

# The Structure of a Bacterial DUF199/WhiA Protein: Domestication of an Invasive Endonuclease

Brett K. Kaiser,<sup>1</sup> Matthew C. Clifton,<sup>1</sup> Betty W. Shen,<sup>1</sup> and Barry L. Stoddard<sup>1,\*</sup>

<sup>1</sup>Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N. A3-025, Seattle, WA 98006, USA

\*Correspondence: [bstoddard@fhcrc.org](mailto:bstoddard@fhcrc.org)

DOI 10.1016/j.str.2009.08.008

## SUMMARY

Proteins of the DUF199 family, present in all Gram-positive bacteria and best characterized by the WhiA sporulation control factor in *Streptomyces coelicolor*, are thought to act as genetic regulators. The crystal structure of the DUF199/WhiA protein from *Thermatoga maritima* demonstrates that these proteins possess a bipartite structure, in which a degenerate N-terminal LAGLIDADG homing endonuclease (LHE) scaffold is tethered to a C-terminal helix-turn-helix (HTH) domain. The LHE domain has lost those residues critical for metal binding and catalysis, and also displays an extensively altered DNA-binding surface as compared with homing endonucleases. The HTH domain most closely resembles related regions of several bacterial sigma70 factors that bind the -35 regions of bacterial promoters. The structure illustrates how an invasive element might be transformed during evolution into a larger assemblage of protein folds that can participate in the regulation of a complex biological pathway.

## INTRODUCTION

Modern protein folds are believed to have arisen from the expansion of a much smaller complement of proteins that were found in the last common ancestor of prokarya, archaea, and eukarya (Caetano-Anolles et al., 2007; Orengo and Thornton, 2005). Evidence for the ability of early proteins to diversify into increasingly elaborate structures, and acquire novel functions, includes the observations that (i) many modern families of protein folds (such as the TIM barrels) encompass a wide variety of biological activities (Nagano et al., 2002); (ii) individual proteins can exhibit multiple unrelated activities (a property termed “moonlighting”) (Jeffery, 2003); and (iii) certain protein sequences can adopt multiple folded states that each display unique functional properties (Tuinstra et al., 2008).

Among the many protein folds that have been diversified over the course of evolution, those that are associated with DNA cleavage and modification stand out. For example, the retroviral RNase H and integrase folds are utilized by many transposases, nucleotidyl transferases, resolvases, and nucleases (Nowotny, 2009). Similarly, the folds found within homing endonucleases (proteins that promote the genetic mobility of microbial introns

and inteins) are also found in an impressive array of enzymes, including those involved in phage restriction, DNA replication, DNA repair, and recombination (Stoddard, 2005). These proteins are constructed around catalytic core folds containing HNH, PD...(D/E)xK, and GIY-YIG active site motifs.

Homing endonucleases have also been domesticated and employed by their biological hosts for more disparate biological purposes. For example, many homing endonucleases have been adapted by their hosts to assist in RNA folding and splicing (Ho et al., 1997), whereas others (such as the yeast HO endonuclease) have been recast for the purpose of initiating nuclear gene conversion events (Koufopanou and Burt, 2005). In most of these cases, the new host-specific function still involves the ability of these enzymes to catalyze phosphotransfer reactions or to promote the rearrangement of nucleic acid substrates.

However, at least three instances have been documented where the biological function of homing endonuclease scaffolds have been more dramatically altered—with the protein in each case finding new employment as a regulator of a more complex biological pathway. In the first example, the DNA-binding domains found in “Smad” proteins (eukaryotic transcription factors involved in transforming growth factor  $\beta$  signaling) were found to be composed of a “ $\beta\beta\alpha$ -metal” endonuclease fold (resembling the I-Ppol endonuclease [Flick et al., 1998]) that has lost its catalytic activity (Grishin, 2001). In the second case, the DNA binding domain of the AP2/ERF family of plant transcription regulators contains a recognizable HNH endonuclease domain (a structure common to many phage-derived homing endonucleases) (Magnani et al., 2004). Finally, proteins of the DUF199 superfamily, found throughout most if not all Gram-positive bacteria, are postulated to contain a degenerated LAGLIDADG homing endonuclease domain (Knizewski and Ginalski, 2007). The most well-characterized member of this family, the WhiA protein from the soil bacterium *Streptomyces coelicolor* (WhiA<sup>Sc</sup>), is required for sporulation and regulates the expression of multiple sporulation-specific “Whi” genes, including its own reading frame (Ainsa et al., 2000). It is unknown whether this regulation occurs through direct or indirect interaction with DNA promoter elements or other proteins. Regardless, it is likely that WhiA homologs in other Gram-positive bacteria function in a similar manner, because those microbes all contain similar Whi operons including a single recognizable DUF199/WhiA protein (Ainsa et al., 2000).

In order to visualize the structural basis for the creation of a transcriptional regulatory protein from a protein fold typically associated with a mobile endonuclease, we have determined the crystal structure of the DUF199/WhiA protein from *T. maritima* (WhiA<sup>Tm</sup>), and examined similarities and differences of its structure and

**Table 1. Crystallographic Statistics**

Data Collection			
Protein	LAGLIDADG Se-Met	LAGLIDADG	Full length
Beamline (Å)	ALS 5.0.2	Home source	Home source
Wavelength	0.98	1.54	1.54
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Unit cell (Å)	a = 50.9, b = 79.7, c = 115.3	a = 50.5, b = 79.9, c = 115.4	a = 51.65, b = 61.30, c = 97.22
Resolution (Å) <sup>a</sup>	50-2.9 (3.0-2.9)	50-2.6 (2.69-2.60)	50-2.34 (2.42-2.34)
R <sub>merge</sub> (%)	10.8 (37.6)	4.3 (18.5)	5.3 (34.7)
I/σI	13.5 (4.07)	18.1 (6.35)	27.0 (5.0)
Redundancy	9.2 (9.2)	2.5 (2.4)	4.7 (4.6)
Completeness (%)	97.7 (98.0)	93.8 (89.0)	98.6 (98.3)
Unique reflections	10776	14107	13400
Refinement Statistics			
R <sub>work</sub> (%)	—	19.8	22.7
R <sub>free</sub> (%) <sup>b</sup>	—	26.0	27.6
Number of atoms	—	3069	2265
Protein	—	2997	2190
Water	—	57	62
Estimated coordinates error (Å)	—	0.24	0.21
Geometry			
Rmsd bonds (Å)	—	0.012	0.005
Rmsd angles (°)	—	1.456	0.868
Rmsd chiral (Å <sup>3</sup> )	—	0.104	0.05
Average B (Å <sup>2</sup> )	—	45.7	42.1
Protein monomer B factors (Å <sup>2</sup> )	—	45.2 (mol A), 47.1 (mol B)	42.2
Water B factors (Å <sup>2</sup> )	—	19.4	40.4
Ramachandran distribution			
Most favored (%)	—	89.5	94.7
Additionally allowed (%)	—	10.2	4.9
Generously allowed (%)	—	0.3	0.0
Disallowed (%)	—	0.0	0.4

<sup>a</sup> Numbers in parentheses correspond to the highest-resolution shells.

<sup>b</sup> Calculated using a test set corresponding to 5% of the data.

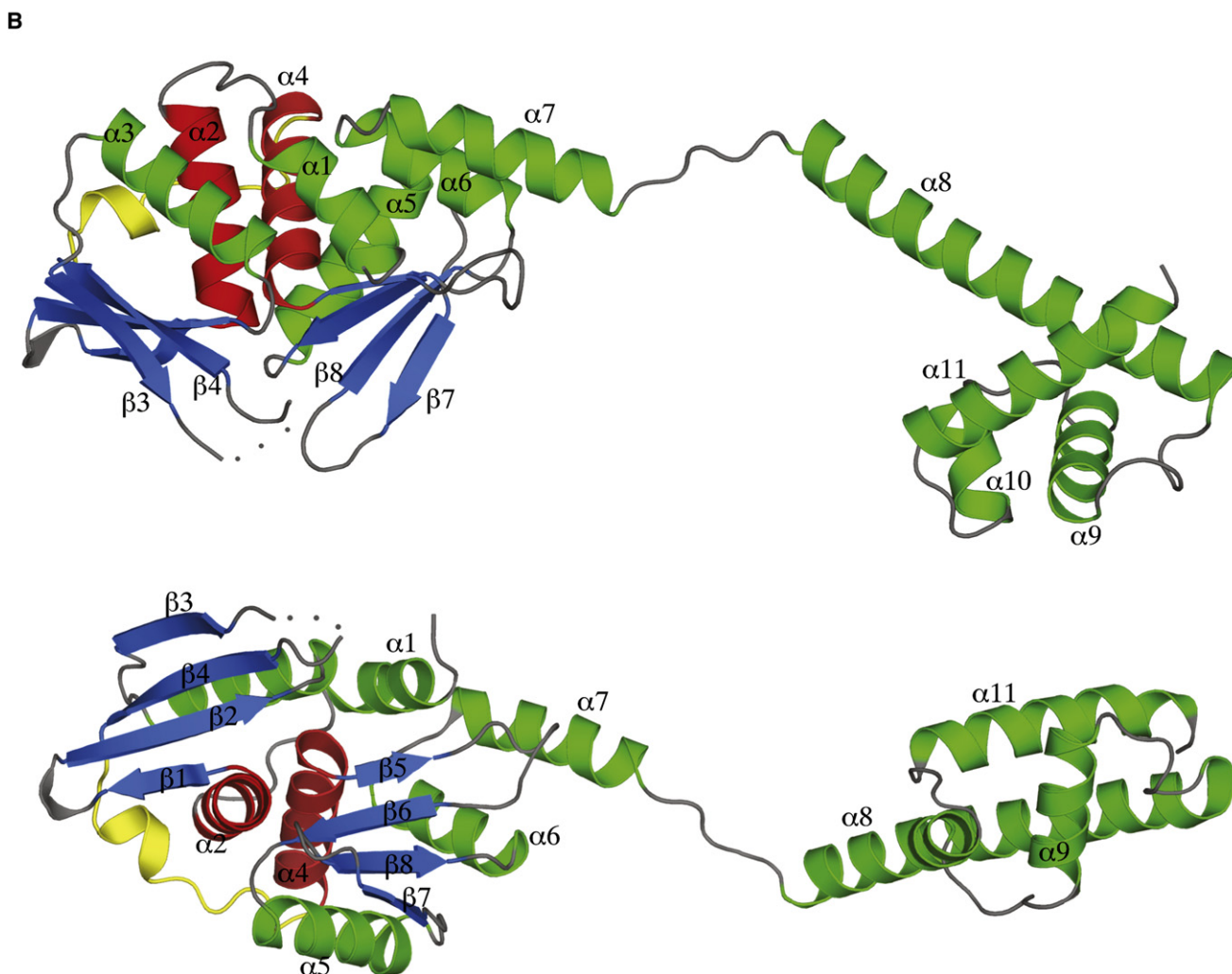
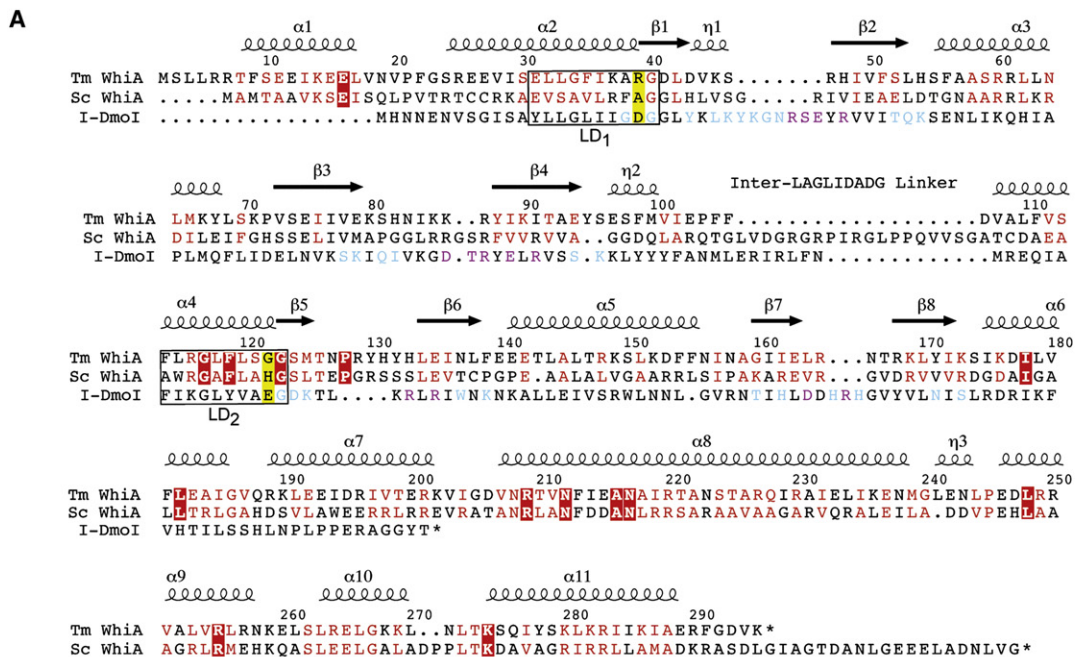
primary sequence relative to its closest bacterial homologs and also to more distantly related LAGLIDADG homing endonucleases. The structure illustrates how the unique evolutionary pressures that are placed upon a genetic regulator, versus those placed on an invasive endonuclease, might produce individually tailored structural and biochemical features that are appropriate for each function. Studies of these proteins also indicate a likely scenario by which an invasive element could be converted into a genetic regulator. This chain of events would involve progression from a simple role as an autoregulator of its own expression, to the subsequent acquisition of novel domains and properties resulting in a more complex protein assemblage that can participate in highly coordinated transcriptional regulation.

## RESULTS

We expressed an untagged, full-length DUF199/WhiA construct originally encoded in *T. maritima* (WhiA<sup>Tm</sup>) as a soluble protein in *E. coli* (see Figure S1A available online). Proteolytic digest exper-

iments with trypsin revealed that the full-length protein could be digested into two stable domains, consistent with a bipartite structural organization (Figure S1B). We were initially able to obtain crystals of the protein's isolated LAGLIDADG domain that diffracted to 2.6 Å resolution. Similar crystals were grown using selenomethionyl-derivatized protein that allowed us to calculate phases to 3.0 Å resolution by single-wavelength anomalous dispersion phasing. The structure was determined and ultimately refined to 2.6 Å resolution, revealing a pair of LAGLIDADG domains in the asymmetric unit (ASU, Figure S2). The resulting model of these LAGLIDADG domains was refined to final values of R<sub>work</sub> and R<sub>free</sub> of 19.0% and 26.1%, respectively (Table 1). The α-carbons of the two individual molecules in the ASU superimpose with a root-mean-square deviation (rmsd) of 0.75 Å.

Subsequently, we obtained crystals of the full-length WhiA<sup>Tm</sup> protein that diffracted to 2.35 Å resolution under different crystallization conditions. This structure (Figure 1) was solved via molecular replacement using the refined coordinates of a single LAGLIDADG domain described above as a search model, and



refined to final values of  $R_{\text{work}}$  and  $R_{\text{free}}$  of 22.7% and 27.6%, respectively (Table 1). The full-length protein contained a single WhiA molecule in the ASU, which was comprised of an N-terminal LAGLIDADG (LHE) domain, a linker region, and a C-terminal helix-turn-helix (HTH) domain. The crystallographic protein-protein contacts in the structure of the full-length protein are completely different from those in crystals of the isolated LAGLIDADG domain, indicating that the contacts between LAGLIDADG domains observed in the two structures represent lattice contacts that differ between different crystal packing arrangements. Overall, the three independent crystallographic views of the LAGLIDADG domain obtained in this study correspond very closely to one other (pairwise  $\alpha$ -carbon rmsd values of 0.75 Å to 1.50 Å). Except where noted below, the remainder of this manuscript describes the structure of the full-length WhiA<sup>Tm</sup> protein.

The structure of WhiA<sup>Tm</sup> reveals that the linker region between its LAGLIDADG and HTH domains consists of two separate  $\alpha$  helices ( $\alpha 7$  and  $\alpha 8$ ) connected by a less structured series of residues (201–206) that may act as a flexible hinge (Figure 1). The overall dimensions of the protein are 105 Å by 30 Å by 25 Å; the distance between the center of the two independent domains is approximately 70 Å. One of the most conserved regions of the WhiA protein corresponds to the N-terminal end of the  $\alpha 8$  helix, which contains 4 invariant residues (209R, 212N, 216A, and 217N) and numerous conservative substitutions among WhiA sequences obtained from 14 divergent bacterial organisms (Figures 1A and S4). This degree of conservation suggests either an essential function (such as interacting with a nucleic acid target region or with an additional protein factor) or a structural role, such as providing rigidity to the long  $\alpha 8$  helix. The putative hinge region likely allows both domains a relatively large range of motion relative to each other; we therefore discuss the structural features of each domain independently.

### The LAGLIDADG Domain

The N-terminal region of WhiA<sup>Tm</sup> contains the same protein fold topology that is observed in monomeric LAGLIDADG homing endonucleases. This region is composed of two structurally similar domains, each containing an  $\alpha\beta\beta\alpha\beta\beta$  core fold, that are connected by a short peptide linker (Figures 1 and 2). The closest structural homolog of this domain, identified using the DALI web-server (Holm et al., 2008), is the I-Dmol homing endonuclease (an archaeal enzyme encoded within a mobile group I intron [Silva et al., 1999]). Despite overall limited sequence homology (13% identity; Figure 1A), both structures superimpose closely, with an  $\alpha$ -carbon rmsd across all aligned residues (Figure 1A) of 2.4 Å (Figure 2A). The most conserved elements within this

region are those residues that comprise the two LAGLIDADG helices ( $\alpha 2$  and  $\alpha 4$  in WhiA) that form the core of the domain interface (Figure 2B). These helices are closely superimposable, including intimate packing between backbone atoms in the helices that is facilitated by the presence of small side chains at positions located near the helical interface (Figure 2B).

A critical difference between WhiA family members and LAGLIDADG homing endonucleases is that the WhiA proteins lack acidic residues at the base of the LAGLIDADG helices that are strongly conserved in homing endonucleases (in I-Dmol these residues are D20 and E117). In the endonucleases, these residues coordinate divalent cations and are required for DNA cleavage. In WhiA<sup>Tm</sup> the corresponding residues are R39 and G123; in the independently determined structures of its LAGLIDADG domain described above, divalent cations are clearly absent. In addition, homing endonuclease active sites contain conserved basic residues that are involved in transition-state stabilization (such as K43 and K120 in I-Dmol). These positions are occupied by a histidine and methionine (H54 and M125, respectively) in the WhiA<sup>Tm</sup> structure, and similarly nonconserved in its closest homologs. Therefore, WhiA family members are almost certainly not endonucleases, a conclusion supported by DNA digest experiments in our lab with WhiA<sup>Tm</sup> and its homolog from *S. coelicolor*, WhiA<sup>Sc</sup> (data not shown).

The crystal structure of the WhiA<sup>Tm</sup> protein also indicates that the mechanism of DNA recognition and binding by its LAGLIDADG domains might differ significantly from that displayed by the same domains in homing endonucleases. Enzymes such as I-Dmol utilize a pair of antiparallel  $\beta$  sheets and associated loops that make extensive contacts with their DNA substrates, via interactions with the DNA backbone and with individual nucleotide base pairs across the entire DNA target. Each LAGLIDADG domain is responsible for recognition of a single DNA half-site, and their DNA-contact surfaces are uniformly positively charged—a feature interrupted only by the presence of conserved acid residues in the active sites at the center of the domain interface (Figure 2C).

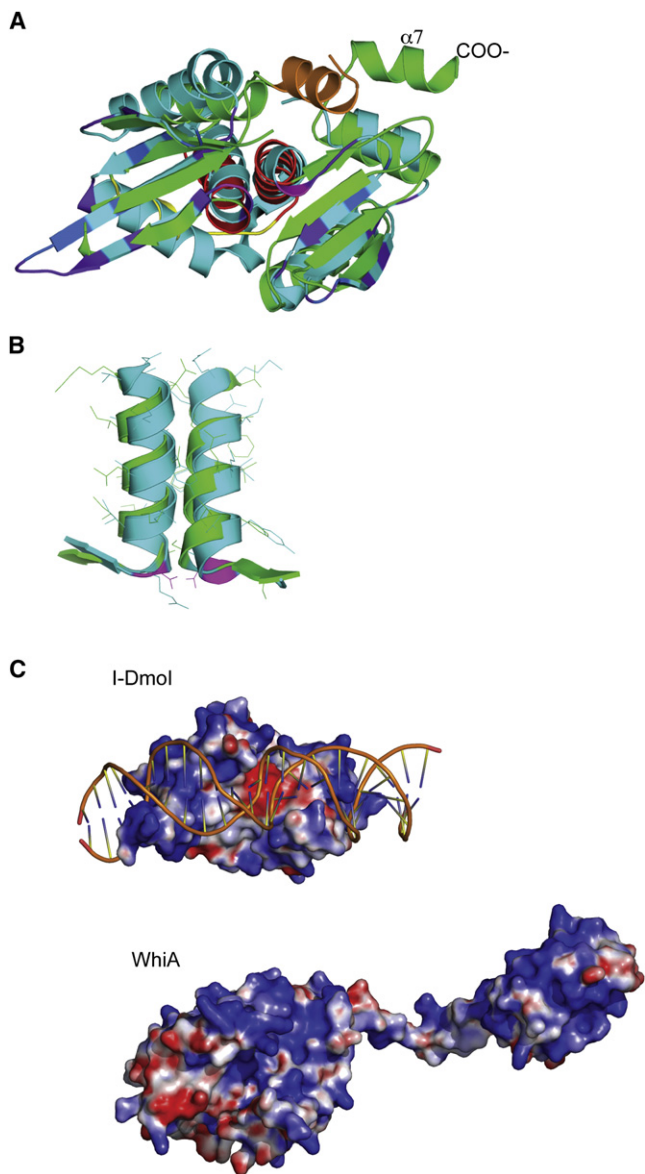
In contrast, a substantial region of the same surface of WhiA<sup>Tm</sup>, corresponding to the N-terminal LAGLIDADG domain, displays significant negative surface charge (Figure 2C). Furthermore, the C-terminal LAGLIDADG domain displays a positively charged surface that extends well beyond its  $\beta$  sheet region. It therefore seems likely that the DUF199/WhiA protein family interacts with its DNA target in a manner unique from the mode of DNA binding exhibited by LAGLIDADG homing endonucleases such as I-Dmol.

There are several additional differences between the LAGLIDADG folds found in WhiA proteins versus homing endonucleases. First, the WhiA family contains an additional

### Figure 1. Structure and Sequence Conservation of Duf199/WhiA

(A) A multiple sequence alignment of 14 different WhiA homologs (the same subset described by Knizewski and Ginalski [2007]) was generated using MUSCLE (Edgar, 2004). From this group alignment we show the alignment of WhiA<sup>Tm</sup> and WhiA<sup>Sc</sup>. We also performed a structure-based alignment of I-Dmol (PDB code 2VS7) compared with WhiA<sup>Tm</sup>. The residue numbering and location of  $\alpha$  helices and  $\beta$  sheets shown above the alignment correspond to WhiA<sup>Tm</sup> and were generated with ESPript (Gouet et al., 1999). Residues blocked in red are identical among the WhiA family members used in the alignment; residues highlighted in red indicate conservative substitutions. I-Dmol residues highlighted in cyan contact the DNA backbone; residues highlighted in purple make base contacts. The yellow-highlighted residues indicate the position of acidic residues present in all LAGLIDADG homing endonucleases that are required for DNA cleavage and absent in the WhiA family. Boxes indicate the two LAGLIDADG motifs (LD<sub>1</sub> and LD<sub>2</sub>).

(B) Structure of full-length WhiA<sup>Tm</sup> shown in two different orientations. The  $\beta$  strands of the LAGLIDADG domain are highlighted in blue; the LAGLIDADG helices ( $\alpha 2$  and  $\alpha 4$ ) are red; the interdomain linker is yellow. The disordered loop connecting  $\beta$  strands 3 and 4 is represented by dots. Labeled secondary structural elements correspond to the alignment in A.



**Figure 2. Structural Comparison of the WhiA<sup>Tm</sup> LAGLIDADG Domain with its Closest Structural Homolog, the I-Dmol Homing Endonuclease**

(A) Superposition of the LAGLIDADG domain of WhiA<sup>Tm</sup> (green) with I-Dmol (cyan, PDB code 2VS7), with the DNA binding domain of I-Dmol oriented toward the viewer. In WhiA<sup>Tm</sup>, the N-terminal  $\alpha 1$  helix present in the WhiA family but absent in I-Dmol is orange; the LAGLIDADG helices are red; and the interdomain linker is yellow. In I-Dmol, the conserved catalytic acidic residues at the base of the LAGLIDADG helices are magenta; residues that contact DNA bases are purple, and residues that contact the DNA backbone are dark blue.

(B) Superposition of the LAGLIDADG helices of WhiA<sup>Tm</sup> (green;  $\alpha 2$  left,  $\alpha 4$  right) with I-Dmol (cyan).

(C) Surface electrostatic charge potential (red: negative charge; blue: positive charge) of I-Dmol bound to its cognate DNA (top); and of full-length WhiA<sup>Tm</sup> (bottom; the LAGLIDADG domain is to the left, and the HTH domain to the right). The LAGLIDADG domains from the endonuclease and from WhiA are shown in the same relative orientation, looking into the  $\beta$  sheet surfaces of each. The electrostatic potentials were calculated using APBS (Baker et al., 2001). The large negative charge in I-Dmol corresponds to the pair of acidic

N-terminal  $\alpha$  helix ( $\alpha 1$ , Figure 2A) that is not present in homing endonucleases. The function of this helix is not clear, although it makes extensive contacts with both LAGLIDADG helices and with helix  $\alpha 7$  in the linker region. In addition, the length, sequence and structure of the peptide that connects the two LAGLIDADG domains in monomeric homing endonucleases and in the WhiA family (Figures 1B and 2A) is highly variable, ranging from 12 residues in WhiA<sup>Tm</sup> to 30 residues in the orthologous WhiA protein from *S. coelicolor*. In I-Dmol this region contains an  $\alpha$ -helix spanning almost three full turns, whereas in WhiA<sup>Tm</sup> this region consists of largely random coil architecture.

Finally, WhiA<sup>Tm</sup> and its most closely related homologs are unique in containing an additional 5 residues at the N terminus that are not present in other WhiA members. In the structure of the isolated LAGLIDADG domains (Figure S2), these additional residues form an interchain  $\beta$  strand interaction with  $\beta 3$  of its crystallographic dimeric partner, but in the full-length structure (a monomer in the asymmetric unit) these residues are disordered.

### The Helix-Turn-Helix Domain

The C-terminal region of WhiA<sup>Tm</sup> forms a canonical three-helical bundle, termed a “helix-turn-helix” (HTH) domain, composed of the  $\alpha 9$ ,  $\alpha 10$ , and  $\alpha 11$  helices of the full-length protein (Figures 1 and 3). Although WhiA<sup>Tm</sup> does not display any additional elaborations upon this core fold, other WhiA homologs contain additional C-terminal residues; for example, WhiA from *S. coelicolor* contains 23 additional residues not present in WhiA<sup>Tm</sup> (Figure 3A). The closest structural homologs of the WhiA<sup>Tm</sup> HTH domain, identified by a three-dimensional similarity search using the DALI webserver (Holm et al., 2008), are similar HTH domains comprising “domain 4” of the bacterial sigma70 protein family (Figure 3A, B). The most similar structure, domain 4 from the *E. coli* SigmaE protein, superposes on the Tm WhiA HTH domain with an  $\alpha$ -carbon rmsd of 1.96 Å over 65 residues.

The HTH domains from bacterial sigma70 factors typically bind the  $-35$  region of bacterial promoters, and the structure of two of these factors (*E. coli* SigmaE, Protein Data Bank [PDB] code 2H27; *T. aquaticus* RNA Polymerase Sigma subunit, PDB code 1KU3) have been solved bound to DNA. Superposition of the WhiA<sup>Tm</sup> HTH domain onto these structures indicates that the third ( $\alpha 11$ ) helix of the WhiA<sup>Tm</sup> HTH domain might make significant DNA contacts (Figure 3D), which would be consistent with the principal mode of DNA binding most commonly displayed by HTH domains (Aravind et al., 2005). In addition, the electrostatic potential of the HTH domain would be compatible with DNA binding (Figure 2C). However, in this orientation, the  $\alpha 8$  linker helix of WhiA<sup>Tm</sup> would closely approach the minor groove of DNA (Figure 3C). Therefore, if the HTH domain of WhiA<sup>Tm</sup> does bind DNA, it might do so in a manner requiring distortion of its DNA target, and might use additional residues from this region to make further contacts with the DNA backbone, as has been observed for other HTH-containing proteins (Khare et al., 2004).

An additional important structural feature of the HTH domain is a cleft formed between the  $\alpha 9$  and  $\alpha 11$  helices, into which

residues (D21, E117) in the active site. The metal ions that these residues normally coordinate (and thus neutralize the active site residues) are not present in the electrostatic calculation.

the C-terminal end of the  $\alpha 8$  helix is docked (Figure 3C). A combination of hydrophobic interactions and hydrogen bonds stabilize this interaction, which includes nearly half (14 out of 32 residues) of the  $\alpha 8$  helix. A cleft in this configuration is typical of HTH domains and is often utilized to pack additional stabilizing structural elements. It is noteworthy that the N-terminal half of the long  $\alpha 8$  helix is well conserved throughout bacterial species, whereas the C-terminal half seems to be less constrained (Figure S4). Two possible explanations are that (i) WhiA<sup>Tm</sup>'s function requires the helical conformation of N-terminal end of  $\alpha 8$  to be stable in the absence of additional packing interactions (which may require a conserved sequence of amino acids), or (ii) this region is involved in additional molecular interactions in order to support the biological function of WhiA.

## DISCUSSION

### Evolution of Transcription Regulatory Activity from a Homing Endonuclease?

There is no direct experimental evidence demonstrating either DNA-binding or direct transcriptional activation functions for the WhiA proteins. However, previous biochemical and genetic studies of WhiA from *S. coelicolor* (Knizewski and Ginalski, 2007) and its homolog from *S. ansochromogenes* (originally termed *sawC* in that organism) (Xie et al., 2007) indicate that those particular proteins are involved in septation and sporulation, and effect the expression of several genes, including their own, that are involved in those processes. Combined with the observation that these proteins contain two separate domains known for their DNA-binding activity, it is reasonable to hypothesize that the WhiA proteins might function as transcriptional regulators.

After invasion of a genomic target by a homing endonuclease, there is little selective pressure imposed by the host for the maintenance of a functional enzyme, leading to the gradual accumulation of mutations that reduce its activity. This evolutionary degradation eventually leads to loss of the endonuclease gene and its associated intervening sequence (Burt and Koufopanou, 2004). Homing endonucleases often avoid this fate by acquiring novel activities that are beneficial to their host during evolution, a situation that places them under selective pressure to maintain a well-behaved protein fold (Stoddard, 2005). Several examples of domesticated LAGLIDADG-containing proteins have been documented (such as the HO endonuclease and maturase intron splicing factors); however, the WhiA family is particularly noteworthy because it is the first example of a putative transcription factor containing a LAGLIDADG fold, and (if this functional annotation is found to be true) would be the third known example of a domesticated transcription factor derived from any homing endonuclease (Grishin, 2001; Knizewski and Ginalski, 2007).

In a manner analogous to the hypothesized recruitment of the LAGLIDADG protein scaffold for transcriptional regulation by the DUF199/WhiA proteins in bacteria, a variety of additional protein folds that are primarily associated with enzymatic activity have also been co-opted and employed as genetic regulators. For example, the eukaryotic Gal80, TAFII150, and Cdc68/Spt16 transcription factors are derived from (or share common ancestors with) oxidoreductase, aminopeptidase N, and aminopeptidase P enzyme families, respectively (Aravind and Koonin,

1998). During this evolutionary transformation, these proteins are often observed to sacrifice their catalytic activity and adopt novel transcriptional regulatory functions.

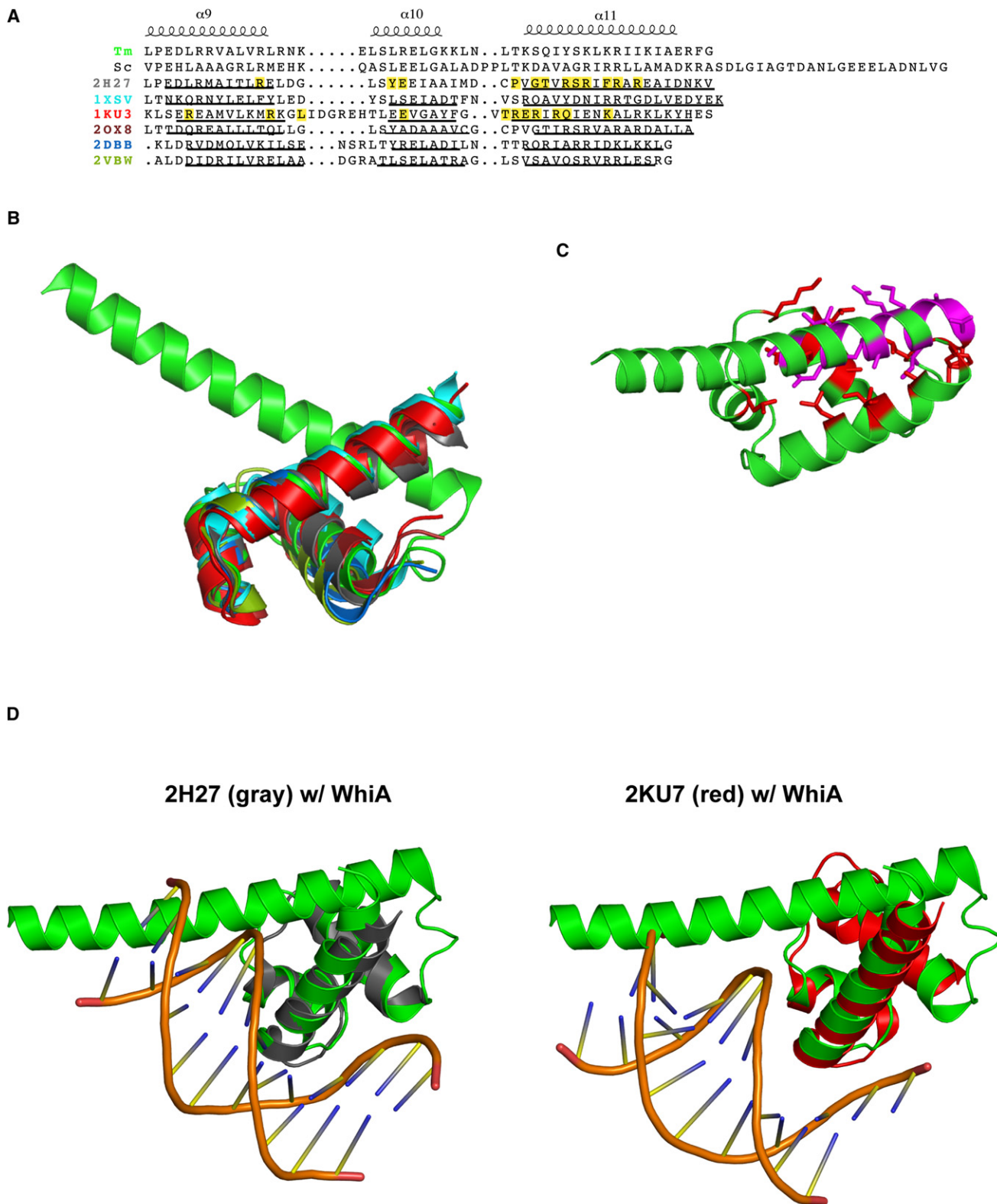
Assuming that the WhiA family of transcription regulators was derived from mobile endonuclease ancestors, several key events would have occurred during the evolutionary creation and expansion of this protein family. First, an ancestor to modern day bacteria would have acquired an active LAGLIDADG homing endonuclease. At some point subsequent to this initial genetic transfer, the homing endonuclease gained an additional HTH protein domain, lost its ability to cleave DNA, and became a completely domesticated transcription factor.

A key question regarding such an evolutionary scenario is whether a bifunctional intermediate might have existed during this process, in which the endonuclease activity and the ability to act as a transcriptional regulator were shared by a single protein scaffold (a relationship commonly termed moonlighting). Many modern homing endonucleases require relatively tight regulation of their own expression, primarily to avoid toxicity that might be associated with its own overexpression. At least one such endonuclease (the phage-derived I-TevI enzyme) also serves as its own transcriptional autorepressor (Edgell et al., 2004). A scenario in which a homing endonuclease first adopted a very simple form of transcriptional regulatory activity to regulate its own expression, followed by subsequent incorporation into more complex forms of gene regulation and loss of its original endonuclease activity, seems attractive.

### WhiA Proteins and Transcriptional Regulation

Genetic studies in *S. coelicolor* suggest that WhiA<sup>Sc</sup>, an essential sporulation factor, functions as a transcriptional activator by regulating the expression of numerous genes, including its own (Ainsa et al., 2000). WhiA<sup>Sc</sup> contains two promoters, a low-level upstream promoter that is expressed independently of WhiA<sup>Sc</sup>, and a sporulation-specific promoter more proximal to the WhiA transcriptional start site that requires WhiA<sup>Sc</sup> for expression (Ainsa et al., 2000). Bacterial transcriptional activators often function by binding on or near the  $-35$  region of promoters and recruiting the bacterial RNA polymerase holoenzyme to the promoter. It is therefore noteworthy that the closest structural homologs of the WhiA<sup>Tm</sup> HTH domain are HTH domains (i.e., domain 4) of bacterial sigma70 factors, which bind to  $-35$  promoter elements.

Given the data summarized above, it seems likely that the bacterial Duf199/WhiA proteins might be involved in interactions with a DNA target and also with additional protein factors within the transcriptional apparatus. The structure of WhiA from *T. maritima* reveals a striking combination of two individual domains that are each well known for their abilities to facilitate both DNA recognition and protein-protein association. LAGLIDADG endonucleases typically recognize long (20 or more base-pair) targets with variable fidelity, whereas HTH domains recognize shorter (6 to 8 base-pair) targets. However, these domains are also capable of facilitating packing interactions with additional structural protein domains: LAGLIDADG endonucleases are often fused to protein splicing domains (termed "inteins"), and HTH domains can facilitate a variety of protein binding and dimerization interactions, primarily in eukaryotes (Aravind et al., 2005). Both the LAGLIDADG and



**Figure 3. The HTH Domain of WhiA<sup>Tm</sup> Is Structurally Related to Domain 4 of Bacterial  $\sigma$  Factors**

(A) Structure-based alignment of WhiA<sup>Tm</sup> (top), and its closest structural homologs identified using the DALI server. The secondary structure elements shown above the alignment correspond to WhiA<sup>Tm</sup>. Residues highlighted yellow in 2H27 and 1KU3 correspond to amino acids that contact DNA. Underlined regions

HTH domains would therefore seem to be well-suited to adopting novel functions as part of a transcription factor complex, because they are both often found in multidomain architectures and can apparently facilitate a wide variety of macromolecular interactions.

Given the broad spectrum of environmental niches of organisms that contain WhiA, it is likely that during evolution WhiA has been utilized to regulate diverse biological pathways. For example, whereas WhiA from *S. coelicolor* has a well-established role in sporulation, in other nonsporulating bacteria WhiA homologs likely regulate distinct pathways. In the course of adopting novel functions in various organisms, WhiA would have also likely developed novel protein interaction partners. A possible WhiA-binding candidate from *S. coelicolor* is the iron-sulfur protein WhiB, which shares a genetic phenotype similar to WhiA in sporulation. However, this interaction would only occur in a subset of organisms that contain WhiA because the WhiB family is only present in *Actinomycetes*.

## EXPERIMENTAL PROCEDURES

### Protein Production

The WhiA<sup>Tm</sup> gene was amplified from *T. maritima* genomic DNA (ATCC) using polymerase chain reaction, cloned into the pET24 expression vector (Novagen), and expressed in BL21(DE3)RIL bacteria (Novagen) in LB media supplemented with 1% glucose and antibiotics (kanamycin and chloramphenicol). A starter culture was grown overnight at 37°C, and diluted 1:50 the next morning into media with antibiotics. WhiA expression was induced with the addition of 1 mM IPTG when the OD<sub>600</sub> reached 0.6–0.8, and incubated for an additional 3 hr at 37°C. Cells were then centrifuged and stored at –20°C. Pellets were thawed and lysed by sonication on ice in 300 mM NaCl, 50 mM Tris (pH 8.0), 1 mM PMSF. After centrifuging for 30 min in a SS34 rotor (Sorvall) at 43,000 × g, the supernatant was incubated in a 70°C water bath for 15 min followed by a 60 min spin at 43,000 × g in an SS34 rotor. The cleared lysate was then loaded onto a 1 ml Heparin HiTrap column (GE Healthcare) at room temperature (using a Bio-Rad peristaltic pump), and eluted on a Pharmacia AktaPrime FPLC with a 300 mM to 1 M NaCl gradient in 25 mM Tris (pH 8.0) buffer over 30 column volumes. The peak fraction typically eluted at ~600–700 mM NaCl. Peak fractions were pooled, precipitated with 25% (w/v) ammonium sulfate and centrifuged in 2 ml Eppendorf tubes at 4°C, 13,000 rpm in an Eppendorf tabletop centrifuge. The pellet was resuspended in 25 mM Tris (pH 7.5) to ~5 mg/ml and dialyzed against 25 mM Tris (pH 7.5), 150 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>. Protein concentration was estimated using optical absorbance at 280 nm with a calculated molar extinction coefficient of 10,430 M<sup>-1</sup>cm<sup>-1</sup>.

For production of selenomethionyl-containing protein, we used BL21(DE3)-RIL bacteria and followed the method of Double (1997), in which the methionine biosynthesis pathway is inhibited prior to induction by the addition of Ile, Lys, and Thr, and supplemented with selenomethionine (Fisher, Acros). Briefly, an O/N culture of BL21(DE3)-RIL bacteria transformed with pET24\_WhiA<sup>Tm</sup> was grown in a 10 ml overnight culture in LB/1% glucose, kanamycin, and chloramphenicol. The next morning the starter culture was pelleted and resus-

ended in 10 ml minimal media, diluted into 1 l minimal media with antibiotics, and grown to OD<sub>600</sub> of 0.6. Amino acids that shut down cellular methionine production and selenomethionine were added, the culture incubated for 15 min, and then induced with 1 mM IPTG for 3 hr. The remainder of the protein purification was performed as described above.

### Crystallization

Crystals of the isolated LAGLIDADG were grown from a solution of the full-length protein that had been subjected to in situ proteolysis with trypsin (Dong et al., 2007). Trypsin (Sigma/Aldrich) was added to the full-length protein solution on ice in a range of ratios (1:1000 to 1:10,000 w/w) just before setting up crystallization trials. Drops were set using the hanging drop vapor diffusion method (1 μl protein/protease plus 1 μl mother liquor) and incubated at 18°C. Trypsin-treated WhiA<sup>Tm</sup> crystallized in 20% ethanol, 0.1 M Tris (pH 9.0), and 200 mM NaCl. Selenomethionine-containing crystals (also treated with trypsin) were grown in 20% ethanol, 0.1 M Tris (pH 9.3), and 0.2 M KCl. For cryopreservation, crystals were transferred to mother liquor containing 25% glycerol and flash frozen in liquid nitrogen.

Crystals of the full-length WhiA<sup>Tm</sup> protein were generated by the hanging drop method in 0.1 M Tris (pH 8.0), 0.2 M NaCl, and 10% PEG8000 at 18°C. Crystals typically grew in clusters after about 2 weeks, and had to be separated for cryopreservation (as described above).

### Data Collection

X-ray data sets on native crystals were collected using an in-house rotating anode HF-007 X-ray generator equipped with a RAXIS IV++ imaging plate area detector (both instruments from Rigaku, Inc.). A single-wavelength data set at the peak energy (12.661 keV) with inverse-beam geometry was collected for a SeMet-containing crystal (trypsin form only) at the Advanced Light Source synchrotron facility (Berkeley, CA), beamline 5.0.2. Data were indexed and scaled using HKL2000 software (Otwinowski and Minor, 1997). Phases for the crystal containing the isolated LAGLIDADG domain were solved using SOLVE (Terwilliger and Berendzen, 1999) and solvent flattened using RESOLVE (Terwilliger and Berendzen, 1999). The model was built using COOT (Emsley and Cowtan, 2004) and was refined using TLS restrained refinement in Refmac5 (Murshudov et al., 1997) while monitoring R<sub>free</sub> (Kleywegt and Brunger, 1996), and also monitoring the overall geometric quality of the model using PROCHECK (Laskowski et al., 1993). TLS parameters were defined using the TLS online server (<http://skuld.bmsc.washington.edu/~tlsmdl/>) (Painter and Merritt, 2006). The resulting models of the WhiA LAGLIDADG domain (of which there were two in the asymmetric unit) were then used as search models in molecular replacement to solve the phases of the full-length crystals using PHASER (McCoy et al., 2007). The density for the HTH domain of WhiA<sup>Tm</sup> was initially built using ARP/WARP (Perrakis et al., 2001) and unaccounted density was manually built in COOT (Emsley and Cowtan, 2004). The CCP4i suite of programs (CCP4, 1994) was extensively used throughout the structure solving process to implement programs and adjust files.

### ACCESSION NUMBERS

The coordinates for the refined models of the isolated LAGLIDADG domain and the full-length WhiA<sup>Tm</sup> protein have been deposited in the RCSB protein structure database (PDB ID codes 3HYI and 3HYJ).

indicate  $\alpha$  helices. PDB codes correspond to the following structures: 2H27: domain 4 of *E. coli* SigmaE (1.96 Å  $\alpha$ -carbon rmsd over 65 equivalent positions); 1XSV: *S. aureus* hypothetical UPF0122 protein SAV123 (2.59 Å over 67 equivalent positions); 1KU3: *T. aquaticus* RNA Polymerase Sigma subunit, domain 4 (2.44 Å over 57 equivalent positions); 2OX8: *M. tuberculosis* SigC –35 element promoter recognition domain (2.57 Å over 69 equivalent positions); 2DBB: *P. horikoshii* HTH-type transcriptional regulator (4.09 Å over 119 equivalent positions); 2VBW: *M. tuberculosis* Rv3291c, transcriptional regulator (2.31 Å over 96 equivalent positions).

(B) Superposition of the structures (color coded as in A). The superposition was created by FATCAT (Ye and Godzik, 2003).

(C) The  $\alpha 8$  “linker” helix binds to a cleft formed between the WhiA<sup>Tm</sup>  $\alpha 9$  and  $\alpha 11$  helices. Residues on the  $\alpha 8$  helix that make contacts with the cleft are in magenta; residues of the HTH domain that contact  $\alpha 8$  are colored red.

(D) Structures of 2H27 (left, gray) and 2KU7 (right, red) bound to DNA. WhiA<sup>Tm</sup> (green) is superposed over the structures, demonstrating a potential clash of  $\alpha 8$  with the minor groove.



## SUPPLEMENTAL DATA

Supplemental Data include four figures and can be found with this article online at [http://www.cell.com/structure/supplemental/S0969-2126\(09\)00328-1](http://www.cell.com/structure/supplemental/S0969-2126(09)00328-1).

## ACKNOWLEDGMENTS

The authors thank the staff of the Advanced Light Source (ALS) beamline 5.0.2 and members of the FHCRC structural biology program for technical assistance, advice and discussion. Funding provided by the NIH (GM49857 and CA133833) and the FHCRC Division of Basic Sciences. The authors declare no conflicts of interest.

Received: June 28, 2009

Revised: August 18, 2009

Accepted: August 18, 2009

Published: October 13, 2009

## REFERENCES

- Ainsa, J.A., Ryding, N.J., Hartley, N., Findlay, K.C., Bruton, C.J., and Chater, K.F. (2000). WhiA, a protein of unknown function conserved among gram-positive bacteria, is essential for sporulation in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* *182*, 5470–5478.
- Aravind, L., and Koonin, E.V. (1998). Eukaryotic transcription regulators derive from ancient enzymatic domains. *Curr. Biol.* *8*, R111–R113.
- Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* *29*, 231–262.
- Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA* *98*, 10037–10041.
- Burt, A., and Koufopanou, V. (2004). Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr. Opin. Genet. Dev.* *14*, 609–615.
- Caetano-Anolles, G., Kim, H.S., and Mittenthal, J.E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc. Natl. Acad. Sci. USA* *104*, 9358–9363.
- CCP4 (Collaborative Computational Project, Number 4). (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* *50*, 760–763.
- Dong, A., Xu, X., Edwards, A.M., Chang, C., Chruszcz, M., Cuff, M., Cymborowski, M., Di Leo, R., Egorova, O., Evdokimova, E., et al. (2007). In situ proteolysis for protein crystallization and structure determination. *Nat. Methods* *4*, 1019–1021.
- Doublet, S. (1997). Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* *276*, 523–530.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* *32*, 1792–1797.
- Edgell, D.R., Derbyshire, V., Van Roey, P., LaBonne, S., Stanger, M.J., Li, Z., Boyd, T.M., Shub, D.A., and Belfort, M. (2004). Intron-encoded homing endonuclease I-TevI also functions as a transcriptional autorepressor. *Nat. Struct. Mol. Biol.* *11*, 936–944.
- Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* *60*, 2126–2132.
- Flick, K.E., Jurica, M.S., Monnat, R.J., Jr., and Stoddard, B.L. (1998). DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-Ppol. *Nature* *394*, 96–101.
- Goet, P., Courcelle, E., Stuart, D.I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* *15*, 305–308.
- Grishin, N.V. (2001). Mh1 domain of Smad is a degraded homing endonuclease. *J. Mol. Biol.* *307*, 31–37.
- Ho, Y., Kim, S.J., and Waring, R.B. (1997). A protein encoded by a group I intron in *Aspergillus nidulans* directly assists RNA splicing and is a DNA endonuclease. *Proc. Natl. Acad. Sci. USA* *94*, 8994–8999.
- Holm, L., Kaariainen, S., Rosenstrom, P., and Schenkel, A. (2008). Searching protein structure databases with DALI-Lite v.3. *Bioinformatics* *24*, 2780–2781.
- Jeffery, C.J. (2003). Moonlighting proteins: old proteins learning new tricks. *Trends Genet.* *19*, 415–417.
- Khare, D., Ziegelin, G., Lanka, E., and Heinemann, U. (2004). Sequence-specific DNA binding determined by contacts outside the helix-turn-helix motif of the ParB homolog KorB. *Nat. Struct. Mol. Biol.* *11*, 656–663.
- Kleywegt, G.J., and Brunger, A.T. (1996). Checking your imagination: applications of the free R value. *Structure* *4*, 897–904.
- Knizewski, L., and Ginalski, K. (2007). Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. *Cell Cycle* *6*, 1666–1670.
- Koufopanou, V., and Burt, A. (2005). Degeneration and domestication of a selfish gene in yeast: molecular evolution versus site-directed mutagenesis. *Mol. Biol. Evol.* *22*, 1535–1538.
- Laskowski, R.A., Moss, D.S., and Thornton, J.M. (1993). Main-chain bond lengths and bond angles in protein structures. *J. Mol. Biol.* *231*, 1049–1067.
- Magnani, E., Sjolander, K., and Hake, S. (2004). From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *Plant Cell* *16*, 2265–2277.
- McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. *J. Appl. Crystallogr.* *40*, 658–674.
- Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* *53*, 240–255.
- Nagano, N., Orengo, C.A., and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* *321*, 741–765.
- Nowotny, M. (2009). Retroviral integrase superfamily: the structural perspective. *EMBO Rep.* *10*, 144–151.
- Orengo, C.A., and Thornton, J.M. (2005). Protein families and their evolution—a structural perspective. *Annu. Rev. Biochem.* *74*, 867–900.
- Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* *276*, 307–326.
- Painter, J., and Merritt, E.A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D Biol. Crystallogr.* *62*, 439–450.
- Perrakis, A., Harkiolaki, M., Wilson, K.S., and Lamzin, V.S. (2001). ARP/wARP and molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* *57*, 1445–1450.
- Silva, G.H., Dalgaard, J.Z., Belfort, M., and Van Roey, P. (1999). Crystal structure of the thermostable archaeal intron-encoded endonuclease I-Dmol. *J. Mol. Biol.* *286*, 1123–1136.
- Stoddard, B.L. (2005). Homing endonuclease structure and function. *Q. Rev. Biophys.* *38*, 49–95.
- Terwilliger, T.C., and Berendzen, J. (1999). Automated MAD and MIR structure solution. *Acta Crystallogr. D Biol. Crystallogr.* *55*, 849–861.
- Tuinstra, R.L., Peterson, F.C., Kutlesa, S., Elgin, E.S., Kron, M.A., and Volkman, B.F. (2008). Interconversion between two unrelated protein folds in the lymphotactin native state. *Proc. Natl. Acad. Sci. USA* *105*, 5057–5062.
- Xie, Z., Li, W., Tian, Y., Liu, G., and Tan, H. (2007). Identification and characterization of sawC, a whiA-like gene, essential for sporulation in *Streptomyces ansochromogenes*. *Arch. Microbiol.* *188*, 575–582.
- Ye, Y., and Godzik, A. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* *19* (Suppl 2), ii246–ii255.