Crystal structure of *Thermotoga maritima* TM0439: implications for the mechanism of bacterial GntR transcription regulators with Zn^{2+}-binding FCD domains

Meiying Zheng[1,§], David R. Cooper[1, §], Nickolas E. Grossoehme[2], Minmin Yu[3], Li-Wei Hung[3,4] , Marcin Cieslik[1], Urszula Derewenda[1], Scott A. Lesley[5,6], Ian A. Wilson[5], David P. Giedroc[2], Zygmunt S. Derewenda[1,*]

[1]Integrated Center for Structure-Function Innovation, Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA 22908-0736; [2]Department of Chemistry, Indiana University, Bloomington, Indiana 47405-7102; [3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, MS4R0230, Berkeley, CA 94720; Physics Division, MS D454, Los Alamos National Laboratory, Los Alamos, NM 87545; [5]The Scripps Research Institute, North Torrey Pines Road, La Jolla, CA 92037; [6]Genomics Institute of the Novartis Research Foundation, 10675 John Jay Hopkins Dr., San Diego CA, 92121

§ The contribution of these authors was equally important

Correspondence e-mail: zsd4n@virginia.edu

PDB reference: 3FMS

The GntR superfamily of dimeric transcription factors, with more than 6200 members encoded in bacterial genomes, are characterized by N-terminal winged helix (WH) DNA-binding domains and diverse C-terminal, regulatory domains, which provide a basis for the classification of the constituent families. The largest of these families, FadR, contains nearly 3000 proteins with all α-helical regulatory domains classified into two related Pfam families: FadR_C and FCD. Only two crystal structures of the FadR family members, i.e. the *E. coli* FadR protein and the LldR from *C. glutamicum,* have been described to date in literature. Here we describe the crystal structure of TM0439, a GntR regulator with an FCD domain, found in the *Thermotoga maritima* genome. The FCD domain is similar to that of the LldR regulator, and contains a buried metal binding site. Using atomic absorption spectroscopy and Trp fluorescence, we show that the recombinant protein contains bound Ni^{2+} ions, but it is able to bind Zn^{2+} with K_D <70 nM . We conclude that Zn^{2+} is the likely physiological metal, where it may perform either or both structural and regulatory roles. Finally, we compare the TM0439 structure to two other FadR family structures recently deposited by Structural Genomics consortia. The results call for a revision in the classification of the FadR family of transcription factors.

Keywords: Protein structure; transcription regulation; GntR family; protein crystallization; structural genomics; surface entropy reduction

1. Introduction

Transcription regulators play a critical role in the biology of microorganisms (Huffman & Brennan, 2002). They repress, de-repress, and activate gene transcription through tightly regulated, direct interactions with cognate DNA sequences, mediated by a variety of unique domains or motifs, such as helix-turn-helix domains, zinc-fingers, homeodomains, leucine zippers and β-sheet DNA-binding proteins. Within the helix-turn-helix (HTH) regulators, numerous superfamilies have been identified based on sequence similarities in the DNA-binding module. The GntR superfamily, Pfam PF00392 (Bateman *et al.*, 2002), first described in 1991 and named after the gluconate operon repressor in *B. subtilis* (Haydon & Guest, 1991), currently comprises over 6200 proteins found in diverse eubacterial genomes. The DNA-binding domains in this family share a significant level of similarity and all exhibit the winged helix-turn-helix (WH) topology with the canonical HTH motif followed by a β-hairpin. By contrast, the C-terminal regulatory ligand-binding domains vary significantly among individual proteins, providing a basis for the current classification of major families, i.e. HutC, MocR, YtrA, AraR, PlmA and—the largest family comprising ~40% of all GntRs—FadR. (Rigali *et al.*, 2002; Lee *et al.*, 2003; Franco *et al.*, 2006). By far the best characterized GntR regulator is the *fadR* gene product, the founding member of the FadR family. It functions as a repressor of the *fad* regulon which includes genes responsible for transport, activation and β-oxidation of long and medium-length fatty acids (DiRusso *et al.*, 1992; DiRusso *et al.*, 1993). The crystal structure of the apo-repressor, as well as structures of complexes with the dsDNA operon oligonucleotide, and with an effector, myristoyl-CoA, have been determined (van Aalten *et al.*, 2001; van Aalten *et al.*, 2000; Xu *et al.*, 2001). These studies revealed the mechanism by which the effector-induced conformation changes in the regulatory domain are transmitted to the WH domain, and consequently disrupt the repressor-operon interaction, thereby relieving repression (van Aalten *et al.*, 2001).

All known FadR family transcription regulators are predicted to contain all α-helical, C-terminal domains, with either seven or six α-helices. An accurate alignment has been elusive because of low levels of amino acid similarities. However, the predicted number of helices serves as a basis for one classification scheme into the FadR (seven helices) and VanR (six helices) groups (Rigali *et al.*, 2002). Both groups appear to be involved at the crossroads of metabolic pathways, e.g. galactonate (DgoR), gluconate (GntR), vannilate (VanR), malonate (MalR), etc. An alternative classification of regulatory domains of FadR members is offered by the Pfam database (Bateman *et al.*, 2002). The smaller FadR_C family (Pfam07840), represented by the C-terminal domain from FadR itself, comprises only ~70 members exhibiting high amino acid similarity. All proteins in this family have C-terminal domains of the FadR group, i.e. with seven helices. Interestingly, in the vast majority of cases there is one gene of this type per bacterial genome. The larger and more diverse FCD family (Pfam007729) has over 2800 known members in more than 400 species. It includes domains with both six and seven predicted α-helices, i.e. members of both FadR and VanR groups.

Recently, atomic coordinates for three new structures of putative FadR-like transcription regulators were deposited in the PDB. Two of these were reported by Structural Genomics groups, without accompanying publications: RO03477 from *Rhodococcus sp* RHA1 (entry 2hs5) and PS5454 protein from *Pseudomonas syringae pv. tomato* str. DC3000 (entry 3c7j). Both structures contain C-terminal domains with six α-helices, making them VanR group members. The third structure, that of CGL2915 protein from *Corynebacterium glutamicum* (2di3) is a FadR group member as judged by the seven helices in its C-terminal domain (Gao *et al.*, 2008). However, in spite of the size difference, all three proteins are annotated in the Pfam database as containing FCD domains.

In this paper, we describe the structure of the TM0439, a putative transcriptional regulator from *Thermotoga maritima*. Based on amino acid sequence, its regulatory domain was also annotated as an FCD family member. We have compared the structure of TM0439 to FadR and the three newly deposited related transcriptional regulators and herein we show that, together with CGL2915 and PS5454, TM0439 is a member of a distinct, yet previously unrecognized group of metal-binding transcription regulators in which a distinct variant of the FCD domain contains a metal binding site. This domain is identified by a conserved fingerprint sequence motif: Arg-X_3-Glu-X_{40}-Asx-X_4-His-$X_{\sim 50}$-His-$X_{\sim 20}$-His. Although the metal in the TM0439 crystal structure is Ni^{2+}, we determined experimentally that the protein can bind both Ni^{2+} and Zn^{2+}, with K_D values in the nM (or lower) range, making Zn^{2+} a more probable biological ligand. Our study sets the stage for an improved annotation of the FadR family of transcription regulators, and offers a structural rationale for the strict conservation of a unique sequence motif in a subset of these proteins.

2. Materials and Methods

2.1. Protein Expression and Purification

The TM0439 gene has been cloned as a part of the structural genomics project of the *T. maritima* proteome (Lesley *et al.*, 2002). Like in other JCSG (Joint Center for Structural Genomics) expression vectors, there is a non-cleavable, N-terminal tag (MGSDKIHHHHHH) as well as both arabinose and T7 promoters. The wild-type protein, expressed and purified using routine methods, did not crystallize. To circumvent this problem, three mutants with reduced surface entropy, E118A,K119A,K122A (variant 1A), K2A,K3A (variant 2A) and E30A,K31A (variant 3A) were designed using the Surface Entropy Reduction prediction (SERp) server (http://nihserver.mbi.ucla.edu/SER/) and created using the Quikchange$^{(TM)}$ protocol (Stratagene, Inc.). Expression was carried out in *E.coli* BL21 strain in M9 media with added SeMet for labeling. The protein was purified using nickel affinity chromatography (Ni-NTA agarose column, Qiagen). Pure fractions were pooled together and dialyzed overnight against a buffer consisting of 20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 2.5 mM β-mercaptoethonal (β-ME). Protein samples were concentrated to 15 mg/ml and stored at -80° C.

2.2. Crystallization and Data Collection

The mutant proteins were screened using a standard JCSG+ screen from Qiagen, Inc., using reservoirs containing either the screen solution or 1.5 M NaCl (Newman, 2005). The triple mutant 1A yielded diffraction quality crystals directly from the screen, i.e. 0.1 M acetate buffer, pH 4.5, 35% v/v MPD. The crystals displayed C2 symmetry, a=85.09 Å, b=72.72 Å, c=43.32 Å, β = 104.6°. A MAD data set was collected at beam line 8.2.1 at ALS equipped with an ADSC Q315R detector. All data were processed using HKL2000 (Otwinowski & Minor, 1997) with data statistics shown in Table 1.

2.3. Structure Solution and Refinement

The asymmetric unit contains one protein molecule, corresponding to solvent content of 58.0%. Using MAD data, 3 selenium sites were located and phase calculations were carried out using SOLVE/RESOLVE (Terwilliger, 2003). Approximately two-thirds of the structure was built automatically. Model building and refinement of the SeMet structure were carried out using the data set collected at the remote high-energy wavelength, truncated at 2.2 Å to ensure completeness in the high resolution shell (Table 1). Iterative refinement and model building were performed using RESOLVE and REFMAC5 (Murshudov *et al.*, 1997). This process dramatically improved the maps and the missing fragments were identified in intermediate models. A combination of “cut and paste” model building and manual refinement resulted in a complete structure. This iterative process allowed the refinement, which had previously stalled with an R_{free} around 0.32, to converge with a crystallographic R and R_{free} values of 0.17 and 0.23, respectively. The final model was refined with PHENIX. (Zwart *et al.*, 2008). using the TLS (translation/libration/screw) approximation of thermal motion (Winn *et al.*, 2001). The validation of the model was carried out using MOLPROBITY(Lovell *et al.*, 2003). The corresponding refinement statistics are shown in Table 1. Figures were prepared with Pymol (http://pymol.sourceforge.net/). The analysis of dimer interface was done using PISA v1.15 (Krissinel & Henrick, 2007). Cavity volumes were calculated using VOIDOO (Kleywegt & Jones, 1994). For CGL2915 our cavity volume calculation yields results different from those reported in literature (Gao *et al.*, 2008).

2.4. Metal analysis

Stock metal concentrations and metal content of TM0439 were determined using a Perkin Elmer AAnalyst 400 atomic absorption spectrometer (AAS) with standard curves generated from NIST standards from Alfa Aesar (Ward Hill, MA). Initial metal content data were verified by ICP-OES (inductively coupled plasma – optical emission spectroscopy) at Dartmouth College Elemental Analysis Lab (Hanover, NH). Complete removal of metal was accomplished by several rounds of extensive dialysis with 10 mM EDTA (ethylenediamine tetracetic acid) and 2 mM DTT (dithiotheitol) in 25 mM Tris and 100 mM NaCl at pH 8.0 and 4ºC and was verified by AAS. Removal of DTT and EDTA was accomplished by four rounds of dialysis under an inert Ar atmosphere with thoroughly degassed buffer (25 mM Tris and 100 mM NaCl at pH 8.0). Zn^{2+} and Ni^{2+} binding assays were done by monitoring tryptophan fluorescence (λ_{ex} 287) on an ISS PC1 spectrofluorimeter under strictly anaerobic conditions. The concentration of TM0439 was 5.3 μM (25 mM Tris and 100 mM NaCl at pH 8.0 and 25 ºC). The data were fit to appropriate chemical models (2:1 and 1:1, respectively) using DynaFit (Kuzmic, 1996) with metal-buffer interactions (log K_{ZnTris} = 2.27; log K_{NiTris} = 2.67; log $\beta_{2,Ni(Tris)2}$ = 4.6) (NIST Standard Reference Database 46, 2003) included in the model.

3. Results and Discussion

3.1. Design of crystallizable mutant

TM0439 was originally selected as one of the targets for a high-throughput pipeline at the Joint Center for Structural Genomics (Lesley *et al.*, 2002). However, the wild-type protein did not yield X-ray quality crystals. To overcome this problem, we used surface entropy reduction (Derewenda, 2004) to generate variants of the protein with enhanced crystallizability. We used the SERp server (Goldschmidt *et al.*, 2007) to predict suitable mutations to generate surface patches with reduced conformational entropy and enhanced ability to mediate crystal contacts and generate X-ray quality crystals (Derewenda & Vekilov, 2006; Derewenda, 2004). Three mutants were suggested by the server; in the order of ranking they were: a triple mutant E118A, K119A, K122A, a double mutant K2A, K3A, and another double mutant E30A, K31A. All three were expressed and screened for crystallization as described in Methods. The triple mutant gave crystals with excellent morphology and diffraction properties directly from the crystallization screen, and this crystal form was used in the subsequent analysis.

3.2. Overview of the structure and comparison to other FadR family members

The crystal structure of TM0439 was determined by MAD (multiwavelength anomalous dispersion) using a SeMet-labeled protein. The atomic model was refined to 2.2 Å resolution (Table 1; see Methods). The protein has a canonical domain architecture of the GntR family, with an N-terminal WH-domain and a C-terminal, all α-helical putative regulatory domain. The presence of only 6 α-helices within the C-terminal domain classifies TM0439 as a VanR member. Gel filtration experiments (not shown) indicated that the protein is an obligate dimer in solution. The C2 space group symmetry allows for a head-to-head dimer *via* the crystallographic two-fold axis, so that a large interface is buried between two C-terminal regulatory domains, with a resulting quaternary structure very close to that of FadR (van Aalten *et al.*, 2000). By contrast, the two WH domains do not interact with one another, although they make limited crystal contacts with neighboring molecules in the unit cell. A comparison of TM0439 with FadR, and with the recently deposited structures CGL2915, RO03477 and PS5454, shows dramatic differences in local tertiary and quaternary architectures, even though the individual domains are remarkably similar (Fig. 1).

As pointed out above, TM0439, RO03477 and PS5454, can be classified in the VanR group, based on secondary structure prediction which identifies only six α-helices in their C-terminal domains (Rigali *et al.*, 2002). In all three structures, a short linker connects the second β-strand of the WH domain directly to α_1-helix of the regulatory domain, so that the α_0-helix seen in FadR is absent. In the TM0439 and RO03477 structures, the mutual disposition of the WH and regulatory domains is similar, with the two WH domains in close proximity; in contrast, the structure of PS5454 is distinctly different, with the two WH domains at opposite ends of the homodimer. The two FadR group proteins (i.e. FadR and CGL2915) contain an extra α_0-helix at the N-terminus of the regulatory domain. In FadR, this helix contains a sharp kink which reverses its course in the center, wedging between the WH and the regulatory domains. Consequently, the mutual disposition of the two domains of FadR is distinctly different from both TM0439 and RO03477, due to a rotation of the regulatory domain relative to the WH domain**.** In CGL2915, the α_0-helix is straight, and as a consequence, the two regulatory domains are swapped between the monomers (Gao *et al.*, 2008).

The site of the three mutations made to enhance crystallizability is located in the loop between helices α_2 and α_3 of the C-terminal domain, and is involved in a heterologous contact with a WH domain of a symmetry related molecule. The site of the mutations is distant from functionally important structural elements.

3.3. The WH domain

The N-terminal portion of TM0439 (residues Val6-Val71) constitutes the winged-helix, dsDNA binding domain, with a canonical order of secondary structure elements α1, α2, α3, β1, β2 (n.b., we refer to them henceforth as a1, a2, a3, b1, b2, to differentiate from the helices α_0 - α_6 in the regulatory domain). The HTH (helix-turn-helix) motif is made up of helices a2 and a3 with the connecting loop, and the antiparallel, two-stranded β-sheet makes up the 'wing'. Helix a1 provides a critical interface with the C-terminal regulatory domain in the same monomer. The WH domain is a hallmark of the GntR family. Not surprisingly, a structural comparison using DALI (Holm *et al.*, 2006) identified a number of known WH domains with similar structure. The top hits, with Z>8.0, include all of the known putative GntR structures, but also the Zα domain of the viral E3L protein (1sfu), double-stranded RNA specific adenosine deaminase (1qbj), catabolite gene activator proteins (CAP) (1i6f), and LEXA repressor (1jhf). The pairwise r.m.s.d. values for the Cα atoms are around 2.0 Å. The highest amino acid sequence identity among proteins of known structure is observed for 3c7j (PS5454) and 2di3 (CGL2915), at 35 % and 32 % respectively.

Although all known structures of WH domains are very similar, their mode of interaction with dsDNA can vary considerably. While most of them use the second helix of the HTH motif to bind to the major grove of cognate DNA sequence (Gajiwala & Burley, 2000), the FadR WH domain uses only the N-terminal fragment of that helix (Xu *et al.*, 2001). Interestingly, residues Arg35, Arg45, Arg49 and Gly66, which are indispensable for DNA binding in FadR are completely conserved in CGL2915. These observations suggest that CGL2915 may bind to DNA in a manner similar to FadR which binds to the TGGTN$_3$ACCA (Xu *et al.*, 2001). In fact, an identical sequence was identified in the *C. glutamicum* genome, in the promoter of *cgl2917* (Gao *et al.*, 2008). However, in TM0439 the residue equivalent to Arg45 of FadR is Phe45, suggesting that the target DNA sequence for this protein is different. Both RO03477 and PS5454 also show differences from the putative dsDNA binding consensus sequence (Fig. 2).

3.4. The regulatory FCD domain

The FCD domain of TM0439, encompassing residues Glu76 – Glu212, contains six α-helices, as predicted for the VanR group, arranged into an antiparallel bundle. The same tertiary fold is observed in the regulatory domains of RO03477 (2hs5) and PS5454 (3c7j), both VanR group members. The C-terminal domains of CGL2915 (2di3) and FadR (1hw1) also show a very similar fold, with the sole exception of the additional α_0-helix characteristic of the FadR group (Fig. 3). Pairwise r.m.s. differences between Cα positions range from 2.2 Å to 2.9 Å. This structural similarity is particularly striking, given the limited amino acid sequence similarities: 18% between TM0439 and RO03477, 13% against PS5454, 17% for CGL2915 and only 11% for FadR. The FadR C-terminal domain is classified as a member of the FadR_C family (PF07840), while the remaining four are in the FCD family (PFam 07729). Thus, the FadR and VanR groups are not equivalent to the

FadR_C and FCD families, respectively, creating confusing classification. We suggest that the FadR and VanR distinction should be discontinued.

Although a fold comprising a six-helix, antiparallel bundle is topologically simple, the FCD/FadR_C fold constitutes a unique family to the extent that DALI (Holm *et al.*, 2006) shows no other structurally related domains with a Z score higher than 6. It seems that the distinction between the FadR_C and FCD families made in the Pfam database is insignificant, and a single family, e.g. FCD, should comprise all these proteins, and in the following discussion, the term FCD shall refer to all members of the FCD/FadR_C fold.

An interesting structural feature of the FCD fold is a conserved kink in the α_4 helix**.** This helix is noteworthy because its N-terminal part is intimately involved in the dimerization of the domain (see below), while the C-terminal portion constitutes the main interface with the WH domain of the same monomer. In TM0439, the α_4 helix has six full turns, and the kink occurs approximately after the first three. The kink results in a strained secondary conformation of Ile153 (φ=-107°, ψ=11°) which leaves the amides of Asp155 and Arg156, as well as the carbonyl of Lys164, free from intra-helical H-bonds. Instead, the side chain Glu58 from the WH domain positions itself so that Oε1 'caps' both chain amides of both Asp155 and Arg156 (Fig. 3). An almost identical structural perturbation occurs in the corresponding α-helix in CGL2915, in which the kink at Leu167 (φ =-86° and ψ =-12°) leaves the amides of Leu169 and Ser170, as well as the carbonyl of Ala166 free; here, Ser81 from the WH domain performs the capping function (Fig. 3). A similar stereochemistry is reproduced in FadR, where Met168 is at the center of the kink (φ =-78°, ψ =-23°) leaving the amides of Gly170 and Leu 171, and carbonyl of Gly167 uncapped, but with no substitute H-bonding partners from the WH domain (Fig. 3). In RO03477 a similar kink occurs after the first two turns, not three as in the previous structures. Met168 is at its center (φ =-84° and ψ =-8°) and the free amides of Ser170 and

Val171, as well as the carbonyl of Val167 are not involved in any H-bonds (Fig 3). The PS5454 structure is the only one in which the α_4 helix is straight. It is also the only structure in which the WH domains are set apart. We will return to this point later.

3.5. FCD domain as a dimerization module

The FCD domains are responsible for the dimeric architecture of the FadR transcription factors. The crystal structures of FadR and CGL2915 show an almost identical disposition of the FCD domains in the homodimers and suggest that the mode of dimerization is conserved (Gao *et al.*, 2008). The TM0439 protein conforms to this paradigm. It forms a homodimer in which the interface is mediated exclusively by the α_1-helix and the N-terminal portion of the α_4-helix of the FCD domain. In each chain, 23 residues bury a surface of ~950 Å^2. The hydrophobic core of the interface is formed by Ile87, Met88, Met89, Phe92, Leu145, Leu146, Leu149, and Ile153. Residues that bury the largest solvent exposed surface are Glu81, Glu84, Met88, Phe92, Asn143, Leu145, Leu149 and Lys152. A total of 14 H-bonds and four salt bridges span the interface at its periphery (Fig. 4). Both the RO03477 and PS5454 structures have topologically very similar interfaces, mediated by the α_1- and α_4- helices, albeit the buried solvent accessible surfaces are smaller than in TM0439 (~780 Å^2 and ~730 Å^2, respectively)**.** The same overall architecture is also seen in FadR and the CGL2915, but their FCD domains contain the additional α_0-helix, which contributes significantly to the dimer contact. In FadR, the surface buried on dimerization is ~780 Å^2 per monomer, of which 112 Å^2 is contributed by Leu80, Ile82 and Leu 83 from the α_0-helix. In CGL2915, these buried surfaces are ~950 Å^2, and ~145 Å^2, respectively; the latter surface is contributed by Ala79, Leu80, Ser83, Val84 and Gln87.

Thus, the mode of dimerization of all FCD domains is highly conserved, notably in the absence of any significant amino acid sequence similarities between the individual proteins. The unique nature of each interface suggests that heterodimerization is not possible within this family.

3.6. A novel metal-binding subfamily of FCD

Based on the FadR paradigm, it is thought that the regulatory domains of the FadR family bind small organic ligands and, as a consequence, undergo conformational changes that reorient the WH domains and affect their binding to cognate DNA. We were, therefore, interested if the structure of TM0439 might reveal a putative binding site for such a ligand. Indeed, we find an internal polar cavity in the FCD domain, at the bottom of which are three histidines (His134, His174 and His196) with imidazole groups arranged in a three-blade propeller, with the Nε2 atoms pointing towards a strong peak of positive electron density. When a dummy atom was placed in this density and refined, it was found to be 2.0-2.2 Å from the three Nε2 atoms, consistent with the coordination stereochemistry of a metal ion.

Histidines coordinate metal ions primarily *via* the Nε2 atoms (Chakrabarti, 1990b), even though in solution they are preferentially protonated on these atoms (Reynolds *et al.*, 1973). Thus, histidines within metal binding sites typically donate hydrogen bonds through their Nδ1 atoms to carboxyl side chains or other H-bond acceptors (e.g. main chain carbonyls), to stabilize the less favorable tautomeric form that is unprotonated on Nε2 (Argos *et al.*, 1978; Christianson & Alexander, 1989). In concert with this paradigm, two of the metal binding histidines, i.e. His134 and His196, are stabilized in this form by H-bonds to neighboring carboxylic acids (Oε1 of Glu173 acts as acceptor for Nδ1 of His196, and Oε1 of Glu90 for Nδ1 His134). In addition, His134 donates a Cδ2(H)…O bond to the main-chain carbonyl of Asp130 (3.1 Å) (Fig. 5). Similar CH…O bonds involving the

Cε1(H) group, which is modestly acidic, are commonly observed for histidines in proteins (Derewenda *et al.*, 1994), but those involving Cδ2(H) are rare.

The three imidazols form a triangular propeller, with the angles at each Nε2 close to 60°. Further, the putative metal ion is elevated ~1.25 Å above the plane defined by the Nε2 atoms, as expected for tetrahedral coordination. The putative fourth position in the coordination sphere is unoccupied, and above it we find electron density consistent with a carbonate or an acetate ion, which may have originated from the crystallization mix. The refined B-value for the metal (36 $Å^2$) was consistent with a divalent ion, such as Zn^{2+} or Ni^{2+}. To identify the metal, we employed atomic absorption spectroscopy on the SeMet samples used for crystallization and found stoichiometric amounts of Ni^{2+}. Metal removal was found to be kinetically impaired, as it required greater than 48 hours of dialysis against 10 mM EDTA and 2 mM DTT to be completely removed at 4ºC. This slow removal may be a consequence of the inherently slow Ni^{2+} ligand exchange kinetics as well as the relatively buried nature of the metal binding site. We suspect that Ni^{2+} may have been inadvertently introduced during the purification protocol, i.e. Ni^{2+}-affinity chromatography, and that Zn^{2+} is the physiological ligand, consistent with the tetrahedral coordination geometry, as well as the presence of histidines as coordinating residues, all favoring Zn^{2+} (Dokmanic *et al.*, 2008).

Using tryptophan fluorescence, we measured the metal affinity of TM0439 for both Zn^{2+} and Ni^{2+}. Fig. 6A shows the fluorescent emission spectrum upon excitation at 287 nm, with a characteristic tryptophan peak at λ_{em} = 340 nM. We find that Ni^{2+} binding is stoichiometric, 1:1, with $K = 1.47 \pm 0.01 \times 10^7$ M^{-1} ($K_d = 68 \pm 5$ nM). Unexpectedly, Zn^{2+} binds with a stoichiometry of 2:1 with sequential binding constants of $K_1 \geq 1.4 \pm 0.1 \times 10^7$ M^{-1} ($K_d \leq 71 \pm 5$ nM) and $K_2 \geq 4.5 \pm 0.4 \times 10^5$ M^{-1} ($K_d \leq 2.0 \pm 0.2$ μM), respectively, with an approximately two-fold large increase in the Trp fluorescence (Fig. 6B). The origin of the second binding site is unknown, and it is not clear if the lower affinity site is of functional significance. We note that the protein contains a His_6-tag which, in principle, could influence the apparent metal binding affinities and stoichiometries. However, the N-terminal localization of the polyhistidine sequence virtually rules out any potential influence on the Trp154 quantun yield, which is located at the kink in the α4 helix of the C-terminal regulatory domain. Both Zn^{2+} and Ni^{2+} bind to synthetic histidine-rich sequences with affinities of ~10^4 (Whitehead *et al.*, 1997). Since the measured Zn^{2+} binding constants are lower limits (see legend, Fig. 6B), it is unlikely that there is significant competition by the polyhistidine tail. Since we do not observe a secondary, low affinity Ni^{2+} binding site, it may be possible that it is masked by competition from the His_6-tail. Taken together and considering the relative abundance of Zn^{2+} as compared to Ni^{2+} for most organisms (Outten & O'Halloran, 2001), it is reasonable to hypothesize that TM0439 is a Zn^{2+} binding protein, although our analysis did not include other transition metals, e.g. Co or Mn, which in principle might also be involved.

Interestingly, the structures of both CGL2915 (2di3) and PS5454 (3c7j) also contain metals bound in stereochemically analogous sites. In CGL2915, the coordinating histidines are His148, His196, His218, and their imidazoles are stabilized in the Nδ1-protonated tautomers by Glu106, Gln193, and Glu195, respectively. His148 is additionally stabilized by a CH…O bond *via* its Cδ1, as is the case with His134 of TM0439. However, another protein atom, Oδ1 of Asp144 (analogous to Asp130 in TM0439), serves as an axial ligand (distal to His218), resulting in slightly distorted trigonal bipyramid coordination, with a water molecule completing the equatorial plane (Fig. 5**)**. The same stereochemistry is preserved in the second, crystallographically independent, subunit. It is also interesting to note that the Oδ1 Asp144 approaches the putative metal with the *syn* sp2 orbital, as is usual in metal binding sites (Chakrabarti, 1994, 1990a). The ligand in CGL2915 is annotated as Zn^{2+} based on XAFS data (Gao *et al.*, 2008).

In the *P. syringae* regulator (3c7j), the coordinating histidines are His148, His192, and His214, while the fourth ligand, equivalent to Asp144 in CGL2915, is Asn144. The His214 and Asn144 side chains serve as axial ligands, and the latter is oriented with its side-chain oxygen towards the metal. His192 and His214 are stabilized in the required tautomeric forms by Nδ1 H-bonds to Asp191 and Gln189, respectively. The His148 residue has the same interesting CH…O bond to the carbonyl of Asn144 as its counterparts in CGL2915 and TM0439. In one subunit, a single water molecule is found in an equatorial plane while, in the second independent monomer, two water molecules complete an octahedral coordination sphere (Fig. 5). The metal in this structure is annotated as Ni^{2+}, consistent with the coordination preference and with reasonable B-values.

Neither the FadR nor the RO03477 structures have metal binding sites. In FadR, the three metal-coordinating histidines are replaced by Phe149, Tyr193 and Tyr215. In RO03477, one of the three histidines, His152, is present but the other two are replaced, respectively, by Asn196 and Tyr218, leaving no room for the metal.

An analysis of the genomic data for the FCD domain family (PF07729) reveals that more than 2800 members have been identified to date in 402 species of Eubacteria and 4 Archaea. The amino acid sequences show low, ~21% average identity of full alignment. A majority (>70%) contain a complete set of motifs with all four putative metal binding residues, that together make up a consensus fingerprint: **R**-X_3-Φ**E**-X_{19}- Φ-X_{19}-D/N- X_2-Φ**H**- X_3-Φ-X_2-S/T-X_2-N-X_2-Φ-X_6-Φ-X_{20}-**H**-X_6- Φ-X_3-D-X_3-A-X_6-**H,** where Φ denotes a hydrophobic residue, typically Leu, Met or Ile, and residues involved in metal coordination are shown in bold. Because of poor amino acid sequence conservation in this family, this fingerprint is not readily identifiable by automated sequence alignment.

Numerous examples of bacterial species contain a number of FCD family proteins: *Mycobacterium smegmatis* contains 46 of these regulators, *Rhodococcus sp* RHA1 - 49, *Arthrobacter sp* (FB24) - 28 and *Agrobacterium tumefaciens* - 51. Interestingly, the sequences are very diverse within each species but, in each case, about two-thirds show conservation of all metal-binding amino acids. This situation is in stark contrast to the FadR_C family, for which there are only 71 annotated sequences, in 70 species (with only one gene per organism), and average amino acid identity of 48%.

3.7. Functional implications

The structural evidence presented here strongly suggests that the majority of FCD domains, and, therefore, the majority of the FadR transcription regulators, are metal—most likely Zn^{2+}—dependent. What is not clear is whether these transcription factors are metal-sensing, or if the metal plays a structural role, or perhaps is required for binding of other effector molecules through direct coordination bonds. Metal-sensing transcription factors are ubiquitous in prokaryotes, with seven major families characterized to date (Giedroc & Arunkumar, 2007). Five of these families, i.e. ArsR, MerR, CopY, Fur and DtxR, utilize WH domains, also found in the GntR regulators, for binding to dsDNA. Almost all these proteins are dimeric, and metals bind typically at or near dimer interfaces, enabling the metal-bound form the regulators to repress, de-repress, or activate transcription of operons coding for metal efflux pumps, transporters, redox machinery, etc. (Giedroc & Arunkumar, 2007; Pennella & Giedroc, 2005; Silver & Phung le, 2005). In the FCD domains, the metal binding site is distinctly buried within an individual monomer, and removal by dialysis takes a relatively long time, which would seems to argue against a role in sensing changes in metal concentration. It is, therefore, more plausible that the FCD domains bind carboxylic acids, or small organic compounds containing carboxylic groups, so that the latter are buried and interact directly with the metal at the bottom of the ligand binding cavity. The presence of acetate (or less likely carbonate) in the TM0439 structure is consistent with that hypothesis. However, the polar cavities observed inside the metal-binding FCD domains of TM0439 and CGL2915 are relatively small, and do not appear to be able to bind larger organic compounds: calculations with a 1.4 Å probe result in only ~130 $Å^3$ for TM0439, and ~72 $Å^3$ for CGL2915. Interestingly, in PS5454, the volume of the cavity is difficult to estimate because one of the flanking loops is disordered in the crystal structure, and the cavity appears to be open to bulk solvent. The loop that is disordered links the α_4-helix with the α_5-helix. We note that PS5454 is unique in that the α_4-helix is straight, lacks the characteristic kink, and it is possible that the structure represents an 'active' conformer in which the cavities are open and able to bind a ligand, while the WH domains are ~68 Å apart, i.e. ideally positioned to bind to major grooves separated by two complete turns of the dsDNA.

Further studies will be needed to fully characterize the new metal-binding subfamily of the FadR transcription regulators.

This study was supported by NIH NIGMS grants U54 GM074946-01US (ISFI), U54 GM074898 (JCSG), and R01 GM042569 (to DPG). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. The authors would like to thank the staffs at the BL 5.0.2 managed by the Berkeley Center for Structural Biology (BCSB) at the Advanced Light Source (ALS) for technical support. The BCSB is supported in part by the National Institutes of Health, National Institute of General Medical Sciences. The ALS is supported by the Director, Office of

Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Figure legends:

Figure 1. Overview of the structure and comparison to other FadR superfamily transcription factors. VanR group members are shown in the top of the figure and FadR group members are shown at the bottom of the figure. The PDB codes for the proteins shown are: TM0439, 3fms; *Rhodococcus sp*. Protein RO03477, 2hs5; *Pseudomonas* protein PS5454, 3c7j; FadR, 1e2x; and CGL2915, 2di3. The red and pink colors denote the DNA binding domain, with the HTH motif highlighted in red. The FCD domain has been painted with a spectrum from blue to red, with the α0 helix of the FadR subfamily highlighted in magenta. The grey chain represents the second monomer in the dimer.

Figure 2 The overall architecture of the HTH domain of TM0439, with putative DNA binding residues shown. The DNA is modeled into this figures based on the superposition of the FadR / DNA comples (1hw2) onto the HTH domain of TM0439.

Figure 3. The regulatory domain of TM0439 and comparison with other FCD/FadR domains. The overall domain structure and a close up of the kinked helix α4 is shown for each protein are shown on the right and left, respectively. In each domain the kinked, α4 helix is shown in red. The seventh helices of the FadR group members, α0, are shown in yellow. The wire cages are the cavities calculated by VOIDOO (Kleywegt & Jones, 1994). The metals have been displayed with a radius of 2.0 Å to highlight their position.

Figure 4. The dimerization interfaces of the FCD and FadR_C domains. For TM0439, two complete FCD domains are shown, with one monomer colored as in Figure 3. Residues described in the text are represented as sticks. In B-D only the helices that participate in dimerization are shown.

Figure 5. Metal binding sites of TM0439 (3fms); CGL2915 (2di3); and PS5454 (3c7j). An omit map contoured at 5σis shown for TM0439. This was generated by deleting the metal and acetate and truncating the histidines back to the Cβ atoms, shaking the coordinates to yield an rmsd of 0.3 Å, and performing a round of refinement in PHEXIX.REFINE.

Figure 6. Metal binding by TM0439 monitored by Trp fluorescence: (A) 200 μM Ni^{2+}; and (B) 200 μM Zn^{2+} titrated into 5.3 μM Tm0439. The inset plots the emission (λ = 340 nm) vs. metal/protein molar ratio and the red line indicates the best-fit according to a one-site (Ni^{2+}; $K = 1.47 \pm 0.01 \times 10^{7}$ M^{-1}) or two-site (Zn^{2+}; $K_1 = 1.4 \pm 0.3 \times 10^{7}$ M^{-1} and $K_2 = 4.5 \pm 0.4 \times 10^{5}$ M^{-1}.) sequential binding model in DynaFit (Kuzmic, 1996) accounting for appropriate metal-buffer interactions. Note that that the best fit shown in the inset of **B** represents a lower limit of Ki values, because as long as ${}^{K_1}\!/_{K_2}$ remains constant, larger K_1 and K_2 fit the data equally well (simulations not shown).

Table 1. Crystallographic data

Data Collection Statistics			
	Peak	**Edge**	**Remote**
Wavelength	0. 97960	0.97980	0.95370
Resolution (Å)	40 - 2.10 (2.18 -2.10)*	40 - 2.10 (2.18 -2.10)	40 - 2.10 (2.18 -2.10)
Total Reflections	77,866	101,252	94,823
Unique Reflections	12,020	14,439	14,002
Redundancy	6.5 (3.6)	7.0 (5.1)	6.8 (4.2)
Completeness (%)	81.7 (27.4)	97.8 (84.2)	94.5 (64.9)
R_{merge} (%)**	6.3 (35.8)	5.4 (20.9)	5.3 (28.6)
Average I/σ (I)	31.2 (2.5)	52.6 (5.5)	42.4 (3.4)
Wilson B Factor (Å^2)	29.7	34.0	33.5
Refinement Statistics			
Wavelength			0.95370
Resolution (Å)			40 - 2.2 (2.42 - 2.20)
Completeness			97.6 (91.0)
Reflections (working)			12,586
Reflections (test)			620
R_{work} (%)§			15.7 (16.7)
R_{free} (%)§			22.8 (27.7)
Number of waters			81
R.m.s. deviation from ideal geometry			
Bonds (Å)			0.017
Angles (°)			1.31
Average B Factors(Å^2)†			
Main Chain			38.9
Side Chain			38.1
Waters			50.2
Molprobity Results			
Overall clashscore			4.89 (98th percentile)
Ramachandran - favored			203 (98.1%)
Ramachandran - outliers			1 (0.5%)

* The numbers in parentheses describe the relevant value for the highest resolution shell.

** $R_{merge} = \sum |I_i - \langle I \rangle| / \sum I$ where I_i is the intensity of the i-th observation and <I> is the mean intensity of the reflections. The values are for unmerged Friedel pairs.

§ $R = \sum||F_{obs}| - |F_{calc}|| / \sum|F_{obs}|$, crystallographic R factor, and $R_{free} = \sum||F_{obs}| - |F_{calc}|| / \sum|F_{obs}|$ where all reflections belong to a test set of randomly selected data.

† B-factors were refined using TLS approximation (see Methods)

References

Argos, P., Garavito, R. M., Eventoff, W., Rossmann, M. G. & Branden, C. I. (1978). *J Mol Biol* **126**, 141-158.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002). *Nucleic Acids Res* **30**, 276-280.

Chakrabarti, P. (1990a). *Protein Eng* **4**, 49-56.

Chakrabarti, P. (1990b). *Protein Eng* **4**, 57-63.

Chakrabarti, P. (1994). *J Mol Biol* **239**, 306-314.

Christianson, D. W. & Alexander, R. S. (1989). *J Am Chem Soc* **111**, 6412-6419.

Derewenda, Z. S. (2004). *Structure (Camb)* **12**, 529-535.

Derewenda, Z. S., Derewenda, U. & Kobos, P. M. (1994). *J Mol Biol* **241**, 83-93.

Derewenda, Z. S. & Vekilov, P. G. (2006). *Acta Cryst D* **62**, 116-124.

DiRusso, C. C., Heimert, T. L. & Metzger, A. K. (1992). *J Biol Chem* **267**, 8685-8691.

DiRusso, C. C., Metzger, A. K. & Heimert, T. L. (1993). *Mol Microbiol* **7**, 311-322.

Dokmanic, I., Sikic, M. & Tomic, S. (2008). *Acta Cryst D* **64**, 257-263.

Franco, I. S., Mota, L. J., Soares, C. M. & de Sa-Nogueira, I. (2006). *J Bacteriol* **188**, 3024-3036.

Gajiwala, K. S. & Burley, S. K. (2000). *Curr Opin Struct Biol* **10**, 110-116.

Gao, Y. G., Suzuki, H., Itou, H., Zhou, Y., Tanaka, Y., Wachi, M., Watanabe, N., Tanaka, I. & Yao, M. (2008). *Nucleic Acids Res*.

Giedroc, D. P. & Arunkumar, A. I. (2007). *Dalton Trans* 3107-3120.

Goldschmidt, L., Cooper, D. R., Derewenda, Z. S. & Eisenberg, D. (2007). *Protein Sci* **16**, 1569-1576.

Haydon, D. J. & Guest, J. R. (1991). *FEMS Microbiol Lett* **63**, 291-295.

Holm, L., Kaariainen, S., Wilton, C. & Plewczynski, D. (2006). *Curr Protoc Bioinformatics* **Chapter 5**, Unit 5 5.

Huffman, J. L. & Brennan, R. G. (2002). *Curr Opin Struct Biol* **12**, 98-106.

Kleywegt, G. J. & Jones, T. A. (1994). *Acta Cryst D* **50**, 178-185.

Krissinel, E. & Henrick, K. (2007). *J Mol Biol* **372**, 774-797.

Kuzmic, P. (1996). *Anal Biochem* **237**, 260-273.

Lee, M. H., Scherer, M., Rigali, S. & Golden, J. W. (2003). *J Bacteriol* **185**, 4315-4325.

Lesley, S. A., Kuhn, P., Godzik, A., Deacon, A. M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H. E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L. S., Miller, M. D., McPhillips, T. M., Miller, M. A., Scheibe, D., Canaves, J. M., Guda, C., Jaroszewski, L., Selby, T. L., Elsliger, M. A., Wooley, J., Taylor, S. S., Hodgson, K. O., Wilson, I. A., Schultz, P. G. & Stevens, R. C. (2002). *Proc Natl Acad Sci U S A* **99**, 11664-11669.

Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins* **50**, 437-450.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst D* **53**, 240-255.

Newman, J. (2005). *Acta Cryst D* **61**, 490-493.

NIST Standard Reference Database 46 (2003).

Otwinowski, Z. & Minor, W. (1997). *Methods in Enzymology* **A276**, 307-326.

Outten, C. E. & O'Halloran, T. V. (2001). *Science* **292**, 2488-2492.

Pennella, M. A. & Giedroc, D. P. (2005). *Biometals* **18**, 413-428.

Reynolds, W. F., Peat, I. R., Freedman, M. H. & Lyerla, J. R., Jr. (1973). *J Am Chem Soc* **95**, 328-331.

Rigali, S., Derouaux, A., Giannotta, F. & Dusart, J. (2002). *J Biol Chem* **277**, 12507-12515.

Silver, S. & Phung le, T. (2005). *J Ind Microbiol Biotechnol* **32**, 587-605.

Terwilliger, T. C. (2003). *Acta Cryst D* **59**, 1174-1182.

van Aalten, D. M., DiRusso, C. C. & Knudsen, J. (2001). *Embo J* **20**, 2041-2050.

van Aalten, D. M., DiRusso, C. C., Knudsen, J. & Wierenga, R. K. (2000). *Embo J* **19**, 5167-5177.

Whitehead, I. P., Campbell, S., Rossman, K. L. & Der, C. J. (1997). *Biochim Biophys Acta* **1332**, F1-23.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Acta Cryst D* **57**, 122-133.

Xu, Y., Heath, R. J., Li, Z., Rock, C. O. & White, S. W. (2001). *J Biol Chem* **276**, 17373-17379.

Zwart, P. H., Afonine, P. V., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., McKee, E., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K., Storoni, L. C., Terwilliger, T. C. & Adams, P. D. (2008). *Methods Mol Biol* **426**, 419-435.

Figure 1

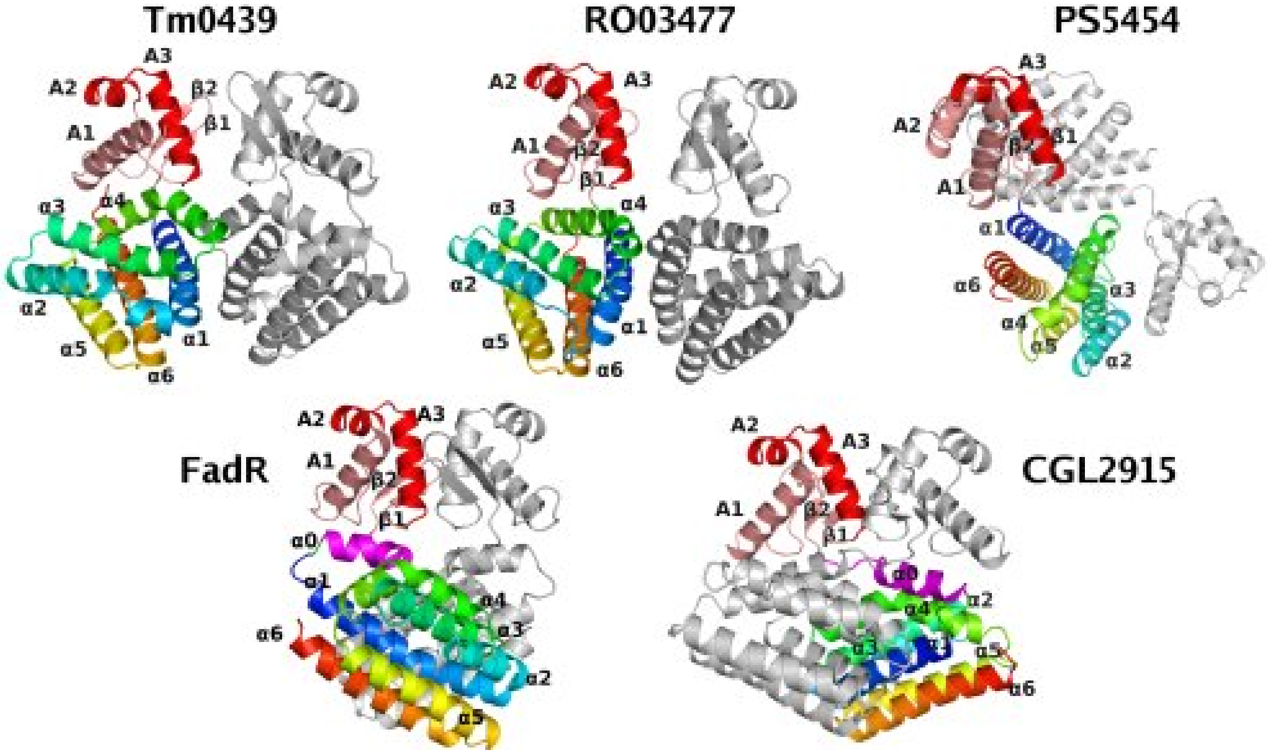

Tm0439
A3
A2
β2
A1
β1
α3
α4
α2
α5
α1
α6
RO03477
A2
A3
A1
β2
β1
α3
α4
α2
α5
α1
α6
PS5454
A3
A2
β2
β1
A1
α1
α6
α3
α4
α5
α2
FadR
A2
A3
A1
β2
β1
α0
α1
α4
α6
α3
α2
α5
CGL2915
A2
A3
A1
β2
β1
α0
α4
α2
α3
α1
α5
α6

Figure 2

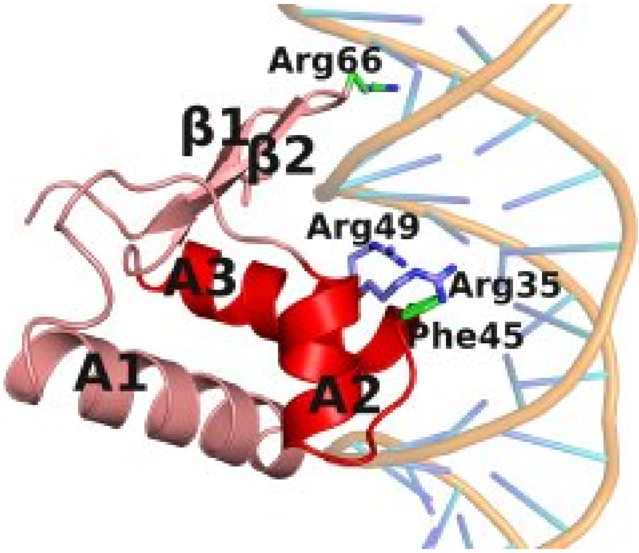


Figure 3

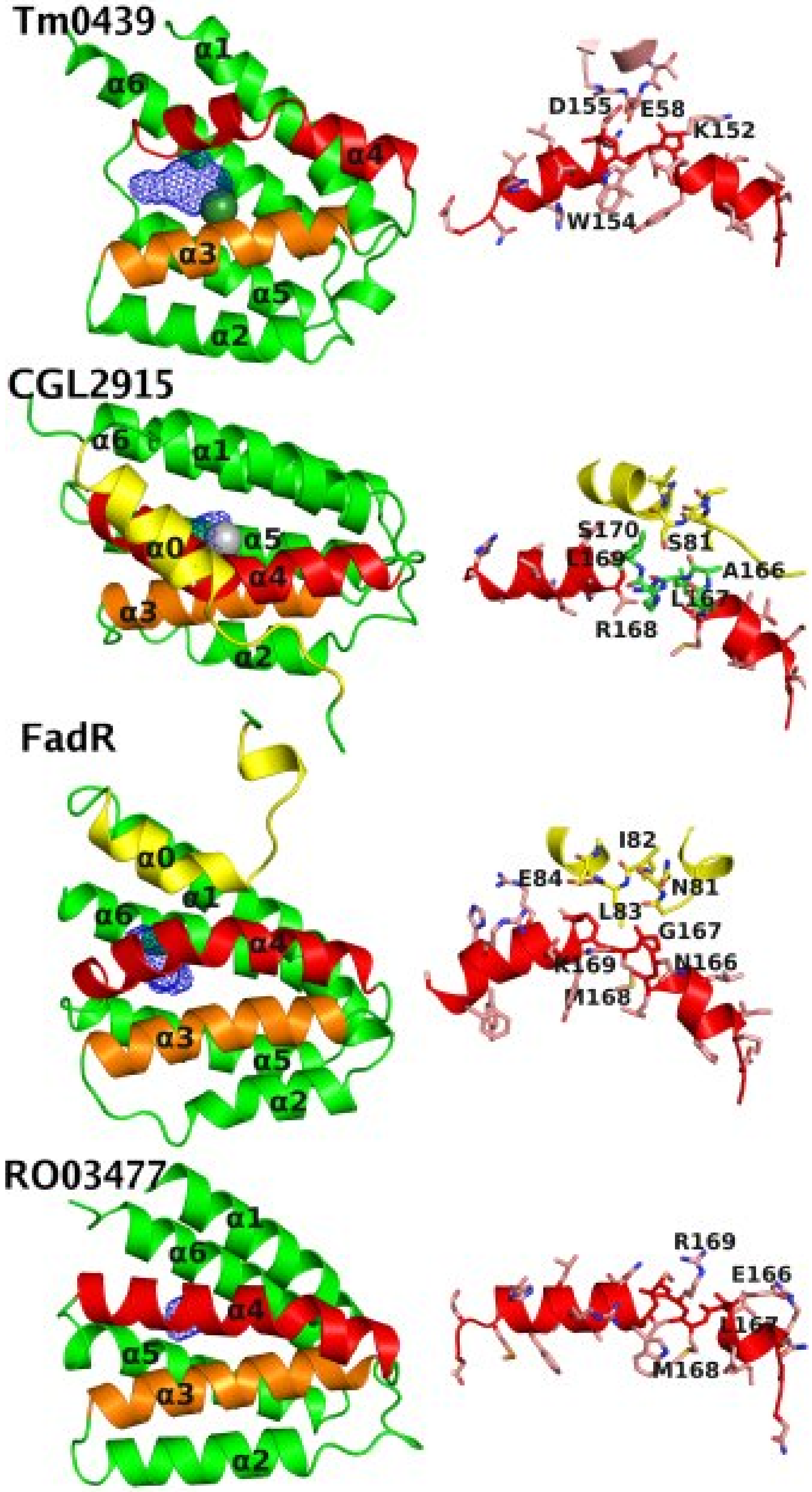

Tm0439
α1
α6
α4
α3
α5
α2
D155
E58
K152
W154
CGL2915
α6
α1
α0
α5
α4
α3
α2
S170
S81
L169
A166
L167
R168
FadR
α0
α1
α6
α4
α3
α5
α2
I82
E84
N81
L83
G167
K169
N166
M168
RO03477
α1
α6
α4
α5
α3
α2
R169
E166
L167
M168

Figure 4

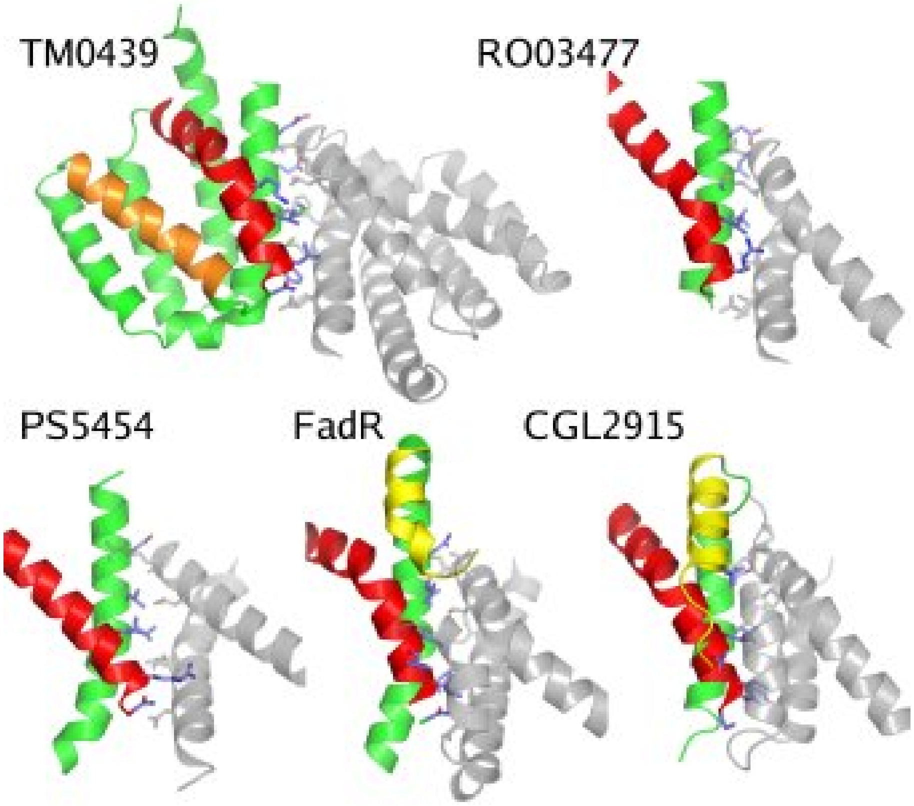


Figure 5

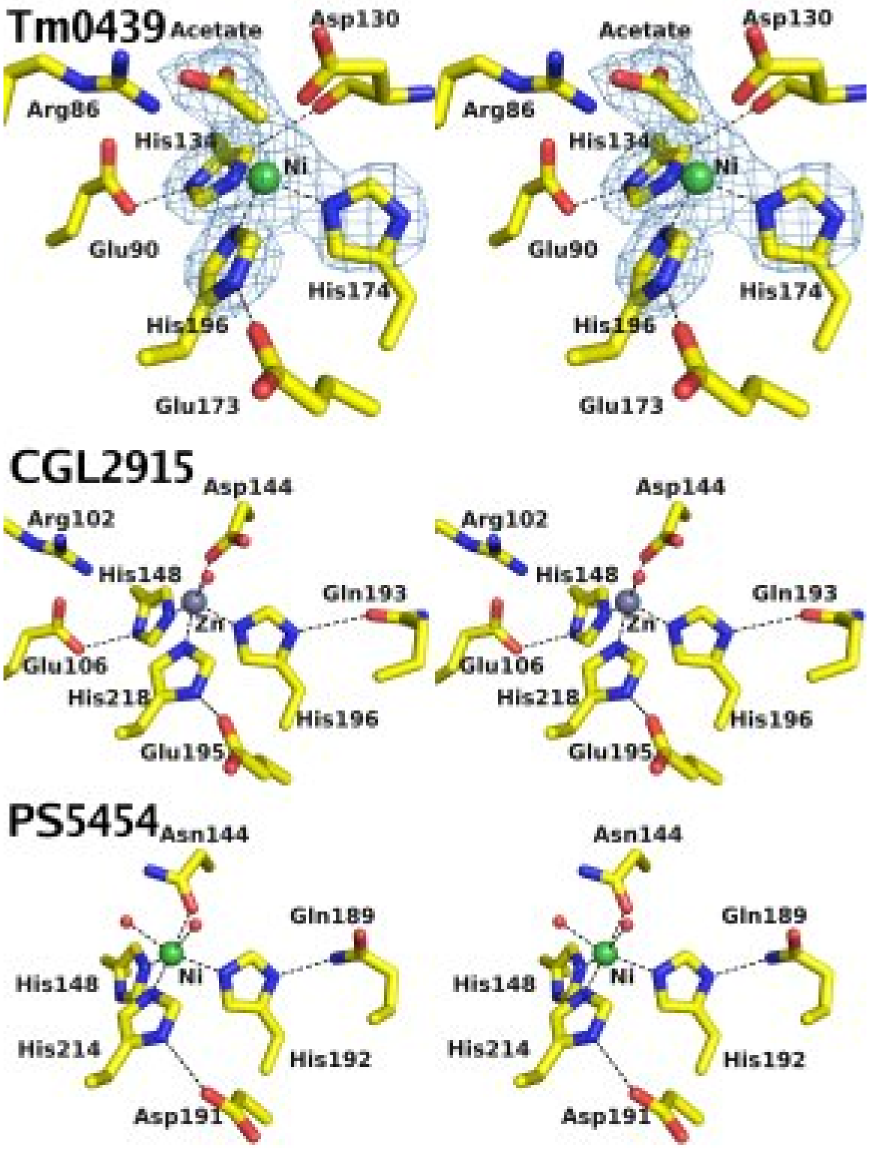
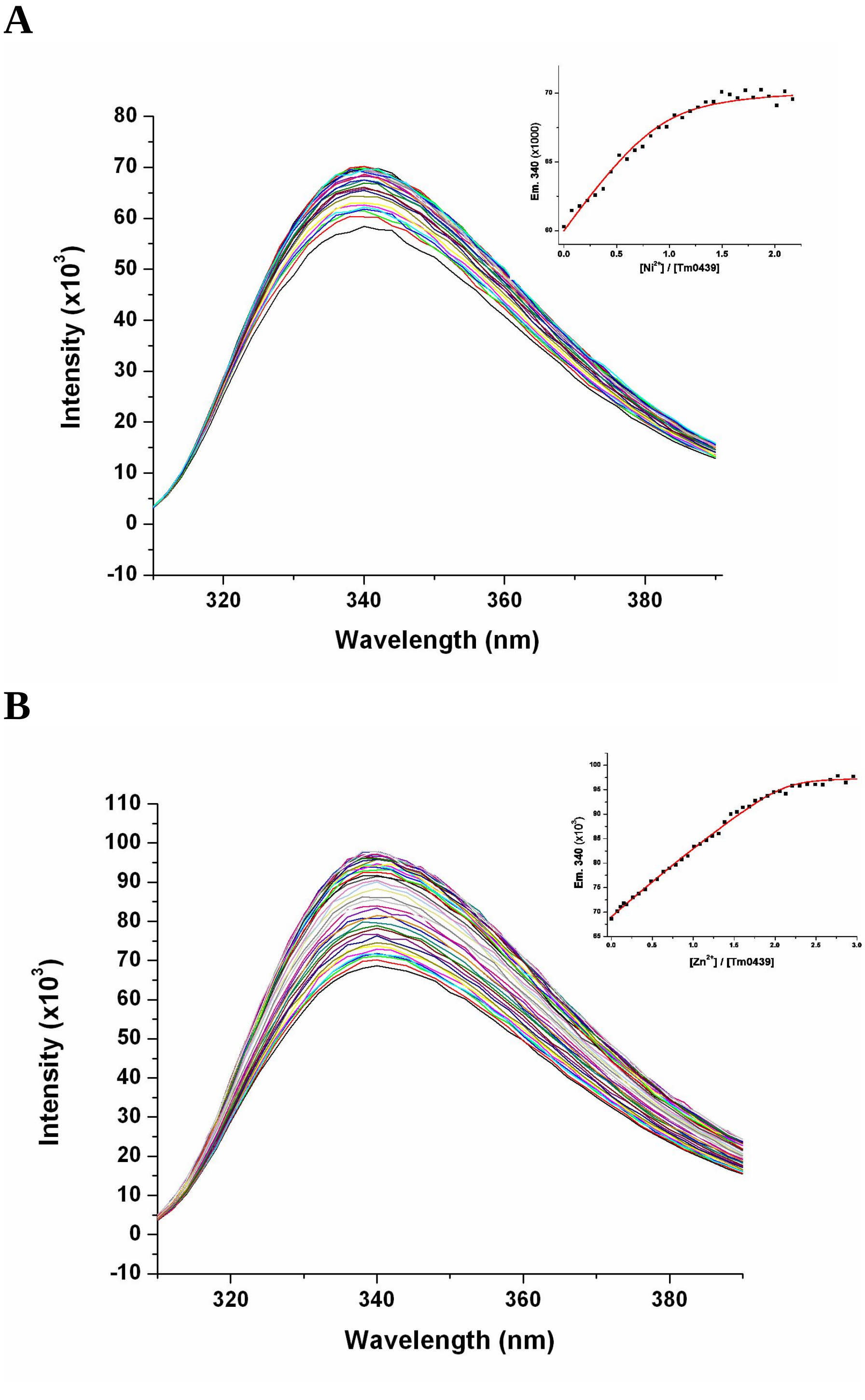

Tm0439
Acetate
Asp130
Arg86
His134
Ni
Glu90
His174
His196
Glu173
Acetate
Asp130
Arg86
His134
Ni
Glu90
His174
His196
Glu173
CGL2915
Asp144
Arg102
His148
Gln193
Zn
Glu106
His218
His196
Glu195
Asp144
Arg102
His148
Gln193
Zn
Glu106
His218
His196
Glu195
PS5454
Asn144
Gln189
His148
Ni
His214
His192
Asp191
Asn144
Gln189
His148
Ni
His214
His192
Asp191

Figure 6


A
Intensity (x10^3)
80
70
60
50
40
30
20
10
0
-10
320
340
360
380
Wavelength (nm)
Em. 340 (x1000)
70
65
60
0.0
0.5
1.0
1.5
2.0
[Ni^{2+}] / [Tm0439]
B
Intensity (x10^3)
110
100
90
80
70
60
50
40
30
20
10
0
-10
320
340
360
380
Wavelength (nm)
Em. 340 (x10^3)
100
95
90
85
80
75
70
65
0.0
0.5
1.0
1.5
2.0
2.5
3.0
[Zn^{2+}] / [Tm0439]