doi:10.1016/j.jmb.2009.11.004 J. Mol. Biol. (2010) 395, 558–567

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Structural Determinants of DNA Binding by a *P. falciparum* ApiAP2 Transcriptional Regulator

Scott E. Lindner[1], Erandi K. De Silva[2], James L. Keck[1] and Manuel Llinás[2]*

[1]*Department of Biomolecular Chemistry, University of Wisconsin School of Medicine and Public Health, 550 Medical Sciences Center, 1300 University Avenue, Madison, WI 53706, USA*

[2]*Molecular Biology and Lewis-Sigler Institute for Integrative Genomics, Princeton University, 246 Carl Icahn Lab Washington Road, Princeton, NJ 08544, USA*

Received 24 August 2009; received in revised form 2 November 2009; accepted 4 November 2009
Available online 10 November 2009

Edited by K. Morikawa

Putative transcription factors have only recently been identified in the *Plasmodium* spp., with the major family of regulators comprising the Apicomplexan Apetala2 (AP2) proteins. To better understand the DNA-binding mechanisms of these transcriptional regulators, we characterized the structure and *in vitro* function of an AP2 DNA-binding domain from a prototypical Apicomplexan AP2 protein, PF14_0633 from *Plasmodium falciparum*. The X-ray crystal structure of the PF14_0633 AP2 domain bound to DNA reveals a β-sheet fold that binds the DNA major groove through base-specific and backbone contacts; a prominent α-helix supports the β-sheet structure. Substitution of predicted DNA-binding residues with alanine weakened or eliminated DNA binding in solution. In contrast to plant AP2 domains, the PF14_0633 AP2 domain dimerizes upon binding to DNA through a domain-swapping mechanism in which the α-helices of the AP2 domains pack against the β-sheets of the dimer mates. DNA-induced dimerization of PF14_0633 may be important for tethering two distal DNA loci together in the nucleus and/or for inducing functional rearrangements of its domains to facilitate transcriptional regulation. Consistent with a multisite binding mode, at least two copies of the consensus sequence recognized by PF14_0633 are present upstream of a previously identified group of sporozoite-stage genes. Taken together, these findings illustrate how *Plasmodium* has adapted the AP2 DNA-binding domain for genome-wide transcriptional regulation.

Keywords: malaria; *Plasmodium falciparum*; AP2 domain; X-ray crystallography; domain swapping

Introduction

The malaria parasite *Plasmodium falciparum* exhibits a complex life cycle that involves infection of both a vertebrate human host and an invertebrate mosquito vector. The various host cell types and morphological transitions involved in the development of the parasite suggest the utilization of fine-tuned regulation of gene expression at all stages.[1] However, dissection of transcriptional regulation in *P. falciparum* and other related apicomplexan parasites has proven difficult because their genomes possess few proteins with detectable homology to established transcription factors in model eukaryotes. To date, only one major family of transcriptional regulators has been identified, the Apicomplexan Apetala2 (ApiAP2) family, which has undergone a lineage-specific expansion in the apicomplexan phylum.[2]

ApiAP2 proteins vary in size from hundreds to thousands of amino acids, and each contains at least one copy of a conserved ~60-residue AP2 domain that is weakly homologous to the plant AP2 domain. In the plant *Arabidopsis thaliana*, AP2 domains are the second most common DNA-binding domain after the Myb domain and are typically found in small proteins on the order of 300 amino acids.[3] Plant AP2 domains mediate specific DNA recognition in all members of the AP2/ethylene response factor (ERF) family, which control development and stress responses in species throughout the plant

*Corresponding author. E-mail address: mllinas@Princeton.edu.

Abbreviations used: AP2, Apetala2; ApiAP2, Apicomplexan AP2; ERF, ethylene response factor; dsDNA, double-stranded DNA; SeMet, selenomethionine; EMSA, electrophoretic mobility shift assay; PBM, protein binding microarray; GST, glutathione *S*-transferase.

kingdom.[3] Typically in plants, the AP2 domains are found as one- or two-domain arrangements in a given protein, and two-domain AP2 proteins have been shown to contact up to 17 contiguous base pairs of DNA.[4] An NMR solution structure of a single *A. thaliana* AP2 domain in complex with DNA reveals a mode of DNA binding where three β-strands contact the DNA backbone and specific bases within the major groove. The β-strands are stabilized by a C-terminal α-helix that does not contact the DNA.[5]

Despite having limited homology to plant AP2 domains, Apicomplexan AP2 domains bind DNA with high sequence specificity and can function as regulators of transcription.[6,7] There are a total of 26 ApiAP2 proteins predicted in the *Plasmodium* genus, and although a given AP2 DNA-binding domain can be highly conserved across related species, there is sparse homology outside of the DNA-binding domain. Presumably, the ApiAP2 proteins must also contain activation domains and additional protein–protein interaction domains, although there are no additional Pfam domains predicted for any of these proteins.[8] Furthermore, paralogous AP2 domains from *Plasmodium* spp. share little sequence identity with each other, in contrast to those from *Arabidopsis* where paralogous AP2/ERF proteins can be functionally grouped on the basis of their primary sequence identity as well as their DNA-binding preferences.[3,9,10] This divergence of AP2 domains in the Apicomplexans is reflected in the diversity of DNA sequences to which these domains can bind.[11]

Recently, an ApiAP2 protein from the murine *Plasmodium berghei* species was shown to be essential *in vivo* for the activation of a set of ookinete-specific genes required during the parasite's replicative stage in infected mosquitoes.[7] This protein, named AP2-O, possesses only a single AP2 DNA-binding domain but appears to require two copies of a TAGCTA DNA sequence for high-affinity binding. Furthermore, all experimentally determined downstream target genes of AP2-O contain two or more copies of the high-affinity site (TAGCTA) in their upstream regions. These data raise the possibility that AP2-O, and other ApiAP2 transcriptional regulators, may dimerize when bound to their target sequences. Dimerization could influence regulation of gene expression, as has been well documented with other eukaryotic specific transcription factors.[12] Although dimerization of plant AP2 domains has not been demonstrated, AP2 domains have been shown to contact other transcription factors and mediate protein–protein interactions through the AP2 domain.[13,14]

The AP2-containing protein PF14_0633 from *P. falciparum* is an 813 amino acid protein that includes a single, highly conserved 60 amino acid AP2 domain and only one other identifiable domain, a short AT-hook DNA-binding motif that directly precedes the AP2 domain.[15] We previously demonstrated that the AP2 domain of PF14_0633 binds a TGCATGCA DNA sequence with high sequence specificity.[6] In this work, we present the crystal structure of a domain-swapped dimer of the AP2 domain from PF14_0633 in complex with double-stranded DNA (dsDNA) containing its consensus binding sequence ("cognate DNA"), as well as solution studies that examine DNA binding and the mechanism of dimer formation. Strikingly, specific binding of the PF14_0633 AP2 domain to its consensus DNA sequence was found to stimulate domain-swapped dimerization. We propose a model whereby DNA-induced dimerization of the AP2 domain of PF14_0633 facilitates the conformational rearrangement of the remainder of the protein or its interaction partners, while simultaneously looping out intervening DNA between pairs of binding sites enriched in the upstream regions of a set of sporozoite-specific genes.

Results and Discussion

Structure of the AP2 domain of PF14_0633 specifically bound to dsDNA

To elucidate the mechanism of DNA binding of an Apicomplexan AP2 domain, we determined the 2.2 Å resolution crystal structure of the AP2 domain of PF14_0633 bound to an 8-bp dsDNA comprising its consensus binding sequence TGCATGCA (Fig. 1 and Supplemental Fig. 1). Single-wavelength anomalous dispersion phasing of a crystal of selenomethionine (SeMet)-incorporated AP2 domain bound to DNA produced a high-quality experimental electron density map into which a partial molecular model was built. This model was used as a molecular replacement search model to phase the higher-resolution diffraction data from a native complex, which yielded electron density maps that allowed building of the final model (Fig. 1; Table 1). The asymmetric unit contains two molecules of the AP2 domain of PF14_0633, with each monomer bound to an individual dsDNA molecule.

The overall structure of the DNA-bound AP2 domain of PF14_0633 retains many of the canonical features of similar DNA-binding domains for which structures have been determined, such as ATERF1 (*A. thaliana* ethylene response factor 1)[5] and the yeast PI-SceI homing endonuclease.[16] Three antiparallel β-strands from the AP2 domain of PF14_0633 wrap into the major groove of the bound dsDNA and contain all residues making base-specific contacts. A flexible loop connects these strands with an extended α-helix, which supports the opposing face of the β-sheet. However, our structure demonstrates that PF14_0633 dimerizes through a three-dimensional domain-swapping mechanism in which the α-helix of one protomer is packed against the β-sheet of its dimer mate (Fig. 1b). Similar dimerization was not previously observed in the NMR solution structure of the related plant AP2 protein ATERF1.[5] The interface derived from domain swapping results in burial of a large surface area (~3140 Å^2) of primarily

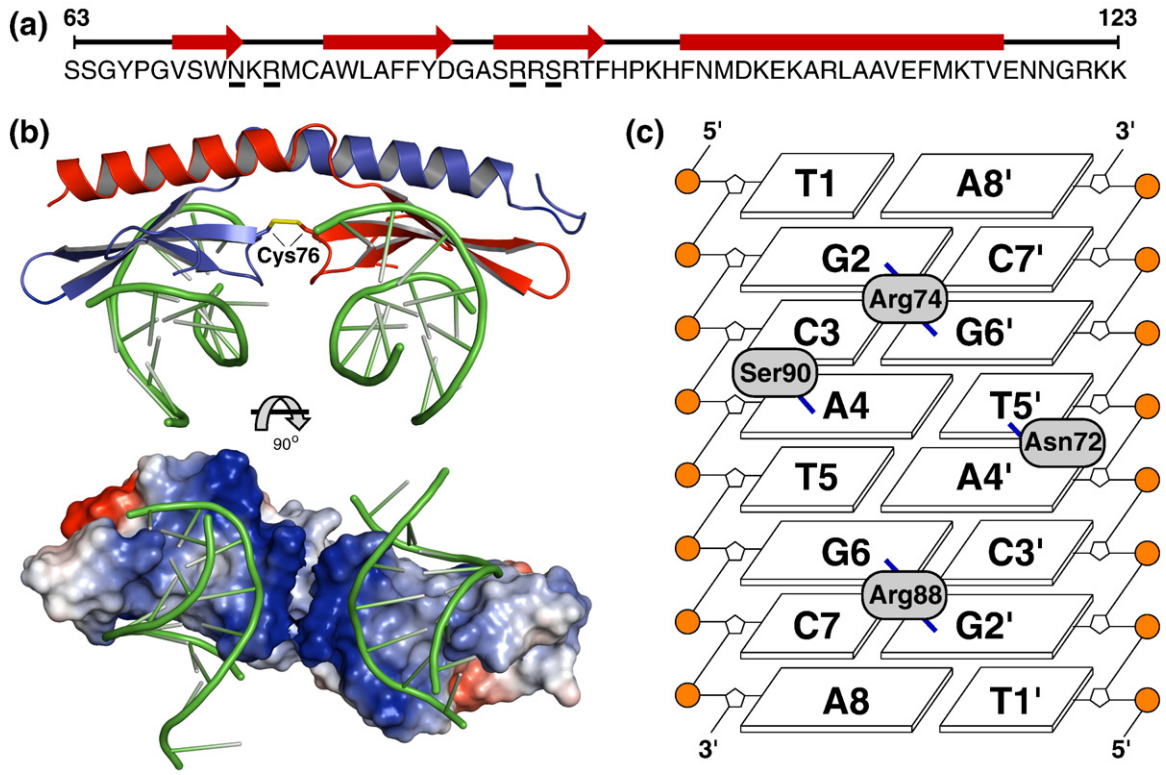

Fig. 1. Structure of the AP2 domain of PF14_0633 specifically bound to dsDNA. (a) Secondary structural elements of the AP2 domain of PF14_0633 are depicted above the primary sequence as arrows (β-strands) and a rectangle (α-helix). Residues that make base-specific contacts to dsDNA are underlined. (b) (Top) Ribbon diagram of the crystal structure of a domain-swapped dimer of the AP2 domain of PF14_0633 (one monomer shaded blue, another red) bound to independent dsDNA molecules (shaded green). Disulfide-bonded Cys76 residues are indicated and shaded yellow. (Bottom) Orthogonal view with protein electrostatic surface potentials depicted (blue, electropositive; red, electronegative). (c) DNA contact map depicting direct base-specific contacts between the Asn72, Arg74, Arg88, or Ser90 side chains and the indicated DNA bases.

hydrophobic residues. Dimerization of the PF14_0633 AP2 domain also aligns Cys76 residues of each monomer with one another with sufficient proximity to allow disulfide bond formation (Fig. 1b). We note that the Cys76 residue is conserved in all orthologues of PF14_0633 in *Plasmodium* spp. but is not conserved in other related apicomplexan species.

Several contacts between the PF14_0633 AP2 domain and its DNA-binding sequence appear to be important to binding specificity (Fig. 1c, summarized in Supplemental Table 1). Residues of particular note are Asn72, Arg74, Arg88, and Ser90, which make base-specific contacts to the consensus binding sequence. Of these, only one residue (Asn72) was previously implicated in DNA binding by computational modeling.[2] The two arginine side chains can each contact both strands of DNA (O6 atoms of dG2 and dG6) (Fig. 1c). The side chains of Asn72 and Ser90 contact a central AT base pair, each to one strand. All four of these residues are invariant among apicomplexan orthologues of PF14_0633 with BLAST expect scores <1e−05, including AP2 domains from *Plasmodium* spp. and *Toxoplasma*, *Theileria*, *Babesia* and *Cryptosporidium* genera. The AP2 domain in this protein therefore appears to have conserved specific binding to the TGCATGCA consensus DNA over a large evolutionary distance separating these apicomplexan species by conservation of four key DNA contact residues. Indeed, the distantly related *Cryptosporidium* PF14_0633 orthologue binds specifically to the same TGCATGCA sequence.[6] In addition to these base-specific contacts, several phosphodiester backbone contacts are made, most notably His94, which is located in the hinge region between the α-helix and the β-sheet of the AP2 domain. His94, Pro95, and several adjacent residues are all well conserved among the orthologous PF14_0633 AP2 domains. A common mechanistic model for domain swapping implicates strain in Pro-containing hinge regions, which relies on Pro isomerization to help drive conformational changes during dimerization.[17,18] The interaction of His94 with the phosphodiester backbone may influence the conformation of the adjacent Pro95 side chain and act as a trigger for domain swapping (discussed further below).

Residues making base-specific DNA contacts in the crystal structure are important in solution for DNA binding

In order to verify the determinants of DNA-binding specificity as predicted in the crystal struc-

Table 1. Data collection, phasing, and refinement statistics of the AP2 domain of PF14_0633 specifically bound to dsDNA

	SeMet anomalous peak	Native
Data collection[a]		
Space group	$C222_1$	$C222_1$
Cell dimensions: *a*, *b*, *c* (Å)	43.65, 59.18, 173.93	43.81, 58.82, 177.30
Wavelength (Å)	0.97886	0.97886
Resolution (Å)	50.0–2.40 (2.44–2.40)	50.0–2.20 (2.26–2.20)
R_{sym}[b]	0.107 (0.477)	0.075 (0.381)
$I/\sigma I$	21.9 (2.5)	32.1 (4.5)
Completeness (%)	96.7 (71.3)	91.9 (72.4)
Redundancy	11.4 (4.2)	11.0 (9.0)
Phasing		
Figure of merit (before/after solvent flattening)	0.35/0.70	
Refinement		
Resolution (Å)		29.55–2.20 (2.26–2.20)
R_{work}/R_{free}		0.230/0.279
No. of atoms		1691
Protein		971
DNA		623
Water		97
⟨*B*-factors⟩		
Protein		59.1
DNA		65.4
Water		61.0
r.m.s. deviations		
Bond lengths (Å)		0.009
Bond angles (°)		1.37

Values in parentheses are for the highest-resolution shell.
[a] Data statistics from one crystal of native or SeMet PF14_0633.
[b] $R_{sym} = \sum\sum_j | I_j - \langle I \rangle | / SI_j$, where I_j is the intensity measurement for reflection j and $\langle I \rangle$ is the mean intensity for multiply recorded reflections.

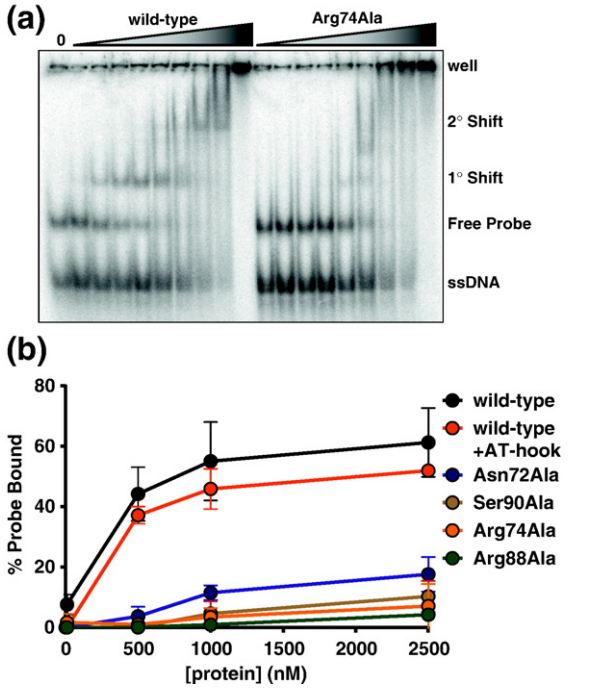

Fig. 2. EMSA of the AP2 domain of PF14_0633 reveals the DNA-binding contributions of several residues. (a) A representative EMSA gel comparing DNA binding by wild-type (left) and a variant (Arg74Ala) PF14_0633 (right) AP2 domain. Ten femtomoles of end-labeled 28-bp probe containing the consensus binding sequence of the PF14_0633 AP2 domain was incubated in the absence or in the presence of 0.1–50 μM AP2 domain. (b) Graph of percentage of specific DNA binding for the wild-type PF14_0633 AP2 domain, four alanine substitution variants, and a fifth variant that includes a native AT-hook located N-terminal of the AP2 domain, as calculated by quantification of the primary EMSA gel shift.

ture, we generated four variants of the PF14_0633 AP2 domain by substituting alanine for the residues observed to make base-specific DNA contacts (Asn72Ala, Arg74Ala, Arg88Ala, and Ser90Ala). These variants were subjected to electrophoretic mobility shift assays (EMSAs; Fig. 2) and protein binding microarrays[19,20] (PBMs; Supplemental Fig. 4) to assess whether these amino acid substitutions altered DNA-binding stability and specificity in solution. All four variants were found to have circular dichroism spectra highly similar to that of the wild-type AP2 domain (Supplemental Fig. 2), indicating that functional differences were not due to differences in protein folding.

We assessed the affinity of the purified wild-type PF14_0633 AP2 domain for its consensus sequence using EMSA (Supplementary Fig. 3). Binding of this domain to two probes (28 and 122 bp in length), each containing a core 28-bp region derived from the upstream region of the *pfi0540w* gene of the *P. falciparum* 3D7 genome, as used previously,[6] was measured (see Materials and Methods). The $K_{d\,app}$ of the wild-type domain for its TGCATGCA target sequence was determined to be 0.5–1.0 μM. Mutation of the central two AT base pairs of the consensus sequence to GC significantly reduced the binding affinity of PF14_0633, confirming that these bases are critical for specific DNA recognition.

The ability of the variant AP2 domains (Asn72Ala, Arg74Ala, Arg88Ala, and Ser90Ala) to bind specifically to dsDNA was evaluated next. The 28-bp dsDNA probe (described above) was incubated with increasing concentrations of the wild-type protein or its variants. A representative gel image for wild-type and Arg74Ala AP2 domains is shown in Fig. 2a, where the wild-type protein forms a specific complex with dsDNA with a $K_{d\,app}$ of 0.5 μM (primary shift). A slower-mobility form of the wild-type protein is observed at concentrations of 2.5 μM and above (secondary shift). Quantification of the percentage of shifted probe for each Ala-substituted variant AP2 domain demonstrated that all bind dsDNA very poorly, binding specifically only at concentrations just below those where nonspecific binding to both single-stranded DNA and dsDNA occurs (Fig. 2b). This is consistent with the protein–DNA interface observed in the crystal structure being used in solution. We also assessed the contribution of the AT-hook element (residues 39–50) to DNA binding by extending the expressed

portion of PF14_0633 to include residues 38–123. AT-hooks are short auxiliary elements that contact DNA in a non-sequence-specific manner and are commonly found adjacent to specific DNA-binding domains.[15] The addition of the AT-hook motif did not significantly affect the binding affinity of wild-type PF14_0633 to the 28-bp probe (Fig. 2b). These results demonstrate that specific DNA binding was weakened in the variant AP2 domains, consistent with elimination of base-specific contacts provided by each side chain.

We reasoned that alteration of the individual residues that make base-specific contacts to the DNA-binding site might alter the sequence specificity of the PF14_0633 AP2 domain. To address this, we assayed glutathione *S*-transferase (GST)-tagged forms of these variants of PF14_0633 using PBMs, and sequence-specific binding was determined as previously described.[19] As expected, the wild-type AP2 domain bound the consensus binding sequence TGCATGCA. Surprisingly, two of the four variants (Asn72Ala and Ser90Ala) also bound this same sequence, albeit more weakly (Supplemental Fig. 4). This result implies that these residues are not necessary for DNA-binding specificity. Furthermore, it also suggests that the PBM methodology is far more sensitive for detecting lower-affinity DNA-binding events, which we could not detect by EMSA. In contrast to the Asn and Ser variants, DNA binding was completely abrogated with both Arg variants.

Taken together with the results from the EMSAs comparing wild-type and mutant DNA probes (Fig. 2b; Supplemental Fig. 3), the observation that Asn72Ala and Ser90Ala variants still recognize the consensus binding sequence with high statistical significance indicates that base-specific contacts to the central AT base pairs are not necessary for specific DNA binding. Instead, an appropriate balance of DNA rigidity/flexibility provided by central AT base pairs in the palindromic sequence may be important for specific binding. In contrast, Arg74 and Arg88 are required for sequence-specific DNA binding. These data imply that the identified consensus binding sequence represents the minimal binding sequence for the AP2 domain of PF14_0633, which is largely determined by the base-specific contacts of both Arg74 and Arg88.

DNA binding stimulates dimerization by domain swapping within the PF14_0633 AP2 domain

Two striking features of the structure of the AP2 domain of PF14_0633 bound to dsDNA are that the two monomers have domain-swapped α-helices to form a dimer and that the dimeric structure aligns the single native Cys residues within proximity to form a disulfide bond. To explore whether these observations are representative of the behavior of this domain in non-crystallographic conditions, we employed several solution-based experiments to examine the DNA dependence of disulfide bond formation, dimerization, and domain swapping.

We first addressed these questions by assessing whether a disulfide bond is formed at the dimerization interface upon binding to dsDNA. Under nonreducing conditions, no significant difference in disulfide bond formation was observed between the apo and dsDNA-bound forms of the AP2 domain of PF14_0633 (Fig. 3a, nonreducing gel: lanes 1, 3, and 5). As an alternate approach, we tested whether the homobifunctional cross-linker $BM(PEG)_3$, which can covalently react with free thiols on exposed Cys residues, has a differential effect on the apo and DNA-bound forms of the AP2 domain. As shown in Fig. 3a, the presence of cognate dsDNA (containing the consensus binding sequence of PF14_0633's AP2 domain, see Materials and Methods for details) stimulated cross-linking of two monomers to form a dimer approximately 50-fold relative to the apo domain (Fig. 3a, reducing gel: lane 4 *versus* lane 2). This effect is likely due to the proximal localization of the Cys residues by dimerization upon DNA binding. Incubation with non-cognate dsDNA also stimulated cross-linking (Fig. 3a, reducing gel: lane 6), but with a reduced efficiency relative to cognate dsDNA. In agreement with the observation that these Cys residues are buried and/or disulfide bonded upon DNA binding, accessibility of these Cys residues was stimulated 9- to 13-fold by the presence of dsDNA as observed by an iodoacetamide protection/2-nitro-5-thiocyanatobenzoic acid cleavage assay (Supplemental Fig. 5). These data indicate that DNA binding induces dimerization of the PF14_0633 AP2 domain.

Size-exclusion chromatography of the wild-type AP2 domain in the absence or presence of cognate or non-cognate DNA was used as an independent means of testing whether DNA binding stimulates dimerization in solution (Fig. 3b). The wild-type AP2 domain eluted at a volume consistent with the molecular size of a monomer (8.8 kDa) in the absence of DNA. In the presence of cognate (8 bp, 4.8 kDa) or non-cognate (17 bp, 10.6 kDa) dsDNA, the AP2 domain eluted as two distinct peaks each, at volumes consistent with an AP2 dimer bound to one or two dsDNA molecules. The presence of DNA in these elution peaks was confirmed by a decrease in the A280:A260 ratio from 1.56 (apo) to 0.56–0.59 (DNA-bound), as well as by the large increases in peak UV absorbance due to the relatively large extinction coefficients of dsDNA. Samples from both peaks of the cognate dsDNA-bound AP2 complexes were compared to the apo form via reducing and nonreducing 20% SDS-PAGE (Fig. 3b, inset). The faster-eluting species (marked as "∗", elution peak at 60.0 ml, 7% of total) contained a disulfide-bonded dimer, whereas the slower-eluting species (marked as "∗∗", elution peak at 67.5 ml, 93% of total) was composed of a non-covalently bonded dimer (Fig. 3b, inset). Taken together, these experiments indicate that DNA binding stimulates the dimerization of the wild-type AP2 domain in solution.

In order to "capture" the domain-swapped form of the PF14_0633 dimer in solution, we engineered a triple substitution variant AP2 domain for fur-

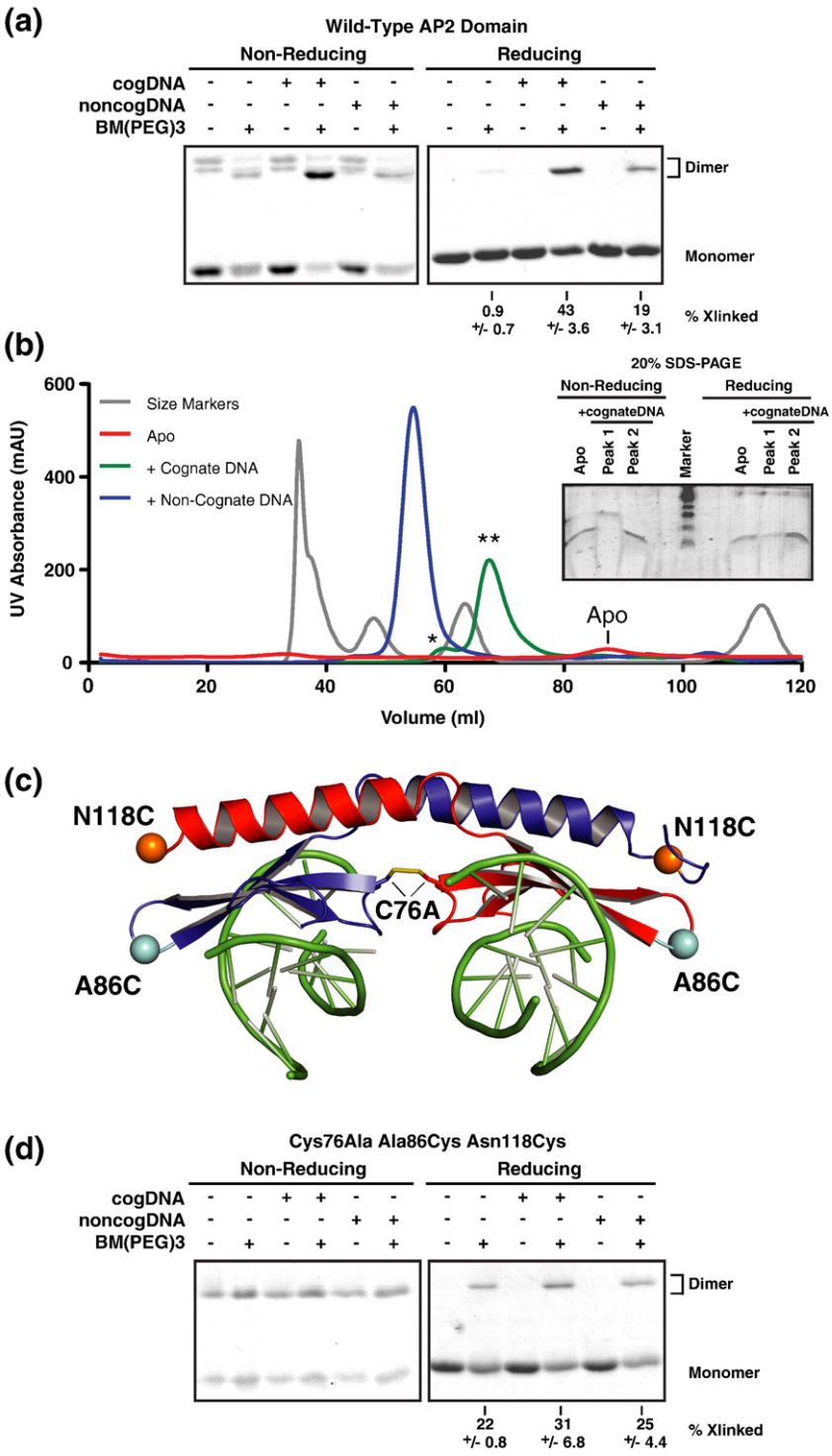

Fig. 3. DNA binding stimulates dimerization of the AP2 domain in solution. (a) PF14_0633 AP2 domain was incubated in the absence or presence of cognate or non-cognate dsDNA and subjected to cross-linking with $BM(PEG)_3$ or vehicle alone. Samples were analyzed by nonreducing (left) or reducing (right) 20% SDS-PAGE. Monomer and dimer forms of the AP2 domain are indicated at the right. The average percentage of total protein cross-linked and standard deviation from three replicates are given below each respective lane. (b) The wild-type AP2 domain of PF14_0633 was incubated in the absence (apo) or presence of cognate or non-cognate dsDNA and applied to a calibrated S100 size-exclusion column. The elution volumes of standard marker proteins of masses 44, 17, and 1.35 kDa were used to interpolate the molecular mass of resultant elution peaks. The peak containing proteins that eluted in the void volume (~35–40 ml) was excluded from this analysis. The elution peak for the apo AP2 domain is indicated due to its low UV absorbance value. Samples from the two elution peaks of the wild-type AP2 domain bound to cognate dsDNA or from the apo AP2 domain were analyzed by nonreducing (left) and reducing (right) 20% SDS-PAGE **(**inset**)**. The more quickly eluting complex (marked as "*", elution peak at 60.0 ml, "Peak 1") contains a disulfide-bonded dimer that is resolved to a monomer upon inclusion of reducing agent. The sample from the more slowly eluting complex (marked as "**", elution peak at 67.5 ml, "Peak 2") contains no disulfide-bonded complexes. (c) Ribbon diagram of the PF14_0633 AP2/DNA complex indicating the positions of Cys76, Ala86 (teal), and Asn118 (orange). A triple substitution variant (Cys76Ala, Ala86Cys, Asn118Cys) is also used in subsequent experiments. (d) The triple substitution variant [depicted in (c)] was analyzed as was the wild-type AP2 domain in (a).

ther cross-linking studies. This protein variant (Cys76Ala, Ala86Cys, Asn118Cys) lacks the single native Cys residue in the dimerization interface but introduces one Cys residue on the domain-swapped α-helix and another in a spatially proximal loop between β-strands (illustrated in Fig. 3c). The introduced Cys residues allow the assessment of domain swapping between dimer mates whereby the β-sheet of one monomer can cross-link to the α-helix donated by the other monomer. As observed above with the wild-type PF14_0633 AP2 domain, the presence of dsDNA containing the consensus binding sequence stimulated cross-linking of the triple point variant (Fig. 3d, reducing gel: lane 4 *versus* lane 2). Again, a reduced level of dimerization is seen in the presence of nonspecific DNA (Fig. 3a,

reducing gel: lane 6). These observations imply that domain swapping also occurs in solution conditions. We hypothesize that DNA binding to the β-sheet may stabilize this portion of the domain, thus allowing the α-helix to adopt the extended conformation and form a domain-swapped dimer.

In light of these data, we propose a model in which the quaternary structure of the AP2 domain is influenced by specific DNA binding (Fig. 4). The DNA trigger could act by stabilizing the dimeric state in a preexisting monomer–dimer equilibrium of the AP2 domain, or, alternatively, DNA binding by an AP2 monomer could induce a conformational change that attracts a second monomer to bind. The AP2 dimer can be stabilized by several factors, including hinge region effects and stabilizing interactions made by the dimer interface or by ligand binding. In the case of the AP2 domain of PF14_0633, both of these factors could be important. First, Pro95 in the hinge region is in a trans conformation in the dimer crystal structure. If the closed monomeric form requires Pro95 to be in a cis conformation, isomerization could both relieve strain and provide energy to produce a conformational change. Notably, other residues in the hinge make backbone contacts with the DNA that may help drive DNA-binding stabilization of the dimeric form of the AP2 domain. Additionally, the structural and biochemical evidence presented herein indicates that the formation of a disulfide bond at the dimerization interface (Cys76) could also increase the stability of the dimer form. Lastly, observations from our three solution-based experiments confirm that binding dsDNA specifically stimulates the formation and/or stabilization of the domain-swapped dimer, as observed in the crystallographic structural model. The influence of these structural features indicates that the composition of the AP2 domain of PF14_0633 is well suited for the formation of a *bona fide* domain-swapped dimer, which may be relevant to its role in transcriptional regulation.

Many eukaryotic specific transcription factors form requisite dimers or higher-order multimers to assemble into their functional conformation. Some, such as Oct-1, Pit-1, and Ets-1, become competent for transcriptional regulation by differential homodimerization induced by the sequence and architecture of their DNA-binding sites.[21,22] Notably, the human FOXP2 transcription factor forms a domain-swapped dimer, with known disease-associated mutations affecting residues of the domain-swapping interface.[23] Similarly to the AP2 domain of PF14_0633, the orientation of the DNA-binding region of the domain-swapped FOXP2 dimer would only allow the binding of two distal DNA-binding sites, thus looping out intervening DNA. The precedence of such a structural explanation of transcriptional regulation by FOXP2 thus lends credence to a similar mechanism for the function of PF14_0633.

Multiple PF14_0633 AP2-binding sites are found upstream of a set of sporozoite stage genes

Previous work by Young *et al.* identified a statistically enriched DNA motif (PfM24.1) located upstream of 30 genes expressed in the mosquito sporozoite stage of the *P. falciparum* life cycle.[24] This motif is highly similar to the consensus TGCATGCA motif recognized by PF14_0633,[6] suggesting that the sporozoite-expressed ApiAP2 protein PF14_0633[25] may be the transcription factor responsible for the transcriptional regulation of these genes. Our structural model implies that binding to multiple TGCATGCA binding sites may facilitate the function of PF14_0633 as a transcriptional regulator by inducing protein dimerization. To test this possibility, we used the PBM-derived position weight matrix for the TGCATGCA consensus sequence to scan the 2-kb upstream regions of these 30 sporozoite genes for secondary occurrences of the TGCATGCA sequence (see Materials and Methods). Of the 30 sporozoite genes, 27 contained two or more matches to the motif (Supplemental Table 2). Furthermore, 24 of these upstream regions contained pairs of sites that are spaced sufficiently distant (>100 bp apart) to permit dimer-induced DNA looping (Supplemental Table 2).

The occurrence of multiple DNA-binding sites upstream of this subset of well-established sporozoite genes suggests that dimer binding may be important for the function of PF14_0633 in transcriptional regulation. Co-occurrence of motifs was also reported for the ookinete-specific *Plasmodium*

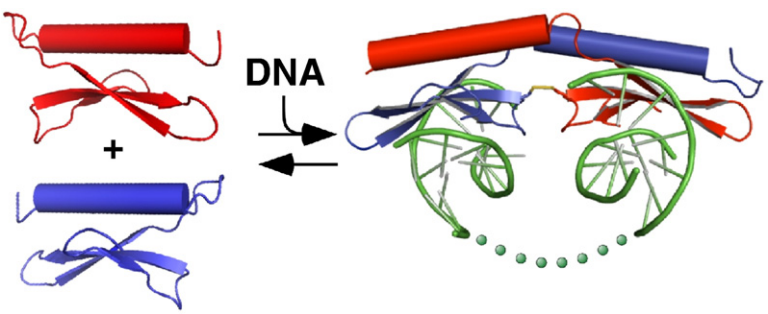

Fig. 4. A model of DNA-induced, domain-swapped dimerization and DNA looping. DNA binding stimulates formation and/or stabilization of the domain-swapped dimer, which, in turn, loops out intervening DNA between the two binding sites (illustrated as dots). Dimerization may allow portions of PF14_0633 not observed in the crystal structure (e.g., the ~700 C-terminal residues) to undergo functional rearrangement of protein interaction/transcriptional regulatory domains.

transcriptional regulator AP2-O.[7] One model is that the positioning of multiple binding sites in a promoter mediates a higher-affinity interaction of PF14_0633 with DNA, facilitated by dimer formation. It remains to be tested whether higher-order structures are essential for function *in vivo*. Structural rotations of the C-terminal α-helix, upon dimer formation, may expose anticipated additional domains of PF14_0633 that activate/repress transcription or modulate interaction with other auxiliary proteins required for regulation. Activation and repression domains that mediate transcriptional regulation have been defined only for a few examples of AP2/ERF domain proteins in the plant kingdom,[26] and to date, no such characterization exists for any ApiAP2 protein. However, the wide range of predicted sizes of ApiAP2 proteins (200–4000 amino acids in *Plasmodium* spp.) presumably speaks to the likely diversity of sequence space available to serve as necessary functional activation and repression domains.

We anticipate that the AP2 DNA-binding domain structure presented herein will serve as a model for assessing the binding interaction of other ApiAP2 proteins. PF14_0633 orthologues from two highly diverged Apicomplexans, *Cryptosporidium parvum* and *P. falciparum*, have conserved their DNA-binding preference for the TGCATGCA sequence,[6] the most over-represented DNA sequence in all sequenced Apicomplexa.[27] Notably, the four key DNA-binding residues identified in the PF14_0633 structure are conserved in the *C. parvum* orthologue *cgd2_3490*. This conservation suggests that insights gained on transcription factor domains in one apicomplexan species will be broadly applicable to understanding how related Apicomplexan AP2 domains contact DNA and regulate gene expression. Although it could be possible to model the structure of other AP2 domains upon that of PF14_0633, the prediction of DNA-binding specificity and which residues make base-specific DNA contacts would likely not be similarly predictable. Important differences that should be taken into consideration are that many AP2 domains do not contain conserved cysteine residues that are present at the dimerization interface; neither do they possess appropriately placed proline residues that might facilitate the domain swapping identified here. Additional structural and biochemical characterization of these domains will be needed to elucidate these functional details. In addition, Apicomplexan AP2 domains possess affinities for a much more diverse set of DNA sequences than in plants where an overwhelming majority of AP2/ERF proteins bind the canonical GCC box (A/GCCGCC).[9,11] This difference from plant AP2 domains demonstrates the need for structural and biochemical studies tailored towards Apicomplexan-derived proteins. Such studies are of great importance given that the ApiAP2 proteins are key regulators of parasitic developmental progression, and the absence of any homologous proteins in their mammalian hosts makes these proteins notable candidate drug targets.

Materials and Methods

Purification of the PF14_0633 AP2 domain and variants

The PF14_0633 AP2 domain (residues 63–123) or its variants were expressed as a GST-fusion protein as previously described.[6] Cells were lysed by sonication in GST lysis buffer (50 mM Tris, pH 8.0, 150 mM NaCl, and 10% glycerol) containing 1 mM PMSF and 1 mM benzamidine. Proteins were purified from the soluble fraction by binding to glutathione Sepharose 4B resin (GE Healthcare) and eluting with GST lysis buffer containing 20 mM reduced L-glutathione (Acros). The GST epitope tag was removed by thrombin cleavage. Proteins were dialyzed exhaustively to remove glutathione and were further purified by cation-exchange and size-exclusion chromatography, which yielded >95% pure samples in monomeric form. SeMet-incorporated PF14_0633 was expressed in *Escherichia coli* grown in M9 minimal media containing SeMet (Acros) and purified as above, except that 2 mM fresh dithiothreitol (DTT) was included in buffers. Protein concentrations were determined by measurement of the absorbance at 280 nm.

Crystallization and structure determination of DNA-bound PF14_0633 AP2 domain

PF14_0633 AP2 domain (15 mg/ml in GST lysis buffer) was mixed at a 1:1.2 molar ratio (protein:dsDNA) with a palindromic dsDNA (5′-TGCATGCA-3′, annealed in 10 mM Tris, pH 8.0, and 50 mM NaCl). Equivalent volumes of the complex and mother liquor (10 mM $CoCl_2$, 100 mM sodium acetate, pH 5.1, 0.5–1.25 M 1,6-hexanediol, and 10–20% v/v ethylene glycol) were mixed, and crystals of the complex formed in 1–3 days by hanging-drop vapor diffusion. Crystals were cryoprotected in mother liquor adjusted to 25% ethylene glycol prior to being flash-frozen in liquid nitrogen. Diffraction data were collected at the Advanced Photon Source at Argonne National Laboratory (Argonne, IL) at the LS-CAT 21-ID-D beamline (Table 1). Data were indexed and scaled with HKL2000,[28] and an experimental electron density map of the SeMet-incorporated AP2 domain of PF14_0633 bound to dsDNA was generated by single-wavelength anomalous dispersion phasing using SOLVE and solvent flattening using DM.[29] A partial model of the complex was generated and then used as a molecular replacement search model to determine the structure of the native complex using Phaser.[30] The native structure model was improved by rounds of manual building using Coot[31] and refinement with REFMAC.[32]

Electrophoretic mobility shift assay

Oligonucleotides containing a 28-bp fragment of the genomic sequence from *P. falciparum* located upstream of *pfi0540w* that contains the PF14_0633 AP2 consensus binding sequence (GCATGC) were synthesized by the University of Wisconsin Biotechnology Center to create a DNA substrate ("wt"). A mutant substrate with the central AT base pairs of the consensus binding site changed to GC (GCGCGC, "mut") was also produced. Both substrates were inserted into the StuI site of the pCR-Blunt vector (Invitrogen) to confirm their sequence identity and to produce 122-bp DNA substrates by restriction with HindIII and XhoI and polyacrylamide gel electrophoretic

purification (probe sequences listed in Supplemental Table 3). One picomole of each duplex DNA probe was end-labeled by T4 Polynucleotide Kinase (New England Biolabs) with γ^{32}P-ATP. Ten femtomoles of end-labeled probe was incubated with 0.1–50 μM concentrations of the PF14_0633 AP2 domain (or a variant) in 20 μl GST lysis buffer supplemented with 0.1 mg/ml bovine serum albumin and 5 μg/ml poly(dI–dC) (Sigma) for 30 min at room temperature. Samples were electrophoresed through a 12% polyacrylamide gel and exposed to a phosphor screen (Molecular Dynamics). Signals were visualized by a Storm 860 PhosphorImager, and band intensities were quantified by ImageQuant (Molecular Dynamics).

Protein binding microarrays

PBM experiments using the PF14_0633 AP2 variant domains were performed as described previously.[33] A minimum of two replicate experiments were performed for each variant.

Cysteine cross-linking

The PF14_0633 AP2 domain (or triple variant: Cys76Ala, Ala86Cys, Asn118Cys) was treated with 2 mM DTT to reduce disulfide bonds and subsequently dialyzed exhaustively against GST lysis buffer (pH 7.4) to remove the DTT. Individual proteins (60 μM) were incubated in the absence of dsDNA or in the presence of 75 μM cognate dsDNA (5′-TGCATGCA-3′) or non-cognate dsDNA (5′-TGTGCATAGTGGTGCGA-3′) for 30 min at 4 °C. Dimethylformamide (0.5 μl) or 20 mM $BM(PEG)_3$ (a 17.8-Å homobifunctional maleimide cross-linker, Thermo Scientific) dissolved in dimethylformamide was added to samples for 30 min at 4 °C. In these conditions, thiol-specific cross-linking proceeds 1000-fold more quickly than nonspecific reaction with primary amines. Samples were electrophoresed on reducing or nonreducing 20% SDS-PAGE gels and visualized by Coomassie Blue. Band intensities were analyzed by ImageQuant.

Calibrated size-exclusion chromatography

PF14_0633 AP2 domain (400 μM) was incubated with 500 μM cognate dsDNA, with non-cognate dsDNA (sequences described above), or without DNA for 1–2 h in GST lysis buffer at 4 °C. Samples were applied to a GE Healthcare S-100 HR column, and elution volumes were determined by detection of UV absorbance peaks. Average molecular weights of each peak were determined by comparing the elution volumes to calibrated standard markers (Bio-Rad Gel Filtration Standard) run just prior to and following experimental runs. Peak fractions were also analyzed by reducing and nonreducing SDS-PAGE as above.

Bioinformatic identification of multiple binding sites with transcriptional upstream regions of *P. falciparum*

ScanACE[34] and the PBM-derived PF14_0633 wild-type position weight matrix were used to scan the 2-kb upstream regions of the *P. falciparum* sporozoite-expressed genes (listed in Ref. 24). The 2-kb upstream regions of the 30 sporozoite genes were extracted from the *Plasmodium* genome resource PlasmoDB.org v5.5.[35] Scoring cutoffs were set to capture two standard deviations below the mean score in the alignment file.

Accession numbers

The X-ray crystallographic structure and structure factors have been deposited in the Protein Data Bank and have been assigned the identifier 3IGM.

Acknowledgements

We thank Spencer Anderson of LS-CAT for beamline assistance, Ken Satyshur for assistance with crystallographic data collection and model building, and Darrell R. McCaslin of the University of Wisconsin-Madison Biophysics Instrumentation Facility, which is supported by the University of Wisconsin-Madison, National Science Foundation grant BIR-9512577 and National Institutes of Health (NIH) grant S10 RR13790, for assistance with circular dichroism spectroscopy. This investigation was supported by the NIH under a Ruth L. Kirschstein National Research Service Award (1F32GM083438) to S.E.L.; by a Shaw Foundation award to J.L.K.; and by NIH grant R01 AI076276, the Arnold and Mabel Beckman Foundation, and, in part, by NIH grant P50 GM071508 to M.L. S.E.L. and E.D.S. performed research; S.E.L., E.D.S., J.L.K., and M.L. designed research, analyzed data, and wrote the paper.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.11.004

References

1. Horrocks, P., Wong, E., Russell, K. & Emes, R. D. (2009). Control of gene expression in *Plasmodium falciparum*—ten years on. *Mol. Biochem. Parasitol.* **164**, 9–25.
2. Balaji, S., Babu, M. M., Iyer, L. M. & Aravind, L. (2005). Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res.* **33**, 3994–4006.
3. Riechmann, J. L. & Meyerowitz, E. M. (1998). The AP2/EREBP family of plant transcription factors. *Biol. Chem.* **379**, 633–646.
4. Krizek, B. A. (2003). AINTEGUMENTA utilizes a mode of DNA recognition distinct from that used by proteins containing a single AP2 domain. *Nucleic Acids Res.* **31**, 1859–1868.
5. Allen, M. D., Yamasaki, K., Ohme-Takagi, M., Tateno, M. & Suzuki, M. (1998). A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J.* **17**, 5484–5496.

6. De Silva, E. K., Gehrke, A. R., Olszewski, K., Leon, I., Chahal, J. S., Bulyk, M. L. & Llinas, M. (2008). Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc. Natl Acad. Sci. USA*, **105**, 8393–8398.
7. Yuda, M., Iwanaga, S., Shigenobu, S., Mair, G. R., Janse, C. J., Waters, A. P. *et al.* (2009). Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol. Microbiol.* **71**, 1402–1414.
8. Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R. *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288.
9. Sakuma, Y., Liu, Q., Dubouzet, J. G., Abe, H., Shinozaki, K. & Yamaguchi-Shinozaki, K. (2002). DNA-binding specificity of the ERF/AP2 domain of *Arabidopsis* DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochem. Biophys. Res. Commun.* **290**, 998–1009.
10. Shigyo, M., Hasebe, M. & Ito, M. (2006). Molecular evolution of the AP2 subfamily. *Gene*, **366**, 256–265.
11. Llinas, M., Deitsch, K. W. & Voss, T. S. (2008). *Plasmodium* gene regulation: far more to factor in. *Trends Parasitol.* **24**, 551–556.
12. Amoutzias, G. D., Robertson, D. L., Van de Peer, Y. & Oliver, S. G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. *Trends Biochem. Sci.* **33**, 220–229.
13. Buttner, M. & Singh, K. B. (1997). *Arabidopsis thaliana* ethylene-responsive element binding protein (AtEBP), an ethylene-inducible, GCC box DNA-binding protein interacts with an ocs element binding protein. *Proc. Natl Acad. Sci. USA*, **94**, 5961–5966.
14. Chandler, J. W., Cole, M., Flier, A., Grewe, B. & Werr, W. (2007). The AP2 transcription factors DORNROSCHEN and DORNROSCHEN-LIKE redundantly control *Arabidopsis* embryo patterning via interaction with PHAVOLUTA. *Development*, **134**, 1653–1662.
15. Aravind, L. & Landsman, D. (1998). AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic Acids Res.* **26**, 4413–4421.
16. Moure, C. M., Gimble, F. S. & Quiocho, F. A. (2002). Crystal structure of the intein homing endonuclease PI-SceI bound to its recognition sequence. *Nat. Struct. Biol.* **9**, 764–770.
17. Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285–1299.
18. Newcomer, M. E. (2002). Protein folding and three-dimensional domain swapping: a strained relationship? *Curr. Opin. Struct. Biol.* **12**, 48–53.
19. Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., III & Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* **24**, 1429–1435.
20. Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M. *et al.* (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339.
21. Lamber, E. P., Vanhille, L., Textor, L. C., Kachalova, G. S., Sieweke, M. H. & Wilmanns, M. (2008). Regulation of the transcription factor Ets-1 by DNA-mediated homo-dimerization. *EMBO J.* **27**, 2006–2017.
22. Remenyi, A., Tomilin, A., Pohl, E., Lins, K., Philippsen, A., Reinbold, R. *et al.* (2001). Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. *Mol. Cell*, **8**, 569–580.
23. Stroud, J. C., Wu, Y., Bates, D. L., Han, A., Nowick, K., Paabo, S. *et al.* (2006). Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14**, 159–166.
24. Young, J. A., Johnson, J. R., Benner, C., Yan, S. F., Chen, K., Le Roch, K. G. *et al.* (2008). In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics*, **9**, 70.
25. Le Roch, K. G., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D. *et al.* (2003). Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, **301**, 1503–1508.
26. Mitsuda, N. & Ohme-Takagi, M. (2009). Functional analysis of transcription factors in *Arabidopsis*. *Plant Cell Physiol.* **50**, 1232–1248.
27. Ling, K. H., Rajandream, M. A., Rivailler, P., Ivens, A., Yap, S. J., Madeira, A. M. *et al.* (2007). Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.* **17**, 311–319.
28. Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326.
29. Cowtan, K. (1994). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallograph*, vol. 31, pp. 34–38.
30. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674.
31. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **60**, 2126–2132.
32. Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **57**, 122–133.
33. Berger, M. F. & Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat. Protoc.* **4**, 393–411.
34. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945.
35. Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B. *et al.* (2009). PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–D543.