NMR Structure of the Extended Myb Cognate Sequence and Modeling Studies on Specific DNA−Myb Complexes‡

P. K. Radha, P. K. Patel, and R. V. Hosur*

Department of Chemical Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai 400 005, India

Received March 24, 1998; Revised Manuscript Received May 4, 1998

ABSTRACT: The recognition sequence of the Myb protein has been recently described to be *py*AACKGHH (where *py* = T/C, K = G/T, and H = A/C/T), modifying the earlier identification as pyAACKG [Ording, E., et al. (1994) *Eur. J. Biochem. 222*, 113−120]. We had earlier determined the solution structure of the minimal cognate sequence TAACGG, choosing *py* = T and K = G, embeded in a 12-mer DNA duplex by NMR and related computational techniques [Radha, P. K., et al. (1995) *Biochemistry 34*, 5913−5912]. To understand the structural significance of the above modification and the role of the variability in the recognition sequence, we have investigated here the solution structure of a different DNA segment, d−ACAACTGCAGTTGT, which contains the extended Myb cognate site, CAACTGCA. The three-dimensional structure of the 14-mer duplex has been determined from NMR data by relaxation matrix and restrained molecular dynamics calculations. The structure of the above cognate sequence in the 14-mer duplex has been compared with that of the cognate sequence, TAACGG, in the 12-mer duplex and also with that in the NMR structure of the Myb DNA binding domain (R2R3)−DNA complex determined by Ogata et al. recently [Ogata, K., et al. (1994) *Cell 79*, 639−648]. The comparison highlighted differences in several structural parameters for the cognate sites in the DNA segments. Modeling studies by taking out the protein from the complex and presenting it with 12-mer and 14-mer DNA structures indicated that the protein induces structural alterations to drive the cognate site to a reasonably conserved structure. The extent of similarity of the derived structures was, however, dependent on the base sequences. Base changes in the minimal cognate sequence in the 12-mer−protein complex and in the 14-mer−protein complex so as to match the sequence of Ogata et al. produced a more conserved structure of the complex. A reverse exercise, in which the Ogata DNA in the complex was mutated to match the 12-mer and 14-mer minimal cognate sequences, complemented the above observations of the subtle sequence dependence of the structure in the complex. On the other hand, base changes in the extension did not influence the DNA−protein complex structure significantly. We also observed that the structural changes in the protein were very minor when different DNA sequences or different DNA structures were presented to it. These observations would be of interest from the point of view of DNA−Myb recognition.

Interactions between proteins and DNA are responsible for a variety of functions in a living cell. Among these is the transcriptional regulation, which plays an essential role in the central dogma. The process of transcription is controlled by various proteins such as repressors and activators, and these factors act by sequence specific interactions with the DNA. The Myb protein is one such transcriptional activator which is known to control the growth and differentiation of hematopoietic cells (*1*−*3*). The protein has a DNA-binding domain at its N-terminus, which is evolutionarily conserved, and is known to contain three imperfect repeats (labeled as R1, R2, and R3) of 51−52 residues each (*2*, *4*−*8*). The cognate sequence of this DNA-binding domain had been identified to be *py*AACKG (where *py* = T/C and K = G/T) (*4*, *5*). More recently, Ording et al. (*9*) have extended the cognate site to *py*AACKGHH (where H = A/C/T) and have shown that the presence of the extension increases the binding affinity of the protein. Furthermore, they suggested that, while R3 repeat participates mostly in specific interactions with the core sequence, AAC, the R2 repeat is largely involved in nonspecific interactions with the extension. Binding studies with DNAs of different lengths indicated that the length of the DNA influences the binding affinity to a significant extent (*9*−*11*). The studies also indicated that within the cognate sequence itself the different nucleotide units have varying degrees of importance in bringing about specific recognition. In this sense, the AAC stretch is seen to be the most critical part of the recognition sequence.

In view of the above interesting observations on Myb−DNA interactions and recognition, we had determined earlier the solution structure of a self-complementary DNA segment d-ACCGTTAACGGT containing the minimal cognate site (underlined) at the 3′-end (*12*, *13*) by NMR methods. The structure showed some interesting features, like enlargement of major groove, kinks, 3-centered hydrogen bonds etc., which were thought to have direct relevance to the Myb binding. This DNA did indeed interact with the DNA-binding domain of *Drosophila* Myb protein in a very specific

‡ PDB file name: 1bdz.

* Author to whom correspondence should be addressed.

S0006-2960(98)00675-8 CCC: $15.00
Published on Web 06/20/1998

manner (*11*) and was also found to bind at a hydrophobic site on the protein (*14*). Recently, Ogata et al. have determined the solution structure of the mouse R2R3−DNA complex by NMR and molecular dynamics methods (*15*−*17*). The DNA was a nonsymmetrical 16-mer duplex and contained the cognate sequence TAACTG in the center. They identified several contact points between the DNA and the protein, and it was observed that a number of contacts were made with the core recognition sequence by the R3 domain of the protein. The structure of the cognate DNA in their complex showed, however, some differences from the structure of the DNA studied in our laboratory. Is this difference a consequence of the difference in the DNA sequences? Are there induced structural changes in the DNA on complex formation? Does complex formation induce a conformational change in the protein? Ogata et al. had also determined earlier the solution structure of the R3 domain alone (*16*) and concluded by comparison that the structure of the R3 domain was unchanged on complex formation. In contrast, it has been suggested on the basis of gel mobility assays and circular dichroism studies that the conformation of the cognate DNA changes on binding to the protein (*18*). We have tried to extend the above studies in the present paper. First, we describe the solution structure of another self complementary DNA, d-ACAACTGCAGTTGT which contains the extended cognate site (underlined). This sequence contains the AACTG stretch as in the R2R3−DNA complex of Ogata et al. (*15*) and a comparison of the two would reflect on the effects of complex formation on the cognate DNA structure. A comparison of the structure of the CAACTG stretch with that of TAACGG (*12*, *13*) would throw light on the effects of sequence changes on the structure of the minimal cognate site. Further, a comparison of all the three structures is expected to show the effects of the cognate site extension on the topology of the Myb cognate site. The NMR structure of the 14-mer DNA has been determined using complete relaxation matrix-based calculations, and the conformational space has been searched by simulated annealing calculations.

Next, we describe model-building studies starting from the solution structure of the R2R3−DNA complex determined by Ogata et al. earlier (*15*). The DNA in the complex was replaced by the 12-mer and 14-mer DNA structures, and the complexes were energy minimized to see possible structural alterations. Mutations were created in the DNAs in the complexes to match the different cognate sequences, and then the complexes were again energy minimized to see the possible role of the sequence on the structure of the DNA in the complex and also to see possible structural changes in the protein. These model-building studies indicated that the structure of the cognate site was preserved in the complex. That means, depending upon the structure of the free DNA molecule in aqueous solution, which is dependent on the base sequence, complex formation induces appropriate conformational changes in the DNA. The protein structure itself did not seem to change significantly when different DNA sequences were presented to it.

MATERIALS AND METHODS

(a) Sample Preparation. The oligonucleotide d(ACAACTGCAGTTGT) was synthesized on an automated synthesizer (ABI 381A) at 10 μmol synthetic level using phosphoramidite chemistry (*19*). After the synthesis, the DNA was cleaved from the column and deprotected using liquid ammonia. The product was purified to homogeneity by gel electrophoresis using 20% denaturing polyacrylamide gels, followed by electroelution of the bands (*20*). The DNA thus extracted was repeatedly dialyzed against water and 4 M NaCl to remove small molecular weight impurities. The samples for NMR were prepared by repeated lyophilization and dissolution in 99.99% D_2O for the D_2O samples. Phosphate buffer (10 mM) with 100 mM NaCl salt concentration at pH 7.0 was used for the final samples.

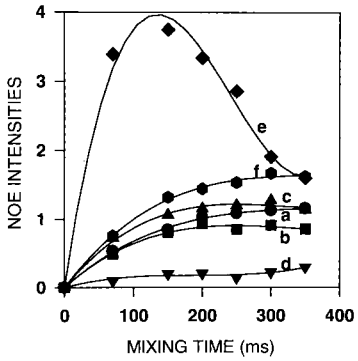

FIGURE 1: Illustrative NOE build-up curves for selected cross-peaks in the NOESY spectra: (a) A9H1′−G10H8, (b) T6H2′−G7H8, (c) T6H2″−G7H8, (d) C2H3′−A3H8, (e) C5H2′−C5H2″, (f) C5H2′−C5H1′.

(b) NMR Experiments. All NMR experiments were performed on a BRUKER AMX 500 NMR spectrometer. Water suppression was achieved by presaturation. 2D spectra recorded included clean TOCSY (*21*) with 50 ms mixing time and NOESY (*22*, *23*) spectra with 70, 150, 200, 250, 300, and 350 ms mixing times under identical conditions. In all the cases, the data set consisted of 2048 t_2 and 512 t_1 time domain points and quadrature detection was achieved by TPPI (*24*) procedure. 2D heteronuclear inverse detected ^{13}C-^{1}H HMQC spectrum (*25*) was recorded at natural abundance using 1024 t_2 and 400 t_1 points. The spectra were processed using 2048 × 1024 points after apodization with a $\pi/2$ shifted sine bell window function along both the dimensions. Separate E. COSY spectra (*26*) with multiple quantum orders up to 3 and 4 contributing to the fine structures of the peaks were recorded. For these, the data sets consisted of 4096 t_2 and 1024 t_1 points and the data was Fourier transformed after zero filling to 2048 points along t_1 and multiplying by sine functions shifted by 45° along both the axes. Spin lattice (T_1) and spin−spin (T_2) relaxation times for the base protons were measured by the inversion recovery and spin−echo based methods, respectively, and these were used to obtain rough estimates of the correlation times (*27*). ^{31}P NMR spectra were recorded to see the dispersion in the resonances which reflect on the heterogeneity in the backbone structure of the molecule.

(c) NOESY Intensities and Relaxation Matrix Calculations. We first obtained NOE build-up curves for all types of peaks in the spectra so as to choose the best mixing time for the NOE intensity based iterative structure determination to be described in the later paragraphs. Figure 1 shows a few illustrative curves, and we notice that at a mixing time of 200 ms most peaks have very good intensities. Consequently, this spectrum was chosen for the simulations. We have described earlier that such an approach is more

Table 1: Table of ^{13}C and ^{1}H Chemical Shifts for the 14-mer DNA Duplex[a]

	H8/H6	C8/C6	H2/H5/HMe	C2/C5/CMe	H1′	C1′	H2′	H2″	C2′	H3′	C3′	H4′	C4′
A1	8.21	142.0	8.01	155.2	6.21	87.0	2.63	2.79	40.2	4.88	78.7	4.29	89.7
C2	7.44	143.0	5.51	98.9	5.22	86.7	2.05	2.28	39.3	4.84	77.2	4.13	86.7
A3	8.25	142.0	7.53	154.5	5.95	85.2	2.81	2.93	40.6	5.11	80.0	4.44	87.5
A4	8.18	140.7	7.69	155.2	6.16	84.7	2.70	2.86	40.5	5.06	79.2	4.53	87.3
C5	7.24	139.5	5.16	97.7	5.75	86.7	1.91	2.44	40.2	4.66	76.2	4.22	85.7
T6	7.28	139.7	1.55	14.75	5.79	86.2	2.08	2.46	39.5	4.89	78.0	4.15	86.0
G7	7.85	138.5			5.82	84.3	2.61	2.68	40.2	4.99	79.2	4.38	87.2
C8	7.34	142.0	5.39	98.6	5.51	86.5	1.97	2.36	39.0	4.84	77.2	4.17	85.7
A9	8.20	142.0	7.42	154.5	6.09	85.2	2.78	2.94	40.3	5.04	78.7	4.43	87.5
G10	7.61	137.2			5.88	85.0	2.48	2.75	39.7	4.95	78.5	4.43	87.5
T11	7.20	138.5	1.28	14.2	6.02	86.0	2.06	2.55	39.4	4.87	78.7	4.26	86.0
T12	7.32	142.7	1.67	14.7	5.88	85.2	2.05	2.41	39.7	4.92	77.5	4.12	87.3
G13	7.92	138.8			6.11	85.0	2.74	2.74	40.3	5.01	79.2	4.44	87.5
T14	7.42	139.7	1.93	14.6	6.28	86.7	2.30	2.30	41.4	4.58	72.2	4.12	87.3

[a] Chemical shifts are with respect to TSP [sodium 3-trimethylsilyl-(2,2,3,3-$^{2}H_4$) propionate].

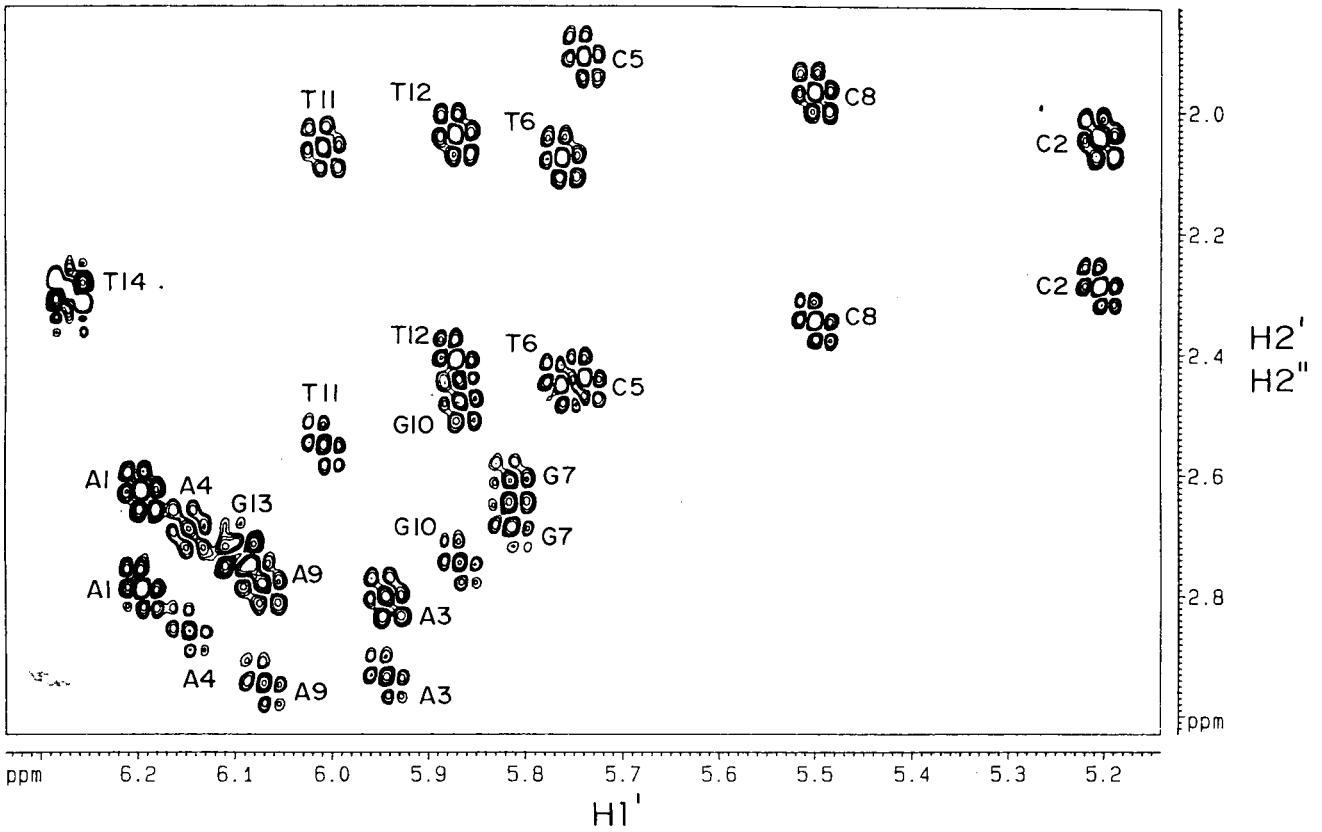

FIGURE 2: H1′−(H2′, H2″) cross-peak region of the E. COSY spectrum at 27 °C of the 14-mer DNA. The experiment was optimized to contain multiple quantum contributions up to third order. The peak assignments to specific nucleotide units are indicated, and in every case the downfield peak along the H2′/H2″ axis corresponds to the H2″ proton.

satisfactory for quantitative fitting of NOE intensities (*12*), because at different mixing times the errors in intensity estimations are different and consequently an attempt to fit quantitatively the NOE data at all mixing times to the same degree of accuracy is prone to yield unacceptable or erroneous structures. A total of 256 experimental peaks from the 200 ms NOESY spectrum were integrated for the simulations, and these included most of the base−base, base−sugar, and sugar−sugar interproton cross-peaks. For the symmetrical duplex, the peak list was duplicated to generate a final list of 512 peaks which were referenced with a cytosine H5−H6 peak (interior C5 and C8 have similar intensities).

Relaxation matrix-based simulations of the NOE intensities (*28*−*30*) were performed using the SIMNOE algorithm (*31*−*33*) as in the case of the 12-mer containing the minimal cognate site (*12*). Although this algorithm assumes a single isotropic correlation time for the whole molecule, it has proved successful in obtaining good NOE fits in many DNA segments (*12*, *34*, *35*). Several other authors have also found a single correlation time to be sufficient for obtaining good NOE fits in DNA segments longer than 10 base pairs (*29*, *36* and references therein). The isotropic correlation time required for the relaxation matrix calculations was roughly estimated from ^{1}H relaxation time measurements as described earlier (*27*), and after slight adjustments, a value of 5.5 ns was found to be suitable for obtaining good NOE fits.

(d) Derivation of Distance Constraints from NOE Intensities. Distance constraints constitute the most essential input for NMR structure calculations by different computational

Table 2: NMR Restraints and Convergence Statistics of 24 Structures

constraints	
NOE intensities	512 (256 per strand)
H−bond distances	68
maximum violations	
NOE intensities	<25%
bond lengths	<0.05 Å
bond angles	< 5°
pairwise rmsds	
backbone	0.75 + 0.5
side chain (without H)	0.78 + 0.4
R-factor	0.206 + 0.013
total energy (kcal/mol)	723.45 + 14.7

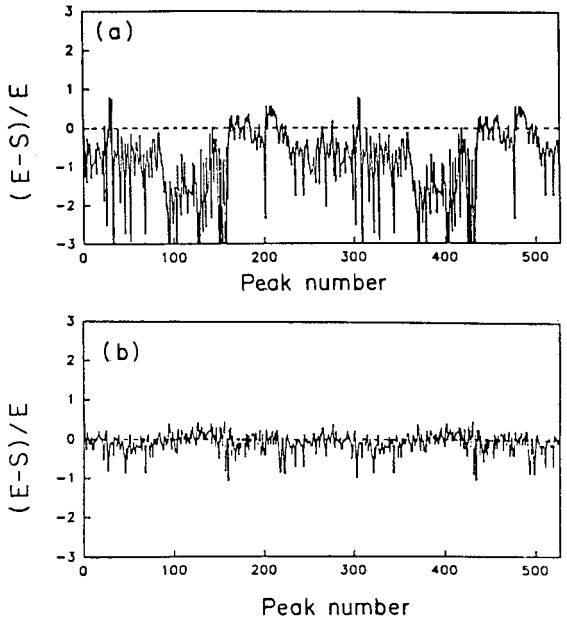

FIGURE 3: Peakwise fits of the experimental (E) and simulated (S) intensities for (a) the starting B-DNA structure and (b) the final structure having a *R*-factor of 0.19. As seen from the fits, a high degree of convergence is obtained for most of the peaks, with deviations less than 25%. The 18 peaks (9 per strand) showing greater deviations belong to the overlapping cases. The 512 peaks are ordered, first according to the atom types (e.g., H1′−H2′, H1′−H2″, H1′−H8, etc.) and then in each case according to the residue number along the sequence.

algorithms. These, as a first approximation, may be determined from the intensities of cross-peaks in a low mixing time NOESY spectrum using the so-called initial rate approximation (*28*). However, there are difficulties associated with this method, which are very well documented (*28*−*31*). Hence, we used a larger mixing time NOESY spectrum which had better intensities and larger number of peaks and coupled its analysis with relaxation matrix calculations which take care of the spin diffusion effects. The procedure involved an iterative approach in which the constraints and the structures were iteratively changed, guided by relaxation matrix calculations and restrained energy minimizations. To begin with, relaxation matrix calculations were performed for A-DNA and B-DNA duplex models to determine the fits of the individual experimental peaks. The fits were seen to be better for the B-DNA, and therefore, subsequent calculations were restricted to the structure being in the B-domain only. Depending on the fits of the 512 individual peaks in the above calculation, a constraint set of 512 distances with center points taken from the structure and bounds of ±0.2 Å for peaks with good fits and ±0.4 Å for poorer fits (deviation >25%) was generated. These constraints were then applied as square well potentials with a scale factor of 20 for restrained energy minimization (for approximately 100 steps) using the X-PLOR force field (*37*). Complete relaxation matrix calculation was again performed for this minimized structure, and again depending on the fits of the calculated and experimental intensities, a new distance constraint set was generated from the structure. This procedure was repeated iteratively till a well converged structure resulted. The restrained energy minimization included also the H-bond constraints for base pairing. Each H-bond constraint consisted of two distance constraints (H−O and N−O distances in a N−H−O bond and H−N and N−N distances in a N−H−N bond). A total of 34 H bonds lead to 68 distance constraints. These constraints were, however, not changed during the iterative procedure described above.

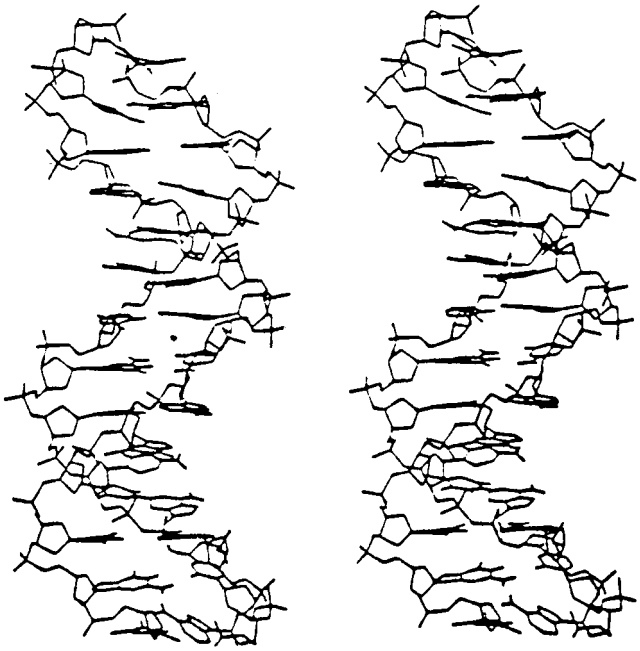

FIGURE 4: Stereoview of the final structure corresponding to the lowest energy and lowest *R*-factor of 0.19.

While the peak-to-peak intensity fits guided the refinement of the distance constraints, the overall quality of the fits was monitored by the *R*-factor defined by the following equation:

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

where *N* is the total number peaks in the list.

(e) Restarined Molecular Dynamics and Simulated Annealing. To examine the uniqueness of the structure defined by the relaxation matrix calculations, a scan through the conformational space was performed by a combined restrained molecular dynamics and simulated annealing protocol (*38*) using X-PLOR 3.0. The good-fit structure as obtained by the restrained energy minimization was heated to 1000 K, and at this temperature, molecular dynamics was

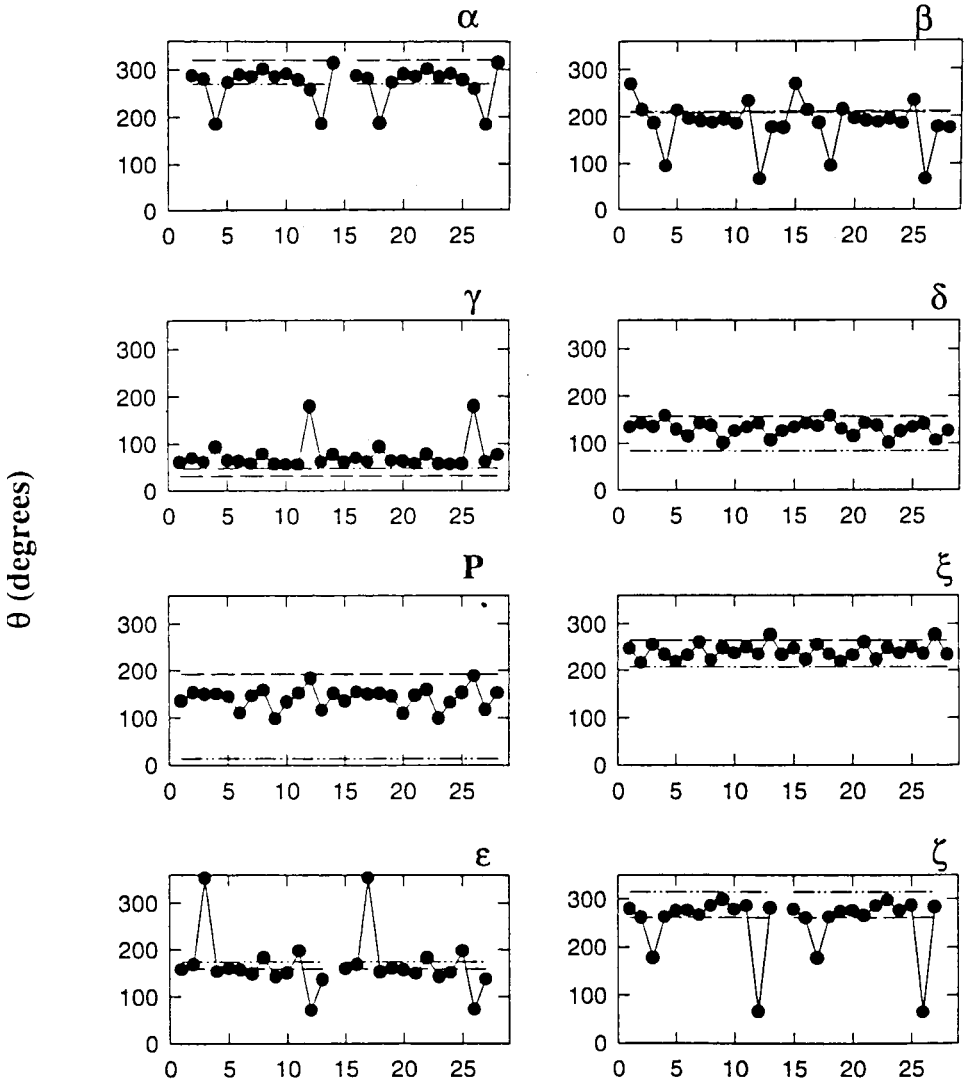

FIGURE 5: The backbone torsion angles for the final structure along with the values for the canonical B-DNA (- - -) and A-DNA (-···-) models. Equivalence among the two strands is evident from the plots. Various sequence dependent variations can be identified from these plots as discussed in the text.

run for 100 ps using 1 fs time steps, and the coordinates were written every 2 ps. All these structures at 1000 K which were very different in terms of pairwise rmsd were subjected to 200 steps of minimization, and each one of them was slowly cooled to 300 K. The cooling was done in steps of 50 K in a total of 5.5 ps for each of the 50 structures. All the 50 annealed structures were then energy minimized for 8000 steps and tested for NOE fitting and covalent geometry violations. This resulted in 24 acceptable structures with good *R*-factors (<0.21) and least deviations in covalent geometry parameters.

(f) Model Building Studies. Myb−DNA interactions were investigated by model building using INSIGHTII of Biosym Technologies, starting from the solution structure of the Myb R2R3−DNA complex reported by Ogata et al. recently (*15*). Although their study used a 16-mer DNA, the structure was well determined only for the central 11 base segment 5′-CCTAACTGACA (the minimal cognate sequence is underlined). Henceforth, we refer to their DNA as 11 mer or the Ogata DNA for our comparisons. The structure of the 11 mer in the complex was compared with the DNA studied here and also with the one studied in our laboratory earlier (*12*, *13*). Next, the protein and the DNA in the complex were separated and energy minimized separately. We observed that the protein structure did not change significantly. On the other hand, the DNA structure showed some variations in the backbone. The cognate sequence in the complex was mutated to the two sequences we have studied and the complexes were then energy minimized. We observed that the mutations caused small changes in the structure of the DNA in the complex. Subsequently, the free DNA structures were docked into the complex so as to superimpose their cognate sequences as closely as possible to the cognate sequence already present in the complex, the Ogata DNA was removed, and then the new R2R3-DNA complexes were energy minimized. The effect of base sequence in these complexes was also investigated by appropriate base mutations to match the Ogata DNA. In all these calculations, a distance-dependent dielectric constant was used with the AMBER force field. Energy minimizations were carried out using steepest descent followed by conjugate gradient methods until a predefined convergence

criterion (energy rms derivative < 0.001) was satisfied.

RESULTS AND DISCUSSION

(a) Resonance Assignments. Sequence-specific resonance assignments for the nonexchangeable protons in the 14-mer were obtained by a combination of NOESY, TOCSY, and HMQC experiments using the well-established protocols (*39*−*43*). The HMQC experiments complemented the information from the TOCSY and helped in differentiating the H2 protons from the H8/H6 protons (*44*). The distinction between the H2′ and the H2″ protons was obtained on the basis of the intensities of the H1′-H2′/H2″ cross-peaks in a low mixing time NOESY experiment (*42*). Table 1 summarizes the complete set of the proton and carbon assignments obtained for the DNA. All protons and carbons except the 5′/5″, which were extensively overlapping, were assigned.

(b) Sugar Geometries from E. COSY Spectra. Sugar geometry information for the individual nucleotide units was obtained from the analysis of peak patterns in E. COSY spectra having contributions from different multiple quantum orders (*12*, *13*, *31*, *34*). The H1′−(H2′, H2″) cross-peak region from one such E. COSY spectrum is shown in Figure 2 as an illustartion. The spectrum shows good resolution and, on the basis of simulations reported earlier (*31*), the patterns could be readilty identified as belonging to sugar geometries in the S-domain. In slightly more quantitative terms, the pseudorotation values for all but two (G13 and T14) nucleotide units were fixed in the range 90−180°. Further, we observed that the H2″−H3′ and H3′−H4′ cross-peaks were very weak, which suggested that these coupling constants were quite small in comparison to the line widths and consequently to the resolution in the spectra. On these considerations, we placed a conservative upper limit of 4 Hz for the above two coupling constants, and the pseudorotation values were further narrowed down to the range 110−160°. In the case of G13 and T14, the H2′ and H2″ protons were equivalent, but even so, a C3′-endo geometry was eliminated from the patterns of the peaks in Figure 2. This is possible because, for a C3′-endo geometry, the H1′−H2′ coupling constant is zero and the H2″−H3′ coupling constant is very large (9−10 Hz), and this results in a H1′−H2″ cross-peak which is narrow along the H1′ axis and elongated along the H2″ axis (*31*). We did not attempt any further finer restriction of the geometries because of the absence of the precise estimates of the H2″−H3′ and H3′−H4′ coupling constants. Qualitative ranges derived for the sugar geometries from the patterns of peaks in Figure 2 were considered sufficient as constraints, and any selection of the dominant geometries from NOE simulations would have to satisfy these limits. The sugar geometry constraints were, however, not explicitly included in the NOE simulations.

(c) The Best-Fit Structure and the Final NOE Fit. As discussed in the Materials and Methods, a structure giving the best NOE fit to the experimental intensities was obtained by an iterative procedure involving structure alterations and relaxation matrix calculations. The final well-fitted structure had an NOE *R*-factor of 0.19. Figure 3 shows the peak-to-peak fits of the 512 (2 × 256) experimental intensities with the corresponding simulated intensities for (a) the starting B-DNA structure and (b) the final structure. As seen for the final structure, all but 18 peaks fit within 25%, thereby

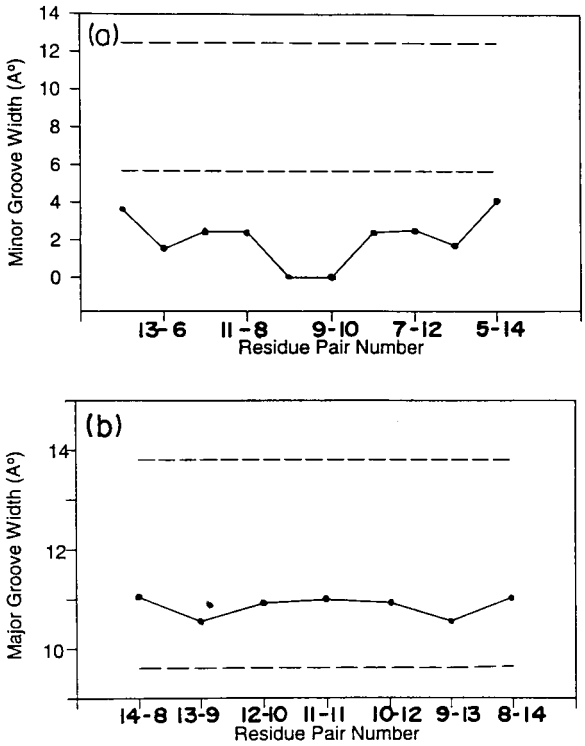

FIGURE 6: Plots of minor (a) and major (b) groove widths for the final structure. The groove widths were calculated from the phosphate−phosphate distance reduced by the sum of the van der Waals radii of two phosphate groups (∼5.6 Å) (*49*−*51*).

indicating good convergence. The deviations of 25% in the intensities would lead to a deviation of less than 5% in the distances, and it would be unreasonable to expect a better fit because of the possible errors in the experimental intensities and the assumptions in the relaxation matrix calculations. The 18 peaks for which the deviations are larger than 25% are those which are overlapping, and hence, quantification is not precise.

(d) Conformational Search by Restrained Moleculaer Dynamics and Simulated Annealing. The NOE intensity-based structure is clearly an acceptable structure for the 14-mer duplex. However, to check if this is a unique structure, it becomes necessary to systematically scan the conformational space of multidimensions and pick out structures which also satisfy the NOE data and then see how different they are from each other. Earlier, in the case of the 12-mer duplex, we had used distance geometry calculations for scanning the conformational space and had observed that the convergent structures satisfying the experimental constraints were very similar, with the pairwise rmsd's being in the range 0.2−1.2 and the individual torsional angles agreeing within 10°. In the present case, we have adopted the simulated annealing approach for the same purpose. It has been observed by several authors (*37*, *45*) that, for linear molecules such as DNA, an ab initio simulated annealing or an unrestricted distance geometry search does not lead to convergence because of insufficient long-range experimental constraints, and therefore, the conformational search has to be guided in a judicious manner. In this sense, a realistic bias will have to be introduced and within those constraints many different starting structures can be used to scan the conformational space. We had followed these guidelines earlier, and presently too, the success of the NOE based

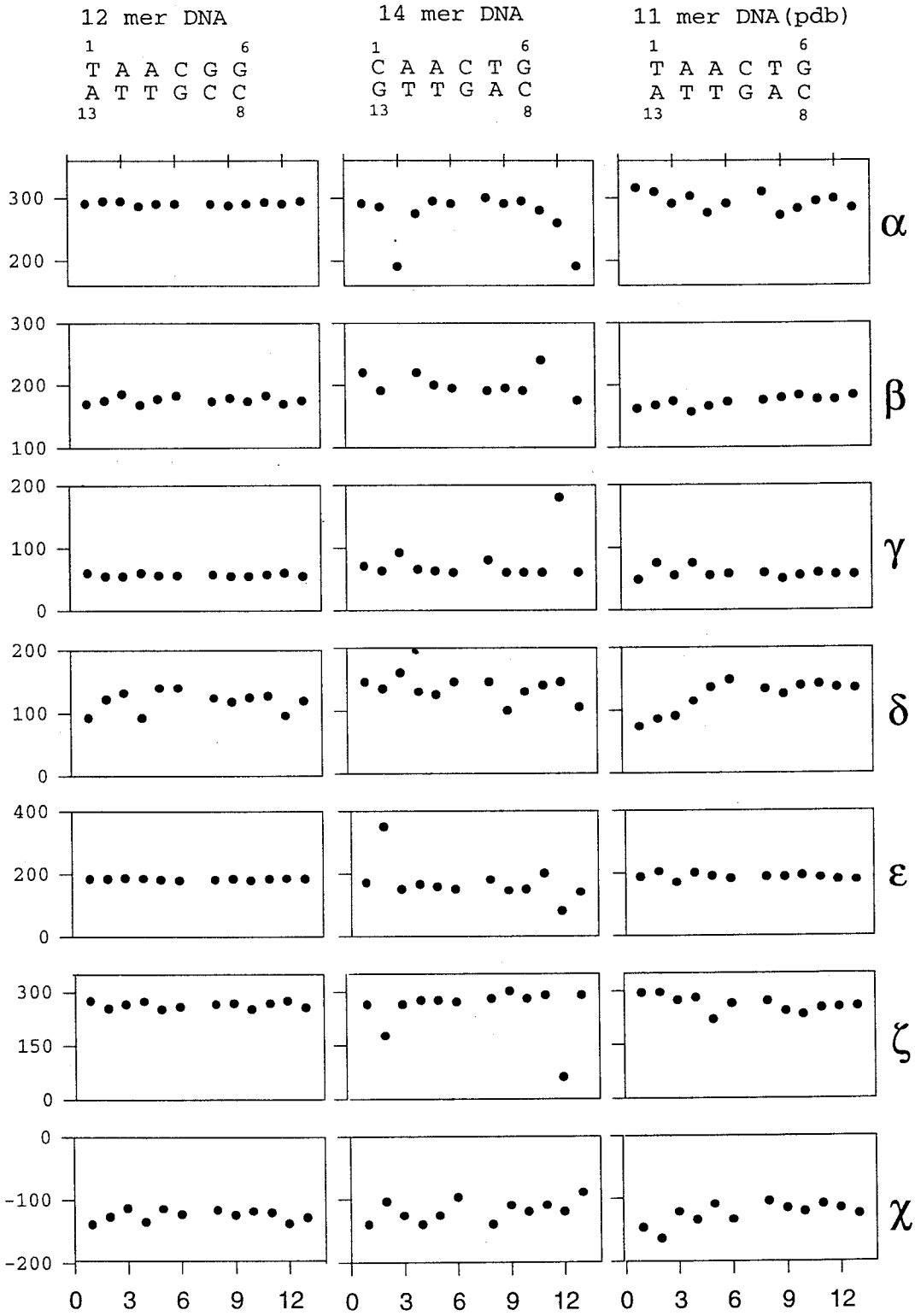

FIGURE 7: Comparison of the torsion angles of the cognate sites in the 12-mer DNA (a), the 14-mer DNA (b), and the DNA in the complex (c). Differences in many of the torsion angles can be seen.

resonance assignment strategies allowed us to restrict our search to the A- and B-DNA families (*39*−*43*). As mentioned in the Materials and Methods, for the A family, the NOE fits were very poor. On the other hand, several initial structures which had very different rmsds within the B-DNA family (note the best NOE-fitted structure belonged to this family), resulting in 24 convergent structures which had good NOE fits, did not violate the bond length and bond angle geometries and had comparable energies. Table 2 gives the convergence statistics of the derived structures in terms of rmsds, *R*-factors, and energies. It is clear that the structures are similar, showing that the data describes a single dominant structure for the molecule in solution. The torsion angle variations among the 24 structures were again within 10°.

(e) Analysis of the Lowest Energy 14-mer Structure. As the conformational search lead to very similar structures, the structure having the best *R*-factor of 0.19, which also had

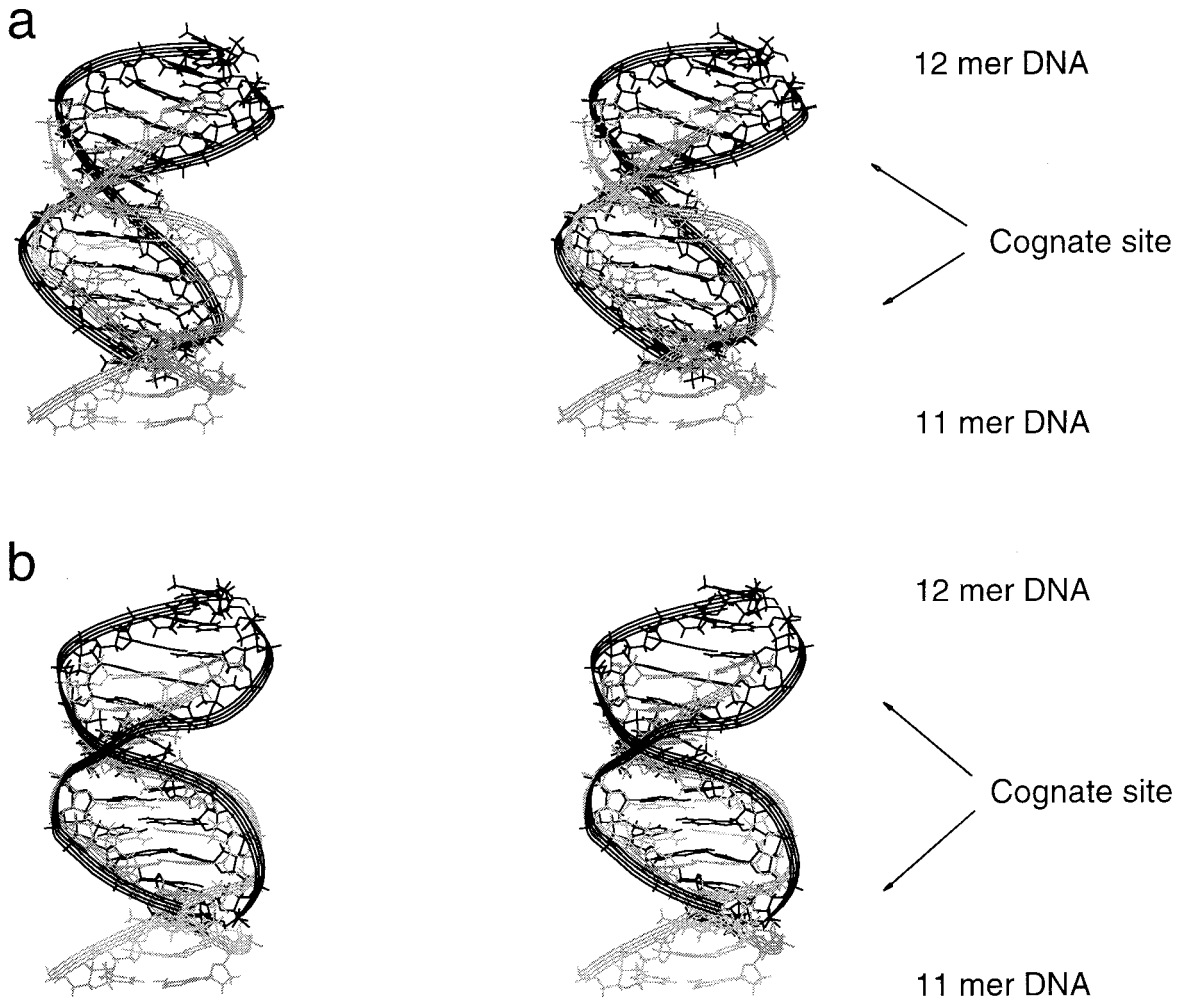

FIGURE 8: (a) Overlay of structures of cognate sites in the 12-mer free DNA and in the Ogata DNA in the complex. (b) Overlay of the cognate sites in the 12-mer−R2R3 complex after minimization and in the Ogata DNA in the complex. The two DNA sequences have been labeled. The two structures are distinguished by thick and light shaded displays. Stereoviews are shown in both cases.

the lowest energy, was analyzed in more detail to understand the sequence dependent variations. Figure 4 shows the stereoview of this final structure, and the details of the structure are described in the following paragraphs.

(1) Torsion Angles. Figure 5 shows the backbone torsion angles for the 14-mer DNA along with the values for the canonical B-DNA and A-DNA models (*46*). The torsion angles show high degree of equivalence of the two strands, as must be expected for a self-complementary DNA. It is worthwhile emphasizing here that this symmetry is a direct consequence of the NOE distance restraints, and no other restraints were added to achieve it. This is a reflection on the adequacy of the constraints (for convergence within the B-family as discussed above) and, consequently, on the quality of the structure. The figure shows that while most of the residues have backbone torsion angles around those for a B-DNA, A3 and T12 are exceptions, and interestingly, these form a part of the recognition site. It must be noted that A3 and T12 are base paired, and thus, any change in the geometry of one must reflect on the other as well. Among the phosphate backbone torsion angles, all the residues are found to be in the normal B_I conformation with ϵ in the trans range and ζ in the −gauche range. Again exceptions to this are the A3 and T12 in which the angles are in the B_{II} conformation. The sugar psuedorotation angles (P) for all the nucleotides fall around the C1′−exo conformation except for T12, which falls in the C3′−exo range. This is consistent with the results from E. COSY analysis. The glycosidic torsion angles are in the range of −anticlinical (−ac, −gauche) in all cases.

(2) Helicoidal Parameters. The intrabase pair and global helicoidal parameters (*47*), the course of the helix axis, and the course of the base-pair normals in the 14-mer DNA were calculated using the NEWHELIX program (*48*) and are shown in Figures S1−S3 of the Supporting Information.

Among the intrabase parameters (Figure S1), the propeller twist shows a well-defined pattern. The *propeller twists* for residues in the terminal and in the center are largely negative whereas those of the remaining are positive. The *buckle* values are positive for the first half of the residues and negative for the remaining. The *opening* values are mostly in the B-DNA range except for the residues 6 and 9, which show peculiarly low values.

Among the global helicoidal parameters (Figure S2) the *inclination* (η) values are negative and not zero as in a regular B-DNA. The *tip* (θ) shows large fluctuations on both sides of the B-DNA values, and the *twists* (Λ) in most cases are slightly larger than in B-DNA. Among the *displacements*, the *x* and *y* displacements are around those for B-DNA for nearly all except the terminal three residues. The *rise* values deviate quite a lot and are much larger than in a regular B-DNA. This clearly indicates that the present DNA is much

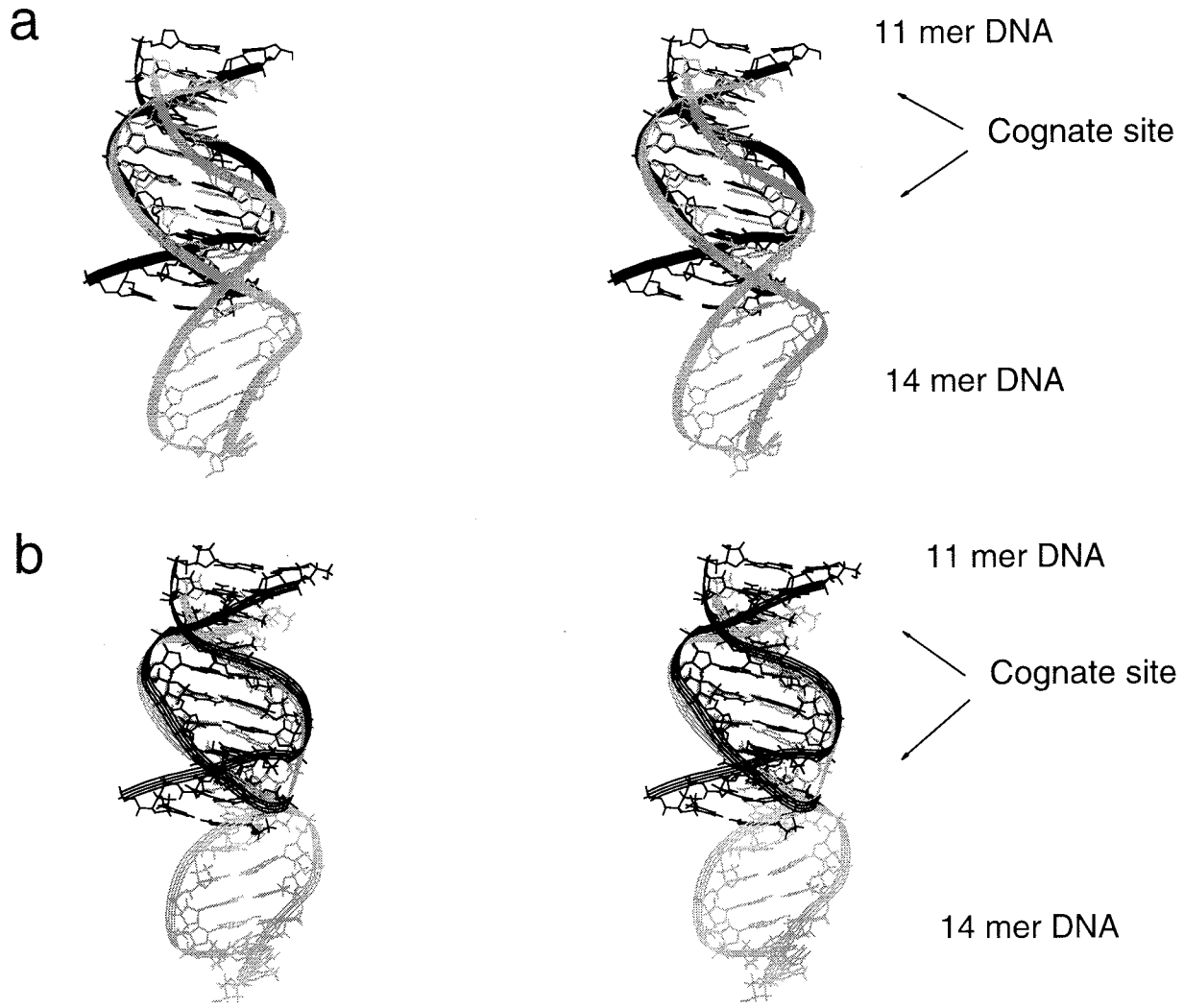

FIGURE 9: (a) Overlay of structures of cognate sites in the 14-mer free DNA and in the Ogata DNA in the complex. (b) Overlay of the cognate sites in the 14-mer−R2R3 complex after minimization and in the Ogata DNA in the complex. The two DNA sequences have been labeled. The two structures are distinguished by thick and light shaded displays. Stereoviews are shown in both cases.

longer and thinner than a B-DNA and the bases must be having weaker stacking interactions. The course of the *local helix axis* and the *base normals* relative to the first base pair are shown in Figure S3. The large deviations of the two vectors with respect to the first base pair indicate substantial displacements of the nucleotides from the *global helix axis*.

(3) Groove widths. Figure 6 shows the plot of the minor and major groove widths for the DNA (*49*−*51*). The groove widths were calculated from the phosphate−phosphate distances reduced by sum of the van der Waals radii of two phosphate groups (∼5.6 Å). It is seen that there is a substantial variation in the minor groove width along the sequence of the molecule. In contrast, the variation in the major groove width is relatively smaller.

Thus, the above results indicate that the NOE-derived structure, while being in the B-DNA family, has substantial sequence-dependent variations. This conclusion is further supported by the observation that the ^{31}P spectra exhibited good dispersion of the resonances (data not shown).

(f) Comparison of the NMR Structures of the Free and Complexed Myb Cognate DNAs. In this section, we first compare the structure of the minimal cognate sequence (CAACTG) in the 14-mer DNA with that of the cognate sequence (TAACGG) in the previously determined 12-mer DNA (*12*) and also with the TAACTG in the R2R3−DNA complex of Ogata et al. (*15*). Since the NMR data always generate an ensemble of structures, a meaningful comparison should take care of the ranges of variation of the individual parameters among the structures. In this sense, we had observed that the torsion angles varied within a range of 10° in both the 12-mer and the 14-mer molecules, suggesting that differences greater than this degree only should be interpreted as true differences. Figure 7 shows a comparison of the torsion angles spanning the cognate sequences in the minimum energy structures of the 12-mer and 14-mer DNA segments and the energy minimized average structure of the 11 mer DNA in the complex. We observe that for many torsion angles there are large differences. These differences will manifest in differences in the trace of the backbones and, thus, would have significant implications for the Myb−DNA recognitions.

(g) Structural Changes Due to Complex Formation. To investigate further the variability of DNA structure in the R2R3−DNA complex, we have carried out modeling and energy minimization studies as discussed in the Materials and Methods. The important results of these are presented below.

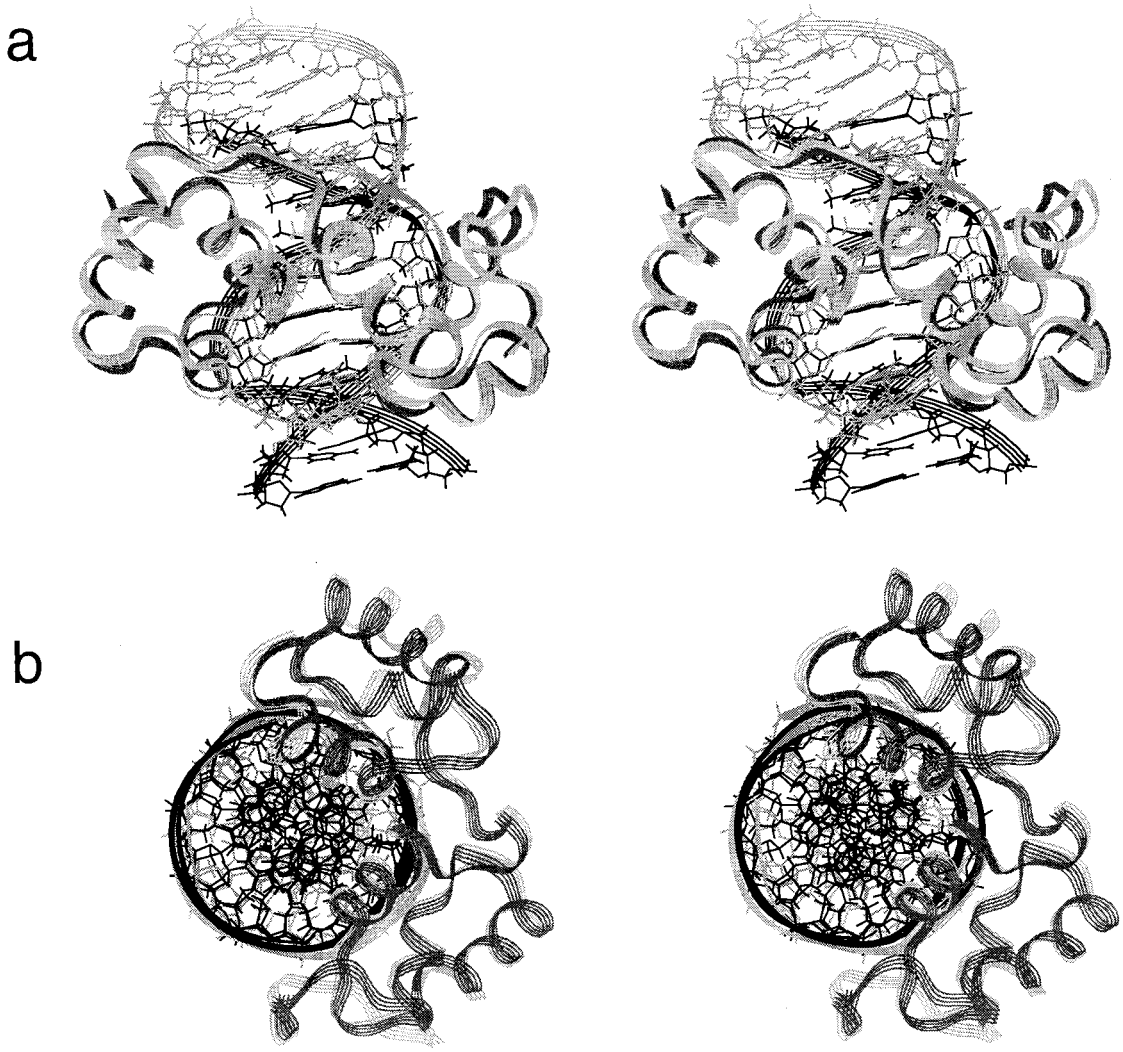

FIGURE 10: Two orthogonal stereoviews of the R2R3−DNA complexes showing overlays of R2R3−12-mer DNA and R2R3−Ogata DNA structures. The protein structure shows some differences in the two cases. The two structures are distinguished by thick and light shaded displays.

11 mer−R2R3 Complex. Specific mutations in the DNA sequence in the complex at the variable sites in the cognate sequence, followed by energy minimization of the resultant 11 mer−R2R3 complexes, caused minor changes in the backbone structure of the DNA. This indicated that the protein is fairly tolerant to these mutations.

12-mer−R2R3 Complex. Figure 8 displays an overlay of structures: (a) free 12-mer over Ogata DNA in the complex and (b) 12-mer in the complex after energy minimization over the Ogata DNA in the complex. It is evident that the 12-mer whose structure is initially very different from that of the Ogata DNA approaches the latter structure on complex formation. In either case, the overlays were optimized for the best possible superposition of the cognate sites. An analysis of the torsion angles in the cognate sequences is presented in Figure S4 of the Supporting Information. We notice that most torsion angles in the 12-mer approach the values in the Ogata DNA to within 10°. Also the two strands in the duplex are seen to behave slightly differently with regard to the extent of structural adjustments.

14-mer−R2R3 Complex. Figures 9 and S5 display comparisons similar to those in Figures 8 and S4 for the 14-mer DNA studied in this paper. Again, we see that the cognate DNA structure, which in the free form is very different from the structure of the Ogata DNA in the complex, approaches the latter on complex formation with the R2R3 protein. Here again, the torsion angles in the 14-mer approach those in the Ogata DNA to within 10°, although for some residues, the approach is not as close. Also, as in the 12-mer, the two strands in the duplex behave slightly different with regard to the extent of structural adjustments.

Considering that the cognate DNAs in the 14-mer and the 12-mer have differences within themselves, their approach toward the structure of the Ogata DNA on complex formation suggests that the protein has a significant influence on the structure of the DNA in the complex. At the same time, the cognate structures of the 12-mer and 14-mer on complex formation do possess some differences, i.e., they do not approach the Ogata DNA to the same degree of equivalence. This indicates that the sequence of the DNA in the minimal cognate site has a certain, though minor, role to play. This was further supported by the observation that when a C to T mutation [C(T)AACTG] was introduced in the 14-mer to match the cognate sequence of Ogata DNA, the energy minimized complex structure had greater resemblance to the Ogata structure. The same observation was made when a G to T [TAACG(T)G] mutation was introduced in the 12-mer.

Finally, Figure 10 displays a comparison of the R2R3−DNA complex of Ogata et al. with that obtained after replacing their DNA by the 12-mer DNA and subjecting it to energy minimization. We see that the protein largely retains its structure except for small changes in the loop portions. The same was true when the 14-mer DNA was used. This observation is supported by the NMR observation by Ogata et al. (*16, 17*) that the R3 domain had identical conformation in both the free and the DNA-bound forms.

Next, we turn our attention to the extension in the cognate sequence (*9*). While the 12-mer contains a part extension (underlined T in TAACGGT), the 14-mer and the 11 mer sequences contain the full extensions: underlined CA in CAACTGCA cognate sequence of 14-mer and underlined AC in TAACTGAC cognate sequence of 11 mer. The extensions in all the three are thus different, and yet, we observe from the modeling studies described above that the structure of the core DNA in the complex is conserved. We made base mutations in the extensions in the R2R3−DNA complex and energy minimized the resultant complexes. We did not observe significant changes in the structure of the core DNA or of the protein. Both the above observations seem to indicate that the extension does not have a major role to play in Myb recognition per se.

Our observations that base changes at the variable sites do not cause major structural changes in the DNA in the complex provide a molecular rationale for the tolerance of the protein for the cognate site variabilities. Examination of the contacts between the DNA and the protein in the NMR structure of the complex (*17*) reveals that most of the contacts are with the phosphate groups in the DNA, and there are a few with the bases directly. This is possibly the reason the backbone structure of the DNA in the cognate site is conserved and individual DNA molecules which may have very different sequence dependent structures in the free state in solution approach a specific structure in the complex.

The modeling studies based on energy minimization described above constitute a modest attempt to see induced conformational changes in the Myb cognate DNA and the Myb R2R3 protein on complex formation and reflect on the shallowness of the potential energy minima. We observed that, in the DNA, substantial conformational changes were induced on complex formation, while in the protein, the changes were rather small. Both these observations are consistent with the experimental data, namely, gel shift assays, in case of DNA, and NMR data in the case of the R3 repeat, as discussed in the introductory portion of this paper. This indicates that the protein in the complex is in a deep potential well and energy minimization cannot take it out; on the other hand, the DNA has a shallow potential surface. Thus, the protein−DNA complex approaches a nearly conserved structure. The energy minimization calculations have also provided useful insights into the R2R3−DNA interactions which dictate the structure of the complex. More detailed characterization of the induced structural changes, both in the R2R3 protein and in different cognate DNAs must await NMR analyses of more DNA−R2R3 complexes.

CONCLUSIONS

We have presented in this paper the solution structure of the extended Myb cognate site. The structure of the DNA has been determined to high precision from quantitative interpretation of 2D NOE data. Restrained molecular dynamics and simulated annealing calculations have indicated that the data define a single dominant structure for the molecule. In the dominant structure, all the sugars were seen to adopt a C1′-exo conformation. The helicoidal parameters and the helix axis indicate large variations along the sequence leading to distortions in the structure.

The structure of the cognate site in the present DNA has been compared in detail with that of the minimal cognate site in the 12-mer DNA previously determined in our laboratory (*12*) and also with the structure of the DNA in the R2R3−DNA complex determined by Ogata et al. (*15*). Many differences are seen in the structures, both in the torsion angles as well as in the helicoidal parameters. The unique kinks and rolls which were observed in the 12-mer are absent in the 14-mer. Also, the enlargement of the major groove width and the three centered hydrogen bonds present in the 12-mer are absent in the 14-mer. The structures of both the 12-mer and the 14-mer are different from the structrure of the DNA in the complex of Ogata et al. All these differences could be due to differences in the sequence, their different placements along the length of the DNA, and could also suggest protein induced structural alterations.

Our model-building studies with the three different DNA sequences and appropriate mutations indicate that R2R3−DNA complex formation induces structural changes in the DNA and that the DNA structure in the complex is nearly conserved. The sequence in the minimal site also plays a role in determining the extent to which the protein alters the structure. This is in accordance with the previous experimental observations where a bend in the DNA was predicted on Myb binding by gel shift assay techniques (*18*). The extension in the cognate sequence, however, does not seem to have a major influence on the Myb−DNA interaction. The modeling studies also indicate that the R2R3 conformation is largely unaltered on complex formation, which is in accord with the NMR observation by Ogata et al. (*16*, *17*) that the free R3 structure was similar to the structure of the R3 repeat in the R2R3-DNA complex.

ACKNOWLEDGMENT

We thank the National Facility for high-field NMR and TIFR and Mumbai for experimental and computational facilities. We thank Dr. M. V. Hosur and K. K. Kannan of Bhabha Atomic Research Center for the X-PLOR usage.

SUPPORTING INFORMATION AVAILABLE

Five figures (S1−S5) as described in the text (6 pages). Ordering information is given on any current masthead page.

REFERENCES

1. Gewirtz, A. M., and Calabretta, B. (1988) *Science 242*, 1303−1306.
2. Luscher, B., and Eisenmann, R. N. (1990) *Gene Dev. 4*, 2235−2241.
3. Thompson, M. A., and Ramsay, R. G. (1995) *BioEssays 17*, 341−350.
4. Biedenkapp, H., Borgemeyer, U., Sippel, A. E., and Klempnauer, K. H. (1988) *Nature 335*, 835−837.
5. Weston, K. M. (1990) *Semin. Cancer Biol. 1*, 371−382.

6. Seng-Ong, G. L. (1990) *Biochim. Biophys. Acta 1032*, 39−52.
7. Howe, K. M., Reakes, C. F. L., and Watson, R. J. (1990) *EMBO J. 9*, 161−169.
8. Oehler, T., Arnold, H., Biedenkapp, H., and Klempnauer, K. H. (1990) *Nucleic Acids Res. 18*, 1703−1710.
9. Ording, E., Kvavik, W., Bostad, A., and Gabrielson, O. S. (1994) *Eur. J. Biochem. 222*, 113−120.
10. Tanikawa, J., Yasukawa, T., Enari, M., Ogata, K., and Nishimura, Y. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 9320−9324.
11. Madan, A., Radha, P. K., Hosur, R. V., and Padhy, L. C. (1995) *Eur. J. Biochem. 232*, 150−158.
12. Radha, P. K., Madan, A., Nibedita, R., and Hosur, R. V. (1995) *Biochemistry 34*, 5913−5922.
13. Radha, P. K., Nibedita, R., Ajay Kumar R., and Hosur, R. V. (1995) *Methods Enzymol. 261*, 73−89.
14. Madan, A., Hosur, R. V., and Padhy, L. C. (1994) *Biochemistry 33*, 7120−7126.
15. Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., and Nishimura, Y. (1994) *Cell 79*, 639−648.
16. Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S., and Nishimura, Y. (1992) *Proc. Natl. Acad. Sci. U.S.A. 89*, 6428−6432.
17. Ogata, K., Morikawa, S., Nakamura, H., Hojo, H., Yoshimura, S., Zhang, R., Aimoto, S., Ametani, Y., Hirata, Z., Sarai, A., Ishii, S., and Nishimura, Y. (1996) *Nature Struct. Biol. 2*, 309−319.
18. Saikumar, P., Gabriel, J. L., and Reddy, E. P. (1994) *Oncogene 9*, 1279−1287.
19. McBride, L. J., and Caruthers, M. J. (1983) *Tetrahedron Lett. 24*, 245−248.
20. Gait, M. J. (1984) *Oligonucleotide Synthesis—a practical approach*, IRL Press Ltd., Oxford.
21. Griesinger, C., Otting, G., Wuthrich, K., and Ernst, R. R. (1988) *J. Am. Chem. Soc. 110*, 7870−7873.
22. Jeener, J., Meier, B. H., Bachmann, P., and Ernst, R. R. (1979) *J. Chem. Phys. 71*, 4546−4554.
23. Kumar, A., Ernst, R. R., and Wuthrich, K. (1980) *Biochem. Biophys. Res. Commun. 95*, 1−6.
24. Redfield, A. G., and Kunz, S. D. (1975) *J. Magn. Reson. 19*, 250−254.
25. Muller, L. (1979) *J. Am. Chem. Soc. 101*, 4481−4484.
26. Griesinger, C., Sorensen, O. W., and Ernst, R. R. (1987) *J. Magn. Reson. 75*, 474−492.
27. Woessner, D. E. (1962) *J. Chem. Phys. 36*, 1−4.
28. Macura, S., and Ernst, R. R. (1980) *Mol. Phys.* 41, 95−117.
29. Borgias, B. A., Gochin, M., Kerwood, D. J., and James, T. L. (1990) *Prog. NMR Spectrosc. 22*, 83−100.
30. Landy, S. B., and Rao, B. D. N. (1988) *J. Magn. Reson. 83*, 29−43.
31. Majumdar, A., and Hosur, R. V. (1992) *Prog. NMR Spectrosc. 24*, 109−158.
32. Nibedita, R.,. Ajay Kumar, R.; Majumdar, A., and Hosur, R. V. (1992) *J. Biomol. NMR 2*, 467−476.
33. Nibedita, R., Ajay Kumar, R., Majumdar, A., and Hosur, R. V. (1992) *J. Biomol. NMR 2*, 477−484.
34. Nibedita, R., Ajaykumar, R., Majumdar, A., Hosur, R. V., Govil, G., Mazumder, K., and Chauhan, V. S. (1993) *Biochemistry 32*, 9053−9064.
35. Singh, S., Patel, P. K., and Hosur, R. V. (1997) *Biochemistry 36*, 13214−13222.
36. Reid, B. R., Banks, K., Flynn, P., and Nerdal, W. (1989) *Biochemistry 28*, 10001−10007.
37. Brunger, A. T. (1992) *X-PLOR (Version 3.0) Manual*, Fellows of Harvard University, New York.
38. Nilges, M., Clore, G. M., and Gronenborn, A. M.(1988) *FEBS Lett. 229*, 317−324.
39. Scheek, R. M., Russo, N., Boelens, R., Kaptein, R., and van Boom, J. H. (1983) *J. Am. Chem. Soc. 105*, 2914−2916.
40. Hare, D. R., Wemmer, D. E., Chou, S. H., Drobny, G. R., and Reid, B. R. (1983) *J. Mol. Biol.* 171, 319−336.
41. Ravikumar, M., Hosur, R. V., Roy, K. B., Miles, H. T., and Govil, G. (1985) *Biochemistry 24*, 7703−7711.
42. Hosur, R. V., Govil, G., and Miles, H. T. (1988) *Magn. Reson. Chem. 26*, 927−944.
43. Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York.
44. Radha P. K. (1996) *Magn. Reson. Chem. 34*, S18 − S22.
45. Metzler, W. J., Hare, D. R. and Pardi, A. (1989) *Bichemistry 28*, 7045−7052.
46. Saenger, W. (1984) *Principles in Nucleic Acid Structure*, Springer-Verlag, New York.
47. Dickerson, R. E. (1989) *J. Biomol. Struct. Dyn. 6*, 627−634.
48. Prive, G. G., Yanagi, K., and Dickerson, R. E. (1991) *J. Mol. Biol. 217*, 177−199.
49. Narayana, N., Ginell, S. L., Russu, I. M., and Berman, H. M. (1991) *Biochemistry 30*, 4449−4455.
50. Bhattacharyya, D., and Bansal, M. (1992) *J. Biomol. Struct. Dyn. 10*, 213−226.
51. Mujeeb, A., Kerwin, S. M., Kenyon, G. L., and James, T. L. (1992) *Biochemistry 32*, 13419−13431.

BI9806753