Solution Structure of the Conserved Segment of the Myb Cognate DNA Sequence by 2D NMR, Spectral Simulation, Restrained Energy Minimization, and Distance Geometry Calculations

P. K. Radha, Anup Madan, R. Nibedita, and R. V. Hosur*

Chemical Physics Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400 005, India

Received October 14, 1994; Revised Manuscript Received February 17, 1995®

ABSTRACT: Solution structure of a self-complementary DNA duplex d-ACCGTTAACGGT containing the TAACGG recognition segment of Myb protein has been obtained by NMR spectroscopy. Complete resonance assignments of all the protons (except H5′, H5″ protons) have been obtained following standard procedures based on two-dimensional NMR techniques. Using a total of 72 coupling constants, and 95 NOE intensities, restrained energy minimization has been carried out, with the X-PLOR force field. The distance constraint set has been iteratively refined, for better fits with experimental NOE intensities. Using the final constraint set thus obtained, and explicit H-bond constraints for A·T, G·C base pairs in the duplex, distance geometry calculations have been carried in the torsion angle space with the program TANDY-2S to identify the family of structures consistent with the NMR data. We observe that the constraint set does indeed define a unique structure for the DNA segment. The structural details have been analyzed, and the sequence-dependent variations in torsion angles, base pair geometries, and helicoidal parameters have been documented. We observed that the helix axis displays a nonregular path, and three centered H-bonds have been seen at AA, AC, and CC steps in the major groove of the helix. Substantial variations have been observed for the helix axis and the groove widths at the recognition site. The base pairs exhibit high negative propeller twists. The structure is characterized by O4′-endo geometry for all the sugar rings (expect G10), and the other torsion angles belong to the B-DNA families.

Sequence-specific DNA recognition by a class of proteins called transcriptional factors plays an essential role in regulating the transcriptional activity in eukaryotic cells. These proteins contain a DNA binding module that functions independent of the rest of the protein, and the mechanism of specific recognition involves "reading" of the local DNA topography at the specific binding sites. Sequence-specific variations, in the geometric disposition of the functional groups of the bases, in the DNA helical geometry, and in local deformability of the DNA, define a unique surface which is complemented by the protein partner in the formation of specific protein−DNA complex. The specific protein−DNA interactions can be summarized as arising from two important sources (Steitz, 1990): (1) direct hydrogen-bonding and van der Waals interactions between protein side chains and exposed bases and (2) sequence-dependent bendability or deformability of the duplex DNA to accommodate the interacting protein side chains. Clearly, for any particular system the knowledge of the three-dimensional structure of the DNA cognate sites to high resolution is very essential for understanding the specific protein−DNA interaction and thus the mode of specific recognition.

Cellular Myb protein, an oncogene product, is a transcriptional activator which functions by binding to a specific site in the DNA. The protein contains a DNA binding domain at the N-terminus which folds into an autonomous structure entity and recognizes the cognate site in the DNA. The nucleotide consensus sequence of myb recognition had been identified to be YAACKG (Beidenkapp et al., 1988), where Y = T/C and K = G/T. More recently, the consensus sequence has been extended to 8 bases, and the derived sequence is represented as YAACKGHH, where H = A/C/T (Weston, 1992). The extension, however, has substantial variability, and thus its role is unlikely to be in making specific contacts. We have been interested for the last few years in understanding the specific myb−DNA interactions, and with this view, we report here a detailed NMR investigation of the three-dimensional solution structure of a self-complementary DNA dodecamer d-ACCGTTAACG-GT containing the most conserved segment TAACGG of the whole recognition sequence. Besides TAACGG and its complement CCGTTA, the sequence contains two additional nucleotides at the ends to minimize the end-fraying effects on the structure of the cognate site.

MATERIALS AND METHODS

(a) DNA Synthesis, Purification, and Sample Preparation. The 12 mer DNA sequence d-ACCGTTAACGGT was synthesized using β-cyanoethyl phosphoramadite chemistry (McBride et al., 1983) on an automated DNA synthesizer (Applied Biosystems, Model 381 A). The 5′-(dimethyltrityl) group was retained which helped in purification by reverse phase HPLC[1] using a PRP-I column. The trityl group was

* Author for correspondence.

® Abstract published in *Advance ACS Abstracts,* April 1, 1995.

[1] Abbreviations: NOE, nuclear Overhauser effect; HPLC, high performance liquid chromatography; EDTA, ethylenediaminetetraacetate; E.COSY, exclusive correlation spectroscopy; NOESY, nuclear Overhauser effect spectroscopy; TSP, 3-(trimethylsilyl)[2,2,3,3-$^2H_4$]-propionate; SICOS, simulation of correlated spectra; SIMNOE, simulation of NOESY spectra; TANDY, torsion angle approach to nucleic acid distance geometry; *rmsd,* root mean square deviation.

0006-2960/95/0434-5913$09.00/0

then cleaved by acid cleavage and the DNA passed through a DOWEX column to get the sodium salt of the oligomer.

The purified product was dried, dissolved in phosphate buffer, pH 7.0, containing 0.1 mM EDTA, and lyophilized 2−3 times from ^2H_2O to remove the exchangeable protons. It was finally dissolved in 0.5 mL of 99.99% ^2H_2O to yield a concentration of approximately 3 mM on single strand basis. For the experiments in ^1H_2O, a 90% ^1H_2O and 10% ^2H_2O solvent mixture was used.

(b) NMR Experiments. NMR spectra were recorded on a BRUKER AMX 500 spectrometer operating at 500 MHz frequency for ^1H. Data for the two-dimensional (2D) phase sensitive E.COSY (Griesinger et al., 1986) spectrum were collected with 2048 points along t_2 and 512 points along t_1 directions, employing time proportional phase incrementation, TPPI (Redfield & Kunz, 1975), for quadrature detection along the F_1 dimension. Two types of E.COSY spectra were recorded with multiple quantum filtering contributions extending to 3 and 4 quantum filters. The former are useful for analyzing H1′−(H2′,H2″) cross peaks in the 2D spectrum, while the latter simplifies the H2′−H2″ cross peaks in the spectrum. For the experiment extending to 3 quantum filters, the 12 step basic phase cycle (Griesinger et al., 1987) was coupled with a 2 step 180 phase alternation to cancel effects of pulse imperfections, yielding a total of 24 step phase cycle; likewise, for the experiment extending to 4 quantum filters, a (32 × 2) step phase cycle was used. The time domain data were zero filled to 1024 points along t_1 before window multiplication by $\pi/4$ shifted sine function and subsequent 2D Fourier transformation. Digital resolutions were 1.97 and 3.94 Hz/pt along F_2 and F_1 dimensions, respectively. Five two-dimensional NOESY spectra (Jeener et al, 1979; Anil Kumar et al., 1980) were recorded with mixing times of 100, 200, 250, 300, and 400 ms in a single day without removing the sample from the spectrometer. For these, the data set consisted of 1024 t_2 and 400 t_1 points each. For each t_1 the signal was averaged for 32 scans. These spectra were used for obtaining NOE buildup curves, and it was observed that for most peaks NOE intensity was maximum at a mixing time of 300 ms. Consequently, a NOESY spectrum with this mixing time and higher signal averaging and resolution with 2048 t_2 and 512 t_1 points was recorded for simulations and quantitative interpretation in terms of 3D structure. A NOESY spectrum in 90% ^1H_2O and 10% ^2H_2O was recorded with a mixing time of 150 ms at 25 °C for identification of exchangeable protons and also nonexchangeable AH_2 protons. For this, the last pulse in the NOESY sequence was replaced by a Jump and Return sequence (Plateau & Gueron, 1982). The data set consisted of 2048 t_2 and 300 t_1 points. In all cases the data were zero filled to 1024 points along t_1 before apodization by $\pi/4$ shifted sine functions and 2D Fourier transformation. Chemical shifts are expressed with respect to internal TSP in the spectra, in an indirect manner by calibrating the HDO signal.

(c) Simulation of Cross Peak Patterns in E.COSY Spectra. The fine structures of cross peaks in E.COSY spectra were simulated using the software package SICOS (Majumdar, 1990; Majumdar & Hosur, 1992). The time domain data processed with UXNMR software on an X-32 data station were transferred to an IRIS 4D 70/G work station, and simulated peaks were superimposed on calculated peak patterns for examination of the fits. The chemical shifts and coupling constants were altered iteratively until a satisfactory

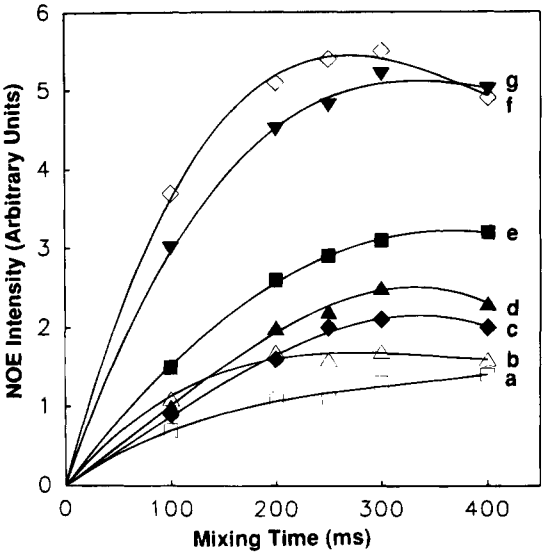

FIGURE 1: NOE buildup curves for a few of the cross peaks in the NOESY spectra. The various curves have been labeled and correspond to the cross peaks: (a) A8H1′−C9H6, (b) A7H4′−A7H1′, (c) A8H2′−C9H6, (d) C9H2″−C9H6, (e) C9H2′−C9H1′, (f) C3H5−C3H6, and (g) A7H2″−A7H1′. We notice that for most of the peaks the NOE intensity is maximum at 300 ms mixing time, and therefore this mixing time has been used for structure optimization by NOESY simulation.

fit resulted between experimental and calculated patterns. We observed that, for well-resolved peaks, the calculated patterns were sensitive to 0.1 Hz variations in the coupling constants, and therefore the precision of the coupling constants derived can be considered to be of that order. A constant line width of 5 Hz was used along both dimensions. Overlapping peaks were simulated together.

(d) Simulation of NOESY Spectra. Relaxation matrix based simulation of NOE spectra has proved to be the best method for structure optimization of DNA segments (Macura & Ernst, 1980; Borgais et al., 1990; Nibedita et al., 1993; Landy & Rao, 1988). A common trend in the literature has been to use several mixing times for NOESY simulation, expecting to improve the statistics and confidence level in the structure determination. We too have attempted this before (Nibedita et al., 1993) but observed that, due to different S/N ratios of cross peaks in different spectra, the intensity measurements were not accurate to the same extent in all the cases, and thus all the spectra could not be fitted simultaneously to the same level of accuracy. Considering that it is best to have an accurate fit with best intensities, we decided here to simulate one spectrum as accurately as possible, for a reliable structure.

Figure 1 shows the NOE buildup curves for a few distinct cross peaks. We have actually obtained such curves for 27 peaks from different regions of the spectrum and observed that for most of the peaks the NOE intensity is maximum around 300 ms mixing time. Therefore, this mixing time has been used for structure optimization by NOESY simulation. Simulations have been carried out using the SIMNOE algorithm, whose details have been discussed elsewhere (Nibedita et al, 1992). The algorithm employs the relaxation matrix approach (Ernst et al., 1987) for calculation of NOE cross peak intensities, and the current version assumes a single isotropic correlation time for the macromolecule. Whether this is a valid assumption or not is a debatable question, and the conclusions can vary from system to

system. Some authors have used multiple correlation times for DNA segments, optimizing them to fit the NOE data (Baleja et al., 1990). Some others have used a single correlation time and have optimized it for proper fit. In either case the bottom line has been the quality of fit. In a particular study, Reid and co-workers (Reid et al., 1989) have measured the base and sugar correlation times and found them to be the same. Reid and co-workers have also justified use of a single isotropic correlation time for such a large DNA duplex on the basis of hydrodynamical theory (Tirado & Garsia de la Torre, 1980; Wang et al., 1992). Our experience has been similar to that of Reid and co-workers, and we were also able to obtain good fits with NOE data with a single correlation time model. The correlation time was optimized for best fit, and we observed that the effect of changing correlation time was essentially to shift the average line of $(E_i - S_i)$ vs i plot, where E_i and S_i are the experimental and calculated intensities for peak number i. We noted that for a value of 5.5 ns the average was around zero and hence was considered most satisfactory. In the following, we briefly describe the individual steps in the NOE simulations, and this we feel is necessary since our method of data treatment is somewhat different in many respects from that of most other workers in the field.

(i) Integration of NOE Cross Peak Intensities. NOE cross peaks were integrated by summing up the points spanning the peaks above selected threshold values. The threshold values as percentages of peak heights were different for different peaks because of the different fine structures and different signal/noise ratios in different regions of the spectra. These integrals were then normalized with respect to some preselected peak and stored in an array along with the respective thresholds. In such a situation it is important to note that the intensities of two different peaks cannot be compared to draw conclusions such as "stronger peak," "shorter distance," etc. These are, however, ideal for comparison with respective calculated intensities, wherein the effects of different thresholds, different fine structure, etc., have been taken into account explicitly. This also minimizes the errors associated with peak integrations.

(ii) Data Base of Simulated Intensities above Specific Thresholds. The SIMNOE algorithm of NOESY simulation, and iterative comparison with experimental intensities, requires the creation of a data base of all the peak shapes, calculated by taking into consideration the widths and fine structure, and with an intensity of 1.0 for each peak. The calculated peaks are integrated above the same thresholds as is done for the respective experimental peaks. Figure S-1 (supplementary material) displays a few of the data base simulations and the respective integrals obtained above particular threshold values. We observe that, even though the total "analog" intensity (Nibedita et al., 1992) is the same for all these peaks, the "digital" intensities are different depending upon the fine structure or the threshold values or both. These digital intensities are then normalized, which then constitute the data base of calculated peak shapes. The structure optimization process which requires fast comparison of calculated and experimental intensities makes use of the data base entries and scales them appropriately.

(iii) Scaling of Calculated NOEs and Comparison with Experimental Intensities. The relaxation matrix calculated NOE intensities are analog numbers and cannot be directly compared with experimental intensities, since the experimental peaks are "digital", have widths, and have fine structures. Therefore, for reliable comparison, the calculated intensities must be cast in the form of spectra with peak shapes, fine structures, etc., and the "digital" calculated intensities must be obtained by integrating the peaks, above the proper thresholds as in the experimental spectra. As we have shown earlier (Nibedita et al., 1992), this can be achieved by making use of the data base described above. If $\mathbf{A}_{ij}$ is the relaxation matrix calculated intensity for a particular peak between spins i and j, and if $\mathbf{C}_{ij}$ is the corresponding data base intensity (which corresponds to an analog intensity of 1.0), then the "digital" intensity $\mathbf{D}_{ij}$ for the particular peak can be computed as

$$\mathbf{D}_{ij} = \mathbf{A}_{ij} \times \mathbf{C}_{ij} \tag{1}$$

Since the experimental and calculated intensities will have to be compared repeatedly at every step of structure refinement, the above procedure renders digitization of calculated peaks very fast and efficient. It results in considerable saving of computational time.

The above procedure is straightforward to use, when the peaks are well resolved and easily integratable. However, difficulties arise when two or more peaks overlap to different extents, and then the scaling equation 1 is not applicable. For example, if three peaks with intensities I_1, I_2, and I_3 overlap, there will be no separate thresholds for the individual peaks, and there are no separate data base entries for the three peaks. In that case, conversion of the corresponding analog intensities a_1, a_2, a_3 into "digital" intensities becomes difficult. The sum $(a_1 + a_2 + a_3)$ is anyway not comparable with the total experimental intensity. To circumvent this problem, we have adopted the following procedure.

First, the overlapping peaks are simulated together, with appropriate chemical shifts and coupling constants, assuming an intensity of 1.0 for each of the peaks. Using a common threshold, the total peak is integrated to give a value, say, S_t. Next, the peaks are separated artificially by changing their positions, and again the peaks are simulated assuming individual intensities of 1.0 and integrated above the same threshold to obtain the individual digital intensities S_1, S_2, S_3, ..., etc. Then the total intensity S_t is apportioned in the proportion $S_1{:}S_2{:}S_3$..., etc., to obtain the relative contributions of the individual peaks to the total "digital" intensity. The resultant values, say, b_1, b_2, b_3, are then entered into the data base. The intensity scaling equation for overlapping peaks, the equivalent of eq 1, is then given as

$$D^\circ = a_1 b_1 + a_2 b_2 + \ldots \tag{2}$$

where D° represents the digital intensity of the calculated overlapping peak. We have successfully used this approach to calculate intensities of up to 4−5 overlapping peaks. The procedure is illustrated in Figure S-2, in the supplementary material.

(e) Distance Constraints and Restrained Energy Minimization. NOE based refinement of the structure of a DNA segment involves iterative alterations of the structure, followed by relaxation matrix calculations and comparisons with experimental NOE intensities. Here, alteration of the structural model has to be guided in some way to drive the changes toward a better fit. For this purpose we have used restrained energy minimization (Nilsson et al., 1986; Nilges

et al, 1987a,b) using X-PLOR software package (Brunger, 1990) which allows modification of a structure so as to satisfy specified constraints. Starting from a particular structural model, a set of NOE intensities is calculated and compared with the experimental intensities. Looking at the peak-to-peak fits, we judge in which direction a particular interproton distance has to change, and depending upon the quality of fit, a rough estimate of the expected change is obtained. Thus, for each experimental peak a distance constraint for the proton pair represented by the peak is derived. This yielded a set of 190 distance constraints (95 per strand). These constraints are incorporated as inputs for the NOE square potential in X-PLOR, defined by E_{NOE} in the following equation, and rigid body minimization is carried out:

$$E_{NOE} = S\Delta^2 \tag{3}$$

where Δ is constraint violation defined as

$$\Delta = \begin{cases} R - d_{plus} & d_{plus} < R \\ 0 & d_{minus} < R < d_{plus} \\ d_{minus} - R & R < d_{minus} \end{cases} \tag{4}$$

S is the restrained force constant; d_{plus} and d_{minus} define the upper and the lower distance bounds, respectively; R is the actual distance in the molecule. The minimization process included also the sugar geometry constraints (five torsion constraints per ring) derived from the coupling constant data and defined by a similar square potential for dihedral angle violations.

The force constant for NOE potential was varied at regular intervals in the range 20−400 kcal/(mol·Å^2), monitoring the progress of minimization. When there was a reasonable improvement in the NOE potential, a new refined distance constraint set for the same experimental peak list was derived from the current structure and minimization performed again to obtain further improvement in the NOE fit and the potential. This process was repeated until no further improvement occurred. The number of steps of energy minimization varied between 20 and 200 for each constraint set, and about 20 times the distance constraint set was refined. The force constant for dihedral angle potential was kept fixed at 60 kcal/(mol·rad^2). During this iterative procedure, the following points were given due attention: (i) The bond length and bond angle energy did not vary much, indicating that the covalent bond criteria as defined in the X-PLOR parameter file (version 2.1) were not violated; the default force constant was used for this purpose. (ii) Since the distance constraint was the same for the two strands in the duplex, the derived structure should show equivalence of the strands with respect to all torsion angles.

(f) Distance Geometry in Torsion Angle Space. NOE based refinement in conjunction with energy minimization certainly provided one of the most satisfactory structures for the DNA segment. However, to further investigate the uniqueness of the distance constraint set, with regard to structure, we have used distance geometry calculations. The calculations have been carried out using the TANDY-2S program developed in our laboratory (Ajay Kumar, 1993); this is an extension of TANDY for double stranded DNA and incorporates a novel angle geometry algorithm for formation of H-bonds (Ajay Kumar, 1992). It uses explicit distance and angle constraints to maintain H-bonds in the base pairs. Several structures were calculated using the final distance constraint set (190 constraints obtained as described above) and pseudorotation angles corresponding to the final structure, and starting from different initial structures generated by randomly flexing the various torsion angles. For the distance constraints a range of ±0.5 Å around the distance value in the final energy minimized structure was used. The acceptable minimum for the target function value was initially set to 1.0 and was progressively decreased to 0.1 for achieving better convergence. The convergence limit was 0.01 for two successive iterations. For every converged structure, the NOE fit was calculated to see if there was any degeneracy of structures with regard to NOE fits.

RESULTS AND DISCUSSION

(a) Resonance Assignment. Sequence-specific assignments of the various hydrogens in the DNA molecule were obtained following well-established procedures based on 2D *J*-correlated and NOESY spectra (Scheek et al., 1983; Hare et al., 1983; Ravikumar et al., 1985; Wuthrich, 1986). Stereospecific assignments of the H2′, H2″ protons have been obtained on the basis of the relative intensities of H1′−H2′ and H1′−H2″ cross peaks in a low mixing time NOESY spectrum; the latter is always more intense than the former, irrespective of sugar geometry (Hosur et al., 1988). Sequence-specific assignments are illustrated in Figures S-3 and S-4, and the chemical shifts of assigned protons have been provided in Table S-1 of the supplementary material.

(b) $^1H-^1H$ Coupling Constants. We have analyzed the H1′−(H2′, H2″) and H2′−H2″ cross peak fine structures in E.COSY spectra to derive all the coupling constants in the sugar rings of the DNA molecule. The cross peak fine structures are a sensitive function of the sugar geometries, and an extensive survey of the H1′−(H2′,H2″) fine structures has been recently reported (Majumdar & Hosur, 1992). The H2′−H2″ cross peak, which also has a wealth of information, has however not been exploited until now. Figures S-5 shows best fit simulations along with their experimental counterparts for several H1′−(H2′, H2″) cross peaks and a particular H2′−H2″ cross peak as illustrations. From the progress of the simulation and matching exercise, we estimate that all the coupling constants except H2″−H3′ are precise to 0.1 Hz, while the precision of *J*(2″,3′) is approximately 0.3 Hz. All the derived coupling constant values are listed in Table 1. These coupling constant values were translated into specific ranges for individual sugar geometries in terms of pseudorotation angles using Karplus type relations (see review: Majumdar & Hosur, 1992), and these are also listed in Table 1. Such an interpretation is driven by considerations that the Karplus type relations are empirical and there will be local fluctuations in the geometries in duplex DNA segment. The measured coupling constants will therefore be averages over these ranges. There was no specific attempt to quantitatively fit the data to a two-state or a three-state model and a particular Karplus type relation.

The coupling constants *J*(4′,5′) and *J*(4′,5″) could not be measured from E.COSY spectra, since the relevant H4′−(H5′,H5″) cross peaks lie very close to the diagonal and also overlap heavily with each other. However, from the overall widths along H4′ axis—which represent sums of *J*(3′,4′), *J*(4′,5′), *J*(4′,5″), and *J*(4′,P)—in the H1′−H4′ cross peaks

Table 1: Various Coupling Constants As Obtained from the Simulation of the E.COSY Spectra and Ranges of the Sugar Pseudorotation Angle (P)

residue	$J(1',2')$ (Hz)	$J(1',2'')$ (Hz)	$J(2',2'')$ (Hz)	$J(2',3')$ (Hz)	$J(2'',3')$ (Hz)	Range of P (deg) lower	upper
A1	7.8	6.4	−14.0	7.5	3.0	97.0	132.0
C2	8.0	6.2	−14.0	7.5	3.0	97.0	132.0
C3	8.4	6.3	−14.5	7.8	3.0	89.0	128.0
G4	8.3	6.8	−14.5	8.0	3.0	103.0	124.0
T5	7.5	7.0	−15.0	7.5	4.0	93.0	132.0
T6	8.5	6.1	−14.0	7.8	3.8	106.0	127.0
A7	8.6	6.0	−13.5	8.0	3.0	108.0	124.0
A8	8.0	7.0	−14.0	8.0	3.0	99.0	124.0
C9	*a*	*a*	*a*	*a*	*a*	*a*	*a*
G10	9.4	5.8	−14.5	8.7	1.0	113.0	162.0
G11	8.0	7.0	−14.0	8.0	3.0	99.0	141.0
T12	*b*	*b*	*b*	*b*	*b*	*b*	*b*

[a] These could not be evaluated due to weak cross peaks. [b] These could not be evaluated due to severe overlap of peaks.

in the NOESY spectra (Mukhopadhyay et al., 1992), we estimated that these coupling constants are less than 5 Hz in all the nucleotide units. $J(4',P)$ values are usually less than 2 Hz, and $J(3',4')$ values are known indirectly from sugar geometries derived above.

(c) $^1H-^{31}P$ Coupling Constants. The $^1H-^{31}P$ coupling constants of significance for structure determination are $J(3',P)$ and $J(5',5''\text{-}P)$, which are related to torsion angles ϵ and β, respectively. Of these, the latter are extremely hard to obtain due to poor dispersion of H5′ and H5″ protons in 2D spectra. In the present case, we observed that the chemical shift dispersions for all the three proton types H3′, H5′, H5″ and also for ^{31}P were very poor. Thus, explicit measurements of the heteronuclear coupling constants were not possible. However, with the knowledge of $^1H-^1H$ coupling constants and the total widths along H3′ axis of different cross peaks in E.COSY and NOESY spectra, we estimated that H3′−P coupling constants are less than 5 Hz in all the nucleotide units. This information, although crude, certainly puts bounds on the acceptable values of the ϵ torsion angle. In fact, the ϵ ranges turn out to be very narrow because of the steep relation between $J(H3'-P)$ and the ϵ torsion angle (see reviews by Govil & Hosur, 1982; Hosur et al., 1988; Van de ven & Hilbers, 1988; Majumdar & Hosur, 1992).

(d) Experimental NOE Peak Intensities. Table S-2 lists the normalized intensities and the corresponding threshold values for 95 peaks, obtained as described in Materials and Methods. The overlapping peaks were integrated together, and thus for each group of peaks, there is a common threshold value. We must mention here that, because of the different thresholds for different peaks, the normalized peak intensities of the various peaks are not directly comparable to draw inferences such as "a particular peak is stronger than the other peak", etc. The peak list does not include the peaks belonging to thymine methyls, since the simulation process does not include effects of methyl rotation. Further, the peak list includes mostly those peaks which are far from the diagonal, and the peaks are integrated several times with different box sizes and averaged to minimize manual errors; the residual errors could be estimated to be about 10%.

(e) Iterative Comparison of Calculated and Experimental Peak Intensities. This represents the final and most crucial step of structure optimization in solution media. The structure of the DNA segment is continuously changed until its

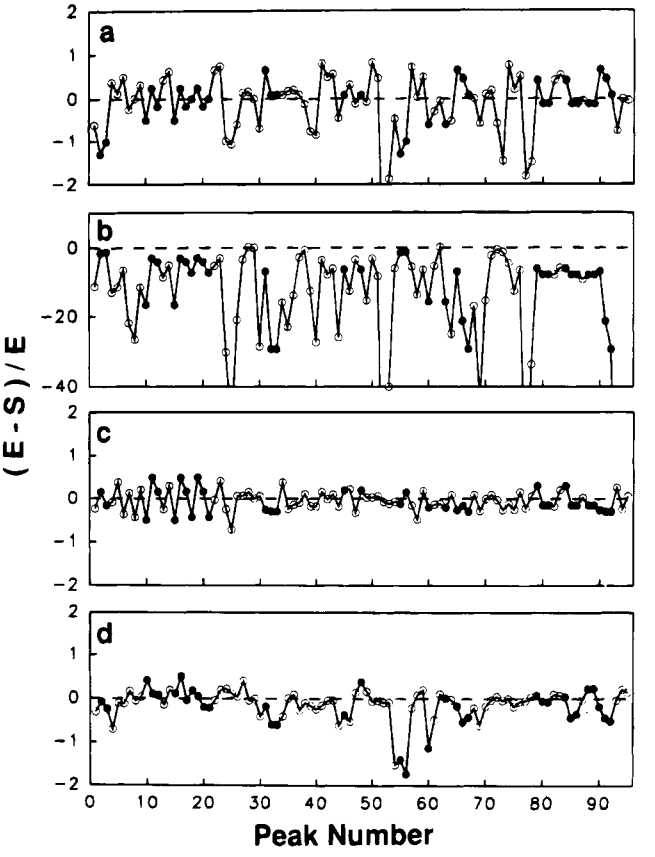

FIGURE 2: (a) Comparison of experimental and calculated NOE peak intensities for the B-DNA model. Open circles represent nonoverlapping peaks, and filled circles represent overlapping peaks. The overlapping peaks have been simulated together. The peaks have been arranged sequence wise from left to right, and the peak numbers are same as used in Table S-2. (b) Peak-to- peak fit of experimental and calculated NOE peak intensities for the A-DNA model. Note that the scale for this plot is much higher than that for the others. The large deviations in the fit here indicate that the structure of the oligonucleotide would be far away from that of the standard A-DNA form. (c) Peak-to-peak fit of experimental and calculated NOE peak intensities for the final structure S1. The *R* factor for the fit was calculated to be 0.27 Å. (d) Comparison of simulated NOE intensities for the final structure S1 and another S2 which differs from S1 to an extent of 0.68 Å on an all-atom *rmsd* scale.

calculated spectrum satisfactorily fits the experimental spectrum. In such an exercise, the starting structural models and the methodology of structure refinements may have an influence in achieving a proper convergence. A good initial guess would lead to a faster convergence and to a better fit with experimental data. Therefore, we first calculated the intensities of all the peaks listed in Table S-2, for B- and A-DNA models, and Figure 2a,b shows peak-to-peak comparison of simulated (S) and experimental (E) intensities for the two models. The initial fit with the B-DNA model is seen to be better than that with the A-DNA model. We then derived separate distance constraint sets (190 distances) from the two models and carried out restrained energy minimization using the X-PLOR package. We observed that, starting from the B-DNA model, the convergence was better and faster. Then, starting from energy minimized B-DNA and corresponding constraint set, the DNA structure and distance constraints were iteratively refined to obtain a better fit and an energetically well favored structure (see Materials and Methods). Figure 2c shows the final peak-to-peak fit

Table 2: Energy Statistics for the Initial B-DNA and for the Final Structure after Restrained Energy Minimization[a]

energy terms	values of the initial structure (kcal/mol)	values for the final structure (kcal/mol)	energy change ΔE (kcal/mol)
total	**3077.16**	**−34.33**	**−3111.49**
grad (E)	71.85	0.28	−71.57
bond	20.05	110.16	90.11
angle	172.35	580.33	407.98
dihedral	556.31	25.24	−531.07
improper	0.005	25.24	25.235
van der Waals	−102.79	−316.53	−213.74
electrostatic	−648.94	−737.25	−88.31
hydrogen bond	−80.09	−95.35	−15.26
sugar torsion angle violation	44.93	0.44	−44.49
NOE distance violation	3115.32	3.54	−3111.78

[a] Final force constant is 20 kcal/(mol·Å^2) for NOE potential and 1 kcal/(mol·rad^2) for sugar torsion potential.

Table 3: Convergence Characteristics of 24 DG Generated Structures

distance constraints	
no. of constraints	190
no. of violations	15 ± 11
av violation	0.108 ± 0.045
greatest violation	0.25 ± 0.6
H-bond constraints	
no. of constraints (3 per H-bond)	90
no. of violations	0
NOE R factor	0.3063 ± 0.06

between the calculated and the experimental intensities. A total of 1650 steps of minimization was carried out, and the constraint set was updated every 20−200 steps as explained in Materials and Methods. The total energy of the initial (B-DNA) structure was 3077.16 kcal/mol and finally approached a stable value of −34.33 kcal/mol. The energy statistics for the initial and final structures are listed in Table 2.

To establish the relation between the goodness of the fit and the goodness of the structure, we compared the calculated intensities of the final structure (labeled S1) with those of another structure (labeled S2) which was different from the final structure to the extent of 0.68 Å on the *rmsd* scale. The results are shown in Figure 2d in the same manner as in Figure 2c. We observe that, for many peaks, the deviations range between 25% and 50%, and for a few, it is even higher. This indicates that the fit in Figure 2c, where the deviations are less than 20% for most peaks, is highly satisfactory and consequently the structure derived therefrom is a true representation of the solution structure of the oligomer in aqueous solutions; it would be unreasonable to expect a better fit because of the errors in the experimental intensities and the assumption involved in the relaxation matrix calculations.

(f) Distance Geometry Search of Conformational Space. In order to find out if the energy minimized structure, which is most acceptable in terms of both energy and fit with experimental data, is unique with regard to the latter, we carried out extensive distance geometry calculations in the torsion angle space (Ajay Kumar et al, 1991) using the program TANDY-2S. The convergence statistics for 24 converged structures is summarized in Table 3. In Figure S-6 the backbone torsion angles for all the nucleotides of the 24 converged distance geometry generated structures have been shown in the form of dials. We observe from this analysis that all the DG structures are very similar, with torsion angles varying mostly within 1−5°, and in a few cases, the deviation is up to 10°. The pairwise *rmsd* on all atom scale ranges from 0.2 to 1.2 Å for the 24 structures, and the NOE R factors (Table 3) calculated using the following equation:

$$R \text{ factor} = [\sum_{i=1}^{N}(E_i - S_i)^2]^{1/2}/(\sum_{i=1}^{N}E_i) \qquad (5)$$

are also similar. This shows how, in spite of the widely different initial structures, the final DG generated structures are highly similar. The conformational energies of all the final structures varied within 5%, with the X-PLOR generated structure having the lowest energy. Thus, we are convinced that the structure we have obtained is unique and that the NOE data type and size are fairly adequate for uniqueness when used in conjunction with the H-bond and sugar geometry constraints in TANDY-2S program. In the following, we analyze the features of the energy minimized structure in greater detail.

(g) Analysis of the Energy Minimized Structure. The conformational parameters of the energy minimized structure have been extracted using the NUPARM package (Bhattacharya et al., 1989), and these are described below.

(i) Torsion Angles. Values of the torsion angles for both the strands along with the values for canonical B-DNA and A-DNA are presented in Figure 3. It is satisfying to see that all the torsion angles are equivalent in both the strands of the duplex; this must be expected as the DNA is self-complementary. We may also mention that such an equivalence is not seen in well-solved crystal structures, indicating that crystal forces do sometimes contribute to structural artifacts.

The results in Figure 3 indicate that the *backbone torsion angles* do not vary significantly from one residue to another and belong to the B-DNA family. The *phosphate backbone* is found to be in the common B_I state with ϵ angles in the trans and ζ angles in the gauche conformation. The B_I conformation in general is energetically more favored as compared to the B_{II} conformation, as in B_I, the C3′−O3′−P elbow is flat on the surface of the DNA cylinder unlike in B_{II}, where it points inwards, leading to steric hindrances (Fratini et al., 1982; Prive et al., 1991).

The *sugar pseudorotation phase angles P*, for all the nucleotides, are seen to be in the O4′-endo conformation, except for G10 which is in the C1′-exo conformation. The *glycosidic torsion angles* χ are all in the −anticlinal (−ac, −gauche) conformation. The χ for G10 residue was found to be much higher than that for the other residues. This is expected since, with an increase in the P value, the χ value also increases due to decrease in steric clashes between sugar and the base atoms (Saenger, 1984).

(ii) Helicoidal Parameters. The helicoidal parameters are classified into three categories: intrabase helicoidal parameters, the local step helicoidal parameters, and the global helicoidal parameters. For the present structure, these are displayed in Figures 4 and 5. In all the three cases, we observe a symmetry in the values along the sequence of the molecule.

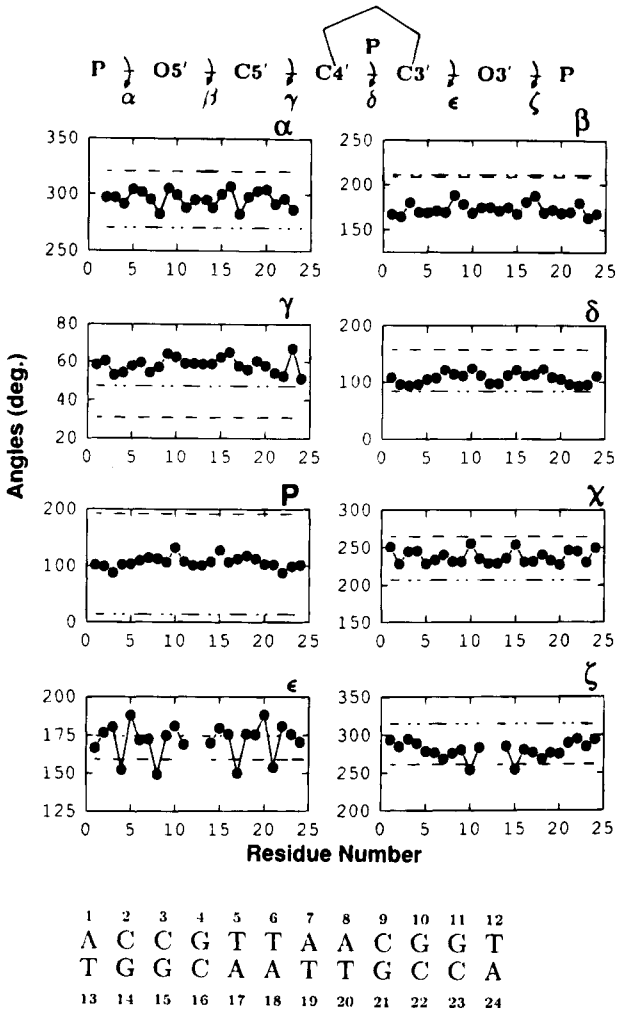

FIGURE 3: Analysis of the energy minimized structure in terms of various torsion angles. Torsion angles in both the strands have been plotted to show the equivalence of the two strands in the final structure. The standard values for B- (- - -) and A- (— • • —) DNA models have also been indicated.

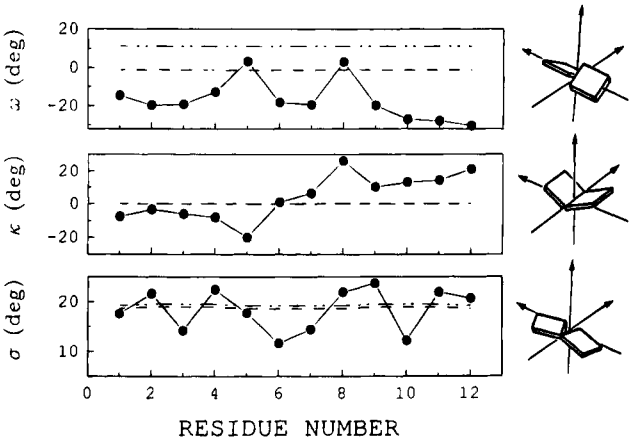

FIGURE 4: Helicoidal parameters for the energy minimized structure. (a) Intra base pair helicoidal parameters which define the relative orientations of the two bases along the x axis (buckle, κ); y axis (propeller twist, ω); and z axis (inclination, η).

Among the *intrabase parameters* shown in Figure 4, the value of propeller twist (ω) is largely negative, being much more negative than for B-DNA. The propeller twists for T5•A17 and A8•T20 are nearly zero, indicating flattening of these base pairs. The buckle (κ) has negative values for the initial five base pairs and positive for the last five base pairs. The numerical values of buckle for T5•A17 and A8•T20 are much higher than for the others. The opening (σ) for all the nucleotides is positive and varies in a zig-zag manner. The smallest σ is found for T6•A18 and G10•C22 base pairs.

The *local step helicoidal parameters* also show a systematic variation along the sequence, as seen in Figure 5a. The tilt is positive for the initial five steps, 0 for the T6•A18—A7•T19 step, and negative for the last five steps. This indicates that for the first five base pair steps the opening is toward strand I and it is toward strand II for the last five steps. The roll (ϱ) is positive or around zero for most of the base steps except G4-T5 and A8-C9 steps, where it is largely negative. These negative values indicate widening of the major groove and narrowing of the minor groove at these steps. The large roll at these two positions is compensated by a decrease in the propeller twist so as to prevent destacking of the bases. The twist (Λ) shows a zig-zag pattern all through the base steps.

The shift (Dx) is positive for the first half of base steps and negative for the rest. The slide (Dy) is more negative for purine—pyrimidine compared to pyrimidine—purine steps, which is in accordance with the Callidine's rules (Calladine, 1982; Weisz et al., 1994) and prevents clashes between interstrand purines. The rise (Dz) is below 3 Å for the terminal residues, and around 3.2 Å for the inbetween ones, which is close to that in the B-DNA structure. The rise in the G4-T5 step is much larger, ~3.5 Å, which may be to avoid steric clashes due to sudden bending at this base step.

Among the *global helicoidal parameters*, shown in Figure 5b, the inclination (η) and twist (Λ) do not show much variation and lie around 10° and 35°, respectively. The tip (θ) varies largely from +15° to −15°, which is a direct reflection of the local step parameter tilt. The x, y, and z displacements are directly dependent on the shift, slide, and rise, respectively, and are seen to exhibit fraying effects. The z displacement for the G4 is much higher than for the others as seen also for the rise.

The *helix axis* (Figure 6a) shows a smooth curvature from T6-A7 to G10-G11 steps, and a sudden kinking is observed at the G4-A5 step. The total curvature of the helix axis is of the order of 30—40°. The path of the *base pair normals*, defined by the mean of the two base normals in the base pair, is shown in Figure 6b. The base normals are seen to exhibit a non-even path with fluctuations at the C3, G4, T5, C9, and the end positions. These effects are directly reflected on the propeller twists and the values of buckle at these positions as seen earlier.

(iii) Groove Widths. The groove widths of the double helix are defined using the coordinates of the phosphate atoms. The smallest separation between the phosphate atoms in the two antiparallel strands reduced by the sum of the van der Waals radii of the two phosphate atoms (=5.8 Å) is used to define the minor groove width in the B-DNA and major groove width in the A-DNA. The minor groove generally occurs between i and $(i - 4)'$ phosphates for the B-DNA and i and $(i - 3)'$ phosphates for the A-DNA (Mujeeb et al., 1993; Bhattacharya & Bansal, 1992).

The minor and major groove (Naryana et al., 1991) widths for the present DNA along with those for the canonical A-DNA and B-DNA were calculated and are plotted in Figure 7. As seen, the minor groove is compressed for the T6—A8 stretch along with a simultaneous widening of the

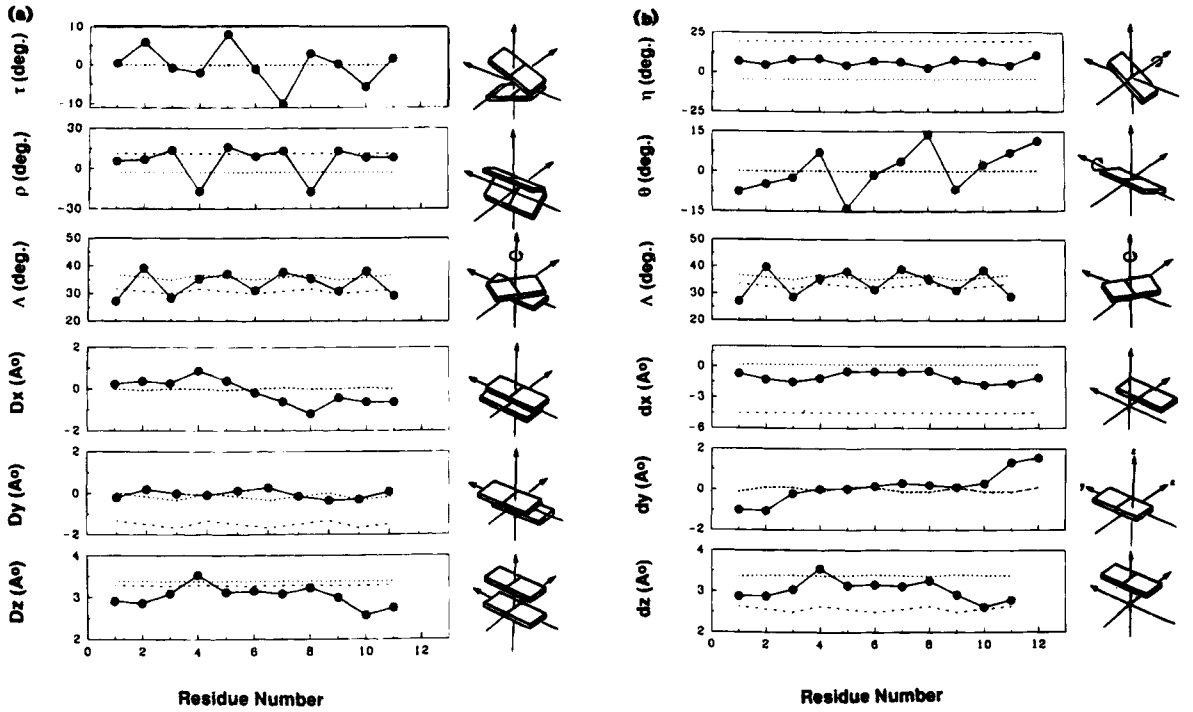

FIGURE 5: (a) Local step helicoidal parameters for the energy minimized structure. The first three panels describe the rotation and the latter three the translations along the three axes. (b) Global helicoidal parameters of the structure. The definitions of the parameters are indicated in their respective icons. The standard values for B- (•••) and A- (— • —) DNA models have also been indicated.

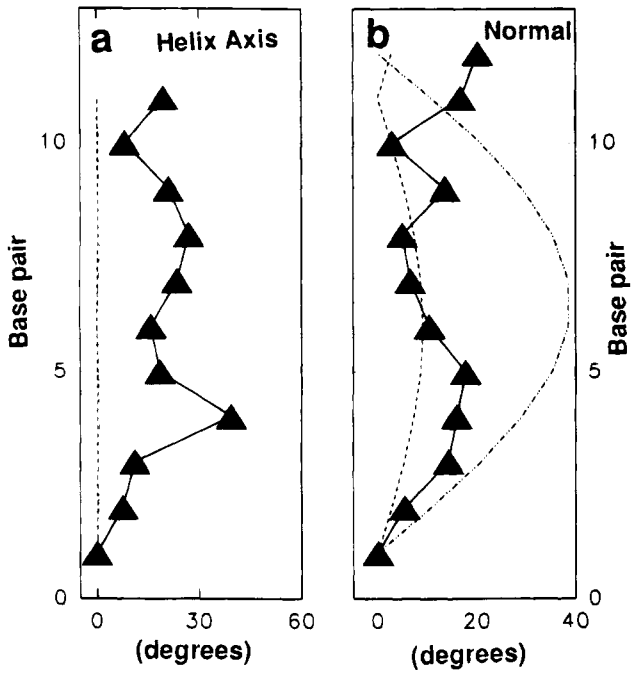

FIGURE 6: Course of the local helix axis (a) and the base pair normals (b) with respect to the first base pair as the reference. The standard values for B- (•••) and A- (— • • —) DNA models have also been indicated.

major groove at these positions. The minor groove widths show a symmetry along the sequence as expected for a self-complementary DNA.

(iv) Three-Center H-Bonds. In view of the observed variations in base pair geometries and helicoidal parameters, we analyzed the structure for the possible presence of three-center H-bonds, which have been seen in many crystal structures (Nelson et al., 1987; Coll et al., 1987). The detailed analysis of several crystal structures indicated that three-center H-bonds are possible, only for the steps AA, AC, CC, and CA in the major groove and at CC and CT steps in the minor groove. Various criteria for the validity of the three-centered bonds are dictated as (i) the planarity of the three heavy atoms involved in the bonding, (ii) the two H-bond distances should be in accepted ranges and (iii) a high negative propeller twist so as to orient the bases involved accordingly (Jeffrey & Mitra, 1984; Heinmann & Alings, 1989).

In the present case, we found valid three-center H-bonds at the following steps: A1-C2, C2-C3, and A7-A8 steps on one strand and also at the A24-C23, C23-C22, and A18-A17 steps in the complementary strand. The detailed geometries of these three-center H-bonds are shown in Figure 8.

CONCLUSIONS

We have described in this paper precise determination of solution structure of a biologically important DNA sequence, namely, the recognition sequence of cellular myb protein. The structure determination has relied on careful interpretation of 2D NMR data in a quantitative manner employing many of our own approaches and indicates that the NOE data do indeed define an accurate structure for the DNA segment. The extensive distance geometry calculations with TANDY-2S algorithm have demonstrated that the structure determination is indeed unique, and they have also allowed scanning of the structures which are in the vicinity of this final structure, in the conformational space of multidimensions.

The structure of the myb recognition DNA sequence is seen to be different from canonical B- and A-DNA models. All the torsion angles are found to exist in the well-preferred orientations, with the phosphate backbone in the B_I conformation. The molecule exhibits three-center H-bonds at AC,

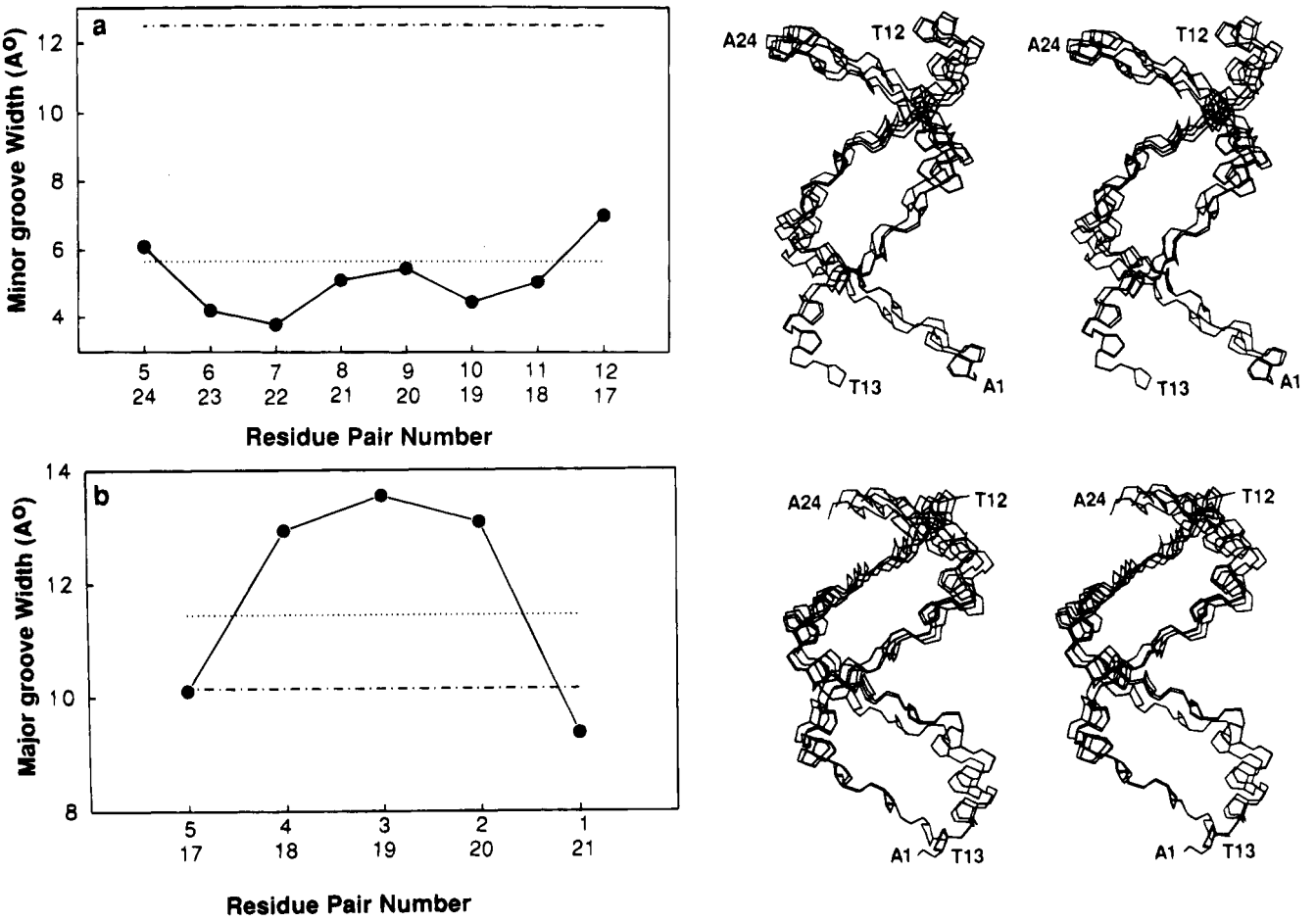

FIGURE 7: Plots of the minor groove widths (a) and the major groove widths (b) along the sequence of the DNA as measured from the interstrand P—P distances. The phosphorus labels along which the distances are measured are indicated on the x axis.

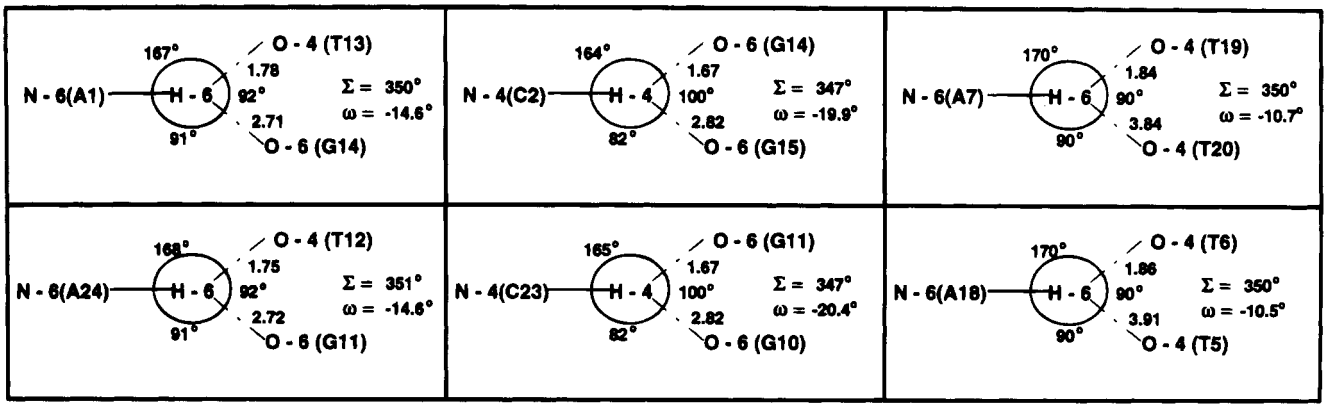

FIGURE 8: Geometries of the three-centered H-bonds observed in the molecule. Σ gives the angle sum around the hydrogen position taken as a measure of coplanarity of the four atoms involved, and ω gives the propeller twist at the base step.

CC, and AA steps, on both the strands. The helix axis shows substantial variations from the B-DNA with a smooth roll from T6 to G11 which constitutes the myb recognition site. The analysis of the groove widths shows the widening of the major groove between T6 and A8 with simultaneous narrowing of the minor groove. This length forms the TAA segment of the recognition site TAACGG, and the above structural observations are significant in view of the recent observation by Ording et al. (1994), using DNA footprinting techniques, that the sequence-specific recognition occurs via direct interactions of the R3 domain of the myb protein with the TAAC segment; widening of the major groove at these positions and the kink at the G4 position (complementary of C in TAAC) would lead to facilitation of the specific interactions.

The structure of the present DNA segment may be compared with that of GCCAAT recognition box recognized by many other transcriptional regulators, the structure being determined by us employing similar procedures (Nibedita et al., 1993). The latter had positive propeller twists as against many negative propeller twists in the present case. The latter also exhibited a higher degree of variations in sugar geometries and some of the backbone torsion angles. Such differences indicate that different transcriptional regulators may have different mechanisms of DNA recognition and this recognition may have different roles in their function.

ACKNOWLEDGMENT

We thank the National Facility for High Field NMR supported by Government of India and located at the Tata Institute of Fundamental Research for providing excellent NMR and computational facilities. We thank Bhabha Atomic Research Center for the use of the X-PLOR package. We thank Dr. M. Bansal for the NUPARM package.

SUPPLEMENTARY MATERIAL AVAILABLE

Two tables of chemical shifts and NOE peak intensities; six figures illustrating the resonance assignments, spectral simulations, and distance geometry generated structures (14 pages). Ordering information is given on any current masthead page.

REFERENCES

Ajay Kumar, R. (1992) *J. Biomol. NMR 2*, 519−526.
Ajay Kumar, R. (1993) Ph.D. Thesis, University of Bombay, Bombay.
Ajay Kumar, R., Hosur, R. V., & Govil, G. (1991) *J. Biomol. NMR 1*, 363−378.
Anil Kumar, Ernst, R. R., & Wuthrich, K. (1980) *Biochem. Biophys. Res. Commun. 95*, 1−6.
Baleja, J. D., Pon, R. T., & Sykes, B. D. (1990) *Biochemistry 29*, 4828−4839.
Beidenkapp, H., Borgmeyer, U., Sippel, A. E., & Klempnauer, K. H. (1988) *Nature 335*, 835−837.
Bhattacharyya, D., & Bansal, M. (1989) *J. Biomol. Struct. Dyn. 6*, 635−653.
Bhattacharyya, D., & Bansal, M. (1992) *J. Biomol. Struct. Dyn. 10*, 213−226.
Borgias, B. A., Gochin, M., Kerwood, D. J., & James, T. L. (1990) *Prog. NMR Spectrosc. 27*, 83−100.
Brunger, A. T. (1990) *X-PLOR (Version 2.1) Manual*, Fellows of Harvard University, New York.
Calladine, C. R. (1982) *J. Mol. Biol. 161*, 343−352.
Coll, M., Fredrick, C. A., Wang, A. H. J. & Rich, A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 8385−8389.
Dickerson, R. E. (1989) *J. Biomol. Struct. Dyn. 6*, 627−634.
Ernst, R. R., Bodenhausen, G., & Wokaun, A. (1987) *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*, Clarendon Press, Oxford.
Fratini, A. V., Kopka, M. L., Drew, H. R., & Dickerson, R. E. (1982). *J. Biol. Chem. 257*, 14686−14707.
Govil, G., & Hosur, R. V. (1982) *Conformation of Biological Molecules: New Results from NMR*, Springer Verlag, Heidelberg.
Griesinger, C., Sorenson, O. W., & Ernst, R. R. (1986) *J. Chem. Phys. 85*, 6837−6852.
Griesinger, C., Sorenson, O. W., & Ernst, R. R. (1987) *J. Magn. Reson. 75*, 474−492.
Hare, D. R., Wemmer, D. E., Chou, S. H., Drobny, G. R., & Reid, B. R. (1983) *J. Mol. Biol. 171*, 319−336.
Heinemann, U., & Alings, C. (1989) *J. Mol. Biol. 210*, 369−381.
Hosur, R. V., Govil, G., & Miles, H. T. (1988) *Magn. Reson. Chem. 26*, 927−944.
Jeener, J., Meier, B. H., Bachmann, P., & Ernst, R. R. (1979) *J. Chem. Phys. 64*, 4546−4554.
Jeffrey, G. A., & Mitra, J. (1984) *J. Am. Chem. Soc. 106*, 5546−5553.
Landy, S. B., & Rao, B. D. N. (1988) *J. Magn. Reson. 83*, 29−43.
Macura, S., & Ernst, R. R. (1980) *Mol. Phys. 41*, 95−117.
Majumdar, A. (1990) Ph.D. Thesis, University of Bombay, Bombay.
Majumdar, A., & Hosur, R. V. (1992) *Prog. NMR Spectrosc. 24*, 109−158.
McBride, L. J., & Caruthers, M. J. (1983) *Tetrahedron Lett. 24*, 245−248.
Mujeeb, A., Kerwin, S. M., Kenyon, G. L., & James, T. L. (1993) *Biochemistry 32*, 13419−13431.
Mukhopadhyay, N., Majumdar, A., & Hosur, R. V. (1992) *Spectrochem. Acta 48A*, 1731−1737.
Narayana, N., Ginell, S. L., Russu, I. M., & Berman, H. M. (1991) *Biochemistry 30*, 4449−4455.
Nelson, H. C. M., Finch, J. T., Luisi, B. F., & Klug, A. (1987) *Nature (London) 330*, 221−226.
Nibedita, R., Ajay Kumar, R., Majumdar, A., & Hosur, R. V. (1992) *J. Biomol. NMR 2*, 467−476.
Nibedita, R., Ajay Kumar, R., Majumdar, A., Hosur, R. V., Govil, G., Majumdar, K., & Chauhan, V. S. (1993) *Biochemistry 32*, 9053−9064.
Nilges, M., Clore, G. M., Gronenborn, A. M., Brunger, A. T., Karplus, M., & Nilsson, L. (1987a) *Biochemistry 26*, 3718−3733.
Nilges, M., Clore, G. M., Gronenborn, A. M., Pile, N., & McLaughlin, L. W. (1987b) *Biochemistry 26*, 3734−3744.
Nilsson, L., Clore, G. M., Gronenborn, A. M., Brtnger, A. T., & Karplus, M. (1986) *J. Mol. Biol. 188*, 455−475.
Ording, E., Kvavik, W., Bostad, A., & Gabrielsen, O. S. (1994) *Eur. J. Biochem. 222*, 113−120.
Plateau, P., & Gueron, M. (1982) *J. Am. Chem. Soc. 104*, 7310−7311.
Prive, G. G., Yanagi, K., & Dickerson, R. E. (1991) *J. Mol. Biol. 217*, 177−199.
Ravikumar, M., Hosur, R. V., Roy, K. B., Miles, H. T., & Govil, G. (1985) *Biochemistry 24*, 7703−7711.
Redfield, A., & Kunz, S. D. (1975) *J. Magn. Reson. 19*, 250−254.
Reid, B. R., Banks, K., Flynn, P., & Nerdal, W. (1989) *Biochemistry 28*, 10001−10007.
Saenger, W. (1984) *Principles of Nucleic Acid Structure*, Springer, New York.
Scheek, R. M., Russo, N., Boelens, R., Kaptein, R., & van Boom, J. H. (1983) *J. Am. Chem. Soc. 105*, 2914−2916.
Steitz, T. A. (1990) *Q. Rev. Biophys. 23* (3), 205−280.
Tirado, M. M., & Garcia de la Torre, J. (1980) *J. Chem. Phys. 73*, 1986−1993.
Van de Ven, F., & Hilbers, C. W. (1988) *Eur. J. Biochem. 178*, 1−38.
Wang, A. C., Kim, S. G., Flynn, P. F., Sletter, E., & Reid, B. R. (1992) *J. Magn. Reson. 100*, 358−366.
Weisz, K., Shafer, R. H., Egan W., & James, T. L. (1994) *Biochemistry 33*, 354−366.
Weston, K. (1992) *Nucleic Acids Res. 20*, 3043−3049.
Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York, NY.

BI942398F