The solution structure of a specific GAGA factor–DNA complex reveals a modular binding mode

James G. Omichinski[1], Paolo V. Pedone[2], Gary Felsenfeld[2], Angela M. Gronenborn[1] and G. Marius Clore[1]

The structure of a complex between the DNA binding domain of the GAGA factor (GAGA-DBD) and an oligonucleotide containing its GAGAG consensus binding site has been determined by nuclear magnetic resonance spectroscopy. The GAGA-DBD comprises a single classical Cys_2–His_2 zinc finger core, and an N-terminal extension containing two highly basic regions, BR1 and BR2. The zinc finger core binds in the major groove and recognizes the first three GAG bases of the consensus in a manner similar to that seen in other classical zinc finger–DNA complexes. Unlike the latter, which require tandem zinc finger repeats with a minimum of two units for high affinity binding, the GAGA-DBD makes use of only a single finger complemented by BR1 and BR2. BR2 forms a helix that interacts in the major groove recognizing the last G of the consensus, while BR1 wraps around the DNA in the minor groove and recognizes the A in the fourth position of the consensus. The implications of the structure of the GAGA-DBD–DNA complex for chromatin remodelling are discussed.

The GAGA factor of *Drosophila melanogaster* is a TFIIIA-like zinc finger protein which was originally identified on the basis of its ability to bind to $(GA)_n$ rich sites in the *Ultrabithorax* promoter[1,2]. It has recently been shown that the GAGA factor is encoded by the *Thrithorax-like* gene[3], is required for normal expression of homeotic genes, and acts as a modifier of position-effect variegation (see ref. 4 for a review). Both *in vivo*[5,6] and *in vitro*[7–11] experiments suggest that the GAGA factor acts as an anti-repressor by helping to disrupt nucleosomes associated with gene regulatory sequences. Putative *Drosophila* target genes for the GAGA factor now include genes such as the heat shock (hsp26 and hsp70), histone h3/h4, homeotic and housekeeping/constitutive genes[4]. Although it is clear that the GAGA factor binds to $(GA)_n$ rich sites, the minimal DNA binding site for the protein is not clearly established and the length of the GA repeat is highly variable[4].

The GAGA factor is 519 residues in length and comprises three domains: an N-terminal POZ/PTB protein interaction domain, a central DNA binding domain and a polyglutamine-rich carboxy terminus[4]. The minimal DNA binding domain (DBD) of the GAGA factor has recently been delineated and shown to bind specifically to DNA derived from the h3/h4 promoter and containing the sequence GAGAGAG with a dissociation constant of ~5nM[12]. The GAGA-DBD comprises residues 310–372 and consists of a single classical (TFIIIA-like) Cys_2-His_2 zinc finger[13,14] preceded by two highly basic regions (BR1 and BR2) (Fig. 1). (Note that there is also a basic region C-terminal to the zinc finger domain, but its presence or absence has no effect on DNA binding affinity[12]). While the genes for several proteins containing only a single classical zinc finger have now been identified[12,15,16], the GAGA-DBD is unique in that it is, so far, the only single classical zinc finger containing protein for which DNA binding by the zinc finger region has been experimentally demonstrated. In all other reported TFIIIA-like zinc finger proteins, a tandem array comprising a minimum of two zinc binding domains is required for sequence specific high-affinity DNA binding[13,14]. In this paper we define the minimal DNA binding site for the GAGA-DBD and present the determination of the three-dimensional structure of a complex of the GAGA-DBD with DNA using multidimensional NMR spectroscopy. The structure shows how a classical zinc finger complemented by an N-terminal extension comprising a basic helix and tail can recognize DNA in a sequence specific manner, making base specific contacts with every base in the pentanucleotide consensus sequence GAGAG. In addition, the presence of both major and minor groove contacts in the complex immediately suggests special constraints on interactions of the GAGA factor and DNA targets within nucleosomes.

Structure determination

We have solved the structure of the complex of the DNA binding domain of the GAGA factor (Fig. 1) with DNA by means of multidimensional double and triple resonance heteronuclear-filtered and heteronuclear-edited NMR spectroscopy[17–20]. The identified DNA binding sites for the GAGA factor vary considerably in length and sequence, but the majority contain the consensus GAGAGAG. In the first complex studied, the 16 bp oligonucleotide employed was derived from the *Drosophila* h3/h4 promoter and contained the seven base pair consensus (Fig. 1*a*). This complex, which exhibited slow exchange on the chemical shift time scale, yielded good spectra for the GAGA-DBD and numerous intermolecular NOEs. Careful analysis of the spectrum of the bound DNA, however, suggested the presence of two overlapping GAGA-DBD binding sites on this oligonucleotide, with each site comprising a minimal pentanu-

Laboratories of [1]Chemical Physics and [2]Molecular Biology, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA

Correspondence should be addressed to G.M.C. and A.M.G. *email: clore@vger.niddk.nih.gov* and *gronenborn@vger.niddk.nih.gov*

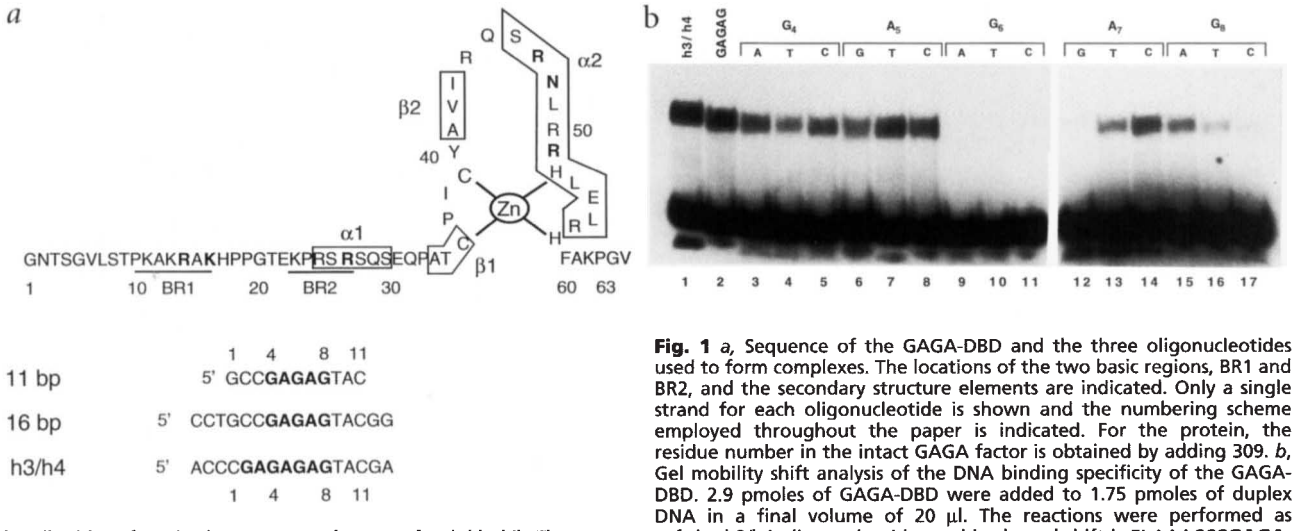

Fig. 1 *a*, Sequence of the GAGA-DBD and the three oligonucleotides used to form complexes. The locations of the two basic regions, BR1 and BR2, and the secondary structure elements are indicated. Only a single strand for each oligonucleotide is shown and the numbering scheme employed throughout the paper is indicated. For the protein, the residue number in the intact GAGA factor is obtained by adding 309. *b*, Gel mobility shift analysis of the DNA binding specificity of the GAGA-DBD. 2.9 pmoles of GAGA-DBD were added to 1.75 pmoles of duplex DNA in a final volume of 20 μl. The reactions were performed as described in ref. 12 in the presence of 50 ng of poly(dI-dC). The sequence of the h3/h4 oligonucleotide used in the gel shift is 5'-AAACCC**GAGAGAG**TACGAAC, slightly larger than the oligonucleotide containing the same motif used in the NMR study. The sequence of the GAGAG oligonucleotide is 5'-AAAACCC**G$_4$A$_5$G$_6$A$_7$G$_8$**TACGAACG. The sequences of the other oligonucleotides involve replacing each base in turn from G4 to G8 with the three other bases.

cleotide GAGAG sequence. The resulting complexity, arising from small differences in chemical shifts between the two potential complexes made the interpretation of the bound DNA spectrum difficult. This observation was confirmed by gel retardation assays where no significant difference in binding affinity could be detected for oligonucleotides containing GAGAGAG or GAGAG sites (Fig. 1*b*, lanes 1 and 2). The DNA spectra for the complexes of the GAGA-DBD with the 16 bp and 11 bp oligonucleotides containing the minimal GAGAG sequence (Fig. 1*a*) were interpretable and all subsequent analysis was carried out on these two complexes, which also exhibited in slow exchange on the chemical shift time scale. While the observed intra- and intermolecular NOEs were essentially identical, we found it useful to interpret spectra from both the 11 bp and 16 bp complexes, since ambiguities could be resolved by taking advantage of small differences in chemical shifts. An example of the quality of the NMR data illustrating intermolecular contacts between the GAGA-DBD and DNA is shown in Fig. 2*a*.

It will be seen from the data in Fig. 2*a* that the resonances of the H4', H5' and H5" sugar protons exhibit extensive overlap. In nearly all cases we were able to differentiate the H4' resonances from the H5' and H5" resonances by the combined analysis of the 2D ^{12}C-filtered HOHAHA (through-bond correlation) and NOE (through-space correlation) spectra. We did not, however, attempt to stereoassign the H5' and H5" resonances, with the single exception of those for C3, where the H5" resonance was significantly downfield shifted (see below). Any ambiguities in the assignment of NOEs involving the H4', H5' and H5" sugar protons were treated as $\Sigma(r^{-6})^{-1/6}$ sums[21]. As the structure determination proceeded through iterative refinement, the number of such ambiguities decreased. All NOEs, however, involving the H5' and H5" sugar protons, with the exception of those of C3, were treated as $\Sigma(r^{-6})^{-1/6}$ sums in the final round of structural calculations.

The structures were calculated by simulated annealing[22,23]. The target function that is minimized during simulated annealing comprises only quadratic harmonic potential terms for covalent geometry, base pair planarity, $^3J_{HN\alpha}$ coupling constant[24], and secondary $^{13}C\alpha$ and $^{13}C\beta$ chemical shift[25] restraints; square-well quadratic potentials for the experimental distance and torsion angle restraints; and a quartic van der Waals repulsion term together with a conformational database potential[26] for the non-bonded contacts. The latter is based on the populations of various combinations of torsion angles observed in a database of 70 high-resolution (1.75 Å or better) protein X-ray structures[27] and biases sampling to conformations that are energetically possible by effectively limiting the choice of dihedral angles to those that are known to be

Table 1 Summary of distance restraints employed for the introduction of interfacial protein-base hydrogen bonds

Protein	DNA
Negative distance restraints (lower bound 4.0 Å and upper bound unrestrained)[1]	
Arg 51 Nη1, Nη2, Nε	N4 of C3 and C19; N6 of A5
Arg 47 Nη1, Nη2, Nε	N4 of C19; N6 of A3 and A5
Arg 27 Nη1, Nη2, Nε	N4 of C15; N6 of A7 and A14
Arg 14 Nη1, Nη2, Nε	N2 of G6 and G8
Asn 48 Nδ2	N4 of C17 and C19; N6 of A5
Asn 48 Oδ1	N7 of G4, A5 and G6; O4 of C18
$(\Sigma r^{-6})^{-1/6}$ ambiguous hydrogen bonding restraints (distance range 2.4–3.3 Å)[1]	
Arg 51 Nη1	O6 of G4 and G20; N7 of G4, A5 and G20; and O4 of T18
Arg 51 Nη2	O6 of G4 and G20; N7 of G4, A5 and G20; and O4 of T18
Arg 47 Nη1	O6 of G6; N7 of A5, G6 and A7; and O4 of T16 and T18
Arg 47 Nη2	O6 of G6; N7 of A5, G6 and A7; and O4 of T16 and T18
Arg 27 Nη1	O6 of G8; N7 of A7, G8 and A14; and O4 of T9 and T16
Arg 27 Nη1	O6 of G8; N7 of A7, G8 and A14; and O4 of T9 and T16
Arg 14 Nη1	N3 of G6, A7 and G8; and O2 of C15, T16 and C17
Arg 14 Nη1	N3 of G6, A7 and G8; and O2 of C15, T16 and C17
Asn 48 Nδ2	N7 of G4, A5 and G6; and O6 of G4 and G6
Asn 48 Oδ1	N6 of A5; and N4 of C17 and C19

[1]Essentially the same results are obtained using a lower bound of 3.5 Å for the 'repulsive' distance restraints and a range of 2.4–3.5 Å for the $(\Sigma r^{-6})^{-1/6}$ ambiguous hydrogen bonding restraints.

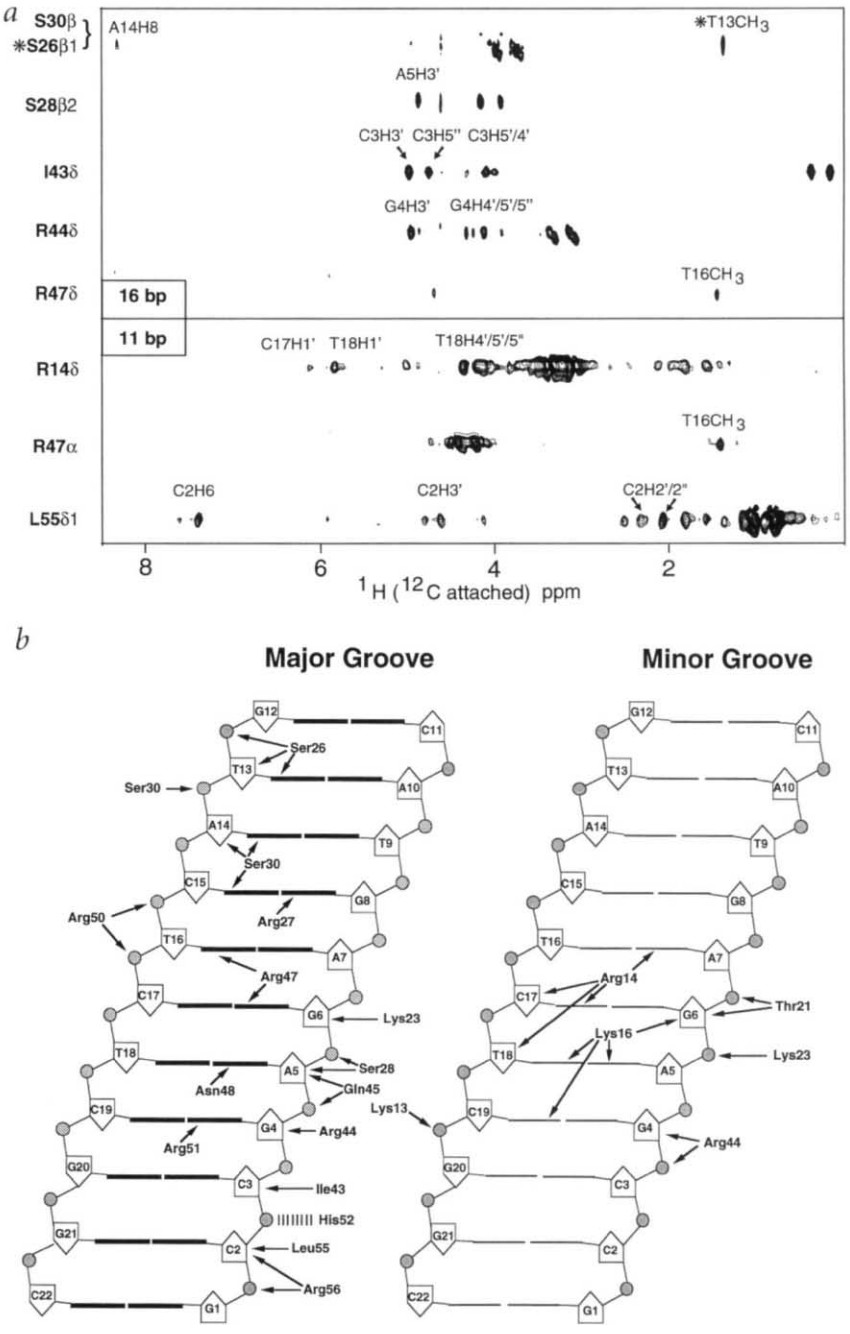

Fig. 2 *a*, A composite of ^{13}C-H strips taken from 3D ^{13}C-separated/^{12}C-filtered NOE spectra (mixing times ~150 ms) recorded at 39.5 °C on the GAGA-DBD complexes with the 16mer (top) and 11mer (bottom) oligonucleotides, illustrating NOEs between protons of the protein (attached to ^{13}C) and protons of the DNA (attached to ^{12}C). The spectrum on the 11mer complex was recorded using a single ^{13}C purge pulse, while that on the 16mer complex used two ^{13}C purge pulses and hence has a higher degree of ^{12}C-filtering. (Note Hβ1 of Ser 26 overlaps with Hβ1/Hβ2 of Ser 30, and the asterisk indicates NOEs involving Hβ1 of Ser 26). *b*, Summary of the contacts between the GAGA-DBD and DNA. The DNA is represented as a cylindrical projection viewed from either the major or minor grooves as indicated. The bases are indicated as thick lines in the major groove and thin lines in the minor groove, with the deoxyribose sugar rings as pentagons and the phosphates as stippled circles. The hydrogen bond between the imidazole ring of His 52 and the phosphate of C3 is shown as a broken line.

physically realizable[26]. No classical empirical energy terms for electrostatic or hydrogen bonding interactions[28] are present in the target function.

In the very final stage of the calculations intermolecular hydrogen bonding restraints between protein side chains of Arg 14, Arg 27, Arg 47, Asn 48 and Arg 51 to the DNA bases were added (Table 1). The side chains of these five residues are buried at the protein–DNA interface and interact with the DNA bases. Since the target function for simulated annealing, however, contains no electrostatic or hydrogen bonding potential terms, and since no intermolecular NOEs involving the NεH (with the exception of Arg 14) and guanidino protons of the arginines, or the side-chain amide of Asn 48 were observed (see below), the exact location of the donor and acceptor groups could not be determined *a priori*. We therefore decided to introduce interfacial hydrogen bonding restraints between protein side chains and DNA bases with the rationale that the energetic cost of burying a polar group with unsatisfied hydrogen bonds is very high[29–31]. In addition, this energetic cost is further increased if donor groups are inappropriately close to other donor groups, and likewise for acceptor groups. Prior to the introduction of interfacial protein–base hydrogen bonding restraints, the χ_1 and χ_2 rotamers of Arg 14, Arg 27, Arg 47 and Arg 51, and the χ_1 rotamer of Asn 48 were defined by the experimental NMR data in combination with the conformational database potential[26] (Fig. 3). In addition, it was evident that Arg 14 was centred around base pairs 5, 6 and 7 in the minor groove, while Arg 27, Arg 47, Asn 48 and Arg 51 were centred about base pairs 8, 6, 5 and 4 respectively in the major groove.

Evidence that the functional groups of the interacting arginines are either buried at the interface or involved in some sort of electrostatic interaction comes from the observation of the NεH and guanidino protons in the complex. Nε-NεH cross peaks for all the arginines (with the exception of Arg 25 which does not interact with the DNA) are observed in the ^{1}H-^{15}N correlation spectrum of the complex with ^{1}H and ^{15}N shifts in the range 7.1–8.2 and 82–86 p.p.m., respectively. We were also able to observe four broad, partially overlapping cross peaks in the region from 6.5–7.0 p.p.m. in the ^{1}H dimension and 71–73 p.p.m. in the ^{15}N dimension which arise from the guanidino groups of the arginines[32]. In addition, weak NOEs (due to the broad linewidths of the guanidino proton resonances) between the NεH and guanidino protons were observed in the 3D ^{15}N-separated NOE spectrum. In contrast, only a single Nε-NεH cross peak (with ^{1}H and ^{15}N chemical shifts of 7.51 and 84.6 p.p.m., respectively, similar to those for a free arginine) and no guanidino cross peaks were observed in the ^{1}H-^{15}N correlation spectrum of the free GAGA-DBD. Line broadening of the guanidino cross peaks is mainly due to rotation about the Nε-

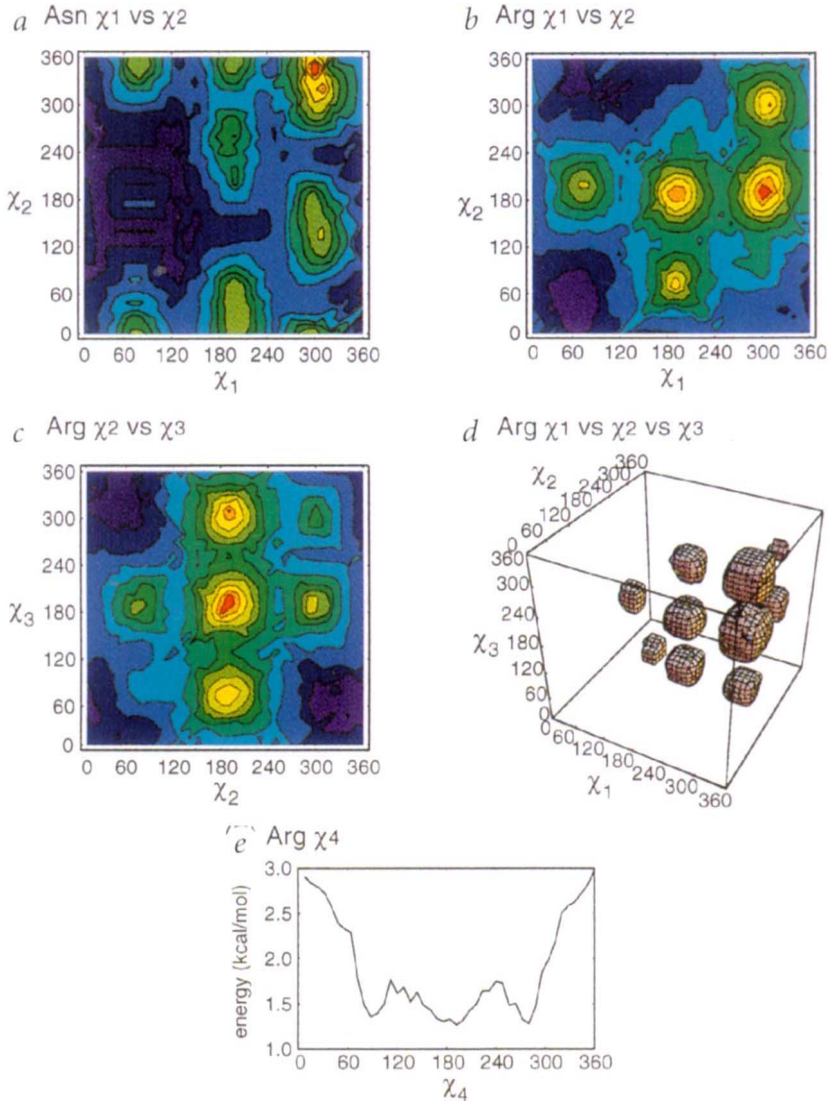

Fig. 3 The conformational database potential energy surface for the side chains of Asn and Arg. Two-dimensional potential energy surfaces for *a*, χ_1 *versus* χ_2 of Asn, *b*, χ_1 *versus* χ_2 of Arg, *c*, χ_2 *versus* χ_3 of Arg. *d*, Three-dimensional potential energy surface for χ_1 *versus* χ_2 *versus* χ_3 of Arg. *e*, One-dimensional potential energy for χ^4 of Arg. The contours in (*a*), (*b*) and (*c*) are colour-coded from red (minimum energy) to violet (maximum energy). (*d*) is contoured at a level 40% higher than the minimum energy (that is, $0.4[E_{max}-E_{min}] + E_{min}$). The conformational database potential energy surfaces were derived as described in ref. 26 using a database of 70 high resolution (1.75 Å or better) protein crystal structures[27]. The potential of mean force at any grid position (10°, 10°×10°, and 10°×10°×10° for the one-, two-, and three-dimensional potential energy surfaces respectively) is given by the negative logarithm of the probability of occurence of a particular conformation times a scale factor[26]. The value of the scale factor used to generate the potential energy surfaces in this figure, as well as the final value of the scale factor employed for the conformational database potential in the simulated annealing calculations is 1.0. The values for the minimum and maximum energies in the potential energy surfaces are 3.3 and 8.7 kcal mol^{-1}, respectively, in (*a*), 2.8 and 10.7 kcal mol^{-1}, respectively, in (*b*), 3.33 and 9.6 kcal mol^{-1}, respectively, in (*c*), and 5.8 and 13.1 kcal mol^{-1}, respectively in (*d*).

Cζ partial double bond[32], as well as flipping about the Cζ-Nη bonds. Chemical exchange line broadening was ascertained by observing an increase in the intensity of the guanidino cross peaks in a ^{1}H-^{15}N HSQC spectrum recorded with CPMG-derived pulse trains during the INEPT transfer steps to reduce loss of dephasing of spin coherence due to chemical exchange[33,34]. A similar chemical exchange mechanism involving rotation about a C-N bond is responsible for broadening of the proton resonances of the 6-NH_2 and 2-NH_2 groups of adenine and guanine[33], respectively, both of which are involved in Watson–Crick base pairing. In addition, the Nδ2-Hδ2' and Nδ2-Hδ2" cross-peaks for Asn 48 are also broad, impeding the observation of any intermolecular NOEs from the Hδ2' and Hδ2" protons of Asn 48 in the 3D ^{15}N-separated NOE spectrum. Line broadening of side-chain amides involved in hydrogen bonding with DNA bases has been previously observed in a complex of *lac* repressor headpiece with DNA[34,35].

The calculational strategy employed for the interfacial protein–base hydrogen bonding restraints involved two types of distance restraints: (i) 'repulsive' distance restraints with a lower bound of 4 Å (and an unrestrained upper bound) to prevent donor groups of the side chains of Arg 14, Arg 27, Arg 47, Arg 51 and Asn 48 from coming closer than 4 Å to donor groups of the DNA bases, and similarly to prevent the Oδ1 acceptor group of Asn 48 from coming closer than 4 Å to acceptor groups of the DNA bases; and (ii) ambiguous distance restraints (with a range of 2.4–3.3 Å), represented by an $(\Sigma r^{-6})^{-1/6}$ sum[21], between donor groups of these five side chains and all possible acceptor groups of the DNA bases, and between the Oδ1 acceptor group of Asn 48 and all possible donor groups of the DNA bases. These distance restraints are summarized in Table 1. The effect of such ambiguous $(\Sigma r^{-6})^{-1/6}$ distance restraints is to permit a hydrogen bond from a specified group to be formed to whichever of the listed potential partners is closest in the evolving calculated structures, the restraint being satisfied providing only that at least one of these potential partners is close.

The introduction of the 'repulsive' distance restraints alone, indicated that the side chain carboxyamide of Asn 48 interacts with the N7 and N6 atoms of A5, the guanidino group of Arg 14 with the O2 atom of C17, the guanidino group of Arg 27 with the N7 and O6 atoms of G8, the guanidino group of Arg 47 with the O6 atom of G6, and the guanidino group of Arg 51 with the N7 and O6 atoms of G4. Following the introduction of the ambiguous $(\Sigma r^{-6})^{-1/6}$ hydrogen bonding restraints we were able to ascertain some additional electrostatic interactions, in particular between the guanidino group of Arg 14 and the N3 atom of A5, and possibly also between the guanidino group of Arg 47 and the N7 atom of G6.

Asn 48 provides a simple example of the effects of introducing the 'repulsive' interfacial distance restraints. For χ_1 in the –60° rotamer, the χ_2 angle of Asn has two conformational minima centred around –30±30° and 150±30° which are the result of using the conformational database potential[26] in the target function that is minimized during simulated annealing (Fig. 3). In the absence of the 'repulsive' distance restraints the two rotamers occur with equal probability. In the presence of the 'repulsive' distance restraints, however, only the χ_2=–30±30° rotamer is populated. For the

χ_2=–30±30° conformer appropriate hydrogen bonding between the side-chain amide group of Asn 48 and the N7 atom of A5 and between the Oδ1 atom of Asn 48 and the 6-NH_2 group of A5 is observed. For the χ_2=150±30° conformer, on the other hand, two hydrogen bond acceptors are opposed (the Oδ1 atom of Asn 48 and the N7 atom of A5), and similarly two hydrogen bond donors are in very close proximity (the Nδ2 atom of Asn 48 and the N6 atom of A5). Such an arrangement, which is excluded by the use of the 'repulsive' distance restraints, is clearly very unlikely based on chemical reasoning, and would also not be able to account for the observation that substitution of Asn by Asp at position 48 reduces DNA binding by greater than ten-fold (data not shown).

It should be noted that the introduction of the interfacial hydrogen bonding restraints had no effect on the overall structure of the complex but improved local convergence and precision for the Nε and guanidino groups of Arg 51, Arg 47 and Arg 27, the guanidino group of Arg 14, and the carboxyamide of Asn 48. Thus, excluding the latter atoms, the precision of the coordinates (protein backbone, DNA and side chains) remains unchanged and the atomic r.m.s. shift in the mean coordinate positions obtained with and without the interfacial hydrogen bonding restraints is only 0.23 Å for all atoms, which is well within the errors of the coordinates. (The atomic r.m.s. shift for all atoms including the complete side chains of Arg 14, Arg 27, Arg 47, Asn 48 and Arg 51 is only 0.26 Å).

The final ensemble of structures of the GAGA-DBD–DNA complex was determined on the basis of 1475 experimental NMR restraints, in conjunction with the hydrogen bonding restraints discussed above and listed in Table 1. A summary of the structural statistics is given in Table 2, a summary of the distribution of NOE restraints is displayed in Fig. 4, and a superposition of the 50 final simulated annealing structures is shown in Fig. 5. Residues 1–13 and 60–62 at the N and C termini, respectively, appear to be disordered in solution since only intraresidue and sequential interresidue NOEs are observed. Excluding these residues, the precision of the coordinates is 0.51±0.06 Å for the protein backbone (N, Cα, C, O atoms) plus the DNA, and 0.72±0.07 Å for all protein atoms plus the DNA. The precision of the backbone coordinates of the N-terminal tail (residues 14–23, 0.88±0.23 Å) is a factor of two worse than that for the rest of the protein (residues 24–58, 0.42±0.11 Å). This is due to the fact that with the exception of two NOEs between the CαH protons of Gly 20 and the NH proton of Glu 22, the remaining intraprotein NOEs are either sequential or intraresidue. The fact that the conformation of the N-terminal tail can be determined at all is due to three factors: (i) the sizeable number of intermolecular NOEs involving the N-terminal tail

Table 2 Structural statistics[1]

	<SA>	($\overline{SA}$)r
R.m.s. deviations from NOE interproton distance restraints (Å)[2]		
All (1024)	0.024±0.002	0.028
Protein		
interresidue sequential (|i– j| = 1) (188)	0.021±0.004	0.034
interresidue short range (1 < |i - j| ≤5) (134)	0.022±0.006	0.023
interresidue long range (|i - j|) > 5) (95)	0.026±0.005	0.029
intraresidue (275)	0.014±0.004	0.020
DNA		
intraresidue (124)	0.014±0.005	0.010
sequential intrastrand (112)	0.035±0.005	0.037
interstrand (21)	0.024±0.010	0.010
Protein–DNA (75)	0.041±0.015	0.040
R.m.s. deviations from hydrogen bonding restraints (Å)		
Protein (24)[3]	0.025±0.009	0.009
DNA (58)[3]	0.019±0.006	0.018
Protein–DNA (10)[4]	0.049±0.013	0.048
R.m.s. deviations from expt.		
dihedral restraints (deg.) (242)[2]	0.428±0.074	0.516
R.m.s. deviations from expt.		
$^3J_{HN\alpha}$ coupling constants (Hz) (33)[2]	0.77±0.06	0.79
R.m.s. deviations from expt ^{13}C shifts		
$^{13}C\alpha$ (p.p.m.) (48)	1.06±0.05	1.03
$^{13}C\beta$ (p.p.m.) (46)	0.99±0.05	0.95
Deviations from idealized covalent geometry		
bonds (Å) (1645)	0.005±0.0002	0.005
angles (deg.) (2992)	1.019±0.008	1.046
impropers (deg.) (846)[5]	0.488±0.047	0.641
E_{L-J} (kcal mol^{-1})[6]	–523±8	–521
Coordinate precision (Å)[7]		
protein backbone plus DNA	0.51±0.06	
All protein atoms plus DNA	0.72±0.07	
Protein backbone	0.44±0.08	
All protein atoms	0.87±0.10	
DNA	0.45±0.07	

[1]The notation of the NMR structures is as follows: <SA> are the final 50 simulated annealing structures; $\overline{SA}$ is the mean structure obtained by averaging the coordinates of the individual SA structures best fitted to each other (with respect to residues 14–58 of the protein and base pairs 1–11 of the DNA); ($\overline{SA}$)r is the restrained regularized mean structure obtained by restrained regularization of the mean structure $\overline{SA}$. The number of terms for the various restraints is given in parentheses. The average PROCHECK[63] values for the percentage of residues in the most favourable region of the Ramachandran plot, the number of bad contacts per 100 residues, the hydrogen bonding energy and the dihedral angle G factor are 82±3%, 9.5±1.8, 0.88±.08, 0.08±0.04 and 0.88±0.04 respectively.

[2]None of the structures exhibited distance violations greater than 0.5 Å, dihedral angle violations greater than 5°, or $^3J_{HN\alpha}$ coupling constant violations greater than 2 Hz. The torsion angle restraints for the whole complex comprise 140 and 102 torsion angles for the protein and DNA, respectively.

[3]Two distance restraints were employed for each backbone hydrogen bond. The hydrogen bond restraints within the DNA were used to maintain Watson–Crick base pairing[40,57].

[4]Intermolecular hydrogen bonding restraints between protein side chains of Arg 14, Arg 27, Arg 47, Asn 48 and Arg 51 to the DNA bases were only added in the very final stage of the calculations (see Table 1 and text). Only the statistics for the attractive $(\Sigma r^{-6})^{-1/6}$ ambiguous hydrogen bonding distances are quoted in the Table, since there are no violations of the 'repulsive' distance restraints.

[5]The improper torsion restraints serve to maintain planarity and chirality.

[6]E_{L-J} is the Lennard–Jones van der Waals energy calculated with the CHARMM PARAM19/20 protein and PARNAH1ER1 DNA parameters[28] and is not included in the target function for simulated annealing or restrained minimization.

[7]The precision of the coordinates is defined as the average atomic r.m.s. difference between the 50 individual simulated annealing structures and the mean coordinates SA. The values refer to residues 14–58 of the protein, the zinc atom and basepairs 1–11 of the DNA.

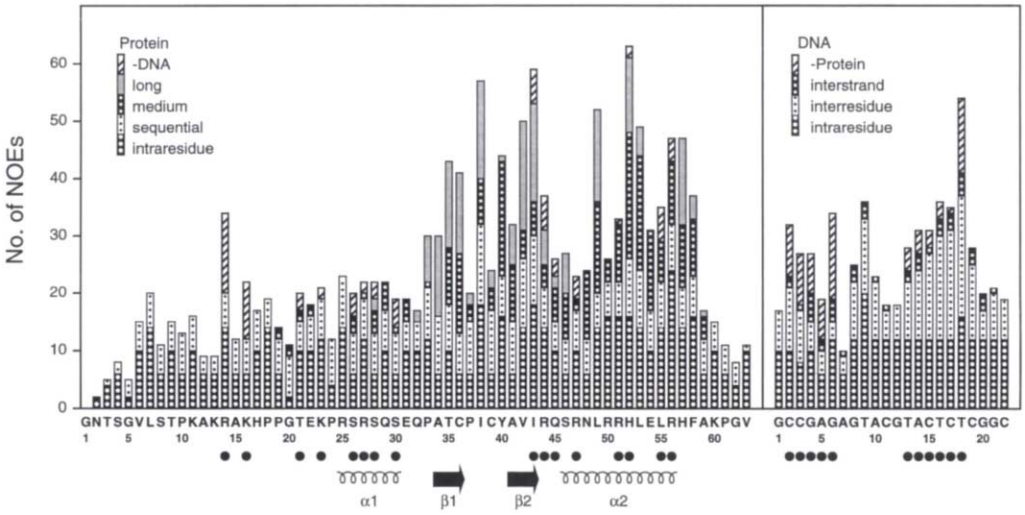

Fig. 4 Summary of the distribution of observed NOEs for the GAGA-DBD–DNA complex. Residues of the protein and nucleotides of the DNA for which intermolecular NOEs were observed are indicated by filled-in circles below the figure. Also indicated, below the figure, is the location of the secondary structure elements.

(specifically 14 for Arg 14, 10 for Lys 16, 3 for Thr 21, and 2 for Lys 23); (ii) direct refinement against $^{3}J_{HN\alpha}$ couplings constants and $^{13}C\alpha$ and $^{13}C\beta$ secondary chemical shifts[24,25]; and (iii) the use of the conformational database potential[26].

Description of the structure

The key feature of the GAGA-DBD–DNA complex is the presence of base specific contacts to every base of the G_4AGAG_8 pentanucleotide consensus (Figs 2*b*, 6, 7). Major groove recognition of G4, A5 and G6 is provided by the helix of the zinc finger, major groove recognition of G8 by the helix of BR2, and minor groove recognition of A7 by BR1.

The DNA in the complex is essentially B-type and displays a smooth bend of ~10° (Fig. 6*c*). A best fit superposition of the DNA in the complex onto classical B- and A-DNA yields atomic r.m.s. differences of 1.8 and 5.7 Å, respectively. The average values for the local helical twist and rise are ~35.5° and ~3.6 Å, respectively. The propeller twist, local inter-base pair tilt angles and local inter-base pair roll angles vary from ~0° to ~–20°, ~+3° to ~–6°, and ~0° to ~+15°, respectively, with average values of ~–10°, ~0° and ~0°, respectively.

The zinc finger core (residues 33–59) of the GAGA-DBD is centred around a tetrahedrally coordinated zinc atom, comprises a ββα motif (Figs 6 and 7*a*), and is very similar to that of other classical zinc fingers with Cα atomic r.m.s. differences ranging from 0.8 Å relative to finger 2 of Tramtrack[36] to 1.6 Å relative to finger 3 of Zif268[37]. Arg 51, Asn 48 and Arg 47 at positions 6, 3 and 2 of the zinc finger helix (α2, residues 46–56) contact the bases of G4, A5 and G6, respectively. The long axis of the helix α2 is oriented at approximately 60° to the long axis of the DNA. The guanidino group of Arg 51 recognizes the O6 and N7 atoms of G4, the guanidino group of Arg 47 recognizes the O6 atom of G6 and possibly the N7 atom of G6, and the side chain carboxyamide group of the Asn 48 recognizes the 6-NH_2 group and N7 atom of A5 (Fig. 6*a*). The Cδ of Arg 47 is also in close contact with the methyl group of T16 (Fig. 2*a*). The base specific interactions are supplemented by hydrophobic interactions with the sugars involving Arg 56, Leu 55, Arg 44 and Ile 43, and electrostatic interactions with the sugar phosphate backbone involving Arg 56, His 52, Arg 50, Arg 44 and Gln 45. As in other classical zing finger–DNA complexes[36–38] the Hδ1 proton of the imidazole ring of His 52 is hydrogen-bonded to the phosphate of C3. (The distance between the Hδ1 atom of His 52 and the closest oxygen atom of the phosphate group of C3 is ~1.8 Å). As a result the Hδ1 resonance is visible in the ^{1}H NMR spectrum at ~14.5 p.p.m., and this interaction is probably responsible for the ~0.7 p.p.m. downfield shift in the H5" resonance of C3. BR2 forms a short helix (α1, residues 25–29) which is located in the major groove and oriented at ~80° to the long axis of the DNA. The guanidino group of Arg 27 of the BR2 helix recognizes the base of G8, while Ser 28, Ser 30 and Ser 26 anchor this helix to the phosphates of G6, A14 and T13, respectively (Fig. 7*b*). The orientation of the BR2 helix with respect to the zinc finger core is further stabilized by an electrostatic interaction between Ser 46 and Glu 31 (Fig. 7*a*). In the minor groove, base specific contacts are made by Arg 14 and Lys 16 of BR1. The guanidino group of Arg 14 interacts with the N3 and O2 atoms of A7 and C17, respectively, while Lys 16 interacts with the O2 atoms of T18 and/or C19 (Fig. 7*c*). From an examination of the ensemble of structures, it is also

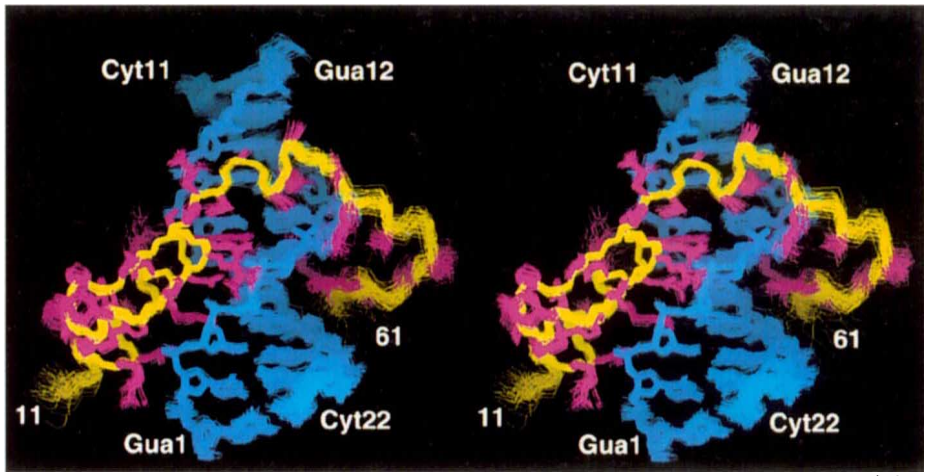

Fig. 5 Stereoview showing a best fit superposition of the 50 simulated annealing structures of the GAGA-DBD–DNA. The backbone (N, Cα, C) atoms of the GAGA-DBD are shown in yellow, side chains in magenta and the DNA in blue.

possible that Lys 13 may interact with the phosphate of G20. However, we were unable to unambiguously identify intermolecular NOEs involving Lys 13. From the structure, it is also apparent that only part of the full BR1 region (residues 11–16), as defined by Pedone *et al.*[12], is required for binding. Thus, the first two residues of BR1, Lys 11 and Ala 12, do not make any contacts with the DNA in the present structure.

Relationship to other zinc finger–DNA complexes

To date, X-ray structures of three classical zinc finger–DNA complexes have been solved: Tramtrack[36], Zif268[37] and GLI[38] which consist of a tandem array of two, three and five zinc finger repeats, respectively, separated by seven amino acids. The two fingers of Tramtrack[36] and the three fingers of Zif268[37] make specific contacts with the DNA, and the orientation of each finger with respect to the DNA is very similar in all cases. For GLI[38], only four (fingers 2–5) of the five zinc fingers contact the DNA. The number of interactions between the zinc finger domains is small and each finger recognizes DNA in a largely independent manner[14]. Specific recognition is achieved by contacts between the helix of each zinc finger and the major groove of the DNA. The residues of the helix involved in specific DNA contacts are located principally at positions 6, 3, 2 or –1 (numbered relative to the N-terminus of the helix and corresponding to Arg 51, Asn 48, Arg 47, and Gln 45, respectively, of the GAGA-DBD; Fig. 1*a*), with usually two of the four positions making one-to-one amino acid-to-base contacts with a base triplet[13,14]. The crystal structures of these three zinc finger–DNA complexes together with extensive data from site-directed mutagenesis, screening/selection and protein design experiments has led to a generalized consensus zinc finger recognition code in which the residues at position 6, 3, and either 2 or –1 recognize the first, second and third base of the triplet (reading in the 5'→3' direction), respectively[13,14].

Recognition of the $G_4A_5G_6$ triplet by the zinc finger helix of the GAGA-DBD involves residues at positions 6 (Arg 51), 3 (Asn 48) and 2 (Arg 47). This mode of interaction is identical to Finger 1 of Tramtrack with the difference that the residue at position 2 (Ser) recognizes a T[36]. Indeed, the side-chain conformations and the nature of the contacts with the DNA bases for the Arg and Asn residues at positions 6 and 3 are essentially identical for the zinc finger of the GAGA-DBD and Finger 1 of Tramtrack[36] (Figs. 6*a* and 7*a*). It is also worth noting that the conformation of the Arg at position 6 in the GAGA-DBD, Finger 1 of Tramtrack[36], and Fingers 1 and 3 of Zif268[37] are similar and in each case the guanidino group recognizes the O6 and N7 atoms of a G base with essentially the same hydrogen bonding geometry.

The number of intermolecular contacts, and hence the strength of the interaction of each zinc finger with DNA, is not sufficient in its own right to yield high affinity DNA binding[13,14]. The latter can be achieved in a number of ways. The simplest design approach, as exemplified by the three zinc finger–DNA crystal structures, is to have a modular protein with a tandem array of two or more zinc fingers. Alternatively, a single zinc finger domain can be employed in conjunction with another DNA recognition motif. This is the path chosen by the GAGA-DBD where the additional contacts required for sequence specific binding are provided by an N-terminal extension comprising two basic regions, BR1 and BR2. In this light, it is of interest to compare the GAGA-DBD–DNA complex with the two zinc finger Tramtrack–DNA complex[36]. This is illustrated in Fig. 8 where the zinc finger core of the GAGA-DBD has

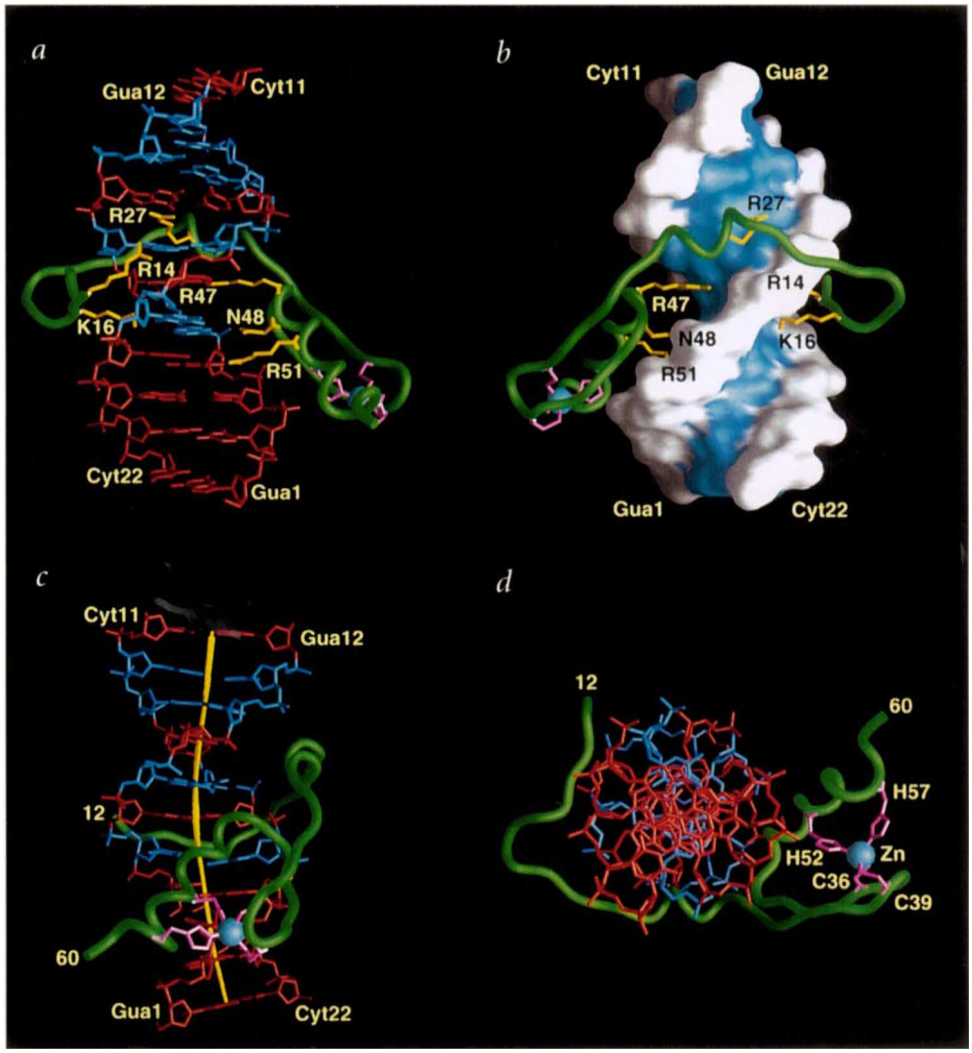

Fig. 6 Four views (*a*–*d*) illustrating the interaction of the GAGA-DBD with DNA. The backbone of the protein is depicted as a green tube, side chains making base specific contacts are in yellow, the histidine and cysteine side chains coordinating the zinc (blue sphere) in magenta, GC base pairs in red, and AT base pairs in blue. In *b* the DNA is depicted as a molecular surface with the bases coloured in blue and the sugar-phosphate backbone in white. *c*, The path of the long axis of the DNA helix, calculated with the program Curves[60] is shown in yellow. The structure shown is the restrained regularized mean structure. This figure was generated with the program GRASP[61].

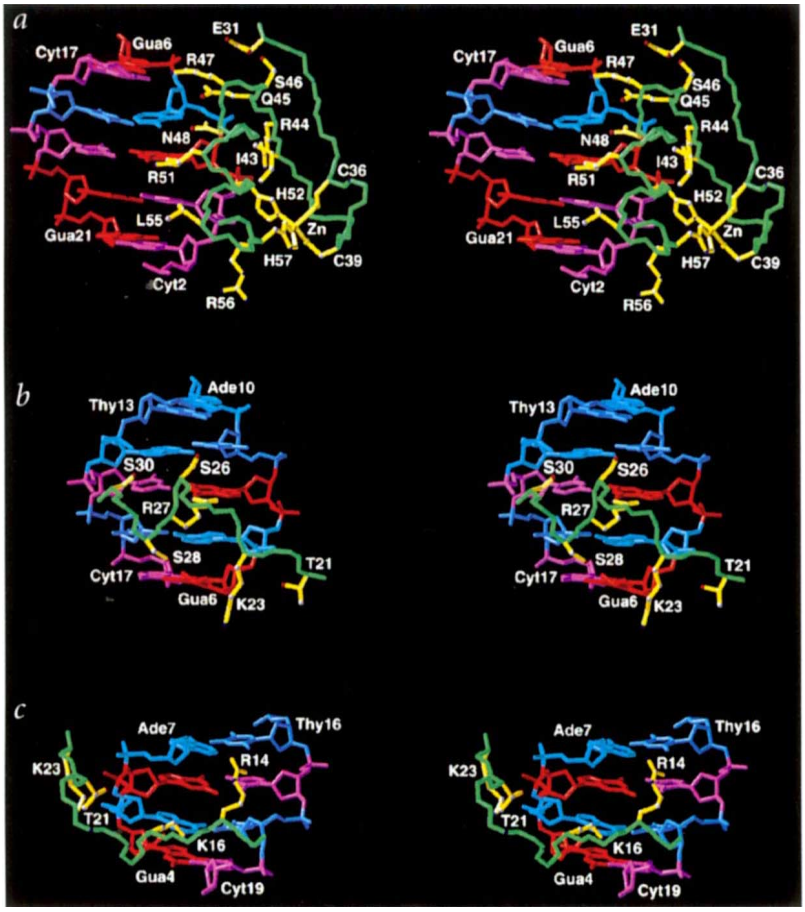

Fig. 7 Stereoviews illustrating the interactions of *a*, the zinc finger core, *b*, the BR2 helix and *c*, the BR1 N-terminal tail of the GAGA-DBD–DNA complex. The protein backbone is in green, the side chains in yellow, T in dark blue, A in lighter blue, C in magenta and G in red. The structure shown is the restrained regularized mean structure. This figure was generated with the program VISP[62].

been superimposed on Finger 2 of Tramtrack. The orientation of the zinc finger core of the GAGA-DBD and Finger 2 of Tramtrack with respect to the DNA are essentially identical. Moreover, the positioning of the BR2 helix of the GAGA-DBD approximately matches that of the helix of Finger 1 of Tramtrack. Unlike the Tramtrack-DBD–DNA complex, however, the GAGA-DBD–DNA complex has a significant minor groove component provided by the basic BR1 region which wraps around the minor groove and is involved in several base specific contacts (Fig. 2*b*, 6 and 7*c*).

N-terminal extensions to the classical zinc finger structure have been previously observed for the first of the zinc fingers of Tramtrack (a β-strand)[36] and SWI5 (a helix and β-strand)[39]. In contrast to the N-terminal BR1 and BR2 regions of the GAGA-DBD, these extensions are in intimate contact with the zinc finger core and are thought to contribute to the stability of the domains. For SWI5, the N-terminal extension also enhances DNA binding, and although no structure of the SWI5 first finger–DNA complex has yet been published, model building suggests the involvement of the additional helix in contacts with the phosphate backbone[39].

Correlation with biochemical data

The impact of the various base specific interactions with the GAGAG consensus in stabilizing the GAGA-DBD–DNA complex is provided by gel retardation experiments (Fig. 1*b*). Although it is not possible to deduce binding constants from single gel shift data points, the data in Fig. 1*b* illustrates that specific DNA binding is reduced to a greater or lesser extent in every case. Consistently the largest effects are seen for the G6→A,T,C substitutions which remove the potential hydrogen bond between the guanidino group of Arg 47 and the O6 atom of G6. No complexes at all were observed for any of the G6→A,T,C substitutions, indicating a reduction in affinity of at least an order of magnitude. This demonstrates the central importance of the Arg 47–G6 interaction involving the major recognition helix of the zinc finger (Fig. 7*a*). The A5→G,T,C substitutions, which decrease the affinity somewhat, remove one of the two potential hydrogen bonds between the carboxyamide of Asn 48 and A5, specifically that between the Oδ1 atom of Asn 48 and 6-NH_2 group of A5 (Fig. 7*a*). Consistent with the role of Asn 48 is the observation that mutation of this residue to an Asp decreases the DNA binding affinity by more than 10-fold (data not shown). This mutation removes the potential hydrogen bond between the side chain NH_2 group of Asn 48 and the N7 atom of A5, and introduces electrostatic repulsion between the negatively charged carboxylate of an Asp and the N7 atom of A5 which has partial negative charge (acting generally as a hydrogen bond acceptor). Similarly, the G4→A,T,C substitutions remove one of the two potential hydrogen bonds between the guanidino group of Arg 51 and G4, specifically that involving the O6 atom of the G4 base. The effects of these substitutions are not as dramatic as those observed for the G6 substitutions, possibly because a hydrogen bond between Lys 16 and the complementary base (C19→T,A,G) could still be formed in the minor groove (Fig. 7*c*). The G8→T,C substitutions markedly reduce the binding, highlighting the importance of the G8–Arg 27 interaction involving the BR2 helix (Fig. 7*b*). Interestingly, complex formation is still

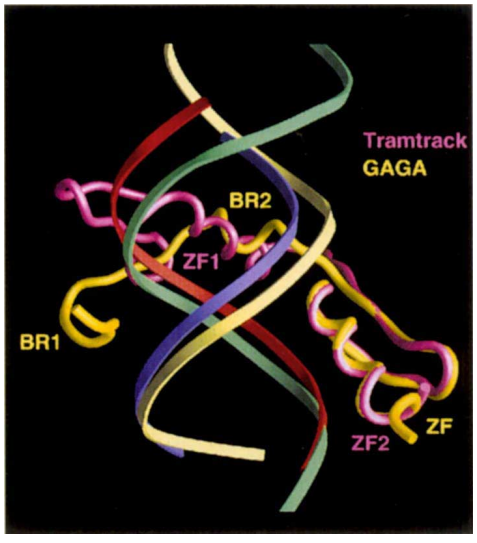

Fig. 8 Comparison of the GAGA-DBD–DNA and Tramtrack-DBD–DNA[36] complexes. The zinc finger core of the GAGA-DBD (residues 33–58) is superimposed on finger 2 (residues 140–165) of Tramtrack (Cα r.m.s. difference of 0.8 Å). The protein backbones are shown as tubes with the GAGA-DBD in yellow and the Tramtrack-DBD in magenta. The two DNA duplexes are shown as ribbons through the phosphate backbone; the shorter of the two duplexes with red and purple strands belongs to the GAGA-DBD–DNA complex, while the longer one with light green and pale yellow strands belongs to the Tramtrack-DBD–DNA complex. The coordinates of the GAGA-DBD–DNA complex shown are those of the restrained regularized mean structure. This figure was generated with the program GRASP[61].

observed for the G8→A substitution, probably because one hydrogen bond could still be formed between the guanidino group of Arg 27 and the N7 atom of the A base. The A7→G,T,C substitutions reflect the base interactions of Arg 14 of BR1 in the minor groove (Fig. 7*c*). The A7→T,C substitutions still bind the GAGA-DBD, albeit with reduced affinity, since the O2 atoms of T and C occupy almost the same position as that of the N3 atom of A. The absence of complex formation for the A7→G substitution is probably due to steric hindrance and electrostatic repulsion arising from the presence of the bulky 2-NH_2 group in G which replaces the H2 proton of A in the minor groove. The importance of Arg 14 for sequence specific recognition is further supported by the observation that a truncated GAGA-DBD starting at Ala 15 exhibits no detectable complex formation as probed by gel shift experiments (unpublished data).

Other DNA binding proteins

The general mode of DNA binding (as opposed to the specifics of side chain-base interactions and recognition) observed for the GAGA-DBD is reminiscent of that seen for the DBD of the GATA-1 transcription factor[40]. The GATA-1-DBD also clamps the DNA with a helix and loop interacting in the major groove, and a basic C-terminal tail, instead of an N-terminal one as in the GAGA-DBD, wrapping around the minor groove. The size of the DNA binding site is similar, and, just as for the GAGA-DBD, the major groove interactions involve a zinc binding domain. The zinc binding module of GATA-1, however, is not a classical TFIIIA-like zinc finger but is structurally related to the N-terminal zinc module of the glucocorticoid receptor[41]. Thus, both the GAGA-DBD and the GATA-1-DBD employ two structural motifs to interact with the DNA. In this regard it is interesting to note that the zinc binding domain and basic tail of chicken GATA-1 are encoded on separate exons[42]. At this stage, the exon/intron structure for the gene encoding the GAGA factor is unknown.

Relationship to chromatin remodelling

The present structure of the GAGA-DBD–DNA complex shows how a single classical Cys_2-His_2 zinc finger complemented by an N-terminal extension comprising a basic helix (BR2) and tail (BR1) can recognize DNA in a sequence specific manner. The requirement for binding of the zinc finger core and BR2 in the major groove, together with the simultaneous binding of BR1 in the minor groove opposite the recognition helix of the zinc finger core, suggests that there may be special constraints on interactions between the GAGA factor and DNA targets within nucleosomes, where a part of the binding site may be relatively inaccessible. Such interactions could favour (and be favoured by) disruption of the nucleosome core by displacement or weakening of DNA–histone contacts, particularly in the presence of additional factors such as the nucleosome remodelling factor NURF[9,43].

The minimal binding site for the GAGA-DBD, as deduced from the present structural study, comprises nine base pairs with a central pentanucleotide G_4AGAG_8 motif (employing the numbering scheme shown in Fig. 1*a*). From the results presented in Fig. 1*b*, it is clear that specific binding with only a relatively small decrease in affinity (that is, less than a factor of 10), can be achieved with modifications, albeit one base pair at a time, of G4 and A5 to any other base, A7 to T or C, and G8 to A, with only G6 being invariant. The promoters that are known to be targets for the GAGA factor frequently contain a high density of GA repeats[4], generating a large number of potential GAGA factor binding sites in close proximity. For example, the *hsp26-1* promoter contains at least 10 partially overlapping potential GAGA factor binding sites[4,44]. Likewise, a subset of highly repetitive DNA sequences found in heterochromatin of *Drosophila* are GA rich and found to be associated with the GAGA factor at all stages of the cell cycle, thereby possibly modifying heterochromatin in these regions[8]. This suggests that binding of multiple copies of the GAGA factor, possibly interacting with each other through protein–protein contacts, may be required to achieve efficient disruption of the nucleosome. The proximity of so many potential binding sites also suggests that the GAGA-factor may possibly migrate along the DNA, diffusing along a linear lattice from one site to the next adjacent site. In the *in vivo* situation, the presence of such a large number of adjacent GAGA factor binding sites is likely to ensure that at least some sites would be located in a linker region on a nucleosome array, serving as an anchor point for GAGA factor binding. Consistent with that mechanism is the observation that the GAGA factor binds weakly to mononucleosomes, and even in the presence of the nucleosome remodelling factor NURF does not cause complete disruption of nucleosome structure under these conditions[43]. In contrast, the GAGA factor binds tightly to a nucleosome array containing multiple GAGA factor sites, even in the absence of NURF[9,43].

The GAGA factor is able to alleviate the repression of transcription by linker histones (H1 or H5)[7]. Conversely, the linker histone H1 decreases NURF/ATP dependent stimulation of GAGA factor binding[9,45]. The globular domain of the linker histones specifically recognizes and binds to the nucleosome core and is thought to be located asymmetrically inside the superhelical gyre of DNA, just inside the nucleosome core region[46,47]. The globular domain of the linker histones belongs to the winged helix-turn-helix family of DNA binding proteins[48,49], exemplified by HNF-3/*fork head* [50]. While the structure of a linker histone–DNA complex has not yet been determined, presumably because linker histones do not bind naked DNA in a sequence specific manner, it seems likely that the mode of binding of the linker histones would be similar to that seen in the complex of HNF-3/*fork head* with DNA. In the latter complex, the recognition helix of the helix-turn-helix motif binds in the major groove, and the two basic wings contact the adjacent minor grooves on either side of the interacting major groove[50]. The C-terminal wing also makes an arginine mediated base specific contact in the minor groove[50]. Thus, it is possible that the combination of major and minor groove binding by the GAGA-DBD, coupled with its significantly higher DNA sequence specificity relative to the linker histones, permits the GAGA-DBD to directly displace a linker histone under certain circumstances (for example, when the linker histone is directly bound to a GAGA factor site).

As the GAGA-DBD effectively clamps the DNA by binding in the major and minor grooves, there are also limitations on the location of the N-terminal POZ domain and the polyglutamine-rich C-terminal tail in the DNA complex with the full-length GAGA factor. These must be free to interact with other components of the transcription machinery, including the nucleosome remodelling factor NURF which facilitates the complete displacement of the histone octamer from the complex in an ATP-dependent manner[9].

Methods

Sample preparation. The coding sequence for the GAGA-DBD (amino acids 310–372 of the GAGA factor) was generated as an NcoI-BamHI DNA fragment using the polymerase chain reaction as described previously[12]. This DNA fragment was cloned into the Escherichia coli vector pET11A and expressed in host strain BL21(DE3). Uniform (>95%) ^{15}N and ^{13}C labelling was obtained by growing the cells in a modified minimal medium containing $^{15}NH_4Cl$ and/or $^{13}C_6$-glucose as the sole nitrogen and carbon

sources, respectively. The cells were grown at 37 °C, after which protein expression was induced for 4 hrs with 0.5 mM isopropyl-d-thiogalactoside. The cells were harvested, resuspended in 50 mM Tris buffer, pH 8.0, 5 mM EDTA, 5 mM benzamidine and 5 mM dithiothreitol (DTT), lysed by passage through a French press, and centrifuged at 100,000g for 1 hr. The supernatant was applied to a DEAE-Sepharose Fast Flow (Pharmacia) column (200 ml bed volume) equilibrated with Buffer A (50 mM Tris, 5 mM DTT, 5 mM EDTA), and the fractions containing the GAGA-DBD were eluted using a gradient (0-100% over 1500 ml) of Buffer B (50 mM Tris, 5 mM DTT, 5 mM EDTA, 1 M NaCl). The pooled fractions were applied to a SP-Sepharose Fast Flow (Pharmacia) column (200 ml bed volume) equilibrated with Buffer A (50 mM Tris, 5 mM DTT, 5 mM EDTA), and the GAGA-DBD was eluted using a gradient (0–100% over 1500 ml) of Buffer B (50 mM Tris, 5 mM DTT, 5 mM EDTA, 1 M NaCl). The pooled fractions containing the GAGA-DBD were subjected to further purification on a C-4 reversed phase (Vydac) high performance liquid chromatography (HPLC) column using a 1 to 100% acetonitrile gradient in 0.05% (v/v) trifluroacetic acid. The product was characterized by mass spectrometry. The HPLC purified product from HPLC was lyophylized, reconstituted with 1.1 equivalents of zinc, and the final pH adjusted to 6.0 with NaOH.

The DNA oligonucleotides used for NMR were purchased from Midland Certified Reagent Co., Texas, purified by anion exchange chromatography and characterized by mass spectrometry.

The GAGA-DBD–DNA complex was prepared by slowly adding the GAGA-DBD (with zinc bound) to a DNA solution (~200 μM DNA, 10mM NaCl) until a 1:1 ratio of DNA to GAGA-DBD was attained. The solution was then concentrated using a Centriprep-3 (Amicon) concentrator to give a final complex concentration of ~2 mM at pH 6.6 with 10 mM NaCl and 2.2 mM $ZnCl_2$. For each complex, three samples were prepared containing ^{15}N GAGA-DBD+DNA in 90% H_2O/10% D_2O, $^{15}N/^{13}C$ GAGA-DBD+DNA in 90% H_2O/10% D_2O, and $^{15}N/^{13}C$ GAGA-DBD+DNA in 100% D_2O.

NMR spectroscopy. Spectra for the complexes were recorded on AMX500, AMX600 and DMX600 Bruker spectrometers equipped with *z*-shielded gradient triple resonance probes. Most of the data on the h3/h4 complex were recorded at 30 °C, and on the 11mer and 16mer complexes at 39.5 °C. Some data on the h3/h4 complex and the 11mer complex were also recorded at 47 °C. Details of the multidimensional experiments used, together with the original references, are provided in the reviews cited in refs 15–17. 3D double and triple resonance through-bond correlation experiments were used to assign the spectra of the protein; 2D ^1H-^1H NOE, ^{12}C-filtered homonuclear Hartmann-Hahn and ^{12}C-filtered NOE experiments were used to assign the spectrum of the bound DNA; couplings constants ($^3J_{HN\alpha}$, $^3J_{\alpha\beta}$, $^3J_{NH\beta}$, $^3J_{C\gamma N}$, $^3J_{C\gamma CO}$, $^3J_{COCO}$) were obtained by 2D and 3D quantitative *J* correlation spectroscopy[20]. Intramolecular NOEs within the protein were obtained from 3D ^{15}N- and ^{13}C-separated NOE spectra; intramolecular NOEs within the DNA from 2D ^1H-^1H NOE (for the imino, amino and H2 protons) and ^{12}C-filtered NOE spectra; and intermolecular NOEs between the protein and the DNA from 3D $^{13}C(F_2)$-separated/ $^{12}C(F_3)$-filtered NOE spectra recorded using either two 90° ^{13}C purge pulses[51] or a single 90° ^{13}C purge pulse[52]. Spectra were processed with the NMRPipe package[53], and analysed using the programs PIPP, CAPP and STAPP[54].

Structure calculations. Approximate interproton distance restraints were derived from the NOE spectra as described previously[40]. NOEs within the protein and between the protein and DNA were grouped into four distance ranges, 1.8–2.7 Å (1.8–2.9 Å for NOEs involving NH protons), 1.8–3.3 Å (1.8–3.5 Å for NOEs involving NH protons), 1.8–5.0 Å and 1.8–6.0 Å, corresponding to strong, medium, weak and very weak NOEs. NOEs within the DNA were classified into five ranges 1.8–2.5 Å, 1.8–3.0 Å, 1.8–3.5 Å, 2.3–5.0 Å and 3.5–6.0 Å, corresponding to strong, medium-strong, medium, weak and very weak NOEs, respectively. 0.5 Å was added to the upper distance limits for NOEs involving methyl protons to account for the higher apparent intensity of methyl resonances. Distances involving methyl groups, aromatic ring protons and non-stereospecifically assigned methylene protons were represented as a $(\Sigma r^{-6})^{-1/6}$ sum[21]. Protein backbone hydrogen bonding restraints within areas of regular secondary structure were introduced during the last stages of refinement. Hydrogen bonding restraints within the DNA were used to maintain base pairing. In the very last stage of refinement, 10 ambiguous $(\Sigma r^{-6})^{-1/6}$ sum distance restraints for potential intermolecular hydrogen bonds, together with 'repulsive' distance restraints to prevent hydrogen bond donors from coming unusually close to donors, and hydrogen bond acceptors to acceptors were introduced for the five residues (Arg 51, Asn 48, Arg 47, Arg 27 and Arg 14) that contact the DNA bases and are buried at the protein-DNA interface (see Table 1 and Section on Structure Determination). ϕ, ψ, χ_1 and χ_2 torsion angle restraints were derived from the NOE and coupling constant data[55,56]. Broad torsion angle restraints, covering the values characteristic for both A and B-DNA, were also employed for the DNA backbone to prevent problems associated with local mirror images[57]: α=60±50°, β=180±50°, γ=60±35°, ε=180±50° and ζ=–85±50°. The structures were calculated by simulated annealing[22,23] using the program XPLOR[58], modified to incorporate pseudo-potentials for $^3J_{HN\alpha}$ coupling constants[24], secondary $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts[25], and a conformational database potential for the protein[26]. The target function that is minimized during simulated annealing and restrained regularization comprises only quadratic harmonic potential terms for covalent geometry, base pair planarity, $^3J^{HN\alpha}$ coupling constant, and secondary $^{13}C\alpha$ and $^{13}C\beta$ chemical shift restraints; square-well quadratic potentials for the experimental distance and torsion angle restraints; and a quartic van der Waals repulsion term together with the conformational database potential for the non-bonded contacts. The covalent geometry restraints include terms for maintaining the tetrahedral coordination geometry of the zinc[59]. There were *no* hydrogen-bonding, electrostatic or 6–12 Lennard–Jones empirical potential energy terms in the target function. All structure calculations were carried out using residues 10–63 of the protein and base pairs 1–11 in the numbering scheme of the 11 bp duplex (see Fig. 1*a*). (Since residues 1–9 are disordered and display no non-sequential NOEs, they were not included in the calculations).

Two different approaches were used to calculate the structure of the complex. In the first method, the complete complex was calculated using a variant of the hybrid distance geometry-simulated annealing protocol[22], followed by further simulated annealing as described below. In the second, only the structure of the protein component was calculated with the hybrid distance geometry-simulated annealing protocol, and the DNA, in an initial regular B conformation, was docked by simulated annealing. The DNA was placed in a number of different orientations and positions with respect to the protein: this included starting structures with the DNA placed 30 Å away from the protein, as well as with the DNA partially docked on the protein in several orientations differing by up to 180° from the final orientation without taking care of atomic overlap. In all cases, convergence to the same ensemble of structures was achieved. The simulated annealing protocol employed permits chains to pass through one another during the early part of the cooling (annealing) process, so that even if the protein and DNA are intertwined in the initial coordinates, convergence is readily achieved[22,23]. Briefly the protocol is as follows. The initial phase comprises 10 ps of dynamics (5000 integration time steps of 2 fs each) at 2000 K with the force constants for the NOEs, dihedral angles, secondary carbon shifts, coupling constants, bonds, angles, impropers, conformational database, and van der Waals terms set to 2 kcal mol^{-1} Å^{-2}, 10 kcal mol^{-1} rad^{-2}, 0.5 kcal mol^{-1} p.p.m.$^{-2}$, 1 kcal mol^{-1} Hz^{-2}, 1000 kcal mol^{-1} Å^{-2}, 200 kcal mol^{-1} rad^{-2}, 50 kcal mol^{-1} rad^{-2}, 0.001, and 1 kcal mol^{-1} Å^{-4} (with the van der Waals radius scale factor set to 1.2), respectively. In the second phase, the system is slowly cooled down (annealed) from 2000–100 K over a total duration of 24 ps. There are 76 cycles of cooling, each comprising 0.316 ps (158 integration time steps of 2 fs each), with a reduction in temperature of 25 K per cycle. During this period the force constants for the dihedral angle, secondary carbon shifts, coupling constants, and bond terms are held constant at 200 kcal mol^{-1} rad^{-2}, 0.5 kcal mol^{-1} p.p.m.$^{-2}$, 1.0 kcal mol^{-1} Hz^{-2}, and 1000 kcal mol^{-1} Å^{-2}, respectively, while the other force constants are slowly increased at each cycle by a factor given by (initial value–final value)$^{1/76}$. The force constants for the NOEs, angles, impropers, conformational database, and van der Waals terms were increased from 2 to 30 kcal mol^{-1} Å^{-2}, 200 to 500 kcal mol^{-1} rad^{-2}, 50 to 500 kcal mol^{-1} rad^{-2}, 0.001 to 1, and 0.004 to 4 kcal mol^{-1} rad^{-4}, respectively. Concomitantly, the van der Waals radius scale factor was decreased from 0.9 to 0.8. Finally, 500 cycles of Powell minimization were performed with the values of the various force constants set to their final values attained at the end of the cooling phase. In addition to the various terms mentioned, a

weak planarity restraint of 10 kcal mol^{-1} Å^{-2} was applied throughout the protocol for each base pair with the exception of the two terminal ones where a strong base pair planarity restraint of 500 kcal mol^{-1} Å^{-2} was applied to prevent artefactual distortions due to end effects. The planarity force constant of 10 kcal mol^{-1} Å^{-2} is sufficiently weak to allow extensive propeller twisting and base roll to occur, while ensuring good stereochemistry.

The coordinates of the 50 final simulated annealing structures of the GAGA-DBD–DNA complex, together with the coordinates of the restrained regularized mean structure, $(\overline{SA})r$, and the complete list of experimental NMR restraints have been deposited in the Brookhaven Protein Data Bank (accession codes 1YUJ, 1YUI and R1YUIMR, respectively).

Acknowledgements

We thank J. Kuszewski for help with Fig. 3; D. Garrett and F. Delaglio for software support; R. Tschudin for technical support; K. Sakaguchi for MS analysis; and A. Bax, S. Grzesiek, J. Huth, J. Kuszewski, and N.Tjandra for useful discussions. This work was supported by the AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health (to G.M.C. and A.M.G.).

1. Biggin, M.D. & Tjian, R. Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. *Cell* **53**, 699–711 (1988).
2. Soeller, W.C., Oh, C.E. & Kornberg, T.B. Isolation of cDNAs encoding the *Drosophila* GAGA transcription factor. *Mol. Cell. Biol.* **13**, 7961–7970 (1993).
3. Farkas, G., Gausz, J., Galloni, M., Reuter, G., Gyurkovics, H. & Karch, F. The *trithorax-like* gene encodes the *Drosophila* GAGA factor. *Nature* **371**, 806–808 (1994).
4. Granok, H., Leibovitch, B.A., Shaffer, C.D. & Elgin, S.C.R. Ga-ga over GAGA factor. *Curr. Biol.* **5**, 238–241 (1995).
5. Lu, Q., Wallrath, L.L. & Elgin, S.C.R. $(CT)_n$•$(GA)_n$ repeats and heat shock elements have a distinct roles in chromatin structure and transcriptional activation of the *Drosophila* hsp26 gene. *Mol. Cell. Biol.* **13**, 2802–2814 (1993).
6. O'Brien, T., Wilkins, R.C., Giardina, C. & Lis, J.T. Distribution of GAGA protein on *Drosophila* genes in vivo. *Genes & Dev.* **9**, 1098–1110 (1995).
7. Croston, G.E., Kerrigan, L.A., Lira, M., Marshak, D.R. & Kadonaga, J.T. Sequence specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. Science **251**, 643–649 (1991).
8. Raff, J.W., Kellum, R. & Alberts, B. The *Drosophila* GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle. *EMBO J.* **13**, 5977–5983 (1994).
9. Tsukiyama, T., Becker, P.B. & Wu, C. ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. *Nature* **367**, 525–532 (1994).
10. Wall, G., Varga-Weisz, Sandaltzopoulos, R. & Becker, P.B. Chromatin remodeling by GAGA factor and heat shock factor at the hypersensitive *Drosophila* hsp26 promoter in vivo. *EMBO J.* **14**, 1727–1736 (1995).
11. Shopland, L.S., Hirayoshi, K., Fernandes, M. & Lis, J.T. HSF access to heat shock elements in vivo depends critically on promoter architecture defined by GAGA factor, TFIID, and RNA polymerase II binding sites. *Genes & Dev.* **9**, 2756–2769 (1995).
12. Pedone, P.V. *et al.* The single Cys_2-His_2 zinc finger domain of the GAGA protein flanked by basic residues is sufficient for high-affinity specific DNA binding. *Proc. Natl. Acad. Sci. USA* **93**, 2822–2826 (1996).
13. Klug, A. & Schwabe, J.W.R. Zinc fingers. *FASEB J.* **9**, 597–604 (1995).
14. Berg, J.M. & Shi, Y. The galvanization of biology: a growing appreciation for the roles of zinc. *Science* **271**, 1081–1085 (1996).
15. Sakai, H., Medrano, L.J. & Meyerowitz, E.M. Role of SUPERMAN in maintaining Arabdopsis floral whorl boundaries. *Nature* **378**, 199–203 (1995).
16. Tague, B.W. & Goodman, H.M. Characterization of a family of Arabidopsis zinc finger protein cDNAs. *Plant Mol. Biol.* **28**, 267–279 (1995).
17. Clore, G.M. & Gronenborn, A.M. Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science* **252**, 1390–1399 (1991).
18. Bax, A. & Grzesiek, S. Methodological advances in protein NMR. *Acct. Chem. Res.* **26**, 131–138 (1993).
19. Gronenborn, A.M. & Clore, G.M. Structures of protein complexes by multidimensional heteronuclear magnetic resonance spectroscopy. *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 351–385 (1995).
20. Bax, A. *et al.* Measurement of homo- and heteronuclear J couplings from quantitative J correlation. *Meth. Enzymol.* **239**, 79–106 (1994).
21. Nilges, M. A calculational strategy for the structure determination of symmetric dimers by ^{1}H NMR. *Proteins Struct. Funct. Genet.* **17**, 295–309 (1993).
22. Nilges, M., Clore, G.M. & Gronenborn, A.M. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing. *FEBS Lett* **229**, 129–136 (1988).
23. Nilges, M., Gronenborn, A.M., Brünger, A.T. & Clore, G.M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints: application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Engng.* **2**, 27–38 (1988).
24. Garrett, D.S. *et al.* The impact of direct refinement against three-bond HN-CαH coupling constants on protein structure determination by NMR. *J. Magn. Reson.Series B* **104**, 99–103 (1994).
25. Kuszewski, J., Qin, J., Gronenborn, A.M. & Clore, G.M. The impact of direct refinement against ^{13}Cα and ^{13}Cβ chemical shifts on protein structure determination by NMR. *J. Magn. Reson.Series B* **106**, 92–96 (1995).
26. Kuszewski, J., Gronenborn, A.M. & Clore, G.M. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci.* **5**, 1067–1080 (1996).
27. Karplus, P.A. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **5**, 1406–1420 (1996).
28. Brooks, B.R. *et al.* CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187–217 (1993).
29. Eisenberg, D. & McLaghlan, A.D. Solvation energy in protein folding and binding. *Nature* **319**, 199–203 (1986).
30. Janin, J & Chothia, C. The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030 (1990).
31. Ippolito, J.A., Alexander, R.S. & Christianson, D.W. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.* **215**, 457–471 (1990).
32. Henry, G.D. & Sykes, B.D. Determination of the rotational dynamics and pH dependence of the hydrogen exchange rates of the arginine guanidino group using NMR spectroscopy. *J. Biomolec. NMR* **5**, 59–66 (1995).
33. Mueller, L., Legault, P. & Pardi, A. Improved RNA structure determination by detection of NOE contact to exchange-broadened amino proteins. *J. Am. Chem. Soc.* **117**, 11043–11048 (1995).
34. Mulder, F.A.A., Spronk, C.A.E.M., Slijper, M., Kaptein, R. & Boelens, R. Improved HSQC experiments for the observation of exchange broadened signals. *J. Biomol. NMR* **8**, 223–228 (1996).
35. Chuprina, V.P., Rullmann, J.A.C., Lamerichs, R.M.N.J., van Boom, J.H., Boelens, R. & Kaptein, R. Structure of the complex of the lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectrscopy and restrained molecular dynamics. *J. Mol. Biol.* **234**, 446–462.(1993)
36. Fairall, L., Schwabe, J.W.R., Chapman, L., Finch, J.T. & Rhodes, D. The crystal structure of a two zinc-finger peptide reveals an extension to the rules of zinc-finger/DNA recognition. *Nature* **366**, 483–487 (1993).
37. Pavletich, N.P. & Pabo, C.O. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**, 809–816 (1991).
38. Pavletich, N.P. & Pabo, C.O. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science* **261**, 1701–1707 (1993).
39. Dutnall, E.N., Neuhaus, D. & Rhodes, D. The solution structure of the first zinc finger domain of SWI5: a novel structural extension to a common fold. *Structure* **4**, 599–611 (1996).
40. Omichinski, J.G. *et al.* NMR structure of a specific DNA complex of a Zn-containing DNA binding domain of GATA-1. *Science* **261**, 438–446 (1993).
41. Luisi, B. *et al.* Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**, 497–505 (1991).
42. Hannon, R., Evans, T., Felsenfeld, G. & Gould, H. Structure and promoter activity of the gene for the erythroid transcription factor GATA-1. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 3004– (1991).
43. Tsukiyama, T. & Wu, C. Purification and properties of an ATP-dependent nucleosome remodelling factor. *Cell* **83**, 1011–1020 (1995).
44. Lis, J. & Wu, C. Protein traffic on the heat shock promoter: parking, stalling and trucking along. *Cell* **74**, 1–4 (1993).
45. Kingston, R.E., Bunker, C.A. & Imbalzano, A.N. Repression and activation by multiprotein complexes that alter chromatin structure. *Genes & Dev.* **10**, 905–920 (1996).
46. Hayes, J.J. Site-directed cleavage of DNA by a linker histone-Fe(II) EDTA conjugate: localization of a globular domain binding site within a nucleosome. *Biochemistry* **35**, 11931–11937 (1996).
47. Pruss, D., Bartolomew, B., Persinger, J., Hayes, J.J, Arents, G., Moudrianakis, N. & Wolffe, A. P. An asymmetric model for the nucleosome: a binding site for linker histones inside DNA gyres. *Science* **274**, 614–617 (1996).
48. Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L. & Sweet, R.M. Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature* **362**, 219–223 (1993).
49. Cerf, C.*et al.* Homo- and heteronuclear two-dimensional NMR studies of the globular domain of histone H1: full assignment, tertiary structure, and comparison with the globular domain of histone H5. *Biochemistry* **33**, 11079–11086 (1994).
50. Clark, K.L., Halay, E.D., Lai, E. & Burley, S.K. Co-crystal structure of the HNF-3/*fork head* DNA-recognition motif resmbles histone H5. *Nature* **364**, 412–420 (1993).
51. Ikura, M. *et al.* Solution structure of a calmodulin-target peptide complex by multi-dimensional NMR. *Science* **256**, 632–638 (1992).
52. Vuister, G.W., Kim, S.-J., Wu, C. & Bax, A. 2D and 3D NMR study of phenylalanine residues in proteins by reverse isotopic labeling. *J. Am. Chem. Soc.* **116**, 9206–9210 (1994).
53. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomolec. NMR* **6**, 277–293 (1995).
54. Garrett, D.S., Powers, R., Gronenborn, A.M. & Clore, G.M. A common sense approach to peak picking in two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.* **95**, 214–220 (1991).
55. Nilges, M., Clore, G.M. & Gronenborn, A.M. ^{1}H-NMR stereospecific assignments by conformational database searches. *Biopolymers* **29**, 813–822 (1990).
56. Powers, R. *et al.* The high resolution three-dimensional solution structure of human interleukin-4 determined by multi-dimensional heteronuclear magnetic resonance spectroscopy. *Biochemistry* **32**, 6744–6762 (1993).
57. Gronenborn, A.M. & Clore, G.M. Analysis of the relative contributions of the nuclear Overhauser interproton distance restraints and the empirical energy function in the calculation of oligonucleotide structures using rsetrained molecular dynamics. *Biochemistry* **28**, 5978–5984.
58. Brünger, A.T. X-PLOR Manual, Version 3.1 (New Haven, Connecticut: Yale University, 1993).
59. Omichinski, J.G., Clore, G.M., Appella, E., Sakaguchi, K. & Gronenborn, A.M. High-resolution three-dimensional structure of a single zinc finger from a human enhancer binding protein in solution. Biochemistry **29**, 9324–9334 (1990).
60. Lavery, R. & Sklenar, H. Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomolec. Struct. Dyn.* **6**, 655–667 (1989).
61. Nichols, A.J., Sharp, K. & Honig, B. Protein folding and association: insights from interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).
62. de Castro, E. & Edelstein, S. VISP 1.0 User's Guide (Geneva, Switzerland: University of Geneva, 1992).
63. Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).