The EMBO Journal vol.6 no.6 pp.1833–1842, 1987

The polypeptide fold of the globular domain of histone H5 in solution. A study using nuclear magnetic resonance, distance geometry and restrained molecular dynamics

G.Marius Clore, Angela M.Gronenborn, Michael Nilges, Dinesh K.Sukumaran and Jutta Zarbock

Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG

Communicated by R.Huber

The polypeptide fold of the 79-residue globular domain of chicken histone H5 (GH5) in solution has been determined by the combined use of distance geometry and restrained molecular dynamics calculations. The structure determination is based on 307 approximate interproton distance restraints derived from n.m.r. measurements. The structure is composed of a core made up of residues 3–18, 23–34, 37–60 and 71–79, and two loops comprising residues 19–22 and 61–70. The structure of the core is well defined with an average backbone atomic r.m.s. difference of 2.3 ± 0.3 Å between the final eight converged restrained dynamics structures and the mean structure obtained by averaging their coordinates best fitted to the core residues. The two loops are also well defined locally but their orientation with respect to the core could not be determined as no long range ($|i-j| >$ 5) proton–proton contacts could be observed between the loop and core residues in the two-dimensional nuclear Overhauser enhancement spectra. The structure of the core is dominated by three helices and has a similar fold to the C-terminal DNA binding domain of the cAMP receptor protein.

Key words: histone H5/globular domain/solution conformation/nuclear Overhauser effect/interproton distances/distance geometry/restrained molecular dynamics

Introduction

Histone H5 is a lysine-rich chromosomal protein present in the nucleated erythrocytes of birds, reptiles and fish which, like the related protein histone H1, is involved in the generation, maintenance and control of higher-order chromatin structure (McGhee and Felsenfeld, 1980). These two histones resemble each other in their primary structure (Yaguchi *et al.*, 1977, 1979; Briand *et al.*, 1980), and have been shown to be composed of a central globular domain of ~80 residues and disordered N- and C-terminal tails (Hartman *et al.*, 1977; Aviles *et al.*, 1978). Further, the globular domain of both histones is able to close two full turns in the nucleosome and to protect from nuclease digestion an extra ~20 bp of DNA present in the chromatosome above that in the core particle (Allan *et al.*, 1980).

In a recent paper (Zarbock *et al.*, 1986), we presented the sequential resonance assignment of the ^{1}H-n.m.r. spectrum of the 79-residue globular domain of histone H5 (GH5) and the elucidation of its secondary structure based on a qualitative interpretation of nuclear Overhauser effects (NOEs). In this paper we extend our previous study to the determination of the tertiary structure of GH5 based on 307 approximate interproton distance restraints derived from NOE data and calculations combining metric matrix distance geometry (Crippen and Havel, 1978; Havel *et al.*, 1983; Havel and Wüthrich, 1984, 1985) and restrained molecular dynamics (Levitt, 1983; Kaptein *et al.*, 1985; Clore *et al.*, 1985, 1986a,b, 1987a,b; Brünger *et al.*, 1986; Nilsson *et al.*, 1986). We show that GH5 is a globular protein whose architecture is dominated by three helical segments and whose topology is similar to that of the C-terminal DNA binding domain of the cAMP receptor protein of *Escherichia coli*. Possible implications of this structure with respect to the interaction of GH5 with DNA as well as with other GH5 molecules are discussed.

Results and Discussion

Tertiary structure computation

The basis for the tertiary structure computation consisted of a set of 307 approximate interproton distance restraints comprising 153 sequential ($|i-j| = 1$), 75 medium ($1 < |i-j| \leq 5$) and 77 long ($|i-j| > 5$) range interresidue distances. This interproton distance data set was derived from pure phase absorption two-dimensional NOE spectroscopy (NOESY) spectra (Jeener *et al.*, 1979) recorded with a mixing time of 100 ms. No intraresidue interproton distance restraints were included in the present calculations. The sequential NOEs were classifed into three distance ranges, 1.8–2.7 Å, 1.8–3.2 Å and 1.8–5.0 Å, corresponding to strong, medium and weak NOEs, while the medium and long range NOEs were grouped into the single distance range, 1.8–5.0 Å (Williamson *et al.*, 1985). A summary of the distance

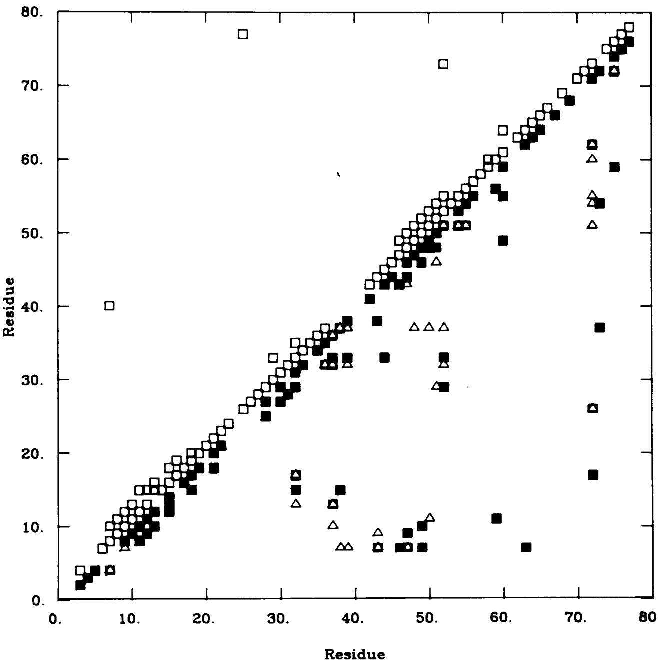

Fig. 1. Diagonal plot of the interproton distance restraints used in the determination of the solution structure of GH5. Backbone–backbone NOEs (□) are shown above the diagonal whereas backbone–sidechain (■) and sidechain–sidechain (△) NOEs are shown below the diagonal.

© IRL Press Limited, Oxford, England

Table I. Protocols of the restrained molecular dynamics refinements

Phase	Method A	Method B
1	3 ps, short $(<r^{-6}>)^{-1/6}$ averaging c = 0.1→20 kcal/mol/Å^2 T = 400−1000K	3 ps, short $<r_c>$ centre averaging c = 0.1→20 kcal/mol/Å^2 T = 400−1000K
2	4.2 ps, all $(<r^{-6}>)^{-1/6}$ averaging c = 0.1→20 kcal/mol/Å^2 T = 400−1000K	4.2 ps, all $<r_c>$ centre averaging c = 0.1→20 kcal/mol/Å^2 T = 400−1000K
3	−	4.2 ps, all $(<r^{-6}>)^{-1/6}$ averaging c = 0.1→20 kcal/mol/Å^2 T = 400−1000K
4	1.25 ps, all $(<r^{-6}>)^{-1/6}$ averaging c = 20 kcal/mol/Å^2 T cooled to 300K	1.25 ps, all $(<r^{-6}>)^{-1/6}$ averaging c = 20 kcal/mol/Å^2 T cooled to 300K
5	600 cycles restrained energy minimization $(<r^{-6}>)^{-1/6}$ averaging c = 20 kcal/mol/Å^2	600 cycles restrained energy minimization $(<r^{-6}>)^{-1/6}$ averaging c = 20 kcal/mol/Å^2

In phases 1−3 the temperature of the system was adjusted to lie between 400 and 1000K by scaling the velocities of the atoms upwards by a factor of 1.5 (phase 1) or 1.25 (phase 2) if the temperature fell below 400K and downwards by a factor of 0.75 if the temperature increased above 1000K. The velocity scaling was carried out every 0.25 ps where appropriate. The NOE restraints force constants (c) were increased from 0.1 kcal/mol/Å^2 up to a maximum value of 20 kcal/mol/Å^2 by doubling their value every 0.25 ps in phases 1−3. In phase 1 only short range ($|i-j| \leq 5$) NOE restraints (i.e. sequential and medium) were included in the calculation (denoted by 'short'); in all other phases all the NOE restraints were included (denoted by 'all').

restraints is shown in Figure 1. The NOE data were not supplemented by ϕ backbone torsion angle restraints as the linewidths of the NH proton resonances were too large to enable one to obtain reliable estimates of the $^3J_{HN\alpha}$ coupling constants from a DQF-COSY spectrum (Neuhaus *et al.*, 1985).

The computation of the tertiary structure employed the same two-stage approach that we used previously for α1-purothionin (Clore *et al.*, 1986b), phoratoxin (Clore *et al.*, 1987a) and hirudin (Clore *et al.*, 1987b): namely, a structure generation phase using the distance geometry program DISGEO (Havel, 1986) followed by a refinement stage using a combination of restrained energy minimization and restrained molecular dynamics in which the NOE interproton distances were incorporated into the total energy function of the system in the form of effective potentials (Kaptein *et al.*, 1985; Clore *et al.*, 1985, 1986a; Brünger *et al.*, 1986).

In the structure generation stage, interproton distances involving methyl and methylene protons were corrected for the pseudo-atom representation used by DISGEO as described by Wüthrich *et al.* (1983). These corrections were kept in some parts of the refinement stage (Table I) where interproton distances involving these protons were referred to single centre average distances, $<r_c>$, which is essentially equivalent to referring them to centrally placed pseudo-atoms. Otherwise, the interproton distances involving these protons were referred to single $(<r^{-6}>)^{-1/6}$ average distances so that no corrections were required (Table I). The form of the effective restraints potential was a square well (Clore *et al.*, 1986b).

Although the four DG structures generated by DISGEO converged to the same polypeptide fold (Table II), they exhibited some large deviations with respect to the NOE interproton distance restraints (see Tables III and IV), particularly for those involving medium- and long-range NOE restraints (Table III), and were poor in stereochemical terms as evidenced by the high values of the non-bonded energy terms (Table IV). Some of these

Table II. Atomic r.m.s. distributions and shifts

		Atomic r.m.s. difference (Å)							
		All residues		Residues 3−18, 23−34, 37−60 and 71−79		Residues 19−22		Residues 61−70	
		Backbone atoms	All atoms	Backbone atoms	All atoms	Backbone atoms	All atoms	Backbone atoms	All atoms
A. R.m.s. distributions									
<DG>	versus $\overline{\mathrm{DG}}$	2.1 ± 0.4	2.8 ± 0.7	1.9 ± 0.4	2.5 ± 0.5	1.0 ± 0.2	1.5 ± 0.6	1.9 ± 0.4	2.5 ± 0.4
<DGm>	versus $\overline{\mathrm{DGm}}$	2.1 ± 0.5	2.6 ± 0.6	1.9 ± 0.5	2.5 ± 0.6	1.0 ± 0.2	1.5 ± 0.5	1.9 ± 0.4	2.5 ± 0.4
<RDDG>	versus $\overline{\mathrm{RDDG}}$	3.0 ± 0.3	3.4 ± 0.3	2.3 ± 0.3	2.9 ± 0.3	1.4 ± 0.2	2.3 ± 0.3	2.3 ± 0.3	2.9 ± 0.5
<RDDG>	versus <RDDG>	4.3 ± 0.7	5.0 ± 0.7	3.4 ± 0.6	4.2 ± 0.6	1.4 ± 0.4	2.5 ± 0.5	3.1 ± 0.5	4.1 ± 0.7
B. R.m.s. shifts									
<DG>	versus <DGm>	0.8 ± 0.06	1.1 ± 0.05	0.8 ± 0.06	0.9 ± 0.04	0.5 ± 0.1	0.7 ± 0.2	0.7 ± 0.1	0.8 ± 0.1
<DGm>	versus <RDDG>	3.9 ± 0.5	4.6 ± 0.5	3.3 ± 0.4	4.2 ± 0.3	1.3 ± 0.3	2.1 ± 0.6	2.7 ± 0.4	3.6 ± 0.6
<DG>	versus <RDDG>	3.9 ± 0.5	4.7 ± 0.5	3.4 ± 0.4	4.3 ± 0.3	1.3 ± 0.3	2.1 ± 0.5	2.7 ± 0.4	3.6 ± 0.6
$\overline{\mathrm{DG}}$	versus $\overline{\mathrm{DGm}}$	0.5	0.6	0.5	0.5	0.3	0.5	0.4	0.5
$\overline{\mathrm{DGm}}$	versus $\overline{\mathrm{RDDG}}$	2.4	2.9	2.1	2.7	1.2	1.8	1.7	2.3
$\overline{\mathrm{DG}}$	versus $\overline{\mathrm{RDDG}}$	2.5	3.0	2.2	2.8	1.2	1.7	1.6	2.2
($\overline{\mathrm{RDDG}}$)m	versus $\overline{\mathrm{RDDG}}$	1.6	1.8	1.4	1.6	1.3	1.8	1.6	1.8

The notation of the structures is as follows: <DG> comprise the four converged distance geometry structures, <DGm> the four structures derived from the DG structures by restrained energy minimization, and <RDDG> the eight structures derived from the DGm structures by restrained molecular dynamics using the two methods outlined in Table I. $\overline{\mathrm{DG}}$, $\overline{\mathrm{DGm}}$ and $\overline{\mathrm{RDDG}}$ are the mean structures obtained by averaging the coordinates of the DG, DGm and RDDG structures, respectively, best fitted to the 'core' residues (3−18, 23−34, 37−60 and 71−79; see text). The standard atomic r.m.s. error of these mean structures is given by ~r.m.s.d./$\sqrt{n}$ where r.m.s.d. is the average atomic r.m.s. difference between the n structures and the mean structure. ($\overline{\mathrm{RDDG}}$)m is the structure obtained by restrained energy minimization of the mean $\overline{\mathrm{RDDG}}$ structure. Residues 19−22 and 61−70 comprise two loops whose local structure is reasonably well defined but whose orientation with respect of the 'core' residues could not be defined.

problems could be partially corrected by 1000 cycles of restrained energy minimization (with force constants of 20 kcal/mol/Å^2 for the NOE restraints with $<r_c>$ centre averaging) to generate the DGm structures. Thus the deviations in all three classes of NOE restraints were significantly reduced (Table III) with a concomitant reduction of ~1700 kcal/mol in the NOE $<r_c>$ restraints energy (Table IV). In addition, the total non-bonding energy was reduced by 3000–70 000 kcal/mol spanning the lowest to highest energy DG structures (Table IV). These improvements, however, were achieved by only minor structural changes as the backbone atomic r.m.s. shifts produced by this procedure were small (≤1 Å; see Table II). Examination of the DGm structures on an interactive molecular graphics display revealed that some impossibly close contacts were still present and that the structures exhibited several features that were unusual for protein structures. These features, however, were restricted to irregular structural elements such as loops and turns. Our conclusion at this stage was that the overall polypeptide fold of the structures generated by the distance geometry calculations and refined by restrained energy minimization was approximately correct but that some local structural features were clearly incorrect. Thus, in energy terms, the distance geometry calculations had located the global minimum energy region but had then got trapped in high energy local subminima. This is not entirely surprising as GH5 represents a difficult case not only because of its size but, more importantly, because of the nature of its secondary structure: namely, there are helices and irregular elements present but not β-sheets. As a result there are very few long-range backbone–backbone NOEs: namely, only three out of a total of 77 long-range NOEs. Whereas the conformation of an α-helix is well defined by the short- and medium-range backbone–backbone NOEs and that of a β-sheet by the intra- and inter-strand backbone–backbone NOEs (Wüthrich *et al.*, 1984), the definition of the conformation of an irregular structure element afforded by interproton distances of <5 Å is considerably less precise.

To overcome these problems we adopted two protocols of restrained molecular dynamics (see Table I for details) designed to overcome large local energy barriers on the path towards the lowest energy local subminima within the global minimum re-

Table III. Interproton distance deviations and radii of gyration

Structure	R.m.s. difference between calculated and target interproton distance restraints (Å)								Radius of gyration (Å)
	$<r^{-6}>)^{-1/6}$ averaging				$<r_c>$ averaging				
	All (307)	Sequential (153)	Medium (75)	Long (77)	All (307)	Sequential (153)	Medium (75)	Long (77)	
<DG>	1.30 ± 0.09	0.51 ± 0.05	1.40 ± 0.11	1.90 ± 0.17	0.57 ± 0.06	0.29 ± 0.07	0.81 ± 0.10	0.69 ± 0.12	11.31 ± 0.08
<DGm>	0.95 ± 0.07	0.42 ± 0.03	0.94 ± 0.10	1.50 ± 0.15	0.20 ± 0.03	0.11 ± 0.17	0.28 ± 0.09	0.18 ± 0.04	11.35 ± 0.11
<RDDG>	0.23 ± 0.14	0.19 ± 0.02	0.20 ± 0.03	0.32 ± 0.03	0.15 ± 0.01	0.13 ± 0.02	0.14 ± 0.04	0.18 ± 0.03	10.96 ± 0.13
$\overline{\text{DG}}$	0.93	0.28	0.90	1.60	0.34	0.19	0.50	0.38	11.06
$\overline{\text{DGm}}$	0.67	0.30	0.58	1.10	0.19	0.19	0.21	0.15	11.13
$\overline{\text{RDDG}}$	0.29	0.34	0.09	0.31	0.23	0.31	0.04	0.12	10.45
($\overline{\text{RDDG}}$)m	0.26	0.21	0.26	0.33	0.13	0.14	0.09	0.16	11.32

The notation of the structures is the same as that in Table II. The r.m.s. difference [r.m.s.d.] between the calculated (r_{ij}) and target restraints is calculated with respect to the upper (r_{ij}^{u}) and lower (r_{ij}^{l}) limits such that

$$\text{r.m.s.d.} = \begin{cases} [\Sigma(r_{ij} - r_{ij}^{u})^2/n]^{1/2} & \text{if } r_{ij} > r_{ij}^{u} \\ 0 & \text{if } r_{ij}^{l} \leq r_{ij} \leq r_{ij}^{u} \\ [\Sigma(r_{ij} - r_{ij}^{l})^2/n]^{1/2} & \text{if } r_{ij} < r_{ij}^{l} \end{cases}$$

In the case of $(<r^{-6}>)^{-1/6}$ averaging, no corrections to the upper limits for distances involving methyl and methylene protons are used. For $<r_c>$ centre averaging, on the other hand, the upper limits for these distances are corrected in the same way as those for the pseudo-atom representation (Wüthrich *et al.*, 1983) used in the distance geometry calculations.

Table IV. Energies of the structures

Structure	Potential	Bond (1238)	Angle (2238)	Dihedral (589)	Improper (342)	Van der Waals	Electrostatic	H-bond	NOE restraints (307)	
									$(<r^{-6}>)^{-1/6}$ averaging	$<r_c>$ averaging
DG	5675– >10^5	336– >10^5	2078 ± 291	642 ± 31	2.9 ± 4.5	2250–69933	–60 ± 49	–17 ± 3	9937 ± 1396	1982 ± 377
DGm	1099 ± 346	228 ± 59	997 ± 186	472 ± 17	5.8 ± 1.6	247 ± 90	–815 ± 70	–49 ± 6	5558 ± 889	237 ± 82
RDDG	–365 ± 150	112 ± 15	648 ± 57	383 ± 24	3.7 ± 1.0	–72 ± 56	–1355 ± 32	–84 ± 10	323 ± 32	133 ± 14
($\overline{\text{RDDG}}$)m	415	145	925	460	6.3	17	–1078	–60	401	108

The notation of the structures is the same as that in Table II. The number of terms for the bond, angle, dihedral and improper dihedral potentials and for the effective NOE interproton distance restraints potential are given in parentheses. The potential energy is the sum of all energies excluding the NOE restraints energy. The effective restraints potentials are represented by a square well potential (see Clore *et al.*, 1986b) with restraints force constants of 20 kcal/mol/Å^2. Two values for the NOE restraints potential are given: one calculated using $(<r^{-6}>)^{-1/6}$ averaging with no corrections to the upper limits for distances involving methyl and methylene protons, the other calculated using $<r_c>$ centre averaging together with an appropriate correction for the upper limits. The DG and DGm structures were generated using a pseudo-atom representation and $<r_c>$ centre averaging, respectively, for distances involving methyl and methylene protons. (Note that the pseudo-atom representation and $<r_c>$ centre averaging are approximately equivalent.) The RDDG and $\overline{\text{RDDG}}$)m structures, on the other hand, were generated using $(<r^{-6}>)^{-1/6}$ averaging.

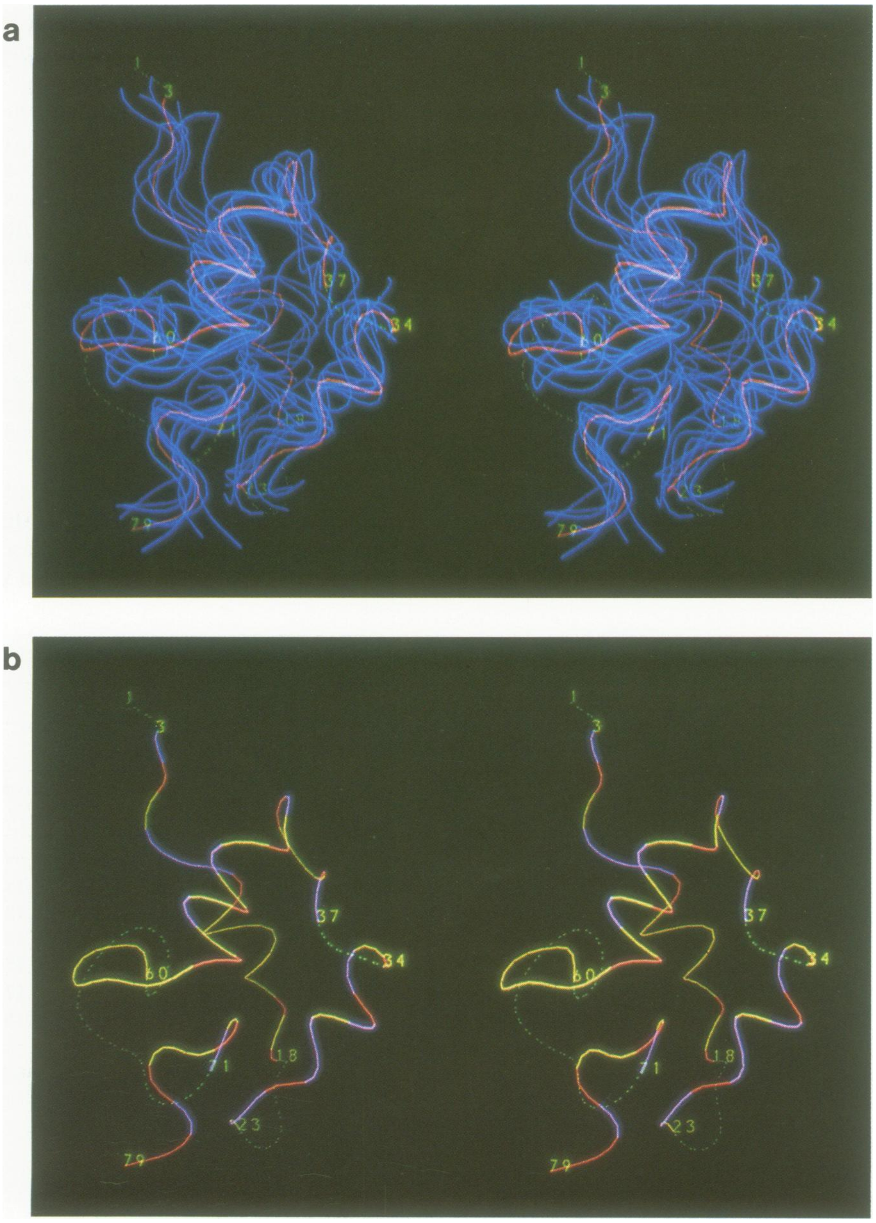

Fig. 2. Smoothed backbone (N, C$^\alpha$, C) atom representation of the final restrained molecular dynamics structures of GH5. **(a)** Superposition of the 'core' residues (3–18, 22–34, 37–60 and 71–79) of the eight RDDG structures (blue) on the restrained energy minimized mean structure ($\overline{\text{RDDG}}$)m (red). The two loops (residues 19–22 and 61–70) which are reasonably well defined locally but whose orientation with respect to the 'core' residues cannot be defined, as well as the two ill-defined regions (residues 1–2 and 35–36) are shown as dashed lines (green) in the case of ($\overline{\text{RDDG}}$)m. **(b)** Distribution of the charged (red), polar (lilac) and hydrophobic (yellow) residues of the 'core' region of ($\overline{\text{RDDG}}$)m. As in (a) the two-loops (residues 19–22 and 61–70) and the two ill-defined regions (residues 1–2 and 35–36) are shown as dashed lines.

gion. These two protocols generated eight RDDG structures from the four DGm structures, and the average r.m.s. difference between pairs of RDDG structures generated from the same DGm structure using different protocols was approximately the same as that between pairs of RDDG structures generated from different DGm structures. Restrained molecular dynamics resulted in large backbone atomic r.m.s. shifts (>2 Å) as well as an increase in the atomic r.m.s. distribution of the structures, while maintaining the overall polypeptide fold (Table II). Concomitant with these structural changes was a considerable reduction in all the non-bonding energies (Table IV). Indeed the total non-bonding energy of the RDDG structures is ~800 kcal/mol lower than that of the DGm structures. In addition, the $<r_c>$ centre average and $(<r^{-6}>)^{-1/6}$ average NOE restraints energies are reduced by factors of ~2 and ~20, respectively, with respect to those for the DGm structures (Table IV). The main source for the reduction in the NOE restraints energy comes from an improvement in the medium- and long-range NOEs (Table III) and is achieved principally by alterations in the local conformations of the loops and turns.

Examination of the RDDG structures on an interactive molecular graphics display revealed the presence of a well-defined 'core' comprising residues 3–18, 23–34, 37–60 and 71–79, and two locally well-defined loops comprising residues 19–22 and 61–70 whose orientations with respect to the 'core' could not be determined (due to the absence of any long-range $|i-j| > 5$ NOEs between loop and 'core' residues). In addition, there were two ill-defined regions comprising residues 1–2 and 35–

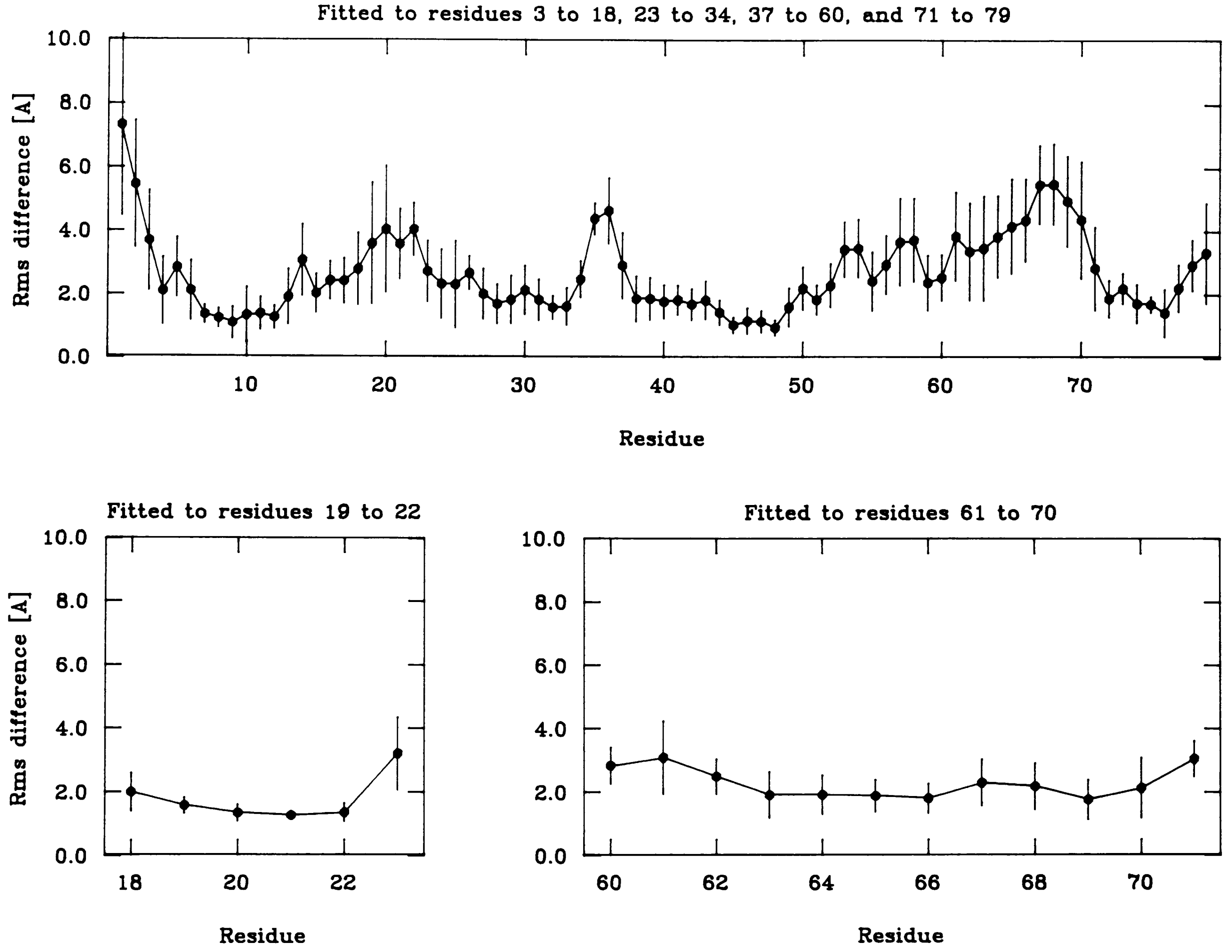

Fig. 3. Atomic r.m.s. distributions of the backbone atoms (N, C^{α}, C, O) of the eight RDDG structures about the mean $\overline{\mathrm{RDDG}}$ structure best fitted to (**a**) the 'core' residues (3–18, 23–34, 37–60 and 71–79), (**b**) the first loop (residues 19–22) and (**c**) the second loop (residues 61–70). The filled-in circles (●) represent the average r.m.s. difference at each residue between the RDDG structures and the mean $\overline{\mathrm{RDDG}}$ structure, and the bars represent the SDs in these values.

36. For this reason the mean structure $\overline{\mathrm{RDDG}}$ was generated by averaging the coordinates of the individual structures best fitted to the 'core' residues. This average structure represents the mean about which the 'core' residues of the RDDG structures are randomly distributed. $\overline{\mathrm{RDDG}}$ is poor with respect to all energy terms and is stereochemically a bad structure. For this reason $\overline{\mathrm{RDDG}}$ was subjected to 1500 cycles restrained energy minimization slowly increasing the van der Waals radii from a quarter of their usual values to their full values (Clore *et al.*, 1986a) to generate the structure ($\overline{\mathrm{RDDG}}$)m. This structure is closer to $\overline{\mathrm{RDDG}}$ than any of the individual RDDG structures (Table I). At the same time ($\overline{\mathrm{RDDG}}$)m is reasonable in stereochemical and energetic terms, and although its energy is not as low as that of the individual RDDG structures, it is lower than that of any of the individual DGm structures (Table IV). The best fit superposition of the 'core' residues of the RDDG structures on ($\overline{\mathrm{RDDG}}$)m is shown in Figure 2a. The atomic r.m.s. distributions of the RDDG structure about the mean structure $\overline{\mathrm{RDDG}}$ best fitted either to the core residues or to the loop residues are plotted in Figure 3 as a function of residue number and the results are summarized in Table II.

The local atomic positions of the backbone atoms are relatively well defined throughout as shown by the local backbone atomic r.m.s. distributions of tripeptide segments of the RDDG structures about the mean structure $\overline{\mathrm{RDDG}}$ (Figure 4). As expected the variation of the ϕ and ψ backbone torsion angles is somewhat larger (Figure 5) with an average value of 75 ± 8° for the average angular r.m.s. difference between pairs of structures. Within the helical regions (residues 7–17, 26–34 and 43–54), however, the variation in the ϕ,ψ angles is smaller with an average angular r.m.s. difference value of 50 ± 10°.

Structural features of GH5

The structural features of GH5 are illustrated by the stereoviews of the smoothed backbone atom traces shown in Figure 2a and 2b and by the diagrammatic representation shown in Figure 6. GH5 is a globular three-helix structure stabilized by hydrophobic interactions involving Tyr-32, Tyr-37 and Phe-72 as well as long-chain aliphatic residues (e.g. Leu and Ile). The total number of helical residues in the present structure is 32 which is in good agreement with the prediction of 28 residues from circular dichroism measurements (Aviles *et al.*, 1978; Giancotti *et al.*, 1981). In addition, the absence of any β-sheet structure is in agreement with the circular dichorism data. The average radius of gyration of the RDDG structures is 10.96 ± 0.13 Å and that of the restrained energy minimized mean structure ($\overline{\mathrm{RDDG}}$)m is 11.32 Å. These values are in agreement with the value of 11.4 ± 0.7 Å determined by small-angle neutron scattering (Aviles *et al.*, 1978). The distance between residues 1 and 79 is ~33 Å. This

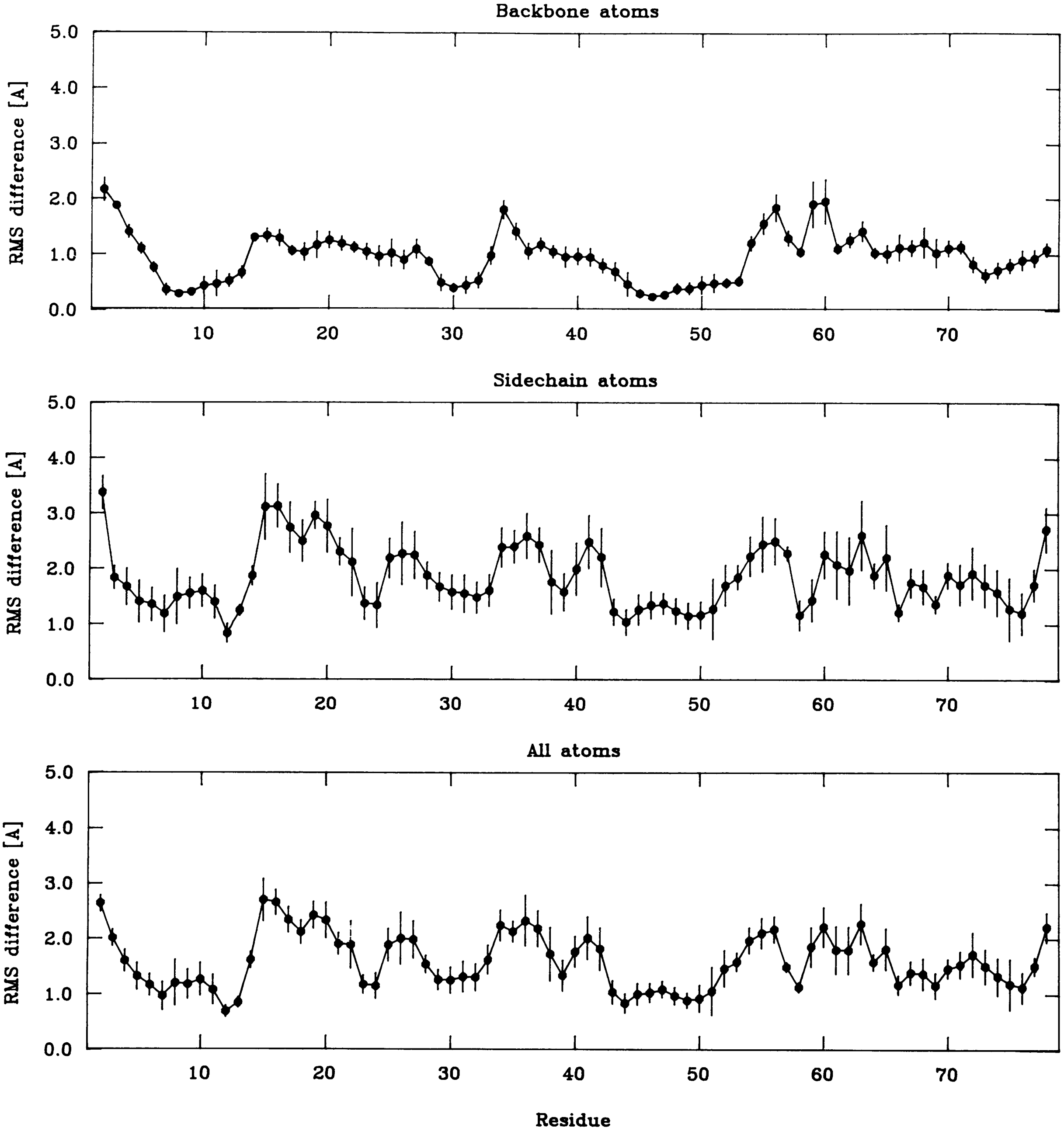

Fig. 4. Local atomic r.m.s. distributions of the eight RDDG structures about the mean structure $\overline{\text{RDDG}}$ for the backbone atoms, sidechain atoms and all atoms. The filled-in circles (●) represent the average best fit atomic r.m.s. differences between tripeptide segments along the chain as a function of the sequence number of the middle residue, and the bars represent the SDs in these values.

value is in accord with the model of Allan *et al.* (1980) in which a cage of three double helical DNA strands binds histone H5 at the exit points of the chromatosome with the globular domain interacting with the central strand and the inner surface of the two outer strands, and the N- and C-terminal tails wrapping around one outer strand each.

The N-terminal strand of GH5 leads into helix I (residues 7–17). This helix is not entirely regular and exhibits a small deformation at residue 14. This is followed by a loop which reverses the direction of the chain and leads into helix II (residues 26–34). The angle between helices I and II is ~130°. Helix III is connected to helix II by an irregular strand with a 'half-turn' and lies at an angle of ~180° to helix II and ~60° to helix I. The rest of the structure is somewhat irregular, consisting principally of loops and turns.

How do the location of the three helices compare with those predicted on the basis of a qualitative interpretation of the short-range NOE data involving the NH and C$^\alpha$H protons given in our previous paper (Zarbock *et al.*, 1986)? In addressing this question two factors must be borne in mind. Namely, while helices can readily be identified on the basis of such a qualitative interpretation, the exact start and end of a helix is difficult to ascertain in this manner (Wüthrich *et al.*, 1984). Further, such a qualitative interpretation does not allow one to readily distinguish turns at the beginning or end of a helix from the helix itself (Wüthrich *et al.*, 1984).

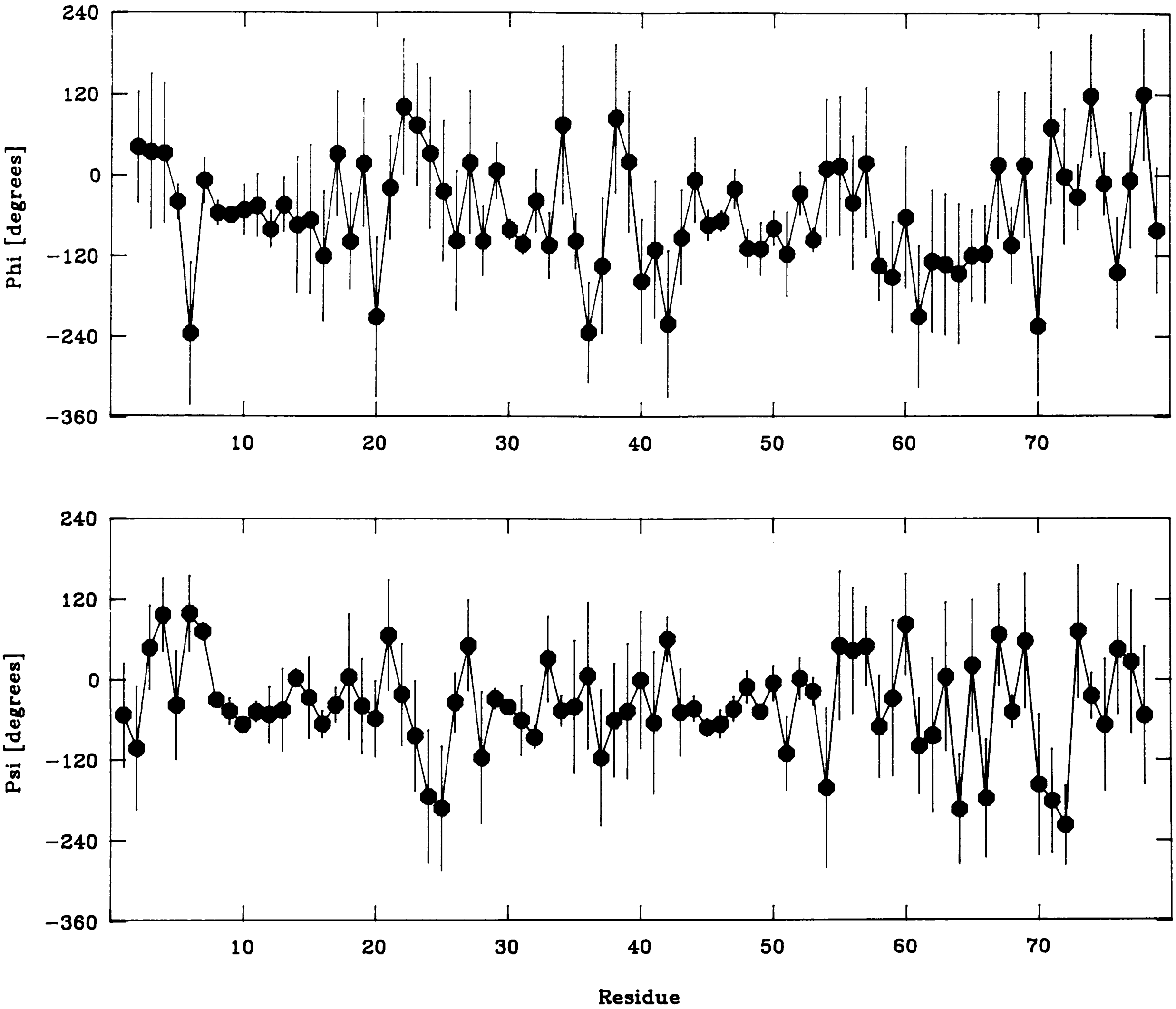

Fig. 5. Angular r.m.s. distributions of the ϕ and ψ backbone torsion angles of the RDDG structures. The filled-in circles (●) are the values of the ϕ and ψ angles of the restrained energy minimized structure ($\overline{\text{RGGD}}$)m and the bars represent the average angular r.m.s. deviations between all pairs of RDDG structures.

We previously proposed four helices comprising residues 7−13, 15−19, 29−35 and 44−55 (Zarbock *et al.*, 1986). The first two helices from residues 7−13 and 15−19 correspond to helix I (residues 7−17) and the adjacent turn (residues 18−19) leading into the loop region (residues 18−25) in the present three-dimensional structure. The reason that two helices were proposed for the region comprising residues 7−19 was that no NH(i)−NH($i+1$) NOE characteristic of a regular α-helix was observed between residues 14 and 15, while a strong C$^{\alpha}$H(i)−NH($i+1$) NOE between residues 14 and 15, characteristic of a more extended conformation, was detected. We did, however, observe a C$^{\alpha}$H(i)−NH($i+2$) NOE between residues 13 and 15 and a C$^{\alpha}$H(i)−NH($i+4$) NOE between residues 11 and 15, both of which point to some sort of helical structure. This NOE data is clearly indicative of some irregularity around residue 14 and it was originally suggested that the helix axis of residues 15−19 could be at a slight angle to that of residues 7−13. In fact this is not the case as in all the structures generated that satisfy the NOE data, there is only one helix comprising residues 7−17 followed by a turn (residues 18−19). Helix I, however, is not entirely regular in the stretch from residues 14 to 15, exhibiting a slight deformation at residue 14, as predicted from a qualitative interpretation of the short-range NOE data. This deformation, though, is not large enough to warrant the division of the segment from residues 7 to 17 into two helices.

The helix from residues 29 to 35 proposed by Zarbock *et al.* (1986) corresponds to helix II (residues 26−34) and the subsequent ill-defined region comprising residues 35−36. Thus, this is an example where the start of the N terminus of a helix was underestimated and the end of the C terminus overestimated by the qualitative interpretation. The size and location of helix III, on the other hand, agrees well with the qualitative deductions we made previously (Zarbock *et al.*, 1986), namely the segment extending from residues 44 to 55.

Thus, helices I, II and III in this paper correspond to helices I/II, III and IV, respectively, in our previous paper on the sequential resonance assignment and identification of secondary structure elements (Zarbock *et al.*, 1986). It must be emphasized that these differences are very minor and simply reflect the deficiencies involved in the exact delineation of the beginning and ends of regular secondary structure elements, particularly helices, based on a qualitative interpretation of short-range NOE data. These ambiguities, however, are easily resolved when the tertiary structure of the protein is determined on the basis of all the available interproton distance data using distance geometry and restrained molecular dynamics calculations.

In the Zarbock *et al.* (1986) paper we also made a limited attempt to define the spatial relationships of the helices using manual model building on the basis of the limited long-range ($|i-j| > 5$) NOE data available at that time. We correctly ascertained that the N-terminal ends of helices I and III and the central region of helix II were close to each other in space. The angle, however, between the long axes of helices II and III was slightly underestimated, illustrating the limitations of a manual model building approach in determining the polypeptide fold of proteins: model building suggested an angle in the range 100–140° (using the standard sign convention which gives information on the relative directionality of the helices), whereas the two helices are in fact somewhat more anti-parallel to each other with an angle of ~180° between the two helical axes (see Figure 6).

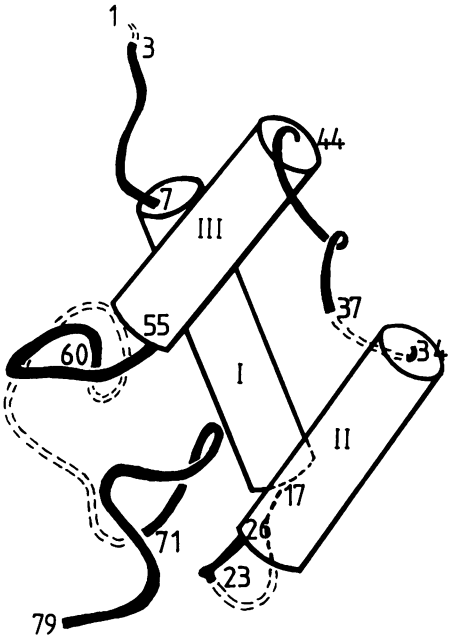

Fig. 6. Diagrammatic representation of the $(\overline{\text{RDDG}})$m structure of GH5. The helices are represented as cylinders and the other residues as thick lines. The two loops (residues 19–22 and 61–70) and the two ill-defined regions (residues 1–2 and 35–36) are shown as dashed lines.

It is also of interest to see how the present structure relates to some of the conclusions derived from a very early one-dimensional n.m.r. study at low (270 MHz) field strength (Chapman *et al.*, 1978). These authors suggested on the basis of one-dimensional NOE measurements that His-4, Tyr-7 and Tyr-32 were close in space. His-4 and Tyr-7 are indeed close with a separation of ~4 Å but far away (~10 Å) from Tyr-32, the discrepancy arising from artefacts of the one-dimensional NOE measurements. They also suggested on the basis of pH titration data that of the three tyrosine residues only Tyr-7 is buried. In fact Tyr-32 and Tyr-37 are also buried. Further, Tyr-7 is only partially buried, its side chain being located in a cleft formed by the junction of helices I and III and shielded from solvent by the side chain of Gln-46.

While looking at the GH5 structure on the graphics display, we noticed that the polypeptide fold was remarkably similar to that of the C-terminal DNA binding domain of the cAMP receptor protein (known as CRP or CAP) solved by McKay and Steitz (1981). For comparison the C^{α} backbone of the C-terminal domain of CRP (residues 138–206) is superimposed on GH5 (Figure 7) and the structural alignment of residues together with their sequences is shown in Figure 8. The orientation of helices I and III of GH5 are similar to those of helices D and F of CRP. Further, the path, but not the local structure, of the polypeptide chain from the end of helix I to the beginning of helix III in GH5 is similar to that from the end of helix D to the beginning of helix F in CRP. The helix–turn–helix motif comprising helices E and F in CRP, however, is not present in GH5 where helix E is replaced by a somewhat irregular structure. Further, helix II in GH5 is replaced by a small anti-parallel β-sheet in CRP. Thus the correspondence of residues in structural terms is as follows: residues 138–152 (helix D) of CRP correspond to residues 5–19 in GH5 (which include helix I), residues 152–161 of CRP to residues 19–24 of GH5, residues 161–168 of CRP to residues 24–34 (helix II) of GH5, residues 168–176 (helix E) of CRP

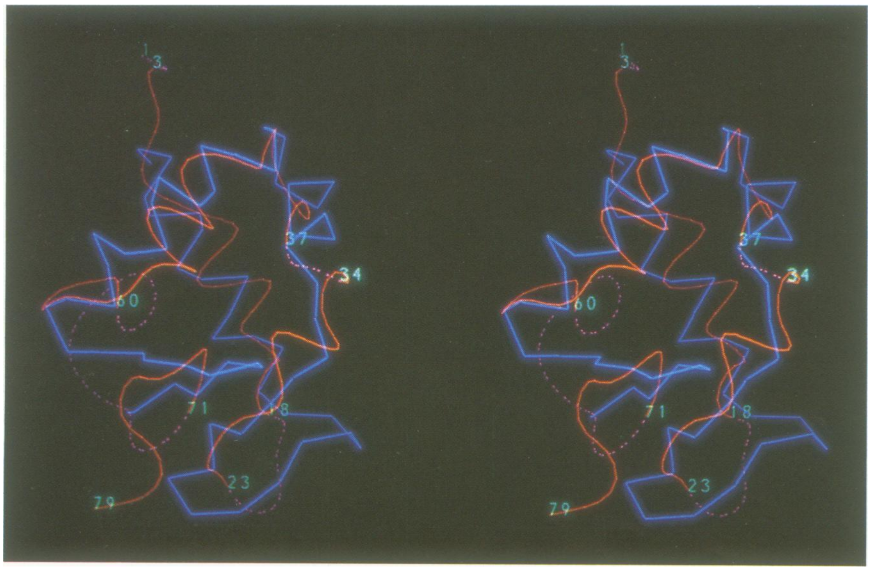

Fig. 7. Superposition of the smoothed backbone of GH5 (red) and the C^{α} backbone (blue) of the C-terminal DNA binding domain of CRP (residues 138–206). The structure of CRP was solved by McKay and Steitz (1981) and the coordinates are taken from the Brookhaven protein data bank.

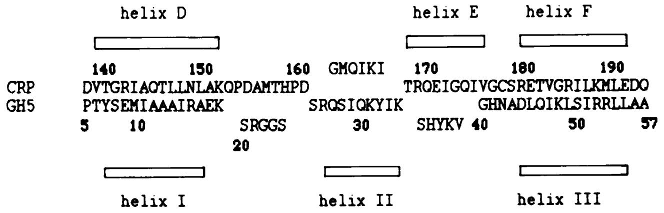
helix D
helix E
helix F
140 150 160 GMQIKI 170 180 190
CRP DVTGRIAQTLLNLAKQPDAMTHPD TRQEIGQIVGCSRETVGRILKMLEDQ
GH5 PTYSEMIAAAIRAEK SRQSIQKYIK GHNADLQIKLSIRRLLAA
5 10 SRGGS 30 SHYKV 40 50 57
20
helix I
helix II
helix III

Fig. 8. Alignment of residues 5–57 of GH5 with those of residues 138–193 of the C-terminal DNA binding domain of CRP (Aiba *et al.*, 1982; Cossart and Gicquel-Sanzey, 1982) according to the structural alignment shown in Figure 7. The one-letter amino acid code is used.

to residues 34–40 of GH5, residues 176–178 (the turn of the helix–turn–helix motif) of CRP to residues 40–43 of GH5, and finally residues 179–193 (helix F) of CRP to residues 43–57 of GH5 (which inclues helix III). There is not a straightforward correspondence for the C terminus, although the path of residues 68–75 of GH5 is slightly similar to that of residues 206–196 of CRP (note the reverse orientation in the direction of the polypeptide chain here). This structural homology does not appear to be related to any sequence homology as is easily ascertained from Figure 8. We also note that the polypeptide fold of GH5 is quite different from that of cro (Anderson *et al.*, 1981) which, like CRP, also has a helix–turn–helix motif associated with DNA binding.

These findings could be interpreted to suggest that similar polypeptide folds have evolved convergently to fulfil similar functions, in this case DNA binding. A similar example which immediately comes to mind is the recently discovered similarity in the polypeptide fold of plasma retinol-binding protein with β-lactoglobulin (Papiz *et al.*, 1986).

The present structure permits some speculation with respect to the location of the DNA binding sites of GH5 that could be tested by biological experiments (e.g. using site-specific mutagenesis) and, hopefully, by X-ray crystallography if suitable crystals are obtained. By analogy with CRP (Ebright *et al.*, 1984; Ebright, 1986; Gent *et al.*, 1987) we predict that the polar residues at the N-terminal end of helix III (e.g. Asp-44, Gln-46 and Lys-48) may be involved in protein–DNA contacts. We would also point out that helix II has a large number of exposed polar and charged residues (Figure 2b) and may therefore also be involved in protein–DNA contacts. These postulated DNA binding sites are supported by the recent work of Thomas and Wilson (1986) who identified a number of lysine residues that are protected from selective radio-labelling by reductive methylation on the binding of histone H5 to the nucleosome. Within the globular domain six out of eight lysines were protected, namely the lysines at positions 31, 34, 38, 48, 61 and 64 (numbered according to the residue numbering in GH5). Lys-31 and Lys-34 are located in helix II, Lys-48 in helix III and Lys-38 in the stretch of polypeptide just before the start of helix III. Lys-64, on the other hand, was even more protected than the other lysines so that the locally well defined loop from residues 61 to 70 in which both Lys-61 and Lys-64 are located could also be involved in DNA binding. Alternatively, Lys-61 and Lys-64 could be protected as a result of a conformational change in the position of this loop upon DNA binding. In contrast to the other two helices, helix I is principally composed of hydrophobic residues and is therefore unlikely to be the site of protein–DNA contacts. The hydrophobic nature of helix I, however, would make it a suitable site for protein–protein contacts of the type proposed by Losa *et al.* (1984) between adjacent histone H1 molecules in chromatin. Further, the fact that the exposed face of helix I lies on the opposite side of the molecule to the exposed faces of helices II and III presents an ideal geometry for such dual DNA–protein and protein–protein interactions.

Materials and methods

GH5 was a gift of Drs J.Gunning and S.Neidle who prepared GH5 by tryptic digestion of chicken erythrocyte H5 and purified it as described by Aviles *et al.* (1978). Samples for n.m.r. contained 7 mM GH5 in either 90% H_2O/10% D_2O or 99.96% H_2O buffer consisting of 500 mM KCl, 50 mM phosphate buffer (pH 3.7) and 0.1 mM EDTA. NOESY spectra (Jeener *et al.*, 1979; Macura *et al.*, 1981) were recorded in the pure phase absorption mode (Marion and Wüthrich, 1983) using the experimental conditions reported previously (Zarbock *et al.*, 1986). All measurements were carried out at 25°C.

Metric matrix distance geometry calculations were carried out using the program DISGEO (Havel and Wüthrich, 1984; Havel, 1986). All energy minimization and restrained molecular dynamics calculations were carried out as described previously (Clore *et al.*, 1986a,b, 1987a,b; Brünger *et al.*, 1986) on a CRAY-XMP using a CRAY version (A.T.Brünger, unpublished data) of the program CHARMM (Brooks *et al.*, 1983). Displaying of the structures was carried out using a modified version of the function network of FRODO (Jones, 1978) interfaced with CHARMM on an Evans and Sutherland PS330 colour graphics system. The smooth backbone (N,C^α,C) atom representations shown in Figures 2 and 7 were obtained as described previously (Feldmann *et al.*, 1986).

Acknowledgements

We thank Drs S.Neidle and J.Gunning for the gift of GH5. This work was supported by the Max-Planck Gesellschaft and Grant No. Cl 86/1-1 of the Deutsche Forschungsgemeinschaft (G.M.C. and A.M.G.).

References

Aiba,H., Fujimoto,S. and Ozaki,N. (1982) *Nucleic Acids Res.*, **10**, 1345–1362.

Allan,J., Hartman,P.G., Crane-Robinson,C. and Aviles,F.J. (1980) *Nature*, **288**, 675–679.

Anderson,W.F., Ohlendorf,D.H., Takeda,Y. and Mattehws,B.W. (1981) *Nature*, **290**, 754–758.

Aviles,F.J., Chapman,G.E., Kneale,G.G., Crane-Robinson,C. and Bradbury, E.M. (1978) *Eur. J. Biochem.*, **88**, 363–371.

Briand,G., Kmiecik,D., Sautiere,P., Wouters,D., Borie-Loy,O., Biserte,G., Mazen,A. and Champagne,M. (1980) *FEBS Lett.*, **112**, 147–151.

Brooks,B.R., Bruccoleri,R.E., Olafson,B.D., States,D.J., Swaminathan,S. and Karplus,M. (1983) *J. Comput. Chem.*, **4**, 187–217.

Brünger,A.T., Clore,G.M., Gronenborn,A.M. and Karplus,M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 3801–3805.

Clore,G.M., Gronenborn,A.M., Brünger,A.T. and Karplus,M. (1985) *J. Mol. Biol.*, **186**, 435–455.

Clore,G.M., Brünger,A.T., Karplus,M. and Gronenborn,A.M. (1986a) *J. Mol. Biol.*, **191**, 523–551.

Clore,G.M., Nilges,M., Sukumaran,D.K., Brünger,A.T., Karplus,M. and Gronenborn,A.M. (1986b) *EMBO J.*, **5**, 2729–2735.

Clore,G.M., Sukumaran,D.K., Nilges,M. and Gronenborn,A.M. (1987a) *Biochemistry*, **26**, 1732–1745.

Clore,G.M., Sukumaran,D.K., Nilges,M., Zarbock,J. and Gronenborn,A.M. (1987b) *EMBO J.*, **6**, 529–537.

Chapman,G.E., Aviles,F.J., Crane-Robinson,C. and Bradbury,E.M. (1978) *Eur. J. Biochem.*, **90**, 287–296.

Cossart,P. and Gicquel-Sanzey,B. (1982) *Nucleic Acids Res.*, **10**, 1363–1378.

Crippen,G.M. and Havel,T.F. (1978) *Acta Crystallogr.*, **A34**, 282–284.

Ebright,R.H. (1986) In Oxender,D. (ed.), *Protein Structure, Folding and Design.* Alan R.Liss, New York, pp. 207–219.

Ebright,R.H., Cossart,P., Gicquel-Sanzey,B. and Beckwith,J. (1984a) *Nature*, **311**, 232–235.

Ebright,R.H., Cossart,P., Gicquel-Sanzey,B. and Beckwith,J. (1984b) *Proc. Natl. Acad. Sci. USA*, **81**, 7274–7278.

Feldmann,R.J., Brooks,B.R. and Lee,B. (1986) Tools for each age: understanding protein architecture through simulated unfolding. Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD.

Gent,M.E., Gronenborn,A.M., Davies,R.W. and Clore,G.M. (1987) *Biochem. J.*, **242**, 645–653.

Giancotti,V., Russo,E., Cosimi,S., Cary,P.D. and Crane-Robinson,C. (1981) *Biochem. J.*, **197**, 655–660.

Hartman,P.G., Chapman,G.E., Moss,T. and Bradbury,E.M. (1977) *Eur. J. Biochem.*, **77**, 45–51.

Havel,T.F. (1986) DISGEO, Quantum Chemistry Program Exchange Program no. 507, Indiana University.

Havel,T.F. and Wüthrich,K. (1984) *Bull. Math. Biol.*, **46**, 673–698.

Havel,T.F. and Wüthrich,K. (1985) *J. Mol. Biol.*, **182**, 281–294.

Havel,T.F., Kuntz,I.D. and Crippen,G.N. (1983) *Bull. Math. Biol.*, **45**, 665–720.

Jenner,J., Meier,B.H., Bachmann,P. and Ernst,R.R. (1979) *J. Chem. Phys.*, **71**, 4546–4553.

Jones,T.A. (1978) *J. Appl. Crystallogr.*, **11**, 268–272.

Kaptein,R., Zuiderweg,E.R.P., Scheek,R.M., Boelens,R. and van Gunsteren,W.F. (1985) *J. Mol. Biol.*, **182**, 179–182.

Losa,R., Thoma,F. and Koller,T. (1984) *J. Mol. Biol.*, **175**, 529–551.

Levitt,M. (1983) *J. Mol. Biol.*, **170**, 723–764.

McGhee,J.B. and Felsenfeld,G. (1980) *Annu. Rev. Biochem.*, **49**, 1115–1156.

McKay,D.B. and Steitz,T.A. (1981) *Nature*, **290**, 744–749.

Macura,S., Huang,Y., Suter,D. and Ernst,R.R. (1981) *J. Magnetic Resonance*, **43**, 259–281.

Marion,D. and Wüthrich,K. (1983) *Biochem. Biophys. Res. Commun.*, **113**, 967–974.

Neuhaus,D., Wagner,G., Vasak,M., Kagi,J.H.R. and Wüthrich,K. (1985) *Eur. J. Biochem.*, **151**, 257–273.

Nilsson,L., Clore,G.M., Gronenborn,A.M., Brünger,A.T. and Karplus,M. (1986) *J. Mol. Biol.*, **188**, 455–475.

Papiz,M.Z., Sawyer,L., Elipoulos,E.E., North,A.C.T., Findlay,J.B.C., Sivaprasadaro,R., Jones,T.A., Newcomer,M.E. and Kraulis,P.J. (1986) *Nature*, **323**, 393–385.

Thomas,J.O. and Wilson,C.M. (1986) *EMBO J.*, **5**, 3531–3537.

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

Wüthrich,K., Billeter,M. and Braun,W. (1983) *J. Mol. Biol.*, **160**, 949–961.

Wüthrich,K., Billeter,M. and Braun,W. (1984) *J. Mol. Biol.*, **180**, 715–740.

Yaguchi,M., Roy,C., Dove,M. and Seligy,V. (1977) *Biochem. Biophys. Res. Commun.*, **76**, 100–106.

Yaguchi,M., Roy,C. and Seligy,V.L. (1979) *Biochem. Biophys. Res. Commun.*, **90**, 1400–1406.

Zarbock,J., Clore,G.M. and Gronenborn,A.M. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7628–7632.

Received on February 17, 1987; revised on March 18, 1987