THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 284, NO. 9, pp. 5827–5835, February 27, 2009

Structural Analysis of MED-1 Reveals Unexpected Diversity in the Mechanism of DNA Recognition by GATA-type Zinc Finger Domains*[S]

Received for publication, November 18, 2008, and in revised form, December 17, 2008 Published, JBC Papers in Press, December 18, 2008, DOI 10.1074/jbc.M808712200

Jason A. Lowry[‡1], Roland Gamsjaeger[‡1], Sock Yue Thong[‡], Wendy Hung[§], Ann H. Kwan[‡], Gina Broitman-Maduro[§], Jacqueline M. Matthews[‡], Morris Maduro[§], and Joel P. Mackay[‡2]

From the [‡]School of Molecular and Microbial Biosciences, University of Sydney, New South Wales 2006, Australia and the [§]Department of Biology, University of California, Riverside, California 92521

MED-1 is a member of a group of divergent GATA-type zinc finger proteins recently identified in several species of *Caenorhabditis*. The *med* genes are transcriptional regulators that are involved in the specification of the mesoderm and endoderm precursor cells in nematodes. Unlike other GATA-type zinc fingers that recognize the consensus sequence (A/C/T)GATA(A/G), the MED-1 zinc finger (MED1zf) binds the larger and atypical site GTATACT(T/C)$_3$. We have examined the basis for this unusual DNA specificity using a range of biochemical and biophysical approaches. Most strikingly, we show that although the core of the MED1zf structure is similar to that of GATA-1, the basic tail C-terminal to the zinc finger unexpectedly adopts an α-helical structure upon binding DNA. This additional helix appears to contact the major groove of the DNA, making contacts that explain the extended DNA consensus sequence observed for MED1zf. Our data expand the versatility of DNA recognition by GATA-type zinc fingers and perhaps shed new light on the DNA-binding properties of mammalian GATA factors.

Members of the GATA family of transcription factors contain either one or two type IV zinc fingers (1). In mammals, six of these proteins, namely GATA-1–6, have been studied in some detail (2–7), and functional data reveal that they are essential for the development of a range of different tissues. These six proteins are highly conserved in vertebrates, and functional GATA family transcriptional regulators have also been described in more distant organisms: for example, ELT-1 is essential for epidermal specification in *Caenorhabditis elegans* (8), whereas AreA and Nit2 regulate nitrogen metabolism in fungi (9, 10).

It is generally recognized that GATA-type zinc fingers bind to double-stranded DNA that contains the consensus sequence (A/C/T)GATA(A/G) (1, 11–13). To date, the three-dimensional structures of three GATA-type zinc fingers in complex with DNA have been determined: the C-terminal zinc finger (C-finger)[3] of chicken GATA-1 (14), the C-finger of murine GATA-3 (15), and the zinc finger of *Aspergillus nidulans* AreA (16). Each has revealed a protein core that consists of two N-terminal β-hairpins and an α-helix; both the helix and the loop connecting the two hairpins make multiple contacts with the major groove of the DNA. In addition, each finger contains a C-terminal tail of ~20 residues that is rich in basic residues. These tails take up similar conformations in the three structures, "folding back" and wrapping around the DNA in an extended fashion. In each case, the tail makes a number of mostly nonspecific contacts with the minor groove and/or the sugar phosphate backbone.

MED-1 is a *C. elegans* GATA family transcription factor that plays a central role in the specification of the mesoderm and endoderm tissues by activating the expression of a range of downstream target genes (17). The protein contains a single zinc finger that shares ~53% sequence similarity with the chicken GATA-1 C-finger over the core zinc finger structure and 46% similarity when the basic tail regions are included (see Fig. 1*A*). Although it was anticipated that MED-1 would bind the consensus GATA sequence, Broitman-Maduro *et al.* (17) demonstrated that the MED-1 zinc finger (MED1zf) specifically recognizes a larger and somewhat divergent site, GTATACT(T/C)$_3$.

We have examined the molecular basis for this unusual DNA-binding specificity. Our data indicate that a single arginine residue in the core zinc finger domain drives a specificity switch from AGATA to GTATA and that, unexpectedly, the C-terminal basic tail of MED1zf forms an additional well ordered α-helix that interacts with the 3′-pyrimidine-rich extension in the DNA consensus sequence. These results underscore the versatility of DNA recognition by zinc finger domains and have implications for the functions of the many GATA family proteins for which DNA-binding specificities have not yet been defined in detail.

* This work was supported by a program grant and a senior research fellowship (to J. P. M.) from the Australian National Health and Medical Research Council. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Figs. S1–S3 and Tables 1 and 2.

The atomic coordinates and structure factors (code 2KAE) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

[1] Both authors contributed equally to this work.

[2] To whom correspondence should be addressed. Tel.: 61-2-9351-3906; Fax: 61-2-9351-4726; E-mail: j.mackay@usyd.edu.au.

[3] The abbreviations used are: C-finger, C-terminal zinc finger; MED1zf, MED-1 zinc finger; ITC, isothermal titration calorimetry; RDCs, residual dipolar couplings; HSQC, heteronuclear single quantum coherence; NOESY, nuclear Overhauser effect spectroscopy; dsDNA, double-stranded DNA; DTT, dithiothreitol.

EXPERIMENTAL PROCEDURES

Preparation of Proteins—Residues 108–174 of MED-1 were overexpressed as a His-tagged fusion protein in pET15b. The use of *Escherichia coli* Rosetta2 cells was essential for obtaining reliable overexpression. Cell pellets were lysed in buffer containing 50 mM sodium phosphate, pH 8.0, 300 mM NaCl, 10 mM imidazole, 1 mM dithiothreitol (DTT), and one complete EDTA-free protease inhibitor mixture tablet (per 50 ml of lysis buffer; Roche Applied Science). His_6-MED1zf was recovered from the soluble fraction and purified on nickel-nitrilotriacetic acid-agarose (Qiagen). The fusion tag was cleaved using thrombin (3–4 h at 25 °C) in 50 mM Tris, pH 8.0, 150 mM NaCl, 2.5 mM $CaCl_2$, and 1 mM DTT, and MED1zf was further purified by size exclusion chromatography (HiLoad 16/60 Superdex 75, Amersham Biosciences). The purified polypeptide contained an additional four N-terminal amino acids (GSHM) derived from the thrombin cleavage site. ^{15}N- and ^{15}N,^{13}C-labeled MED1zf were prepared following the procedure of *Cai et al.* (18) and purified as described above. A range of MED1zf point mutants was created using site-directed mutagenesis and purified as described above, and their correct folding was confirmed by one-dimensional ^1H NMR spectroscopy.

Preparation of DNA—For isothermal titration calorimetry (ITC), complementary 20-mers containing the MED1zf-binding site (5′-CAAAAGGTATACTTTTCCGT-3′ and its complement; Sigma Genosys) were resuspended in 10 mM Tris, pH 8.0; combined at equimolar concentrations; annealed at 95 °C for 10 min; and then slowly cooled to room temperature. The annealed duplex was further purified by size exclusion chromatography (Superdex 75). For NMR, a palindromic DNA oligonucleotide was designed (5′-CGGAAAAGTATACTTTTCCG-3′), annealed, and purified as described above. For surface plasmon resonance, a biotinylated 25-mer (5′-GGACCCCGTATACTTTTCCGGAGAG-3′; Sigma Genosys) was annealed with a complementary 20-mer (5′-CGGAAAAGTATACGGGGTCC-3′) and purified as described above.

Isothermal Titration Calorimetry—Purified MED1zf and double-stranded DNA (dsDNA) were dialyzed against 20 mM sodium phosphate, 40 mM NaCl, and 1 mM DTT, pH 6.5. The MED1zf concentration was determined by absorbance at 280 nm using the calculated extinction coefficient (ϵ = 9770 M^{-1} cm^{-1}). The DNA concentration was determined by absorbance at 260 nm using ϵ = 219,410 M^{-1} cm^{-1}. The dsDNA (90 μM) was titrated into MED1zf (14 μM, 1.4 ml) at 25 °C in a series of 10 μM injections, with a 4-min interval between injections, using a MicroCal VP-ITC microcalorimeter. The evolved heats were integrated and normalized for protein concentration. After base-line correction (using data from a titration of dsDNA into buffer), the data were fitted to a single-site model using Origin 5.0 (MicroCal).

Surface Plasmon Resonance—Kinetic analysis was performed on a Biacore 3000 surface plasmon resonance instrument (Biacore AB, Uppsala, Sweden). Biotinylated dsDNA was immobilized on a streptavidin-coated SA sensor chip (Biacore AB). The buffer used for all experiments was 30 mM sodium phosphate, 200 mM NaCl, 1 mM DTT, and 0.01% surfactant P20. The chip was pretreated according to the manufacturer's instructions with conditioning solution (3× 100-μl injections at 50 μl/min with 50 mM NaOH and 1 M NaCl). The biotinylated dsDNA was diluted to 10 nM and injected onto one of the sensor chip channels (Fc-2) at a flow rate of 5 μl/min for 5 min, resulting in an immobilization level of ~120 response units. The sensor chip was then washed with running buffer. Upstream unmodified channel surfaces were used for reference subtraction. Kinetic measurements at protein concentrations across the range 1 nM to 1 mM (40 μl) were performed at 25 °C with a KINJECT protocol and a flow rate of 20 ml/min. Wild type and mutant protein samples were sampled alternately, zero concentration samples were included for double-referencing, and three cycles were performed. The chip surface was regenerated with 10 mM Tris, pH 7.5, 500 mM NaCl, 1 mM EDTA, and 0.005% SDS between each set of protein samples. Data analysis was performed with BIAevaluation software (Biacore AB).

NMR Spectroscopy—Purified MED1zf was dialyzed into buffer containing 20 mM sodium phosphate, 40 mM NaCl, and 1 mM DTT, pH 6.5, to final concentrations of up to ~0.8 mM. NMR samples also contained 10 μM 2,2-dimethyl-2-silapentane-5-sulfonic acid as a chemical shift reference and 5–10% (v/v) D_2O. Spectra were recorded at 298 K on Bruker 600- and 800-MHz spectrometers equipped with cryoprobes. Samples of MED1zf mutants were prepared similarly. All homonuclear two-dimensional data were collected and analyzed as described (19). Mixing times were 70 and 100 ms for total correlation and nuclear Overhauser effect spectra, respectively. ^{15}N and ^{13}C chemical shift assignments were made from the standard suite of triple resonance experiments as described previously (20). NOE-derived distance restraints were obtained from two-dimensional NOESY and three-dimensional ^{15}N-separated NOESY. ϕ and ψ restraints were included on the basis of an analysis of backbone chemical shifts in the program TALOS (21) and from $^3J_{HNHA}$ scalar couplings measured in HNHA (22). One-bond HN residual dipolar couplings (RDCs) were recorded for the MED1zf·DNA complex in 6 mg of Pf1 phage (ASLA Biotech) using the in-phase/anti-phase pulse sequence (23). Correct alignment of the complex was checked by measuring the D_2O splitting (12–15 Hz). The program PALES (24) was used for the calculation of the magnitude and orientation of the sterically induced alignment tensor (see below for details). All NMR data were processed using TOPSPIN (Bruker, Karlsruhe, Germany) and analyzed with SPARKY 3 (25).

Structure Calculations—The program HADDOCK (26, 27) was used to calculate a data-driven model of the MED1zf·DNA complex. The starting structures for the docking were a B-form model of the DNA fragment (5′-CGGAAAAGTATACTTTTCCG-3′) constructed by using the program 3DNA (28) running on the three-dimensional Dart server located at the Centre for Biomolecular Research at Utrecht University (The Netherlands). A model of the MED1zf structure (residues 111–166, excluding the unstructured N and C termini) was constructed based on the structure of chicken GATA-1. Residues 152–164, which are largely extended in the GATA-1 C-finger, were formed into an additional helix based on TALOS predictions, NOE data, and $^3J_{HNHA}$ coupling constants. This structure was energy-minimized in CNS (three rounds of 200 steps with the backbone fixed) (29) and used as input for HADDOCK. During HADDOCK runs, MED1zf residues 114–146 (zinc finger core)

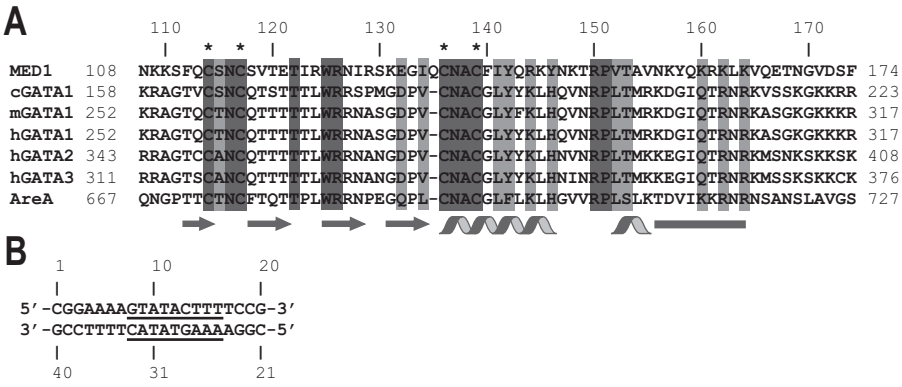

FIGURE 1. **Sequences used in this study.** *A*, amino acid sequences of the GATA-type zinc fingers from *C. elegans* MED-1, chicken (*c*) GATA-1, murine (*m*) GATA-1, human (*h*) GATA-1/2/3, and *A. nidulans* AreA. *Light* and *dark gray* areas indicate conserved and identical residues, respectively. The zinc ligands are indicated with *asterisks*, and the secondary structure of the chicken GATA-1 zinc finger is indicated. Numbering is for full-length proteins. *B*, sequence of the 20-bp oligonucleotide used in this work. The MED-1-binding site is *underlined*.

and 152–162 (helix 2) were defined as semiflexible, and residues 111–113 (N terminus), 147–151 (linker between the two helices), and 163–166 (C terminus) were defined as flexible. Within the semiflexible regions, the following residues were selected as "active" for the definition of ambiguous interaction restraints: 120–125, 138–146 (with the exception of Cys^{139}, a zinc-ligating residue), and 152–162. These definitions were based on chemical shift perturbation data (see Fig. 4, *A* and *B*). For the DNA, ambiguous interaction restraints were defined solely from the unique base atoms of bases 7–16 and 25–34 of the core region and the 3′-flanking thymine stretch to suitable atoms (*i.e.* from hydrogen to nitrogen or oxygen and vice versa) (30, 31) of the above-mentioned residues of MED1zf based on the gel-shift experiments (see Fig. 3) and the chemical shift changes observed (see Fig. 4, *C* and *D*). Because the used DNA sequence is palindromic, it is not immediately clear on which strand the binding occurs; however, a few intermolecular NOEs (supplemental Fig. S1*A*) allowed us to unambiguously define the binding site on the DNA (3′-flanking thymine stretch of bases 14–16 and 25–27 rather than 5–7 and 34–37). In addition to the ambiguous interaction restraints, dihedral angle restraints, all unambiguously assigned intra- and intermolecular NOEs (upper distance limit of 6 Å), restraints for the tetragonal coordination of the zinc atom (32), and restraints to maintain the conformation of the DNA were used as inputs into HADDOCK (supplemental Table 1). After rigid-body minimization, semiflexible annealing, and water refinement, the best 10 structures within the lowest energy cluster (cutoff = 7.5 Å) (26, 27) were used as input for another semiflexible annealing and water refinement step. The resulting best 10 structures (cutoff = 1.5 Å) of the lowest energy cluster overlaid with a root mean square deviation of 0.72 Å (backbone of MED1zf residues 114–162 and DNA). At this stage, H_N-N RDCs were introduced as direct restraints (using the SANI statement); axial and rhombic components of the alignment tensor (D_a and D_r) were calculated using the ensemble of 10 structures and the software PALES (24). Semiflexible annealing and water refinement, as described above, were then carried out once using the RDCs as well as all other mentioned restraints, and the alignment tensor was recalculated based on the resulting best 10 structures (cutoff = 1.5 Å), and HADDOCK was run again (see above) using these new values. After this run, the tensor components of the final 10 structures (cutoff = 1.0 Å) were not significantly different from the ones from the previous stage. The root mean square deviation dropped from 0.72 Å to a final value of 0.40 Å. The final family of 10 structures was analyzed using standard HADDOCK protocols (supplemental Table 2); no major violations were observed. The structures have been deposited in the Protein Data Bank with accession code 2KAE.

RESULTS

The MED-1 Zinc Finger Binds DNA with 1:1 Stoichiometry—To examine the DNA-binding properties of MED1zf, we overexpressed and purified a peptide corresponding to residues 108–174 of *C. elegans* MED-1 (Fig. 1*A*), which includes the core zinc finger region and the basic tail. Given the large predicted size of the recognition site for MED1zf compared with GATA-1, as well as its partially palindromic nature, we first sought to establish the stoichiometry of the MED1zf·DNA complex. We therefore subjected the purified protein to size exclusion chromatography incorporating a multiangle laser light-scattering detector. The observed peak (Fig. 2*A*, *upper panel*) yielded a molecular mass of 8.4 kDa (theoretical mass of 8.3 kDa), demonstrating that this domain is monomeric in solution. We performed the same procedure with a 20-bp dsDNA molecule that contained the MED1zf-binding site (Fig. 2, *middle panel*). To determine the stoichiometry of the protein-DNA interaction, we mixed equimolar amounts of MED1zf and dsDNA. A multiangle laser light-scattering analysis estimated a molecular mass of 20 kDa (theoretical mass of 20.5 kDa), indicating a 1:1 interaction (Fig. 2*A*, *lower panel*).

Having established the stoichiometry of the interaction, we used ITC to determine the affinity of the complex. The interaction is exothermic ($\Delta H = -11$ kcal/mol), and the data fit well to a 1:1 binding model with a dissociation constant of 13 nM (Fig. 2*B*). This affinity is very similar to the dissociation constant measured for the GATA-1·DNA complex (8 nM) (33).

MED-type Zinc Fingers Recognize a Different Consensus Sequence—Although the MED-1 zinc finger has a substantial degree of sequence similarity to the mammalian GATA-type zinc fingers, we previously demonstrated that MED1zf specifically recognized the variant consensus sequence $GTATACT(T/C)_3$ (17). To determine which bases were critical for the MED1zf-DNA interaction, we carried out electrophoretic mobility shift assays using ^{32}P-labeled wild type DNA and mutant DNA oligonucleotides as competitors.

The first issue we addressed was whether the pyrimidine-rich region that flanks the core GTATAC sequence was important for binding. Because the core of the binding site is palindromic (GTATAC), there is the potential for MED1zf to bind the DNA

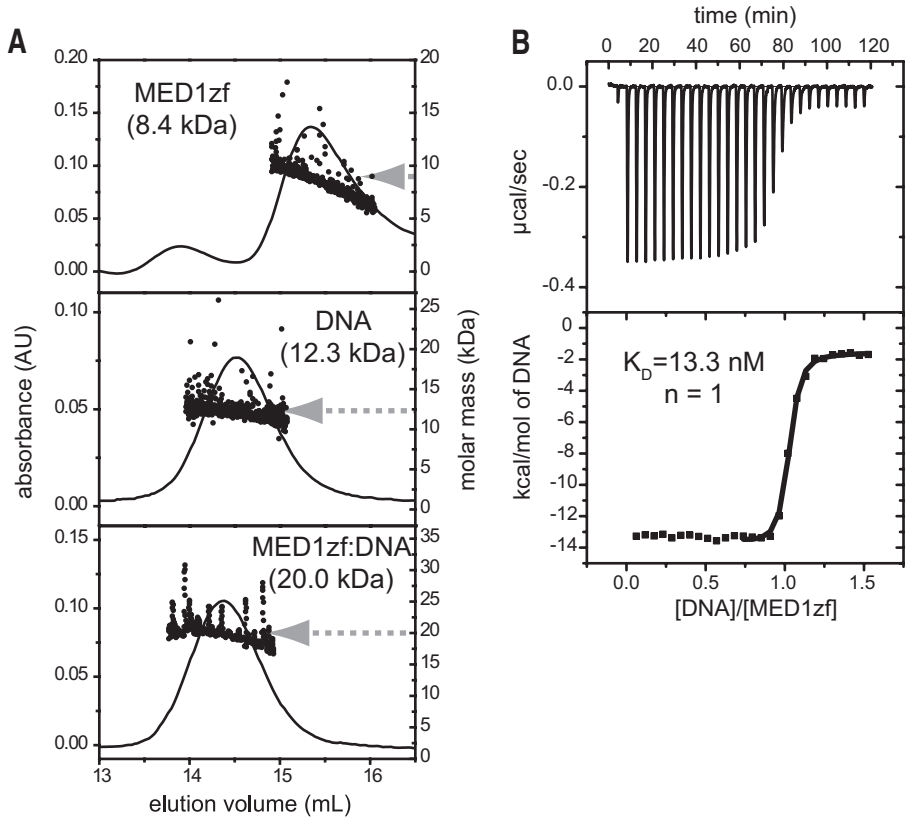

FIGURE 2. **MED1zf binds DNA in a 1:1 fashion with nanomolar affinity.** *A*, Size exclusion and multiangle laser light-scattering data. Each panel shows a size exclusion chromatography trace monitored by A_{280} and multiangle laser light-scattering data. *Upper panel*, MED1zf alone; *middle panel*, DNA alone; *lower panel*, 1:1 mixture of MED1zf and DNA. *B*, ITC data for the MED1zf·DNA complex. The fit to a simple 1:1 binding model is shown. *AU*, absorbance units.

site in two orientations. We therefore created two completely palindromic sequences in which the core GTATAC on both strands was flanked on the 3′ side by either TTTT (the preferred sequence from the published site selection data) or GGGG. As shown in Fig. 3*A*, only AAAAGTATACTTTT could compete with the labeled probe (*lane 4*), whereas the other unlabeled probe (CCCCGTATACGGGG; *lane 5*) failed to compete, even at a 150-fold molar excess. This indicated that MED1zf binds the labeled probe (CCCCGTATACTTTT) in only one orientation and that the 3′-pyrimidine-rich stretch is an important component of the MED1zf recognition site. This latter finding is consistent with the observation that MED1zf target genes could be predicted based on the DNA consensus sequence (17). It is also notable that probes containing a consensus GATA site (TTATCA, corresponding to TGATAA) were unable to compete with the MED1zf sequence (Fig. 3, *A*, *lane 6*; and *B*, *lane 4*).

Further mutagenesis was carried out within the core and flanking regions, whereby we substituted at each position thymine for guanine and adenine for cytosine and vice versa (Fig. 3*A*, *lanes 7–21*). In a second experiment (Fig. 3*B*), an oligonucleotide with a slightly different sequence 5′ of the GTATA site (thymines instead of cytosines) was used, and each thymine was substituted by cytosine and each adenine by guanine and vice versa (*lanes 5–16*). Mutagenesis of each of the three 3′-flanking thymines resulted in slight to moderate reductions in binding affinity, whereas mutagenesis of the flanking 5′-cytosines (Fig. 3*A*) or 5′-thymines (Fig. 3*B*) gave rise to essentially no changes. All substitutions within the core palindrome, with the exception of the first thymine in Fig. 3*A* (*lane 16*), had a large impact on binding. Overall, these data suggest that the core region (GTATAC) as well as the 3′-flanking thymines constitute the binding site of MED1zf.

Mapping the MED1zf-DNA Interaction by NMR Spectroscopy—To gain insight into the structural basis of the MED1zf-DNA interaction, we analyzed the chemical shift changes that occur upon titration of 1 mol eq. of dsDNA into ^{15}N-labeled MED1zf. The 20-bp DNA sequence shown in Fig. 1*B* was used for the titration. Comparison of the ^{15}N HSQC spectrum of ^{15}N-labeled MED1zf before and after the addition of dsDNA reveals numerous shift changes (Fig. 4*A*), and the quality of the data are consistent with the formation of a stable 1:1 complex. The interaction is in slow exchange on the chemical shift time scale, consistent with the measured binding affinity.

We used triple resonance NMR data to assign MED1zf in both the absence and presence of DNA, and a summary of the backbone chemical shift changes is also shown in Fig. 4*B*. Fig. 4*B* shows the secondary structure content of MED1zf, both in the free form and in the presence of DNA, based on TALOS (21) and medium-range NOE data. It is clear that the secondary structure of the free zinc finger closely matches that of other GATA-type zinc fingers, with four short β-strands and a single α-helix (helix 1). Within the zinc finger core (Ser^{112}–Tyr^{146}), the largest chemical shift changes following DNA binding are observed for the second and third β-strands and the loop connecting these strands. Some significant changes also occur in the center of the α-helix. However, the most striking effects occur across a 14-residue stretch of the basic tail, and an analysis of chemical shift and NOE data reveals that a substantial conformational change takes place in MED1zf upon DNA binding, with the formation of a second 11-residue α-helix within the formerly disordered basic C-terminal tail (helix 2, Val^{152}–Lys^{162}).

Assignments of DNA resonances were also made in both the absence and presence of MED1zf. Fig. 4*C* shows that significant changes were observed for the imino protons of Thy^{9} and Thy^{29} of the core GTATAC motif. An analysis of chemical shift changes of imino (thymines and guanines), amino (adenine and cytosines), and methyl (thymines) protons (Fig. 4*D*) reveals major changes in the core motif; however, no significant shifts were observed in the flanking regions.

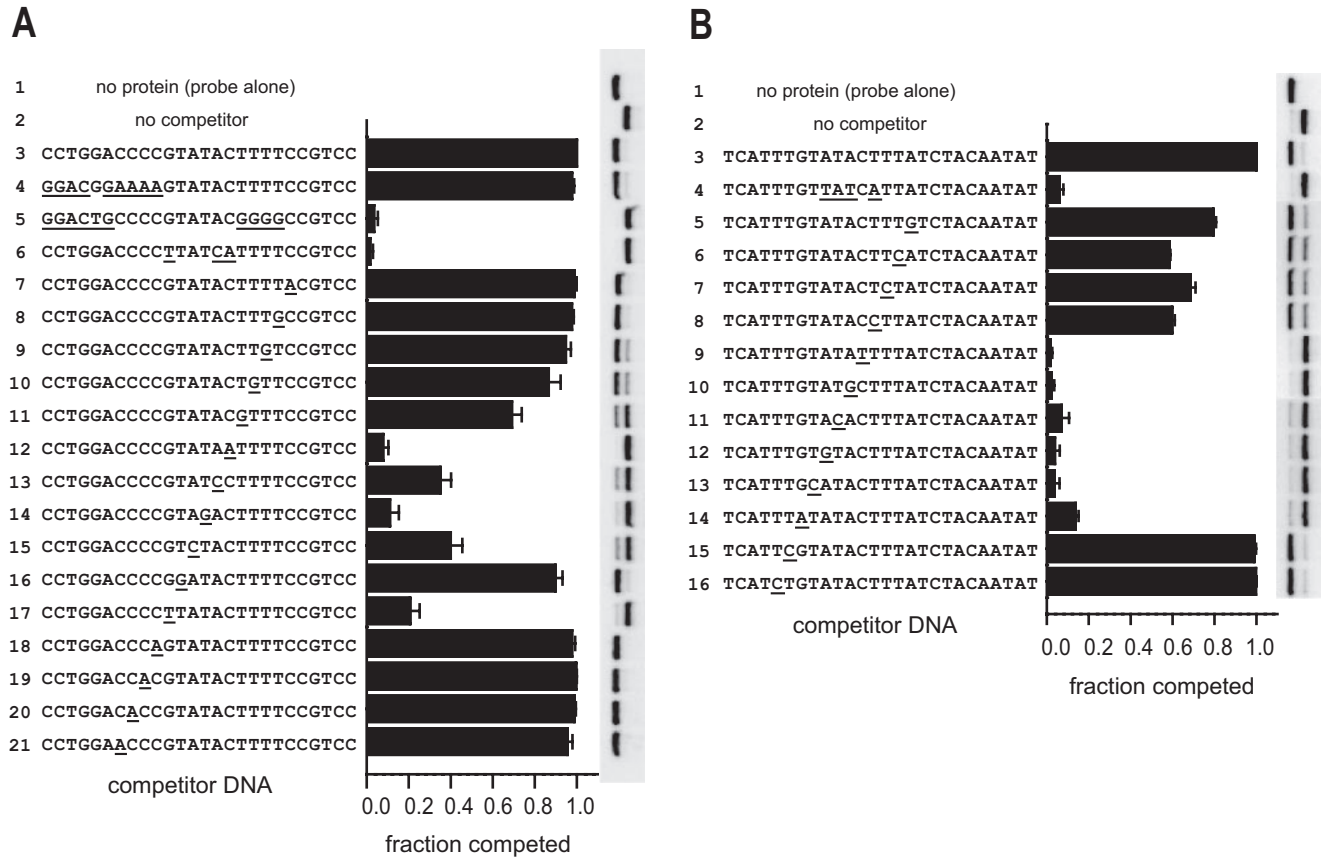

FIGURE 3. **The pyrimidine-rich sequence is important for high-affinity MED1zf recognition shown by electrophoretic mobility shift assay.** *A*, electrophoretic mobility shift assays were carried out as competition experiments; the wild type DNA sequence (*lane 3* in both panels) was labeled with ^{32}P, and the ability of unlabeled competitor DNAs containing mutations to compete for binding to MED1zf was assessed. A shifted band (and a low number on the graph) indicates that the given mutant is a poor competitor. Three independent experiments were performed, and *error bars* (S.E.) are shown. *B*, a slightly different oligonucleotide compared with *A* was used.

Model of the MED1zf·DNA Complex—Attempts to crystallize a MED1zf·DNA complex were unsuccessful,[4] and so we used our NMR and gel-retardation data to derive a structural model for the complex. Only 13 intermolecular NOEs could be assigned from heteronuclear edited/filtered NOESY spectra (supplemental Fig. S1*A*), as well as 458 unambiguous intramolecular NOEs within MED1zf, which is well below the limit of 8–10 NOEs per residue usually required for a full structure determination. In part, these difficulties were caused by the tendency of the complex to undergo degradation starting at 7 days after purification, as evident from the appearance of additional peaks in the ^{15}N HSQC spectrum. The NMR and gel-retardation experiments described above, together with the limited number of intermolecular NOEs, do, however, provide a significant amount of data on the structure of the complex, and we supplemented these data with measurements of one-bond H^N-N RDCs for the protein in the complex.

We first created a homology model of the zinc finger core structure using the GATA-1 C-finger structure (Protein Data Bank code 1GAT) (14) as a template. Val^{152}–Lys^{164} in this model were further defined as being in a helical conformation, based on TALOS and medium-range NOE data, together with $^3J_{HNHA}$ coupling constants. We used HADDOCK (26, 27) to dock this model onto an idealized B-form dsDNA oligonucleotide, having established the B-form nature of the DNA from medium-range NOE patterns (34). For the data-driven docking process, 25 residues in MED1zf were defined as active based on chemical shift perturbation data, together with specific DNA base atoms from both the GTATAC core region and the 3′-pyrimidine-rich stretch. Ambiguous interaction restraints which were used to direct the docking calculations were defined between these groups. A combination of other restraints, including inter- and intramolecular NOEs from two-dimensional and three-dimensional ^{15}N NOESY spectra (see supplemental Fig. S1*A*), dihedral angles from TALOS predictions, and RDCs were used in the HADDOCK calculations. All side chains were allowed to undergo free motion, together with the backbone of the protein in regions that were not deemed to be part of regular secondary structure (see "Experimental Procedures" for details).

Following several HADDOCK docking runs that included simulated annealing and water refinement protocols, the 10 conformers with the lowest energies were taken to represent a model for the MED1zf·DNA complex. The family of structures is well converged, overlaying with a root mean square deviation of 0.40 Å over all protein and DNA backbone atoms, excluding the flexible termini of MED1zf (Fig. 5*A* and supplemental Table 1). Overall, there are few violations of the experimental data,

[4] M. Luo, personal communication.

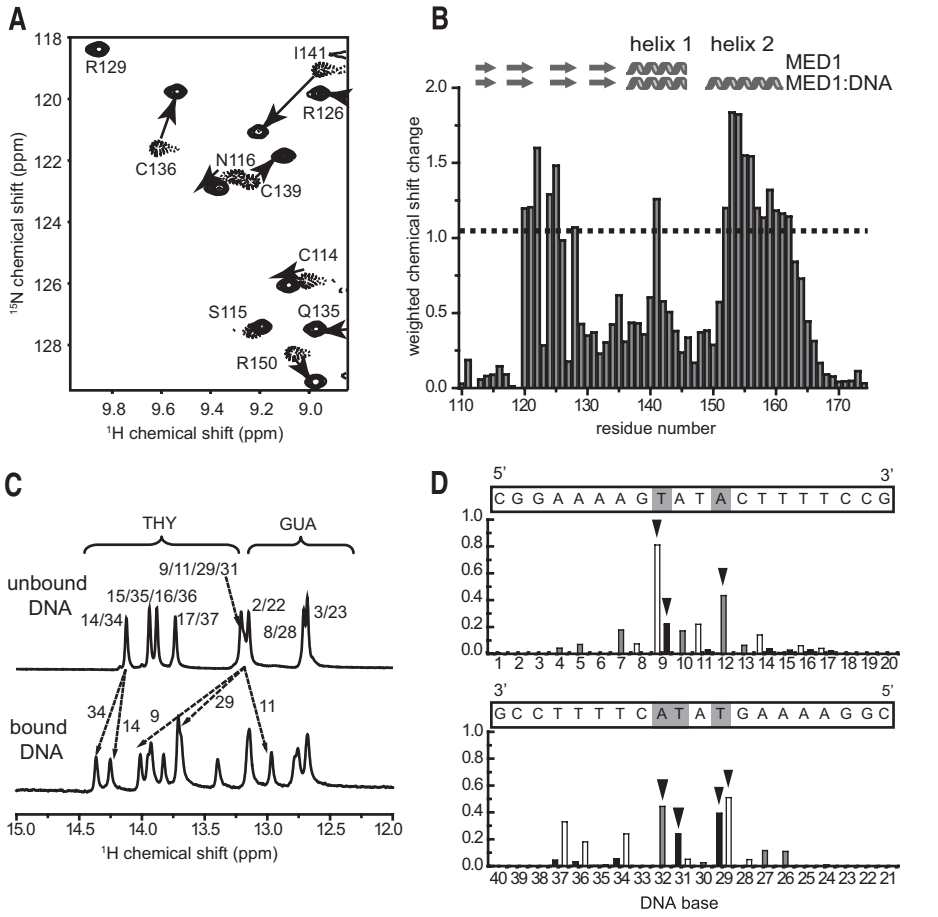

FIGURE 4. **NMR analysis of the MED1zf-DNA interaction.** *A*, section of the ^{15}N HSQC spectrum of MED1zf alone (*dashed lines*) and a 1:1 mixture of MED1zf and dsDNA (*solid lines*). Assignments and directions of movement are indicated. *B*, secondary structure and backbone chemical shift changes (weighted average of H^N, N, C^α, and C′ data) (43) for MED1zf upon binding to DNA. The *dashed line* indicates the mean change plus 1 S.D. *C*, imino region of a one-dimensional ^{1}H spectrum of the DNA in the absence (above) and presence (below) of MED1zf. *D*, summary of chemical shift changes for DNA protons following binding of MED1zf. *Black*, *gray*, and *white bars* represent imino (thymines and guanines), amino (cytosines and adenines), and methyl (thymine) protons, respectively. Significant (average plus S.D.) changes are indicated by an *arrowhead*, and the corresponding DNA bases are marked *gray* in the sequence at the *top*.

and the correlation between the predicted and observed H^N-N RDCs is very good ($r = 0.98$) (supplemental Fig. S1*C*). The zinc finger core packs against the major groove of the DNA at the GTATA site via the loop that connects β-strands 2 and 3 as well as residues in the α-helix (Fig. 5, *B* and *C*). In total, 27 hydrogen bonds and 34 hydrophobic contacts are observed in at least 50% of the structures (supplemental Table 2); of these, roughly one-third are base-specific.

In the zinc finger core, Arg^{124} forms hydrogen bonds with Gua^8, as well as hydrophobic contacts with Thy^{31}. These interactions, together with contacts made by Ile^{123} and Arg^{126}, define the sequence specificity of the protein at one end of the recognition site (the GTA in GTATAC). The next three bases are specified predominantly by hydrophobic interactions involving residues in the first helix of the zinc finger (namely Ile^{141}, Tyr^{142}, and Arg^{144}), consistent with the high-thymine content of the DNA. Arg^{144} also forms hydrogen bonds with Thy^{29}, explaining the significant chemical shift changes seen for this base (Fig. 4*D*). Helix 2 forms an extensive interface with the major groove of the pyrimidine-rich stretch, with Ala^{154}, Tyr^{158}, and Arg^{161} forming specific interactions with DNA bases (Fig. 5, *blue*). We also used the program 3DNA (28) to assess the conformation of the DNA in the complex (supplemental Fig. S1*B*). Although some deviations from ideal geometry are observed, for example between Ade^{10} and Thy^{11} where Ile^{141} contacts the DNA, a B-form-like conformation is largely maintained across the sequence.

Validation of the Data-driven Model by Site-directed Mutagenesis—To assess the accuracy of our data-derived model of the MED1zf·DNA complex, we created a series of 22 MED1zf point mutants and used surface plasmon resonance to assess their ability to bind to DNA containing the MED1zf consensus sequence (Fig. 6 and supplemental Fig. S2). Biotinylated wild type DNA was bound to a streptavidin-derivatized Biacore chip and treated with solutions of either wild type MED1zf or MED1zf mutants. The K_D for wild type MED1zf determined in this way (8 nM) (supplemental Fig. S2*A*) agrees well with the value measured by ITC (K_D = 13 nM) (Fig. 2*B*). Approximately half of the mutations reduced folding by >10-fold (Fig. 6, *gray bars*), and in excellent agreement with our structural model, mutations that affected binding were largely confined to those residues that contact the DNA (Fig. 6, *boxed residues*). Of particular note, helix 2 residues Ala^{154}, Tyr^{158}, and Arg^{161}, which contact the DNA specifically, show significant reductions in K_D, whereas Asn^{156} and Gln^{159}, which point away from the DNA, can be mutated without effect (Fig. 6). These results serve to confirm the orientation of helix 2 about its long axis with respect to the DNA.

DISCUSSION

Structural Basis for the Unusual Sequence Specificity of MED1zf—MED-1 is the first GATA-type zinc finger protein to be shown to recognize a consensus DNA sequence in which the core element differs from GAT. Furthermore, the recognition site for MED-1 is substantially longer than the canonical GATA site. Our data reveal the structural basis for these differences, showing that the GTATA core recognized by MED-1 corresponds to the AGATA motif in terms of the recognition mechanism. In all cases, recognition of DNA by these domains can be divided into three distinct regions. First, the β2 loop-β3 region contacts bases at the 5′ end. Here, Arg^{124} makes base-specific hydrogen bonds with Gua^8; indeed, arginine-guanine interactions are the most common of all base-specific interac-

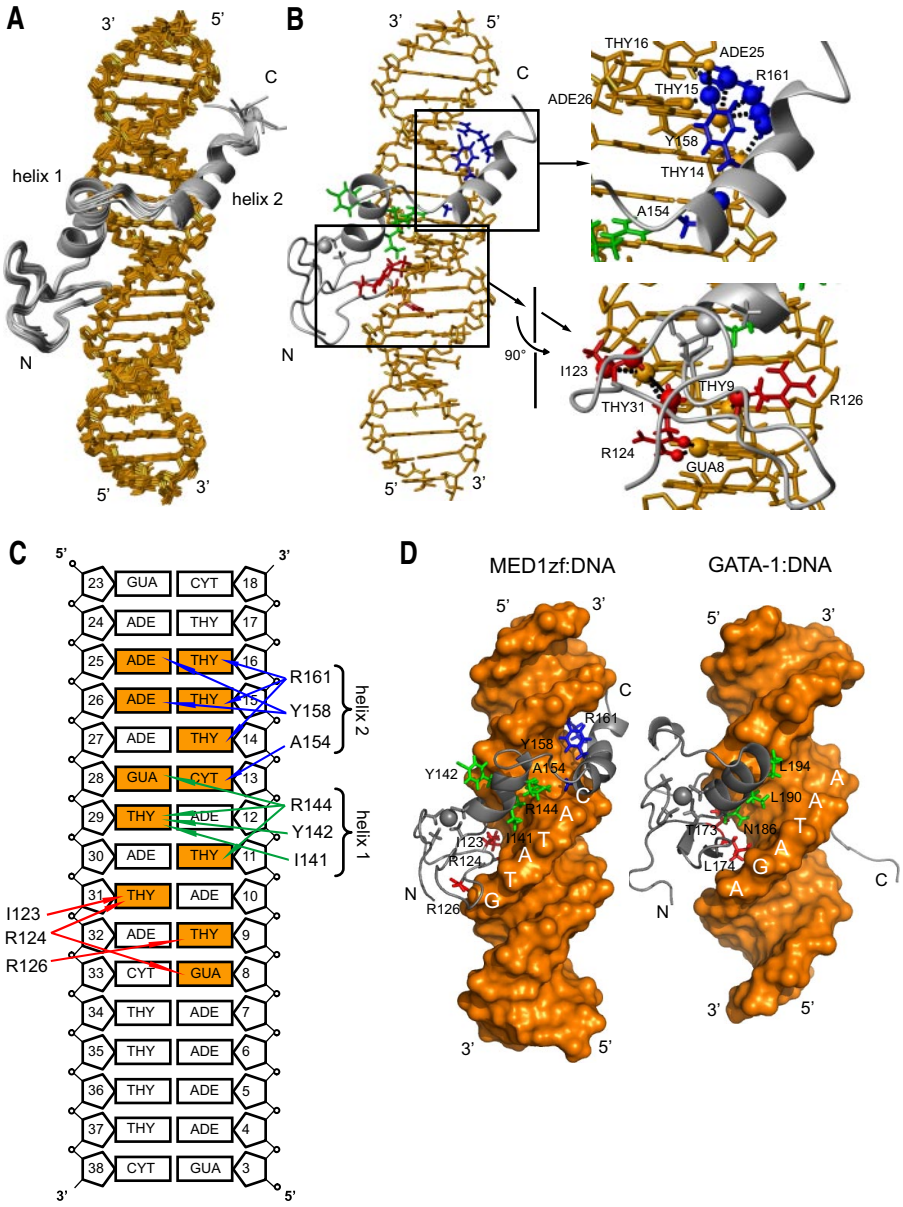

FIGURE 5. **Data-driven model of the MED1zf·DNA.** *A*, overlay of the 10 lowest energy HADDOCK structures of the MED1zf·DNA. *B*, ribbon diagram of the lowest energy structure from HADDOCK and detailed view of helix 2 and the N-terminal part ($\beta2$-$\beta3$ region) of MED1zf. Residues colored in *red*, *green*, and *blue* of the $\beta2$-$\beta3$ region, helix 1, and the newly formed helix 2 of MED1zf, respectively, make specific interactions with the DNA. The zinc atom and coordinating cysteines are shown in *gray*. *C*, schematic summary of the specific protein-DNA contacts. Bases that are specifically recognized by MED1zf are colored *orange*; interactions are indicated by *arrows* in the same colors as in *A* and *B*. *D*, comparison of the DNA-binding modes of MED1zf and GATA-1 (Protein Data Bank code 1GAT) (14). Residues that specifically contact DNA are colored as in *A* and *B*.

tions in protein·DNA complexes (35). This residue is replaced with a leucine in GATA proteins that recognize AGATA, and it appears that the specificity difference is driven in part by this substitution. Second, Ile[123], together with residues in helix 1 of MED1zf, contacts that ATA sequence in GTATAC. These residues are highly conserved between MED1zf and other GATA zinc fingers (Fig. 1), explaining the observation that this part of the consensus sequence is completely conserved.

The greatest differences are observed in the C-terminal tail region. In structures of other GATA-type zinc fingers bound to DNA, this tail "doubles back" and wraps in an extended fashion around the DNA, contacting the backbone and/or the minor groove. However, in MED1zf, an α-helix is induced upon DNA binding, and this helix contacts the pyrimidine-rich sequence that is 3′ to the GTATAC site. It is notable that this helix is not predicted by secondary structure prediction tools such as JUFO (36) or PORTER (37), highlighting one of the difficulties that is presented in computational efforts to predict protein structure without a knowledge of cognate binding partners.

Residues in helix 2 make base-specific contacts with three of the four thymines: Thy[14–16] (Fig. 5 and supplemental Table 2). This observation is consistent with our competition binding data (Fig. 3), which show that mutation of the most distal thymine has a much smaller effect on binding. Thus, our data cannot readily explain the selection of this last thymine in previous site selection experiments (17). We have recently observed a similar phenomenon in a separate study, in which a site selection experiment revealed a six-base consensus sequence for a single-stranded RNA-binding protein, and yet mutation of either of the two outer bases had no measurable effect on binding.[5] It is possible that subtle sequence-dependent conformational effects within the oligonucleotide can affect binding without requiring direct protein-nucleic acid contacts.

Diversity in Nematode GATA Factors—Whereas *C. elegans* contains two *med* genes, it is notable that the *med* family has undergone significant expansion in several other nematode species. *Caenorhabditis briggsae* and *Caenorhabditis remanei* have four and seven predicted *med* genes, respectively, whereas the recently sequenced *Caenorhabditis brenneri* displays 25 *med* genes, although little is known about the functions of these genes. An alignment of the sequences of these proteins (supplemental Fig. S3) reveals substantial variation among, for example, the 25 putative *C. brenneri* genes, even including the residues that we have shown to be important for sequence-specific recognition of DNA in *C. elegans* MED1zf. This might point to the existence of different target sites for these proteins.

[5] J. P. Mackay, unpublished data.

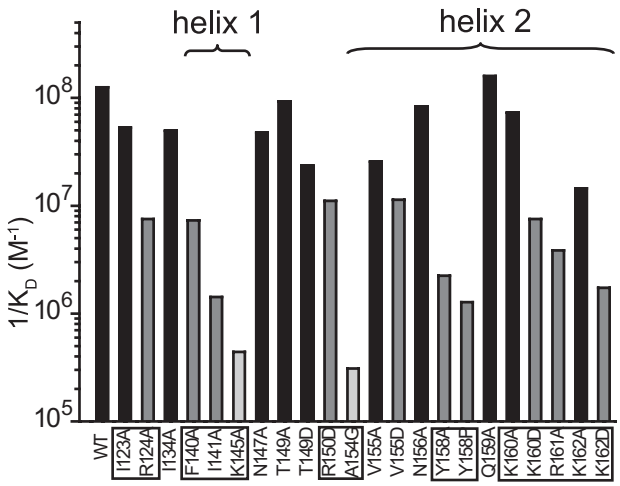

FIGURE 6. **Mutational data corroborate the structural model.** Surface plasmon resonance-derived DNA-binding affinities for MED1zf mutants are shown. *Black*, *dark gray*, and *light gray bars* show mutants with affinities that are altered by less than a factor of 10, 10–100-fold, or >100-fold, respectively. *Boxes* indicate residues that contact DNA in the structural model. *WT*, wild type.

However, it is notable in this context that Arg^{124}, which drives recognition of the G in the GTATAC motif, is a glutamine in three of the *C. briggsae* MED proteins. Our data show that an R124A mutation, which removes almost the entire side chain, only reduces DNA binding by 10-fold, so it is likely that the R124Q substitution, which would still allow the formation of one hydrogen bond to Gua^8, would not significantly compromise binding to a GTATAC site. Indeed, we have previously shown that the *C. briggsae* and *C. remanei* MED proteins can substitute for *C. elegans* MED-1 and MED-2 when introduced as transgenes in *C. elegans med-1,2*(−) strains (38).

Implications for the GATA Family of Transcription Factors—The accepted paradigm holds that GATA family proteins recognize the 6-bp consensus sequence **(A/C/T)GATA(A/G).** Site selection data have been published for GATA-1, -2, and -3 (11, 13), and a close examination of the data reveals that additional sequence preferences might exist outside the limits routinely quoted. For example, chicken GATA-3 shows an identifiable preference for adenine at the underlined positions in the sequence NNAGATANNN and a stronger preference for purines at the italicized position (11). Sequence selection at the first, second, and final positions cannot be explained by the GATA-1 and AreA structures. In the same study, GATA-2 displayed a similar preference for adenine at the final position in the motif above, and a site selection carried out on human GATA-3 showed a bias toward purines at this same position (13). It should be noted that these experiments were carried out with full-length proteins (which contain a second adjacent zinc finger), so it is possible that these additional preferences are induced by structures outside the C-finger.

In this regard, it is notable that the structures of GATA-1 and AreA bound to DNA were determined using 16-and 13-bp oligonucleotides, respectively. Even if the C-terminal tails of these domains had a propensity for helix formation, fraying of the oligonucleotide would compromise the docking site for a new helix and thereby inhibit such a conformational change. It is intriguing that the tails of these two domains nevertheless form a single turn of helix that starts in a region that is highly conserved with MED1zf (and which forms part of the extra helix of the latter protein).

Furthermore, it has been shown that the tail sequence QTRNRK in the C-finger of GATA-1 is important for determining the sequence specificity of this domain (39). However, inspection of the structure of the GATA-1·DNA complex indicates that five of the six side chains do not appear to make any significant contacts with the minor groove. In contrast, the MED1zf·DNA complex places these residues, which form part of helix 2, in the major groove, where they are able to fully engage with the DNA bases.

It is possible that the extended DNA site observed for MED1zf will also be observed for other GATA-type DNA-binding domains, including members of the mammalian and fungal GATA protein families. Manfield *et al.* (40) have recently shown that an extended version of AreA can bind up to two nucleotides on either side of the GATA site, indicating the existence of contacts in addition to those observed in the three-dimensional structure of the complex of the minimal zinc finger protein with DNA. However, their additional C-terminal sequence extends well beyond the sequence used in our study. The formation of a helix at a subset of DNA target sites (ones that displayed the longer recognition sequence) could also potentially constitute a mechanism by which interactions with other cofactor proteins might be modulated. Such proteins might either recognize the tail only in a helical conformation or only in an alternative conformation, providing a mechanism by which GATA family proteins could have different activities at different DNA sites. It is also notable that the C-terminal tail of GATA family proteins can be post-translationally modified by, for example, serine phosphorylation and lysine acetylation (41, 42). Such modifications could act as switches that regulate conformational changes in the tail region. Overall, the data in this study expand our understanding of the structural basis for DNA recognition by zinc finger domains and suggest new mechanisms through which these proteins might regulate gene expression in organisms ranging from fungi to mammals.

Acknowledgments—We thank Gottfried Otting for access to the 800-MHz NMR spectrometer at the Australian National University. We also thank Ming Luo (University of Alabama at Birmingham) for sharing unpublished results.

REFERENCES

1. Teakle, G. R., and Gilmartin, P. M. (1998) *Trends Biochem. Sci.* **23,** 100–102
2. Arceci, R. J., King, A. A., Simon, M. C., Orkin, S. H., and Wilson, D. B. (1993) *Mol. Cell. Biol.* **13,** 2235–2246
3. Evans, T., and Felsenfeld, G. (1989) *Cell* **58,** 877–885
4. Kelley, C., Blumberg, H., Zon, L. I., and Evans, T. (1993) *Development* (*Camb.*) **118,** 817–827
5. Laverriere, A. C., MacNeill, C., Mueller, C., Poelmann, R. E., Burch, J. B., and Evans, T. (1994) *J. Biol. Chem.* **269,** 23177–23184
6. Tsai, S. F., Martin, D. I., Zon, L. I., D'Andrea, A. D., Wong, G. G., and Orkin, S. H. (1989) *Nature* **339,** 446–451
7. Yamamoto, M., Ko, L. J., Leonard, M. W., Beug, H., Orkin, S. H., and Engel, J. D. (1990) *Genes Dev.* **4,** 1650–1662
8. Page, B. D., Zhang, W., Steward, K., Blumenthal, T., and Priess, J. R. (1997)

Genes Dev. **11,** 1651–1661
9. Feng, B., Xiao, X., and Marzluf, G. A. (1993) *Nucleic Acids Res.* **21,** 3989–3996
10. Kudla, B., Caddick, M. X., Langdon, T., Martinez-Rossi, N. M., Bennett, C. F., Sibley, S., Davies, R. W., and Arst, H. N., Jr. (1990) *EMBO J.* **9,** 1355–1364
11. Ko, L. J., and Engel, J. D. (1993) *Mol. Cell. Biol.* **13,** 4011–4022
12. Lowry, J. A., and Atchley, W. R. (2000) *J. Mol. Evol.* **50,** 103–115
13. Merika, M., and Orkin, S. H. (1993) *Mol. Cell. Biol.* **13,** 3999–4010
14. Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. J., and Gronenborn, A. M. (1993) *Science* **261,** 438–446
15. Bates, D. L., Chen, Y., Kim, G., Guo, L., and Chen, L. (2008) *J. Mol. Biol.* **381,** 1292–1306
16. Starich, M. R., Wikstrom, M., Arst, H. N., Jr., Clore, G. M., and Gronenborn, A. M. (1998) *J. Mol. Biol.* **277,** 605–620
17. Broitman-Maduro, G., Maduro, M. F., and Rothman, J. H. (2005) *Dev. Cell* **8,** 427–433
18. Cai, M., Huang, Y., Sakaguchi, K., Clore, G. M., Gronenborn, A. M., and Craigie, R. (1998) *J. Biomol. NMR* **11,** 97–102
19. Liew, C. K., Kowalski, K., Fox, A. H., Newton, A., Sharpe, B. K., Crossley, M., and Mackay, J. P. (2000) *Structure* (*Camb.*) **8,** 1157–1166
20. Deane, J. E., Mackay, J. P., Kwan, A. H. Y., Sum, E. Y., Visvader, J. E., and Matthews, J. M. (2003) *EMBO J.* **22,** 2224–2233
21. Cornilescu, G., Delaglio, F., and Bax, A. (1999) *J. Biomol. NMR* **13,** 289–302
22. Bax, A., Vuister, G. W., Grzesiek, S., Delaglio, F., Wang, A. C., Tschudin, R., and Zhu, G. (1994) *Methods Enzymol.* **239,** 79–105
23. Cordier, F., Dingley, A. J., and Grzesiek, S. (1999) *J. Biomol. NMR* **13,** 175–180
24. Zweckstetter, M., Hummer, G., and Bax, A. (2004) *Biophys. J.* **86,** 3444–3460
25. Goddard, T., and Kneller, D. (2006) *SPARKY 3*, University of California, San Francisco
26. de Vries, S. J., van Dijk, A. D., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V., Wassenaar, T., and Bonvin, A. M. (2007) *Proteins* **69,** 726–733
27. Dominguez, C., Boelens, R., and Bonvin, A. M. (2003) *J. Am. Chem. Soc.* **125,** 1731–1737
28. Lu, X. J., and Olson, W. K. (2003) *Nucleic Acids Res.* **31,** 5108–5121
29. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54,** 905–921
30. Gamsjaeger, R., Swanton, M. K., Kobus, F. J., Lehtomaki, E., Lowry, J. A., Kwan, A. H., Matthews, J. M., and Mackay, J. P. (2008) *J. Biol. Chem.* **283,** 5158–5167
31. Kamphuis, M. B., Bonvin, A. M., Monti, M. C., Lemonnier, M., Munoz-Gomez, A., van den Heuvel, R. H., Diaz-Orejas, R., and Boelens, R. (2006) *J. Mol. Biol.* **357,** 115–126
32. Liew, C. K., Simpson, R. J., Kwan, A. H., Crofts, L. A., Loughlin, F. E., Matthews, J. M., Crossley, M., and Mackay, J. P. (2005) *Proc. Natl. Acad. Sci. U. S. A.* **102,** 583–588
33. Omichinski, J. G., Trainor, C., Evans, T., Gronenborn, A. M., Clore, G. M., and Felsenfeld, G. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90,** 1676–1680
34. Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York
35. Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A., and Brasseur, R. (2005) *Proteins* **61,** 258–271
36. Meiler, J., and Baker, D. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100,** 12105–12110
37. Pollastri, G., and McLysaght, A. (2005) *Bioinformatics* (*Oxf.*) **21,** 1719–1720
38. Coroian, C., Broitman-Maduro, G., and Maduro, M. F. (2006) *Dev. Biol.* **289,** 444–455
39. Ghirlando, R., and Trainor, C. D. (2003) *J. Biol. Chem.* **278,** 45620–45628
40. Manfield, I. W., Reynolds, L. A., Gittins, J., and Kneale, G. G. (2000) *Biochim. Biophys. Acta* **1493,** 325–332
41. Boyes, J., Byfield, P., Nakatani, Y., and Ogryzko, V. (1998) *Nature* **396,** 594–598
42. Hung, H. L., Lau, J., Kim, A. Y., Weiss, M. J., and Blobel, G. A. (1999) *Mol. Cell. Biol.* **19,** 3496–3505
43. Ayed, A., Mulder, F. A., Yi, G. S., Lu, Y., Kay, L. E., and Arrowsmith, C. H. (2001) *Nat. Struct. Biol.* **8,** 756–760