Hardy, J. R., Harvey, V. J., Paxton, J. W., Evans, P., Smith, S., Grove, W., Grillo-Lopez, A. J., & Baguley, B. C. (1988) *Cancer Res. 48*, 6593-6596.
Karle, J. M., Cysyk, R. L., & Karle, I. L. (1980) *Acta. Crystallogr., Sect. B B36*, 3012-3016.
Liu, L. F. (1989) *Annu. Rev. Biochem. 58*, 351-375.
Low, C. M. L., Drew, H. R., & Waring, M. J. (1984a) *Nucleic Acids Res. 12*, 4865-4879.
Low, C. M. L., Olsen, R. K., & Waring, M. J. (1984b) *FEBS Lett. 176*, 414-420.
Muller, W., & Crothers, D. M. (1968) *J. Mol. Biol. 35*, 251-290.
Muller, W., Bunemann, H., & Dattagupta, N. (1975) *Eur. J. Biochem. 54*, 279-285.
Neidle, S., Webster, G. D., Baguley, B. C., & Denny, W. A. (1986) *Biochem. Pharmacol. 35*, 3915-3921.
Nelson, E. M., Tewey, K. M., & Liu, L. F. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1361-1364.
Prakash, A. S., Denny, W. A., Gourdie, T. A., Valu, K. K., Woodgate, P. D., & Wakelin, L. P. G. (1990) *Biochemistry 29*, 9799-9807.
Sakore, T. D., Reddy, B. S., & Sobell, H. M. (1979) *J. Mol. Biol. 135*, 763-785.
Searle, M. S., Hall, J. G., Denny, W. A., & Wakelin, L. P. G. (1988) *Biochemistry 27*, 4340-4349.
Smith, J. M., & Thomas, D. J. (1990) *CABIOS 6*, 93-99.
Suck, D., & Oefner, C. (1986) *Nature 321*, 620-625.
Suck, D., Lahm, A., & Oefner, C. (1988) *Nature 332*, 465-468.
Wadkins, R. M., & Graves, D. E. (1989) *Nucleic Acids Res. 16*, 9933-9946.
Wakelin, L. P. G., & Denny, W. A. (1990) *Molecular Basis of Specificity in Nucleic Acid-Drug Interactions* (Pullman, B., & Jortner, J., Eds.) pp 191-206, Kluwer Academic Press, Dordrecht, The Netherlands.
Wakelin, L. P. G., McFadyen, W. D., Walpole, A., & Roos, I. A. G. (1984) *Biochem. J. 222*, 203-215.
Wakelin, L. P. G., Atwell, G. J., Rewcastle, G. W., & Denny, W. A. (1987) *J. Med. Chem. 30*, 855-861.
Wakelin, L. P. G., Chetcuti, P., & Denny, W. A. (1990) *J. Med. Chem. 33*, 2039-2044.
Waring, M. J. (1976) *Eur. J. Cancer 12*, 995-1001.
Williams, L. D., Egli, M., Ughetto, G., van der Marel, G. A., van Boom, J. H., Rich, A., Wang, A. H.-J., & Frederick, C. A. (1990) *J. Mol. Biol. 215*, 313-320.
Wilson, W. D., Wang, Y. H., Kusuma, S., Chandrasekaran, S., Yang, N. C., & Boykin, D. W. (1985) *J. Am. Chem. Soc. 107*, 4989-4995.
Wilson, W. D., Tanious, F. A., Barton, H. J., Jones, R. L., Fox, K. R., Wydra, R. L., & Strekowski, L. (1990) *Biochemistry 29*, 8452-8461.
Wilson, W. R., Baguley, B. C., Wakelin, L. P. G., & Waring, M. J. (1981) *Mol. Pharmacol. 20*, 404-414.

A Simple Spectral-Driven Procedure for the Refinement of DNA Structures by NMR Spectroscopy†

Howard Robinson and Andrew H.-J. Wang*
Department of Physiology and Biophysics, University of Illinois at Champaign-Urbana, Urbana, Illinois 61801
Received October 28, 1991; Revised Manuscript Received January 30, 1992

ABSTRACT: We have developed a simple and quantitative procedure (SPEDREF) for the refinement of DNA structures using experimental two-dimensional nuclear Overhauser effect (2D NOE) data. The procedure calculates the simulated 2D NOE spectrum using the full matrix relaxation method on the basis of a molecular model. The volume of all NOE peaks is measured and compared between the experimental and the calculated spectra. The difference of the experimental and simulated volumes is minimized by a conjugated gradient procedure to adjust the interproton distances in the model. An agreement factor (analogous to the crystallographic *R*-factor) is used to monitor the progress of the refinement. The procedure is an iterative one. The agreement is considered to be complete when several parameters, including the *R*-factor, the energy associated with the molecule, the local conformation (as judged by the sugar pseudorotation), and the global conformation (as judged by the helical *x*-displacement), are refined to their respective convergence. With the B-DNA structure of d(CGATCG) as an example, we show that DNA structure may be refined to produce calculated NOE spectra that are in excellent agreement with the experimental 2D NOE spectra. This is judged to be effective by the low *R*-factor of ~15%. Moreover, we demonstrate that not only are NOE data very powerful in providing details of the local structure but, with appropriate weighting of the NOE constraints, the global structure of the DNA double helix can also be determined, even when starting with a grossly different model. The reliability and limitations of a DNA structure as determined by NMR spectroscopy are discussed:

There has been a growing awareness that DNA molecules are highly polymorphic. It has been unequivocally shown that in addition to B-DNA a number of stable alternative DNA conformations also exist (Rich et al., 1984; Palecek, 1991). They may play important roles in biological systems. A fuller understanding of the biological function requires the detailed knowledge of the three-dimensional structure of DNA and its potential in adopting various conformations, some of which are yet to be uncovered.

In the past, X-ray crystallography has been the principal tool for determining the three-dimensional structure of DNA molecules. Its greatest virtue is that a carefully refined crystal

† This work was supported by grants from the NIH (GM-41612 and CA-52506) to A.H.-J.W.
* Corresponding author.

0006-2960/92/0431-3524$03.00/0

structure of oligonucleotides with defined nucleotide sequence provides a reliable and detailed view of not only the three-dimensional structure of the molecules but also of many solvent molecules and ions surrounding the oligonucleotides. A high-resolution (1.0–2.0-Å) crystal structure can provide distance information between atoms of the order of 0.02–0.05-Å accuracy. However, a major drawback of X-ray crystallography is that a suitable crystal has to be produced. This limitation has hampered the investigation of the 3D structure of many important biological molecules.

Recently, high-resolution nuclear magnetic resonance (NMR) spectroscopy has emerged as a powerful tool in the analysis of molecular structure in solution (Wüthrich, 1986). In contrast to the X-ray diffraction method, NMR spectroscopy does not require a crystal, which means that NMR may be applied to many more biological systems. In this method, the information about the three-dimensional molecular structure is implicity encoded in the hundreds of nuclear Overhauser effect (NOE) relaxation cross peaks that are observed in a proton two-dimensional NOE spectrum. The size of the NOE cross peaks may be used to deduce the interproton distances from which a three-dimensional structure is built. Until recently, a common approach has been to classify the NOE cross peaks according to their peak strength into three categories: strong, medium, and weak, corresponding to interproton distances in the range of 2–3, 3–5, and over 5 Å, respectively. The distance information is used as the input for various constrained refinement procedures (Oppenheimer & James, 1989). However, those approaches do not compare the interproton distance information of the model that is being refined with the experimental NOE data in a direct and quantitative manner.

It has been shown (Solomon, 1955) that the rate (ρ_{ij}) of the spin cross relaxation between protons i and j is inversely proportional to sixth power of the distance between the two protons, r_{ij}.

$$\rho_{ij} \propto r_{ij}^{-6} F(\tau_c) \tag{1}$$

where $F(\tau_c)$ is a function of the rotational correlation time of the molecule. In a molecule, all proton pairs are relaxing each other simultaneously during the mixing time, τ_m. Thus, the relation between the NOE cross-peak intensity matrix **I** and the relaxation rates can be expressed in matrix notation as

$$\mathbf{I} \propto e^{-R\tau_m} \tag{2}$$

where the **R** matrix contains all the cross- and self-relaxation rate terms (Macura & Ernst, 1980). The matrix **R** can be calculated for a particular set of coordinates with a correlation time. Thus, the NOE cross-peak intensity matrix **I** can also be calculated for a particular mixing time. The simulated intensities derived from a model can then be compared with the experimental set of intensities, and a residual error factor can be produced which evaluates the match between these intensities (Gonzalez et al., 1991). In an inverse manner, the relaxation matrix can be derived directly from the experimental intensities. Equation 1 can then be used to calculate r_{ij}'s from the NOE-derived relaxation matrix. Calculations of these r_{ij}'s from just experimental data have been found to not work well, and several strategies have been developed that circumvent this difficulty such as IRMA (Boelens et al., 1989), MORASS (Post et al., 1990), and MARDIGRAS (Borgais et al., 1990). These programs calcualte the r_{ij}'s from an intensity matrix that is a hybrid combination of simulated and experimental intensities. Those experimental intensities that are within a certain tolerance of the simulated intensities are included in the calculation with the simulated intensities. In effect, the analysis of the simulated intensities is perturbed by the inclusion of experimental NOE intensities. By adjusting the model on the basis of the hybrid-calculated r_{ij}'s, the model converges toward the available experimental data. These issues have been elegantly reviewed recently (Clore & Gronenborn, 1989; James, 1991).

We have chosen an alternative to this hybrid-intensity matrix procedure. Our approach is to keep the iterative refinement procedure returning to reevaluate the entire observed 2D NOE spectrum as outlined in Figure 1. Analysis of the two spectra drives adjustments in the model until convergence. With the B-DNA structure of d(CGATCG) as an example, we show that DNA structure may be refined to produce calculated NOE values that are in excellent agreement with the experimental NOE values. This is judged to be effective by the good agreement factor (analogous to the crystallographic R-factor) of ~15%. Moreover, we demonstrated that not only are the NOE data very powerful in providing the detailed local structure but with sufficient weighting of NOE strengths the global structure of DNA double helix can largely be determined even starting with a grossly different model.

Experimental Procedures

Sample Preparation. DNA oligonucleotides were prepared according to a published procedure (van der Marel et al., 1981). The NMR sample was prepared by dissolving the lyophilyzed powder (13.3 mg) of the ammonium form of d(CGATCG) in 500 μL of H_2O with 50 mM sodium phosphate buffer (pH 7.0) and 150 mM NaCl. The solution was lyophilyzed on a Speedvac, and the dried material was dissolved in 500 μL of 99.8% D_2O and lyophilyzed again. This step was repeated. The sample was then dissolved in 300 μL of 99.8% D_2O, centrifuged to remove any precipitate, and the supernate loaded into an NMR tube (Wilmad no. 528PP). The solution in the NMR tube was dehydrated under a slow stream of dry argon gas, delivered to 5 mm above the liquid level by a Teflon capillary tube. Finally, the 6.7 mM duplex sample was prepared by adding 500 μL of freshly opened 99.96% D_2O (Aldrich Chemical Co.), and the tube was capped and sealed. Another sample with a duplex concentration of 2.4 mM was prepared in a similar manner. Unless otherwise stated, the sample with 6.7 mM duplex concentration was used.

Data Collection. The 1D and 2D proton NMR spectra were collected on a GN-500 NMR spectrometer. For the 2D data sets, 512 FIDs were collected twice by the method of States et al. (1982) to allow the extraction of phase information from the t_1 time dimension. Each FID consisted of 2048 complex points and was the average of 32 scans. The 2D NOE data sets were collected with a mixing time of 200 ms and a recycle time of 4 s, which is about twice the T_1 relaxation times for all the spins except A3H^2 (Figure 1S). The data were transferred to a Silicon Graphics IRIS 4D/25 computer and processed by the program FELIX (version 1.1, Hare Research Inc.). Linear prediction was used to correct the first data points in t_1 and t_2. In both the t_1 and t_2 time domains, the NOE data was apodized to reduce truncation affects by having the last quarter of the FID smoothly attenuate to zero with a sine bell squared curve. The resulting FID was exponentially multiplied with a constant of 4 Hz. Special attentions were paid to the polynomial baseline correction in both frequency dimensions. This is particularly important as erroneous baselines would affect the accuracy in the measurement of the intensity of NOE cross peaks. All resonances have been assigned by the sequential assignment procedure (Hare et al., 1983) using the combination of COSY and NOESY infor-

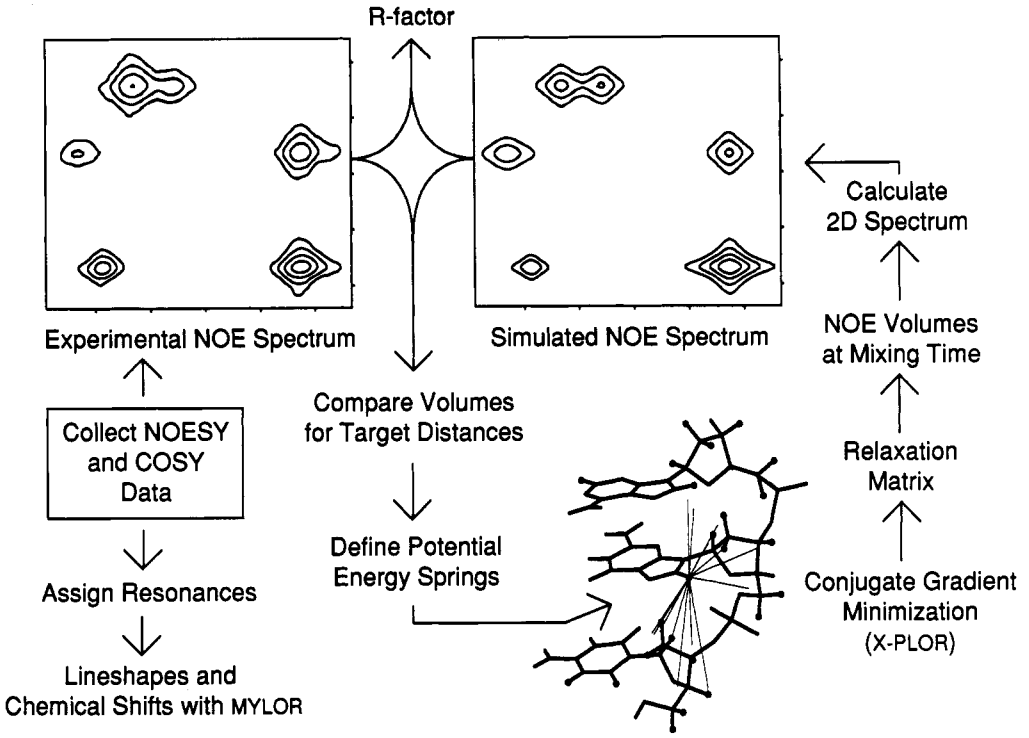

FIGURE 1: Steps involved in the cyclic refinement of the simulated 2D NOE spectra toward convergence with the observed spectrum. Deconvolution of the experimental cross peaks is based on the known convolution of the simulated 2D NOE spectrum. The model is adjusted with conjugate gradient minimization by attaching springs that represent the errors in the match between the observed and simulated NOE intensities. The balance between the NOE defined forces and the chemistry defined forces should be toward the NOE's so that the observed data can be expressed over the influences of chemistry potentials. A new simulated 2D NOE spectrum is calculated for the new model, and the refinement continues until convergence of the observed and simulated 2D NOE spectra.

mation. The chemical shifts (they are referenced to DSS through the DHO resonance) of all resonances are listed in Table IS in the supplementary material. Our assignments are in agreement with those of Lown et al. (1984) studied previously under similar conditions. At the mixing time of 200 ms, we have measured 703 NOE cross peaks for the 6.7 mM sample and 491 at 2.4 mM sample (out of $54 \times 53/2 = 1431$ possible peaks) whose intensity is considered to be above the background level. If the mixing time were very short, the signal to noise ratio would be low and also there would be very little spin diffusion. If the mixing time had been very long, the starting spins would migrate equally to all other spins. At some mixing time between these two extremes, there is a maximum in terms of the structural information encoded in the measured NOE's.

Refinement Strategy. In our approach (Figure 1), the goal is to derive models whose simulated 2D NOE spectra agree with the observed spectrum as well as possible, while retaining reasonable chemistry in the model. During the development of the refinement strategy, it was clear that the greatest limitation was that the observed data seriously underdetermined the structure (compared to the number of degrees of freedom). So any additional information that can be gleaned from the experimental data is of great importance. At each iteration of the refinement, we calculate a simulated 2D NOE spectrum with accurate chemical shifts and line shapes for all resonances. This allows automatic reevaluation of the deconvolution of the experimental 2D NOE spectrum. With the greatest loss of information occurring due to cross-peak overlap, this model-dependent deconvolution maximizes the information retrieval from the experimental data. Rather than using the NOE intensities to directly calculate the r_{ij}'s, we have used a procedure similar to that of Nilges et al. (1991), where an energy-based error function is established for each proton pair based on a comparison of the simulated and experimental intensities. The calculation of these energy functions can be rather sophisticated, as in the procedure of Nilges et al. (1991) where they calculate gradients of the errors. But the simplified first-order error estimates used in our study also work, since the procedure is a convergent one. The following is the description of various steps in the refinement procedure.

Model Building. A starting model, which may be derived from the crystal structure or by other means, is used for the calculation of the simulated NOE spectra. For the DNA hexamer (d(CGATCG) at neutral pH, it is reasonable to assume that the molecule adopts a conformation close to that of B-DNA. However, we have also used A-DNA as a starting model for comparison. Both the B-DNA and A-DNA models were constructed on the basis of the coordinates of the fiber DNA model (denoted as B_F and A_F) (Arnott et al., 1982). The models were initially relaxed without NOE restraints using program XPLOR [version 2.0 (Brünger, 1990)] with conjugate gradient minimization to obtain good geometries. XPLOR's all-atom force field for DNA was used as is, with explicit hydrogen-bond potentials (except that acceptor antecedent terms were excluded). No solvent or cations were used, and the net charges on each of the 10 phosphate groups were left at the default value of −0.33 eV. To compensate for the lack of solvent and cations, a half-distance dependence was used for the nuclear-centered monopolar electrical potentials, i.e., the Coulomb potential was multiplied by $1/(2r)$ with ϵ of 1. This compensating factor was found to be optimal for maintaining the gross shape of B-DNA with this force field. These fiber-derived energy-minimized models are denoted as B_{FEM} and A_{FEM} as shown in Figure 2A,B.

Simulated NOE Spectrum. The simulated NOE relaxation rates and NOE intensities for a set of atomic coordinates are calculated by the program MORASS developed by Gorenstein

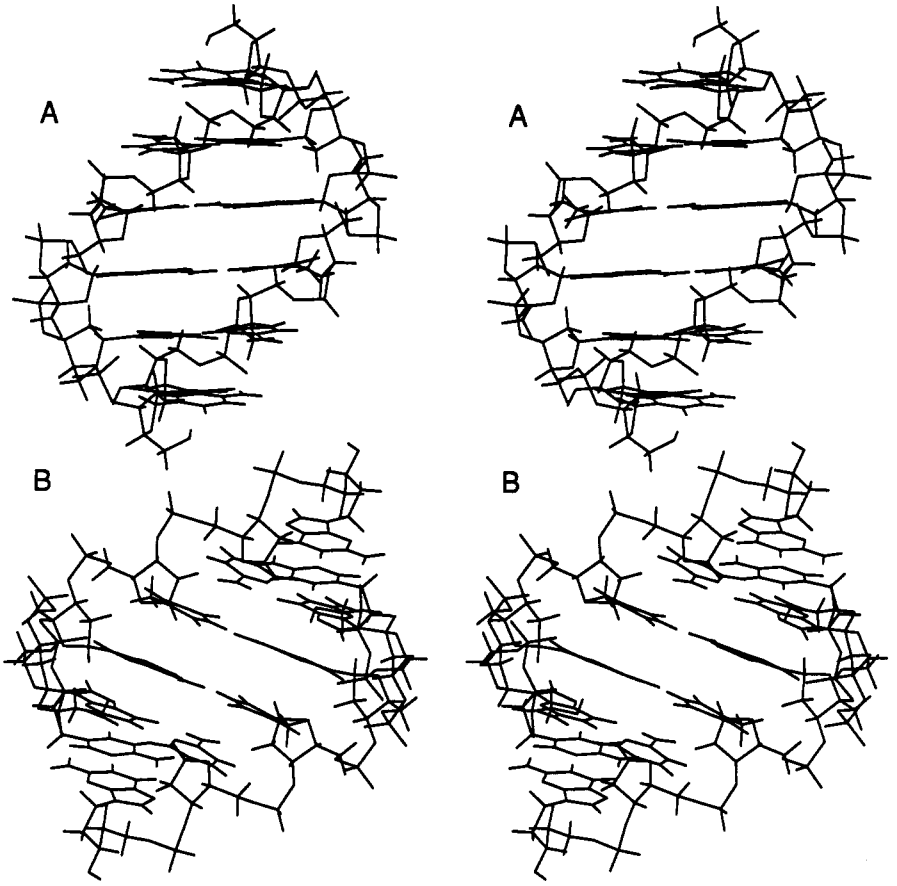

FIGURE 2: Stereodrawing of the d(CGATCG) duplex in the idealized fiber B-DNA (top) and A-DNA (bottom) conformations which have been energy-minimized using the program XPLOR. These models are denoted as B_{FEM} and A_{FEM}, respectively.

and his colleagues (Post et al., 1990). This full-matrix relaxation calculation uses a single model with rigid isotropic tumbling and generates relaxation rates between all pairs of spins. In this version of MORASS, methyl groups were handled as three protons located at their geometric mean. The rate matrix is then used to calculate the resulting NOE intensities at a particular mixing time, τ_m.

To calculate the simulated 2D NOE spectrum on the basis of these intensities, we must know accurate line shapes and chemical shifts for each spin. The program MYLOR has been written by us to facilitate an interactive determination of these characteristics for each spin. Since the observed data are apodized to retain Lorentzian line shapes, MYLOR allows the user to arrange multiple symmetric Lorentzians interactively on top of the observed line shape of each spin. The columns and rows of the observed 2D NOE spectrum are used as the source of the observed resonances for fitting. Although there often is severe overlap in the 1D spectrum, usually, slices through a 2D cross-correlation can be found where overlap from other cross-correlations is minimized. Mouse-driven actuators and sliders facilitate manual adjustments to the component Lorentzians and baselines. A visual overlay of the composite shape on the observed line helps direct these adjustments. Line shapes and chemical shifts are defined in this way for each spin in each frequency dimension. Another program (CSL) uses these line shapes and chemical shifts and the NOE intensities calculated from the simulation to produce a simulated 2D nOE spectrum for the model in FELIX-readable matrix format (see Figure 1).

Comparison of the Observed and Simulated 2D NOE Spectra. The 2D NOE spectrum is a superposition of all the cross-peak shapes with associated volumes centered at the pair-wise intersections for every spin. The dispersion of chemical shifts gives rise to considerable overlap of the cross peaks in the 2D spectrum. This means that the NOE intensity at a particular point in the 2D spectrum may be attributed to multiple NOE cross-correlations due to overlap. Given the known shape and position of each cross-correlation from the MYLOR program, the entire observed spectrum in principle could be deconvoluted into the sum of the component cross-correlation and diagonal volumes. We are developing a procedure along this line, but this effort is complicated by the difficulty of handling the evaluation of the indeterminacy where there is overlap.

As an alternative, we have devised an approach in which the differences of the NOE volumes between the observed and simulated 2D spectra are directly compared at each spin-pair intersection (see Figure 1) as the basis for adjusting the interproton distances. To reduce the influence of overlap, the NOE intensities are evaluated only for a very small region (one line width in each dimension) centered at each spin intersection. Thus, a rectangular integration pad of one line width is defined for every spin intersection, and the volume integral is evaluated. Since both the observed and simulated 2D NOE spectra are identically integrated at all spin intersections, the extreme truncation of the integrals can be accurately accounted for. In addition, the small integration limits provide for an improvement in the signal to noise ratio since areas of the 2D spectrum where the signals are likely to be very small relative to the noise are not included. A linear regression of all off-diagonal volume elements is used to produce a single scalar between the two sets of volume integrals (excluding the intrageminal volumes). We then compare the magnitude of the observed (the average of the two observations is taken) and

simulated volumes. If the observed integral is bigger, then those two protons should be pulled together; if the observed integral is smaller, then those two protons should be pushed apart. The sixth root of the ratio of the two volumes is used to establish a target distance from the current model's distance. Each spin pair's target distance is determined independently from the comparison of its volumes.

Through cycling of the refinement, this procedure should eventually cause the simulated spectrum to converge toward the observed spectrum. Where there is overlap, the stronger signals will overwhelm the small signals at the beginning. Thus, the overwhelmed signals will not refine until the larger signals have refined. Because of the possibility of oscillations, we diminish the target distance by 10% back toward the model to provide a damping effect.

Producing a New Model. The target distances so obtained are used as the equilibrium distances for potential energy springs as defined by Hook's law. In fact, we define two different springs (biharmonic) for each pair of protons, one for tension and one for compression. The compression spring is stronger to reflect the inverse sixth power distance dependence of the rate of NOE relaxation. In addition, the spring constants are loosened slightly for longer target distances to reflect the diminished signal-to-noise ratio with the weaker NOEs (force constants used for the biharmonic NOE's are listed in Table IIS in the supplementary material). The conjugate gradient minimizer of XPLOR is used with those potential energy springs as inputs for 50 minimization cycles, and the resulting new model becomes the starting model for the next cycle of the refinement. In the refinement cycle, the weighting of the NOE data relative to the chemistry contributions such as those from bond length, bond angles, dihedrals, Lenard-Jones and eletrostatics, may be varied. Several NOE strength regimes were tested in order to determine the optimum weighting for the NOE biharmonic potential springs relative to the chemistry potentials. We have tested three NOE strength regimes: strong, medium, and weak, with XPLOR biharmonic spring NOE scalars of 10, 20, and 60, respectively. With the strong NOE scalar, the NOE restraints far outweigh the chemical potentials; for the weak NOE scalar, the NOE restraints are about on a par with the chemical potentials. By using different regimes, we are able to reveal details of how the models respond to the NOE springs that have been placed on them and thus obtain a deeper understanding of the nature of information provided by the NOE's. To facilitate comparisons between the refinement strength regimes, weak weighting was used in the medium NOE strength regime for cycles 120–140. This change is evident as a discontinuity in Figures 3, 6, and 9 at cycle 120 for the medium NOE strength refinements.

Monitoring the Refinement Progress. The progress of the refinement is monitored by an NMR residual discrepancy factor (*R*-factor), analogous to the crystallographic *R*-factor:

$$R\text{-factor (NMR)} = \sum(|N_o - N_c|)/\sum N_o \qquad (3)$$

Here the N_o and N_c are the respective observed and calculated off-diagonal NOE volumes (cross-correlations within geminals and methyls have been excluded). A reasonable model should produce an *R*-factor near 20% or below. We have used this measure of the residual errors because it retains an analogy to the crystallographic *R*-factor.

A critical issue in the refinement procedure is how to decide that a globally refined "converged" structure has been achieved. In addition to the overall *R*-factors and the sugar puckers (both are sensitive to local structure but much less sensitive to global conformation), we have examined the *x*-displacement of the base pair. The *x*-displacement of the base pair position relative to the helix axis is a good indicator of the global structure of a nucleic acid double helix. In B-DNA, the *x*-displacement is 0.3 Å, i.e., the center of the base pair is very close to the helix axis. In contrast, the *x*-displacement in A-DNA is −4.8 Å, i.e., the center of the base pair is shifted 4.8 Å toward the minor groove. By monitoring the *x*-displacement, we can follow the change in the global structure.

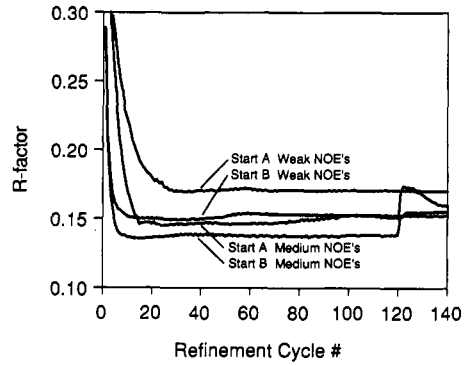

FIGURE 3: Progressive improvement of the overall *R*-factor (based on agreement between observed and simulated NOE's) during the refinement of the solution structure of d(CGATCG) with B_{FEM} and A_{FEM} as starting models. The refinements were carried out using the weak and medium NOE strength regimes.

There are other useful tools which we have developed to help locate the troublesome regions of the molecule being refined. The first is to calculate an individual *R*-factor associated with each spin. These individual *R*-factors should be evenly distributed about the mean (overall *R*-factor). Those that are not deserve more attention, as this may indicate problems with the resonance assignment or molecular conformation, or difficulties caused by relaxation or internal motion. The *R*-factor as computed above is dominated by the stronger shorter NOE's. To circumvent this, we calculate a distance-based dispersion of the residual errors based on the distances from the model. This removes the dominance caused by the discrepancy with the magnitudes of the NOE's, while maintaining the conventional definition of the *R*-factor.

The programs MYLOR and CSL and other programs needed to attach in FELIX, MORASS, and XPLOR for the SPEDREF procedure were written in C for the Silicon Graphics 4D series computers running IRIX 3.3 and will be made available to interested users.

RESULTS

Refinement Starting with the B-DNA Model. Using the procedure described above, the B_{FEM} model (Figure 2A) was subjected to 140 cycles of refinement using three NOE strength regimes. The progress of the refinements using weak and medium NOE strength constraints are indicated by the changes in *R*-factor shown in Figure 3. Only the results from the medium and weak NOE strength regimes are shown (as the refinement with strong NOE strength did not significantly improve the structure and only tended to distort the chemistry of the molecule). For the weak NOE scalar, it can be seen that the *R*-factor rapidly decreased from the initial 28% to ~15% within 10 cycles of refinement and stayed at the value thereafter until cycle 50 when there is a small increase in *R*-factor to 15.2%. There was no appreciable change in *R*-factor between cycle 50 and cycle 140.

In this refinement, the τ_c was set at 9.5 ns (an optimal value determined by the method described later). Figure 4A shows the refined model B_{RF140}, which is very similar to the starting B_{FEM} model as evident by the small RMSD (0.87 Å) between

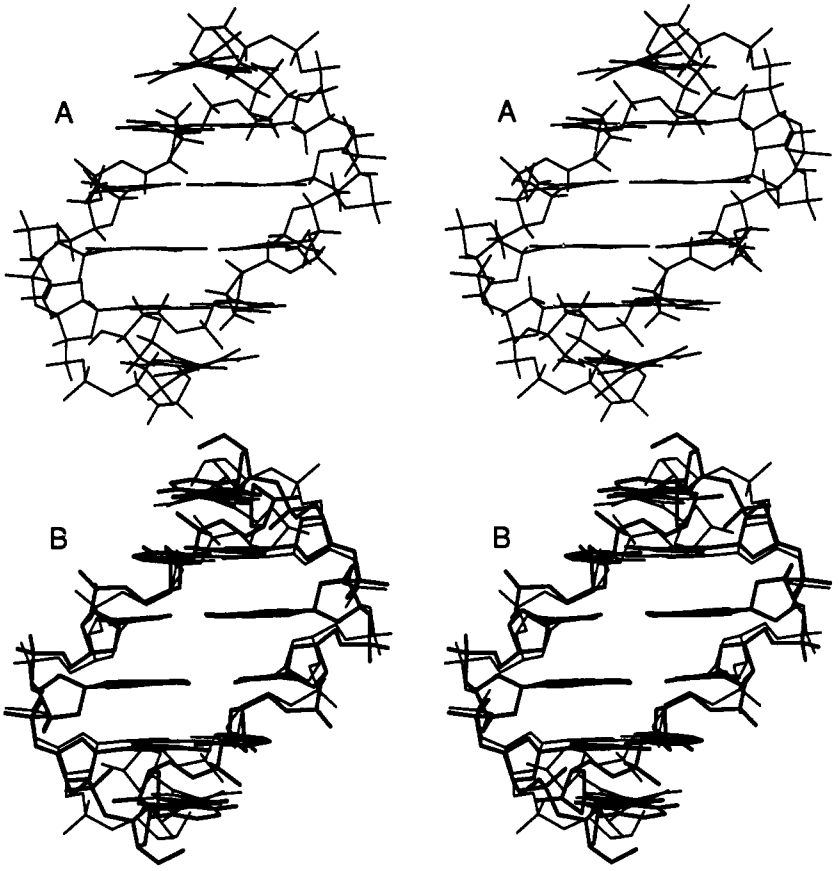

FIGURE 4: (A) Stereodrawing of the weak NOE-refined d(CGATCG) duplex (denoted B_{RF140}) with B_{FEM} as starting model. (B) Least-squares superposition of B_{FEM} and B_{RF140} (RMSD = 0.87 Å).

the two models. The agreement between the experimental NOE data and the simulated NOE data derived from the refined model can be inspected by their corresponding NOE spectra shown in Figure 5. It is obvious that after refinement, the simulated NOE pattern is in excellent agreement with the observed data. However, there are some notable observed NOE cross peaks which cannot be accounted for in the simulated NOE spectrum. These extra NOE peaks can only be accounted for by an end-to-end stacking of the DNA duplexes.

Figure 4B compares the starting B_{FEM} and B_{RF140} models by superimposing the two models using a least-squares fitting. The major changes in the structure appear to be the adjustment of the sugar puckers from the initially uniform C2′-endo (P = 180°) to a somewhat more variable range that results from the applied NOE constraints (P = 180° for A3 to P = 120° for C5) (Figure 6).

Determination of the Rotational Correlation Time. At the onset of the analysis, we assumed that the DNA hexamer d(CGATCG) would adopt a conformation closely related to canonical B-DNA. We first refined this model with a range of values for the isotropic rotational correlation time (τ_c) to empirically determine an optimal value. Figure 7A shows that the resulting R-factor was minimized with a τ_c of 9.5 ns for the duplex at 6.7 mM. Figure 7C shows that, for a lower hexamer duplex concentration of 2.4 mM, the R-factor was minimized at a τ_c of 4.0 ns. We can also judge the optimization of τ_c by examining the target distances from the two cytosine ring H5 to H6 NOE's. This distance should be close to an idealized distance of 2.458 Å. In Figure 7B this distance is matched for cytosine C1 at 9.25 ns and for cytosine C5 at 7.50 ns for the 6.7 mM solution. At a lower concentration of 2.4 mM duplex, this distance is matched for both cytosines around 4.0 ns (Figure 7D). This is consistent with the observation that, at higher concentration, longer end-to-end aggregation is occurring, which would cause the deduced τ_c to be longer.

Refinement Starting with the A-DNA Model. To test the dependency of the refinement procedure on different starting models, we initiated the refinement with an A-DNA conformation. The A-DNA starting model has much larger deviations between its simulated 2D NOE spectrum and the observed 2D NOE spectrum than the starting B-DNA model (Figure 8A). When starting with the A-DNA model, both the local and gross conformations are very different from those indicated by the observed data. The sugars in A-DNA all start in the C3′-endo conformation, whereas the observed data strongly indicates C2′-endo conformations (Figure 5A). In addition, the gross helix shape of the A-DNA model is very different from that of a B-DNA model (Figure 3). In A-DNA, the center of the base pairs are −4.5 Å from the helix axis (x-displacement), and the bases are inclined 20° to the plane normal to the helix axis. In contrast, in B-DNA the base pairs are centered near the helix axis with very small inclination angles (~3°).

So, will starting with either a B-DNA or A-DNA both converge to structures with the same conformation? Figure 6 shows that, during the first 20 cycles, the sugar pseudorotation angle for the two refinements have converged, with the exception of the terminal base, G6. This is probably not surprising as NOE data encode interproton distances with considerable redundancy for the local structure. Concomitant with these changes in sugar conformation, the R-factor (Figure 3) also shows dramatic changes during the first 20 refinement cycles. However, while the gross conformation is subsequently changing from A- to B-DNA-like, (as judged by the x-displacement values, see Figure 9), there is no significant change

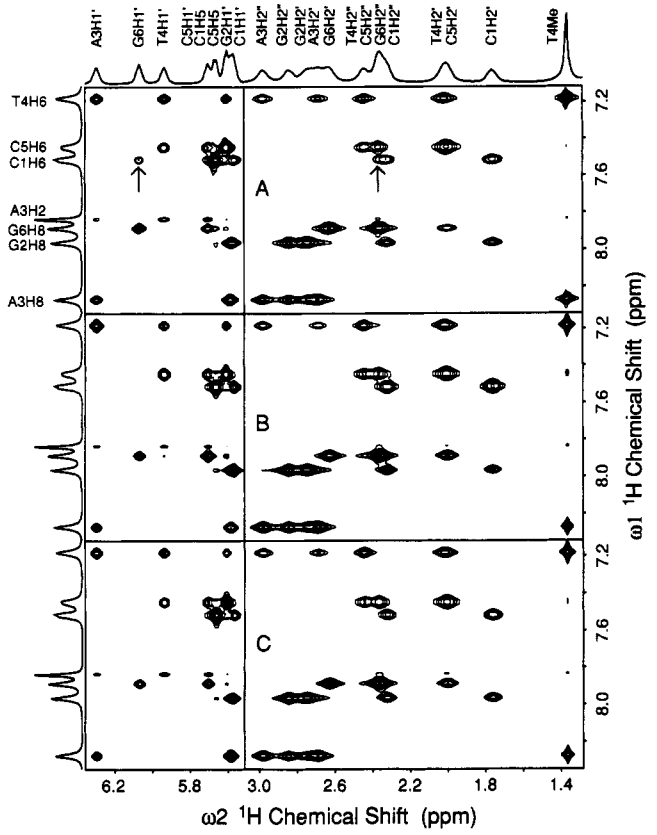

FIGURE 5: Comparison of the experimental and simulated NOE spectra as illustrated by the expanded sections of the 2D NOE spectrum. (A) Experimental NOE spectra. Note that there are interesting cross peaks (at the arrows) between C1–H6 and G6–H1′ and G6–H2″ which are due to the transient stacking of two hexamer helixes, bringing the terminal base pairs close to each other. (B) Simulated 2D NOE spectra based on the starting B_{FEM} model. (C) Simulated 2D NOE spectra based on the refined B_{RF140} model (*R*-factor = 15.2%).

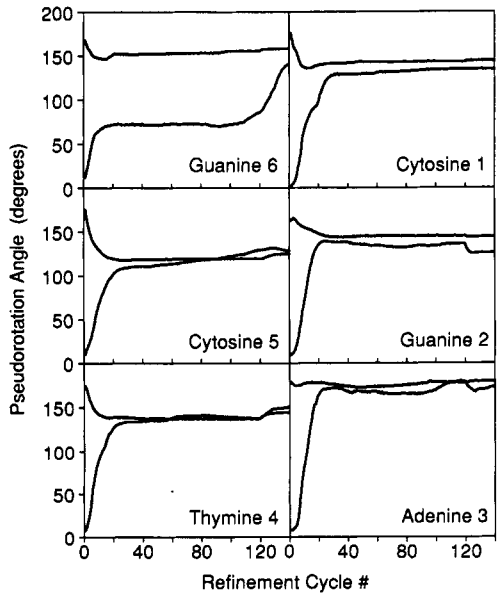

FIGURE 6: Changes of sugar pseudorotation angles of the six nucleotide residues in the d(CGATCG) hexamer duplex during the refinement cycles using medium NOE strength (scalar = 20). The curves starting at about 175° are for B-DNA, and curves starting at about 10° are for the A-DNA. The pseudorotation angle was calculated by the method of Altona and Sundaralingam (1972).

in the *R*-factor. This may be because the *R*-factor is dominated by the comparatively larger NOE's of the shorter range

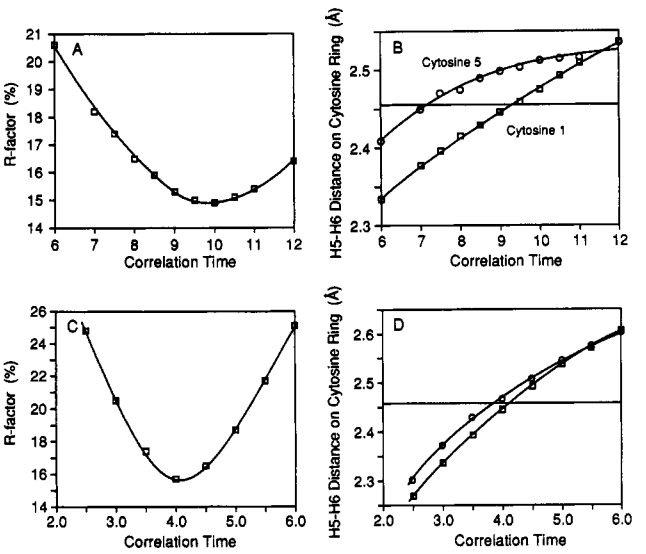

FIGURE 7: Determination of optimum rotation correlation time τ_c for the d(CGATCG) hexamer duplex (at 6.7 mM, panels A and B, and 2.4 mM, panels C and D). The starting B_{FEM} model was refined 40 cycles with weak NOE strengths. In panels A and C, the resultant *R*-factor is shown, and in panels B and D the target distance for the H5 and H6 vectors of cytosines 1 and 5 is shown. The horizontal line is at the idealized distance of 2.458 Å.

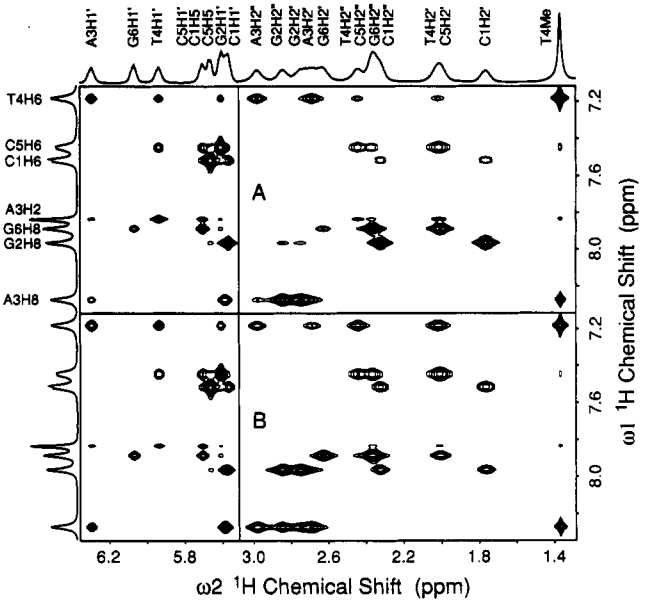

FIGURE 8: A-DNA model-based simulated 2D NOE spectra of same regions shown in Figure 5. (A) Simulated 2D NOE spectra based on the starting A_{FEM} model (*R*-factor = 48%). (B) Simulated 2D NOE spectra based on the A_{RF140} model (*R*-factor = 15.6%).

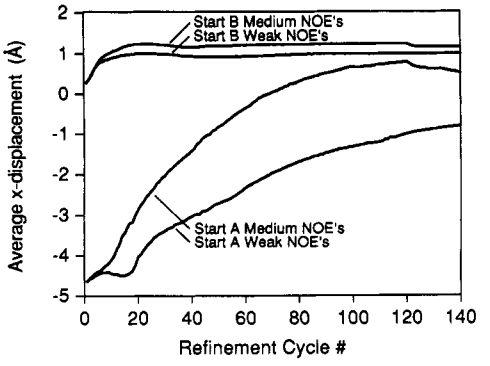

FIGURE 9: Average *x*-displacement [as defined by Dickerson et al. (1989)] of the refinements of the d(CGATCG) hexamer duplexes.

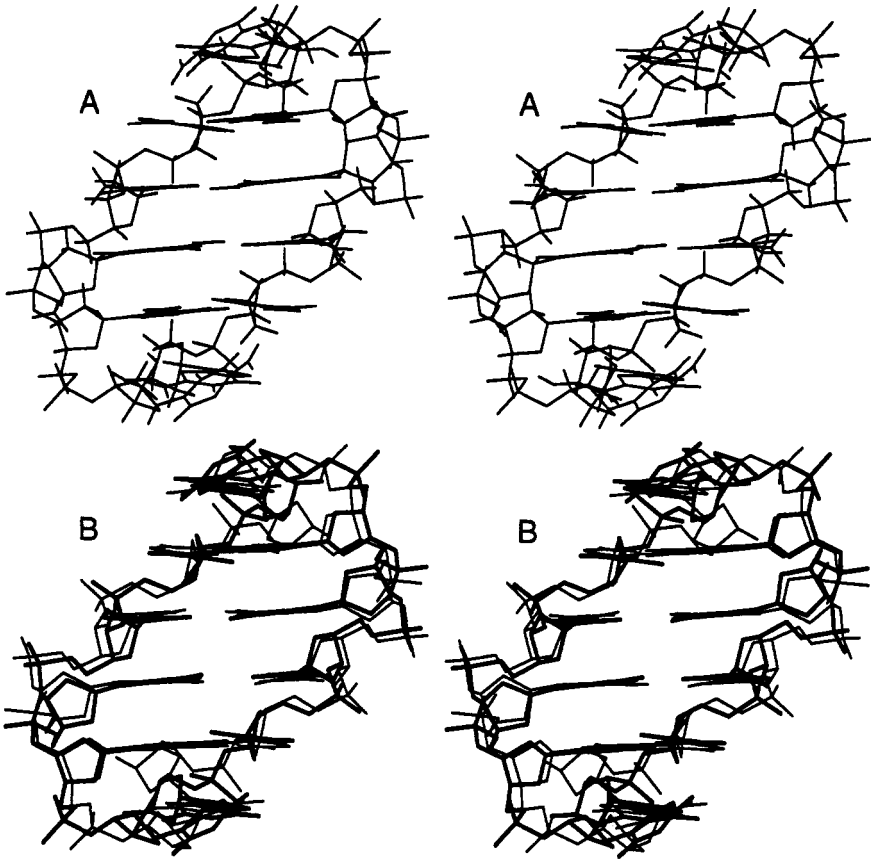

FIGURE 10: (A) Stereodrawing of the NOE-refined d(CGATCG) duplex (denoted A_{RF140}) with A-DNA as starting model and medium NOE's. (B) Least-squares superposition of B_{RF140} with medium NOE's (thick lines) and A_{RF140} refined with medium NOE's (RMSD = 1.57 Å) (thin lines).

spin pairs including those NOE's that predominantly define the sugar conformation.

In the refinement starting with the A-DNA model using the medium constraints, the overall shape of the refining model changes to a B-DNA shape within about 100 cycles. With the same starting model and weak NOE strengths, the gross conformation does not start to change until after 20 refinement cycles and does not converge even after 140 cycles (Figure 9). Making the NOE scalar any weaker would eliminate the change in the gross conformation toward B-like; however, the sugars would continue to converge (data not shown). Figure 10 shows the refined structure when starting with A-DNA and with medium the NOE scalar. This is A_{RF140}, which has an RMSD with B_{RF140} of 1.57 Å. It is clear that the refinement procedure has successfully converted the gross structure on the basis of NOE data only.

Residual Errors. The overall *R*-factor for the refinement using weak NOE strength with the B-DNA model is 15.2%. Figure 11 shows the *R*-factor for each individual spin. In general, the distribution of errors is quite well distributed with the exception of part of cytosine C1 sugar protons, in particular for the H5′ and H5″ protons. Here nearly all of the simulated NOE's are larger than the experimental. The T_1's for these protons are considerably shorter (1 s for H5′ and H5″) than those of the other sugar protons (~1.9 s). Clearly these errors associated with the C1 H5′ and H5″ protons (R ~60%) suggest that the theory within the refinement is unable to produce a structure consistent with the observed NOE's. The observed T_1's suggest that the simplified model for internal motion of the molecule will misrepresent the relaxation rates for the C1 H5′ and H5″ protons relative to the rest of the protons on the molecule. In addition, the end-to-end stacking of the duplexes observed at high concentration (6.7 mM)

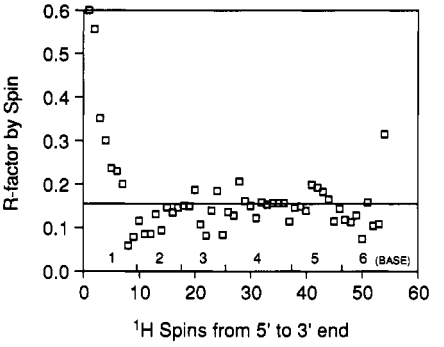

FIGURE 11: *R*-factor for each individual spin in the d(CGATCG) molecule, which has a total of 54 nonexchangeable spins. Within each base, the order is H5′, H5″, H4′, H3′, H2′, H2″, H1′, and then H5, H6 (cytosine) or H8, H2 (adenine, guanine) or Me H6 (thymine).

suggests that the molecule may tumble more like a cylinder than a sphere. Consequently, the use of an isotropic correlation time τ_c may limit the interpretation of the measured NOE's. This may be related to the nonequivalent τ_c values deduced on the basis of the fixed H5′–H6 distance for the two cytosine residues at 6.7 mM DNA concentration. Upon dilution of the solution to 2.4 mM, the difference between the τ_c values for the two cytosines diminishes from 23% to 10% (Figure 7B,D). The inclusion of an anisotropic correlation tensor and a periodic boundary condition may account for some of the errors, for the majority of spins. However, the disparity in T_1's will still require a more sophisticated treatment to account the behavior of the terminal sugars.

Discussion

We have developed a simplified procedure to refine DNA structures in solution using 2D NOE data in a quantitative

manner. This method incorporates a cyclic process where the assessment of the correctness of the structure is based on the original experimental 2D NOE spectrum and the evaluation of the target distances derived thereof. Our results show that this procedure allows a well-behaved refinement process to proceed even when starting with grossly disparate models. The agreement between the experimental and the simulated data appeared to be very respectable, with an *R*-factor of 15% at the current stage of refinement.

An important issue regards the treatment of the weighting scheme with respect to different types of NOE cross peaks in the refinement. As the NOE strength is inversely proportional to the $(r_{ij})^6$, there is a very large dynamic range for the NOE intensities within a molecule. For example, the NOE's between the H2′ and H2″ geminal protons are very strong because of the fixed close distance (1.79 Å) between the two protons, and they would dominate the *R*-factor calculation. However, those NOE's do not really provide useful structural information. In contrast, weak NOE's between protons that are far apart (~6–7 Å), although less reliable, are useful in defining the tertiary structure. Should they be given more weight in the contribution of NOE constraint in the refinement step? This question is currently being explored.

The inverse sixth power dependence on the distance of eq 1 expresses the idea that the spin interaction is between two randomly oriented nuclear spin dipoles. The rate also depends upon a function of the correlation time, $F(\tau_c)$, because the relaxation processes rely on the presence of a background of oscillators provided by the molecular tumbling in the magnetic field. We assume that the molecule is isotropically tumbling such that this molecular diffusion can be expressed by a single number, the correlation time, τ_c. The precise meaning of r_{ij} is in need of clarification here, since the molecules of interest are likely to experience considerable internal motion. It turns out that if the internal motions of the two nuclei are relatively fast, uncorrelated, and isotropic, then r_{ij} means the time-average center-to-center distance (LeMasters et al., 1988). In addition, the r_{ij} for a proton to a methyl group can be accurately considered as the distance from the geometric center of the three protons of methyl group to the other proton. This is the context for the internal motion in which the static appearing molecular models of Figures 4 and 10 should be viewed. The point is that the measured NOE cross-correlations represent average distances, but there is little or no information about rapid fluctuations due to internal motion of the molecule. The next step is trying to decide when and why these simplified assumptions are no longer valid and must be supplemented by additional theoritical considerations. When the internal motion becomes slow, the situation becomes more complex (Landy et al., 1989; Koning et al., 1990); when the shape of the molecule becomes very oblate, more complex molecular tumbling models may need to be considered (Duben et al., 1990; Withka et al., 1990). For our purposes here, we have used simplifying assumptions about motion and shape and have considered anomalies of the analysis to point out where more complex relations are needed. For instance, the analysis of both the T_1's and residual *R*-factors (by spin) for the terminal bases (C1 and G6) suggest that a more sophisticated treatment of these bases is needed. A first step would be to incorporate the disparity in the observed T_1's into the calculation of the relaxation matrix, as these relaxation pathways can be fairly accurately measured and the relaxation matrix should be consistent with these observations.

The refinement algorithm incorporates the conjugate gradient minimizer in XPLOR (Brünger, 1990). This procedure proves to be sufficiently powerful, as it was able to force the sugar pucker from the C3′-endo conformation to the C2′-endo conformation in our refinement of d(CGATCG) when starting with a divergent A-DNA model. However, it is possible that the molecular dynamics simulated annealing procedure, in conjunction with NOE constraints, can overcome large local energy barriers to attain a more accurate model. In our experience, molecular dynamics simulated annealing produced refinement results that were qualitatively similar to the conjugate-gradient minimization procedure employed here.

One method for evaluating the dependence of the refined structure on the experimental observations is by comparison of refinements that started with different conformations. The convergence of conformation among these refinements strengthens the assertion that the observed NOE's have determined the resulting structures. In Figure 6, the convergence of the pseudorotation angles of each bases' sugar indicates these conformations are dictated by the observed NOE data.

The refinement trajectory of guanine G6 suggests that this bases' sugar is heading toward convergence; however, the information directing this transformation is much weaker than that for the other bases' sugars. This is reasonable since important NOE's from the aromatic protons of the next 3′ base to the H1′, H2′, and H2″ are unavailable. We suggest that probably there is sufficient information regarding the sugar conformation of guanine G6; unfortunately it is available only from one side, rather than from a variety of directions. This limited directionality together with the reduced number of constraints reduces the accuracy in defining its structural features from the NOE observations.

There are several other aspects in the refinement procedure that limit interpretation of the NOE data. The results in Figure 5 and 7 suggest that an important limitation on the refinement is due to using an isotropic correlation time rather than an anisotropic correlation tensor that would be related to the molecular moments of inertia. The source of this anisotropic tumbling is due to the transient formation of duplex multimer tubes in solution at this concentration of DNA at 10 °C. The H5–H6 vector of cytosine C5 probably lies nearly in the plane normal to the helix axis. This means that it may experience considerably faster rotation at a correlation time of about 7 ns. This faster rotation effectively reduces the rate of NOE coupling due to the decreased spectral density. On the other hand, the H5–H6 vector of cytosine C1 may be aligned somewhat away from this plane and thus be more representative of the average rotational correlation time where we would expect the isotropic correlation time refinement to minimize the *R*-factor (9.5 ns). Those NOE's that are aligned with the helix axis will experience a higher density of oscillations due to the slower tumbling, thus larger NOE relaxation rates. The surprisingly long isotropic correlation time of 9.5 ns is consistent with the duplexes forming into tubes, and the degree of anisotropy in the correlation time can be interpreted in term of the time-averaged extent of multimerization.

The result in Figure 11 that the residual errors by spin are all very well distributed suggests that the limitations imposed by the use of an isotropic correlation time are spread uniformly among all the spins. However, the cytosine C1's H5′ and H5″ are clearly exceptional cases with errors greater than 50%. These two spins also have T_1 relaxation times (see Figure 1S) that are considerably shorter than that of the other spins (1.0 vs 1.9 s). The observed NOE's are considerably smaller than any reasonable model can account for. This group is rather unconstrained conformationally since it is at the terminal 5′ end of the DNA molecule and can easily rotate about the

C4′–C5′ bond. We suggest that this group may adopt different conformations at a high rate, and thus its NOE's are greatly diminished with the T_1 relaxation accelerated due to autorelaxation processes.

Despite these limitations, under optimum conditions NMR spectroscopy can provide another powerful tool for elucidating 3D molecular structure in addition to the well-established X-ray diffraction method. In fact, they are highly complementary to each other. We believe the combination of the two methods would provide an extremely powerful approach to the structural and dynamic aspects of many biological problems, including the interactions of antitumor drugs with nucleic acids or higher order structures (such as hairpins).

Acknowledgments

We are indebted to Professors Jacques H. van Boom and Gijs A. van der Marel for their continuous support. We gratefully acknowledge the contributions to this work in its early stages by Dr. Y.-C. Liaw. Technical help from Drs. M. Leopold, Z. Gan, and V. Mainz of the School of Chemical Sciences Molecular Spectroscopy Lab is gratefully acknowledged.

Supplementary Material Available

Four tables listing chemical shifts, force constants for NOE springs, torsion angles, and coupling constants and two figures showing T_1's by spin and the 2D NOE spectrum (10 pages). Ordering information is given on any current masthead page.

References

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

Arnott, S., Chandrasekaran, R., Hall, I. H., Puigjaner, L. C., Walker, J. K., & Wang, M. (1982) *Cold Spring Harbor Symp. Quant. Biol. 47*, 53–64.

Boelens, R., Koning, T. M. G., van der Marel, G. A., van Boom, J. H., & Kaptein, R. (1989) *J. Magn. Reson. 82*, 290–308.

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

Brünger, A. T. (1990) XPLOR, version 2.1, The Howard Hughes Medical Institute and Yale University; New Haven, CT.

Clore, C. M., & Gronenborn, A. M. (1989) *Crit. Rev. Biochem. Mol. Biol. 24*, 479–564.

Dickerson, R. E., et al. (1989) *EMBO J. 8*, 1–4.

Duben, A. J., & Hutton, W. C. (1990) *J. Am. Chem. Soc. 112*, 5917–5924.

Gonzalez, C., Rullmann, J. A. C., Bonvin, A. M. J. J., Bolenz, R., & Kaptein, R. (1991) *J. Magn. Reson. 91*, 659–664.

Hare, D., Wemmer, D. E., Chou, S. H., Drobny, G., & Reid, B. R. (1983) *J. Mol. Biol. 171*, 319–336.

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

Koning, T. M. G., Boelens, R., & Kaptein, R. (1990) *J. Magn. Reson. 90*, 111–123.

Landy, S. B., & Rao, B. D. N. (1989) *J. Magn. Reson. 81*, 371–377.

LeMaster, D. M., Kay, L. E., Brünger, A. T., & Prestegard, J. H. (1988) *FEBS Lett. 236*, 71–76.

Lown, J. W., Hanstock, C. C., Bleackley, R. C., Imbach, J.-L., Rayner, B., & Vasseur, J. J. (1984) *Nucleic Acids Res. 12*, 2519–2533.

Macura, S., & Ernst, R. R. (1980) *Mol. Phys. 41*, 95–117.

Metzler, W. J., Wang, C., Kitchen, D. B., Levy, R. M., & Pardi, A. (1990) *J. Mol. Biol. 214*, 711–736.

Nilges, M., Habazettl, J., Brünger, A. T., & Holak, T. A. (1991) *J. Mol. Biol. 219*, 499–510.

Palecek, E. (1991) *Crit. Rev. Biochem. Mol. Biol. 26*, 2861–2875.

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

Rich, A., Nordheim, A., & Wang, A. H.-J. (1984) *Annu. Rev. Biochem. 53*, 791–846.

Solomon, I. (1955) *Phys. Rev. 99*, 559–565.

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

van der Marel, G. A., van Boeckel, C. A. A., Willie, G., & van Boom, J. H. (1981) *Tetrahedron Lett. 22*, 3887–3888.

Withka, J. M., Swaminathan, S., & Bolton, P. H. (1990) *J. Magn. Reson. 89*, 386–390.

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