Three-Dimensional Structure of a DNA Hairpin in Solution: Two-Dimensional NMR Studies and Distance Geometry Calculations on d(CGCGTTTTCGCG)[†]

Dennis R. Hare
Infinity Systems, Seattle, Washington 98117
Brian R. Reid*
Departments of Chemistry and Biochemistry, University of Washington, Seattle, Washington 98195

Received December 8, 1985; Revised Manuscript Received April 25, 1986

ABSTRACT: The three-dimensional structure of d(CGCGTTTTCGCG) in solution has been determined from proton NMR data by using distance geometry methods. The rate of dipolar cross-relaxation between protons close together in space is used to calculate distances between proton pairs within 5 Å of each other; these distances are used as input to a distance geometry algorithm that embeds this distance matrix in three-dimensional space. The resulting refined structures that best agree with the input distances are all very similar to each other and show that the DNA sequence forms a hairpin in solution; the bases of the loop region are stacked, and the stem region forms a right-handed helix. The advantages and limitations of the technique, as well as the computer requirements of the algorithm, are discussed.

In the last few years, high-resolution NMR, especially two-dimensional NMR spectroscopy, has developed into a powerful tool for analyzing the solution structure of biological macromolecules [for a recent review, see Wemmer & Reid (1985)]. The two-dimensional nuclear Overhauser effect spectroscopy (NOESY)[1] experiment is capable of yielding distance information for all protons pairs up to 4 or 5 Å apart in space while COSY experiments contain scalar couplings, which can be interpreted as dihedral angles. It is apparent that a great deal of structural information is present in NMR data but this information is largely in the form of interproton distances, which are not readily converted into three-dimensional structure. This distance information has previously been used qualitatively to deduce some aspects of biopolymer structure such as the A or B conformation of DNA (Reid et al., 1983); this approach has been especially useful in deciphering local secondary structure elements in proteins where extended β-structures and α-helices can readily be distinguished by nearest-neighbor NOE patterns (Wuthrich et al., 1982; Billeter et al., 1982; Wagner & Wuthrich, 1982).

The mathematics of distance geometry was developed in the 1930's (Menger, 1931); 40 years later, Crippen and Havel pioneered the application of distance geometry to biological macromolecules (Crippen, 1977; Havel et al., 1979). The determination of local structure in proteins from NOE distances by distance geometry methods has been reported for segments of the peptide hormone glucagon bound to lipid micelles (Braun et al., 1981) and for the nine-residue C-terminal domain of mellitin (Brown et al., 1982). Very recently, Havel and Wuthrich demonstrated the feasibility of determining the entire structure of a small protein by distance geometry using short proton–proton distances taken from the crystal coordinates of BPTI (Havel & Wuthrich, 1985), and Williamson et al., have calculated the entire structure of the 57-residue bull seminal protease inhibitor BUSI-IIA from NOE data using distance geometry methods (Williamson et al., 1985). The determination of DNA solution structure by distance geometry methods has not yet been attempted, although Clore and Gronenborn have used NOE distances to refine the X-ray fiber coordinates of DNA (Clore & Gronenborn, 1985). This approach is obviously incapable of defining ab initio the structure of new conformations for which X-ray data are unavailable.

We became interested in distance geometry as a potential method for determining the structure of several small synthetic DNA sequences we have studied by 2D NMR. For several such sequences, the quality of the NOESY data was quite sufficient to calculate reasonably accurate interproton distances. The sequence d(CGCGTTTTCGCG) was chosen as our initial test of distance geometry because it has a well-resolved proton spectrum. Furthermore, the presence of the central mismatched TTTT sequence made it likely that this sequence would assume an atypical bulged duplex or unimolecular hairpin conformation in solution. Such structures have not been analyzed by X-ray crystallographic methods, and it was our goal to establish whether the structure could be determined directly from distance constraints by distance geometry methods rather than by using molecular dynamics or thermodynamic arguments. Previous results from this laboratory have shown that the sequence d(CGCGTATACGCG)$_2$ exchanges between duplex and hairpin conformations on a 1-s time scale (Wemmer et al., 1985). The sequence d(CGCGTTTTCGCG) does not exhibit such behavior, and its spectrum is not complicated by problems of chemical exchange. The thermodynamics of T$_n$ loops have been reported previously (Haasnoot et al., 1983).

Materials and Methods

The DNA dodecamer d(CGCGTTTTCGCG) was synthesized by using solid-phase phosphite triester techniques as described previously (Hare et al., 1983; Chou et al., 1984) and was dissolved in 0.4 mL of a buffer containing 10 mM sodium phosphate, pH 7.0. The sample was lyophilized to dryness, and 0.4 mL of 99.996% D$_2$O was added. After one more

[†] Supported in part by National Institutes of Health Program Project Grant GM 32681 to B.R.R. and National Institutes of Health SBIR Grant GM 35620 to D.R.H.

* Address correspondence to this author at the Department of Chemistry, University of Washington.

[1] Abbreviations: 2D NMR, two-dimensional NMR; COSY, two-dimensional correlated NMR spectroscopy; NOE, nuclear Overhauser effect; NOESY, two-dimensional nuclear Overhauser effect spectroscopy; DG, distance geometry; BPTI, bovine pancreatic trypsin inhibitor.

0006-2960/86/0425-5341$01.50/0

lyophilization step, another 0.4 mL of 99.996% D_2O was added and the solution transferred to a 5-mm Wilmad NMR tube.

NMR spectra were acquired on a Bruker WM-500 spectrometer. The sample temperature was regulated at 298 K, and all experiments were performed within a single 5-day period without removing the sample from the spectrometer or changing any frequencies or gain settings. An absolute-magnitude COSY was acquired into 2048 points in t_2 and a total of 366 points in t_1. Four pure absorption, NOESY spectra were collected by using phase-sensitive methods (States et al., 1982) with mixing times of 0.1, 0.2, 0.3, and 0.5 s. The NOESY spectra were acquired into 2048 points in t_2 and 700 points in t_1. For each t_1 value, one real spectrum and one imaginary spectrum were collected to accomplish quadrature in ω_1.

Exchangeable proton spectra were obtained by using a Redfield 214 pulse of about 270-μs length, with the carrier frequency placed at 12 ppm (Redfield, 1978). NOE difference spectra were acquired as described previously (Hare & Reid, 1982; Chou et al., 1984).

After collection, the data were written onto magnetic tape and transferred to a DEC microVAX II for processing by our own software (D. R. Hare, unpublished results). The COSY data were apodized by a skewed sinebell of 600 complex points in length and then Fourier transformed in the t_2 dimension. Because Bruker spectrometers digitize real and imaginary points at different times, it was necessary to Fourier transform all t_2 data using an algorithm developed by Redfield and Kunz (Redfield & Kunz, 1975). The t_1 data were apodized by a skewed sinebell of length 366 points, zero filled to 1024 complex points, Fourier transformed, and converted to absolute magnitude. The NOESY spectra were not apodized in t_2 before Fourier transformation since apodization yields incorrect relative resonance intensities. After Fourier transformation, the data were phase corrected to yield pure absorption line shapes. The phase-corrected real and imaginary spectra for each t_1 were then combined to yield a total of 350 complex points in t_1. The t_1 data were apodized by using a right-shifted sinebell-squared window with a value of 1 for the first 280 points; beyond this, the window dropped smoothly to zero to avoid truncation effects. The data were zero filled to 1024 complex points and Fourier transformed, and a small phase correction was applied. Copies of each transformed matrix were made, and one copy was subjected to diagonal low-point symmetrization. Symmetrized data are less noisy but may contain artifacts; therefore, comparison with unsymmetrized data is important to determine whether any given cross-peak is real. One-dimensional data were Fourier transformed, phase corrected, and base-line corrected with a cubic spline function using a microVAX computer. All data were plotted and annotated on a Hewlett-Packard 7550 digital plotter interfaced to the microVAX.

Results

The one-dimensional spectrum of d(CGCGTTTTCGCG) in D_2O at 298 K is shown in Figure 1, along with the proton types found in each spectral region. The four cytosine C6H doublets and the four thymine C6H singlets (vide infra) are well resolved in the 7.4–7.8 ppm region, and the thymine methyl groups near 1.8 ppm are obvious. By elimination, the remaining four protons near 8 ppm must be guanine H8 resonances. Scalar-coupled proton pairs were identified in the COSY spectrum (Figure 2). All pyrimidine base protons can be identified in the COSY spectrum; the cytidine C5H–C6H cross-peaks (Figure 2, lower left) and the thymine four-bond methyl–C6H cross-peaks (Figure 2, lower right) yield the chemical shifts of the aromatic and methyl protons of the four cytidine and the four thymine bases, respectively.

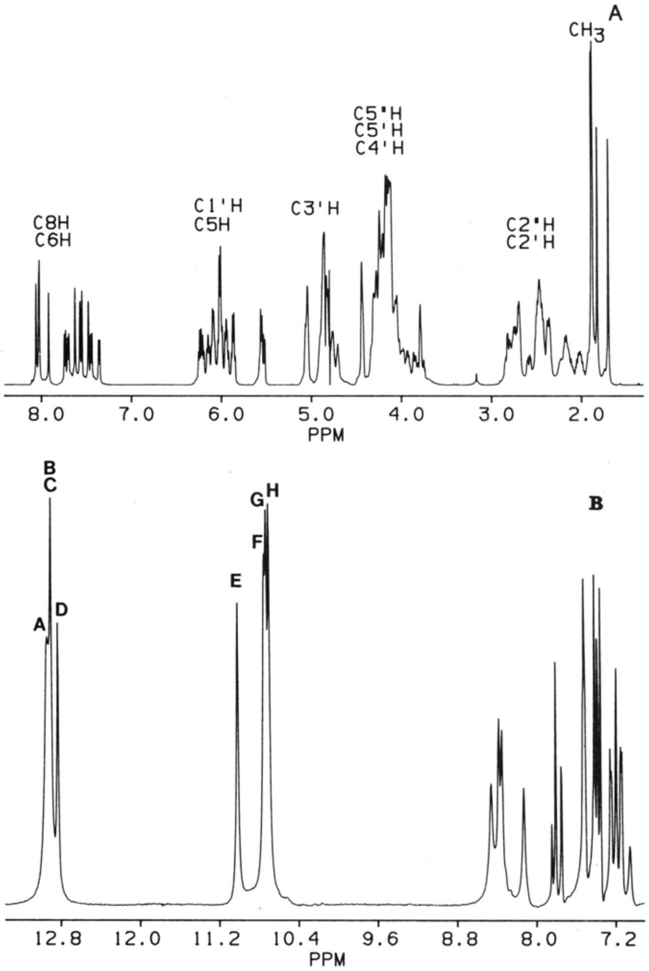

FIGURE 1: (A) Nonexchangeable proton spectrum of d(CGCGTTTTCGCG) in 10 M sodium phosphate/D_2O at 25 °C. (B) Spectrum of d(CGCGTTTTCGCG) in 10 M sodium phosphate/H_2O at 7 °C.

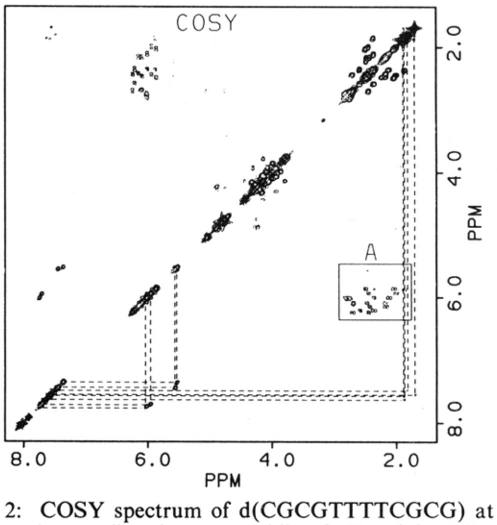

FIGURE 2: COSY spectrum of d(CGCGTTTTCGCG) at 25 °C showing scalar couplings between cytidine C5H and C6H (lower left) and thymine methyl and C6H at lower right. The region labeled A contains deoxyribose C1′H to C2′H and C2″H cross-peaks.

The scalar coupling between the deoxyribose C1′H and the 2′ and 2″ protons on the same sugar yields cross-peaks in region A of Figure 2, shown expanded in Figure 3A. Thus, the chemical shifts of all the 1′, 2′, and 2″ protons can be derived from the COSY, although the 2′H cannot be distin-

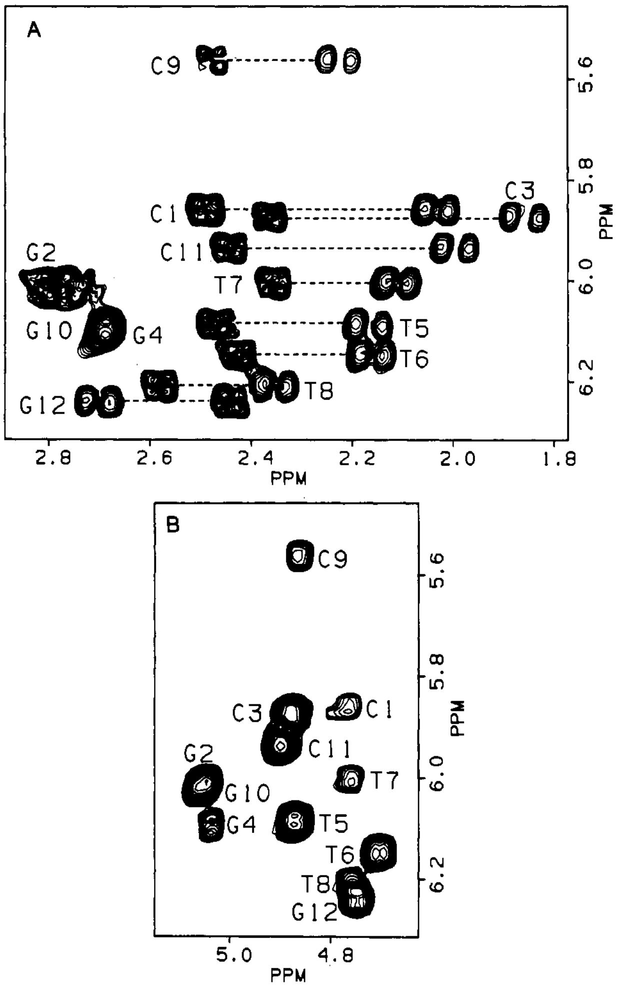

FIGURE 3: (A) Expanded region A of the COSY spectrum in Figure 2 showing cross-peaks between C1′H and C2′ protons. (B) Expanded region of the RELAY spectrum showing cross-peaks resulting from the transfer of coherence from C1′H through the C2′ protons to C3′H. This is a convenient way of connecting the C1′H and C3′H of each sugar.

guished from the 2″H. Due to their extensive coupling, the 2′ and 2″ protons are broad peaks in a crowded spectral region, making the 3′–2′2″ cross-peaks less intense and often ambiguous in terms of assigning 3′ resonances; however, the C3′ protons can easily be assigned by means of relayed coherence transfer from the C1′H through the C2′ protons to the C3′ proton. An expansion of the C1′H to C3′H relay cross-peaks is given in Figure 3B.

At this point, we have identified the aromatic and deoxyribose protons belonging to a single base or sugar but have not yet established the position of each base or sugar in the sequence. To accomplish this, the NOESY data (Figure 4) were used to identify protons that are close to each other in space. In previous 2D NMR studies of double-helical DNA duplexes in solution, the NOEs observed in the NOESY spectrum agreed completely with those expected for a right-handed double helix, leading to a set of assignment rules (Scheek et al., 1983; Hare et al., 1983). Since d(CGCGTTTTCGCG) was not expected to assume a regular duplex conformation in solution, no structural assumptions could be made for assignment purposes except those implicit in the fundamental chemical structure of deoxyribonucleotides. These assumptions were the following: (1) Bond lengths, bond angles, and van der Waals radii do not differ significantly from those observed by X-ray crystallography of small molecules. (2) The bases of DNA are assumed to be planar. (3) Stereochemically, only the D enantiomer of deoxyribose is allowed.

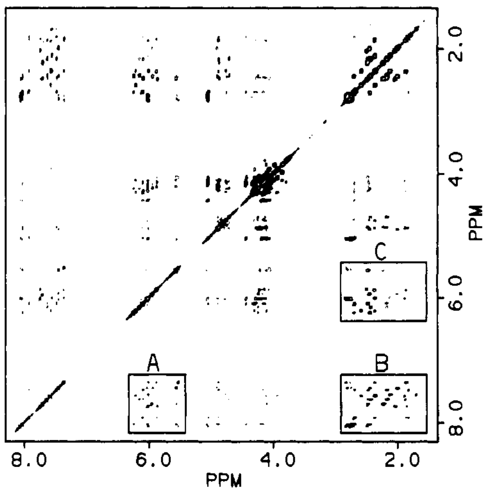

FIGURE 4: Phase-sensitive NOESY spectrum of d-(CGCGTTTTCGCG) at 25 °C collected with a mixing time of 0.4 s. Regions A, B, and C are shown expanded in Figures 5 and 6.

On the basis of these assumptions, one expects to observe an NOE from every purine C8H or pyrimidine C6H to its own C1′H proton since, regardless of the glycosidic dihedral angle, the maximum possible separation between these protons is 3.97 and 3.74 Å, respectively. We observed these intraresidue cross-peaks, and furthermore, most C6H and C8H protons displayed a second NOE to an additional C1′H resonance. The cross-peaks arising from cross-relaxation between the aromatic base protons and deoxyribose C1′ protons are shown in Figure 5. The connectivity pattern within the CGCG segments at each end of the molecule was immediately recognized as being the same as in d(CGCGAATTCGCG)$_2$, i.e., that of right-handed helical DNA (Hare et al., 1983). It was next possible to differentiate between the 5′-C1-G2-C3-G4 end and the C9-G10-C11-G12 3′ end, based on the fact that only one of the G residues and only one of the C residues exhibited cross-relaxation with protons derived from T residues; this also established the identities of T5 and T8 (Figure 5). The identity of T6 and T7 was not immediately obvious from the data (we will refer to them as A and B for now), although several important clues were present: (1) One of the unassigned methyls (thymine B) cross-relaxes both the C6H of thymine A and the C6H of thymine 5. All the other methyls display only one aromatic NOE, to their own C6H (Figure 5). (2) This same methyl group (thymine B) is close to three C1′H protons: one of which is known to be that of T5, one is that of thymine B, and the third is the C1′- of thymine A (Figure 6). (3) The C6H of thymine B cross-relaxes with two C1′H protons, one of which is known to be that of T8 (Figure 5A).

The only assignment possible on the basis of this information is that thymine B is T7 and thymine A is T6. The NOEs from the methyl resonance of TB (T7) to the H6 and 1′H resonances of T5 and TA (T6), combined with the knowledge that all four thymidines are in the anti conformation (relatively weak H6–1′H NOEs), orients the TB (T7) deoxyribose pointing away from T5 and T6; given the constraints of the observed T5 NOEs, the distances are too great for TB to be linked to T5 by a single –CH$_2$–O–P–O backbone linkage. In the case of this particular sequence, these assignments were rather apparent; if they had not been, however, we could have proceeded with distance geometry analysis, leaving the ambiguous residues unbonded to the main body of the polymer, and allowed the distance geometry algorithm to orient them relative to the known residues. Thus, the distance geometry algorithm

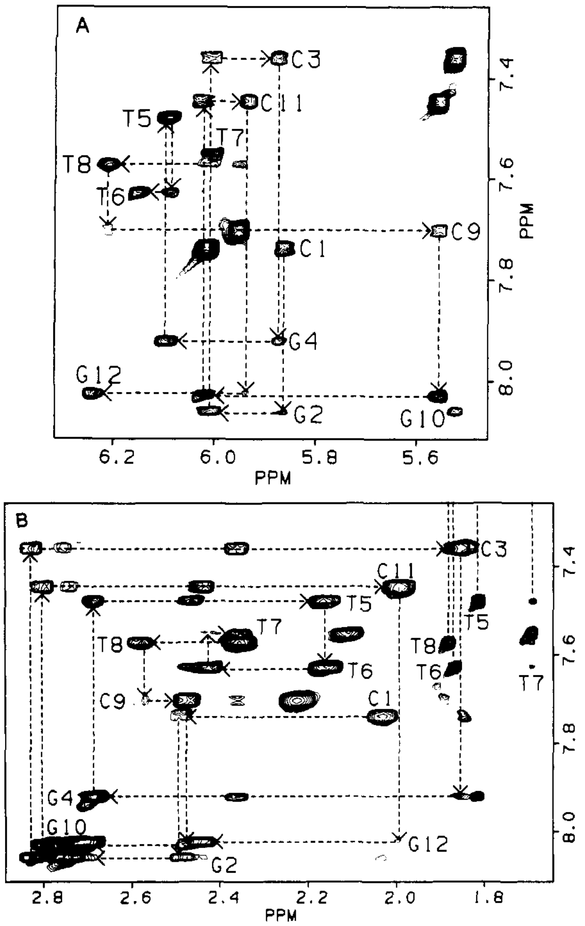

FIGURE 5: (A) Expansion of region A of Figure 4, containing cross-peaks between aromatic protons and the deoxyribose C1′H resonances. (B) Expansion of region B of Figure 4, containing cross-peaks between aromatic protons and deoxyribose C2′ protons as well as methyl protons.

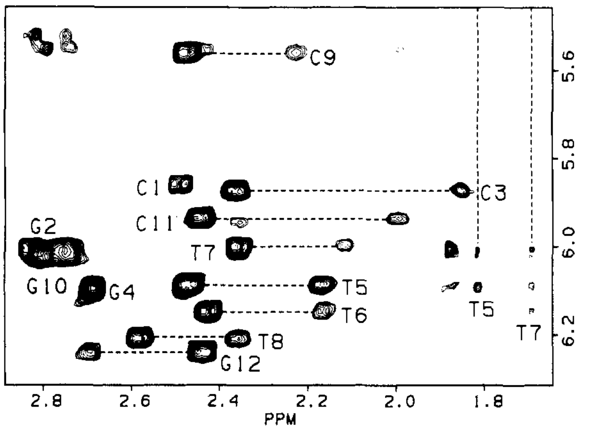

FIGURE 6: Expansion of region C of Figure 4, containing cross-peaks between C1′ protons and the C2′,2″ and methyl protons.

can actually become part of the assignment procedure; the algorithm inevitably detects distance constraints that are inconsistent with standard nucleotide chemistry.

An additional uncertainty in this problem involved the possibility that the sequence might exist in solution either as a hairpin or as a bulged duplex. Due to symmetry, a bulged duplex could appear very similar to a hairpin; careful analysis of the potential loop/bulge resonances was required to resolve the ambiguity. The exchangeable proton spectrum (Figure 1B) shows eight imino proton resonances. Four of these, between 12.5 and 13.0 ppm, are characteristic of GC base pairs and indicate that the CGCG sequence forms stable base pairs. More surprising is the fact that the TTTT imino protons at ca. 10.8 ppm, although not hydrogen bonded, are also narrow, which indicates they are not exchanging rapidly with the water solvent; this leads one to suspect that these imino protons are enclosed in some sort of hydrophobic region (usually a result of base stacking), rather than being exposed to solvent.

The NOE connectivity pattern of these exchangeable protons yields a few additional clues. First, the highest field and the lowest field thymine imino proton resonances (E and H) are both cross-relaxed by a guanine imino proton (almost certainly G4); these two thymine imino protons are also close to each other on the basis of large mutual NOEs. Whether the TTTT sequence forms a bulge structure or a hairpin loop structure, it is safe to assume that these two thymine imino protons are those of T5 and T8, since we have already connected T5 to G4 and T8 to C9 on the basis of NOE connectivities between the nonexchangeable protons on the bases and sugars. Thus, with T5 stacked on G4 and T8 stacked on G9, if the central region were a bulge in a bimolecular duplex, then T6 and T7 would have to loop out without reversing the DNA strand direction. Furthermore, the observed T7 methyl NOEs to the H6 protons of T5 and T6 and to the 1′H resonances of T5 and T6 can only be explained if T7 executes a sharp turn in which it doubles back in order to reside between T5 and T6. However, the NOE pattern between the T7 and T8 residues requires the two bases to be stacked on each other with their deoxyribose moieties oriented similarly. If G4 forms a bimolecular base pair with the C9 of a second molecule rather than pairing with C9 of the same strand, it is not possible, either by computer graphics or by model building, to orient a T5–G4 stack and a T8–C9 stack without violating the observed NOE constraints. This evidence, combined with the additional observation that the T_1 values of the H8 protons in this sequence are about twice as long as those in d(CGCGAATTCGCG)$_2$, convincingly points to a hairpin structure for d(CGCGTTTTCGCG).

On the basis of our assignments (even if some of them are considered to be tentative at this point), we next proceeded to convert the NOE buildup rate for each cross-peak into an approximate distance. Figure 7 shows an example of the time course for magnetization transfer between the assigned aromatic protons of the bases and the sugar 2′ and 2″ protons as well as the thymine methyl resonances. These data give some idea of the precision with which the interproton cross-relaxation rate can be measured. Since the various proton resonances have differing line widths, simple measurement of cross-peak height or integration of a row or column passing through the cross-peak did not give a reliable measure of the cross-relaxation rate. Therefore, we devised a computer method for integrating the volume of cross-peaks numerically and found that distance measurements based on such cross-peak volumes were quite consistent and were entirely reasonable.

In pyrimidines, the thymine CH_3–C6H and the cytosine C5H–C6H cross-peak intensities correspond to known fixed distances, as do those of deoxyribose 2′H–2″H geminal protons. These proton pairs can therefore be used as reference distances, provided there is no differential local motion between residues located in different regions of the molecule. Such differential local motion, if it occurred, would seriously complicate, or even render impossible, the uniform measurement of accurate distances throughout the entire molecule. To guard against this possible artifact, we monitored the rate of the NOE buildup between the C5H and C6H of the four cytosines located in positions 1, 3, 9, and 11, as well as the rate of NOE buildup between the CH_3 and C6H of the four thymines lo-

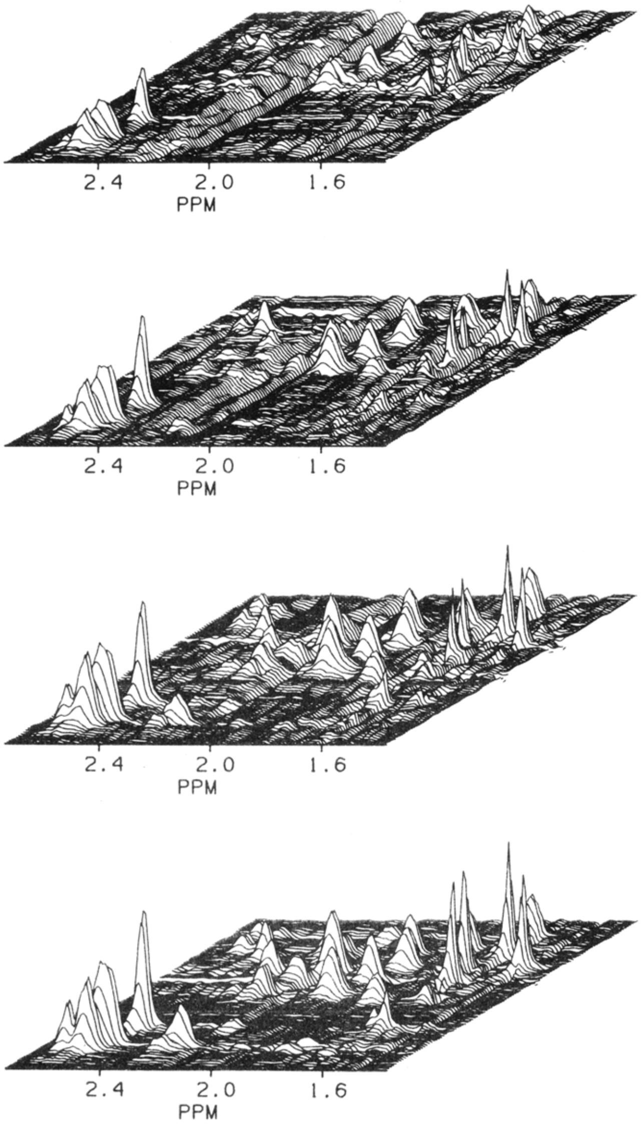

FIGURE 7: Stack plots showing the time course of the NOE buildup between the base protons and the 2′H, 2″H, and methyl resonances. NOESY spectra were collected with mixing times of 0.1, 0.2, 0.3, and 0.5 s (from top to bottom).

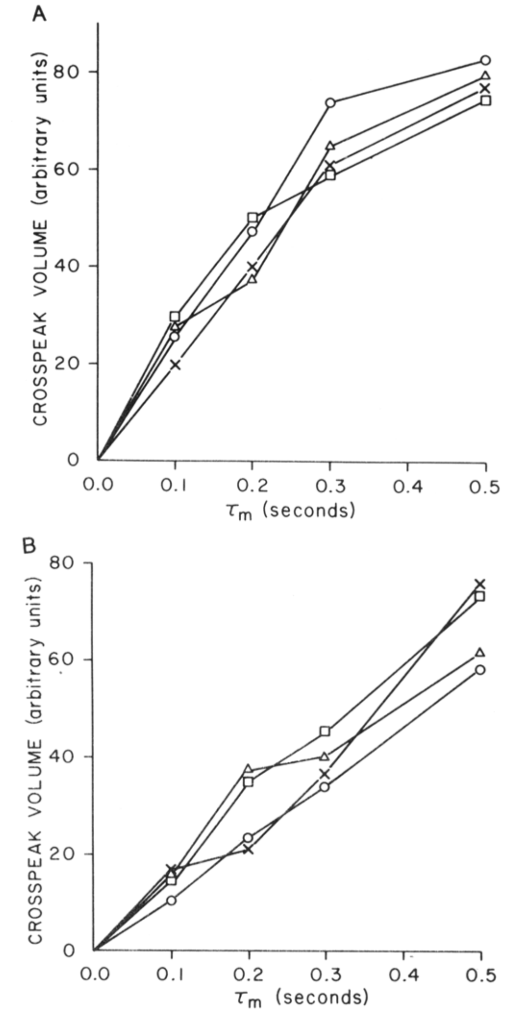

FIGURE 8: Time dependence of the NOE intensity between C6H and C5H in cytosine residues 1, 3, 9, and 11 (A) and between the C6H and methyl protons in thymine residues 5, 6, 7, and 8 (B).

cated in positions 5, 6, 7, and 8. The results are shown in Figure 8. It is apparent that the four cytosines all have the same correlation time and, within experimental error, we cannot detect any differential local motion between the four thymines in the loop. Thus, we feel confident that reasonably accurate distances can indeed be measured from the NOE data.

In generating the distance matrix, we used the cytidine C5H–C6H pair as a reference distance for all aromatic–aromatic and aromatic–ribose proton cross-peaks, the C2′H–C2″H pairs as a reference for all ribose–ribose proton cross-peaks, and the CH_3–C6H pair as a reference distance for all methyl cross-peaks. Measuring cross-peak volumes at each mixing time yielded a time course for each NOE intensity which could then be compared with the appropriate reference to calculate an approximate distance from the known r^{-6} distance dependence of the cross-relaxation process. This process resulted in a set of about 240 distance measurements.

Preparing these approximate distances for input to the distance geometry program involved generating upper and lower bounds for each distance. These bounds were chosen as follows: (1) Initial upper and lower bounds were set to d + 0.2 Å and d − 0.2 Å, respectively, for strong NOEs and d + 0.5 Å to d − 0.5 Å for weak NOEs. (2) These bounds were screened against the maximum and minimum distances possible in the structure of nucleotides; any values outside these bounds limits were corrected. (3) Any distances that were ambiguous due to cross-peak overlap were given only bounds implicit in the primary structure, or implied by other observed NOEs.

The distance geometry algorithm uses known bond lengths, geminal distances (which define bond angles), and chirality properties of nucleotides taken from small-molecule X-ray data. These distances, plus the experimental distance bounds determined from NOEs, are entered into the bounds matrix, after which all unspecified lower bounds between nonbonded atoms are set to the sum of van der Waals radii of the atoms involved. All unspecified upper bounds are set to an arbitrarily large number (300 Å in this case). The bounds matrix is then subjected to smoothing using the triangle inequality for all sets of three atoms (Crippen, 1981; Havel & Wuthrich, 1985) to bring the upper and lower bounds as close together as possible, after which 26 different trial distance matrices were generated in which each distance was randomly chosen between the upper and lower bounds for that distance. Each distance matrix was then embedded in 3-space using the procedure outlined by Crippen (1981) yielding 26 trial structures.

The initial trial structures (three of which are shown in Figure 9) are quite distorted due to the fact that the randomly chosen trial distance matrices contain combinations of distances that cannot occur in 3-space; this is due to the ap-

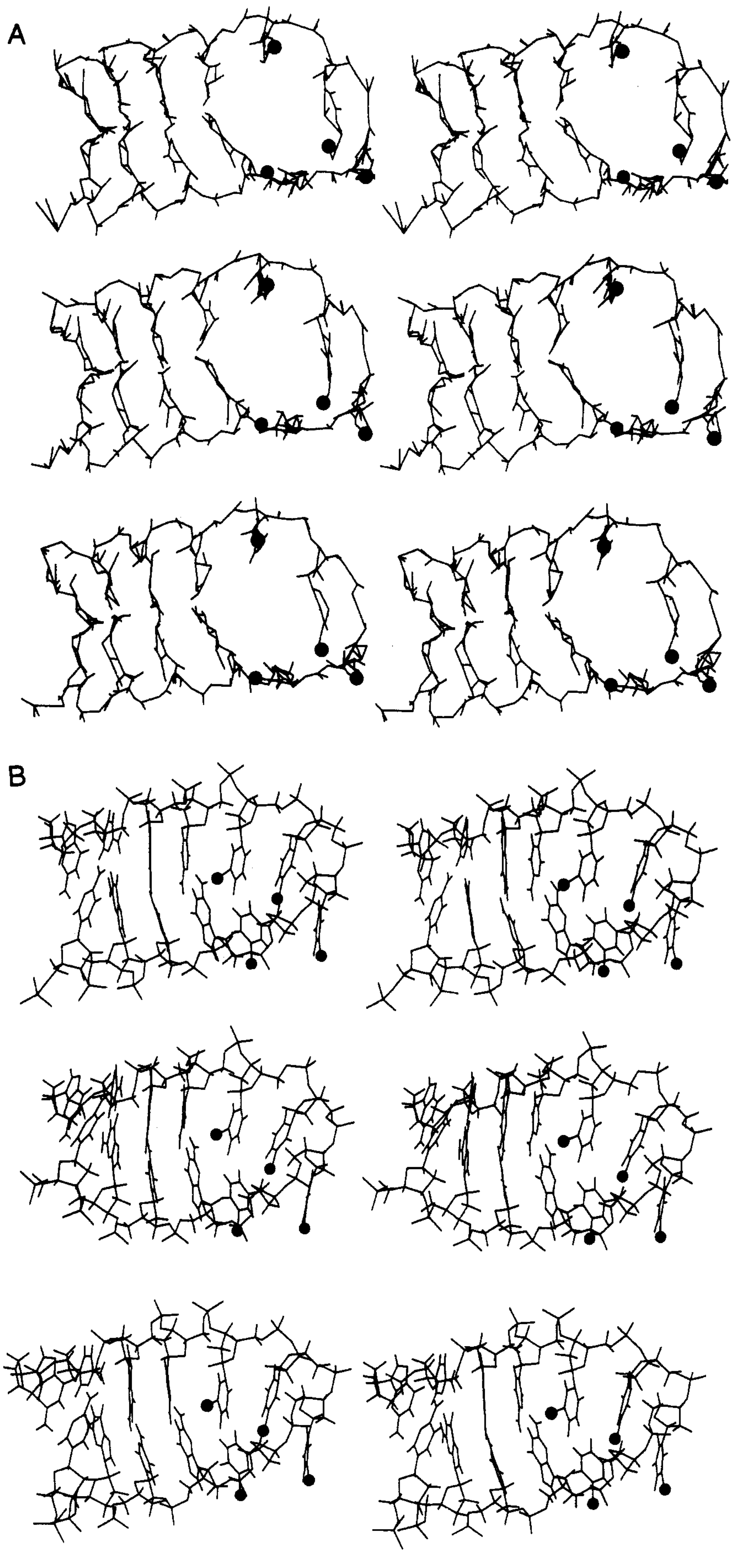

FIGURE 9: (A) Three embedded trial starting structures shown as stereopairs. (B) The three refined structures derived from the trial structures shown in (A). The refined structures are shown as stereopairs with thymine methyl groups represented as large filled spheres.

proximate and incomplete nature of the distance data. In fact, the trial structures violate the bond length, bond angle, and distance bounds, and these structures must be refined to conform to the original distances, including the interresidue NOE distances, in the bounds matrix.

Refining a structure on the basis of distances is an ill-behaved problem, and it was necessary to resort to nonlinear optimization to correct each of the trial structures. Nonlinear optimization has several problems; it is typically slow and converges to minima whether they are local or global. We developed two distinct optimization algorithms which we found to be useful. The first of these (algorithm A) is a common conjugate gradient optimization algorithm which treats the x, y, and z coordinates of each atom as unique optimization parameters. This routine adjusts the coordinates of each atom so as to minimize the penalty function resulting from disagreement between the structure and the original distance bounds.

The second refinement algorithm (algorithm B) manipulates groups of atoms, allowing each group to reorient itself with three rotational and three translational degrees of freedom. Groups may be defined as rigid or nonrigid; nonrigid groups permit relative motion of internal atoms. Rigid groups are always stereochemically correct, and we have found this algorithm to be effective in the initial stage of refinement. Although this algorithm reduces the number of optimization parameters dramatically, it is not as computationally efficient as the conjugate gradient algorithm due to the fact that the analytic gradient of the penalty function with respect to rotations is difficult to generate. Therefore, we have resorted to a refinement algorithm which does not require one to compute a gradient but requires memory on the order of n^2, where n is the number of parameters. This memory requirement currently renders this algorithm incapable of refining the entire structure simultaneously.

Our refinement strategy generally involves initial refinement of atom groups using algorithm B until most large distance errors and chirality errors have been relieved. Algorithm A is next used to refine the entire structure, which usually resulted in rapid convergence to a deep minimum of the penalty function.

During the refinement of the trial structures, false minima were generally encountered. These minima were detected by the program whenever the gradient of the error function became small before the value of the error function was small. When such a minimum was encountered, we examined the error for each residue and usually found one or two residues that needed to move but were trapped by bumping into one another. To extricate the structure from such a trap, a large random vector (5 Å) was added to the coordinates of the offending residues, and the refinement was continued. This allows the structure to reorient itself and usually resulted in an escape from the false minimum. We found it very valuable to detect such phenomena as early as possible in order to avoid wasting the time involved in converging to a false minimum.

Of the 26 trial structures, we picked the 8 refinements that had the smallest error value and subjected them to further refinement. Most of the abandoned trial structures were caught in very deep false minima—a common occurrence was that bases 1 and 12 would pair with their minor and major groove sides reversed with respect to the other base pairs. The 8 best refinements were then further refined for 5 repetitions consisting of a randomization of 1-Å maximum amplitude followed by 64 cycles of conjugate-gradient optimization. This procedure appeared to allow the structures to relax their bounds violations most rapidly. After this point, all eight structures had a total sum-of-squares disagreement with bounds of less than 1.0 Å^2; the largest single errors were less than 0.1 Å, and all stereochemistry was correct.

The final refinements were examined visually and found to be very similar. Small variations in the loop region were observed, but this was expected due to the "looseness" of the input bounds from this part of the molecule. All the stem regions formed right-handed helices, and the same base stacking arrangement was observed in the loop region of all the refined structures.

Figure 9 also shows three fully refined structures alongside the embedded tiral structures from which they are derived. The remarkable similarity between these three structures effectively defines them as the "same structure" within the limitations of the method. Figure 10 shows the superimposition of pairwise combinations of these structures in order to generate a visual basis for comparing structural similarities. Although residues 2 through 11 adopt a remarkably similar conformation, there is much more variation between structures for C1 and G12. These two residues form the terminal base pair at the blunt end of the double helix; it is worth noting that this base pair is subject to "terminal fraying" and its imino proton is not detectable at 20 °C, indicating that this end of the molecule may be less structured, although it is inherently more difficult to fully constrain terminal residues where NOEs are only observed on one side, rather than both sides, of that residue. There are several points worth mentioning in terms of how the four T residues form a hairpin loop. T5 partially stacks with the G4 of the terminal base pair while T8 stacks on the complementary C9 of this GC pair. The sugar–phosphate moiety of T6 effectively "turns the corner" of the loop leading to T7, which stacks on T8. Thus, the 5' strand of the stem (C1-G2-C3-G4) is extended by a T5 stack whereas the 3′ strand of the stem is extended by two additional residues, namely, T8 and T7, that stack above C9; T6 connects the ends of these two stacks. The imino proton of T6 appears to be the most exposed to solvent and should be one of the first resonances to exchange out of the spectrum at elevated temperature. The imino protons of T5, T8 and T7 form a tight cluster in the interior of the loop, with T7 somewhat more accessible to solvent. The methyl group of T7 is buried in the interior of the loop in close proximity to the 1′H of both T6 and T5 and to the imino proton of T8. The bases of T5 and T8 both point inward toward each other, and it is tempting to speculate that they might form a "wobble pair" with each other; in any case, their imino protons are close in space, and the T5 and T8 imino protons are both close to the imino proton of GC4 at the end of the double-helical stem.

Although graphic visual comparisons are useful, more rigorous structure comparisons can be made by analysis of deviations in coordinates, dihedral angles, and local structural parameters between structures. Space limitations preclude tabulating the Cartesian coordinates of every atom in the three structures shown; however, Tables I–III list several structural parameters and structural comparisons. Table I lists the dihedral angles α, β, γ, δ, ϵ, ζ, and χ, as well as the pseudorotation phase angle P for one of the structures. This constitutes sufficient information to recreate this structure by computer graphics. It is worth noting that, in the absence of 5′H and 5″H assignments, the linker region between sugars is not constrained (except by standard bond lengths) and hence varies between structures. The major experimental constraints are sugar–base and base–base distances, and these lead to quite good relative orientations of the nucleoside units in all

Table I: Torsional Angles for Refined Structure E

residue	α	β	γ	δ	ϵ	ζ	χ	P
1			-100.16	135.70	-172.11	-143.81	-115.60	139.77
2	67.81	112.78	-50.54	172.31	-86.79	170.24	-67.52	220.80
3	-62.32	116.46	28.87	142.44	-84.67	149.40	-70.50	144.75
4	136.14	-103.56	98.88	138.40	-84.25	163.54	-84.27	150.53
5	-69.12	132.84	30.35	154.38	-76.54	162.08	-52.83	161.39
6	-20.92	142.22	-9.76	168.17	-80.47	88.52	-89.72	188.67
7	65.71	153.70	178.14	158.05	58.59	41.16	-63.49	238.39
8	-119.03	-164.38	53.81	134.87	-82.36	119.44	-62.99	150.52
9	-6.78	113.72	-11.11	130.10	-85.34	149.03	-71.98	137.50
10	-41.14	114.44	11.67	145.77	-82.37	123.90	-85.88	150.51
11	12.90	103.42	-23.99	147.99	58.62	30.62	-63.09	213.53
12	-157.79	-95.49	4.14				-59.78	213.05

Table II: Local Structure Parameters Rise, Slide, Roll, Pitch, and Yaw for Three Refined Structures G, E, and Y[a]

step	dx	dy	dz	roll	pitch	yaw
			Structure G			
1-2	1.80	-0.85	3.92	20.21	40.05	4.22
2-3	-0.55	-1.20	3.82	8.11	4.76	12.36
3-4	-0.77	-1.00	4.94	39.75	48.96	-18.94
4-5	-0.73	-1.20	4.04	22.18	-5.27	17.74
5-6	-1.03	-1.14	5.15	-13.51	68.96	-37.71
6-7	0.20	-4.41	-2.60	18.21	5.79	-25.76
7-8	-1.03	0.68	4.57	19.16	38.44	33.27
8-9	0.17	0.41	5.46	29.55	34.87	-15.33
9-10	0.60	-3.51	2.59	-4.23	56.92	-0.71
10-11	-0.28	-1.55	3.61	3.85	8.06	8.53
11-12	0.93	0.77	4.04	26.65	36.35	-0.14
			Structure E			
1-2	1.48	-2.01	2.86	-0.36	42.20	2.52
2-3	-0.51	-1.04	4.02	7.19	3.04	14.49
3-4	-1.11	-0.84	4.27	34.66	41.62	-16.98
4-5	-0.30	-1.22	4.09	22.23	-0.43	21.55
5-6	-1.47	-1.38	4.85	-18.26	71.04	-46.46
6-7	-0.27	-4.38	-2.83	17.58	0.36	-23.35
7-8	-1.46	1.00	4.48	20.10	59.26	36.50
8-9	0.51	-0.52	5.48	29.20	15.15	-5.02
9-10	0.25	-2.80	2.57	-13.84	46.85	3.86
10-11	-0.08	-1.39	3.56	3.03	7.74	12.09
11-12	0.65	0.97	4.31	33.75	34.07	-3.75
			Structure Y			
1-2	1.34	-0.82	4.18	25.55	34.56	1.78
2-3	-0.69	-1.21	3.75	9.67	4.00	11.09
3-4	-0.94	-0.86	4.68	34.78	40.28	-10.46
4-5	-0.54	-1.37	3.84	16.45	4.76	17.89
5-6	-1.41	0.04	5.12	-1.98	60.05	-37.74
6-7	0.06	-4.87	-2.45	17.92	2.88	-15.93
7-8	-2.09	0.62	4.29	22.42	59.78	23.76
8-9	0.09	-0.50	4.97	18.61	12.94	-5.52
9-10	0.23	-2.29	3.13	2.60	48.72	-4.56
10-11	-0.20	-1.51	3.91	18.92	7.60	3.20
11-12	0.65	0.33	4.03	20.40	33.21	1.34

[a] dx, dy, and dz are listed in angstroms; roll, pitch, and yaw are listed in degrees.

Table III: Distance Deviations and Coordinate Deviations between Structures[a]

			E	Y
G	DM	total	187.99	204.80
		av	0.0027	0.0030
	RMSC	total	20.51	23.13
		av	0.055	0.062
E	DM	total		143.46
		av		0.0027
	RMSC	total		14.21
		av		0.038

[a] DM refers to deviations between distance matrices; RMSC refers to deviations between root-mean-square coordinates.

structures. Table II lists the values of dx, dy, and dz as well as base roll, pitch, and yaw between adjacent residues for the three refined structures shown in Figure 9. In this analysis, the origin of the local coordinate system was taken as N9 (purines) or N1 (pyrimidines) with the z axis normal to the base plane and the x axis parallel to the C6 → N3 vector (pyrimidines) or the C4 → N1 vector (purines). Thus, dz and dy are analogous to "rise" and "slide", respectively, for each pair of adjacent bases. Roll (x rotation), yaw (y rotation), and pitch (z rotation) have their usual connotation. The data thus allow numerical comparison of the similarities between the refined structures. Finally, the sum of the deviations between pairwise combinations of these three structures (denoted G, E, and Y) is tabulated in Table III in terms of their distance matrix deviations as well as their root-mean-square coordinate deviations. Structures G and Y are the least similar, but their average root-mean-square coordinate difference is only 0.062 Å.

Discussion

Although we have demonstrated the feasibility of determining the structure of unknown DNA conformations directly from solution NMR data, it is important to understand exactly what the results mean. Since our structure is constructed from a large number of approximate distances, it must be recognized that the structure itself is also approximate. The steep distance dependence of the NOE buildup rate results in remarkable precision in distance measurements at the local level, leading to local structural resolution at least as good as that normally obtained from macromolecule X-ray crystallography. However, the fact that the structure is created by combining many short distances, each with a finite error, led us to presume that these errors would propagate, leading to an exponential loss in structural resolution over longer distances. This was investigated by comparing pairs of structures in the following way. Distances between pairs of atoms were divided into 14 distance groups from 2 to 28 Å, each encompassing the range ±1 Å; i.e., an average interatomic distance of 9.1 Å would be classified in the 10-Å group, as would a distance of 10.9 Å. For each pair of structures, the difference between each specific interatomic distance in the two structures, as well as the average distance, was recorded. For instance, there are 9563 interatomic distances between 9 and 11 Å, and the sum of their deviations between structures G and E is 4696.1 Å, leading to an average difference of 0.491 Å at a distance of 10 Å. Such analyses allowed us to plot the average distance difference in 2-Å increments out to 28 Å (the longest interatomic distance in the molecule). The results are shown graphically in Figure 11. To our surprise, the errors appear to propagate only linearly, being about 0.75 Å at 20 Å and about 1 Å at 28 Å. Thus, even at distances approaching 10 times longer than the experimental NOE distances, the errors are relatively modest, and the overall resolution appears to be better than we anticipated. In terms of the level of determination, this

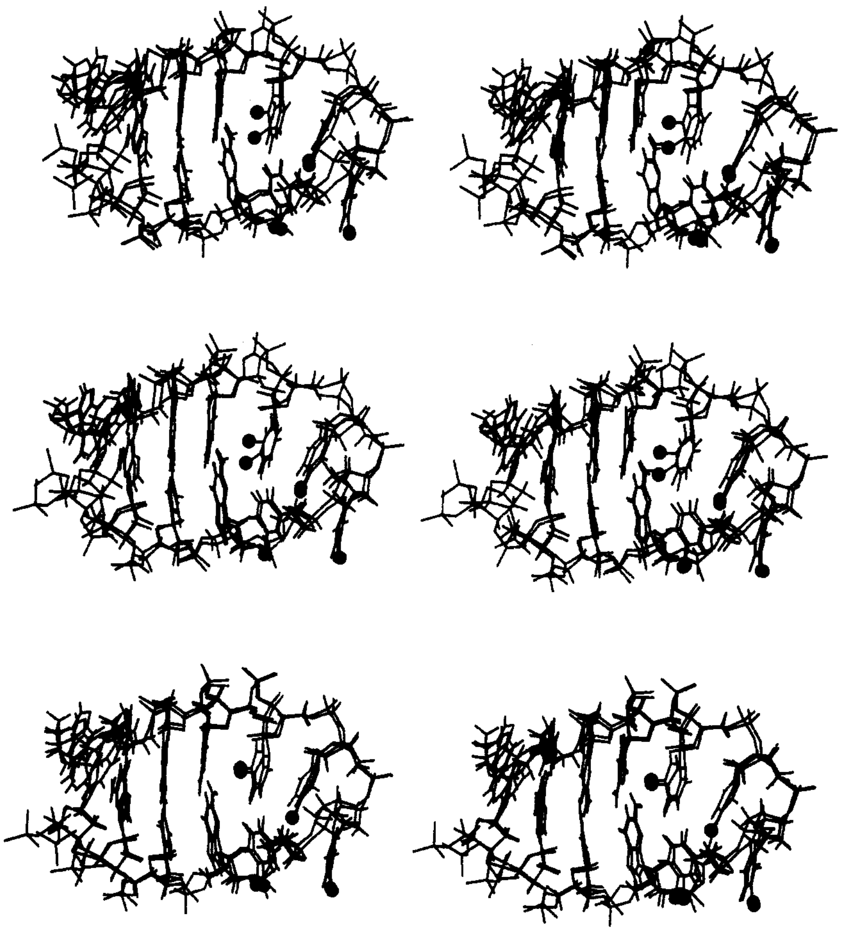

FIGURE 10: Stereoviews of superimposed pairs of the three refined hairpin structures G, E, and Y.

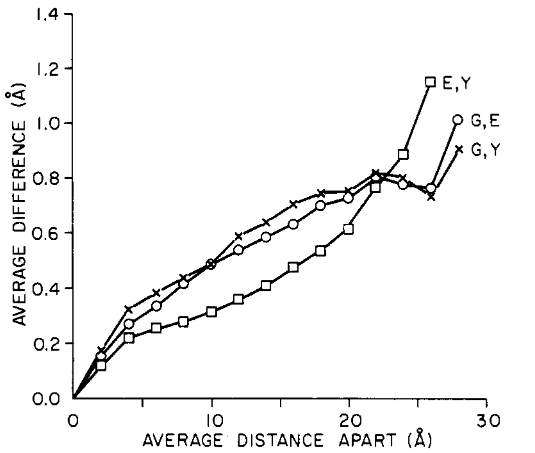

FIGURE 11: Plot of the average difference between interatomic distances in pairs of structures as a function of distance (see text for details).

result reassures us that, far from being underdetermined, the distance constraints are probably overdetermined and the long-range resolution presumably comes from a highly enmeshed network of distance constraints.

Compared with X-ray crystallography, structure determination based on NMR has both advantages and disadvantages. One obvious advantage of the approach outlined here is the ability to determine structure directly in solution, whereas crystallography requires that a sample be crystallized or oriented, often a difficult, and sometimes an impossible, task. The major disadvantage of the approach outlined here is the maximum size of biomolecule that is amenable to study (probably around 15 000 daltons at this time) due to assignment limitations, whereas the structure of significantly larger molecules can be determined by crystallographic techniques. We feel that both techniques are useful for different applications and yield quite different types of information.

Despite the remarkable similarity between the refined structures by visual inspection, a comparison of the structures with respect to coordinates or, more conveniently, dihedral angles is illuminating in terms of the strengths and weaknesses of the technique (coordinates are available upon request). To be useful in DG a proton must be unambigously assignable, yet the spectral region from 4.0 to 4.4 ppm is extremely crowded, containing the 4′, 5′, and 5″ protons of each residue, i.e., 36 resonances in the present case. Because of this crowding, these protons were not assigned and hence could not be used in the distance matrix. Thus, although the base–base, base–sugar, and sugar–sugar orientations are reasonably well-defined, the backbone "linker" between residues, consisting of O3–P–O5–C5′–C4′, is completely undefined. Thus, the sugar pseudorotation angles and base–sugar torsional angles are quite consistent between the various refined structures, but the interresidue backbone torsional angles (especially α, β, γ, and ϵ) exhibit quite large variations from structure to structure; i.e., the linker between sugars is relatively free to bridge the gap between defined nucleoside residues any way it chooses. Hence, in contrast to polypeptide structures, the side-chain positions and conformations of polynucleotide structures calculated by DG are well-defined whereas the precise backbone geometry emerges as a structurally feasible, but not necessarily correct, set of coordinates.

An important feature of the distance geometry technique involves the random selection of trial distances and trial structures to establish the extent to which the experimental

NOE contraints define a unique structure. If our distance data had not been sufficient to adequately determine a structure, i.e., if the system were inherently underdetermined, the resulting trial structures would have been very different, and refinement of these structures would not have yielded a similar set of structures, such as those obtained in this study. Rather, we would have obtained a set of very dissimilar structures, all of which obeyed the input bounds' constraints; i.e., distance geometry gives one a means to determine the extent of structural determination implicit in the input data and allows one to weigh the significance of the resulting structures accordingly. Thus, the distance geometry approach to determining DNA structure differs markedly from the work of Clore and Gronenborn (1985) in which X-ray coordinates are used as a starting point and are subsequently refined on the basis of distances measured by NMR. Although this type of refinement does result in structures, there is no test for completeness of the NMR data. For instance, in the extreme case in which only one or two NMR distances were used as input data, the structure resulting from this technique would differ in some ways from the X-ray structure but could not be construed as "the solution structure". Given a similar lack of data, distance geometry would generate a series of very different structures alerting the user to the fact that the system was underdetermined and that no structure determination was justified.

Finally, it is worth commenting on the computational aspects of the algorithm. Distance geometry requires substantial computer resources but is quite feasible with modern microcomputers. The smoothing, embedding, and refinement techniques used in the present work were performed on a DEC microVAX II computer (which operates at about 80% of the speed of a VAX 11/780) containing 9 Mbytes of memory. The DNA dodecamer studied here contains 372 atoms, and some representative CPU timings for each operation are as follows: (1) bounds smoothing, 60–70 min; (2) random distance selection and embedding, 30–40 min; (3) conjugate gradient refinement, 9 s/cycle; (4) complete refinement, ca. 6 h.

Thus, a dozen trial structures of this size can be embedded and refined in approximately 100 CPU hours on a dedicated micro VAX II. Unfortunately, the computation time requirements increase roughly as the third power of the number of atoms. The use of faster computer hardware, i.e., array processors, should make molecules of a thousand atoms or more, which are still within the range of NMR assignment methods, perfectly amenable to this technique. The structures generated by distance geometry in conjunction with NOESY spectra have not formally been subjected to energy minimization (although some of the constraints on allowed bond angles constitute a form of energy minimization), and it will be interesting to see to what extent energy minimization results in further improvement of the structure. Even if substantial improvement can be achieved, the DG-generated structure is virtually devoid of operator bias (unlike model building) and in the worst case would serve as an excellent starting structure for molecular dynamics and/or energy minimization calculations.

Acknowledgments

D.R.H. gratefully acknowledges support from National Institutes of Health SBIR Grant GM 35620, which provided partial funding of the computer equipment utilized for most of the distance geometry calculations in this project. We thank Professor Verner Schomaker for help and encouragement with methods of structure refinement. We also thank Patrick Chou and Foon Lee for synthesizing the DNA sample.

Registry No. d(CGCGTTTTCGCG), 103639-63-0.

References

Billeter, M., Braun, W., & Wuthrich, K. (1982) *J. Mol. Biol. 155*, 321–346.

Braun, W., Bosch, C., Brown, L. R., Go, N., & Wuthrich, K. (1981) *Biochim. Biophys. Acta 667*, 377–396.

Brown, L. R., Braun, W., Kumar, A., & Wuthrich, K. (1982) *Biophys. J. 37*, 319–328.

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

Clore, G. M., & Gronenborn, A. M. (1985) *EMBO J. 4*, 829–835.

Crippen, G. M. (1977) *J. Comput. Phys. 24*, 96–107.

Crippen, G. M. (1981) *Distance Geometry and Conformational Calculations*, Research Studies Press/Wiley, Chichester.

Haasnoot, C. A. G., de Bruin, S. H., Berendsen, R. G., Janssen, H. G. J. M., Binnendijk, T. J. J.; Hilbers, C. W., van der Marel, G. A., & vanBoom, J. H. (1983) *J. Biomol. Struct. Dyn. 1*, 115–129.

Hare, D. R., & Reid, B. R. (1982) *Biochemistry 21*, 519–5135.

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

Havel, T. F., & Wuthrich, K. (1985) *J. Mol. Biol. 21*, 5129–5135.

Havel, T. F., Crippen, G. M., & Kuntz, I. D. (1979) *Biopolymers 18*, 73–81.

Menger, K. (1931) *Jahres. Deutsch. Math.-Verein 40*, 201–219.

Redfield, A. G. (1978) *Methods Enzymol. 49*, 253–270.

Redfield, A. G., & Kunz, S. D. (1975) *J. Magn. Reson. 19*, 250–254.

Reid, D. G., Salisbury, S. A., Bellard, S., Shakked, Z., & Williams, D. H. (1983) *Biochemistry 22*, 2109.

Scheek, R. M., Russo, N., Boelens, R., & Kaptein, R. (1983) *J. Am. Chem. Soc. 105*, 2914–2917.

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

Wagner, G., & Wuthrich, K. (1982) *J. Mol. Biol. 155*, 347–366.

Wemmer, D. E., & Reid, B. R. (1985) *Annu. Rev. Phys. Chem. 36*, 105–137.

Wemmer, D. E., Chou, S. H., Hare, D. R., & Reid, B. R. (1985) *Nucleic Acids Res. 13*, 3755–3772.

Williamson, M. P., Havel, T. F., & Wuthrich, K. (1985) *J. Mol. Biol. 182*, 295–315.

Wuthrich, K., Wider, G., Wagner, G., & Braun, W. (1982) *J. Mol. Biol. 155*, 311–319.