A-bulge are stacked inside the helix, consistent with the solution structures of shorter SL1 constructs determined previously. The outer 4-base pair stems and, proximal to them, purine-rich 1–3 internal loops, or G-bulges, are the least stable parts of the molecule. The G-bulges display high conformational variability in the refined ensemble of structures, despite the availability of many structural restraints for this region. Nevertheless, most conformations share a similar structural motif: a guanine and an adenine from opposite strands form a GA mismatch stacked on the top of the neighboring stem. The two remaining guanines are exposed, one in the minor groove and another in the major groove side of the helix, consistent with secondary structure probing data for SL1. These guanines may be recognized by the nucleocapsid protein, which binds tightly to the G-bulge in vitro.

As with all retroviruses, two identical copies of genomic RNA are packaged in the viral particles of HIV-1. The non-covalent dimer formed by these two RNA strands is involved in several stages of the viral life cycle, such as reverse transcription, recombination, RNA packaging; its integrity is important for optimal viral infectivity (reviewed in Ref. 1). The packaging signal is located in the 5′ untranslated region of unspliced RNA; it contains four stem-loop structures, SL1, SL2, SL3, and SL4 (Fig. 1A). Three of them, SL1, SL2, and SL3, are specifically recognized in vitro by the NC domain of the viral Gag polyprotein (2), a key component of RNA packaging. SL1 is a 35-nt hairpin; its apical loop contains a 6-nt self-complementary sequence and three flanking residues, most commonly purines. This hairpin was identified as a dimerization initiation sequence (3, 4); RNA constructs containing SL1 are capable of spontaneously dimerizing in vitro via loop-loop interaction, forming metastable kissing dimer structures. Incubation of kissing dimers at 55 °C converts them into a more stable linear, or mature, dimeric form with 1–2 internal loops, or A-bulges, being formed on both flanks of the palindromic sequence (Fig. 1B). A similar effect is achieved by incubation of kissing dimers at 37 °C together with NC protein, in a process possibly reflecting maturation of genomic RNA in vivo (reviewed in Ref. 1).

Another feature of SL1 is a very conserved G-rich 1–3 internal loop (G-bulge) at the bottom of the hairpin, important for the RNA packaging. The sequence of the G-bulge is not essential for the dimerization step (4), but it is important for the packaging and overall infectivity (5). NC protein binds tightly in vitro to both SL1 sites, central palindromic with flanking A-bulges and the G-bulge (2). Delineating structure of SL1 is important for understanding dimerization and packaging of HIV-1 RNA, and it is potentially valuable for structure-based development of therapeutic drugs interfering with these processes. Several structures of the palindromic moiety of SL1 surrounded by the A-bulges have been solved for shorter RNA constructs, i.e., not including the G-bulge, for both kissing (6–8) and linear dimeric isomers (8–11). Structures for the G-bulge have been determined in monomeric constructs that excluded the palindromic sequence and the A-bulges (8, 12, 13). In addition, the full SL1 RNA was modeled using the two shorter structures (8) determined by NMR. Here, we present the NMR structure for the full-length linear dimer of SL1, which has a wild-type sequence of the Lai HIV-1 isolate (subtype B) with the exception of inversions in the terminal and penultimate base pairs (Fig. 1C). The structure has been determined with the help of NOE distance and residual dipolar coupling (RDC) restraints.
**EXPERIMENTAL PROCEDURES**

**Sample Preparation**—35-nt RNA (SL1) was synthesized by *in vitro* transcription with T7 RNA polymerase (14, 15), purified by PAGE and electroelution, and dialyzed. The sample was heated to 90 °C for 3 min and slowly cooled to room temperature at an RNA concentration of \(\frac{1}{10} \text{M}\), then concentrated to 0.75 mM. The formation of the linear isomer of the SL1 dimer was verified by PAGE as described (16). The NMR sample had RNA at concentration of 0.75 mM in 300 mM potassium phosphate buffer with 250 mM NaCl and 0.1 mM MgCl\(_2\) at pH 6.4, either in D\(_2\)O or in a 90/10% H\(_2\)O/D\(_2\)O mixture. A uniformly doubly 13C,15N-labeled SL1 sample was prepared in a similar way but starting with *in vitro* transcription using a mixture of 13C,15N-labeled nucleoside triphosphates (Silantes, München, Germany).

For the measurement of RDC constants, RNA was weakly aligned with the filamentous phage Pf1 as described (17), so that deuterium quadrupolar splitting was 9 Hz. Attempts to align SL1 RNA with \(n\)-alkyl poly(ethylene glycol) C\(_{12}\)E\(_6\) led to severe deterioration of NMR signals (data not shown).

**NMR Spectroscopy**—All NMR spectra were acquired either on Varian Inova spectrometers operating at 600 MHz (for protons) equipped with a cryogenic or a conventional probe at the University of California at San Francisco or on a Varian Unity Inova 900 MHz spectrometer equipped with a conventional probe at National Magnetic Resonance Facility, Madison, Wisconsin. Spectra were processed with nmrPipe/nmrDraw (18) and analyzed with SPARKY (19). Spectra in D\(_2\)O were acquired at 30 °C, and spectra in H\(_2\)O were acquired at 10 or 15 °C. Quadrature detection in the indirect dimensions of multidimensional experiments was achieved using the States-TPPI (time-proportional receiver phase incrementation) method. Water suppression in two-dimensional NOESY experiments in H\(_2\)O was achieved using either the 1-1 echo (20) or a symmetrically shifted shaped pulse (21). Spectra in D\(_2\)O were recorded without any suppression of the residual HDO signal. Decoupling of 13C during acquisition was achieved using GARP and WURST80 decoupling schemes on the 600- and 900-MHz spectrometers, respectively.

Resonances were assigned as described previously (22) with the help of homonuclear two-dimensional NOESY and TOCSY spectra, three-dimensional 13C-edited NOESY-HMQC (acquired separately with optimization for aromatic and ribose correlations) and HCCH-TOCSY spectra, and two-dimensional versions of triple-resonance HCN spectra. NOE intensities for the determination of interproton distances were integrated by line fitting in homonuclear two-dimensional NOESY spectra acquired in isotropic and liquid crystalline solutions. To minimize overlap, these experiments were set up such that only one multiplet component appeared in each spectrum; the two spectra containing the bottom-right and bottom-left components in the non-decoupled constant time \(^1\text{H}-^{13}\text{C}\) TROSY-HSQC spectra (23) acquired in isotropic and liquid crystalline solutions. To minimize overlap, these experiments were set up such that only one multiplet component appeared in each spectrum; the two spectra containing the bottom-right and bottom-left components in the non-decoupled constant time \(^1\text{H}-^{13}\text{C}\) TROSY-HSQC spectra (23) acquired in isotropic and liquid crystalline solutions. To minimize overlap, these experiments were set up such that only one multiplet component appeared in each spectrum; the two spectra containing the bottom-right and bottom-left components in the non-decoupled constant time \(^1\text{H}-^{13}\text{C}\) TROSY-HSQC spectra (23) acquired in isotropic and liquid crystalline solutions. To minimize overlap, these experiments were set up such that only one multiplet component appeared in each spectrum; the two spectra containing the bottom-right and bottom-left components in the non-decoupled constant time \(^1\text{H}-^{13}\text{C}\) TROSY-HSQC spectra (23) acquired in isotropic and liquid crystalline solutions.

**Structural Restraints**—Bounds for interproton distances were calculated based on two-dimensional NOE intensities using a complete relaxation matrix method, MARDIGRAS (24), with the RANDMARDI analysis of experimental errors in the intensities (25), as described previously (22). No correction of NOE intensities was made due to a potentially strong anisotropy of the RNA molecule. Instead, relative NOE errors, used as input for RANDMARDI, were increased to 20%. MARDIGRAS calculations were run for a series of correlation times. In the end, a correlation time was selected that best reproduced fixed and nearly fixed distances (such as H1'-H2' or H1'-H3'). Sugar conformations were restrained to C2'-endo for residues with observed H1'-H2' correlation in TOCSY spectra, as explained below; the remaining sugar...
Linear Dimer of SL1 RNA

conformations were restrained to C3'-endo. All RDC restraints were used as measured in the non-decoupled HSQC spectra, with error bars of ±2 Hz. Idealized hydrogen bond distance restraints were used for all Watson–Crick GU and GA pairs. All glycosidic torsion angles were restrained to the anti-conformation, except G31, which was restrained to the syn-conformation in some calculations, as explained below. Idealized A-type backbone torsion angle restraints were used for stem residues during DYANA calculations. In addition, “non-NOE” distance restraints (lower bounds of 5 Å) were used for proton pairs that did not have any intensity in any of the 2- or 3-dimensional NOESY data sets but that tended to have short distances in preliminary rounds of structure refinement.

Structure Calculations—The NMR structures for SL1 RNA were calculated in two stages, using DYANA (26) and miniCarlo (27). In the first stage, DYANA was used to perform simulated annealing with 20,000 steps of torsion angle dynamics starting with 100 randomized conformations. The weights of NOE-derived distance restraints, non-NOE restraints, and hydrogen bond restraints were 1.0, 1.0, and 5.0, respectively. The weights of sugar pucker-related torsion and bond angle restraints and of glycosidic angle restraints were 10.0, and the weights of backbone torsion angle restraints in stems were 20.0. Additional distance restraints were used to ensure proper closure of the five-member furanose rings in DYANA as described (22). No RDC restraints were used in the DYANA calculations. In the second stage, the 30 best DYANA structures were passed to miniCarlo for further refinement. The miniCarlo program is specialized software for modeling, energy minimization, and Metropolis Monte Carlo simulation of nucleic acids. It uses generalized helical parameters of nucleic bases as internal coordinates. These parameters define relative positions of idealized aromatic bases and sugars, as well as internal conformations of sugar moieties. The atomic coordinates of the sugar–phosphate backbone are then calculated by a chain closure algorithm (28). The force field used in miniCarlo was optimized specifically for nucleic acids (29). To pass the DYANA structures to miniCarlo, helical parameters were calculated from the atomic coordinates as described (30). The structures were then refined using restrained Metropolis Monte Carlo simulation annealing followed with restrained minimization (31, 32). At first, the structures were restrained-minimized, followed by 10,000 Metropolis steps at 1500 K with the weight of restraint energy gradually increasing from zero to its full term. The structures were again restrained-minimized and another Metropolis simulation run for 20,000 steps at 200 K with the full weight of restraints. The structures calculated during this low temperature segment of the Monte Carlo chain were averaged based on the helical parameters (31), and the average structure was again restrained-minimized. The above protocol was applied to the fragment of SL1 comprising residues 1–18 and 17*–35* (Fig. 1), i.e., to one-half plus 1 base pair of the symmetric homodimer. In the end, full homodimer was calculated using a dyad symmetry transformation of helical parameters, and the full SL1 molecule was once more restrained-minimized. Restraints used in miniCarlo calculations included distance restraints as described above, sugar pucker restraints (these were used only during high-temperature Monte Carlo simulations), and RDC restraints. A flat-well potential (33) was used for all restraints. The full force constants for distance and RDC restraint energy terms were 10.0 kcal/(mol-Å²) and 0.2 kcal/(mol-Hz²), respectively. RDC constants are calculated in miniCarlo similarly to AMBER (34, 35), using a symmetric traceless alignment tensor, which is optimized during the refinement.

Molecular graphics were prepared with UCSF Chimera (36) and MidasPlus (37); molecular surface representation was calculated with

the MSMS package (38) running from Chimera; the helical axis was calculated with CURVES (39).

The refined structures, chemical shifts, and NMR restraints will be deposited to PDB (code 2GM0) and BMRB.

RESULTS

Resonance Assignments and Analysis of NOE Data—Imino protons in SL1 RNA were assigned (summarized in Fig. 2) based on two-dimensional NOESY spectra acquired at 10 and 15 °C in water. Imino protons of U11 and G23 had a strong cross-peak, characteristic of a wobble base pair (data not shown). No cross-peaks were observed for imino protons in pairs G1–C35*, G4–C32*, C6–G28*, and G12–C22*; because of that, these imino protons were not assigned in a sequence-specific manner. These base pairs are located at the ends of stems 1 and 2 (Fig. 1), and the exchange with solvent is presumably increased due to end fraying. Nevertheless, these base pairs, even though destabilized, are apparently formed as deduced from the connectivities observed in two-dimensional NOESY spectra acquired in D₂O. These connectivities are characteristic of an A-type helical conformation in each of three stems, including the ends of the stems. In particular, very strong sequential H²-H6/H8 cross-peaks were observed even at a low mixing time of 15 ms (supplemental data) for residues G1–A2, U34–C35 (consistent with the formation of the G1–C35* pair), C3–G4, C32–G33 (G4–C32* pair), C6–U7, A27–G28 (C6–G28* pair), and U11–G12, C22–G23 (G12–C22* pair). In addition, strong cross-strand and sequential cross-peaks involving adenine H₂ protons also indicated an A-type helical geometry (e.g. H2A2–H1*C35*, H2A27–H1*G28). Interestingly, G15 at the end of stem 3 manifests a strong imino proton signal (Fig. 2B), so that palindromic stem 3 appears to be the most stable of the three stems. In contrast, the outer stem 1 appears to be the least stable; imino protons of U34 and G33 were relatively weak, and U34 exhibited species (Fig. 2) suggesting that alternative conformers exist for this stem exchanging slowly on the NMR time scale. Signals, implying multiple species for the residues of stem 1, were observed for non-exchangeable protons as well. Also of note is that imino proton resonances are not very strong in U7 and U8 (Fig. 2). These residues form a UU:AA stretch in stem 2, separated by one GC base pair from the G-bulge; a diminished intensity of the imino proton signals indicates a somewhat increased flexibility of this end of stem 2.
As with other work (12, 13), the imino proton of G5 was not observed in spectra acquired in water due to rapid exchange with solvent. However, D$_2$O NOESY data are consistent with the formation of G5-A29* base pair stacked on the top of stem 3, in particular the observation of strong cross-peaks C6H1'-A29*H2 and G28H2'-A29H8.

Unexpectedly, presaturation of the residual HDO signal in D$_2$O spectra led to a severe diminution in signal-to-noise ratio. This effect is likely explained by an extensive hydration of this RNA and a very effective dipole-dipole relaxation expected for a molecule of this size (70 nucleotides). Indeed, strong cross-peaks between aromatic protons and water were observed in two-dimensional NOESY spectra acquired in H$_2$O (data not shown). Because of this effect, all spectra in D$_2$O were acquired without any suppression of the residual HDO signal.

Broad line widths were observed for non-exchangeable protons, especially in labeled samples, due to the added $^{13}$C-$^1$H relaxation mechanism in such samples (Fig. 3). Because of that, signal-to-noise ratio was very low in three-dimensional carbon-selected spectra; only strong cross-peaks were observed in three-dimensional NOESY. Consequently, most sequence-specific assignments were made using homonuclear spectra acquired with unlabeled samples; heteronuclear spectra were mostly used to confirm the sequence-specific proton assignments obtained by homonuclear methods and to obtain the $^{13}$C assignments. To aid with the assignments, a series of two-dimensional NOESY spectra with a range of mixing times covering more than an order of magnitude were acquired. A data set with short mixing time (15 ms) was very helpful for assigning H2'/H1' cross-peaks (supplemental data). A data set with long mixing time (300 ms) provided a multitude of cross-peaks near anomeric (Fig. 4) and aromatic (supplemental data) diagonals. These cross-peaks correspond to relatively long interproton distances.
Linear Dimer of SL1 RNA

(5.5–6.5 Å); they are weak or not observed at lower mixing times. A long mixing time has an additional advantage for these spectral regions, because diagonal peaks manifest decreased intensity and less overlap with the cross-peaks. All aromatic, H1′ and H2′ protons were assigned; H3′, H4′, H5′ and H6′ protons were assigned for a subset of residues.

Fingerprint connectivities between H1′ and aromatic protons were observed at medium and high mixing times for all residues without exception (supplemental data). However, these NOEs are always affected by spin diffusion, which is expected to be very significant for a molecule of this size. As a rule, these cross-peaks were not observed at a short mixing time of 15 ms. One notable exception was the H1′-H8 cross-peak in G31, which was observed at 15 ms, suggesting that either this residue has a syn-conformation, or the syn- and anti-conformations are in equilibrium for G31. In contrast, sequential H2′(i)-to-aromatic proton (i+1) distances are very short in A conformations, and the corresponding NOEs are less affected by spin diffusion. All these NOEs were observed as very strong cross-peaks at a mixing time of 15 ms for all three stems of SL1 (supplemental data). In addition, G12H2′-A13H8 and C20H2′-A21H8 were observed as strong cross-peaks at the junctions of the A-bulge. At the same time, H2′(i)-H6/H8(i+1) connectivities were observed as relatively weak cross-peaks or were even absent at 15 ms for dinucleotide steps G4-G5, G5-C6, A13-A14, A14-A15, A29-G30, G30-G31, and G31-C32. All these steps are associated with one of the two bulges of SL1, suggesting a non-A conformation of the bulges. Apart from the species multiple for stem 1, each nucleus had a single exception (supplemental data). However, these NOEs are always presented a problem, because only the major species were systematically assigned. The apparent intensities of NOEs were therefore diminished for this stem, because the fractional population of the major species was less than 1.0. To estimate this population, the intensities of pyrimidine H5-H6 peaks were compared between stem 1 and the rest of the molecule. This population was estimated as 0.61, consistently for all mixing times (standard deviation: 0.04). Therefore, the NOE intensities were scaled by a factor 1.0/0.61 for the residues of stem 1 prior to MARDIGRAS calculations; the calculated distances correspond to the distances in the major conformer of stem 1.

As explained under “Experimental Procedures,” MARDIGRAS was run with a series of effective isotropic correlation times in the range between 6 and 48 ns. Calculations with correlation times 30 and 32 ns best reproduced the fixed and nearly fixed distances; for results of these two correlation times were pooled together to produce a set of distance restraints. For the first MARDIGRAS calculations, the initial SL1 structure, also required as MARDIGRAS input, was model-built and energy-minimized using miniCarlo. After preliminary rounds of refinement, the resulting structures were used as new input structures for MARDIGRAS to calculate a new set of interproton distance bounds; the whole procedure was repeated several times. The rationale for these iterations is that there is some residual dependence of MARDIGRAS distances on the initial structure, especially for long correlation and mixing times. For the same reason, the NOE data at a high mixing time of 300 ms were not used in the first round of MARDIGRAS calculations; they were introduced at later iterations. Distances calculated from the 300-ms NOESY data set were used as restraints only if they were not available at lower mixing times. To assess the accuracy of bounds for long distances obtained from the 300 ms data set, the intra-residue adenine H2-H8 distance was used, which has a fixed value of 6.39 Å. There were four residues where the H2-H8 cross-peak was observed at 300 ms: A13, A14, A21, and A26. The distance bounds calculated with MARDIGRAS for these residues were 5.58–7.65, 5.14–7.44, 5.83–7.88, and 4.84–7.35 Å, respectively. In each case, the bounds covered the exact value of 6.39 Å. Even though quite wide, these bounds were accurate, which justified using the long mixing time NOE data with valuable information about long distances (e.g. Fig. 4). Altogether, 294 distance bounds were calculated with MARDIGRAS (588 per SL1 dimer). Although this number is not very great for a molecule of this size, most of these restraints are inter-residue (Table 1), and they are distributed well among the residues (supplemental data); both are important for a better definition of structure.

Residues A13, A14, G30, and G31 exhibited strong H2′-H1′ correlations in the homonuclear two-dimensional TOCSY spectrum; consequently, their riboses were restrained to the Southern-type sugar pucker (pseudorotation phase angle 90–180°). Residues G12 and A29 exhibited medium H2′-H1′ correlations, consistent with flexible sugar conformations. To make sure that all parts of conformational space are explored for these sugars, they were restrained to the southern conformations in the following way: 2′-endo is preferred for G12 and A29. This flexible restraint was especially necessary for the synthesis of the triplex (20).

### Table 1

| restraint type | intra-residue | inter-residue | total | average flat well width |
|----------------|---------------|---------------|-------|-------------------------|
| distance bounds | 168           | 418           | 586   | 2.31 Å                  |
| H-bond restraints | 104           |               |       |                         |
| non-NOE restraints | 22            |               |       |                         |
| RDC restraints | 130           |               |       |                         |
| total restraints | 842           |               |       |                         |
| restraints per residue | 12.0         |               |       |                         |
| 
| Total energy | −332.0 ± 18.7 kcal/mol |
| Conformational energy | −520.1 ± 16.1 kcal/mol |
| Restrained energy | 188.1 ± 11.1 kcal/mol |
| Residual distance restraints deviation | 0.041 ± 0.003 Å |
| Residual RDC restraints deviation | 1.5 ± 0.1 Hz |

### Table 2

| Residue | Run 1 | Run 2 | Run 3 | Run 4 | Run 5 |
|---------|-------|-------|-------|-------|-------|
| G12     | 2′-endo | 2′-endo | 3′-endo | 2′-endo | 2′-endo |
| A29     | 2′-endo | 2′-endo | 3′-endo | 3′-endo | 3′-endo |
| G31     | syn    | anti  | syn   | syn   | anti  |
some refinement runs, and to Northern conformations (pseudorotation phase angle 0–90°) in other runs (Table 2). Terminal residues G1 and C35 exhibited weak H2’-H1’ correlations, and the rest of the residues did not have any H2’-H1’ cross-peaks; all these residues were restrained to the Northern sugar conformations. Residue G31 had a relatively strong H1’-H8 cross-peak in the 15 ms NOESY data set. The corresponding distance bounds calculated with MARDIGRAS were 2.83–3.51 Å, which is shorter than in anti-conformations (typically 3.9 Å) but longer than in syn conformations (2.5 Å). Correspondingly, the glycosidic torsion angle in G31 was restrained to anti-conformations in some refinement runs and to syn in others. Glycosidic angles in the rest of the residues were restrained to anti-conformations. Altogether, five different refinements series were carried out with different combinations of restraints for residues G12, A29, and G31 (Table 2); these restraints were used in DYANA calculations and during the high temperature segment of miniCarlo refinement but not in the final restrained minimization (see “Experimental Procedures”).

Sixty-five RDC C-H constants were measured for the SL1 RNA (130 per SL1 dimer). Despite using a relatively low amount of the Pfu phage for the alignment of RNA (see “Experimental Procedures”), the measured RDC constants had quite large amplitude; they varied in the range $10^2$–$10^3$ Hz, which is shorter than in anti-conformations (typically $10^2$ Hz) but longer than in syn conformations ($10$ Hz). In the rest of the residues, the RDC constants were restrained to anti-conformations in some refinement runs and to syn in others. Glycosidic angles in the rest of the residues were restrained to anti-conformations. Altogether, five different refinements series were carried out with different combinations of restraints for residues G12, A29, and G31 (Table 2); these restraints were used in DYANA calculations and during the high temperature segment of miniCarlo refinement but not in the final restrained minimization (see “Experimental Procedures”).

Analysis of Structures—Structures refined in each of the five runs (Table 2) were sorted according to their total energy, and 10 best structures are discussed here. The refined structures have low average residual distance restraints and RDC restraints deviations (Table 1). However, there are three individual distance restraints violated by more than 1 Å (A13H2-A14H1’ in structures 2, 3, 7, 8, and 10; A29H2-G30H1’ in structures 6, 9, and 10; A29H2-G31H1’ in structure 7), and two RDC restraints violated by more than 5 Hz (A29 C8-H8 in structures 1, 5; A29 C2’-H2’ in structure 6). Note that all violated distances correspond to relatively strong H1’ to adenine H2 cross-peaks, which resulted in relatively tight distance bounds determined by MARDIGRAS (3.21–4.09, 3.30–4.15, and 3.32–5.00 Å for the three violated distances, respectively). One explanation for these violations is some error in experimental distance bounds, which can be due to a strong anisotropy of this RNA. Another explanation is possible dynamics in the bulge regions of SL1. Indeed, restraint violations are localized mostly in the G-bulge and some in the A-bulge of SL1. If multiple conformers are present, exchanging rapidly at the NMR time scale, the average apparent NOE distances may be inconsistent with any single conformation (43). Besides, they may be also inconsistent with average RDC constants, because NOE distances and RDC constants have different averaging properties. Test calculations showed that all violated distance restraints could be satisfied by increasing the distance force constant during the refinement but only at the expense of severe deterioration of structure quality, leading to a significant increase in conformational energy (data not shown).

As explained above, variants of sugar pucker and glycosidic angle restraints were used in the high temperature stage of the refinement for residues G12, A29, and G31 (Table 2). In the 10 highest scoring structures, there are five conformations with G31 anti and five with G31 syn; there are six structures with A29 C3’-endo, three with A29 C2’-endo (sugar pseudorotation phase angle P between 144 and 180°), and one with A29 C1’-exo ($\phi = 117°$). This means that for both the residues, the existing experimental data are consistent with both types of conformations, and both are likely to be present in solution. In contrast, no G12 C2’-endo sugar puckers were observed in the top ten structures. Nine out of ten structures have G12 in an intermediate C4’-exo conformation ($\phi = 62 ± 7°$), and one has G12 C3’-endo ($\phi = 26°$). C4’-exo is a suboptimal sugar pucker; it is possible that this residue is in a C3’-endo-C2’-endo equilibrium, similarly to A29, and the observed C4’-exo conformation is a compromise between conflicting restraints. However, all restraints are satisfied for this residue, including distances H8-H3’, H8-H2’, H1’-H4’, which are sensitive to sugar pucker (44), and RDC C2’-H’2, which is also sensitive to sugar pucker (45). Therefore, it is likely that C4’-exo is an equilibrium conformation of G12, which connects the wobble pair U11:G23 and the A-bulge.

Refined SL1 structures have an elongated shape (Figs. 5 and 6) with two symmetry-related bends distributed between the G-bulge and stem 2. Because the bends are separated by approximately one and a half turns of RNA double helix, the bends are in the opposite direction, so that the axis is roughly planar. Despite the presence of four bulge regions (two symmetrically related G-bulges and two A-bulges), the overall morphology of SL1 is typical of A-RNA, with the deep and mostly narrow major groove and shallow minor groove; however, the major groove becomes noticeably wider near the G-bulge (Fig. 6B). The shape of the molecule is very consistent among all ten refined structures (Fig. 5), with the average overall pairwise atomic r.m.s.d. 2.9 ± 0.9 Å (Table 1). This value is a quite low for an elongated molecule of this size, which was achieved due to using RDC restraints, which define the orientations of C-H vectors (17, 46). Test calculations without RDC restraints led to structures where the degree and direction of bends varied considerably (data not shown). The exact dimensions of SL1 were less consistent among the 10 refined structures. The end-to-end distance was 96.2 ± 3.9 Å (measured along the axis using CURVES) for the 10 structures, but there were two outliers: structures 5 and 7, with the end-to-end distances of 102.0 and 87.6 Å, respectively. While the RDC restraints were mostly responsible for the shape of SL1, distance restraints are likely to contribute predominantly in the definition of the SL1 dimensions, and distance restraints are not very precise for a molecule of this size (average difference between the upper and lower distance bounds is 2.3 Å, Table 1).

Both internal stems of SL1, stem 3 and stem 2, are each defined to a high degree, with a r.m.s.d. below 1 Å (Table 1). All three “flanking” adenines of the A-bulge, A13, A14, and A21, are situated inside the helix, with a partial stacking interaction between each other and with neighboring base pairs. However, there are variations in the relative position of A13 and A14 (Fig. 7), which is correlated with variations in

FIGURE 5. Superposition of 10 best scoring NMR structures of SL1. Here and in other figures, asterisks are used to label residues from the second strand; adenines A13, A14, A21, and A29 are shown in red, wobble Gu pairs are shown in green, G5 in blue, G30 in orange, and G31 in magenta. Also, all sugar ring O4’ atoms are colored yellow.
the distance A13H2-A14H1' (see above). Because of that, the central region of SL1, stem 2 + A-bulge + stem 3 + A-bulge + stem 2, is defined less precisely, with an average r.m.s.d. of 1.8 Å. The terminal base pair of stem 3, G15:C20*, is defined very precisely, but the terminal pair of the stem 2 at the other side of the A-bulge, G12:C22*, is partially disordered and distorted (Fig. 7). This is correlated with the presence of a strong imino proton resonance in water spectra for G15 and with the absence of such resonance for G12 (see above).

Stem 1 and G-bulge are the lesser defined portions of SL1 RNA (Fig. 8); their average r.m.s.d. values of ca. 2 Å are even greater than the r.m.s.d. of the entire central region (Table 1). It is interesting that base pairs flanking the G-bulge, G4:C32* and C6:G28*, are also noticeably disordered and distorted (Fig. 8), which is correlated with the absence of strong imino proton signal in G4 and G28 in water spectra. Importantly, the observed variation in the G-bulge conformation is not due to the paucity of restraints. On the contrary, most of the violations in restraints occur in this region (see above); such internal inconsistencies in experimental restraints are expected when multiple conformers exchange rapidly on the NMR time scale (43). Despite this low degree of definition, certain structural motifs are apparent for the G-bulge. In all structures, G5 and A29* form a mismatched base pair stacked on top of stem 2. In nine out of ten structures, G30 and G31 are not stacked with neighboring base pairs. Instead, they are situated on the different sides of the sugar-phosphate backbone, G31 on the major groove side and G30 on the minor groove side. These nine structures exhibit two structural motifs: one with G31 in the anti-conformation (Fig. 9A) and
another with G31 in the syn-conformation (Fig. 9B). Structures with the G31 anti-conformer are additionally stabilized by an H-bond between G4 O6 and the amino group of G31* (shown with orange line in Fig. 9A). Only one out of ten structures displays a continuous partial stacking of A29, G30, and G31 (Fig. 9C). There is one distance restraint, A29H2–G31H1*, incompatible with this conformation (observed distance of ~6 Å with distance bounds 3.32–5.00 Å). Nevertheless, because of the relatively low conformational and total restraint energy, this structure is ranked number seven and is included in the final ensemble. This structure may represent one of conformers of the G-bulge present in solution.

**DISCUSSION**

Because of the importance of SL1 hairpin for the dimerization of the HIV genomic RNA, this structure has been extensively studied, both by NMR and by x-ray crystallography. Several structures have been proposed for various RNA constructs representing both kissing SL1 isoform (6–8) and its linear alternative. Here we will discuss structures of a linear SL1 isomer as well as structures of various constructs, which include the G-bulge.

![Superposition of all 10 structures showing G-bulge and first base pairs of flanking stems.](image)

**FIGURE 8.** Superposition of all 10 structures showing G-bulge and first base pairs of flanking stems.

The palindromic central stem and the A-bulge structure in the linear SL1 dimer have been determined in the context of shorter RNA constructs lacking the G-bulge by NMR (8–10) as well as by x-ray crystallography (11). Apart from subtle differences in the orientation of the “flanking adenosines,” all structures solved by NMR, including the one presented here, are qualitatively similar. They are characterized by a zipper-like arrangement of adenosines, which are located inside the helix. Still, there is a limited variation in the relative orientation of A13 and A14 (Fig. 7), which may reflect solution conformers of the A-bulge and account for some minor differences in published structures. Note that the structures presented here, unlike structures determined of shorter constructs, were refined using RDC restraints, which help define the orientation of residues more accurately but are also more sensitive to the existence of multiple conformers, due to different averaging properties relative to the NOE distances. The crystal structure of the linear SL1 dimer has been solved for the Mal HIV-1 isolate, which has a substitution of guanine for the A14 position compared with the Lai isolate used in the present work (using the nomenclature of Fig. 11). In the crystal structure, G14 and A21* form a mismatched base pair, while A13 is extrahelical. It is not clear if the extrahelical position of A13 is due to crystal packing forces or due to the differences in RNA sequence.

The G-bulge structure has been determined using NMR by three different groups using monomolecular hairpin constructs stabilized by a stable tetraloop: by Lawrence et al. (12) using GAGA tetraloop (PDB 1N8X), by Yuan et al. (13) using UACG tetraloop (PDB 1OSW), and by Baba et al. (8) using UUCG tetraloop (PDB 2D17). Additionally, in constructs for 1N8X and 2D17, stem 1 was stabilized by 2 or 3 ancillary base pairs (there are only 4 base pairs in stem 1 in the wild-type SL1). In the 1OSW construct, stem 1 had 4 base pairs, but all UU and UG pairs were removed from stem 2, to prevent a potential slippage of the G-bulge into an alternative secondary structure (13); this possibly also had an effect of additionally stabilizing the RNA construct. In contrast to the situation with the A-bulge, all NMR structures of the G-bulge, including ours, differ significantly (Fig. 9). In the 2D17 (Fig. 9D) and 1N8X (Fig. 9E) structures, A29, G30, and G31 are continuously stacked, which is reminiscent of minor conformer 7 in our calculations (Fig. 9C). In the 1OSW structure (Fig. 9F), G30* is stacked on top of the G5:A29* mismatch, while G31* is extrahelical, located from the minor groove side of...
the helix. In most structures refined in the present work, both G31 and G30 are extrahelical, but G31 is always located in the major groove and G30 in the minor groove. To check if distance and RDC restraints we obtained in this work are consistent with other structures shown in Fig. 9, we modeled each of the three conformations of the G-bulge, 2D17, 1N8X, and 1OSW (in each case, structure 1 of the ensemble deposited in the PDB) into the first structure calculated by us (data not shown). In each of the three cases, distance A29H2-G31H1’ was violated by 2.7–4.2 Å (it was also violated in our structure 7 by 1 Å, see above). Most importantly, many RDC restraints were also violated for the bulge residues in these three deposited structures; for example, RDC G30 C8-H8 was violated by 19–64 Hz, and RDC G30 C1’-H1’ was violated by 12–38 Hz. Conversely, we checked if the structural restraints deposited with structures 1N8X and 1OSW are consistent with the structures calculated in this work (there are no structural restraints deposited with structure 2D17). Some of the deposited restraints were consistently violated for the G-bulge residues, in particular, sequential distances A29H2-G30H8, G30H2-G31H8, and G31H2’-C32H6. Apparently, these distances corresponded to strong NOE cross-peaks, a signature of continuous stacking; for example, they were assigned an upper bound of 2.7 Å in the 1N8X restraints, similar to equivalent distances in helical regions. As discussed above, A29H2-G30H8 and G30H2-G31H8 are below the noise level in the 15 ms NOESY data set of our SL1 construct, and the G31H2’-C32H6 cross-peak is significantly weaker than equivalent peaks in helical regions.

This analysis confirms that indeed, the G-bulge conformation in our SL1 construct is significantly different from other published structures. The most likely reason for this appears to be the difference in sequence of RNA constructs. As mentioned above, 1N8X, 1OSW, and 2D17 constructs were stabilized compared with the wild-type SL1 by using an extra-stable tetraloop, lengthening stem 1, or by removing a potentially flexible UU:AA stretch from stem 2. Apart from the inversions of the terminal and penultimate base pairs in stem 1, our SL1 construct has a wild-type sequence, which led to low stability and multiple conformers for stem 1, to a significant loosening of base pairs at each end of stem 2, and to an apparent flexibility of the G-bulge. While conformers with continuous stacking of purines (such as in Fig. 9C) cannot be ruled out, the predominant conformations of the G-bulge appear to have extrahelical G30 and G31. It is noteworthy that another G-bulge construct was studied by Greatorex et al. (47); it had stem 1 of 4 base pairs and a shortened stem 2, which nevertheless had a UU:AA stretch 1 base pair apart from the G-bulge. The NMR line widths for the bulge in this construct were broad to an extent that precluded its structure determination (47).

Baba et al. (8) have also presented a full-length SL1 structure, which included all stems and bulges. However, the structure determination was carried out on two shorter overlapping constructs based on NOE distance restraints. One included stem 1, G-bulge, and stem 2, stabilized by UUGC tetraloop. Another was a homodimer consisting of stem 2, A-bulge, and palindromic stem 3. The full SL1 RNA was then modeled using the two shorter determined structures; this divide-and-conquer approach has typically proven to be quite useful in NMR structure determinations. In the present work, we have directly determined the full-length wild-type SL1 structure based on NOE distances and RDC restraints. The overall shapes of the two structures are distinctly different: the structure determined by Baba et al. (8) has a single bend in the center of the palindromic stem 3, while the structure determined here has two bends associated with two symmetry-related G-bulges (Fig. 6). We attribute this difference to different types of restraints used in the refinements. NOE distances can define accurate local geometries, but they are notoriously inaccurate in defining the overall shape of elongated molecules due to error propagation. RDC restraints, due to their long range nature, are beneficial for defining both the local geometries and the overall shape of molecules. The bend they have observed in the middle of the palindromic duplex, which we do not observe, could conceivably be attributed to the Mal sequence they employed in their experiments (GUGCAC) while we used the Lai sequence (GCGCGC).

The relatively low stability of stem 1 and the associated flexibility of the G-bulge have an apparent biological role in the two-step dimerization of the HIV genome. Overstabilization of the SL1 kissing dimer, by either extending stem 1 or deleting the G-bulge, negatively affects the ability of the NC protein to convert the SL1 kissing dimer into its linear isomorph in vitro. However, the high degree of sequence conservation of the G-bulge suggests that it has roles other than mere destabilization of the SL1 hairpin. Indeed, it has a role in the packaging of genomic RNA separate from its dimerization, as the mutation changing each purine of the G-bulge to pyrimidines reduces the packaging effectiveness to 20% and overall infectivity to 10% of the wild-type sequence, comparable with a complete deletion of SL1 (5). At the same time, this mutation does not affect the ability of RNA to form kissing dimers (4) or the ability of NC to convert kissing dimers to linear dimers in vitro. Fluorescence measurements showed that NC protein binds to the G-bulge with a dissociation constant of 140 nM, about 5–6 times less tightly than to the apical loops of SL2 and SL3 (2). The NMR structures of NC complexes with SL2 and SL3 revealed that the second and fourth guanines of flexible apical loops (GGUG and GGAG, respectively) are specifically recognized by zinc fingers of NC (48, 49). These guanine residues are exposed in the unbound structures of SL2 and SL3 (50, 51). Furthermore, the exposure of guanines appears to be a prerequisite for tight NC binding, as mutating GGAG of SL3 into a tightly stacked stable GNRA tetraloop drastically reduces affinity to NC (52). Residues G30 and G31 are extrahelical in the predominant structural motifs of the G-bulge determined in this work (Fig. 9, A and B), so it is possible that one or both guanines have a similar mode of interaction with NC. It is interesting that these residues were found to be exposed during in vitro probing, as they were accessible to RNase T1 and to modification with kethoxal, while G5 was not (53); G5 is stacked in the NMR structure.

Acknowledgments—We thank Moritz Mayer for preparation of T7 RNA polymerase and Gregory S. Couch for help with the program Chimera.

REFERENCES
1. Paillart, J. C., Shehu-Xhilaga, M., Marquet, R., and Mak, J. (2004) Nat. Rev. Microbiol. 2, 461–472
2. Shubda, M. F., Paolleti, A. C., Hudson, B. S., and Borer, P. N. (2002) Biochemistry 41, 5276–5282
3. Skripkin, E., Paillart, J. C., Marquet, R., Ehresmann, B., and Ehresmann, C. (1994) Proc. Natl. Acad. Sci. U. S. A. 91, 4945–4949
4. Clever, J. L., Wong, M. L., and Parslow, T. G. (1996) J. Viral. 70, 5902–5908
5. Clever, J. L., and Parslow, T. G. (1997) J. Viral. 71, 3407–3414
6. Mujeeb, A., Clever, J. L., Billeci, T. M., James, T. L., and Parslow, T. G. (1998) Nat. Struct. Biol. 5, 432–436
7. Ennifar, E., Walter, P., Ehresmann, B., Ehresmann, C., and Dumais, P. (2001) Nat. Struct. Biol. 8, 1064–1068
8. Baba, S., Takahashi, K., Noguchi, S., Takaku, H., Koyanagi, Y., Yamamoto, N., and Kawai, G. (2005) J. Biochem. (Tokyo) 138, 583–592
9. Mujeeb, A., Parslow, T. G., Zarrinpap, A., Das, C., and James, T. L. (1999) FEBS Lett. 458, 387–392
10. Girard, F., Barbault, F., Gouyette, C., Huynh-Dinh, T., Paolleti, J., and Lancelot, G. (1999) J. Biomol. Struct. Dyn. 16, 1145–1157
11. Ennifar, E., Yusupov, M., Walter, P., Marquet, R., Ehresmann, B., Ehresmann, C., and

A. Mujeeb, S. Georgeritis, I. Smirnov, J. Chung, T. L. James, and T. G. Parslow, unpublished data.

Linear Dimer of SL1 RNA
