J. Mol. Biol. (1998) **277**, 605–620

JMB

The Solution Structure of a Fungal AREA Protein-DNA Complex: An Alternative Binding Mode for the Basic Carboxyl Tail of GATA factors

Mary R. Starich[1], Mats Wikström[1], Herbert N. Arst Jr[2], G. Marius Clore[1]* and Angela M. Gronenborn[1]*

[1]*Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive Kidney Diseases National Institutes of Health Bethesda, MD 20892-0520 USA*

[2]*Department of Infectious Diseases, Imperial College School of Medicine at Hammersmith Hospital Du Cane Road, London W12 0NN UK*

**Corresponding authors*

The solution structure of a complex between the DNA binding domain of a fungal GATA factor and a 13 base-pair oligonucleotide containing its physiologically relevant CGATAG target sequence has been determined by multidimensional nuclear magnetic resonance spectroscopy. The AREA DNA binding domain, from *Aspergillus nidulans*, possesses a single Cys_2-Cys_2 zinc finger module and a basic C-terminal tail, which recognize the CGATAG element *via* an extensive network of hydrophobic interactions with the bases in the major groove and numerous non-specific contacts along the sugar-phosphate backbone. The zinc finger core of the AREA DNA binding domain has the same global fold as that of the C-terminal DNA binding domain of chicken GATA-1. In contrast to the complex with the DNA binding domain of GATA-1 in which the basic C-terminal tail wraps around the DNA and lies in the minor groove, the structure of complex with the AREA DNA binding domain reveals that the C-terminal tail of the fungal domain runs parallel with the sugar phosphate backbone along the edge of the minor groove. This difference is principally attributed to amino acid substitutions at two positions of the AREA DNA binding domain (Val55, Asn62) relative to that of GATA-1 (Gly55, Lys62). The impact of the different C-terminal tail binding modes on the affinity and specificity of GATA factors is discussed.

Keywords: AREA; GATA transcription factors; NMR structure; class IV zinc finger; cGATA-1 DBD

Introduction

The GATA transcription factors represent a major family of zinc-containing, regulatory proteins in a wide range of organisms (Crawford & Arst, 1993; Weiss & Orkin, 1995). Mammalian and avian GATA factors enhance a myriad of gene expression profiles observed during normal cell differentiation. Filamentous fungal and yeast GATA factors include primary regulators of nitrogen metabolism, allowing these organisms to utilize effectively a variety of nitrogen nutrients.

All GATA proteins share a highly conserved DNA-binding domain (DBD) consisting of a Cys_2-Cys_2 type IV zinc finger (Omichinski *et al.*, 1993a) and a variable length basic arm required for recognition of the GATA sequence (Figure 1A). Whereas vertebrate, insect and nematode proteins possess two adjacent homologous fingers, most from yeast and fungi contain a single finger. The zinc fingers in the fungal proteins AREA and NIT2 most closely resemble the carboxyl fingers of the vertebrate proteins (Kudla *et al.*, 1990; Feng *et al.*, 1993). This is not surprising, as deletion studies targeting the tandem fingers of GATA-1 indicate that only the carboxyl finger is required for specific binding (Martin& Orkin, 1990). Further, gel-retardation assays have shown that a 66-residue construct derived from chicken GATA-1 and containing the

Present address: M. Wikström, Pharmacia & Upjohn, Department of Structural Chemistry, S-112 87 Stockholm, Sweden.

The first two authors (M.R.S. and M.W.) contributed equally to this work.

Abbreviations used: DBD, DNA binding domain; NOE, nuclear Overhauser enhancement; HSQC, heteronuclear single quantum coherence; 3D, three-dimensional.

0022–2836/98/130605–16 $25.00/0/mb981625

carboxyl zinc finger and adjacent C-terminal basic region is sufficient for recognition of a single asymmetric GATA site (Omichinski *et al*., 1993b).

Although most GATA factors recognize and bind to a single consensus target of the form (A/T) GATA(A/G), discriminative specificities for the flanking bases and the fourth base of the GATA core sequence have been reported (Ko & Engel, 1993; Merika & Orkin, 1993). Thus, AREA is capable of recognizing a non-consensus CGATAG site which is physiologically relevant to the fungal system (Ravagnani *et al*., 1997), while the amino fingers of GATA-2 and GATA-3 demonstrate higher affinity for sites containing a GATC core sequence (Pedone *et al*., 1997).

While considerable work has focused on identifying the cellular targets of GATA family transcription factors, the molecular mechanisms that determine subtle differences in DNA specificity for each of these regulatory proteins are not well understood. Thus, the structure of only a single complex of the carboxyl finger of the chicken GATA-1 DNA binding domain (cGATA-1 DBD) bound to a consensus AGATAA target has been determined to date (Omichinski *et al*., 1993a). To gain further insight into transcriptional control and specificity within the GATA family, we have initiated structural studies of AREA, the primary nitrogen regulatory protein of *Aspergillus nidulans* (Arst & Cove, 1973; Kudla *et al*., 1990). AREA positively regulates more than 100 structural genes necessary for nitrogen source utilization in the absence of the preferred nitrogen sources of ammonium or glutamine (Kudla *et al*., 1990; Wiame *et al*., 1985). Further, the AREA system represents an ideal eukaryotic model for studying GATA specificity, as extensive sophisticated formal genetics characterizing the system *in vivo* has already been carried out. Here we present the three-dimensional solution structure of the AREA DBD complexed with a 13 bp DNA oligonucleotide containing its physiologically relevant CGATAG target sequence using multidimensional nuclear magnetic resonance (NMR).

Results and Discussion

AREA affinity for GATA

A 66 amino acid construct encompassing the AREA DBD (Figure 1B) was chosen for NMR studies based on sequence alignments with mammalian GATA DBDs (Figure 1A) and deletion studies which indicated that Arg61 is the last residue required for AREA activity (Stankovich *et al*., 1993; Platt *et al*., 1996b). The accompanying 13 bp target site (Figure 1C) was selected to contain the CGATAG binding site identified as critical for regulation of uric acid-xanthine permease expression *in vivo* (Gorfinkiel *et al*., 1993; Ravagnani *et al*., 1997). Gel retardation assays, with the AREA DBD and 13 bp oligonucleotide maintained in the micromolar range, displayed a single distinct band shift corresponding to the formation of a specific AREA DBD·CGATAG complex with an equilibrium dissociation constant of ~3 μM (see Figure 1 of Starich *et al*., 1998). Although this interaction is specific, it is about 300-fold weaker than that for cGATA-1 (M.R.S., M.W., G.M.C. & A.M.G., unpublished data; Omichinski *et al*., 1993b). Interestingly, neither the AREA nor the cGATA-1 DBDs discriminate between CGATAG and AGATAA sites (M.R.S., M.W., G.M.C. & A.M.G., unpublished data).

A ^{1}H-^{15}N correlation spectrum of the AREA DBD·DNA complex displays a unique set of cross-peaks indicating that a specific complex is formed (see Figure 2A of Starich *et al*., 1998). In addition, exchange between the free and bound states is slow on the chemical shift scale. In this regard, we note that although titration of DNA into protein is not feasible since precipitation is observed in the presence of excess protein over DNA, a sample comprising an ~30% excess of DNA over protein reveals two sets of resonances corresponding to free and bound DNA. An upper limit for the overall exchange rate can be obtained from the differences in chemical shift between the free and bound DNA. For example, the frequency difference, $\Delta\delta$, at 600 MHz between the free and bound resonances of the imino proton of T4 is 66 Hz (0.11 ppm) indicating that the overall exchange rate must be much less than $\ll 400\ s^{-1}$ (i.e. $\ll 2\pi\Delta\delta$). Chemical exchange cross-peaks involving the imino protons are readily observed in a ^{1}H-^{1}H NOE spectrum recorded in water. Knowing the equilibrium dissociation constant (~3 μM) for the binding of the 13 bp oligonucleotide duplex to the AREA DBD, the concentrations of free and bound DNA, and the approximate ^{1}H T_1 value ($\sim 1\ s^{-1}$), estimates of the dissociation and association rate constants can be obtained at a single mixing time (150 ms) from the relative intensities of the diagonal and chemical exchange cross-peaks (Ernst *et al*., 1987). On this basis, we estimate that the dissociation and association rate constants lie in the range of 1 to $5\ s^{-1}$ and 5×10^5 to $2 \times 10^6\ M^{-1}\ s^{-1}$, respectively.

Given that the AREA DBD only contacts 9 bp (see description of structure below), four additional non-specific complexes could potentially be formed with the 13 bp oligonucleotides in addition to the single specific complex. The sensitivity of an ^{1}H-^{15}N HSQC spectrum is sufficiently high to easily detect minor species above the 5% level. Since no cross-peaks corresponding to any minor bound protein species are observed, one can conclude that the ratio of affinities for specific to non-specific binding must exceed two orders of magnitude.

Structure determination

The solution structure of the AREA DBD bound to its cognate CGATAG site was solved using multidimensional heteronuclear-filtered and hetero-

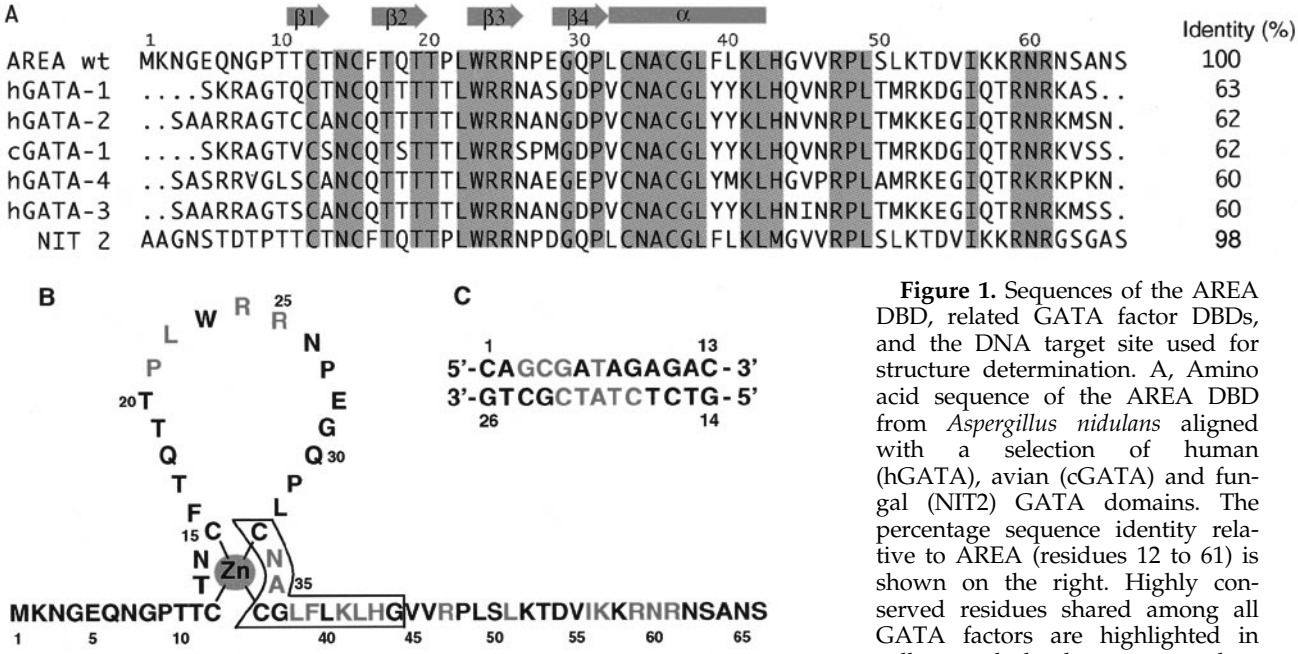

Figure 1. Sequences of the AREA DBD, related GATA factor DBDs, and the DNA target site used for structure determination. A, Amino acid sequence of the AREA DBD from *Aspergillus nidulans* aligned with a selection of human (hGATA), avian (cGATA) and fungal (NIT2) GATA domains. The percentage sequence identity relative to AREA (residues 12 to 61) is shown on the right. Highly conserved residues shared among all GATA factors are highlighted in yellow and the known secondary structure is designated above the sequence. B, The 66-residue AREA DBD construct used to make the protein-DNA complex. The zinc finger module possesses four cysteine residues coordinated to a single zinc atom, highlighted in yellow. Those residues that contact the DNA are shown in red and the α-helical region of the domain is outlined in black. C, The 13 bp DNA duplex used to make the complex contains a CGATAG core element. Bases contacting the AREA DBD in the complex are depicted in red.

nuclear-edited NMR spectroscopy (Clore & Gronenborn, 1991; Gronenborn & Clore, 1995; Bax & Grzesiek, 1993; Bax *et al*., 1994). The structure was determined on the basis of 928 experimental NMR restraints, including 48 intermolecular NOEs. Intermolecular interproton distance restraints were derived unambiguously from a 3D ^{13}C-separated/^{12}C-filtered NOE spectrum which correlates through-space (<6 Å) NOE interactions between protein protons attached to ^{13}C and DNA protons attached to ^{12}C. An example of the quality of the NMR data and the assignment of a number of intermolecular NOEs which uniquely define the orientation of the AREA DBD with respect to the DNA is shown in Figure 1B, a plot displaying the distribution of intermolecular NOEs in Figure 2B, and the superposition of the final 35 simulated annealing structures for the two complexes in Figure 3A. A summary of the structural statistics is provided in Table 1. Residues 1 to 8 and 62 to 66 at the N and C termini, respectively, are highly disordered as evidenced by negative ^{15}N{^{1}H} heteronuclear NOEs, indicative of large amplitude motions (Figure 2D). In contrast, all the remaining residues exhibit heteronuclear ^{15}N{^{1}H}-NOEs greater than 0.6 with the exception of Val45 and Arg61 which have ^{15}N{^{1}H}-NOE values of 0.46 and 0.35, respectively. The mean ^{15}N{^{1}H}-NOE value for residues 9 to 61 is 0.76(±0.12). The precision of the coordinates for the complex (backbone of residues 10 to 61 of the AREA DBD and base-pairs 2 to 11 of the DNA) is ~0.5 Å.

The AREA DBD comprises a protein core (residues 10 to 54) and a C-terminal tail (residues 55 to 61). Although the tail is slightly more mobile than the core (average ^{15}N{^{1}H}-NOE value of 0.62(±0.08) and 0.65(±0.08) excluding residue 61, *versus* 0.78(±0.11)) it is still well ordered due to contacts with the DNA (Figure 2). Within the core there are numerous intramolecular NOEs between residues far apart in the sequence to define the structure. In the case of the C-terminal tail, however, the intramolecular NOEs (with the exception of those involving residues 55 and 56 which display numerous long range NOEs to core residues) are intraresidue and sequential. Nevertheless, the conformation of residues 55 to 61 in contact with the DNA is still well defined by the experimental data owing to the presence of four different types of restraints: (a) intermolecular NOEs between the tail and the DNA which provide long range restraints; (b) $^3J_{HN\alpha}$ and secondary ^{13}C chemical shift restraints which restrict the available ϕ,ψ conformational space (Garrett *et al*., 1994; Kuszewski *et al*., 1995); (c) the conformational database potential which biases sampling during simulated annealing refinement to conformations that are energetically feasible by limiting the choice of dihedral angles to those that are known to be physically realizable (Kuszewski *et al*., 1996, 1997); and (d) the residual one-bond ^{15}N-^{1}H dipolar coupling restraints (Figure 2C). The latter, in contrast to other NMR parameters, provide restraints that characterize long range order *a priori*, since the

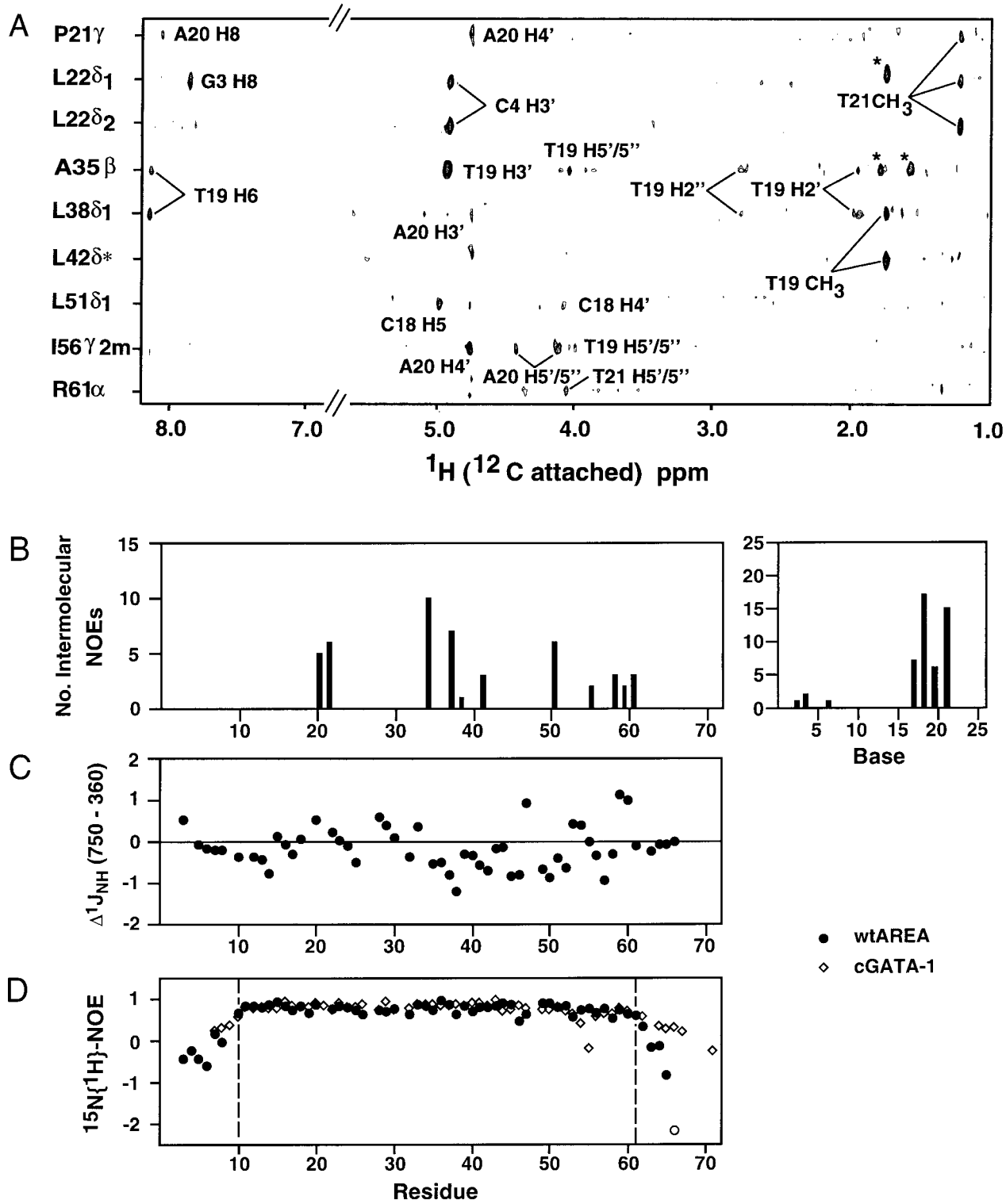

Figure 2. Intermolecular NOEs, residual dipolar couplings and heteronuclear $^{15}N\{^1H\}$-NOEs observed for the AREA DBD·DNA complex. A, Composite of ^{13}C-H strips selected from the 3D ^{13}C-separated/^{12}C-filtered NOE spectrum (mixing time ~ 150 ms) recorded at 25°C for the AREA DBD complexed with a 13 bp oligonucleotide containing a CGATAG target site. This spectrum illustrates the assignment of intermolecular NOEs between protons of the protein (attached to ^{13}C) and protons of the DNA (attached to ^{12}C). Asterisks indicate residual diagonal cross-peaks corresponding to incompletely filtered protons attached to ^{13}C. B, Summary of the distribution of intermolecular NOEs. C, Residual one-bond ^{15}N-^1H dipolar couplings, Δ^1J_{NH}(750-360), obtained by taking the difference in the $^1J_{NH}$ couplings at 750 and 360 MHz. The residual dipolar couplings provide direct information on the orientation of the NH vectors relative to the magnetic susceptibility tensor, which in this case lies approximately parallel with the long axis of the DNA. D, Plot of the heteronuclear $^{15}N\{^1H\}$-NOE values for the AREA (filled circles) and cGATA-1 (open diamonds) DBD·DNA complexes as a function of residue number. Positive NOE values between 0.6 and 0.8 are indicative of the absence of significant internal motions. Negative NOE values indicate the presence of very large amplitude internal motions. The broken lines delimit the ordered region of the AREA DBD when complexed to DNA.

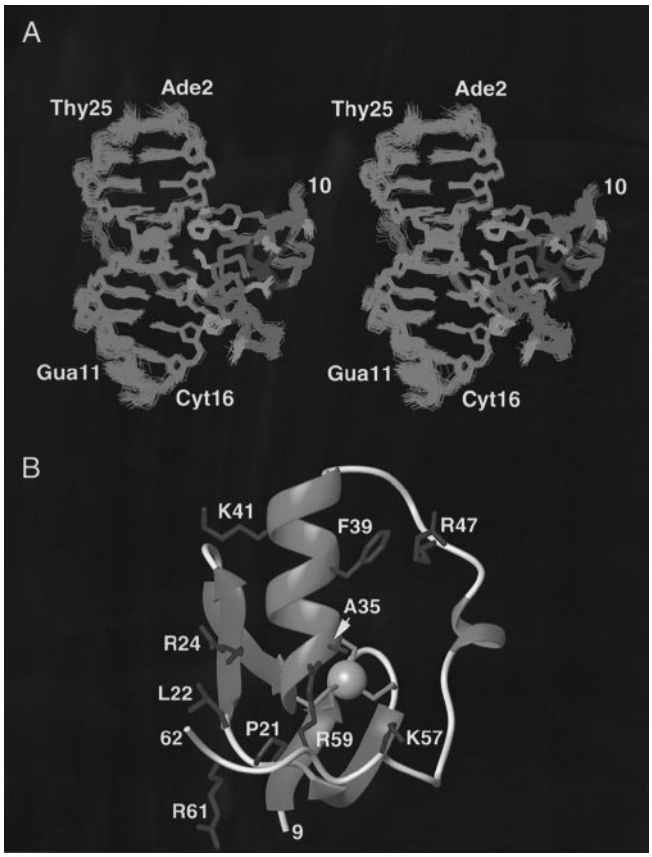

Figure 3. Structure of the AREA DBD·DNA complex. A, Stereoview showing the best fit superposition of the final 35 simulated annealing structures of the AREA DBD·DNA complex. B, Ribbon diagram showing the restrained regularized mean structure of the AREA DBD (residues 9 to 62) in its complexed form. In A, the backbone atoms (N, C^{α}, C) of residues 10 to 61 are shown in red, selected side-chains in yellow, the Zn and coordinating cysteine residues in green, and all non-hydrogen atoms of the DNA (base-pairs 2 to 11), with the exception of the O1P and O2P phosphate oxygen atoms, in blue. In B the core module of the AREA DBD consists of two irregular, antiparallel β-sheets (red) followed by an α-helix (blue) and an extended loop containing an α-helical turn (blue); the zinc atom is represented by a pink ball with the four coordinating cysteine residues shown in yellow; selected side-chains for which *in vivo* mutational data are discussed are shown in green.

magnitude of the residual dipolar couplings is related in a simple geometric manner to the angle between the N-H vectors and the principal axis of the magnetic susceptibility tensor, which in this case is approximately parallel with the long axis of the DNA (Tjandra *et al*., 1997).

Description of structure

The AREA DBD possesses a modular design consisting of a compact protein core (residues 10 to 54) centered around a tetrahedrally coordinated zinc atom, followed by a C-terminal tail (residues 55 to 61), which closely resembles that observed for the cGATA-1 DBD. Superposition of the backbone atoms (C, C^{α}, N) of the respective protein cores yields a backbone atomic rms difference of 1.2 Å. If the superposition is restricted to residues 9 to 44, which comprise all the elements of regular secondary structure (the single helix and the two short antiparallel β-sheets), the backbone rms is only ∼0.7 Å. Given the similarity of these two structures and the detailed description of the domain fold already published for the cGATA-1 DBD (Omichinski *et al*., 1993a), only a brief summary of the AREA DBD fold is presented here.

A schematic illustration of the AREA DBD (Figure 3B) shows that the protein core begins at residue 10 with a short, irregular β-sheet (β1 strand, 10 to 12; β2 strand, 16 to 19) followed by a three-residue extended loop leading into a second irregular β-sheet (β3 strand, 23 to 26; β4 strand, 29 to 32). The α-helix (residues 33 to 43) is terminated by Gly44 which is part of an α_L-type helix capping motif (Aurora *et al*., 1994) that is further stabilized by interactions between His43 and Val45. An extended loop (residues 45 to 56) includes the hallmark helical turn (residues 49 to 51) recognized in cGATA-1 DBD, but lacks the Ω loop observed in the cGATA-1 structure. The remainder of the domain consists of an extended C-terminal tail beginning at Lys57 and ending at Arg61.

Stability of the core module of the AREA DBD relies on the zinc coordination site and efficient hydrophobic packing of the β-sheets about a central tryptophan (residue 23). The zinc atom coordinates with the S^{γ} atoms of four cysteine residues; the first two ligands are donated by the β1 strand (Cys12) and ensuing hairpin turn (Cys15), and the second two ligands are provided by the first turn of the α-helix (Cys33, Cys36). This arrangement orients the first β-sheet relative to the α-helix. The relative positions of the two β-sheets and their interconnecting loop are determined by the hydrophobic packing of Thr10, Thr17, Gln18, Thr20 and Pro31 around the side-chain of Trp23. An additional hydrophobic cluster formed by Thr19, Pro21, Thr53 and Ile56 anchors the C-terminal tail to the loop region between the β-sheets.

The structure of the complexed DNA target closely resembles classical *B*-type DNA (Saenger, 1984), yielding a 2.2 Å atomic rms difference when best-fit to idealized *B*-DNA of the same sequence. A similar best-fit superposition of the complexed DNA and idealized *A*-DNA yields an atomic rms difference of 7.7 Å. The average local helical twist and rise are 35(±4) and 3.8(±0.4) Å, respectively. Values for propeller twist, local inter-base-pair tilt angles and local inter-base-pair roll angles range from ∼0° to $\sim -22°$, $\sim +4°$ to $\sim -4°$, and $\sim +7°$ to $\sim -5°$, respectively, with mean values of $\sim -13°$, ∼0° and ∼0°, respectively.

The overall topology of the complex shows that the zinc finger module of the AREA DBD spans a section of the major groove corresponding to a CGATAG sequence element with the long axis of the helix oriented at ∼60° to the long axis of the

Table 1. Structural statistics

	$\langle SA \rangle$	$(\overline{SA})r$
Structural statistics		
rms deviations from NOE interproton distance restraints (Å)[a]		
All (538)	0.042 ± 0.002	0.030
Protein		
Interresidue sequential ($\|i-j\|=1$) (119)	0.033 ± 0.010	0.029
Interresidue short range ($1<\|i-j\|\leqslant 5$) (49)	0.041 ± 0.008	0.045
Interresidue long range ($1<\|i-j\|\leqslant 5$) (68)	0.037 ± 0.009	0.038
Intraresidue (44)	0.010 ± 0.009	0.010
DNA		
Intraresidue (75)	0.012 ± 0.003	0.012
Sequential intrastrand (115)	0.061 ± 0.004	0.044
Interstrand (20)	0.043 ± 0.008	0.030
Protein-DNA (48)	0.055 ± 0.009	0.060
rms deviation from hydrogen bonding restraints (Å)		
Protein (20)[b]	0.084 ± 0.009	0.068
DNA (66)[b]	0.021 ± 0.004	0.013
Protein-DNA (4)[c]	0.043 ± 0.024	0.028
rms deviations from distance restraints to phosphates (2)[d]	0.001 ± 0.004	0.001
rms deviations from ''repulsive'' restraints (Å) (8)[d]	0.004 ± 0.013	0.007
rms deviations from exptl dihedral restraints (deg.) (294)[a]	0.21 ± 0.07	0.39
rms deviations from exptl $^3J_{HN\alpha}$ (Hz) (41)	0.82 ± 0.04	0.82
rms deviations from exptl residual one-bond ^{15}N-^{1}H dipolar couplings (Hz) (48)	0.123 ± 0.007	0.15
rms deviations from exptl ^{13}C shifts		
$^{13}C^{\alpha}$ (ppm) (41)	0.92 ± 0.03	0.91
$^{13}C^{\beta}$ (ppm) (36)	0.79 ± 0.03	0.76
Deviations from idealized covalent geometry[e]		
Bonds (Å) (1932)	0.005 ± 0.0001	0.006
Angles (deg.) (3502)	0.997 ± 0.006	1.111
Impropers (deg.) (974)	0.471 ± 0.027	0.686
Measures of structural quality		
$E_{L\text{-}J}$ (kcal mol^{-1})[f]	−588 ± 8	−544
PROCHECK[g]		
% residues in most favourable region of Ramachadran map	83 ± 3	79
Number of bad contacts/100 residues	4.3 ± 1.2	7.7
Coordinate precision[h]		
Protein backbone plus DNA (Å)	0.52 ± 0.12	
All protein atoms plus DNA (Å)	0.82 ± 0.17	
Protein backbone (Å)	0.42 ± 0.14	
All protein atoms (Å)	0.99 ± 0.23	
DNA (Å)	0.47 ± 0.13	

The notation of the NMR structures is as follows: $\langle SA \rangle$ are the final 35 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 10 to 61 and the zinc atom of the protein and base-pairs 2 to 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.

[a] 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 3 Hz. The torsion angle restraints comprise 124 (61 ϕ, 8 ψ, 39 χ_1, 15 χ_2 and 1 χ_3) experimentally determined torsion angles for the AREA DBD. In additon, there are 170 broad torsion angle restraints for the DNA, covering (with the exception of the C3′-C4′ δ torsion angle) values characteristic for both *A* and *B*-DNA, to prevent problems associated with local mirror images (Omichinski *et al.*, 1993a): $\alpha = 60 \pm 50°$, $\beta = 180 \pm 50°$, $\gamma = 60 \pm 35°$, $\delta = 145 \pm 30°$, $\varepsilon = 180 \pm 50°$, $\zeta = -85 \pm 50°$; $\chi = -125 \pm 60°$. The C3′-C4′ δ torsion angle restraints were derived from a qualitative interpretation of the ^{12}C-filtered NOE data which indicated that the sugar puckers were unambiguously *B*-like.

[b] Hydrogen bond restraints within the DNA were used to maintain Watson-Crick base-pairing (Gronenborn & Clore, 1989). Protein backbone hydrogen bonding restraints (two per hydrogen bond) within areas of regular secondary structure were introduced during the last stages of refinement using standard criteria.

[c] Intermolecular hydrogen bonding restraints between the protein side-chain of Arg24 and G5 were only added in the final stage of refinement based on the observation of four distinct resonances for the guanidino protons of Arg24 in ^{1}H-^{15}N HSQC and ^{15}N-edited NOE spectra.

[d] In the final stage of the structure calculations, eight ''repulsive'' distance restraints (Omichinski *et al.*, 1997), with a lower bound of 4 Å (and an unrestrained upper bound), were introduced to prevent energetically unfavorable proximity of hydrogen bond donors to other donors, and hydrogen bond acceptor groups to other acceptors. In addition, the N$^{\zeta}$ atom of Lys57 and guanidino groups of Arg25 were restrained within 5.5 Å of DNA phosphate atoms when ^{1}H-^{1}H NOEs from the residue to DNA sugar protons near a phosphate or structure calculations indicated that the side-chain interacted with a DNA phosphate. In each case, $(\Sigma r^{-6})^{-1/6}$ sum distance restraints (Nilges 1993) were used and included a choice of two adjacent phosphate atoms. The effect of such restraints is to permit a chemically sensible distance contact to be obtained between the functional group of the side-chain and at least one of the phosphate groups designated in the restraint.

[e] The improper torsion restraints serve to maintain planarity and chirality. Terms defining the tetrahedral coordination geometry of the zinc are included in the covalent geometry restraints (Omichinski *et al.*, 1993a).

[f] $E_{L\text{-}J}$ is the Lennard-Jones van der Waals energy and is not included in the target function for simulated annealing or restrained minimization.

[g] The PROCHECK (Laskowski *et al.*, 1993) statistics relate to the ordered region of the polypeptide chain comprising the zinc finger core (residues 10 to 54) and the C-terminal tail (residues 55 to 61) in contact with the DNA. There are no residues whose ϕ/ψ angles fall in the disallowed region of the Ramachandran plot.

[h] The precision of the coordinates is defined as the average atomic rms difference between the 35 individual simulated annealing structures of each complex and the mean coordinates $\overline{SA}$. The values refer to residues 10 to 61 of the AREA DBD, the zinc atom and base-pairs 2 to 11 of the DNA.

DNA, while the C-terminal tail parallels the contour of the phosphate backbone (Figure 4). The core module utilizes hydrophobic residues on one face of the α-helix and along the loop between the β2 and β3 strands in order to recognize the major groove surface of the DNA, in a manner that is very similar to that employed by the cGATA-1 DBD (Omichinski *et al*., 1993a). Three leucine residues and a proline account for the majority of these contacts. The side-chain of Arg24, however, also lies in the major groove (Figure 4B and C). A triad of positively charged side-chains (Arg25, Lys41, Arg47) supplements this set of hydrophobic contacts, forming electrostatic interactions with phosphate groups lining the edges of the major groove (Figure 4B and C). Residues in the extended loop (Leu51, Ile56) provide additional hydrophobic contacts with backbone sugars along the antisense strand of the DNA, while orientation of the basic C-terminal tail requires participation of both the polypeptide backbone and charged side-chains (residues Lys57 to Arg61; Figure 4A). Viewing the complex down the long axis of the DNA reveals that the AREA DBD stretches over one half the cir-

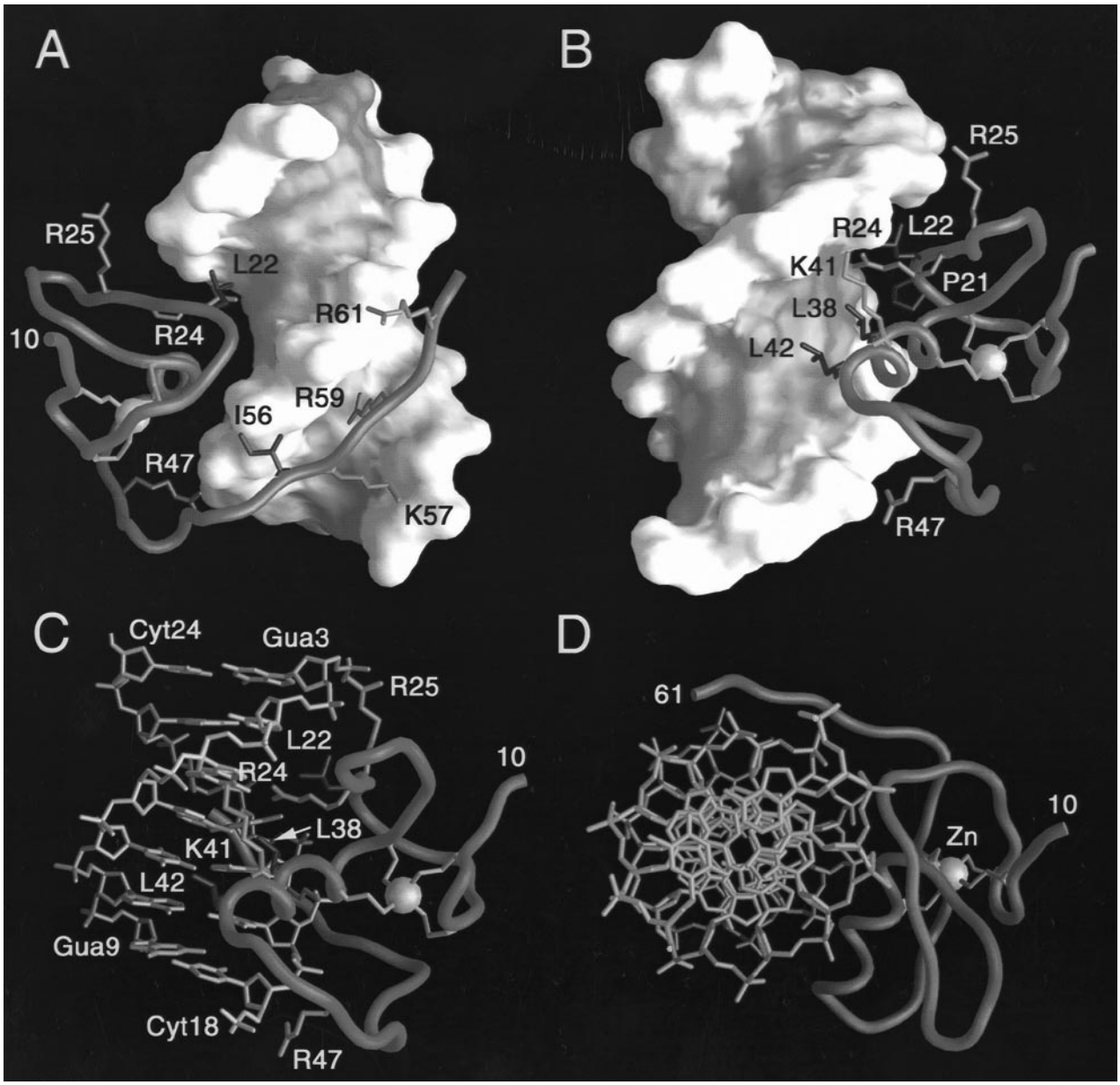

Figure 4. Four views illustrating the interaction of the AREA DBD with its cognate CGATAG target. The protein backbone is depicted as a red worm, while hydrophobic and hydrophilic side-chains participating in DNA recognition are shown in green and blue, respectively. Cysteine side-chains coordinating the zinc atom (pink sphere) are shown in yellow. For A and B the DNA is depicted as a molecular surface with the major groove colored light blue and the minor groove colored light red. In C a bond representation of the DNA is shown with A·T base-pairs in purple and G·C base-pairs in light blue. A bond representation for the DNA is also shown in D with all base-pairs in light blue.

cumference of the double helix (Figure 4D). Upon binding in this fashion, the solvent-accessible surface area of the AREA DBD decreases by ~700 Å^2, representing a ~15% decrease in accessible surface area relative to the free domain.

Protein-DNA contacts

Recognition of the CGATAG element by the AREA DBD requires a network of specific side-chain contacts with nine bases in the major groove (G3, C4, G5, T7, C18, T19, A20, T21 and C22; Figure 5A and B) and numerous non-specific contacts with the sugar-phosphate backbone along the edge of the minor groove (Figure 5C). A diagrammatic summary of the contacts is provided in Figure 6A. The methyl protons of Leu22 exhibit intermolecular NOEs to base and sugar protons (Figure 2A) of G3, C4 and T21, centering residue 22 with respect to the C4/G23 and G5/C22 base-pairs (Figure 5A). This side-chain is preceded by Pro21, which interacts with A20 and T21. The methyl group of Ala35 interacts with the base of T19 and also contacts the sugar rings of T19 and A20 (Figure 5A). The remaining hydrophobic contacts to bases in the major groove are with leucine and phenylalanine side-chains along one face of the α-helix. Thus, the side-chain of Leu38 packs against the bases of T19 and A20, while the side-

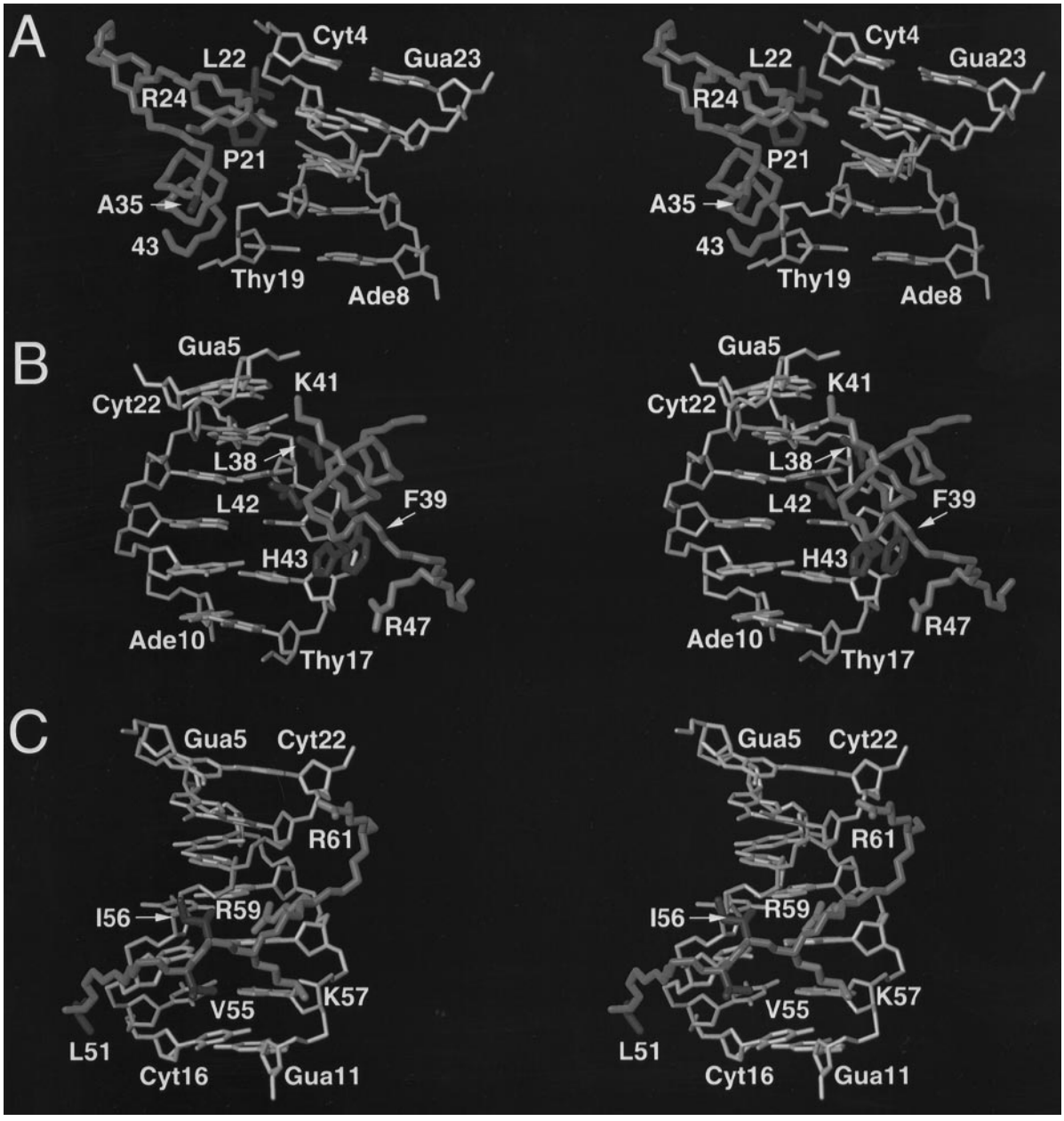

Figure 5. Stereoviews showing specific and non-specific interactions between AREA DBD side-chains and the DNA. The protein backbone is shown in red, hydrophobic side-chains in green and hydrophilic side-chains in dark blue. In each case, A·T base-pairs are colored purple and G·C base-pairs are colored light blue.

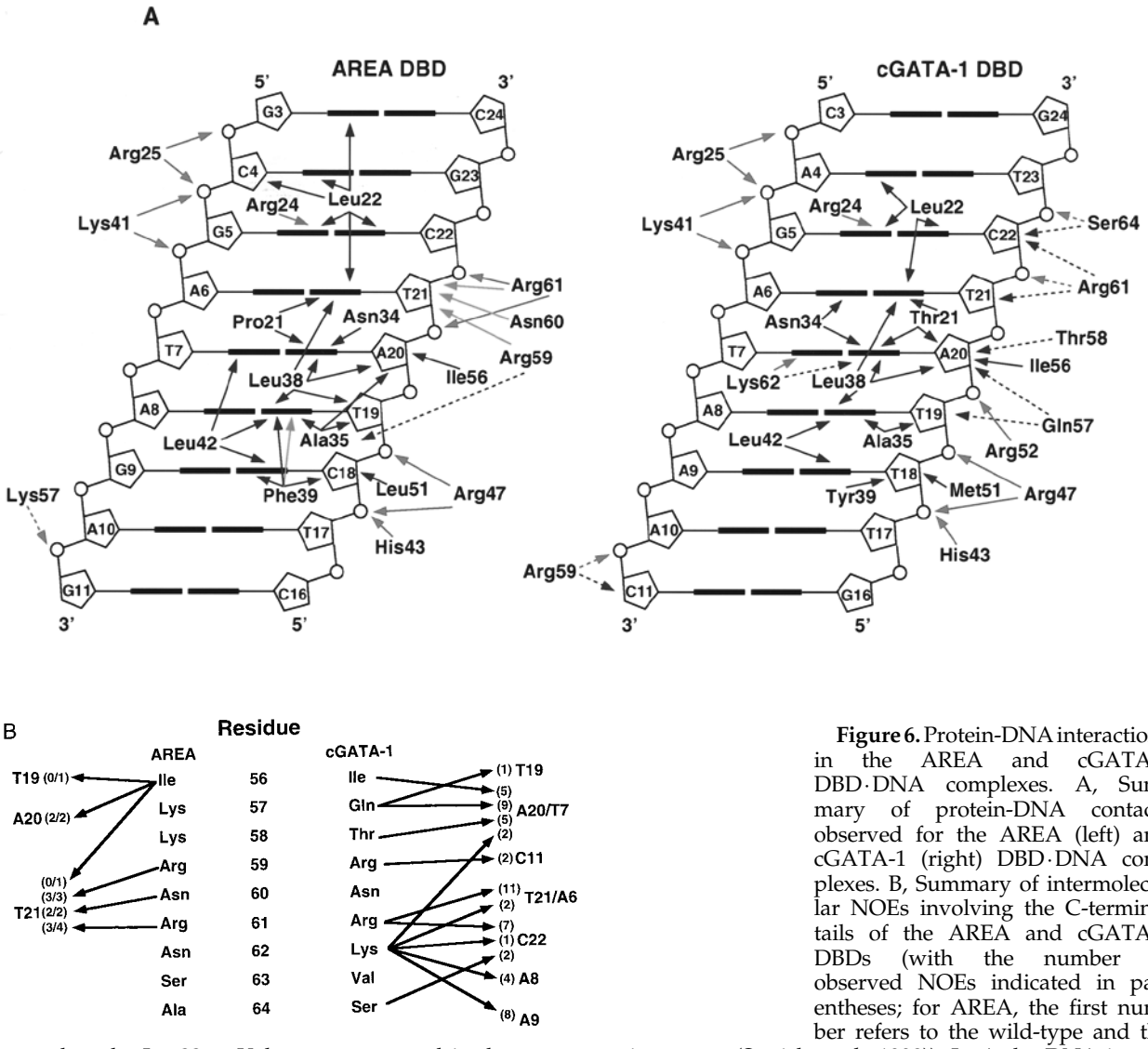

Figure 6. Protein-DNA interactions in the AREA and cGATA-1 DBD·DNA complexes. A, Summary of protein-DNA contacts observed for the AREA (left) and cGATA-1 (right) DBD·DNA complexes. B, Summary of intermolecular NOEs involving the C-terminal tails of the AREA and cGATA-1 DBDs (with the number of observed NOEs indicated in parentheses; for AREA, the first number refers to the wild-type and the second to the Leu22 → Val mutant reported in the accompanying paper (Starich *et al.*, 1998)). In A the DNA is represented as a cylindrical projection viewed from the major groove side, and the numbering scheme for the cGATA-1 oligomer has been altered to match that of the AREA site; bases are indicated as thick lines, the deoxyribose sugar rings as pentagons, and the phosphates as circles; contacts involving amino acid residues are delineated as hydrophobic (green arrows), electrostatic (blue arrows) and H-bonds (magenta); interactions involving backbone amide or alpha protons are designated with yellow arrows and minor groove interactions are designated with broken arrows.

chain of Leu42 interacts with both the T7/A20 and A8/T19 base-pairs (Figure 5B). The aromatic ring of Phe39 interacts with the sugar of C18 and is also packed orthogonally to the base of T19. In this regard, it is interesting to note that the H6 proton of T19 (as well as the equivalent thymidine in the cGATA-1 DBD·DNA complex; Omichinski *et al.*, 1993a) is unusually downfield shifted relative to other thymidine H6 protons (8.15 ppm, compared to 6.9 to 7.4 ppm; cf. Figure 2A). This is due to a ring current shift since the H6 proton of T19 lies directly in the plane of the aromatic ring of Phe39 (and in the case of the cGATA-1 DBD·DNA complex a tyrosine at the equivalent position). Additional contacts with the major groove face of the sugar phosphate backbone are primarily electrostatic and are made by Arg25 (with the phosphate of C4 or G5), Lys41 (with the phosphate of G5 or A6), Arg47 (with the phosphate of C18) and His43 (with the sugar and/or phosphate of C18). Finally, the guanidino NH_2 groups of Arg24 also participate in a base-specific hydrogen bonds with the O6 and N7 atoms of G5. This intermolecular hydrogen bonding pattern is supported by the observation of four distinct ^1H-^{15}N correlations for the guanidino N^{η}-H pairs of Arg24 in the ^1H-^{15}N HSQC spectrum collected at 25°C (data not shown), indicating motional restriction of the functional group (Henry & Sykes, 1995). Given the location of the Arg24 side-chain in the major groove, the most probable reason for this restricted motion is the formation of a buried hydrogen bonding interaction. The assignment of these guanidino protons and nitrogen atoms to Arg24 (N^{ε},

86.7 ppm; H^{ε}, 8.7 ppm; $N^{\eta 1}$, 71.3 ppm; $H^{\eta 11}/H^{\eta 12}$, 5.84/7.07 ppm; $N^{\eta 2}$, 74.8 ppm; $H^{\eta 21}/H^{\eta 22}$, 7.30/9.39 ppm) is based on unambiguous NOEs to surrounding residues (e.g. Leu22 and Leu38) observed in the 3D ^{15}N-separated NOE spectrum. Examination of models calculated before any hydrogen bonding restraints for the guanidino group of Arg24 were introduced indicated that the only possible hydrogen bonding partners were the O6 and N7 atoms of G5.

The majority of non-specific contacts observed in the complex involve residues in the C-terminal basic region of the AREA DBD and the sugar-phosphate backbone of the antisense strand (C18 to C24). In particular, the methyl groups of Leu51 interact with the sugar of C18, while the side-chain of Ile56 interacts with the sugar of A20 (Figure 5C). The positively charged side-chains of Lys57 and Arg61 lie close to the phosphate groups of G11 and C22, respectively, and probably participate in salt bridges (Figure 5C). The side-chain of Arg59 lies in the minor groove (Figure 5C) but the precise location of its guanidino group is not determined by the present data. Based on the regularized mean coordinates for the structure of the complex, the guanidino group of Arg59 lies within close proximity of several hydrogen bond acceptors (the O2 atoms of C18 and T19 and/or the N3 atoms of A8 and G9). Alternatively, the aliphatic portion of the side-chain may simply be providing additional hydrophobic contacts in the minor groove. Interestingly, all of the intermolecular NOEs observed for Arg59, Asn60 and Arg61 involve either H4′ and H5′/H5″ protons of the sugar of T21. Further, each of these residues exhibits NOEs from its backbone NH or H^{α} protons to the sugar protons, suggesting that the polypeptide backbone plays a principal role in orienting and stabilizing the C-terminal tail of the AREA DBD.

Correlation with genetic data

Since extensive mutational data derived from classical genetics performed on *Aspergillus* are available on AREA (Kudla *et al*., 1990; Wiame *et al*., 1985), we have chosen to focus on those residues that interact directly with the DNA, the location of which is illustrated in Figure 3B.

Truncation studies of the C terminus of the AREA protein suggest that the C-terminal limit required for retention of some function coincides with Asn60 or Arg61 of the AREA DBD (Stankovich *et al*., 1993; Platt *et al*., 1996b). The structural results agree with these findings and demonstrate that Arg61 is the last amino acid of the 66-residue AREA DBD to interact with the DNA (Figure 2). Further, the structure provides a rationale for the functional tolerance of the non-conservative amino acid substitutions Lys57 → Glu, Gln or Leu, Arg59 → Leu and Arg61 → Leu (Platt *et al*., 1996a). The lysine at position 57 crosses the minor groove and lies closest to the phosphate group of G11. A Lys57 → Glu substitution would disrupt interactions with the negatively charged phosphate group, but its shorter length could accommodate a potential hydrogen bond with the 2-NH_2 group of G11 in the minor groove. Likewise, introduction of a shorter Gln side-chain at this position might also favor hydrogen bond formation in the minor groove. In the case of both the Lys57 → Glu and Gln mutations no change in backbone conformation would be required to accomodate the proposed alternative contacts. A Lys57 → Leu mutation mimics the aliphatic portion of the Lys57 side-chain and could participate in hydrophobic interactions with the bases and sugars of A10 and G11. In the case of Arg59 and Arg61, intermolecular NOEs are observed between the NH or H^{α} protons of these residues and the sugar-phosphate backbone protons of the DNA (Figures 2A and 6A). Additionally, the side-chain of Arg59 projects into the minor groove, potentially participating in hydrophobic contacts with the A8/T19 or G9/C18 base-pairs; substitution of Arg59 with a leucine would maintain the integrity of these contacts. Collectively, structural and functional studies of AREA suggest that the placement of the C-terminal tail relative to the DNA relies more heavily upon interactions between the protein backbone and DNA than those provided by a particular side-chain.

Mutations associated with loss-of-function include Pro21 → Leu, Arg24 → Leu, Arg24 → Gln, Ala35 → Pro, Leu38 → Phe, Arg47 → His and the deletion of Lys41 (R. A. Wilson, H.N.A. Jr, T. Langdon & K. N. Rand, unpublished; Kudla *et al*., 1990; Platt *et al*., 1996a). Hydrophobic contacts between Pro21 and Ile56 (as evidenced by NOEs between the two methyl groups of Ile56 and the H^{δ} protons of Pro21) are critical for anchoring the C-terminal tail to the zinc finger core. Pro21 also makes important base contacts with A20 and T21 and the Pro21 → Leu mutation might sterically interfere with the side-chain of Leu22, the methyl group of T21 or the phosphate backbone, potentially disrupting DNA binding. Examination of mammalian and avian GATA sequences supports the requirement of a smaller side-chain at this position, which is consistently occupied by the more compact threonine in non-fungal fingers (Figure 1A). Arg24 is another highly conserved residue which lies in the loop region connecting the two β-sheets and is hydrogen bonded to the base of G5 in the major groove (Figure 5A). Given its key role in base-specific recognition, it is not surprising that mutation to Leu, which is incapable of hydrogen bonding, results in loss-of-function. Although Gln retains the capacity to form one of the two hydrogen bonds with either the N7 or O6 of G5, the shorter side-chain would probably require a mediating water to complete a hydrogen bond with G5.

The hydrophobic side-chain of Ala35 stabilizes the α-helix *via* interaction with the methyl groups of Leu38, and participates in DNA recognition *via* contacts with the sugar and base of T19. While the

Ala35 → Pro substitution likely maintains hydrophobic contacts with T19, its propensity for helix formation is low, and it is likely that the geometry of the α-helix and possibly the nearby metal center would be distorted.

Leu38 is highly conserved throughout the GATA family. In the structure of both the AREA and cGATA-1 DBD complexes, Leu38 packs against the bases of T19 and A20 with its χ_1 and χ_2 side-chain torsion angles in the -60° and 180° rotamers, respectively. Given that Leu38 is located in a helix, a phenylalanine in this position can have a χ_1 rotamer of either $\sim -60^\circ$ or $\sim 180^\circ$, with the χ_2 angle centered around $\sim 90^\circ$ (or the stereochemically equivalent -90° conformation; Kuszewski *et al.*, 1997). As a result of the bulky aromatic ring, a χ_1 rotamer of $\sim -60^\circ$ would result in steric clash between the phenylalanine and the side-chain of Arg24 which could only be accomodated by displacement of Arg24 and the concomitant loss of the intermolecular hydrogen bonds between the guanidino group of Arg24 and G5. Similarly, a χ_1 rotamer of $\sim 180^\circ$ would result in steric clash with Leu42, which would perturb the hydrophobic contacts between the methyls of Leu42 and the bases of C18 and T19.

The remaining loss-of-function mutations involve the deletion of Lys41 and an Arg47 to His substitution. Deletion of the highly conserved Lys41 eliminates electrostatic interactions with the phosphate backbone of the DNA and probably interferes with proper termination of the α-helix. Similarly, the Arg47 → His mutation also disrupts the charged triad (Arg25, Lys41, Arg47) which recognizes phosphate groups lining the edges of the major groove. Histidine lacks the charge characteristics typical of arginine at physiological pH, as well as the side-chain length and conformational flexibility to interact effectively with the phosphate backbone.

Comparison to cGATA-1 DBD·DNA complex

For clarity of comparison, all references to the cGATA-1 DBD will refer to the amino acid numbering scheme and sequence alignment presented in Figure 1A. References to DNA interactions for both complexes will refer to the numbering scheme shown in Figure 1C for the CGATAG site.

Direct comparison of the AREA and cGATA-1 (Omichinski *et al.*, 1993a; Tjandra *et al.*, 1997) DBD·DNA complexes indicate that the global fold and gross DNA recognition features of these two class IV zinc finger domains are similar (Figure 7). The 11 hydrophobic residues making key contacts in the major groove are highly conserved with the exception of Thr21 and Tyr39, which are replaced by Pro and Phe, respectively, in the fungal domain (Figure 1A). The replacement of Thr by Pro at position 21 has no apparent effect on the protein backbone conformation and maintains the base contacts with T21 and A20 observed for Thr21 in the cGATA-1 complex. Substitution of Phe for Tyr at position 39 in the AREA complex maintains hydrophobic contacts with the sugar and base of C18, but no longer permits the formation of a hydrogen bond to the phosphate. The remaining sequence substitutions observed for AREA are predominantly found in loop regions away from the major groove.

The most significant difference between the AREA and cGATA-1 DBD·DNA complexes lies in the orientation and backbone dynamics of the C-terminal tails. The C-terminal tail of the AREA DBD (residues 55 to 61) runs parallel with the sugar phosphate backbone along the edge of the minor groove, while that of the cGATA-1 DBD (residues 55 to 64) wraps around the DNA and lies in the minor groove (Figure 7A). A comparative summary of DNA contacts made by both domains (Figure 6A) and of the intermolecular NOEs involving the C-terminal tail (Figure 6B) emphasizes a substantially larger number of minor groove contacts observed for the cGATA-1 DBD relative to the AREA DBD. Indeed, 59 intermolecular NOEs are detected for the C-terminal tail of the cGATA-1 DBD compared to only ten for the C-terminal tail of the AREA DBD (and 13 for the Leu22 → Val mutant AREA DBD; Starich *et al.*, 1998). (Note that NOEs involving residues 57 to 61 of the C-terminal tail are observed for all intermolecular interproton distance contacts less than 3.5 Å in the restrained regularized mean structure of the AREA DBD·DNA, with the exception of a predicted contact of ~2.6 Å between the H^γ proton of Arg59 and the H4′ of A120; observation of this NOE was precluded since the H^γ protons of Arg59 are extensively line broadened and could not be assigned despite the fact that the C^γ resonance of Arg59 could be assigned.) This accounts for most of the difference in the total number of intermolecular NOEs observed for the cGATA-1 DBD·DNA complex (117) *versus* the two AREA DBD·DNA complexes (48 for the wild-type and 58 for the Leu22 → Val mutant). In addition, a substantially smaller number of intermolecular NOEs is observed for Pro21 of the AREA DBD (five and four for the wild-type and Leu22 → Val mutant, respectively) compared to the equivalent Thr in the cGATA DBD (15), owing to the more favorable relaxation properties of the methyl and β-methine groups of Thr relative to the methylene groups of Pro.

A best-fit superposition of the zinc finger core of the AREA and cGATA-1 DBDs (backbone atomic rms of 1.2 Å for residues 10 to 54), results in a backbone atomic rms displacement between the two C-terminal tails (residues 55 to 61) of 3.2 Å. This displacement can be attributed to amino acid substitutions of the highly conserved Gly55 and Lys62 in the mammalian GATA factors to Val and Asn, respectively, in AREA.

The different backbone conformations of the C-terminal tails of the cGATA-1 and AREA DBDs can be pinpointed to a change in the backbone ψ angle of residue 55. Thus, in the cGATA-1 DBD

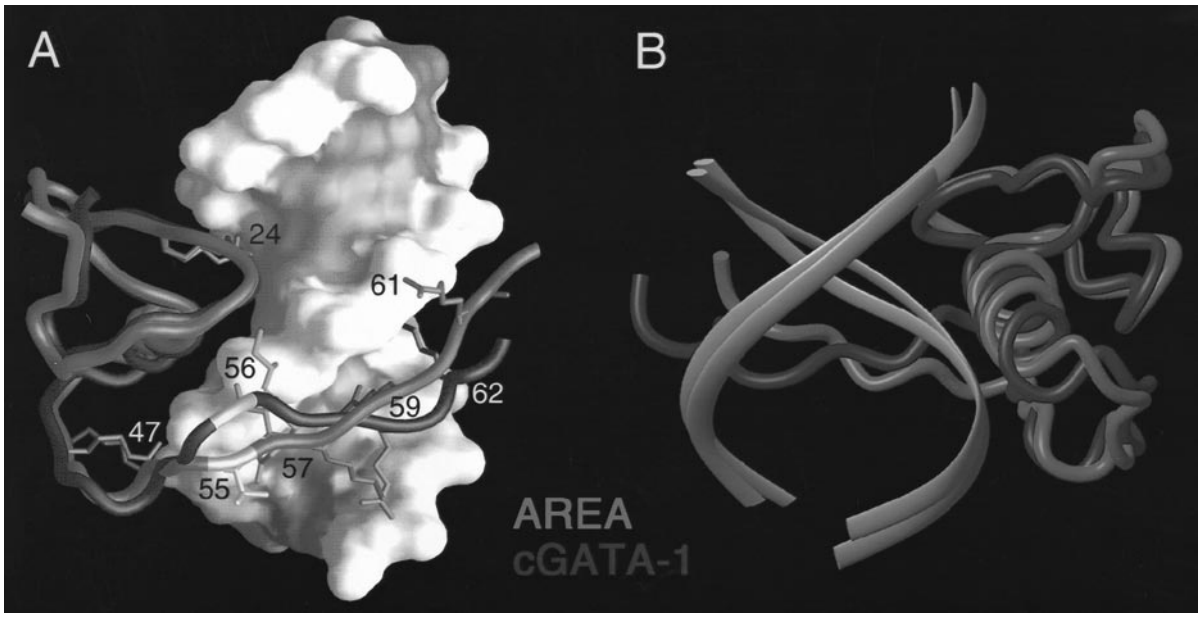

Figure 7. Comparison of the AREA DBD·DNA and cGATA-1 DBD·DNA complexes. A, The core (residues 10 to 54) of the AREA DBD superimposed on the core of the cGATA-1 DBD (backbone atomic rms difference of 1.2 Å). The AREA and cGATA-1 polypeptide backbones are represented as red and green worms, respectively. Side-chains for the AREA and cGATA-1 DBDs are shown in dark blue and yellow, respectively. The surface representation of only one of the DNA models (the CGATAG oligonucleotide calculated for the AREA complex) is shown for clarity. The major groove of the surface is light blue and the minor groove is light red. B, Ribbon representation of the superposition of protein atoms (C, C^{α}, N) and heavy atoms of the DNA (phosphate and sugar atoms) for the AREA and cGATA-1 complexes. The protein backbones are displayed as red (AREA) and green (cGATA-1) worms. The strands of the DNA are shown as ribbons tracing the path of the sugar-phosphate backbone for the AREA target (light red) and the cGATA-1 target (light green). The accession code for the regularized mean coordinates of the cGATA-1 DBD·DNA complex is 2GAT.

Gly55 has ϕ/ψ values of $-95^{\circ}/-180^{\circ}$, while the equivalent residue in AREA, Val55, has ϕ,ψ values of $-110^{\circ}/+128^{\circ}$. In this regard, it is noteworthy that examination of the distribution of ϕ/ψ angles in a database of very high resolution ($\leqslant$1.75 Å or better) crystal structures (Karplus, 1996; Kuszewski *et al.*, 1997), indicates that while the conformations of Gly55 in cGATA-1 and Val55 in AREA occur in highly favorable regions of the Ramachandran plot, the ϕ/ψ region occupied by Gly55 is unpopulated by valine, and likewise the ϕ/ψ region occupied by Val55 is minimally populated by glycine.

Further positioning of the C-terminal tail in the cGATA-1 DBD is provided by the long side-chain of Lys62, which contributes both hydrophobic recognition of the T7/A20 base-pair in the minor groove of the cGATA-1 complex and forms a hydrogen bond between its $N^{\zeta}H_3^+$ and the O2 atom of T7 (Omichinski *et al.*, 1993a). Replacement of this residue with the shorter asparagine side-chain reduces van der Waals contacts and eliminates this crucial hydrogen bond in the AREA complex, as evidenced by the absence of intermolecular NOEs involving the protons of Asn62. Indeed, Arg61 is the last residue of the AREA DBD which exhibits NOEs to the DNA, and these are limited to NOEs from the H^{α}, H^{γ} and H^{δ} protons to the H5/H5″ protons of T21. Direct evidence for a sizable difference in backbone flexibility for the C-terminal tails of the cGATA-1 and AREA DBDs complexed to DNA is afforded by the heteronuclear $^{15}N\{^1H\}$-NOE experiments collected for both complexes (Figure 2D). Thus, while residues 63 to 66 of the AREA DBD exhibit negative $^{15}N\{^1H\}$ NOEs indicative of very large amplitude fast internal motions, the equivalent residues of the cGATA-1 DBD all have positive $^{15}N\{^1H\}$-NOEs. In essence, the additional residue utilized by the cGATA-1 DBD for DNA recognition (Lys62) tethers the C-terminal tail in the minor groove, reducing its overall flexibility and creating slack in the backbone prior to the tethering point.

To accommodate this slack, the backbone of the cGATA-1 DBD buckles near Gly55, forming an Ω loop (Leszcynski & Rose, 1986) comprising residues Arg52 to Gln57. The local flexibility of the backbone in this region is further evidenced by a negative heteronuclear $^{15}N\{^1H\}$-NOE value for Gly55 and a slight decrease in the values for surrounding residues (Asp54, Ile56; Figure. 2D). In contrast, the values of +0.7 to +0.8 for the heteronuclear $^{15}N\{^1H\}$-NOEs exhibited by the equivalent backbone amides in the AREA DBD indicate the absence of large amplitude fast internal motions. This reduction in backbone mobility relative to that observed for the cGATA-1 DBD is attributed to replacement of Gly55 by Val, a side-chain that experiences more van der Waals restrictions on its conformation and participates in a larger number of medium and long-range contacts than glycine.

In addition, the C-terminal tail of the AREA DBD does not form an Ω loop, but instead adopts a more extended conformation that runs parallel with the sugar-phosphate backbone.

Given the high degree of structural similarity observed between the core modules of the AREA and cGATA-1 DBDs, it appears that the reduced affinity of the AREA DBD for CGATA and AGATA elements might be explained, in part, by the amino acid differences and consequent structural changes observed for the basic C-terminal tail. The equilibrium dissociation contants for the binding of the AREA and cGATA-1 DBDs to DNA differ by a factor of about 300, which represents a free energy change of ~3.4 kcal mol^{-1}. The following factors might contribute to this effect: First, a hydrogen bond between the $N^{\zeta}H_3^+$ of Lys62 and the O2 atom of T7 is observed in the cGATA-1 complex, but is absent in the AREA complex, which has an Asn at position 62. Second, the energetic cost of solvating the four methylene groups of Lys (~2.8 kcal mol^{-1}) *versus* the single methylene of Asn (~0.7 kcal mol^{-1}) may drive the longer side-chain into the more hydrophobic environment of the minor groove (Cantor & Schimmel, 1980). Third, the calculated accessible surface area that is buried upon DNA binding for the AREA DBD (~700 $Å^2$) is smaller than that observed for the cGATA-1 DBD (~1000 $Å^2$). This difference of ~300 $Å^2$ is entirely due to the C-terminal tails since the accessible surface area of the zinc finger core (residues 9 to 54) buried upon DNA binding is nearly identical for the two DBDs (~500 $Å^2$).

Concluding remarks

GATA regulatory proteins have been discovered in a wide range of organisms since the founding member of this group (GATA-1) was identified as the major DNA binding protein associated with erythroid development (Weiss & Orkin, 1995). The structural comparison of the single finger AREA DBD·DNA complex with the carboxyl finger complex of cGATA-1 extends our knowledge of the features underlying sequence-specific recognition and the affinity differences observed for this diverse group of transcription factors. The similar global folds observed for the core modules of these domains suggest that the geometric arrangement and types of residues required to distinguish the GATA element from other sequences cannot be compromised. Indeed, only two of the eight residues which recognize major groove bases differ between the AREA and cGATA-1 fingers, highlighting the importance of Leu22, Arg24, Asn34, Ala35, Leu38 and Leu42 for recognition of CGATA and AGATA core elements (Figure 1A). Conservative substitutions for residues 21 and 39 are tolerated well from a structural viewpoint and do not have any apparent effect on the positions of the loop and helix in the major groove.

Recently published biochemical data and the work presented here suggest that variability in affinity and specificity of the GATA factors for their DNA targets may be imparted by the location and characteristics of the basic tail. While the cGATA-1 and AREA DBDs utilize a basic C-terminal tail in addition to the zinc finger core for DNA binding, the N-terminal domains of chicken GATA-2 and GATA-3 possess a zinc finger core module flanked by basic regions on both sides (Pedone *et al.*, 1997). Interestingly, these domains bind only weakly to the consensus AGATAA target but demonstrate specific high affinity binding to AGATCT elements. The N-terminal zinc finger domain of chicken GATA-1, which lacks the N-terminal basic region and shows only 46% sequence identity to the carboxyl cGATA-1 domain, does not bind to either site. Gel-retardation and structural data presented for the AREA DBD represent another example for the general theme of affinity and specificity modulation *via* basic tails attached to the core recognition module. Here the sequence characteristics of the C-terminal tail influence its structure in the complex and hence modulate the affinity of the DBD for GATA sites. The AREA DBD·DNA structure provides molecular insight into how different members of the GATA family might discriminate against available binding sites, and suggests how differential affinity could contribute to GATA factor-mediated transcriptional control.

Materials and Methods

Sample preparation

The coding sequence for the AREA DBD (amino acids 506 to 570 of the AREA regulatory protein) was generated as an *Nde*I-*Bam*HI DNA fragment using the polymerase chain reaction. This DNA fragment was cloned into the *E. coli* vector pET11A and expressed in host strain BL21(DE3). Purification of the AREA DBD followed the same procedure as that used for the cGATA-1 DBD (Omichinski *et al.*, 1993a). Uniform (>95%) ^{15}N and ^{13}C labeling 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 purified AREA DBD was characterized by amino acid analysis and mass spectrometry. The purified AREA DBD was lyophylized, reconstituted with 1.1 equivalents of zinc, and the final pH adjusted to 6.5 with NaOH.

The DNA oligonucleotides used for NMR were purchased from Midland Certified Reagent Co. (Texas) as single-stranded 13 bp oligodeoxynucleotides containing the CGATAG sequence or its complement, characterized by NMR and subsequently annealed at a 1:1 ratio.

The AREA DBD·DNA complex was prepared by slowly adding the AREA DBD (~125 μM protein with zinc bound) to a DNA solution (~125 μM DNA, 12 mM NaCl) until a 1:1 ratio of DNA to AREA DBD was attained. Samples were then concentrated using a Centriprep-3 (Amicon) filtration system to give a final complex concentration of ~2 mM at pH 6.5 with 12 mM NaCl, 2.2 mM $ZnCl_2$ and 5.0 mM NaN_3 in a total volume of 250 μl. Three samples were prepared for NMR studies

and contained ^{15}N AREA DBD·DNA in 90% H_2O/10% ^2H_2O, $^{15}N/^{13}C$ AREA DBD·DNA in 90% H_2O/10% ^2H_2O or $^{15}N/^{13}C$ AREA DBD·DNA in 99.996% ^2H_2O.

NMR spectroscopy

Spectra for the complex were recorded at 25°C on AMX500, DMX500, AMX600, DMX600, DMX750 and AMX360 Bruker spectrometers equipped with *x*, *y*, *z*-shielded gradient triple resonance probes. Details of the multidimensional experiments used, together with the original references, are reviewed elsewhere (Clore & Gronenborn, 1991; Bax & Grzesiek, 1993; Gronenborn & Clore, 1995). 3D double and triple resonance through-bond correlation experiments were used to assign the spectra of the protein (CBCANH, CBCACONH, HBHA-(CO)NH, C(CO)NH, H(CCO)NH, HCCH-COSY, HCCH-TOCSY, HNHA, ^{15}N-separated HOHAHA); and 2D ^{12}C-filtered homonuclear Hartmann-Hahn (in ^2H_2O), ^1H-^1H NOE (with a 1-1 read pulse in H_2O for the imino protons), ^{12}C-filtered NOE (in ^2H_2O) and ^{15}N-filtered NOE (in H_2O) experiments were used to assign the spectrum of the bound DNA using conventional sequential assignment methodology for nucleic acids (Clore & Gronenborn, 1989). Virtually complete assignments for the non-exchangeable protons were obtained for the bound DNA but no distinction was made between the H5′ and H5′′ sugar protons. Three-bond coupling constants ($^3J_{HN\alpha}$, $^3J_{\alpha\beta}$, $^3J_{NH\beta}$, $^3J_{C\gamma N}$, $^3J_{C\gamma CO}$, $^3J_{CC}$) were obtained by 2D and 3D quantitative *J* correlation spectroscopy (Bax *et al*., 1994; Hu & Bax, 1997; Hu *et al*., 1997). Residual one-bond ^{15}N-^1H dipolar couplings with a precision of 0.1 Hz were obtained from a series of *J*-modulated ^1H-^{15}N HSQC spectra recorded in duplicate at 360 and 750 MHz (Tjandra *et al*., 1996, 1997). The experimentally observed values of the dipolar couplings ranged from −1.2 Hz to +1.1 Hz, and the value of the axial component of the magnetic susceptibility tensor χ_a, extracted from the dipolar coulings as described by Tjandra *et al*. (1997), was -22.4×10^{-34} m^3/molecule. Intramolecular NOEs within the protein were obtained from 3D ^{15}N- and ^{13}C-separated NOE spectra and a 3D ^{15}N-separated ROE spectrum; intramolecular NOEs within the DNA from 2D ^1H-^1H NOE (for the imino, amino and H2 protons), ^{15}N-filtered NOE and ^{12}C-filtered NOE spectra; and intermolecular NOEs between the protein and the DNA from a 3D $^{13}C(F_2)$-separated/$^{12}C(F_3)$-filtered NOE spectrum. Heteronuclear $^{15}N\{^1H\}$-NOEs were measured as described by Grzesiek & Bax (1993). Spectra were processed with the NMRPipe package (Delaglio *et al*., 1995), and analyzed using the programs PIPP, CAPP and STAPP (Garrett *et al*., 1991).

Structure calculations

Interproton distance and torsion angle restraints were derived from the NOE and coupling constant data as described by Omichinski *et al*. (1997). Distances involving methyl groups, aromatic ring protons of Tyr and Phe, and non-stereospecifically assigned methylene protons were represented as a $(\Sigma r^{-6})^{-1/6}$ sum (Nilges, 1993). The structures (comprising residues 1 to 66 of the protein, the zinc atom and base-pairs 1 to 13 of the DNA) were calculated by simulated annealing (Nilges *et al*., 1988), exactly as described previously (Omichinski *et al*., 1997), using the program XPLOR-31 (Brünger, 1993), modified to incorporate pseudo-potentials for $^3J_{HN\alpha}$ coupling constants (Garrett *et al*., 1994), secondary $^{13}C^\alpha$ and $^{13}C^\beta$ chemical shifts (Kuszewski *et al*., 1995), residual one-bond $^{15}N^{-1}H$ dipolar couplings (Tjandra *et al*., 1997; Clore *et al*., 1998), and a conformational database potential for both proteins and nucleic acids (Kuszewski *et al*., 1996, 1997). Non-bonded contacts are represented by a quartic van der Waals repulsion term (Nilges *et al*., 1988). No hydrogen bonding, electrostatic or 6-12 Lennard-Jones empirical potential energy terms are present in the target function used for simulated annealing or restrained minimization.

Structural DNA parameters were analyzed using the program CURVES (Lavery & Sklenar 1989). Surface accessible areas were computed with a probe radius of 1.4 Å using XPLOR. Figures were generated with the programs MOLMOL (Koradi *et al*., 1996) and GRASP (Nicholls *et al*., 1991).

The coordinates of the cGATA-1 DBD·DNA complex used for comparison were obtained by simulated annealing refinement of the coordinates of Omichinski *et al*. (1993a) incorporating both residual one-bond ^{15}N-^1H and $^{13}C^\alpha$-^1H dipolar couplings (Tjandra *et al*., 1997) and a conformational database potential for proteins and nucleic acids (Kuszewski *et al*., 1996, 1997; PDB accession codes 2GAT, 3GAT and 2GATMR). The rms difference between the mean coordinate positions for these coordinates relative to those published by Tjandra *et al*. (1997) is small (∼0.4 Å for the protein backbone, ∼0.6 Å for protein heavy atoms, for the DNA heavy atoms and for the protein backbone plus the DNA heavy atoms, and ∼0.7 Å for protein plus DNA heavy atoms, using residues 2 to 59 of the protein and base-pairs 6 to 13 of the DNA in the numbering scheme of Omichinski *et al*. (1993a) which corresponds to residues 7 to 64 and base-pairs 4 to 11 in the current numbering scheme) and within the error of the coordinates (∼0.7 Å for the protein backbone, the DNA, and the protein backbone plus DNA; ∼1 Å for protein plus DNA heavy atoms, and ∼1.2 Å for protein heavy atoms). The only difference relative to the coordinates of Tjandra *et al*. (1997) is a small increase in the percentage of residues in the most favorable region of the Ramachadran plot (from 79% to 83%) and a small decrease in the number of bad contacts per 100 residues (from ∼10 to ∼6).

The coordinates of the 35 final simulated annealing structures of the AREA DBD·DNA complex, together with the coordinates of the restrained regularized mean structure, $\langle SA \rangle r$, and the complete list of experimental NMR restraints have been deposited in the Brookhaven Protein Data Bank (PDB accession codes 4GAT, 5GAT and 4GATMR).

Acknowledgements

The authors thank D. S. Garrett and F. Delaglio for software support; A. Murphy for amino acid analysis; L. Pannell for MS analysis; N. Tjandra for help in measuring the residual dipolar couplings; R. Tschudin for technical support; and A. Bax, J. Huth, J. Kuszewski, J. Omichinski, T. Strzelecka, N. Tjandra and C. Trainor for useful discussions. M.W. gratefully acknowledges a post-doctoral fellowship from the Swedish Natural Sciences Research Council (NFR). 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.).

References

Arst, H. N., Jr & Cove, D. J. (1973). Nitrogen metabolite repression in *Aspergillus nidulans*. *Mol. Gen. Genet.* **126**, 111–141.

Aurora, R., Srinivasan, R. & Rose, G. D. (1994). Rules for α-helix termination by glycine. *Science,* **264**, 1126–1130.

Bax, A. & Grzesiek, S. (1993). Methodological advances in protein NMR. *Acc. Chem. Res.* **26**, 131–138.

Bax, A., Vuister, G. W., Grzesiek, S., Delaglio, F., Wang, A. C., Tschudin, R. & Zhu, G. (1994). Measurement of homo- and heteronuclear J couplings from quantitative *J* correlation. *Methods Enzymol.* **239**, 79–125.

Brünger, A. T. (1993). *XPLOR: A System for X-ray Crystallography and NMR,* Yale University Press, New Haven.

Cantor, C. & Schimmel, P. (1980). *Biophysical Chemistry Part I: The Conformation of Biological Macromolecules,* W. H. Freeman & Co., New York.

Clore, G. M. & Gronenborn, A. M. (1989). Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy. *CRC Crit. Rev. Biochem. Mol. Biol.* **24**, 479–564.

Clore, G. M. & Gronenborn, A. M. (1991). Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. *Science,* **252**, 1390–1399.

Clore, G. M., Gronenborn, A. M. & Tjandra, N. (1998). Direct structure refinement against residual dipolar couplings in the presence of rhombicity of unknown magnitude. *J. Magn. Reson.* In the press.

Crawford, N. M. & Arst, H. N., Jr (1993). The molecular genetics of nitrate assimilation in fungi and plants. *Annu. Rev. Genet.* **27**, 115–146.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR.* **6**, 277–293.

Ernst, R. R., Bodenhausen, G. & Wokaun, A. (1987). *Principles of Nuclear Magnetic Resonance in One and Two Dimensions,* Clarendon Press, Oxford.

Feng, B., Xiao, X. & Marzluf, G. A. (1993). Recognition of specific nucleotide bases and cooperative DNA binding by the *trans*-acting nitrogen regulatory protein NIT2 of *Neurospora crassa*. *Nucl. Acids Res.* **21**, 3989–3996.

Garrett, D. S., Powers, R., Gronenborn, A. M. & Clore, G. M. (1991). 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.

Garrett, D. S., Kuszewski, J., Hancock, T. J., Lodi, P. J., Vuister, G. W., Gronenborn, A. M. & Clore, G. M. (1994). The impact of direct refinement against three-bond HN-CαH coupling constants on protein structure determination by NMR. *J. Magn. Reson. ser. B,* **104**, 99–103.

Gorfinkiel, L., Diallinas, G. & Scazzocchio, C. (1993). Sequence and regulation of the *uap*A gene encoding a uric acid-xanthine permease in the fungus *Aspergillus nidulans*. *J. Biol. Chem.* **268**, 23376–23381.

Gronenborn, A. M. & Clore, G. M. (1989). Analysis of the relative contributions of the nuclear Overhauser interproton distance restraints and the empirical energy function in the calculation of oligonucleotide structures using restrained molecular dynamics. *Biochemistry,* **28**, 5978–5984.

Gronenborn, A. M. & Clore, G. M. (1995). Structures of protein complexes by multidimensional heteronuclear magnetic resonance spectroscopy. *CRC Crit. Rev. Biochem. Mol. Biol.* **30**, 351–385.

Grzesiek, S. & Bax, A. (1993). The importance of not saturating H_2O in protein NMR: application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc.* **115**, 12593–12594.

Henry, G. D. & Sykes, B. D. (1995). Determination of the rotational dynamics and pH dependence of the hydrogen exchange rates of the arginine guanidino group using NMR spectroscopy. *J. Biomol. NMR,* **5**, 59–66.

Hu, J. S. & Bax, A. (1997). χ_1 angle information from a simple two-dimensional NMR experiment that identifies trans $^3J_{NC\gamma}$ couplings in isotopically enriched proteins. *J. Biomol. NMR,* **9**, 323–328.

Hu, J. S., Grzesiek, S. & Bax, A. (1997). Two-dimensional NMR methods for determining χ_1 angles of aromatic residues in proteins from three-bond $J_{C'C\gamma}$ and $J_{NC\gamma}$ couplings. *J. Am. Chem. Soc.* **119**, 1803–1804.

Karplus, P. A. (1996). Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.* **5**, 1406–1420.

Ko, L. J. & Engel, D. (1993). DNA-binding specificities of the GATA transcription factor family. *Mol. Cell. Biol.* **13**, 4011–4022.

Koradi, R., Billeter, M. & Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.

Kudla, B., Caddick, M. X., Langdon, T., Martinez-Rossi, N. M., Bennett, C. F., Sibley, S., Davies, R. W. & Arst, H. N., Jr (1990). The regulatory gene *are*A mediating nitrogen metabolite repression in *Aspergillus nidulans*. Mutations affecting specificity of gene activation alter a loop residue of a putative zinc finger. *EMBO J.* **9**, 1355–1364.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1995). The impact of direct refinement against $^{13}C\alpha$ and $^{13}C\beta$ chemical shifts on protein structure determination by NMR. *J. Magn. Reson. ser. B,* **106**, 92–96.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1996). 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.

Kuszewski, J., Gronenborn, A. M. & Clore, G. M. (1997). Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *J. Magn. Reson.* **125**, 171–177.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallog.* **26**, 283–291.

Lavery, R. & Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dynam.* **6**, 655–667.

Leszcynski, J. F. & Rose, G. D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science,* **234**, 849–855.

Martin, D. I. K. & Orkin, S. H. (1990). Transcriptional activation and DNA binding by the erythroid factor GF-1/NF-E1/Eryf1. *Genes Dev.* **4**, 1886–1898.

Merika, M. & Orkin, S. H. (1993). DNA-binding specificity of GATA family transcription factors. *Mol. Cell. Biol.* **13**, 3999–4010.

Nichols, A. J., Sharp, K. & Honig, B. (1991). Protein folding and association: insights from interfacial and

thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281–296.
Nilges, M. (1993). A calculational strategy for the structure determination of symmetric dimers by ^{1}H NMR. *Proteins: Struct. Funct. Genet.* **17**, 295–309.
Nilges, M., Clore, G. M. & Gronenborn, A. M. (1988). Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry- dynamical simulated annealing calculations. *FEBS Letters,* **229**, 317–324.
Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. J. & Gronenborn, A. M. (1993a). NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. *Science,* **261**, 438–446.
Omichinski, J. G., Trainor, C., Evans, T., Gronenborn, A. M., Clore, G. M. & Felsenfeld, G. (1993b). A small single-''finger'' peptide from the erythroid transcription factor GATA-1 binds specifically to DNA as a zinc or iron complex. *Proc. Natl Acad. Sci. USA,* **90**, 1676–1680.
Omichinski, J. G., Pedone, P. V., Felsenfeld, G., Gronenborn, A. M. & Clore, G. M. (1997). The solution structure of a specific GAGA factor-DNA complex reveals a modular binding mode. *Nature Struct. Biol.* **4**, 122–132.
Pedone, P. V., Omichinski, J. G., Nony, P., Trainor, C., Gronenborn, A. M., Clore, G. M. & Felsenfeld, G. (1997). The N-terminal fingers of chicken GATA-2 and GATA-3 are independent sequence-specific DNA binding domains. *EMBO J.* **16**, 2874–2882.
Platt, A., Langdon, T., Arst, H. N., Jr., Kirk, D., Tollervey, D., Sanchez, J. M. M. & Caddick, M. X. (1996a). Nitrogen metabolite signalling involves the C-terminus and the GATA domain of the *Aspergillus* transcription factor AREA and the 3′ untranslated region of its mRNA. *EMBO J.* **15**, 2791–2801.
Platt, A., Ravagnani, A., Arst, H. N., Jr., Kirk, D., Langdon, T. & Caddick, M. X. (1996b). Mutational analysis of the C-terminal region of AREA, the transcription factor mediating nitrogen metabolite repression in *Aspergillus nidulans*. *Mol. Gen. Genet.* **250**, 106–114.
Ravagnani, A., Gorfinkiel, L., Langdon, T., Diallinas, G., Adjadj, E., Demais, S., Gorton, D., Arst, H. N., Jr & Scazzocchio, C. (1997). Subtle hydrophobic interactions between the seventh residue of the zinc finger loop and the first base of an HGATAR sequence determine promoter-specific recognition by the *Aspergillus nidulans* GATA factor AreA. *EMBO J.* **16**, 3974–3986.
Saenger, W. (1984). *Principles of Nucleic Acid Structure,* Springer-Verlag, New York.
Stankovich, M., Platt, A., Caddick, M. X., Langdon, T., Shaffer, P. M. & Arst, H. N., Jr (1993). C-terminal truncation of the transcriptional activator encoded by *are*A in *Aspergillus nidulans* results in both loss-of-function and gain-of-function phenotypes. *Mol. Microbiol.* **7**, 81–87.
Starich, M. R., Wikström, M., Schumacher, S., Arst, H. N., Jr., Gronenborn, A. M. & Clore, G. M. (1998). The solution structure of the Leu22 → Val mutant AREA DNA binding domain complexed with a TGATAG core element defines a role for hydrophobic packing in the determination of specificity. *J. Mol. Biol.* **277**, 621–634.
Tjandra, N., Grzesiek, S. & Bax, A. (1996). Magnetic field dependence of nitrogen-proton J splittings in ^{15}N-enriched human ubiquitin resulting from relaxation interference and residual dipolar coupling. *J. Am. Chem. Soc.* **118**, 6264–6272.
Tjandra, N., Omichinski, J. G., Gronenborn, A. M., Clore, G. M. & Bax, A. (1997). Use of dipolar ^{1}H-^{15}N and ^{1}H-^{13}C couplings in the structure determination of magnetically oriented macromolecules in solution. *Nature Struct. Biol.* **4**, 732–738.
Weiss, M. J. & Orkin, S. H. (1995). GATA transcription factors: key regulators of hematopoiesis. *Exptl Hematol.* **23**, 99–107.
Wiame, J. M., Grenson, M. & Arst, H. N., Jr (1985). Nitrogen catabolite repression in yeasts and filamentous fungi. *Advan. Microb. Physiol.* **26**, 1–88.

Edited by P. E. Wright

(Received 25 November 1997; received in revised form 30 December 1997; accepted 6 January 1998)