Articles

A Potential Gene Target in HIV-1: Rationale, Selection of a Conserved Sequence, and Determination of NMR Distance and Torsion Angle Constraints[†]

Anwer Mujeeb,[‡] Sean M. Kerwin,[‡,§] William Egan,[∥] George L. Kenyon,[‡] and Thomas L. James[*,‡]

Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94143-0446, and Biophysics Laboratory, Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, Maryland 20892

Received June 5, 1992; Revised Manuscript Received July 17, 1992

ABSTRACT: Recently, the capability for determining the high-resolution, sequence-dependent structure of oligonucleotides in solution via careful analysis of multidimensional NMR spectra and structure refinement procedures has been developed. Consequently, the rationale for selection of a genome sequence as a target for drug design based on the detailed three-dimensional structure of the target is presented. The concept is illustrated by the successful search for a highly conserved region of the HIV-1 genome's long terminal repeat which could serve as a molecular target. A compound which could selectively bind the target sequence could inhibit both RNA transcription from the integrated provirus and the reverse transcription process. Of 148 unique HIV-1 sequences examined, 147 exhibit a 21-base conserved sequence (nucleotides 70–90 in HIVHXB2R) in the R region of the long terminal repeat. The only exception, a minor constituent for one individual, has a change in the penultimate base. A 13 base pair duplex sequence, [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)], from this conserved region was selected and synthesized for NMR structure studies. Phase-sensitive proton two-dimensional nuclear Overhauser enhancement (2D NOE) and double-quantum-filtered correlation (2QF-COSY) spectra were obtained at 500 MHz for the DNA duplex. Exchangeable and nonexchangeable proton resonances were assigned. Quantitative assessments of the 2D NOE cross-peak intensities for different mixing times were carried out using conventional Fourier transform NMR and the maximum likelihood method (MLM). Distance constraints, along with upper and lower bounds, were obtained from the 2D NOE intensities using the iterative complete relaxation matrix algorithm MARDIGRAS. Distances entailing both exchangeable and nonexchangeable protons were determined: 7–11 experimental distance constraints per residue including interresidue and interstrand distances. Simulations of the scalar coupling effects manifest in 2QF-COSY cross-peaks by means of the program SPHINX/LINSHA were compared with experimental data to yield torsion angle constraints for the sugar rings. A single conformer was inadequate to describe any of the sugar puckers, but a rapid two-state equilibrium with one conformer strongly dominant (75–95%) provided a good fit of the 2QF-COSY cross-peaks. The sugar pucker of the major conformer exhibited significant variability for the various nucleotides but was roughly 2′-endo. Though derived independently and subject to different time-averaging effects, the 2QF-COSY and 2D NOE results are in accord.

In this era of "rational drug design", much effort is being made to design diagnostic and therapeutic agents which bind to protein receptors. This effort has been fueled by the increasing availability of protein structures via X-ray crystallography primarily and, more recently, from NMR studies. As yet, however, there has been little or no effort to design drugs rationally on the basis of the sequence-dependent three-dimensional structure of the DNA in a gene. The reason for this perhaps is that while X-ray crystallographic techniques have the capability for determining high-resolution structures in the crystalline state, with a few notable exceptions, there has been limited success in crystallizing DNA fragments in the B-form, i.e., the form in which they are nearly always found in solution. However, the field of NMR structure determination has advanced to the state that a few labs can now determine the solution structure of interesting DNA fragments. And further improvements in methodology which will enhance the quality of the structures determined are certain. Consequently, we can now entertain the concept of designing agents to bind to gene targets on the basis of the detailed three-dimensional structure of the target. While this idea seems logical and is being applied to design ligands for protein receptors, it has not yet been attempted for DNA as a receptor. Obviously, if one could successfully design a drug on the basis of a gene target, rather than a gene product, that drug has the potential to be considerably more efficacious.

Human acquired immunodeficiency syndrome (AIDS) has affected 1.2 million people worldwide, and an additional 5–10 million people worldwide are believed to be infected with the causative agent, the human immunodeficiency virus (HIV)[1] (St. Georgieve & McGowan, 1990). The only drugs currently approved for the treatment of HIV infection are zidovudine (AZT) and didanosine (ddI). These drugs are believed to act by interfering with the virally encoded reverse transcriptase

[†] This work was supported by National Institutes of Health Grants GM 39247 (T.L.J.) and GM 39552 (G.L.K.).

[*] Author to whom correspondence should be addressed.

[‡] University of California, San Francisco.

[§] Present address: Division of Medicinal Chemistry, The University of Texas, Austin, TX 78712-1074.

[∥] Center for Biologics Evaluation and Research.

0006-2960/92/0431-9325$03.00/0

enzyme. While these drugs may offer some palliative treatment for AIDS patients, and may slow the progression of those with AIDS related complex (ARC) to full blown AIDS, there is no cure for this apparently always fatal disease. This lack of effective long-term treatment for HIV infection, as well as the emergence of strains of the virus that are resistant to AZT and ddI, demonstrates the need to develop alternative chemotherapeutic agents (Larder et al., 1989).

We have embarked upon a program of targeting the double-stranded DNA form of the HIV genome. Compounds that recognize viral DNA may serve as leads in the development of a new class of anti-HIV agents (Kerwin et al., 1991). Such compounds are likely to inhibit both RNA transcription from the integrated provirus and the reverse transcription process which apparently utilizes a DNA template (vide infra). Key initial steps in this process are the selection of an appropriate target sequence and the structure determination of the double-stranded DNA form of this target sequence. In this paper we present the selection of a highly conserved region of the HIV-1 long terminal repeat (LTR) as a molecular target for drug design. A synthetic duplex oligonucleotide encompassing a portion of this highly conserved region was synthesized. The ^{1}H NMR spectrum of this duplex oligonucleotide has been assigned, and torsion angle constraints and distance constraints have been determined, respectively, from analysis of proton homonuclear 2QF-COSY and 2D NOE spectra.

Multidimensional NMR has the capability of yielding internuclear distances and bond torsion angles as experimental structural constraints for noncrystalline molecules (James & Basus, 1991; Oppenheimer & James, 1989a,b; Wagner, 1990; Wüthrich, 1986). Internuclear distances and bond torsion angles, however, do not directly constitute a molecular structure. Consequently, various computational methods, e.g., distance geometry (DG) and molecular dynamics (rMD), have been adapted to yield molecular structures which are consistent with the experimental constraints.

Complete relaxation matrix analysis of proton homonuclear 2D NOE spectra enables numerous accurate interproton distances to be calculated (Borgias & James, 1988, 1990; James, 1991; Keepers & James, 1984; Liu et al., 1992). The most effective techniques for calculating accurate distances entail an iterative approach (Boelens et al., 1988; Borgias & James, 1988, 1989, 1990; Liu et al., 1992; Madrid et al., 1991; Post et al., 1990). The program MARDIGRAS (matrix analysis of relaxation for discerning geometry of an aqueous structure) seems to be rather robust and displays little dependence on the initial model while yielding reliable distances with computational efficiency (Borgias & James, 1990; Borgias et al., 1990; Liu et al., 1992; Thomas et al., 1991). The distance information from the 2D NOE analysis can be augmented by torsion angle constraints derived from coupling constants obtained from double-quantum-filtered COSY (2QF-COSY) and exclusive COSY (ECOSY) spectra. Broad lines prevented direct analysis of coupling constants, so we used simulation of 2QF-COSY cross-peaks using the programs SPHINX and LINSHA (Widmer & Wüthrich, 1987); this enables us to extract vicinal coupling constants and subsequently torsion angle constraints (Celda et al., 1989; Gochin et al., 1990; Schmitz et al., 1990). In the case of DNA helices, we are interested in fairly subtle structural details which can influence recognition and stability. These subtle variations demand detailed knowledge of the structure and, therefore, accurate structural constraint determinations as well as many structural constraints.

MATERIALS AND METHODS

The present project entails selection of a potential drug binding site on the HIV-1 gene, synthesis of the oligonucleotide corresponding to the target, and derivation of numerous accurate structural constraints preparative to determination of the detailed solution structure of the double-stranded DNA sequence. The resulting structure will be subsequently examined in the context of its being a receptor for a ligand to be designed, i.e., considering molecular geometry as well as the electrostatic surface and potential hydrogen-bonding partners presented to the ligand. On this basis, one or more potential gene blockers will be designed with the aim of developing a new class of potential diagnostic and chemotherapeutic agents. This approach has the future potential for development of effective chemotherapeutic agents with considerably reduced side effects and enhanced efficacy.

Selection of Target Sequence

A number of compounds that bind to or covalently modify double-stranded DNA have been shown to possess antiviral activity (Dixon et al., 1990; Li et al., 1991; Lown et al., 1989; Nakamura et al., 1987). Some of these compounds have shown activity against the HIV virus (Dixon et al., 1990; Li et al., 1991; Lown et al., 1989; Nakamura et al., 1987). In some cases, it has also been shown that these antiviral compounds selectively inhibit HIV reverse transcriptase (Ajito et al., 1989; Atsumi et al., 1988; Lown et al., 1989; Nakamura et al., 1987). These antiviral DNA-binding compounds presumably inhibit viral reverse transcription by binding to base-paired regions of a DNA template. The inhibition of the process of reverse transcription by a variety of double-stranded DNA binding compounds led us to postulate that a sequence-selective DNA-binding compound that bound to a specific region of double-stranded DNA template would also inhibit the reverse transcription process. A double-stranded DNA template is believed to be important in the strand-displacement synthesis of the minus DNA strand of the 3′-LTR (Gilboa et al., 1979). In addition to interfering with reverse transcription, such a DNA-binding compound could also inhibit RNA transcription from the integrated provirus, particularly if the binding region coincides with enhancer or promoter regions in the LTR (Cooney et al., 1988). Consequently, we sought to find a highly conserved region of the HIV-1 genome within the LTR that could serve as a target for drug design.

To determine the extent of conservation among the various HIV-1 viruses that had been sequenced, a multiple alignment of 20 of the HIV-1 sequences taken from the GenBank database was performed using the IntelliGenetics GENALIGN program.[2] The following sequences were used: HIVHXB2, HIVHIVAT3, HIVHIVC15, HIVHIVXB2, HIVHIVXB3, HIVMNCG, HIVZ2Z6, HIVBH102, HIVANT70, HIVBRUCG, HIVBRVA, HIVELICG, HIVHXB2CG, HIVMAL, HIVNL43, HIVRF, HIVSC, HIVSF2CG, HIVZ6, and HIVZ321. The resulting

[1] Abbreviations: 1D NMR, one-dimensional NMR; 2D NMR, two-dimensional NMR; 2D NOE, two-dimensional nuclear Overhauser effect; COSY, correlation spectroscopy; 2QF-COSY, double-quantum-filtered COSY; FID, free induction decay; PCR, polymerase chain reaction; HIV, human immunodeficiency virus; LTR, long terminal repeat; TAR, trans-activating response element; MLM, maximum likelihood method; rMD, restrained molecular dynamics.

[2] GENALIGN is a copyrighted software product of IntelliGenetics, Inc., Mountain View, CA. The program was developed by Dr. Hugo Martinez of the University of California, San Francisco.

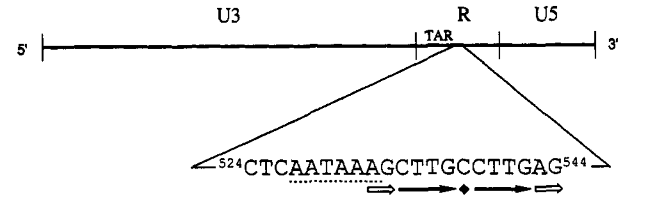

FIGURE 1: Location of the highly conserved sequence within the HIV-1 long terminal repeat. Numbers indicate the position of the sequence relative to the 5′ end of the U3 region. The polyadenylation signal sequence is underlined. Short repeated sequences within the 3′ terminal region of the conserved sequence are denoted by arrows.

consensus sequence was then compared to the consensus sequences published in the 1990 and 1991 versions of the Los Alamos HIV sequence database (Myers, 1990; Myers, 1991). There was only one region of greater than 20 bases within the LTR that was conserved in all of the sequences examined. (Only five such 100% conserved regions of 20 or more bases were noted within the entire HIV-1 genome when comparing all complete HIV-1 sequences in the Los Alamos database.) The conserved region (nucleotides 70–90 in HIVHXB2R) is located in the R region of the LTR and encompasses the polyadenylation signal sequence (Figure 1).

We believe that the probability of finding such a large 100% conserved region in many different HIV-1 sequences is quite low. We note that the HIV-1 genome is quite variable, due primarily to the poor fidelity of the HIV-1 reverse transcriptase (Bebenek et al., 1989; Ji & Loeb, 1992). It has been estimated that a single round of reverse transcription introduces up to 10 mutations into the HIV-1 genome (Preston et al., 1988). The higly variable nature of the viral genome decreases the likelihood of a large 100% conserved region in many different sequences occurring by chance. For a random distribution of mutations throughout the genome, we estimate the probability of such a 100% conserved region happening by chance to be 1 in 5000. This estimate was determined by calculating the probability of a mutation at any site by the formula $P = M/B$, where P is the probability, M is the total number of substitutions, insertions, and deletions in 23 randomly chosen HIV-1 sequences within a 120 base region surrounding the conserved region, and B is the total number of bases in the 120×23 array. While the inclusion of mutational hot spots into such a calculation would serve to increase the probability of finding such a conserved sequence, the probability is, we presume, still quite low.

To verify that this sequence remains conserved in the majority of viral isolates, an additional multiple alignment of all HIV-1 sequences that contain this region was performed. In the resulting alignment of 148 unique HIV-1 sequences, 147 have this sequence completely conserved, and one sequence (HIVNE001) has a single mutation in the penultimate base. The one sequence with a single base mismatch was one of 111 sequences of in vitro and in vivo viral isolates from a single HIV-infected individual obtained as part of a 4-year longitudinal study (Delassus et al., 1991). There is a possibility that the mutation observed in this sequence is an artifact of the PCR reaction used to amplify the sequence prior to cloning and sequencing. The authors estimated a minimum natural-to-artifactual substitution rate of 7:1 (Delassus et al., 1991). Even if this single base substitution is not an artifact, the fact that a viral isolate containing this substitution was isolated only once, as a minor (5%) constituent of the first in vivo sequenced quasispecies, argues for its relative lack of replication competence.

Although the highly conserved region we observed for HIV-1 does not appear in HIV-2, there are a number of highly divergent HIV-1-related sequences that contain this conserved region. These include HIVANT70, the sequence of a highly divergent HIV strain that is related to both types 1 and 2 (De Leys et al., 1990), and the simian immunodeficiency virus sequence SIVCPZ (Huet et al., 1990). ANT-70 is a strain of virus isolated from two persons of west-central African origin that is only 70% homologous to HIV-1.

For subsequent study, we chose to select from the 21-base conserved region a subsequence. This subsequence was selected using two criteria. First, the subsequence should have a base composition and size to facilitate study by NMR spectroscopy while still possessing sufficient length for uniqueness. Our second criterion was more subjective: we felt that the ability to design molecules that recognized this subsequence would be enhanced if there were a symmetry (or pseudosymmetry) to the subsequence. Many DNA-binding proteins are bipartite; each subunit of the protein contributes residues that aid in the recognition of one half of the cognate DNA site (Johnson & McKnight, 1989). By targeting a sequence possessing a degree of symmetry, we hoped to employ a similar strategy in designing our DNA-binding ligand; a single ligand recognizing a relatively long DNA sequence might be assembled from two (similar or identical) moieties, each able to distinguish a much small DNA sequence. The subsequence AGCTTGCCTTGAG, consisting of the final 13 bases of the 21-base conserved region, was chosen because it satisfied our two criteria. The subsequence was predicted to be relatively free from base-pair unraveling at the termini of the duplex. Similar nucleotides within the subsequence are distinct with regard to neighboring residues. This arrangement would help to minimize overlapping signals in the NMR spectrum of the duplex. Finally, the subsequence possesses a degree of symmetry as shown in Figure 1.

Preparation of Sample

The deoxynucleotide trisdecamers d(AGCTTGCCTTGAG) and d(CTCAAGGCAAGCT) were obtained via solid-phase oligonucleotide synthesis, carried out on an Applied Biosystems Inc. model 380B DNA synthesizer using standard β-cyanoethyl chemistry according to the manufacturer's protocol (Stolarski et al., 1992). All reagents used in the synthesis were purchased from Applied Biosystems. The base-protected oligonucleotide bearing the dimethoxytrityl group was purified by reverse-phase HPLC using a Hamilton PRP-1 column, detritylated with acetic acid, dried, rehydrated in water, and extracted three times with ethyl acetate. The resulting oligonucleotide was repurified by reverse-phase HPLC.

The sample of double-stranded DNA for NMR studies was made by mixing the two strands, d(AGCTTGCCTTGAG) and d(TCGAACGGAACTC), in 1:1 stoichiometry, which was determined by titrating one strand with the other and monitoring the titration by UV spectrophotometry. Following titration, no peaks were observed in the proton 1D NMR spectrum corresponding to an excess of either single strand. The duplex sample for NMR was prepared in 50 mM NaCl; the pH was adjusted to 7.0, but no buffer was added. The final concentration of the duplex was about 1.8 mM. The NMR sample was lyophilized several times from 99.95% ^2H_2O and finally dissolved in 99.995% ^2H_2O. For assignments of exchangeable protons, a sample of DNA duplex in 10% ^2H_2O and 90% H_2O was prepared.

NMR Spectroscopy

^1H NMR experiments were run on a 500-MHz General Electric GN500 NMR spectrometer equipped with Nicolet

1280 computer. As the melting point transition of the DNA duplex was found to be ~52 °C via 1D NMR experiments, unless specified otherwise, all subsequent experiments were performed at 35 °C. The 2D NOE experiment in H_2O was carried out at 15 °C. Pure absorption phase 2D NMR spectra were obtained by using State's method of phase cycling and alternating block acquisition with the pulse sequence delay–90°–t_1–90°–τ_m–90°–t_2 (States et al., 1982). The 2D NOE spectra in ^2H_2O were recorded at three mixing times: 80, 120, and 200 ms. A delay time of 11 s between scans was used to allow for complete relaxation of magnetization. A spectral width of 4000 Hz was used, and the carrier frequency was set at the HDO proton resonance. Four hundred free induction decays (16 scans per FID) were collected with 2K data points in the t_2 dimension. Data were processed on a SUN Sparcstation 2 using locally written NMR processing software and the NMR2 software package from NMRi (Syracuse, NY). During processing, data were multiplied by a phase-shifted sine-squared function and were zero-filled in both dimensions to yield a final 2K × 2K spectrum.

The 2D NOE spectra in H_2O were recorded using the three-pulse sequence with the third pulse replaced by the $1\ 3\ \bar{3}\ \bar{1}$ water suppression pulse and a 10-ms homogeneity spoil gradient pulse at the beginning of the mixing period to suppress further the water signal (Hore, 1989). Data were collected for a spectral width of 10 000 Hz with the excitation maximum set between imino and aromatic proton regions at 11.5 ppm. Other experimental conditions were the same as for 2D NOE spectra in ^2H_2O (vide supra).

Pure absorption double-quantum-filtered COSY (2QF-COSY) spectra were recorded by using time-proportional phase incrementation (Marion & Wüthrich, 1983). Signals were collected for 800 t_1 values with 16 scans per t_1 value; 2K data points were acquired in t_2. A square sine bell filter function, shifted by 45°, was used for apodization in both dimensions. The data matrix was zero-filled to have a final size of 2K × 2K and a digital resolution of 1.95 Hz/point in both dimensions.

Measurements of proton spin–lattice relaxation times T_1 were performed using the inversion–recovery method, utilizing a 180° composite pulse, with a repetition time of 30 s (Freeman et al., 1980).

NMR Data Analysis

Simulation of 2QF-COSY Cross-Peaks. SPHINX and LINSHA programs (Widmer & Wüthrich, 1987) were used to simulate 2QF-COSY cross-peaks for the deoxyribose spin system H1′, H2′, H2″, H3′, H4′, and the ^{31}P spin-coupled to H3′. SPHINX calculates stick spectra, treating all proton spins as strongly coupled nuclei. Line shapes, based on digital resolution, apodization functions, truncations of FIDs, and natural line widths, were added to the stick spectra using LINSHA. The experimental cross-peaks were compared with the simulated ones for many different sets of coupling constants and line widths. Conformations of the deoxyribose rings in the trisdecamer were determined using an extensive compilation of reported values of proton–proton coupling constants (Rinkel & Altona, 1987).

Quantitation of 2D NOE Cross-Peaks. Different methods of quantitating the 2D NOE cross-peaks were employed and results compared. This was partially motivated by a desire to improve the spectral resolution, yielding more usable peaks and, consequently, more distance constraints. Various methods of contrast enhancement have been examined in numerous labs in recent years (Hoch, 1989). The maximum likelihood method (MLM) of constrained deconvolution is one which we have recently examined for analysis of our spectra.

The MLM algorithm, in particular a version of program DECO2D (Ni & Scheraga, 1989), has been incorporated into the NMR2 software package for 2D NMR data processing; this was used for our analysis. MLM treats the 2D NMR data sets with a line width sharpening function in both dimensions and deconvolutes the peaks, thus resulting in an apparently better signal-to-noise ratio. Deconvolution greatly simplifies the interpretation of overlapping peaks. The question of whether quantitative use of cross-peaks resulting from the maximum likelihood method could be justified needs to be examined. This question has been addressed recently with the conclusions that quantitation could be reliably carried out if resolution enhancement was not pushed too hard, i.e., if a 2–3-fold improvement was sufficient, and if no sensitivity enhancement was to be expected (Jeong et al., 1992). Our less extensive analysis corroborates those conclusions. MLM works best when applied on small square regions of 128 × 128 or 256 × 256 points containing a small number of cross-peaks; results were unsatisfactory for regions containing strong diagonal peaks or any other comparatively large peak. Values of line sharpening functions used in both dimensions affect the final shape and size of MLM-treated peaks; if efforts to strongly enhance resolution are employed, the excessive resolution enhancement may result in distortion or even removal of some peaks. This problem could be avoided by first estimating line widths of peaks (in both dimensions) in a selected region of the spectrum; line sharpening functions were then chosen in such a way that the values were smaller than the smallest line width value in each dimension. This is a kind of compromise with the deconvolution of large line width peaks which are clustered, as we may not obtain satisfactory resolution in an overlapped region. However, the compromise in spectral resolution enables reliable peak quantitation.

To test the feasibility of using MLM-treated cross-peaks quantitatively, a subset of experimental cross-peaks, with varying degrees of overlap, was examined using MLM and traditional FT analysis. In both cases, following baseline flattening, the peak intensities were quantitated by "boxing" (or "circling") individual peaks, in a manner similar to that employed in a number of programs currently in wide use, and by surface fitting. Analogous to curve fitting in a 1D spectrum, surface fitting allows for more accurate determination of intensities for close or overlapped peaks. Intensities are calculated analytically, on the basis of a line shape model. A convergence criterion is defined which is the percentage intensity difference between experimental and modeled surfaces. It was observed that relative integrals were in agreement for all four sets of intensities, i.e., the ratios of peak volumes were within 10–15% regardless of whether boxing or surface fitting was employed or whether MLM or traditional FT. But in the case of very weak cross-peaks which emerged only after MLM treatment, the intensities were found to depend strongly on processing parameters; none of these intensities was subsequently used. On the basis of our limited experience, we find that MLM constrained deconvolution can greatly help in assignment of resonances. Its application to quantitative estimations is rather limited; cautious use of MLM yields quantifiable peaks with slightly increased resolution. On the oligonucleotide duplex examined here, we estimate that we can quantitate ~10% more peaks by cautiously using MLM. A greater number of peaks should yield a larger number of interproton distances and a better structure. However, only

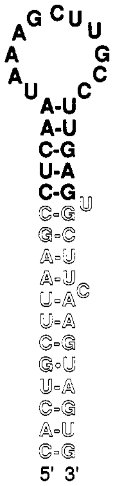

FIGURE 2: Predicted structure of HIV-1 RNA immediately downstream of TAR. The highly conserved region appears in bold letters.

a small gain in number of distances was garnered by the extra effort.

The intensities of cross-peaks from exchangeable imino protons were also determined and included in distance constraint evaluation. These were extracted from 2D NOE spectra obtained for the oligomer in H_2O solution. The excitation profile resulting from use of the $1\ 3\ \bar{3}\ \bar{1}$ water suppression pulse was determined and used to scale and normalize peak intensities. Cross-peaks of imino protons from terminal and penultimate residues of the duplex were not utilized because of significant proton exchange with water. As nucleic acids do not have many nonexchangeable base protons, distances to the exchangeable imino protons should be quite valuable for improving structure quality.

RESULTS AND DISCUSSION

Selection of a Conserved Sequence in the Long Terminal Repeat of the HIV-1 Genome

A consensus sequence in the long terminal repeat of the HIV-1 genome was successfully identified (vide supra). The function that this highly conserved sequence serves in the HIV-1 virus is unknown. The conserved region corresponds to and encompasses the terminal stem-loop of a larger hairpin structure that has been predicted to occur at the 5′ end of the HIV-1 mRNA (Muesing et al., 1987). A calculation indicates that this longer hairpin loop, shown in Figure 2, is relatively stable (19 kcal/mol); the calculation was carried out using the DynPro program, developed by Dr. Hugo Martinez (UCSF). The results of RNase digestion experiments are in accord with the structure shown in Figure 2 (Muesing et al., 1987). Because many RNA-binding proteins bind to stem-loop structures, we may speculate that this highly conserved region is recognized in its RNA form by one or more cellular or viral proteins. The polyadenylation signal, which is a part of the larger conserved region, certainly is recognized as an RNA structure (Keller et al., 1991). It is possible that the stem-loop structure formed by the larger conserved region may play a role in the occlusion of the polyadenylation signal at the 5′ end of the viral RNA transcript (Varmus, 1988). This conserved region does not appear to be required for polyadenylation signal occlusion in HeLa cells, however (Weichs an der Glon et al., 1991). It remains to be seen whether the same is true in T-cell lines. The highly conserved nature of the sequence in Figure 2 argues against its functioning solely as an RNA stem-loop. The TAR RNA structure is another stem-loop immediately upstream of the structure shown in Figure 2. In the HIV-1 sequences we compared, the TAR region is not so well conserved as the downstream conserved region; several mutations in TAR were evident. Experiments by Roy and co-workers have demonstrated that mutations in the stem of TAR that do not disturb the base pairing result in levels of transactivation commensurate with the wild-type structure (Roy et al., 1990). Presumably, similar compensatory mutations in the conserved stem-loop region would obtain. So we suggest that the conserved region serves another function, perhaps in the form of a double-stranded DNA structure. The superposition of RNA- and DNA-acting cognition sequence elements has been established in the case of the HIV-2 TAR element (Jones et al., 1988) and the HTLV-1 Rex-responsive region (Levinger & Lawtenberg, 1987).

Assignment of Proton Resonances

The nucleotides in the two strands were numbered in ascending order from the 5′ → 3′ end as shown below:

5′-A1 G2 C3 T4 T5 G6 C7 C8 T9 T10 G11 A12 G13-3′
3′-T26 C25 G24 A23 A22 C21 G20 G19 A18 A17 C16 T15 C14-5′

Assignments of all nonexchangeable protons were made in a sequential manner from 2D NOE, at mixing times of 80, 120, and 200 ms, and 2QF-COSY spectra in ^2H_2O following the strategy described in detail in earlier work (Broido et al., 1984; Feigon et al., 1983; Scheek et al., 1983). An example of the 2D NOE spectrum in the aromatic–2′/2″ region, labeled with resonance assignments, is shown in Figure 3. Exchangeable imino and amino protons were assigned from the 2D NOE spectrum with the oligomer in 90% H_2O, 10% ^2H_2O solution (Chou et al., 1984; Rajagopal et al., 1988; Zhou et al., 1988). The imino proton (T-N3H and G-N1H) region of the 200-ms 2D NOE spectrum in H_2O at 15 °C is shown in Figure 4.

Due to heavy overlap of signals, assigning 5′ and 5″ protons was not possible. The chemical shifts of all other protons are listed in Table I. Adenine H2 proton resonances are recognizable by their long T_1 values (Tables II). These resonances were specifically assigned to the appropriate residues using imino proton connectivities in the 2D NOE spectrum obtained in H_2O. Cross-peaks between T4-NH and A23-H2, T5-NH and A22-H2, and T9-NH and A18-H2 protons (Figure 4) are intra-base-pair and help in assigning H2 protons of adenine in an AT base pair. The assignment of AH2′ to a specific signal is also verified by its T_1 value. In addition, inter-based cross-peaks, e.g., T5-NH to A23-H2 and T10-NH to A18-H2 are also seen. Since T3-NH and AH2 protons of an AT base pair residue in the minor groove, appearance of these cross-peaks between adjacent AT base pair protons reflects a narrower minor groove than the standard B-DNA helix for this part of the sequence (Gochin et al., 1990).

Terminal residue imino and amino proton signals were not observed due to fast exchange from fraying of the ends under the experimental conditions. G-NH_2 proton signals were also too broad to appear at 15 °C, presumably due to rotation about the C–N bond. Cross-peak arising from guanine imino protons interacting with the H5 and amino protons of its base paired cytosine are evident. Non-hydrogen-bonded and hydrogen-bonded amino protons of cytosines have been marked as (1) and (2), respectively, and were distinguished on the basis that the downfield-shifted signal is due to the hydrogen-

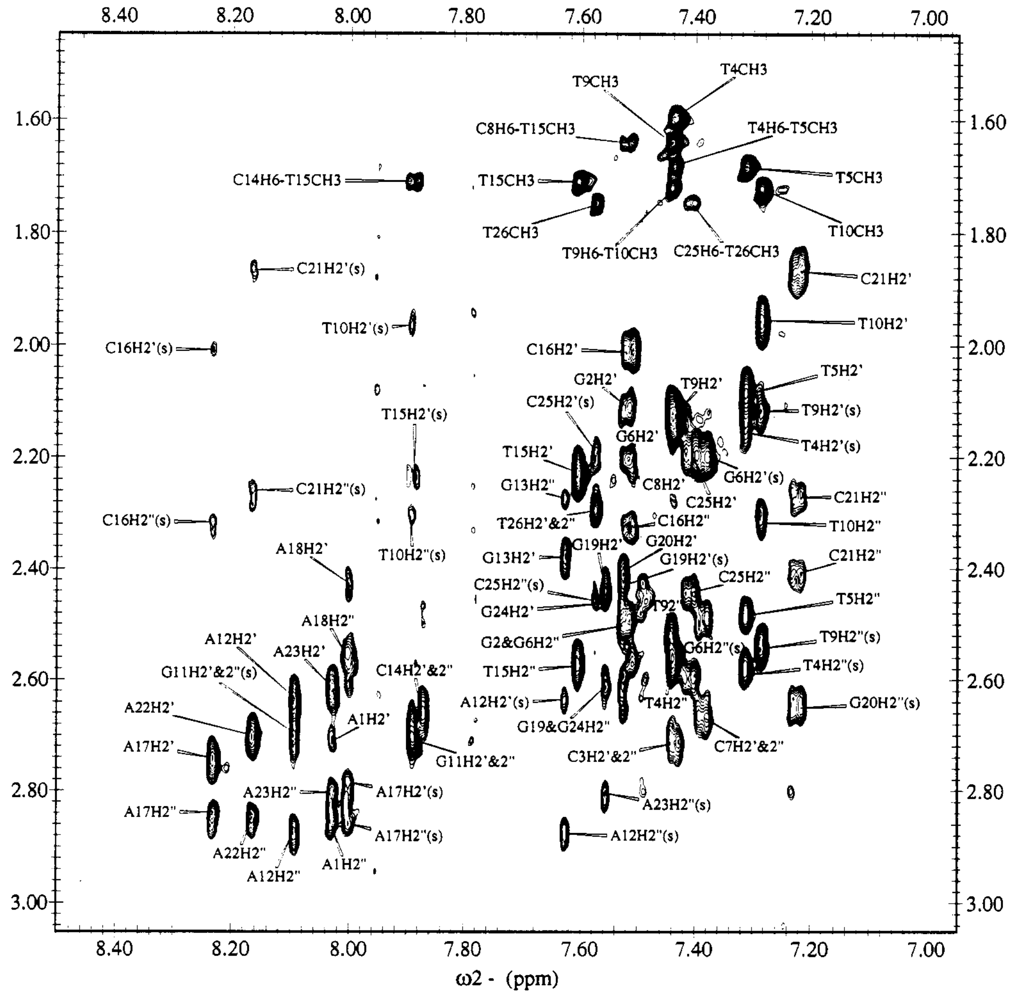

FIGURE 3: Aromatic-to-H2′/H2″ cross-peak region of the 2D NOE spectrum (200-ms mixing time) of [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] in ^2H_2O, at 35 °C; (s) denotes sequential connectivity.

bonded amino proton (Rajagopal et al., 1988; Zhou et al., 1988).

Determination of Distance Constraints from 2D NOE Spectra

To obtain the best possible structures, one should utilize as many structural constraints as possible and determine these constraints as accurately as possible. Here we utilize 2D NMR to determine experimentally both interproton distance constraints and torsion angle constraints. There are different methods of analyzing 2D NOE spectra for internuclear distance and structural information. We have examined the use of the commonly employed two-spin or isolated spin pair approximation (ISPA) to obtain interproton distances from 2D NOE spectra and found that while the approximation leads to sizeable systematic as well as random distance errors (Borgias & James, 1988; Keepers & James, 1984), good protein structures can still be obtained, albeit with occasional local structural distortions when the distances are assumed to be more accurate than is warranted (Thomas et al., 1991). Use of a complete relaxation matrix approach (CORMA) to ascertain interproton distances from 2D NOE peak intensities enables more accurate distance determinations as well as a greater number of constraints to be derived (Borgias & James, 1988, 1990; Keepers & James, 1984); this, consequently, offers the opportunity of determining protein solution structure with greater accuracy and resolution and is essential for determination of DNA helix structures. The most effective techniques employ iterative refinement against experimental 2D NOE spectra by calculating theoretical spectra for the molecular structures during refinement (Boelens et al., 1988; Borgias & James, 1988, 1989, 1990; Liu et al., 1992), in concert with molecular dynamics or distance geometry calculations. For example, the program MARDIGRAS exhibits little dependence on the initial model while yielding reliable distances with computational efficiency; the resulting distances can be used with restrained molecular dynamics or distance geometry calculations.

Use of MARDIGRAS necessitates a starting structure. Distances were calculated with different starting structures and the same 2D NOE intensities. Starting model structures of [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] were generated for A-DNA and B-DNA geometries using the program package AMBER 4.0 (Pearlman et al., 1991). The NUCGEN, PREPARE, LINK, and EDIT modules of AMBER were used to generate the all-atom (including hydrogen) DNA geometries. Hexahydrated sodium ions were added at a distance of 5 Å to neutralize the phosphate charges. Energy minimization of these standard A- and B-DNA structures was carried out using the MINMD module with an initial steep energy descent during the first 200 cycles followed by refinement until an energy gradient value lower than 4.0 kJ·mol^{-1}·nm^{-1} was achieved.

More recent versions of MARDIGRAS enable us to iteratively determine distances to groups for which interproton distances fluctuate due to internal motions such as methyl

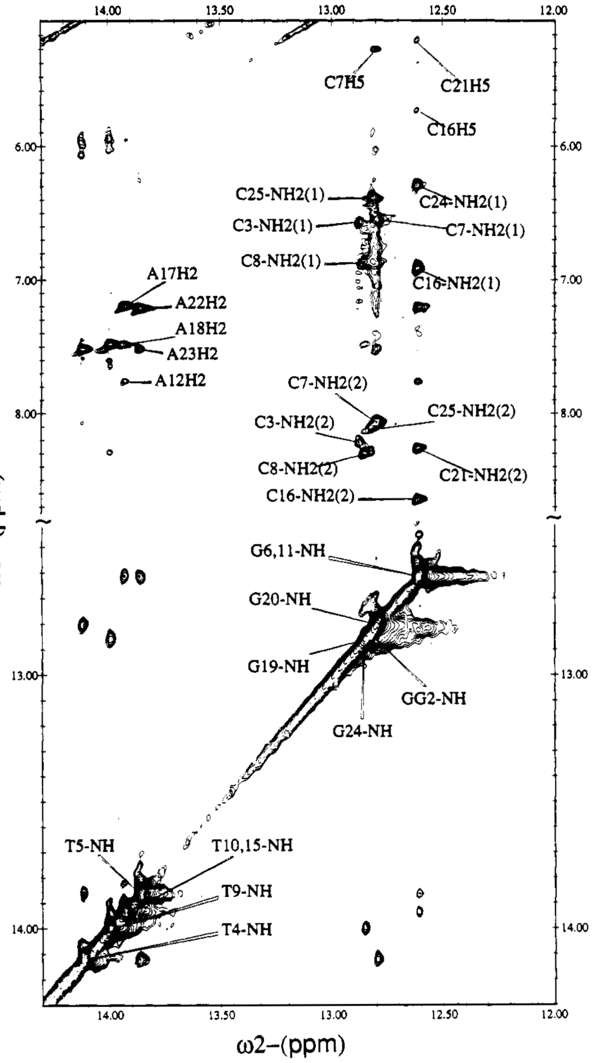

FIGURE 4: Section of the 200-ms 2D NOE spectrum of [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] obtained in H_2O at 15 °C showing the cross-peaks arising from the imino proton region.

rotation (Liu et al., 1992). For example, it has been found that the six fixed H6–methyl distances on the thymines of the DNA duplex [d(GTATAATG)·d(CATATTAC)] were in error by less than 2% when calculated using MARDIGRAS and including methyl rotations but were typically underestimated 10–12% when methyl rotations were ignored (Liu et al., 1992). Methyl group rotation can be modeled in different ways using MARDIGRAS, with choice of model dependent on situation (Liu et al., 1992). For the present studies with [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)], a three-site jump model was used to calculate intraresidue distances, and an 18-site jump model, which is a good approximation to free internal rotation diffusion, was used to calculate interresidue distances.

An overall tumbling correlation time τ_c is also needed for the distance determinations via MARDIGRAS. For a 13-mer, the hydrated length-to-radius ratio is about 2. This degree of anisotropy has a negligible effect on NOE values. Consequently, an overall molecular tumbling is effectively isotropic and is well approximated by a single correlation time. This was estimated from the equation relating the experimentally measured T_1 and T_2 relaxation times for resolved adenine H2 and H8 protons (Suzuki et al., 1986). Values ranging between 1.6 and 2.8 ns were calculated. The calculation assumes a simple isotropic molecular motion with no spin diffusion. However, spin diffusion contributes to H8 signals, making the calculated value of τ_c = 2.8 ns too long, and dissolved oxygen could make the value of 1.6 ns for the H2 proton too short (Gochin et al., 1990). Consequently, a value of 2.0 ns for the isotropic overall tumbling motion was chosen for use in subsequent MARDIGRAS calculations. We note that the distances resulting from the MARDIGRAS calculation are, in fact, not very sensitive to the actual value of τ_c utilized.

MARDIGRAS calculations were performed on three 2D NOE data sets, with mixing times 80, 120, and 200 ms. Standard B-DNA, energy-minimized B-DNA, and energy-minimized A-DNA were utilized as starting structures. For a given mixing time, 148–206 final distances were obtained from the cross-peak intensities after rejection of very low intensity values in the experimental data set. The accuracy of distances calculated increases with the fraction of possible experimental intensities in the data set, so it is of value to extract all possible intensities from the experimental data. A few distances calculated from preliminary MARDIGRAS calculations yielded distances very incompatible with all others; closer examination of the original data revealed these to arise from badly overlapped peaks or peaks with quite low signal-to-noise ratios. These few were culled from the data set for final MARDIGRAS calculations. The final number of distances obtained was 244, corresponding to 7–11 distance constraints per residue including interresidue and interstrand distances, these latter arising mainly from imino contacts in the present case. The quality of interproton distances, derived from MARDIGRAS, is revealed by running the program twice for the same intensity data set but using different starting structures. A comparison of interproton distances between the two starting models, energy-minimized A- and B-DNA, and the distances after two MARDIGRAS runs with energy-minimized A- and B-DNA as starting structures is presented in Figure 5 for the 120-ms 2D NOE data set. This was achieved after five iteration cycles in MARDIGRAS. The deviation in MARDIGRAS-derived distances calculated using any two different starting structures was less than 10%, generally much less.

With MARDIGRAS, it is possible to compare experimental 2D NOE cross-peak intensities with values calculated for the starting structures as well as at each iteration in the MARDIGRAS algorithm (Kerwood et al., 1991). Various numerical indices could be used to evaluate the overall fit of experimental and calculated intensities. Typically, a residual index analogous to a crystallographic *R* factor has been used. But this *R* factor is dominated by cross-peaks corresponding to very short (<2.5 Å) distances. The 2D NOE cross-peak intensities depend on distances with a sixth root dependence; consequently, we have found that a more sensitive monitor of fitting is the sixth-root residual index (James, 1991; Kerwood et al., 1991; Thomas et al., 1991):

$$R_1^x = \frac{\sum_i |I_o^{1/6}(i) - I_c^{1/6}(i)|}{\sum_i I_o^{1/6}(i)}$$

where $I_o(i)$ is the experimental intensity of cross-peak *i*, and $I_c(i)$ is the corresponding cross-peak intensity calculated for a particular structure.

R_1^x ranged from 0.1 to 0.3 for the various canonical and energy-minimized A- and B-DNA starting structures calculated for the three mixing times. But the R_1^x values corresponding to the final converged MARDIGRAS matrix ranged

Table I: Chemical Shift Values (ppm) of Different Protons in d(AGCTTGCCTTGAG)·(CTCAAGGCAAGCT)

proton	A1	G2	C3	T4	T5	G6	C7	C8	T9	T10	G11	A12	G13
H1′	5.80	5.89	5.86	6.07	5.84	5.94	5.82	5.90	6.02	5.70	5.42	6.10	5.99
H2′	2.73	2.10	2.68	2.13	2.07	2.19	2.66	2.23	2.12	1.94	2.71	2.64	2.36
H2″	2.85	2.51	2.72	2.57	2.48	2.48	2.68	2.59	2.54	2.30	2.72	2.87	2.26
H3′	4.90	4.76	4.98	4.87	4.89	4.77	4.77	4.67	4.75	4.86	4.98	5.02	4.60
H4′		4.16	4.97	4.18		4.24	4.24	4.09	4.28	4.07	4.29	4.41	4.13
H2												7.78	
H5			5.32				5.31	5.53					
H6			7.44	7.44	7.31		7.38	7.52	7.44	7.29			
H8	8.03	7.51				7.52					7.89	8.10	7.63
CH3				1.59	1.68				1.64	1.72			
labile protons													
GH1/TH3		12.82		14.12	13.86	12.62			14.00	13.93	12.62		
CH4(1)[a]			6.56				6.54	6.86					
CH4(2)[b]			8.21				8.05	8.28					

proton	C14	T15	C16	A17	A18	G19	G20	C21	A22	A23	G24	C25	T26
H1′	5.81	6.14	5.45	5.81	5.97	5.74	5.81	5.49	5.81	5.97	5.54	6.07	6.25
H2′	2.67	2.22	2.01	2.74	2.43	2.43	2.41	1.87	2.70	2.61	2.44	2.18	2.29
H2″	2.67	2.56	2.33	2.84	2.61	2.59	2.64	2.27	2.84	2.80	2.60	2.43	2.29
H3′	4.98	4.91	4.84	5.04	4.91	4.91	4.89	4.79	5.02	5.02	4.91	4.77	4.55
H4′		4.25	4.08	4.37	4.19	4.33	4.34	4.08		4.19	4.31	4.20	4.07
H2				7.18	7.48				7.20	7.50			
H5	5.99		5.76					5.24				5.34	
H6	7.89	7.61	7.51					7.23				7.41	7.58
H8				8.23	8.00	7.56	7.53		8.16	8.03	7.56		
CH3		1.71											1.75
labile protons													
GH1/TH3		13.93				12.85	12.80				12.87		
CH4(1)[a]			6.92					6.29				6.38	
CH4(2)[b]			8.64					8.25				8.10	

[a] CH4(1) is the non-hydrogen-bonded amino proton of cytosine. [b] CH4(2) is the hydrogen-bonded amino proton of cytosine.

Table II: Spin–Lattice Relaxation Times of Various Protons of d(AGCTTGCCTTGAG)·(CTCAAGGCAAGCT) at 35 °C

proton(s)	T_1 (s)	proton(s)	T_1 (s)
A12,A17,A18H8	1.8	G19H1′	1.8
A1,A23H8	2.2	C3,C25H1′	1.8
A22H8	2.9	C7,C14H1′	2.2
G11,G13,G20H8	2.2	T4,T5H1′	1.8
G24H8	1.5	T9H1′	2.2
C3,C7,C25H6	1.8	T10H1′	2.5
C8,C14,C21H6	2.2	T4,T5,T9CH$_3$	1.8
T4,T5,T9H6	1.8	T10,T15CH$_3$	1.8
T10,T26H6	1.8	T26CH$_3$	1.5
T15H6	2.2	A12H2	5.8
A17,A18,A22,A23H1′	2.2	A17H2	4.7
A12H1′	1.8	A18H2	3.6
G2,G6H1′	2.5	A22H2	4.7
G13,G20H1′	2.2	A23H2	3.6

from 0.001 to 0.009.

Determination of Torsion Angle Constraints from 2QF-COSY Spectra

Distance information from the 2D NOE analysis has been augmented by torsion angle information from vicinal proton coupling constants obtained from double-quantum-filtered COSY (2QF-COSY) spectra. Shown in Figure 6 are pertinent regions of the 2QF-COSY spectrum, obtained at 35 °C (also see Supplementary Material). Assignments of 1′, 2′, 2″, and 3′ proton resonances made by 2D NOE were confirmed via 2QF-COSY. To evaluate coupling constants, the H1′(ω_2)–H2′/H2″(ω_1) region, the H2′/H2″(ω_2)–H1′(ω_1) region (diagonally opposite), and the H2′/H2″(ω_2)–H3′(ω_1) region were employed for simulated fitting with the SPHINX and LINSHA programs (Celda et al., 1989; Gochin et al., 1990; Schmitz et al., 1990; Widmer & Wüthrich, 1987).

Line Widths. Direct extraction of coupling constants from COSY antiphase peaks is difficult due to the large inherent line widths compared to the scalar couplings. As we have previously found (Schmitz et al., 1990), establishing proper line widths for individual proton resonances is crucial for determining coupling constants. During simulation, a change in line width for signals alone ω_2 significantly affects the appearance of the cross-peaks and may also lead to deceptive similarities of cross-peak patterns for different J coupling values. Consequently, establishing the proper line width values for protons involved in scalar coupling becomes necessary for reliable evaluation of coupling constants via simulation of 2QF-COSY cross-peaks. A range of line width values was chosen following a protocol developed in our earlier studies (Gochin et al., 1990), and experimental data were matched with simulations using various line widths. The line width of the H1′ proton resonances was established from the H1′(ω_2)–H2′(ω_1) peak. In experimental H1′(ω_2)–H2′(ω_1) peaks, the pair of in-phase components along ω_2 due to passive coupling are well-resolved, indicating a smaller line width for H1′ peaks; a line width of 4 Hz for H1′ was initially chosen on the basis of the simulations. In subsequent simulations, it was noted that local variations in coupling constant and line width can occur. So line widths were readjusted during final extraction of J couplings for individual nucleotides. However, only minor adjustments of line width are found to be necessary in most cases which did not affect the determined coupling constant values within experimental error.

For 2′ and 2″ proton line width estimations, cross-peak patterns entailing H1′ were followed on changing line width values. In the H1′(ω_2)–H2′(ω_1) peak, the resolution of the outer pair of in-phase components along ω_2, which arises from passive coupling between 2′ and 3′ protons, depends on the H2′ line width. Comparison of simulated and experimental cross-peaks for various residues led to a line width of 8 Hz for most H2′ protons. A 16-peak pattern in all H1′(ω_2)–H2″(ω_1) cross-peaks indicates a slightly smaller line width

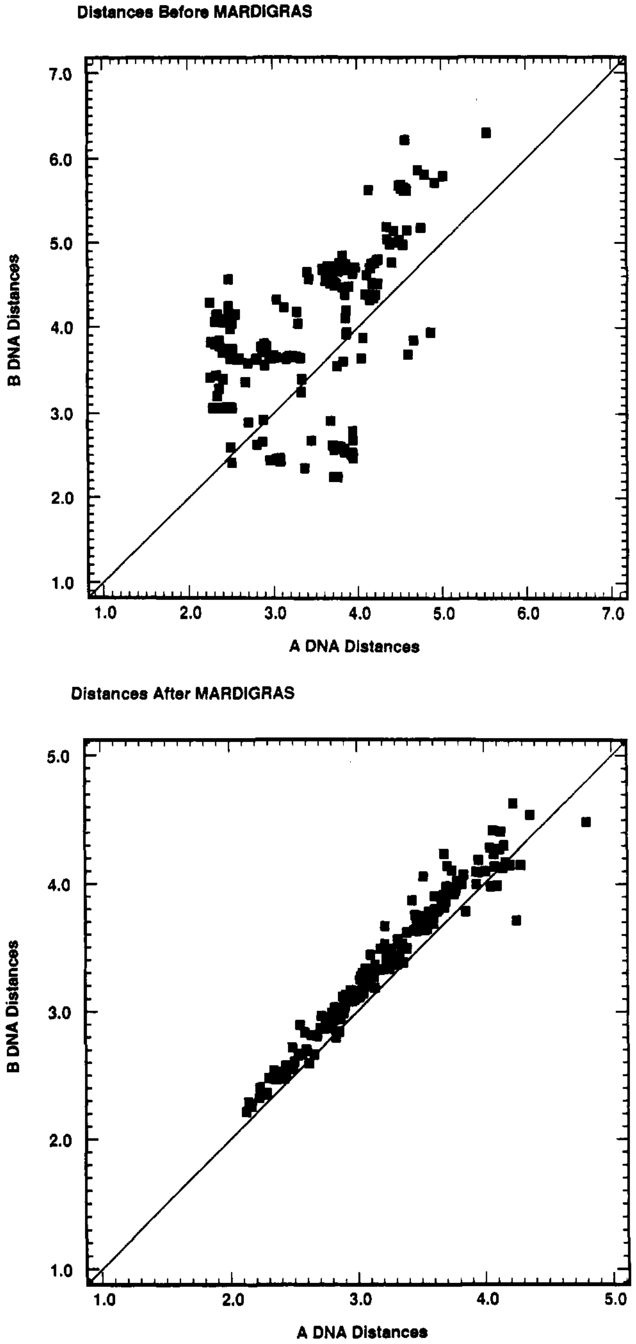

FIGURE 5: Comparison of interproton distances for [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] in energy-minimized A-DNA and B-DNA conformations (A, top) and for corresponding MARDIGRAS-derived distances using energy-minimized A- and B-DNA starting models (B, bottom). The 120-ms 2D NOE data set was used.

for H2″ than H2′ resonances; for H2″ resonances, a line width of 7 Hz was generally determined. However, for C8 and C25, smaller line widths of 6 and 7 Hz were respectively found for H2′ and H2″ resonances. For terminal residues, line widths were observed to be significantly reduced, indicating considerable fraying at the ends of the duplex. Simulated fitting of cross-peaks from the T26 and G13 terminal residues were nevertheless successful.

Simulation of Cross-Peak Fine Structure. The experimental 2QF-COSY cross-peaks (Figure 7; also see Supplementary Material) exhibit fine structure variations between individual residues; these pattern variations can be attributed to local variations in coupling constants and, to a small extent, line widths. Most of the H2′(ω_2)–H1′(ω_1) cross-peaks exhibit four outer components along ω_2, but the intensities of the central components are quite variable; for example, the internal components are completely absent at this noise level for the T9 residue but observable and well-resolved for the T15 residue. Phase changes in the internal components of H2′(ω_2)–H1′(ω_1) cross-peaks are also observable for certain residues. For purines these variations occur less often than for pyrimidines; in general, the H2′(ω_2)–H1′(ω_1) cross-peaks arising from guanidines manifest fewer variations than cross-peaks from thymidines. For H2″(ω_2)–H1′(ω_1) cross-peaks, all residues display similar fine structure, having four resolved components of almost equal intensity along the ω_2 dimension. For signals manifesting three nondegenerate scalar couplings, one expects an eight-component multiplet in the ω_2 dimension. But none of the experimental cross-peaks exhibits this pattern. This is due to amalgamation and cancellation effects which characteristically arise from overlap of positive and negative components of multiplets. Consequently, the observed peak-to-peak separations do not directly yield the coupling constants (Celda et al., 1989). Some H2″(ω_2)-H1′(ω_1) cross-peaks possess additional small signals on both sides of cross-peaks in the ω_1 dimension due to FID truncation effects (Widmer & Wüthrich, 1987). In addition to truncation, the actual line width of individual proton resonances, acquisition times, and apodization functions also contribute to this effect. Since experimental parameters are already defined, these wiggles helped to determine the line width of the proton resonances involved.

Simulations were carried out for a model with a rigid sugar pucker. Coupling constant values for various pseudorotation phase angles *P*, varied in 9° steps from 0 to 360°, and a pucker amplitude ϕ_m of 35° were taken from those reported by Rinkel and Altona (1987). During these simulations, line widths were also varied for different resonances as noted above in the range discussed in the previous section. However, comparison with experimental cross-peaks indicated that a single static sugar conformation cannot account for the experimental data; simulated cross-peaks for a single conformer do not provide a good fit to the experimental set of cross-peaks for any value of *P*. Consequently, rapid interconversion between two sugar conformers, i.e., a dynamic two-state model, was considered. This is commonly employed with DNA using one conformer from the S and one from the N region of the pseudorotation circle. In this model, the two conformation states are defined: a minor conformer (N), with pseudorotation phase angle P_N = 9° and pucker amplitude ϕ_m^N = 35°, i.e., essentially C3′-endo, and a dominant S conformer with the same pucker amplitude and pseudorotation phase angle P_S corresponding to approximately the C2′-endo range of puckers but subject to some variation (Altona & Sundaralingam, 1972; Rinkel & Altona, 1987). Cross-peak simulations were carried out using values of P_S ranging from 117° to 225° in 18° steps. Simulations were performed with varying fractions of N and S conformer as well; the resultant coupling constants were calculated as

$$J_{ab} = X_N J_{ab}(N) + X_S J_{ab}(S)$$

where X_N and X_S are the fractional populations of N and S conformers, respectively, and J_{ab}(N) and J_{ab}(S) are the respective vicinal coupling constants between proton a and proton b for the N and S conformers. The fractional population of S conformer (X_S) was systematically varied in steps of 0.05 (i.e., 5%).

Simulated fits with the mixture of two conformers were successful. Coupling constants and other pucker parameters

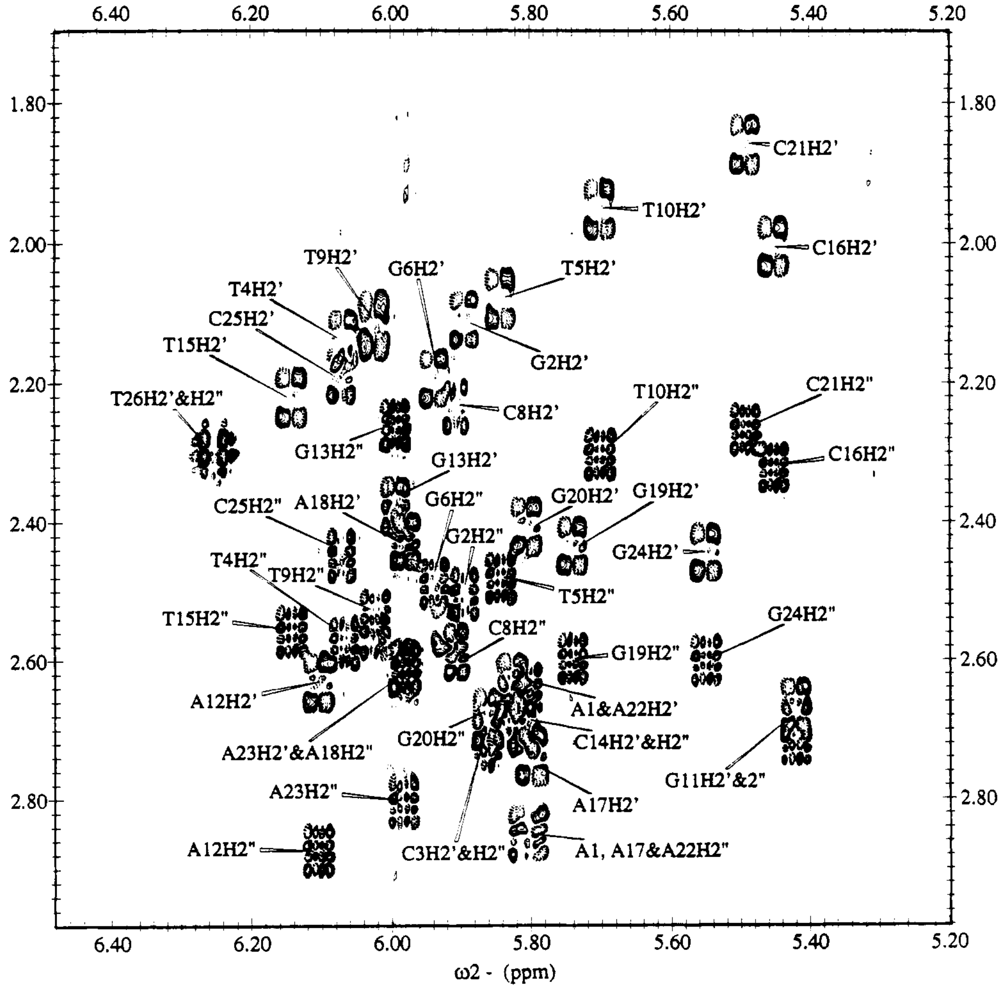

FIGURE 6: Section of the double-quantum-filtered COSY spectrum of [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] showing the H1'(ω_2)–H2'/H2"(ω_1) region.

for different residues are listed in Table III. The listed coupling constant values correspond to the sugar pucker parameters used in the simulation which gave the best fit to experimental cross-peaks. Consequently the accuracy of the sugar pucker parameters can be estimated to be ±9° for P_S and ±5% for conformer populations. The pseudorotation parameters of the minor conformer are only approximations, of course. Simulations could not be made in the case of some resonances due to severe overlapping of their cross-peaks; no values are listed for these in Table III. However, estimates (or at least limits) in their puckers could conceivably be made on the basis of their few observed (or partially overlapped) cross-peak patterns. Figure 7 compares plots of experimental and simulated 2QF-COSY cross-peaks for individual nucleotides. The sugar conformations found in the present study can be categorized into three types with P_S values of 144°, 162°, or 180°. Results indicate a large number of nucleotide puckers centered around P_S = 144–162°. Except for terminal base pairs, the S conformer populations strongly dominates with X_S >75% and generally >85%. These general characteristics in particular pertain to purine residues.

The inner components are nearly absent in the H2'(ω_2)–H1'(ω_1) cross-peak of T4, while T5 displays weak inner components on the right side. During simulations, an increase in line width by 1 Hz for H2' was found to lower the intensity of left side inner components to nearly zero, while an increment in the fractional population of S conformer from 80 to 85% lowers the overall intensity of the right inner components. It should be noted that these effects may also occur due to changes in pucker amplitude (Schmitz et al., 1990). On the other hand, ϕ_m is not expected to deviate much, so the value of ϕ_m was set to 35° for nearly all simulations. Puckers of T9 and T10 were found to be centered around 144° for the S conformer, and the fractional population of S conformer was found to be 75% for T9 and 85% for T10, manifesting small differences in coupling constants. The H1'(ω_2)–H2"(ω_1) cross-peak of C8 is different from other H1'(ω_2)–H2"(ω_1) cross-peaks observed in the trisdecamer, as it is collapsed to an eight-peak pattern, instead of a 16-peak pattern. Collapse of the 16-peak pattern can occur if P_S is 216° or higher. However, patterns arising from other cross-peaks, although suffering from some overlap, are not consistent with a higher pseudorotation angle. Careful simulations with a grid of different coupling constant values, line widths, and pucker amplitude lead to satisfactory matching of cross-peaks. As noted above, in this case, smaller line widths of 6 and 7 Hz were found for H1' and H2" resonances; this, together with an increase in pucker amplitude to 37° gave the best fit to the observed cross-peak patterns. However, partial overlap for some of the C8 cross-peaks leads us to caution that the precise sugar conformation could not well-established. Consequently, a broader range of pucker parameters is reported (Table III).

All nucleotides yielding P_S = 180° are purines: guanines G19, G20, and G24 with X_S = 85% and A18 with X_S = 95%. The terminal nucleotide T26, with a phase angle P_S = 162°, was fit with a lower fraction of S conformer, 65%, typical of terminal nucleotides subject to increased conformational flexibility due to limited fraying (Celda et al., 1989; Schmitz

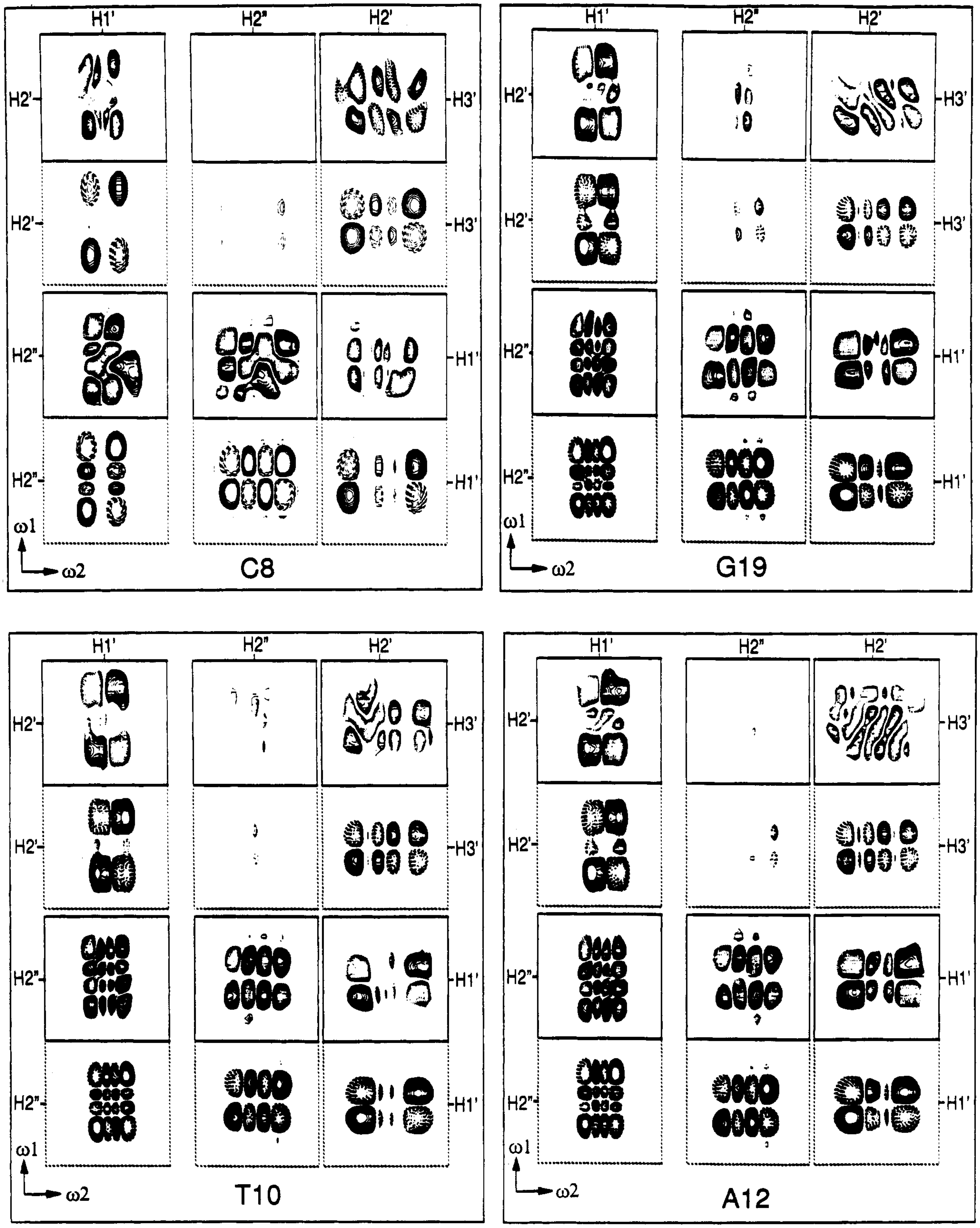

FIGURE 7: Plots of experimental (solid boxes) and simulated (broken boxes) double-quantum-filtered COSY cross-peaks for some representative spectral patterns observed for nucleotides in [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)]. Negative peaks are shaded for experimental peaks and dashed for simulated peaks. Note that the experimental cross-peaks displayed in some cases also exhibit partial cross-peaks from other nearby or overlapping cross-peaks. The pucker parameters and corresponding coupling constants for the best fits are listed in Table III.

et al., 1990).

Comparison of Derived Torsion Angle and Distance Constraints. To a limited extent we can compare the structural constraints derived from 2QF-COSY data to those from 2D NOE data. Among intranucleotide distances, the distance corresponding to the H1′–H4′ NOE contact is most sensitive to pseudorotation angle (Gochin & James, 1990; Lane, 1990; Wüthrich, 1986). Consequently, it should be possible to

Table III: Description of Sugar Puckers Found from 2QF-COSY Cross-Peak Simulations in d(AGCTTGCCTTGAG)·(CTCAAGGCAAGCT)[a]

nucleotide	P_S	%S	%N	$J_{1'-2'}$	$J_{1'-2''}$	$J_{2'-3'}$	$J_{2''-3'}$	$J_{3'-4'}$
A1								
G2	162	75	25	8.0	6.3	6.2	3.3	3.1
C3								
T4	135	85	15	8.9	6.0	7.2	2.5	3.7
T5	135	85	15	8.9	6.0	7.2	2.5	3.7
G6	162	75	25	8.0	6.3	6.2	3.3	3.1
C7								
C8[b]	135–144	70–80	30–20	6.0	6.1	7.0	2.8	3.2
T9	144	75	25	8.1	6.2	6.8	3.3	3.8
T10	144	85	15	8.9	6.0	6.7	2.5	3.2
G11								
A12	162	85	15	8.8	6.0	6.0	2.5	2.4
G13	144	95	05	9.8	5.4	6.6	1.6	2.6
C14								
T15	162	85	15	8.8	6.0	6.0	2.5	2.4
C16	144	95	05	8.9	5.4	6.6	1.6	2.6
A17	162	85	15	8.8	6.0	6.0	2.5	2.4
A18	180	95	05	9.0	5.6	5.6	1.8	1.4
G19	180	85	15	8.3	6.1	5.8	2.6	2.1
G20	180	85	15	8.3	6.1	5.8	2.6	2.1
C21	144	85	15	8.9	6.0	6.7	2.5	3.2
A22								
A23	162	85	15	8.8	6.0	6.0	2.5	2.4
G24	180	85	15	8.3	6.1	5.8	2.6	2.1
C25	153	75	25	7.0	6.0	6.4	3.3	3.3
T26	162	65	35	7.1	6.5	6.3	4.2	3.7

[a] Pseudorotation angle for N conformor, P_N = 9° and pucker amplitude, ϕ_m = 35°. [b] Pucker amplitude, ϕ_m, was increased to 37° for best fits.

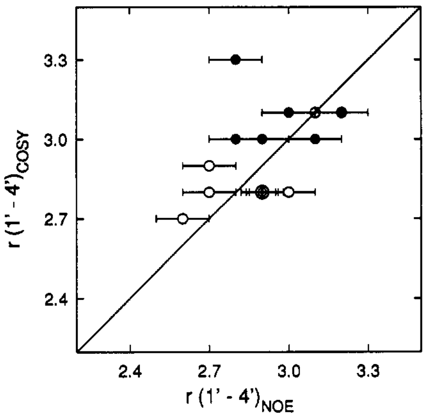

FIGURE 8: Plot of the effective distance (Å) between the H1′ and H4′ sugar protons of [d(AGCTTGCCTTGAG)·d-(CTCAAGGCAAGCT)] estimated from 2QF-COSY results as a function of the H1′–H4′ distance determined via MARDIGRAS analysis of experimental 2D NOE intensities. The effective distance estimated from 2QF-COSY utilized the r^{-6} distance averaging that occurs from conformational jumps between N and S conformers. The distances in each conformer were calculated using the pseudorotation angle for the conformer, P, as well as the sugar pucker amplitude ϕ_m. However, it should be noted that the effective distance calculated assuming conformational exchange differed from that of the major conformer by <4%. Filled and open circles denote purine and pyrimidine nucleotides, respectively. Error bars for the NOE distances are those calculated from MARDIGRAS.

compare H1′–H4′ distances derived from MARDIGRAS analysis of the 2D NOE spectra with H1′–H4′ distances estimated from the conformational parameters extracted from analysis of the 2QF-COSY spectra. Such a comparison is made in the rather expanded plot in Figure 8. Although the 2QF-COSY experiments indicate that the S conformer is strongly dominant and therefore largely establishes the effective H1′–H4′ distance in the plot, we did take into account the populations in the two-state model used as well as the r^{-6} distance averaging that occurs. The effective distance calculated differed by at most 4% from the distance in the major conformer. From Figure 8, it is clear that a correlation exists. The correlation is not wonderful though due to errors associated with both distance and torsion angle determinations. In particular, the most deviant outlier arises from A18; the cross-peaks for that residue are significantly overlapped. Of course, our subsequent use of any structural constraints derived from analysis of such overlapped peaks would be cognizant of that limitation.

It will be noted in Figure 8 that the H1′–H4′ distances for pyrimidines are generally lower than for purines using either 2D NOE or 2QF-COSY analysis; an obvious exception is from the terminal T26 with an H1′–H4′ distance of 3.1 Å. This correlation is basically associated with lower P_S for pyrimidines. This is consistent with results found for other oligonucleotides studied in this lab (Gochin et al., 1990; Schmitz et al., 1992; Stolarski et al., 1992; Weisz et al., 1992).

CONCLUSIONS

We have been successful in our search for a highly conserved region of the HIV-1 genome's long terminal repeat as a molecular target for drug design. However, the function of this highly conserved sequence is not known. The 13-base-pair deoxyribonucleoside duplex, [d(AGCTTGCCTTGAG)·d-(CTCAAGGCAAGCT)], selected for experimental study has proven amenable to quantitative analysis of torsion angle and distance constraints from 2D NMR spectra. The 2D NOE experimental data obtained at three mixing times were subjected to processing via the maximum likelihood method (MLM) in addition to the more common Fourier transform processing to yield quantitative cross-peak intensities with slightly better resolution. The iterative relaxation matrix algorithm MARDIGRAS, used for analysis of the 2D NOE intensity data, has yielded 244 accurate distance constraints, i.e., 7–11 experimental distance constraints per residue including interresidue and interstrand distances. Simulated fitting of cross-peaks in the 2QF-COSY spectrum of the 13-mer using the program SPHINX/LINSHA indicated that a single conformer was inadequate to describe any of the sugar puckers. However, a rapid two-state equilibrium with one conformer strongly dominant (75–95%) did enable of good fit of 2QF-COSY cross-peaks, yielding sugar pucker torsion angle constraints in 19 of the 24 nonterminal nucleotides. The other five nucleotides had too many overlapping cross-peaks for the sugar pucker to be well-described. The sugar pucker of the major conformer exhibited significant variability for the various nucleotides but was roughly 2′-endo in each nucleotide. Though derived independently and subject to different time-averaging effects, the 2QF-COSY and 2D NOE results are in accord.

The large number of structural parameters, torsion angles, and internuclear distances obtained in the present study will be used as constraints in subsequent structural refinement. Restrained molecular dynamics calculations should yield more insight into the dynamic solution structure of the potential drug target.

ACKNOWLEDGMENT

We gratefully acknowledge the help of Dr. Anil Kumar with the initial MLM calculations and useful discussions with Drs. Klaus Weisz and Uli Schmitz regarding extraction of torsion angle constraints.

SUPPLEMENTARY MATERIAL AVAILABLE

Sections of the double-quantum-filtered COSY spectrum of [d(AGCTTGCCTTGAG)·d(CTCAAGGCAAGCT)] showing (Figure A) the H2′/H2″(ω_2)–H1′(ω_1) region and (Figure B) the H2′/H2″(ω_2)–H3′(ω_1) region. Plots (Figure C) of experimental (solid boxes) and simulated (broken boxes) double-quantum-filtered COSY cross-peaks for the nucleotides not shown in Figure 7 (the pucker parameters and corresponding coupling constants for the best fits are listed in Table III) (6 pages). Ordering information is given on any current masthead page.

REFERENCES

Ajito, K., Atsumi, S., Ikeda, D., Kondo, S., Takeuchi, T., & Umezawa, K. J. (1989) *J. Antibiot. 42*, 611–619.

Altona, C., & Sundaralingam, M. (1972) *J. Am. Chem. Soc. 94*, 8205–8212.

Atsumi, S., Muraoka, Y., Nogami, T., Hoshino, H., Takeuchi, T., & Umezawa, K. J. (1988) *Drugs Exp. Clin. Res. 14*, 719.

Bebenek, K., Abbotts, J., Roberts, J. D., Wilson, S. H., & Kunkel, T. A. (1989) *J. Biol. Chem. 264*, 16948–16956.

Boelens, R., Koning, T. M. G., & Kaptein, R. (1988) *J. Mol. Struct. 173*, 299–311.

Borgias, B. A., & James, T. L. (1988) *J. Magn. Reson. 79*, 493–512.

Borgias, B. A., & James, T. L. (1989) in *Methods in Enzymology, Nuclear Magnetic Resonance, Part A: Spectral Techniques and Dynamics* (Oppenheimer, N. J., & James, T. L., Eds.) Vol. 176, pp 169–183, Academic Press, New York.

Borgias, B. A., & James, T. L. (1990) *J. Magn. Reson. 87*, 475–487.

Borgias, B. A., Gochin, M., Kerwood, D. J., & James, T. L. (1990) in *Progress in Nuclear Magnetic Resonance Spectroscopy* (Emsley, J. W., Feeney, J., & Sutcliffe, L. H., Eds.) Vol. 22, pp 83–100, Pergamon Press, Oxford.

Broido, M. S., Zon, G., & James, T. L. (1984) *Biochem. Biophys. Res. Commun. 119*, 663–670.

Celda, B., Widmer, H., Leupin, W., Chazin, W. J., Denny, W. A., & Wüthrich, K. (1989) *Biochemistry 28*, 1462–1470.

Chou, S.-H., Wemmer, D. E., Hare, D. R., & Reid, B. R. (1984) *Biochemistry 23*, 2257–2562.

Cooney, M., Czernuszewicz, G., Postel, E. H., Flint, S. J., & Hogan, M. E. (1988) *Science 241*, 456–459.

Delassus, S., Cheynier, R., & Wain-Hobson, S. (1991) *J. Virol. 65*, 225–231.

De Leys, R., Vanderborght, B., Haedevelde, M., Heyndrickx, L., van Geel, A., Wauters, C., Bernaerts, R., Saman, E., Nijs, P., Wellems, B., Taelman, H., van der Groen, G., Piot, P., Tersmette, T., Huisman, J. G., & von Heuverswyn, H. J. (1990) *J. Virol. 64*, 1207–1216.

Dixon, D. W., Schianazi, R., & Marzilli, L. G. (1990) *Ann. N.Y. Acad. Sci. 616*, 511–513.

Feigon, J., Denny, W. A., Leupin, W., & Kearns, D. R. (1983) *Biochemistry 22*, 5930–5942.

Freeman, R., Kempsell, S. P., & Levitt, M. H. (1980) *J. Magn. Reson. 38*, 453–479.

Gilboa, E., Mitra, S. W., Goff, S., & Baltimore, D. (1979) *Cell 18*, 93–100.

Gochin, M., & James, T. L. (1990) *Biochemistry 29*, 11172–11180.

Gochin, M., Zon, G., & James, T. L. (1990) *Biochemistry 29*, 11161–11171.

Hoch, J. C. (1989) in *Methods in Enzymology, Nuclear Magnetic Resonance, Part A: Spectral Techniques and Dynamics* (Oppenheimer, N. J., & James, T. L., Eds.) Vol. 176, pp 216–241, Academic Press, New York.

Hore, P. J. (1989) in *Methods in Enzymology, Nuclear Magnetic Resonance, Part A: Spectral Techniques and Dynamics* (Oppenheimer, N. J., & James, T. L., Eds.) Vol. 176, pp 64–77, Academic Press, New York.

Huet, T., Cheynier, R., Meyerhans, A., Roelants, G., & Wain-Hobson, S. J. (1990) *Nature 345*, 356–359.

James, T. L. (1991) *Curr. Opin. Struct. Biol. 1*, 1042–1053.

James, T. L., & Basus, V. J. (1991) *Annu. Rev. Phys. Chem. 42*, 501–542.

Jeong, G. W., Borer, P. N., Wang, S. S., Kumar, A., & Levy, G. C. (1992) Quantitation with the Maximum Likelihood Method, *J. Magn. Reson.* (in press).

Ji, J., & Loeb, L. A. (1992) *Biochemistry 31*, 954–958.

Johnson, P. F., & McKnight, S. L. (1989) *Annu. Rev. Biochem. 58*, 799–839.

Jones, K. A., Luciw, P. A., & Duchange, N. (1988) *Genes Dev. 2*, 1101–1114.

Keepers, J. W., & James, T. L. (1984) *J. Magn. Reson. 57*, 404–426.

Keller, W., Brenroth, S., Lang, K. M., & Christofori, G. (1991) *EMBO J. 10*, 4241–4249.

Kerwin, S. M., Kuntz, I. D., & Kenyon, G. L. (1991) *Med. Chem. Res. 1*, 361.

Kerwood, D. J., Zon, G., & James, T. L. (1991) *Eur. J. Biochem. 197*, 583–595.

Lane, A. N. (1990) *Biochim. Biophys. Acta 1049*, 189–204.

Larder, B. A., Darby, B., & Richman, D. D. (1989) *Science 243*, 1731–1734.

Levinger, L. F., & Lawtenberg, J. A. (1987) *Eur. J. Biochem. 166*, 519–526.

Li, M.-X., Yeung, H.-W., Pan, L.-P., & Chan, S. I. (1991) *Nucleic Acids Res. 19*, 6309–6312.

Liu, H., Thomas, P. D., & James, T. L. (1992) *J. Magn. Reson.* (in press).

Lown, J. W., Krowicki, K., Balzarini, J., Newman, R. A., & de Clercq, E. J. (1989) *J. Med. Chem. 32*, 2368–2375.

Madrid, M., Llinas, E., & Llinas, M. (1991) *J. Magn. Reson. 93*, 329–346.

Marion, D., & Wüthrich, K. (1983) *Biochem. Biophys. Res. Commun. 113*, 967–974.

Muesing, M. A., Smith, D. H., & Capon, D. J. (1987) *Cell 48*, 691–701.

Myers, G. (1990) *Human Viruses and AIDS*, Los Alamos National laboratory, Los Alamos, NM.

Myers, G. (1991) *Human Viruses and AIDS*, Los Alamos National Laboratory, Los Alamos, NM.

Nakamura, H., Yamamoto, N., Inoue, Y., & Nakamura, S. (1987) *J. Antibiot. 40*, 396–399.

Ni, F., & Scheraga, H. A. (1989) *J. Magn. Reson. 82*, 413–418.

Oppenheimer, N. J., & James, T. L. (1989a) *Methods in Enzymology, Nuclear Magnetic Resonance, Part B: Structure and Mechanism*, Vol. 177, Academic Press, New York.

Oppenheimer, N. J., & James, T. L. (1989b) *Methods in Enzymology, Nuclear Magnetic Resonance, Part A: Spectral Techniques and Dynamics*, Vol. 176, Academic Press, New York.

Pearlman, D. A., Case, D. A., Caldwell, J., Seibel, G. L., Singh, U. C., Weiner, P. K., & Kollman, P. A. (1991) *AMBER 4.0 (UCSF)*, University of California, San Francisco, CA.

Post, C. B., Meadows, R. P., & Gorenstein, D. G. (1990) *J. Am. Chem. Soc. 112*, 6796–6803.

Preston, B. D., Poiesz, B. J., & Loeb, L. A. (1988) *Science 242*, 1168–1171.

Rajagopal, P., Gilbert, D. E., Marel, G. A., Boom, J. H., & Feigon, J. (1988) *J. Magn. Reson. 78*, 526–537.

Rinkel, L. J., & Altona, C. (1987) *J. Biomol. Struct. Dyn. 4*, 621–649.

Roy, S., Parkin, N. T., Rosen, C., Itovich, J., & Sonenberg, N. (1990) *J. Virol. 64*, 1402–1406.

Scheek, R. M., Russo, N., Boelens, R., Kaptein, R., & Boom, J. H. (1983) *J. Am. Chem. Soc. 105*, 2914–2916.

Schmitz, U., Zon, G., & James, T. L. (1990) *Biochemistry 29*, 2357–2368.

Schmitz, U., Sethson, I., Egan, W., & James, T. L. (1992) *J. Mol. Biol.* (in press).

St. Georgieve, V., & McGowan, J. (1990) *Ann. N.Y. Acad. Sci. 616*, 1–10.

States, D. J., Haberkorn, R. A., & Ruben, D. J. (1982) *J. Magn. Reson. 48*, 286–292.

Stolarski, R., Egan, W., & James, T. L. (1992) *Biochemistry 31*, 7027–7042.

Suzuki, E.-I., Pattabiraman, N., Zon, G., & James, T. L. (1986) *Biochemistry 25*, 6854–6865.

Thomas, P. D., Basus, V. J., & James, T. L. (1991) *Proc. Natl. Acad. SCi. U.S.A. 88*, 1237–1241.

Varmus, H. E. (1988) *Science 240*, 1427–1434.

Wagner, G. (1990) in *Progress in Nuclear Magnetic Resonance Spectroscopy* (Emsley, J. W., Feeney, J., & Sutcliffe, L. H., Eds.) Vol. 22, pp 101–139, Pergamon Press, Oxford.

Weichs an der Glon, C., Monks, J., & Proudfoot, N. J. (1991) *Genes Dev. 5*, 244–253.

Weisz, K., Shafer, R. H., Egan, W., & James, T. L. (1992) *Biochemistry* (in press).

Widmer, H., & Wüthrich, K. (1987) *J. Magn. Reson. 74*, 316–336.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Zhou, N., Manogaran, S., Zon, G., & James, T. L. (1988) *Biochemistry 27*, 6013–6020.