FEBS Letters 347 (1994) 99–103

FEBS 14130

Human chromosomal centromere $(AATGG)_n$ sequence forms stable structures with unusual base pairs

T.N. Jaishree, Andrew H.-J. Wang*

Biophysics Division and Department of Cell & Structural Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received 1 April 1994; revised version received 5 May 1994

Abstract

Nine DNA sequences related to the purine strand of the human centromeric satellite $(AATGG)_n \cdot (CCATT)_n$ repeat have been studied by two-dimensional nuclear magnetic resonance spectroscopy. Earlier studies have suggested that the structure of $(AATGG)_n$ sequence has an equilibrium between the duplex form and a fold-back form. Structural refinement of d(CAATGG) and its related sequences by an NOE-constrained simulated annealing procedure reveals that the duplex form incorporates dynamic type-I G-A base pairs. 1D exchangeable proton NMR data support this model. The reverse sequence motif (GGTAA) destabilizes the structure.

Key words: Centromere sequence; DNA structure; NMR; Unusual base pairs

1. Introduction

Centromeric DNA sequences attach any DNA that contain them to the mitotic spindle during the M phase of the cell cycle [1]. Their role in maintaining the structural integrity of chromosomes is evident. Human centromeric satellite DNA contains a highly conserved repeating sequence $(AATGG)_n \cdot (CCATT)_n$ [2]. This repeat is also found on chromosomes that show high incidence of nondysjunction and other mitotic/meiotic abnormalities [3]. Grady et al. [2] have shown recently that the purine strand of this sequence, $(AATGN)_n$, forms a stable structure which likely incoporates G–A mismatched base pairs. In addition, this sequence behaved anomalously in gel electrophoresis. The $(AATGG)_n$ molecule migrated between the pyrimidine strand $(CCATT)_n$ and the perfect duplex form of $(AATGG)_n \cdot (CCATT)_n$, suggesting a fold-back or multistranded structure (Fig. 1).

A number of recent studies have revealed that simple short repeats of DNA sequences, like $(C–G)_n$ [4], $(C–A)_n$ [5], the telomere repeat, $(TTGGGG)_n$ [6,7], $(CGA)_n$ [8,9] and d(TCCCCC) [10], often are associated with unusual DNA structures that may play important biological roles. Here we analyzed in detail the structures associated with the $(AATGG)_n$ sequence by two-dimensional nuclear magnetic resonance spectroscopy (2D NMR) through a quantitative treatment of the nuclear Overhauser effect (NOE) data. We showed that the $(AATGG)_n$ sequence forms a duplex structure with dynamic G–A base pairs.

2. Materials and methods

The oligonucleotides were synthesized by an Applied Biosystems DNA synthesizer. The deblocked oligonucleotides were purified by Sepharose G-50 column chromatography. Solutions of the DNA oligomers (~ 5 mM duplex) with 150 mM NaCl and 50 mM phosphate buffer (pH 7.0) were prepared in 0.5 ml of D_2O as described earlier [11,12]. NMR spectra were collected on either a GE GN500 or a Varian VXR500 500 MHz spectrometer and the data were processed with FELIX v1.1 [13]. The 2D NOE spectra were collected at 5 °C with a mixing time of 200 ms and a total recycle delay of 4 s. The data were collected by the States technique [14] with 512 t_1 increments and 2048 t_2 complex points, each the average of 16 transients. Apodization of the data in the t_1 and t_2 dimensions consisted of 5 Hz exponential multiplication with one half of a sine-squared function for the last one fourth of the data to reduce truncation artifacts. Integrals from the 2D NOE data set were extracted by evaluation of the known shapes of each spin in the ω_1 and ω_2 dimensions. These shapes were determined by spectral analysis with the program MYLOR. This procedure allowed us to obtain 1088 NOE integrals involving all the non-exchangeable protons for the hexamer. Refinement of the starting model was carried out by the sequence of procedures comprising the SPEDREF package [11]. This included a complete relaxation matrix calculation of the NOEs for the model, with comparison of the experimental and simulated spectra to deconvolute overlapped areas of the spectra. A linear regression of all off-diagonal volume elements was used to generate a single scalar for comparison of the two sets of volume integrals. The target distances derived after the deconvolution and comparison of the two spectra were used to obtain potential-energy springs. The NOE energy term (kcal/mol) was calculated using X-PLOR's biharmonic potential option using a scale factor of 30 [15]. With these energy springs as inputs, we then used conjugate gradient energy minimization to obtain a new model which was taken through the same procedure iteratively until a convergence between experimental and simulated spectra was achieved. The isotropic rotational correlational time used for the calculations was 4.5 ns, obtained empirically by comparison between the theoretical and experimental spectra. The initial models were constructed using X-PLOR [15] and QUANTA [16]. The simulated annealing refinement was performed using X-PLOR by non-equilibrium heating of the starting model to 1000 K and running a molecular dynamics simulation by the Verlet algorithm for 100 ps at this temperature. Coordinates were written every 10 ps along the trajectory and the molecules were slowly cooled by 1 K every 10 fs to a final temperature of 300 K with the above described NOE constraints. Ten models were thus obtained. The NMR *R*-factors (defined as $R(NMR) = \Sigma(|N_o - N_c|)/\Sigma N_o$ where N_o and N_c are observed and calculated NOE volumes, respectively) for these models

*Corresponding author. Fax: (1) (217) 244 3181.

SSDI 0014-5793(94)00516-X

were calculated using the procedure SPEDREF [11]. Different starting models including hexamer duplexes with type I and type II base pairs in them, were tested and they resulted in comparable models (rmsd ~ 0.6 Å and *R*-factor ~ 18%). The exchangeable proton 1D spectra were collected using the selective excitation 1$\bar{3}$3$\bar{1}$ pulses to excite resonances ranging from 9 ppm to 17 ppm as described by Hore et al. [17].

3. Results and discussion

Nine DNA molecules (CAATGG, CCATGG, CAATIG, C[c^{7}A]ATGG, GCAATGGC, TGCAATGGCA, ATGCAATGGAAT, AGGTAACG and AAGGTAACGT) have been subjected to 2D NOESY analyses. The 2D NOESY spectra of d(CAATGG) exhibited characteristic patterns of a helical structure. As can be seen from Fig. 2A, there is no interruption of the connectivity for sequential assignment and every aromatic proton (H^6 and H^8) has crosspeaks to the H$^{2'}$/H$^{2''}$ of its own sugar and the sugar on the 3′ side, confirming the helical nature of the backbone. To have a more definitive view of the possible structure, several models (including blunt-ended 6-bp duplexes with type-I or type-II G–A mismatches, or a 5-bp duplex with a 3′-overhang G, all with an overall B or A conformation) were constructed and subjected to NOE-constrained refinement by the procedure SPEDREF [11] and a simulated annealing (SA) procedure using X-PLOR [12,15]. Our results indicated that the most plausibe model is a 6-bp B-DNA duplex with type-I G–A base pairs. An ensemble of 10 structures, obtained by the SA procedure is shown in Fig. 3. The refined models (R-factors are in the vicinity of 18%) produce simulated NOE spectra (Fig. 2A, lower panels) substantially similar to the experimental NOE spectra. All of the models adopt an overall B-DNA conformation with all residues showing an *anti* glycosidyl angle.

To verify that type-I G–A base pair is used in the CAATGG sequence, we synthesized two hexamers C[c^{7}A]ATGG and CAATIG, where c^{7}A is 7-deaza-A and I is inosine, and analyzed their structures by NMR. Both sequences formed well-defined structures as evident from their 2D NOESY spectra which showed extensive NOE crosspeaks (data not shown). Their structures were also refined by a combined SPEDREF and SA (X-PLOR) procedure. In these two mclecules, only type-I G–A mismatch is possible, since type-II G–A mismatch requires the N^2 of G and N^7 of A for hydrogen bonds.

The type-I G–A base pair is unequivocally demonstrated in the $C_1A_2A_3T_4I_5G_6$ molecule. Fig. 2B shows a 1D slice through the I5H^2 resonance in the 2D-NOESY spectrum. It can be seen that strong *inter*-strand NOE crosspeaks are detected between I5H^2 and A2H^2, A3H^2, A2H$^{1'}$ plus A3H$^{1'}$. These NOE crosspeaks can only occur if the molecule forms a duplex with type-I G–A base pairs (Fig. 1).

Close inspection of the structures derived from the NOE-constrained simulated annealing refinement of CAATGG reveals an interesting structural feature. The size of type-I G–A mismatched base pair is larger than a normal Watson–Crick base pair. When it is embedded among normal Watson–Crick base pairs, structural distorion of the helix is expected, as shown in the crystal structures [18] and such is the case. In the case of CAATGG, the structures appeared to cluster in two groups. In one group of structures, the G5–A2 base pair is highly propellar-twisted (~37°), as in the crystal structure of CCAAGATTGG [18]. Yet in another group of structures, the G5 base shifts to pair with the A3 base. In other words, the guanine base in the (AATGN) motif is either paired with an opposing A, or is paired with the other A on the 3′-side of the opposite A. Such unusually shifted base pairs have been observed in the crystal structures of DNA containing (C–A)$_n$ sequence [19], or in the 434 repressor–DNA complex [20].

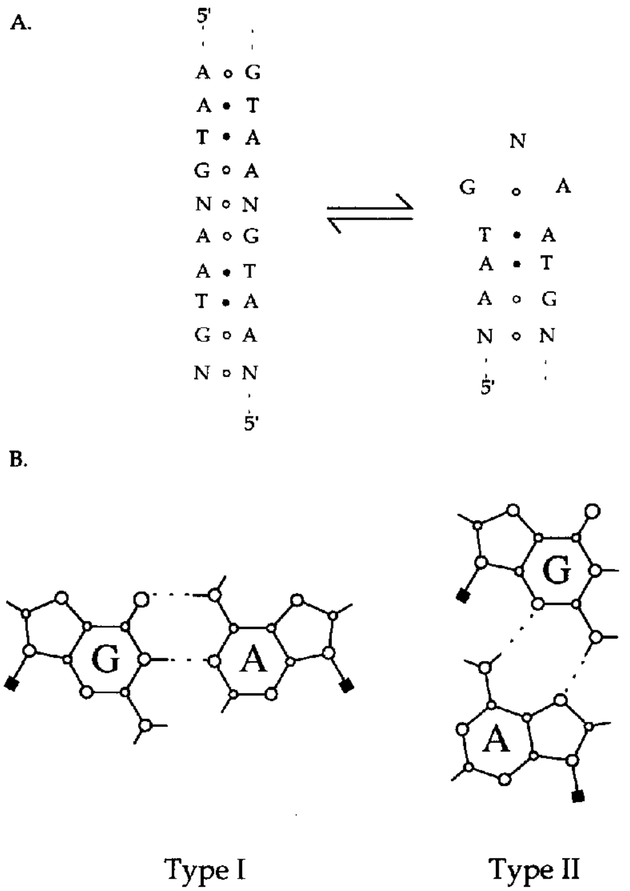

Fig. 1. (A) Schematic diagram showing the proposed equilibrium between the duplex form and the fold-back form of the centromere consensus sequence (AATGN). Solid and open circles denote normal Watson–Crick and mismatched base pairs repectively. (B) The two types of G–A mismatched base pairs are shown.

In the first group, the central two A–T base pairs have high propellar twist and involve weak hydrogen bonds while in the second group, the T imino protons are not involved in hydrogen bonds. The terminal G–C base

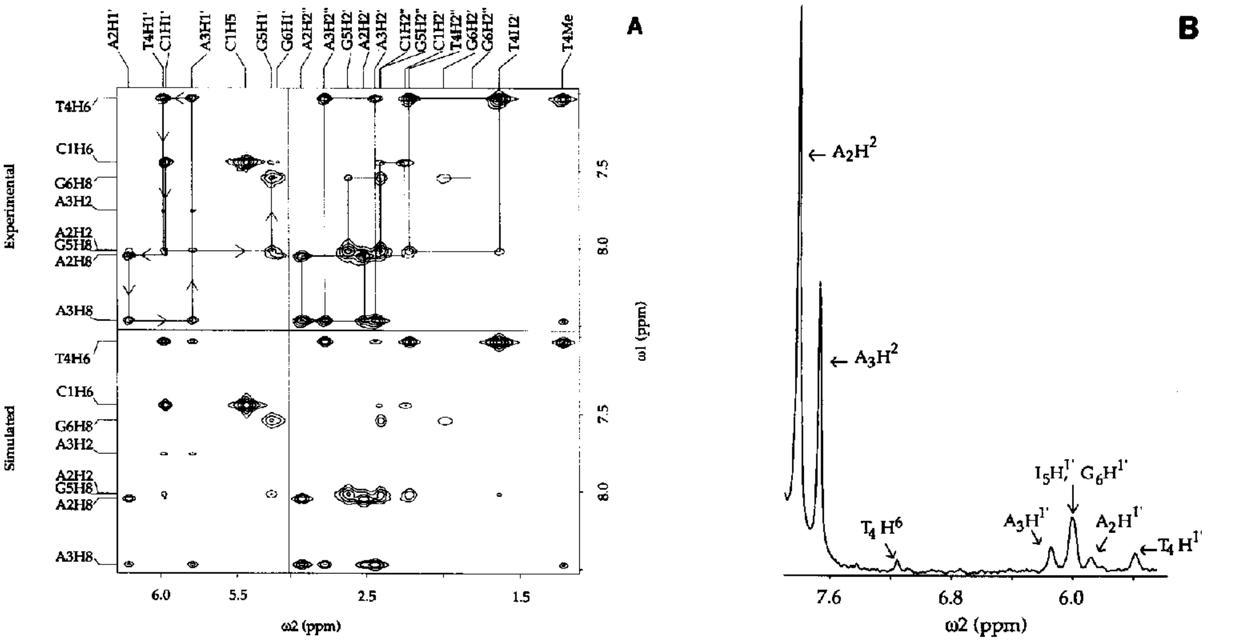

Fig. 2. (A) Portions of the non-exchangeable proton phase-sensitive 2D NOESY spectra of the CAATGG duplex which provide key structural information including the glycosyl conformation, sugar puckers and base-base stack. The top panels are the experimental NOESY cross-peaks between the aromatic (6.9–8.6 ppm) to $H^{1'}/H^5$ protons (5.2–6.2 ppm). The sequential assignment pathway is shown with arrowed lines (left) and the characteristic NOEs between the H^6/H^8 protons and the $H^{2'}/H^{2''}$ protons due to a helical backbone are marked (right). The lower panels are the simulated NOE spectra of the same regions based on the refined model (*R*-factor = 18%). (B) A 1D slice through the $I5H^2$ resonance in the 2D NOESY spectrum of $C_1A_2A_3T_4I_5G_6$ showing interstrand crosspeaks that indicate type-I GA base pairing (see text).

pairs are fraying. Such mobility in base pairs may influence the exchange rate of the imino protons. The 1D exchangeable proton NMR spectra (Fig. 4) were recorded for six molecules. The B-DNA, CCATGG has three well-resolved imino proton signals with the corresponding crosspeaks seen in the 2D NOESY spectrum in H_2O (data not shown) as expected. However, CAATGG has only two broad resonances. The resonance at 13.3 ppm is likely from T-imino of the central A–T base pair, whereas the resonance at 11.9 ppm is likely to be partly from G-imino of the G–A base pair. In 1 M NaCl, this broad resonance splits into two peaks, one of which may be from the fraying G6 residue. The chemical shift and line width of the T imino proton resonance (as compared with that in the control B-DNA sequence) agrees with our contention of two families of

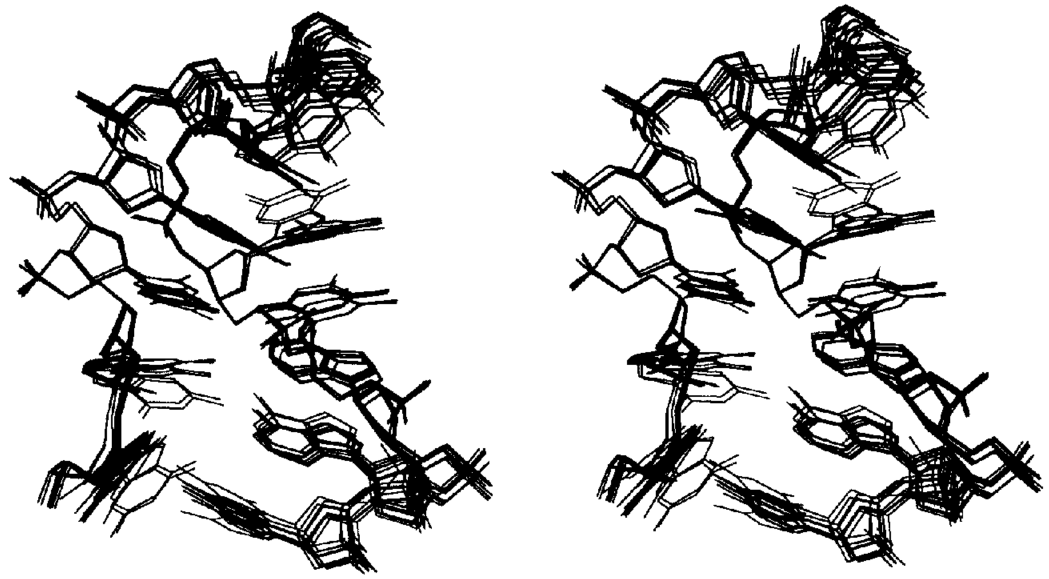

Fig. 3. An ensemble of 10 structures of CAATGG from the simulated annealing procedure described in text. They appear to fall into two families of structures which show the G5 residue base pairing with either the A2 residue or the A3 residue of the opposite strand.

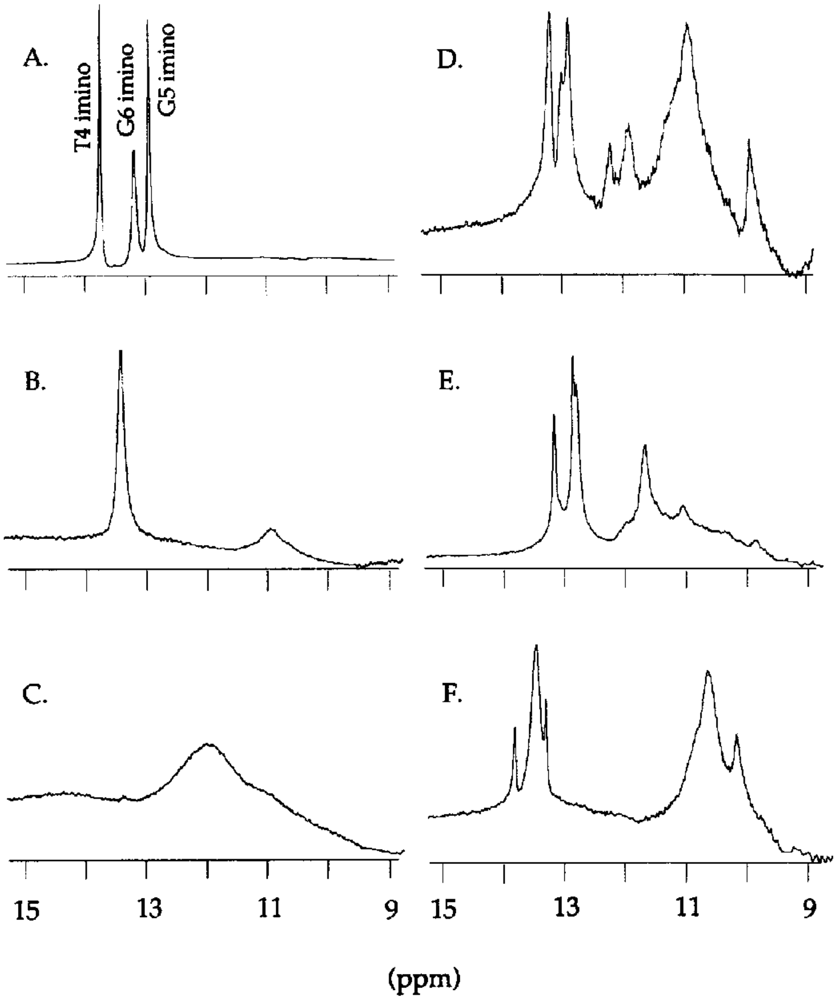

Fig. 4. 1D exchangeable proton NMR spectra of the imino proton region of (A) CCATGG, (B) CAATGG, (C) CAATIG, (D) GCAATGGC, (E) TGCAATGGCA and (F) ATGCAATGGAAT in 150 mM NaCl. In 1 M NaCl, the broad resonance at 11.9 ppm for CAATGG splits into two peaks.

structures existing in equilibrium, with the T imino protons weakly hydrogen bonded with the adenine residues part of the time. The chemical shift and line width of the G imino proton resonance from the G–A base pair are consistent with the dynamic model that we propose. The dynamic exchange may cause the broad resonances which were especially evident in case of CAATIG. Carbonnaux et al. [21] have also observed a similar chemical shift for the G imino proton involved in an unusual G-A base pair. Interestingly, CAATIG has only a very broad resonance centered at 12 ppm. This data is consistent with the NOE-constrained SA refinement of CAATIG which indicates more dynamic A–T base pairs. This suggests that the N^2 amino group of G5 residue is important structurally. In CAATGG, the ability of the G5 residue to base pair with A2 or A3 using its N^2 amino group as a donor may keep the A and T residues more organized.

The dynamic nature of the $(AATGG)_n$ motif is reflected in the fact that the longer molecules with this motif have multiple conformers. We have tested GCAATGGC, TGCAATGGCA and ATGCAATGGAAT and found that their 2D NOESY spectra have crosspeak patterns characteristic of multiple conformers. It is interesting to note that only molecules with correct (AATGN) repeats have the broad A–T imino resonance at 13.3 ppm. This may be related to the desirable property of (AATGG) sequence in switching back and forth between the duplex form and the fold-back form, due to similar energy levels of these two forms. It is important to point out that the stability of the $(AATGG)_n$ sequence is critically dependent on the sequence arrangement. The molecules AGGTAACG and AAGGTAACGT contain exactly the opposite 5′-GGTAA sequence and their 2D NOESY spectra did not show any NOE crosspeaks associated with well-formed structures. Model building studies suggest that the G in the G–A mismatch of GGTAA may not be able to readily pair with the residue on the 5′ side of the opposing A in these cases. This might destabilize the duplex. Therefore the sequence context in which the G–A mismatched base pair exists, may dictate the stability of the particular sequence.

At present we do not have information on the fold-back form of the centromere sequence. Catasti et al. [22] have suggested that a fold back form with a GGA loop exists with type II G–A base pairing in the loop and the stem. They also saw broad imino proton resonances involving the A–T and G–A base pairs at similar chemical shifts. Heus and Pardi have shown that an RNA (GANA) tetra loop used a type-II G–A base pair to stabilize the loop structure [23].

In conclusion, centromere sequences play a very important biological role during cell division. These sequences are known to bind specific proteins which initiate the formation of a multiprotein complex (kinetochore) that binds to the ends of microtubules and helps in chromosomal migration during cell division [1]. Grady et al. [2] have observed some proteins that bind specifically to the human centromeric repeat that we have studied. The dynamic base pairing scheme proposed above may play an important role in the specific protein recognition of this DNA repeat.

Acknowledgements: This work was supported by NIH Grants GM-41612 and CA-52506 to A.H.-J.W. The Varian VXR500 spectrometer was supported by NIH shared Instrumentation Grant 1S10RR06243.

References

[1] Price, C.M. (1992) Curr. Opin. Cell. Biol. 4, 379–384

[2] Grady, D.L., Ratliff, R.L., Robinson, D.L., McCanlies, E., Meyene, J. and Moyzis, R.K. (1992) Proc. Natl. Acad. Sci. USA 89, 1695–1699.

[3] Bove, A., Bove, J. and Gropp, A. (1984) Adv. Hum. Genet. 14, 1–57 .

[4] Wang, A.H.-J., Quigley, G.J., Koplak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G.A. and Rich, A. (1979) Nature 282, 680–686.

[5] Kladdle, M.P., D'Cunha, J. and Gorski, J. (1993) J. Mol. Biol. 229, 344–367.

[6] Kang, C.-H., Zhang, X., Ratliff, R., Moyzis, R. and Rich, A. (1992) Nature 356, 126–131.

[7] Smith, F.W. and Feigon, J. (1992) Nature 356, 164–168.

[8] Robinson, H. and Wang, A.H.-J. (1993) Proc. Natl. Acad. Sci. USA 90, 5224–5228.

[9] Robinson, H., van Boom J.H. and Wang, A.H.-J. (1994) J Am. Chem. Soc. 116, 1565–1566.

[10] Gehring, K., Leroy, J.-L. and Gueron, M. (1993) Nature 363, 561–565.

[11] Robinson, H. and Wang, A.H.-J. (1992) Biochemistry 31, 3524–3533.

[12] Jaishree, T.N., van der Marel, G.A., van Boom, J.H. and Wang, A.H.-J. (1993) Biochemistry 32, 4903–4911.

[13] FELIX, Version 1.1, (1992) Hare Research, Woodinville, WA.

[14] States, D.J., Haberkorn, R.A. and Ruben, D.J. (1982) J. Magn. Reson. 48, 286–292.

[15] Brunger, A.T. (1992) X-PLOR, Version 3.0, The Howard Hughes Medical Institute and Yale University, New Haven, CT.

[16] QUANTA, Version 3.2.3, (1991) Polygen Corporation, 200 Fifth Avenue, Waltham, MA 02254.

[17] Hore, P.J. (1983) J. Magn. Reson. 55, 283–300.

[18] Prive, G.G., Heinemann, U., Chandrasegaran, S., Kan, L.-S., Kopka, M.L. and Dickerson, R.E. (1987) Science 238, 498–504.

[19] Timsit, Y., Vilbois, E. and Moras, D. (1991) Nature 354, 167–170.

[20] Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M. and Harrison, S.C. (1988) Science 242, 899–907.

[21] Carbonnaux, C., van der Marel, G.A., van Boom, J.H., Guschlbauer, W. and Fazakerley, G.V. (1991) Biochemistry 30, 5449–5458.

[22] Catasti, P., Gupta, G., Garcia, A.E., Ratliff, R., Hong, L., Yau, P., Moyzis, R.K. and Bradbury, M.E. (1994) Biochemistry 33, 3819–3830.

[23] Heus, H.A. and Pardi, A. (1991) Science 253, 191–194 .