Summer 8-15-2017

# Specificity Determination by paralogous winged helix-turn-helix transcription factors

Adam Joyce
*Washington University in St. Louis*

WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Developmental Biology

Dissertation Examination Committee:
Timothy S. Schedl, Chair
James J. Havranek
Robi Mitra
Kristen Naegle
James B. Skeath
Gary D. Stormo

# Specificity Determination by Paralogous Winged Helix-turn-helix Transcription Factors

By

Adam P. Joyce

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2017
St. Louis, Missouri

# Table of Contents

LIST OF FIGURES…….. v

ACKNOWLEDGEMENTS……. v

ABSTRACT OF THE DISSERTATION…… vii

CHAPTER 1. INTRODUCTION…….……... 1

2.1 Two-component signaling pathway specificity…… 3

2.2 Structural attributes of the OmpR family ……...…….……... 5

2.2 References ……..……...…….……..... 6

CHAPTER 2. DIVERSE SEQUENCE-SPECIFIC DNA-BINDING BY *E. COLI* OMPR FAMILY RESPONSE REGULATORS…….…… 7

2.1 Context and Motivation for research…… 8

2.2 Introduction…….……...…….. 15

*2.2.1 Modeling TF:DNA specificity*……. 15

*2.2.2 Sequence recognition in the OmpR family*……. 16

2.3 Results…….……. 20

*2.3.1 Variation in protein and DNA structure at wHTH:DNA interface*…… 20

*2.3.2 Multi-specificity in sequence recognition by eRRs*…… 24

*2.3.3 eRR vary in half-site sequence, spacing, and orientation*…….. 26

*2.3.4 LIII specificity determinants include sequence and complex assembly*……. 30

2.4 Discussion…… 34

2.5 Materials and Methods.……..……...……... 38

2.6 Supplementary Figures……. 42

2.7 References……...……. 46

CHAPTER 3. SPECIFICITY-DETERMINING RESIDUES IN DNA-BINDING BY OMPR FAMILY TFS……...…… 52

3.1 Context and Motivation for research……. 53

*3.1.1 Structure-based modeling of protein:DNA interactions*……. 53

*3.1.2 Recent improvements in statistical potentials*……..... 55

*3.1.3 Modeling water in protein-DNA interfaces*……..... 61

*3.1.4 Flexibility at the protein:DNA interface*……..... 65

*3.1.5 Evaluating improvements in protein:DNA modeling*……..... 70

*3.1.6 Prospects for the future of structure-based modeling*……..... 72

*3.1.7 Definition and detection of specificity-determining residues*……... 73

3.2 Introduction……..…….. 77

3.3 Results……...……..... 79

*3.3.1 Specificity-centric MI subnetwork distinguishes the RH as a SDR hub*…... 79

*3.3.2 SDRs alter different aspects of eRR-DNA specificity*……... 82

3.4 Discussion…….. 85

3.5 Materials and Methods.…….....……... 89

3.6 References…….……...…….. 92

CHAPTER 4. CONCLUSIONS AND FUTURE DIRECTIONS……..…….. 101

4.1 Validation and characterization of non-canonical binding mode…….. 102

4.2 Characterization of native targets by Spec-seq……...……... 104

4.3 Expanding SDR prediction to other TF families……..…. 104

4.4 Training structure-based models……...…… 105

4.5 References……. 107

# List of Figures

Figure 2.1: Diverse residue contacts and DNA shape at the protein-DNA interface for OmpR family response regulators……..……. 23

Figure 2.2: Lineage-independent multi-specificity of eRRs…… 26

Figure 2.3: Variation in half-site recognition by OmpR family orthologs…….. 29

Figure 2.4: Determination of DNA binding specificity for CpxR, OmpR and RstA…….. 33

Figure S2.1: Phylogenetic tree of OmpR family represented in *E. coli*…….. 42

Figure S2.2: DNA shape varies between OmpR family RRs…….. 43

Figure S2.3: Deviation of phosphate backbone induced by RR binding. ……...… 44

Figure S2.4: Binding orientation revealed by base preferences adjacent to fixed half-site sequence…….……...…….. 45

Figure 3.1.1. Atomically detailed structures of protein:DNA complexes……..…… 55

Figure 3.1.2. Modeling protein:DNA complexes. …….…… 60

Figure 3.1.3. Water molecules at the protein:DNA interface participate in hydrogen bonding networks……..…. 65

Figure 3.1.4. Protein and DNA adopt diverse backbone conformations and orientations in complex…….. 70

Figure 3.3.1: Sequence- and structure-informed identification of specificity-determining residues at the protein-DNA interface…… 81

Figure 3.3.2: Quantitative determination of binding specificity for recognition helix variants in the CpxR and OmpR proteins. ……. 84

# Acknowledgements

I am deeply grateful to all those who impacted me throughout my graduate studies; the guidance of mentors, the critical input of colleagues, and support of friends and family made my time at Washington University both personally and professionally enriching.

I first and foremost thank my thesis mentor, Dr. James Havranek, for always encouraging me to approach scientific problems with creativity and independence; supporting my work and professional development; and allowing me to develop as a scientist. I also wish convey my appreciation to all of the members of my thesis committee who gave their helpful comments and constructive criticism over the years. Dr. Gary Stormo, especially, was always happy to offer scientific input. Through my outreach work, it was my great pleasure to work with Dr. Tom Woolsey, Rochelle Smith, and Dr. John Edwards, who work tirelessly behind the scenes to ensure the DBBS community is welcoming, diverse, and inclusive for everyone.

It was also a great privilege to work with and learn from so many excellent DBBS students and post-docs over the years. My lab-mates Dr. Benjamin Borgo and Dr. Chi Zhang as well as Kenny Chang, Dr. Zheng Zhuo, David Granas, Dr. Basab Roy, and all the past and present members of the Stormo Lab. I also offer special thanks to all of my great friends, colleagues, and advisors in the *Young Scientist Program* and *DBBS Connections* for broadening my perspective on the research community.

I cannot hope to name all of the wonderful friends I made throughout my time as a graduate student, but there are a few I'd like to single out: Dr. Marie Strand, Dr. Jon Fisher, Dr. Claire Schulkey, Dr. Katie Nutsch, Dr. Chris Fiore, and Dr. Elizabeth Selleck, Jessica Miller, Dr. Francesco Vallania, Dr. Jamie Kwasnieski, Dr. Britta Anderson and Drew Hughes. During my

time in St. Louis I also met many people who helped keep me grounded: the elite footballers of the Sunday Funday Soccer League, Bell Community gardeners, and an unforgettable Pathfinder party.

Last, but certainly not least, I thank my all the members of my family for their encouragement during the course of my degree. I'll always be grateful to my beautiful fiancé GiNell, my favorite person in the world, for her advice, her tireless support, and for giving me all my happiest memories (and, I hope, many more to come). My caring sister Cailin was always ready to share her experience, and made St. Louis truly feel like home. And, even from far away, I knew that I always count on the encouragement of my parents and brother when I needed it.

Adam Joyce

*Washington University*

*August 2017*

ABSTRACT OF THE DISSERTATION

Specificity determination by paralogous winged helix-turn-helix transcription factors

by

Adam Joyce

Doctor of Philosophy in Biology and Biomedical Sciences

Developmental Biology

Washington University in St. Louis, 2017

Professor Timothy Schedl, Chair

Transcription factors (TFs) localize to regulatory regions throughout the genome, where they exert physical or enzymatic control over the transcriptional machinery and regulate expression of target genes. Despite the substantial diversity of TFs found across all kingdoms of life, most belong to a relatively small number of structural families characterized by homologous DNA-binding domains (DBDs). In homologous DBDs, highly-conserved DNA-contacting residues define a characteristic 'recognition potential', or the limited sequence space containing high-affinity binding sites. Specificity-determining residues (SDRs) alter DNA binding preferences to further delineate this sequence space between homologous TFs, enabling functional divergence through the recognition of distinct genomic binding sites.

This thesis explores the divergent DNA-binding preferences among dimeric, winged helix-turn-helix (wHTH) TFs belonging to the OmpR sub-family. As the terminal effectors of orthogonal two-component signaling pathways in *Escherichia coli*, OmpR paralogs bind distinct genomic sequences and regulate the expression of largely non-overlapping gene networks. Using high-throughput SELEX, I discover multiple sources of variation in DNA-binding, including the

spacing and orientation of monomer sites as well as a novel binding 'mode' with unique half-site preferences (but retaining dimeric architecture). Surprisingly, given the diversity of residues observed occupying positions in contact with DNA, there are only minor quantitative differences in sequence-specificity between OmpR paralogs. Combining phylogenetic, structural, and biological information, I then define a comprehensive set of putative SDRs, which, although distributed broadly across the protein:DNA interface, preferentially localize to the major groove of the DNA helix. Direct specificity profiling of SDR variants reveals that individual SDRs impact local base preferences as well as global structural properties of the protein:DNA complex.

This study demonstrates clearly that OmpR family TFs possess multiple 'axes of divergence', including base recognition, dimeric architecture, and structural attributes of the protein:DNA complex. It also provides evidence for a common structural 'code' for DNA-binding by OmpR homologues, and demonstrates that surprisingly modest residue changes can enable recognition of highly divergent sequence motifs. Importantly, well-characterized genomic binding sites for many of the TFs in this study diverge substantially from the presented *de novo* models, and it is unclear how mutations may affect binding in more complex environments. Further analysis using native sequences is required to build combined models of *cis*- and *trans*-evolution of two-component regulatory networks.

# Chapter 1. Introduction

All living organisms employ signal transduction systems that receive, interpret and ultimately determine the appropriate physiological response to environmental conditions, available metabolites, and real-time activity of other cells. In many ways, the architecture of signaling networks in eukaryotes and prokaryotes reflects their opposing evolutionary strategies. Eukaryotes (and other multicellular organism life) typically invest substantial energy in robust developmental processes that ensure a limited range of variation in the extracellular environment, which advantages complex pathways that integrate predictable sensory stimuli and robust developmental transitions [1]. Conversely, prokaryotes must adapt quickly and precisely to navigate the limitless spectrum of environmental conditions (e.g., medium osmolarity, $0_2$ content, etc.), specific chemical compounds (e.g., antibiotics, carbon sources, etc.), and various states of cellular stress (e.g. protein unfolding, membrane disruption, etc.) to ensure survival and a competitive rate of replication. For example, the chemotactic behavior of *E. coli* is governed largely by rapid sensory responses to glucose gradients via the cheY system, which biases its otherwise random motions toward environments with ample food [2]; individual cells often contain dozens of such pathways, the sum total of which optimizes survival in the intended environmental niche [3].

## 1.1 Two-component signaling pathway specificity

Most prokaryotic signaling inputs are transduced via a 'two-component' architecture, comprised of a transmembrane sensor histidine kinase (HK) and a cytoplasmic response regulator (RR). Typically, detection of an appropriate stimulus by extracellular 'sensory' domains of a HK stimulates the phosphorylation of a highly conserved aspartate residue in the 'receiver' domain of the cognate RR, altering its behavior [4].

The high degree of sequence and structural homology among two-component proteins operating in the same cellular space raises the potential for cross-talk between pathways, and multiple strategies have arisen to maintain the fidelity of signal transmission [3]. At each HK:RR interface, cognate components are selected by a subset of 'specificity-determining residues' (SDRs), which, if mutated (or transferred between HK:RR pairs), can re-wire two-component pathways [5; 6]. Additionally, the majority of HKs are bifunctional, displaying both kinase and phosphatase activity toward RRs [1], with strong "kinetic preference" toward their cognate partner. For example, the VanS HK exhibits a $10^4$-fold preference ($k_{cat}/K_M$) *in vitro* for its cognate RR (VanR) in comparison to the non-cognate PhoB [7]; thus, HKs can rapidly suppress any spurious RR phosphorylation (as well as maintain tight control over cognate RR activity). Due to the mild promiscuity of the HK:RR interface, active competition within the cellular RR pool

for HK-occupancy is unavoidable, and theoretical studies have shown that introducing multiple competing RRs reduces the pathway sensitivity [1].

Specificity mechanisms downstream of each two-component pathway are less clear, as there is no formal reason for pathways to maintain distinct outputs. The majority of RRs contain one or more 'effector' domains that serve a regulatory role in cellular processes. Most commonly, RRs contain DNA-binding domains (DBDs) belonging to the winged helix-turn-helix (wHTH) family [8]. These RRs are collectively defined by structural homology to the prototypical osmolarity response protein OmpR, and also share certain functional characteristics [9; 10].

Conventionally, phospho-activation of an OmpR-family RR shifts the monomer-dimer equilibrium toward a predominantly homodimeric state, which co-orients DBDs and promotes cooperative binding at tandem repeat sequences [11–13]. Phosphorylation of OmpR family RRs may also stimulate the coordinated occupancy of multiple adjacent binding sites, often leading to complex regulatory outcomes. For example, *E. coli* ArcA was shown to target genomic regions containing the tandem repeat 'TGTTA$N_5$TGTTA' (distributed in phase with the DNA double helix), and individual repeats yield distinct effects on ArcA occupancy level, repressor activity, and responsiveness to pathway activation [14; 15]. Likewise, OmpR-dependent enhancers at multiple porin genes (e.g, *ompF, ompC*, *ompS1*, etc) exhibit distinct binding affinity and regulatory activity, and their unique, enhancer-specific sensitivity to OmpR mutations further suggests that the conformation of enhancer-bound OmpR is sequence-dependent

[16–18]. Binding studies of $RR_{DBD}$ show little evidence of direct (DBD:DBD) cooperativity [19], but flexible interactions between regulatory domains [20] and DNA-mediated allostery likely play a role [21–23].

## 1.2 Structural attributes of the OmpR family

Helix-turn-helix (HTH) protein domains are distributed throughout all kingdoms of life, wherein they most commonly mediate specific and non-specific DNA-binding in transcriptional regulation, DNA repair, and replication [24]. The HTH domain itself is a simple right-handed, tri-helical bundle, but many structural variations have emerged to carry out increasingly specific functions. In this work, I focus on the winged helix-turn-helix (wHTH); more specifically, the OmpR sub-family.

The basic topology of the wHTH domain consists of the HTH motif, a right-handed, tri-helical bundle that makes up both the hydrophobic core and bulk of the DNA interface, situated between an N-terminal, antiparallel β-sheet and C-terminal hairpin (or 'wing') [9; 10]. OmpR family DBDs are distinguished from other wHTH domains by the presence of an N-terminal β-sheet, the extended HTH 'turn', and a somewhat long alpha α3-helix. Bound to DNA, the latter functions as an archetypical 'probe helix', and projects into the major groove orthogonal to the helical axis. Structural evidence suggests that this sub-family of RRs assembles in a head-to-tail orientation on the DNA, typically embedding the wing residues of one monomer into the intervening minor groove [13].

## 1.3 References


1. Rowland MA, Deeds EJ (2014) Crosstalk and the evolution of specificity in two-component signaling. Proc Natl Acad Sci U S A, 111, 5550–5555

2. Baker MD, Wolanin PM, Stock JB (2006) Signal transduction in bacterial chemotaxis. Bioessays, 28, 9–22

3. Laub MT, Goulian M (2007) Specificity in two-component signal transduction pathways. Annu Rev Genet, 41, 121–145

4. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annu Rev Biochem, 69, 183–215

5. Podgornaia AI, Casino P, Marina A, Laub MT (2013) Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling. Structure, 21, 1636–1647

6. Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. Cell, 133, 1043–1054

7. Fisher SL, Kim SK, Wanner BL, Walsh CT (1996) Kinetic comparison of the specificity of the vancomycin resistance VanSfor two response regulators, VanR and PhoB. Biochemistry, 35, 4732–4740

8. Galperin MY (2010) Diversity of structure and function of response regulator output domains. Curr Opin Microbiol, 13, 150–159

9. Martínez-Hackert E, Stock AM (1997) Structural relationships in the OmpR family of winged-helix transcription factors. Journal of molecular biology, 269, 301–312

10. Kenney LJ (2002) Structure/function relationships in OmpR and other winged-helix transcription factors. Curr Opin in Microbiol, 5, 135–141

11. Toro-Roman A, Wu T, Stock AM (2005) A common dimerization interface in bacterial response regulators KdpE and TorR. Protein Sci, 14, 3077–3088

12. Gao R, Tao Y, Stock AM (2008) System-level mapping of Escherichia coli

response regulator dimerization with FRET hybrids. Mol Microbiol, 69, 1358–1372

13. Blanco AG, Sola M, Gomis-Rüth FX, Coll M (2002) Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. Structure, 10, 701–713

14. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ (2013) The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genet, 9, e1003839

15. Park DM, Kiley PJ (2014) The influence of repressor DNA binding site architecture on transcriptional control. MBio, 5, e01684–14

16. Mattison K, Oropeza R, Byers N, Kenney LJ (2002) A phosphorylation site mutant of OmpR reveals different binding conformations at ompF and ompC. J Mol Biol, 315, 497–511

17. Flores-Valdez MA, Fernández-Mora M, Ares MÁ, Girón JA, Calva E, De la Cruz MÁ (2014) OmpR phosphorylation regulates ompS1 expression by differentially controlling the use of promoters. Microbiology, 160, 733–741

18. Head CG, Tardy A, Kenney LJ (1998) Relative binding affinities of OmpR and OmpR-phosphate at the ompF and ompC regulatory sites. J Mol Biol, 281, 857–870

19. Narayanan A, Paul LN, Tomar S, Patil DN, Kumar P, Yernool DA (2012) Structure-function studies of DNA binding domain of response regulator KdpE reveals equal affinity interactions at DNA half-sites. PLoS One, 7, e30102

20. Walthers D, Tran VK, Kenney LJ (2003) Interdomain linkers of homologous response regulators determine their mechanism of action. J Bacteriol, 185, 317–324

21. Mizuno T (1987) Static bend of DNA helix at the activator recognition site of the ompF promoter in Escherichia coli. Gene, 54, 57–64

22. Ogasawara H, Yamada K, Kori A, Yamamoto K, Ishihama A (2010) Regulation of the Escherichia coli csgD promoter: interplay between five transcription factors. Microbiology, 156, 2470–2483

23. Siggers T, Gordân R (2014) Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res, 42, 2099–2111

24. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev, 29, 231–262

# Chapter 2. Diverse Sequence-specific DNA-binding by OmpR family Response Regulators[1]

[1] This chapter is adapted from the following provisionally accepted manuscript: 'Deciphering the protein-DNA code of bacterial winged helix-turn-helix transcription factors'

## 2.1 Context and motivation for research

Large-scale specificity profiling of large TF families can lead to many important insights in TF function and evolution; probably the clearest example of this is the homeodomain-containing class of TFs. Homeodomain TFs make up one of the largest classes of metazoan TFs, and regulate developmental gene networks essential to cellular differentiation [1], tissue patterning [2], and specification of the vertebrate body plan [3]. In two of the first large-scale efforts to profile a TF family, two groups simultaneously determined the DNA-binding specificities of homeodomain TFs in the *Drosophila melanogaster* [4] and *Mus musculus* [5] genomes, each identifying multiple sub-specificities within the class that could (1) be attributed structurally to specific DNA-contacting residues and (2) be applied to predict the specificities of homeodomain TFs with distant homology. In a follow-up analysis, Chu and colleagues [6] adopted a 'DNA-centric' approach to identify mutant homeodomains that recognized specific sequences not observed in the naturally-occurring specificity profiles. Surprisingly, this analysis revealed a number of "novel combinations of specificity determinants that are uncommon or absent in extant HDs" with novel DNA-binding properties, suggesting that these determinants are disadvantageous or inaccessible to naturally evolving systems. Further profiling of the polymorphic (human) homeodomain TFs has since shown substantial effects on DNA-binding affinity and specificity, in many cases associated with Mendelian disorders [7]. As a result of the aforementioned studies (and many others not

mentioned), the function, evolution, and biochemistry of homeodomain TFs are, to some extent, 'solved problems', providing critical context for complex genomic analysis and biological studies.

Other than sharing a common 'helix-turn-helix' DNA-binding domain fold, the specific biology of the aforementioned homeodomain family is of little importance to this thesis. Rather, it is an example of the power of deep, broad analysis of DNA-binding specificity in TF families to reveal properties crucial for predicting phenotypic variation, understanding TF-related disease states, and decoding developmental gene regulatory networks. However, deep specificity profiling is, at the time of this writing, largely absent for prevalent prokaryotic TF families and DNA-binding domains.

There are many technical reasons for the intense focus on eukaryotic TFs (eTFs) and general avoidance of prokaryotic TFs (pTFs) in large-scale profiling efforts. First, the sequences bound by pTFs (10-20bp) are usually quite long compared to eukaryotic TFs (6-8bp), reflective of the alternative cis-regulatory strategies of eukaryotic and prokaryotic genomes. (A complete discussion of this topic is outside the scope of this thesis.) Popular specificity profiling technologies, such as protein-binding microarrays [8] (PBMs) and bacterial one-hybrid platforms [9], are limited in their coverage of long sequences. PBMs, for example, offer full coverage of all 10mer sequences, sufficient for the vast majority of eTFs but insufficient to describe the longer binding sites of pTFs. Secondly, it is standard in large-scale TF specificity profiling to remove accessory domains, which are assumed not to affect sequence-specific binding. Although this may

be a safe assumption for eTFs, pTFs often require structural domains outside the DBD for dimerization or assembly of the protein:DNA complex; often this process involves the transmission of a signal (e.g. phosphorylation) or the binding of a chemical ligand. It is difficult to produce, handle, and 'activate' such proteins; moreover, the complexity and diversity of DNA-binding mechanisms confounds systematic, large-scale analysis. Notably, recent and future advances in selection-based protocols (e.g. SELEX [10]) and targeted, EMSA-based library screening (e.g., Spec-seq [11; 12]) have and will continue to push these boundaries. Thirdly, and substantively, the motivation for understanding eTFs and pTFs are very different. Profiling eTFs is often aimed at decoding the combinatorial 'regulatory grammar' of biological processes - a research aim in itself - whereas the analysis of pTFs is often a utilitarian step in the identification of regulatory targets (for which there are more direct and reliable approaches than predictions based on specificity models [13]).

Speaking to the basic motivation behind pTF analysis, there are three main areas of research, in particular, that can benefit greatly from a more complete quantitative understanding of these protein families. First and foremost, pTFs harbor vast potential as tools for synthetic biology, particularly in the construction of artificial genetic circuits. For example, the TetR TF family was recently 'genomically mined' *in vitro* to identify orthogonal, or non-overlapping, sequence-specific binding activity; 16 distinct TFs were ultimately discovered and converted to separate 'NOT/NOR' logic gates (theoretically capable of $10^{54}$ circuit combinations) [14]. Secondly, the relative simplicity of

prokaryotic cellular systems (compared to eukaryotes) makes these organisms more amenable to top-down systems modeling; in fact, efforts are already underway to create an '*in silico* cell', a theoretical, complete reconstruction of the metabolic and regulatory processes underlying cellular homeostasis, growth, and information processing [15]. With regard to whole-genome transcriptional regulatory models, projects such as the *Transcription Factor Profiling of Escherichia Coli* (https://shigen.nig.ac.jp/ecoli/tec/top/) show great promise in providing the foundational data to begin the prediction of regulatory networks and TF interactions [13]. Thirdly, the complexity of pTF binding can reveal biochemical properties of protein:DNA interactions that are not prominent in less sophisticated eTFs. For example, early structural analysis of the Trp repressor revealed a sequence-specific protein:DNA interface with no direct residue-base contacts; instead, specificity was mediated through DNA backbone, and base hydration [16]. Examples of DNA-binding specificity through DNA deformation [17], bending [18], and cooperative interactions[19] abound in prokaryotic systems, and will continue to inform models of complex TF:DNA interactions (e.g., multi-protein enhancer complexes).

In this work, I chose to generate a representative profile of DNA-binding specificity of the OmpR sub-family of wHTH TFs in *E. coli* for several reasons. First, as previously discussed, two-component signaling systems are heavily insulated from cross-talk between pathways, and I set out to understand how (or whether) that property was communicated to downstream regulatory networks [20]. Second, in contrast to many eTF families, bacterial response regulators function in the same cellular volume, imposing

unique and stringent constraints to prevent (or promote) cross-reactivity at the transcriptional level. Thirdly, few studies [21] have isolated and explored the role that intrinsic sequence recognition potential might play in sequence recognition in a synthetic environment, away from the influences of native genomic sequence or co-regulatory factors. Finally, despite its prevalence (>50%) in bacteria, the DNA-binding specificity of winged helix-turn-helix TFs is, in general, poorly understood.

## 2.2 Introduction

### 2.2.1 Modeling TF:DNA specificity

TF binding specificity may be modeled in a variety of ways, each striking a balance between simplicity of representation and quantitative resolution for the chosen application [22; 23]. The earliest representation of sequence-specific TFs, for example, was the 'consensus sequence,' a string format in which DNA bases (A/C/G/T) at each position represents that most frequently observed in a set of aligned binding sites [24]. This form of model has significant quantitative shortcomings, as it equates the relative importance of any base in plurality; for example, consensus models will represent two positions as 'T' bases despite the magnitude of their plurality (40% T ~ 100% T). Modifications to the consensus approach include the establishment of arbitrary thresholds (i.e. setting a minimum threshold frequency for a position to contribute to the model) and the use of more complex IUPAC ambiguity codes representing multiple base identities (http://www.bioinformatics.org/sms/iupac.html). Despite the shortcomings of consensus and string-format specificity models, they can be highly useful in certain applications. For example, the most precise functional model of a TF with a single genomic binding site would, in fact be that single sequence; likewise, the genomic distribution of TFs with predominantly low-affinity binding sites can be more accurately predicted from simple consensus models. The most widespread class of TF:DNA recognition models is the position weight matrix (PWM); the functional form of these models is extensively

described elsewhere [25]. Similar to other approaches, PWMs are constructed from a pre-defined frequency matrix based on a set of aligned binding sites; observed base frequencies at each position in the binding site are log-transformed to yield weights that generally approximate true differences in binding energies [25]. From a functional standpoint, these models correct for background probabilities of different DNA bases, which is highly useful when binding sites are derived or predicted from skewed genomic sequences (e.g., A-rich promoters). PWMs may additionally accommodate more complex binding parameters, such as preferences for dinucleotide preferences [26], which may reflect DNA shape recognition [27], base- and base-step geometry [28], or bidentate contacts [29]. Many additional model types have been applied to represent TF binding specificity, including hidden Markov models [30], neural networks [31], and ranked lists [32], as well as structure-based biophysical models, which are the subject of *Chapter 3*.

### 2.2.2 Sequence recognition by the OmpR family

The sequence preferences of several OmpR family members are defined by consensus-based models using a small number of native operator sequences [33], while others have been generated from high-throughput functional assays [34]. Because native sequences are subject to multiple selective pressures, their utility for constructing quantitative specificity models is limited. For example, both CpxR and OmpR bind the *ompF* enhancer in porin gene regulation [35], and both factors can affect different regulatory outcomes depending on the sequence and architecture of binding sites [36]

[37]. Thus genomically-derived models will reflect this complex functional relationship. Models derived entirely *in vitro*, however, should accurately reflect only the intrinsic sequence preferences of a TF, and may provide significant regulatory insight. For example, Park and colleagues [34] have shown that arrayed 'TGTTA' repeats direct binding of *E. coli* ArcA *in vivo*, and a previous *in vitro* binding analysis produced a near-identical model for *Shewanella* ArcA [21]. In contrast, OmpR binding sites are highly divergent from the *in vitro* consensus, indicating that these two factors have evolved distinct, affinity-based strategies for target discrimination [38]. For other OmpR family proteins, it would appear that binding motifs derived *in vitro* and *in vivo* are completely inconsistent. PhoP, for example, has been shown to regulate distinct sub-populations of targets through different 'submotifs' with distinct evolutionary rates and species distribution, most of which exhibit the characteristic repeat architecture of the OmpR family [39]. However, a SELEX-derived model of *Mycobactium tuberculosis* PhoP revealed a novel sequence preference, unrelated to any previously reported for OmpR family RRs [40].

Since the initiation of this work, the *Transcription Factor Profiling of E. coli* Project (and affiliated groups) (TEC) have generated comprehensive catalogs of *in vitro* binding profiles using SELEX with fragmented genomic DNA. Depending on the method of analysis, this technique produces an set of enriched genomic regions with high affinity for the targeted TF; in particular, TEC utilizes an *E. coli* tiling array to measure genomic enrichments, thereby restricting the resolution of putative sites to >60bp (probe length).

More modern techniques based on high-throughput sequencing, such as the recently developed Chip-exo [41], stand to substantially improve the identification of TF target sites; however, significant experimental confirmation by both footprinting and functional assays has revealed a great deal about the binding of many TFs, including several OmpR family RRs. For example, BasR and QseB (of the metals-responsive BasRS and quorem-sensing QseBC systems, respectively) enriched only partially-overlapping sets of genomic sequences in phosphorylated and non-phosphorylated states. Footprinting analysis of 3 phospho-BasR target regions revealed a likely preference for 'TTAA' half-sites in tandem orientation. Similar analyses were performed for OmpR [38], CpxR [42], and YedW [43], all of which validate previous findings *in vitro* and *in vivo* findings. An important limitation of genomic SELEX is its inability to distinguish energetic preferences from functional constraints on naturally evolved binding sites; in the aforementioned example, is 'TTAA' the true, preferred half-site, or one selected for reduced affinity toward BasR or close paralogs? Additionally, compared to modern *in vitro* specificity profiling techniques, such as PBMs or HT-SELEX, naturally-occurring sequences limit the coverage of potential binding sites necessary for accurate biochemical models; as such genomic SELEX is best-suited to large-scale TF target identification.

Despite their role in signaling pathways critical to bacterial homeostasis, disease processes, and bioengineering, sequence-specific DNA binding by the OmpR family remains poorly defined from a quantitative standpoint. Using a combination of techniques, we deeply characterize the recognition potential of ten OmpR homologues from *E. coli*,

and further identify specificity-determining residues both experimentally and computationally. We find that OmpR homologues are capable of multiple modes of sequence-specific DNA binding, and that the balance between these binding modes can be maintained by a single residue position.

## 2.3 Results

### 2.3.1 Variation in protein and DNA structure at wHTH:DNA interface

Using the PFAM database [44], we identified 18 putative winged helix-turn-helix (wHTH) transcriptional regulators (PF00486) present in the *E. coli* K12 genome, 14 of which fall into the archetypical class of bipartite, signal-activated transcriptional response regulators (*E. coli* RR, or eRR). They exhibit high similarity at residue positions presented toward the domain core as well as those in contact with the DNA phosphate backbone, implying the preservation of both wHTH fold structure and DNA-binding ability (**Figure 2.1a**) anticipated from prior structural and functional studies of OmpR family proteins [45; 46]. The protein:DNA interface spans three structural elements within the wHTH domain (α1 N-terminus (α1-N), beta strands β6-7 (wing), and α3 'recognition helix' (RH)), which contact different regions of the double helix (**Figure 2.1b**).

At the protein:DNA interface, highly conserved residues often interact non-specifically with DNA or make sequence-specific contacts important for 'familial' binding specificity [47], whereas narrowly conserved residues are more likely to act as specificity determinants between paralogous proteins. We identified 1925 high-confidence orthologs (see Methods) for the 14 eRRs, which we further subdivided into four distinct lineages (LI-IV) based on protein sequence similarity (**Figure S2.1**). The consensus for each alignment closely matched its corresponding eRR sequence in all

cases except PhoP, which contains two atypical, non-conservative mutations in the RH (**Figure 2.1c**). We immediately observed multiple positions in the DNA-contacting sub-domains exhibiting patterns of variability consistent with paralog-specific functions, especially at the DNA-exposed surface of the RH. [For consistency, we will hereafter reference these positions by their domain context, numbered position, and (if applicable) residue identity (e.g., $RH_{,(12)}$[R] ~ C-terminal Arg in RH).] For example, residue dyads Arg-Asp (LII, LIII, and LIV) and Asn-Glu (LI) were frequently observed at $RH_{(2,5)}$, indicating a coevolutionary relationship. Based on structural analysis, it is known that these residues interact directly at the protein:DNA interface, and $RH_{(5)}$[D/E] serves is a hub of polar interactions between $RH_{(2)}$[R/N], $W_{(7)}$[Y], and certain backbone-proximal residues at the N-terminus of α2. Planar residues are commonly preferred at $RH_{(7)}$, with a strong His prevalence in LIV proteins as well as the LIII protein CpxR; non-valine residues at $RH_{(6)}$ appear additionally to co-occur with $RH_{(7)}$[H]. $RH_{(9\text{-}10)}$ were typically conserved at lower levels than other DNA-contacting residues, but trends were apparent at the lineage level, such as the preference for $RH_{(9)}$ [S] in LIII. Overall, our evolutionary analysis validates the widely held assumption that DNA-contacting residues in the RH are the primary specificity determinants.

Because multiple residues with paralog- and lineage-specific distributions are positioned to interact primarily with the DNA backbone, we performed a comparative structural study to investigate the role of shape-specificity in RR-DNA binding. As previously reported, the DNA minor groove narrowed substantially in the spacer region in

all five protein:DNA structures, favoring DNA curvature toward the bound face of the dimer [48], and we further observed that the DNA major groove expanded in the region occupied by the RH (**Figure S2.2a**). We then superimposed RHs to visualize the relative position of the DNA helix, and found that phosphate backbone trajectories diverge strand-specifically directly over the half-site (**Figure S2.2b**). For each RR, backbone trajectory was similar between upstream and downstream half-sites, leading us to conclude that structural variability is primarily dependent on protein binding, not underlying sequence (**Figure S2.3**).

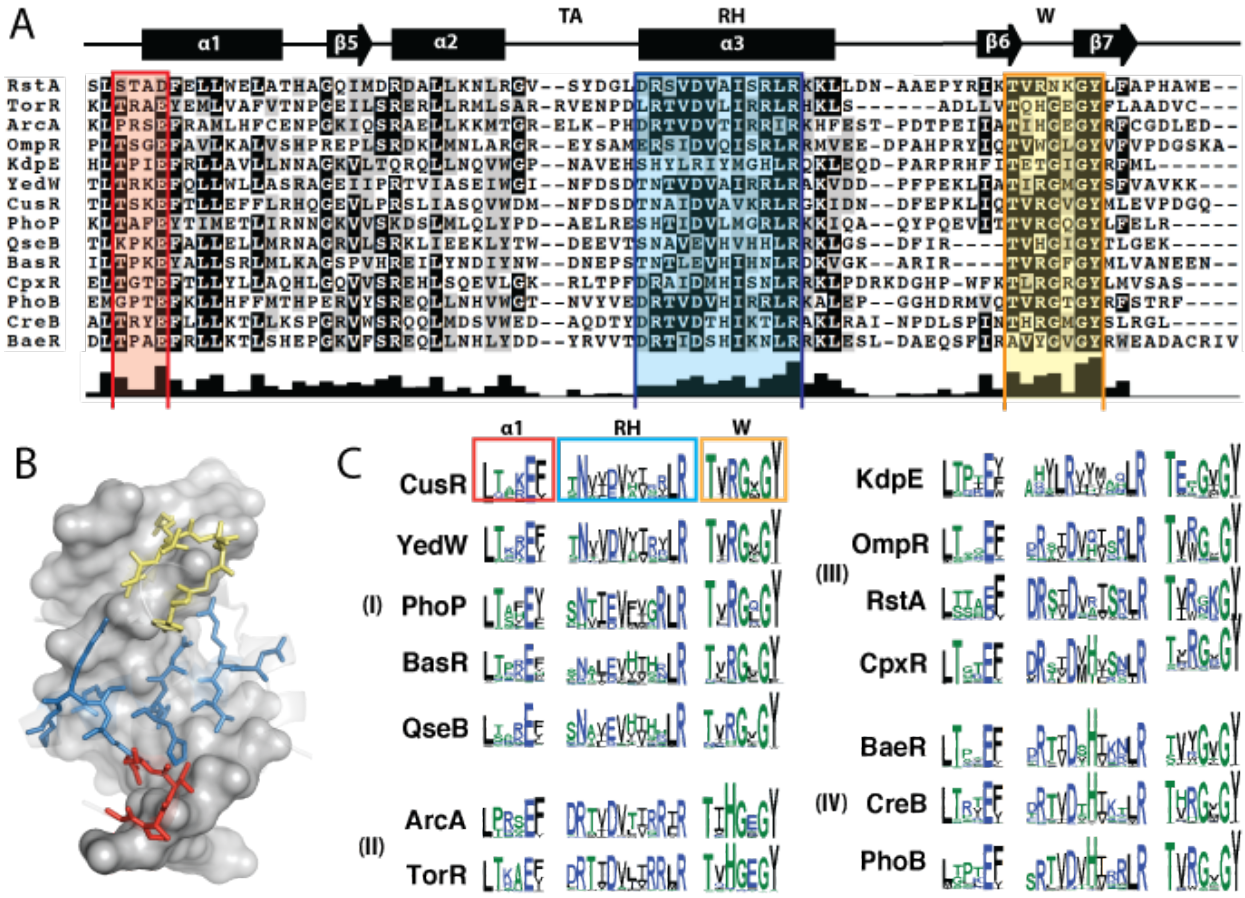


**Figure 2.1: Diverse residue contacts and DNA shape at the protein-DNA interface for OmpR family response regulators**. **A.** The results of a structural alignment of winged helix-turn-helix domains for *E. coli* K12 OmpR homologues are presented. Highly and moderately conserved residues are highlighted by black and grey boxes, respectively, and a histogram of relative entropy is plotted for each position (bottom) from a sampling of 1000 proteins from OmpR homologues identified previously [49]. Colored regions indicate structural sub-domains that contact DNA in representative co-crystal structures of OmpR homologues bound to target DNA sequences. **B.** DNA-contacting residues in the RH, W, and α1 are shown using the crystal structure of a single PhoB monomer bound to a high-affinity half-site (PDB code: 1GXP [48]). Residues are rendered as sticks and colored in correspondence with the alignment in panel A;

DNA is shown in a grey surface representation. **C.** Residue conservation within the RH is displayed in sequence logo format, organized into lineages (I-IV) based on amino acid similarity over the full-length protein.

### 2.3.2 Multi-specificity in sequence recognition by eRRs

To systematically explore the intrinsic sequence recognition potential of the OmpR family, we constructed a randomized (20N) library and performed high-throughput SELEX on phosphorylated and non-phosphorylated eRR. Binding motifs were identified *de novo* for five proteins (KdpE, BasR, QseB, BaeR, and OmpR) from three of the previously identified lineages, revealing two apparent 'modes' of binding (**Figure 2.2**). One mode, characterized by a half-site based on the consensus 'GT-A', was enriched following selection with CpxR (LIII), OmpR (LIII), and QseB (LI). The three 'GT-A'-binding eRR exhibited different responses to chemical phosphorylation. OmpR, for example, yielded near-identical binding motifs (i.e., same specificity) in both phosphorylation states; however, the overall representation of sequences in the selected pools was higher in the in the phosphorylated state. We can thus infer that phospho-OmpR bound its specific targets with greater affinity, as observed in many previous studies, resulting in a greater number of sequences stably bound (and selected) in each successive round. Phosphorylated and non-phosphorylated CpxR, by contrast, yielded *de novo* binding motifs consistent with direct and inverted repeat architectures, respectively. The recognition of inverted repeats has previously been suggested at

genomic binding sites for the copper-responsive RRs CusR and YedW ; however, there is currently no corroborating structural evidence. Additionally, because these assays cannot differentiate between complexes of distinct molecular weight, inverted repeat motifs for CpxR, CusR, and YedW are consistent with both (1) an inverted dimer (→ ←) and (2) adjacently-bound dimers (→→ ←←) exhibiting the standard, direct architecture. Surprisingly, a distinct sequence repeat containing a novel 'GCT' core was also enriched in selections using QseB, KdpE, and BaeR. Half-site recognition was asymmetric ('ACGCTN$_4$TTGCT'), with preferential specificity toward the upstream and downstream sites in the presence and absence of phosphodonor, respectively.

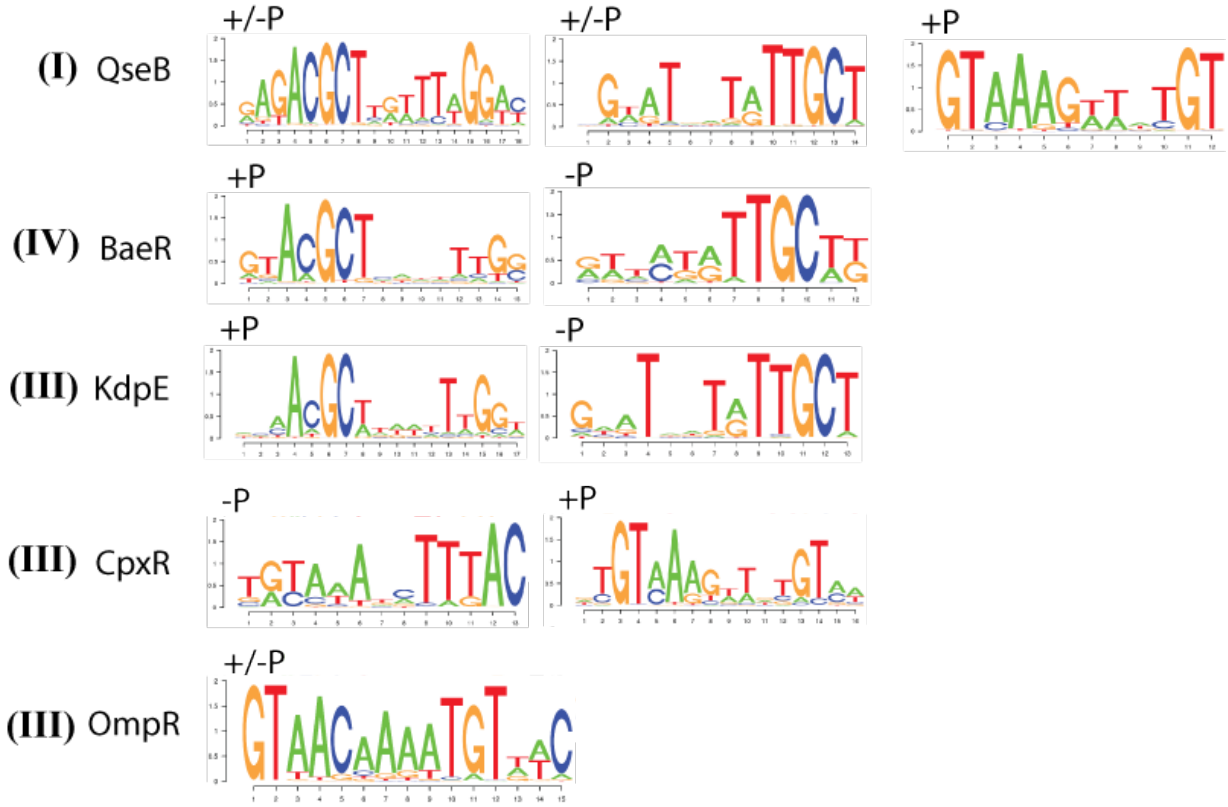


**Figure 2.2: Lineage-independent multi-specificity of eRRs.** DNA sequence logos are derived from *de novo* motif searching of SELEX pools. Binding motifs were discovered following selection of the indicated eRR in the presence (+P) or absence (-P) of phosphodonor, or are representative of the same motif identified in both conditions independently (+/-P).

### 2.3.3 eRR vary in their preference for half-site sequence, spacing, and orientation

The 'GT-A' repeat sequences identified through SELEX bore similarity to binding sequences previously derived from *in vitro* and *in vivo* analyses of OmpR ('TGTAACAAAATGTTTC') [19], CpxR ('GTAA($N_6$)GTAA') [50], RstA ('GTA'/'GTAAC') [51], PhoP ('TGTTta') [52], PhoB (TGTCA) [53], and ArcA

(‘TGTTA’) [21]. We expected that this could represent a familial mode of binding specific to the OmpR family. To explore the diversity of repeat sequence, spacing, and orientation in this mode of binding, we performed three rounds of SELEX on each paralog (+/- phosphodonor) using a partially randomized library flanked on one side by a synthetic half-site, (‘AG**GTAA**$(N)_{20}$’). Binding motifs (each representing thousands of individual sequences) were identified *de novo* for eight eRRs in their phosphorylated and non-phosphorylated states (**Figure 2.3a**). Despite differences in sequence and regulatory function, the OmpR family overall displays a consistent preference for half-sites of the form ‘$t_{(+1)}$GTnAn$_{(+6)}$‘ (on the reverse strand, ‘$n_{(-6)}$TnACa$_{(-1)}$‘), hereafter referred to as the ‘canonical’ mode (**Figure 2.3b, left**). In general, half-site sequences varied over the profiled TFs, but mainly at the weakly selective fourth and sixth positions**.** CreB was a notable exception, adopting a preference for a $G_{(+4)}$, highly similar to the ‘cre tag’ sequence previously observed in promoters of several CreBC TCSP targets [54] (Figure 3B, center). Interestingly, there was evidence of a similar preference for $G_{(+4)}$ for PhoB, also an LIV protein, but the overall specificity was difficult to assign due to the palindromic structure of the half-site [55].

To investigate spacing and orientation preferences for each eRR, we asked whether ‘hits’ to representative half-site position-weight matrices (PWMs) were over-represented at specific positions (in forward or reverse orientation) within the randomized region for the selected sequence pools. Overall, orientation of putative binding events suggested that recognition of direct repeats is a familial trait, and spacing preferences range only

narrowly from 9bp to 10bp center-to-center distance (ctcd). A few notable exceptions to this rule include 1) strong binding of CusR (L1) to a head-to-head inverted repeat, 2) a lack of spacing preference for the non-phosphorylated form of KdpE, and 3) an atypical pattern of spacing preference for CpxR. Importantly, these exceptions reflect known binding activities of these factors *in vivo* [43] [56] [57]. Overall, through *in vitro* SELEX we have observed a surprising similarity in DNA-binding (a 'canonical mode') by OmpR family proteins , and we have also identified non-ubiquitous binding characteristics for individual family members that explain variable behavior *in vivo*.

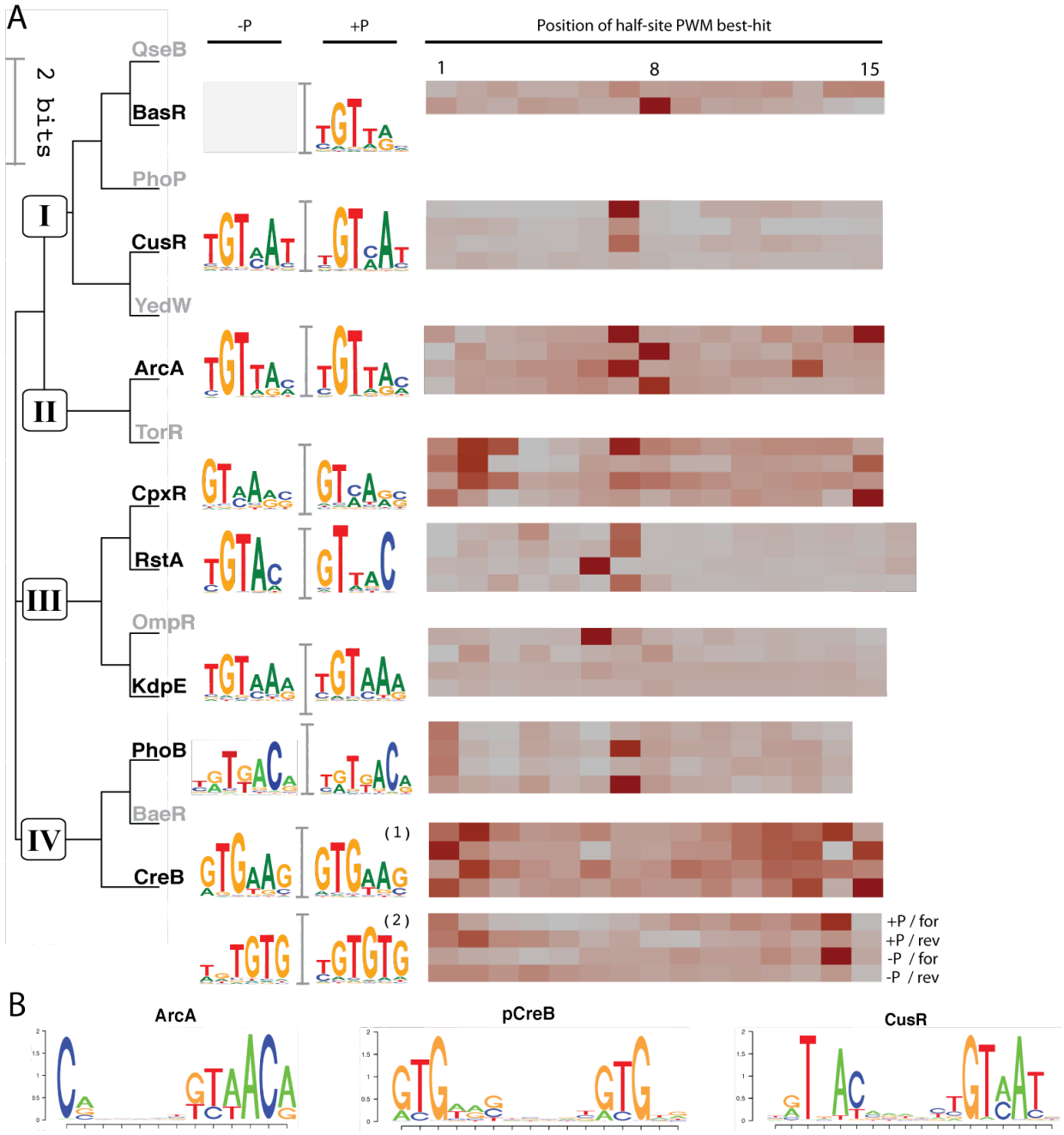


**Figure 2.3: Variation in half-site recognition by OmpR family orthologs**. **A.** Representative half-site sequence logos derive from selection of the 'anchored' GTAA-(20N) degenerate library are shown for nine eRR, with the names of eRR for which logos could not be obtained are in grey. Motifs obtained in the presence or absence of phosphodonor are indicated by +P and -P, respectively. The dendrogram reflects phylogenetic relatedness of the full-length consensus sequences, and distinct lineages (I-IV) correspond to

Figure 2.1c. For each motif, we display heat maps indicating the distribution of start sites for putative half-site binding events (top-scoring weight matrix hits) within the randomized region of the library. Note that longer motifs (for example those of PhoB) have fewer potential starting positions in the 20bp library, giving rise to a shorter heat map. Separate distributions are displayed by row in each heat map for putative binding events by phosphorylated and non-phosphorylated proteins (+P, -P) in either forward or reverse orientation (for, rev) relative to the 'fixed 'GTAA' half-site, ordered as shown for CreB (bottom). **B.** Full-length eRR motifs reflect alternative use of the fixed half-site and 20bp randomized region.

### 2.3.4 LIII specificity determinants include sequence preference and complex assembly

The LIII family members CpxR, OmpR, and RstA target overlapping and/or identical operators in the *E. coli* genome. However, they exhibited varied recognition of asymmetric binding sites and/or mechanisms of multi-meric assembly in a SELEX format suggesting they may have different affinities for their common genomic targets. To further interrogate specificity determinants within this lineage, we performed Spec-seq, a technique permitting the measurement of relative affinities toward thousands of individual sequences while visualizing protein:DNA assembly in an EMSA format [12; 58]. We designed a partially randomized library based on the high-affinity OmpR site, which was similar to SELEX-derived consensus sequences for CpxR and RstA. Each base pair was biased towards the consensus during synthesis (85% native base, with each alternate base present at 5%) to produce a complex library of targeted variants (~5.4%

OmpR consensus, 4.7% single-mutants, and 1.7% double-variants); this design also permitted a broad sampling of sequences with greater similarity to the RstA or CpxR consensus, albeit at lower frequency in the pool. To reduce the prevalence of weak binders in each pool, we selected high-affinity binders over three rounds without replacement; as expected, average library affinity was reduced in each successive round (data not shown). For each TF, the three pools were combined in a 3:2:1 ratio, in ascending order of average affinity.

Strikingly, we observed distinct banding patterns for all three proteins, indicating alternative mechanisms of assembly and/or complex structure. OmpR formed a single complex that migrated identically in both phosphorylation states, consistent with previous findings *in vivo* and *in vitro* that it minimally requires a homodimer – which is stabilized by phosphorylation - for target recognition (**Figure 2.4d-e, upper**). Phospho-CpxR failed to form discrete complexes, but rather shifted the population continuously in proportion to the total protein concentration, producing a prominent 'smear,' albeit one with distinct lower and upper bounds consistent with dimeric and tetrameric assembly states. Interestingly, non-phosphorylated CpxR also migrated continuously, but as a discrete band and at higher protein concentration (**Figure 2.4a-b, upper**). Phospho-RstA formed a distinct banding pattern composed of three closely separated micro-states (and a fourth high-molecular weight state), which depended on protein concentration (**Figure 2.4c, upper**). The upper two micro-states dispersed with increasing phospho-RstA levels, while the lowest band steadily increased in intensity.

One major advantage of Spec-seq over other techniques is the ability to produce high-resolution binding models directly from relative affinity measurements; this contrasts with other techniques that may incur artefacts by inferring binding energies indirectly from other measurable properties [59]. In this experimental context, we saw no evidence of the asymmetric half-site recognition previously observed for RstA using SELEX (Figure 4C, lower). However, this complex may be represented in the high affinity micro-states, which unfortunately did not yield enough material for sequencing. CpxR produced a recognition model that was distinct from OmpR and RstA, and also from its own previously generated by SELEX (Figure 5A,B). Strikingly, this novel mode of recognition specifically altered the recognition of the '**GT**-**A**', resulting in a dramatic departure from the canonical model. Half-site recognition was asymmetric, with the downstream monomer adopting a highly specific '$t_{(+1)}$g**TGA**$a_{(+6)}$' and an upstream binding motif undergoing a concentration-dependent shift from '$t_{(+1)}$g**TGA**$a_{(+6)}$' to '$t_{(+1)}$t**AAA**$n_{(+6)}$'. In summary, this analysis demonstrates unique binding properties and sequence recognition among three LIII family members, suggesting complex binding dynamics.

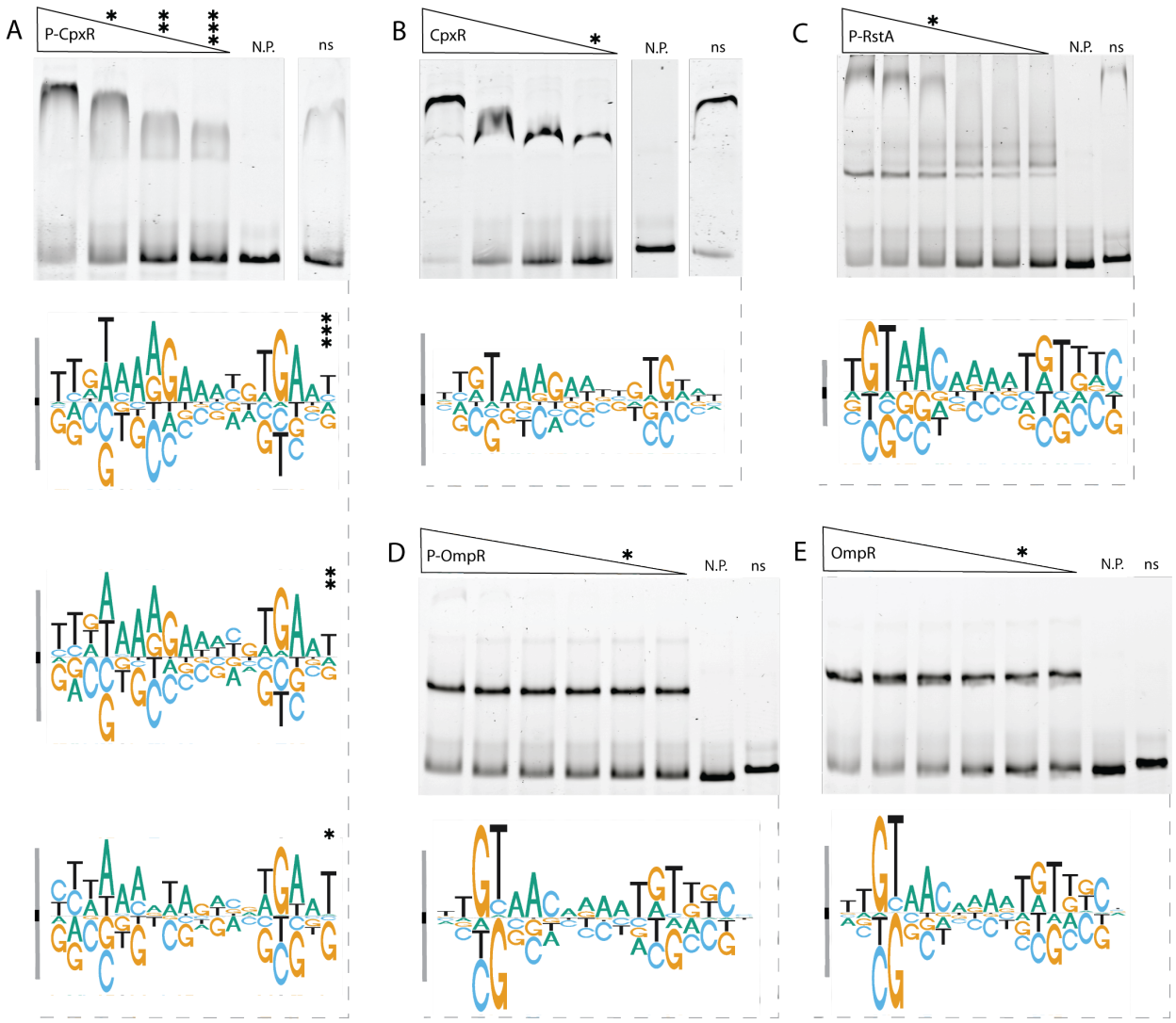


**Figure 2.4: Determination of DNA binding specificity for CpxR, OmpR and RstA**

DNA-bound complexes of full-length, strep-tagged phospho-CpxR **(A)**, CpxR **(B)**, phospho-RstA **(C)**, phospho-OmpR **(D)**, and OmpR **(E)** were obtained by excising bands separated by gel electrophoresis. Proteins were incubated with pooled, partially degenerate DNA libraries based on the OmpR consensus binding sequence (N.P., no protein; ns, non-specific library). Asterisks indicate the lanes that were analyzed. Library members in bound and unbound bands were sequenced and analyzed using Spec-seq. Energy logos below the images represent observed relative affinities of single-variants to the consensus (see Methods).

## 2.4 Discussion

In this chapter, I applied two different high-throughput techniques to profile specificity determinants in the OmpR sub-family of wHTH TFs, which constitute approximately 30% of downstream effectors in two-component signal transduction systems. Two-component signaling pathways (TCSPs) are a critical and predominant sensory modality in Bacteria, and are a classic system for the study of functional specificity of paralogous proteins and pathways. This work is notable for several reasons. First, in contrast to the simple mechanisms of DNA binding employed by most of the profiled eukaryotic TF families, OmpR family proteins bind as multi-mers in response to phosphorylation of a regulatory domain, greatly increasing the complexity of potential sequence interactions. Second, prokaryotic TFs are usually profiled individually with low-throughput methods, whereas we generated high-resolution specificity models from thousands of sequences for a representative majority of the OmpR family. Third, we utilized fully randomized, synthetic binding site libraries that allowed us to challenge OmpR family members with a more complex set of binding partners than they encounter *in vivo*. Our use of these *in vitro* libraries to identify binding motifs, rather than genomic DNA, further ensures that our results are unbiased by native sequence context. Fourth, using a recently developed technique known as Spec-seq, we were able to measure relative affinities toward thousands of sequences directly, while simultaneously visualizing the assembly of distinct protein:DNA complexes in an EMSA format. This

approach provided an unprecedented level of insight into the relationship between DNA-binding specificity and protein:DNA assembly, and added to our understanding of TF interactions important for gene regulation.

A striking result of our initial SELEX experiment, which used a fully-randomized 20N library, was the emergence of two binding motifs based on distinct 'GT-A' and '(AC/TT)GCT' sequences. Interestingly, at least one protein (QseB) appears capable of binding in both modes. Multiple modes of binding specificity within a single structural family have been proposed before, although they have also been shown to arise artefactually from the models used to represent sequence-specific interactions [59]. Nevertheless, bona fide multi-specific binding has been observed for the eukaryotic FOX family of sequence-specific TFs, which notably also contain a DBD belonging to the winged helix-turn-helix class (although distinct in some structural attributes) [60]. Furthermore, the two most common binding motifs for the profiled Forkhead-domain TFs are of the consensus 'GTAAAC' and 'ACGC,' partial matches for two of the binding motifs shared by OmpR homologues in our SELEX experiment. The alternative 'GCT' motif also has a structure somewhat similar to a binding consensus previously generated by *in vitro* SELEX for *Mycobacterium* PhoP (**GCT**GTGA) [40]. Both *Mycobacterium* PhoP and *Klebsiella* PmrA (a BasR homolog) have been crystallized in complex with sequences containing 'GCT'-like motifs occupying equivalent positions relative to the protein, with each overlapping at the canonically conserved $T_{(+3)}$ ('GC**T**' / 'G**T**$_{(+3)}$-A'). [61; 62].

Based on a moderately sized collection of binding sites and representative crystal structures, it is often assumed that OmpR family TFs recognize sequence repeats in tandem with center-to-center distance of 9-10 bp. However, the binding models derived *de novo* in this work cast doubt on the ubiquity of the direct repeat model, as we observed the emergence of binding sites with both tail-to-tail and head-to-head architectures. In the case of CusR, inverted (head-to-head) binding was previously demonstrated in a native operator sequence [43]. Because the solution-state techniques employed in this work do not resolve complexes of different size, however, the discovery of an inverted architecture could also be explained through the adjacent binding of two dimers in the canonical tandem orientation. Additionally, many of the half-site motifs discovered through SELEX are themselves semi-palindromic (e.g., OmpR: '**TGT**A**ACA**', PhoB: 't**GT**g**AC**a'), making it difficult to define site orientation with high confidence. Overall, we conclude that alternative dimeric architectures are possible (though not widespread), but future work is needed to confirm and explore the mechanism of their formation.

This representative collection of high-resolution binding models provides significant insight into the regulatory logic of the OmpR family. For example, the closely related (in primary sequence) BasR and QseB proteins appear to diverge in two different dimensions: half site recognition (TGTAAA vs. TGTTAc) and center-to-center distance (10bp vs 9bp). It is possible that the two parameters are inter-related; that altering the spacing of monomers causes (or requires) changes in the presentation of identical residues to the DNA. Additionally, in this study, *in vitro* SELEX analysis showed that BasR recognizes

‘TGTTAc’ half-sites (an apparently canonical mode of binding); however, a previously defined *in vivo* sequence for BasR ( ‘cTTAAnnTTnncTTAAnnTT’) diverges substantially from that model (losing the ‘G’ base preference broadly conserved across the OmpR family) [63]. In the same study, Ogasawara and coworkers showed that BasR forms multiple distinct complexes (determined by EMSA) with its native operators and occupies regions of DNA far larger than a single homodimer, indicating that BasR operators are tuned to interact with multi-meric protein complexes. This apparent contradiction suggests that higher-order protein:DNA complexes may have a greater influence on target recognition *in vivo* than intrinsic sequence preferences for OmpR family proteins.

In contrast to the ‘assembly-based’ mechanisms for specificity determination, some OmpR family RRs can be distinguished from other paralogs by highly unique site and sequence preferences. For example, CreB exhibited a highly unique variation of the canonical motif with a preference for ‘GTG-’-containing half-sites, which matches closely to the ‘cre-tag’ identified in promoters responsive to the CreBC two-component system [54]. Similarly, the paralog-specific head-to-head repeat preference observed for CusR matches precisely to recently characterized binding sites in several target promoters [43]. From an evolutionary standpoint, the emergence of a unique base preference would reduce the risk of cross-talk at paralog-specific operators, thereby reducing the need for an assembly-based strategy for operator discrimination. However, TFs with highly distinctive sequence preferences lose the ability to recognize common targets.

## 2.5 Materials and Methods

**Cloning, expression, and purification**

Coding sequences of 14 response regulators (RRs) of the OmpR sub-family (ArcA, BaeR, BasR, CpxR, CreB, CusR, KdpE, OmpR, PhoB, PhoP, QseB, RstA, TorR, YedW) were amplified directly from *E. coli* MG1655 genomic DNA. Coding sequence for the StrepTagII affinity tag (WSHPQFEK) was added by PCR amplification along with upstream and downstream restriction sites for MfeI and XhoI, respectively. Strep-RR fusion protein sequences were sub-cloned into the pET-42a(+) expression vector in-frame with N-terminal GST and 6xHis purification tags and a thrombin protease cleavage site, generating triple-tagged constructs. Stock plasmids were stored, purified and handled using standard laboratory techniques. ArcticExpress (DE3) competent cells (Agilent) were chemically transformed with expression plasmids, and single colonies from selective (Kan) LB-agar plates were used to inoculate 5ml LB-Kan starter cultures. After 6-8 hours growing at 37°C, starter cultures were scaled up to 400ml expression cultures in triple-baffled 4L flasks prepared with auto-induction media containing Kanamycin according to the Studier method [64]. Cultures were expanded at 37°C for 3-6 hours, then grown several hours past saturation (24-36 hours total growth time) at 20°C to achieve maximum protein yield. Bacterial pellets were harvested by centrifugation, sonicated, re-pelleted at high speed to remove cellular debris, and lysate (diluted with 1X PBS to reduce viscosity) was passed through a 0.45μm syringe-tip filter (MANUFAC) for

clarification. Lysate was passed over a HiTrap GST affinity column (1 ml capacity, GE Healthcare) and eluted under manufacturer-specified buffer conditions. Fusion protein was cleaved with 5U thrombin protease, and GST-6xHis was removed with two rounds of treatment with Ni-NTA resin (Thermo Scientific). Protein samples were cleared completely of resin by passage through 0.22 μm syringe-tip filters. Purity was assessed by both SDS-PAGE and size-exclusion chromatography, and protein concentration was determined by NanoDrop (Thermo Scientific).

**SELEX and Spec-seq library preparation**

DNA libraries were designed to contain flanking sequences to support PCR amplification and direct sequencing on the Illumina platform, and were obtained as a single-stranded, PAGE-purified oligonucleotides form Integrated DNA Technologies. For SELEX library construction, 250ng single-stranded DNA (ssDNA) were mixed with a reverse primer in two-fold molar excess in 1X NEBuffer 2 (50mM NaCl, 10mM Tris-HCl, 10mM $MgCl_2$, 1mM DTT, pH 7.9 at 25°C), heated to 85°C and slowly annealed to 30°C. Following the addition of 10U Klenow Fragment (NEB) and 1mM dNTPs, extension reactions were incubated at 37°C for 2 hours, and double-stranded DNA (dsDNA) libraries were subsequently purified using Qiaquick PCR Purification columns (Qiagen) and eluted in Qiagen EB (1omM Tris-C1, pH 8.5 at 25°C). Labeled dsDNA libraries for Spec-seq were generated by two-step PCR with Q5 High-Fidelity DNA Polymerase (NEB) using FAM-labeled primers and purified as described above.

**SELEX**

Strep-tagged proteins were pre-incubated 1 hour at 32°C in binding buffer (10mM Tris-Cl, 7.5; 200 mM KCl; 10 mM NaCl, 1 mM $MgCl_2$), 2 μg polydI-dC, 0.1 mg/ml BSA, and either $NH_4Cl$ or ammonium phosphoramidate for non-phosphorylated and phosphorylated conditions, respectively. Incubated protein samples were aliquotted (40 μl) into PCR strip tubes containing 200 ng of the appropriate DNA library, and incubated an additional hour at 32°C. Binding reactions were mixed with a washed suspension of Strep-tactin magnetic beads (Qiagen) and placed on ice for 30 minutes; to prevent bead settling, reactions were mixed by gentle pipetting at 10 minute intervals. Beads were pelleted magnetically and supernatant was removed by gentle pipetting. Pellets were washed once (without disturbance) with a single volume of ice-cold binding buffer. Pellets were resuspended in 20μl elution buffer (Qiagen TE + 150 mM NaCl) and incubated for 20 minutes at 80°C. Eluted DNA was amplified for subsequent selections in a two-step reaction using Phusion High-Fidelity DNA polymerase for 14-18 cycles, and purified using the MinElute PCR purification system (Qiagen).

**Spec-seq**

Binding reactions were prepared on ice in 12μl volumes containing 20ng FAM-labeled dsDNA library in 1X EMSA Buffer (25 mM Tris-Cl, 60 mM KCl, 140 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mg/ml BSA, 5% glycerol, 10 ng/μl salmon sperm DNA, pH 8.3 at 8°C).

Reactions were incubated 2 hours at 32°C with 25mM ammonium phosphoramidate or ammonium chloride for binding of phosphorylated and non-phosphorylated response regulators, respectively. Bound and unbound DNA pools were separated by native PAGE (8% polyacrylamide, 0.8x TBE [72 mM Tris-borate, 1 mM EDTA]) at 8°C. Gels were visualized on a Typhoon FLA 9500 (GE Healthcare) biomolecular imager. Bands containing bound and unbound DNA were excised, and DNA was extracted by the crush and soak method in gel diffusion buffer (.3 M sodium acetate, 1 mM EDTA). Eluted DNA was concentrated using the Qiaex II gel extraction kit (Qiagen).

## 2.6 Supplementary Figures

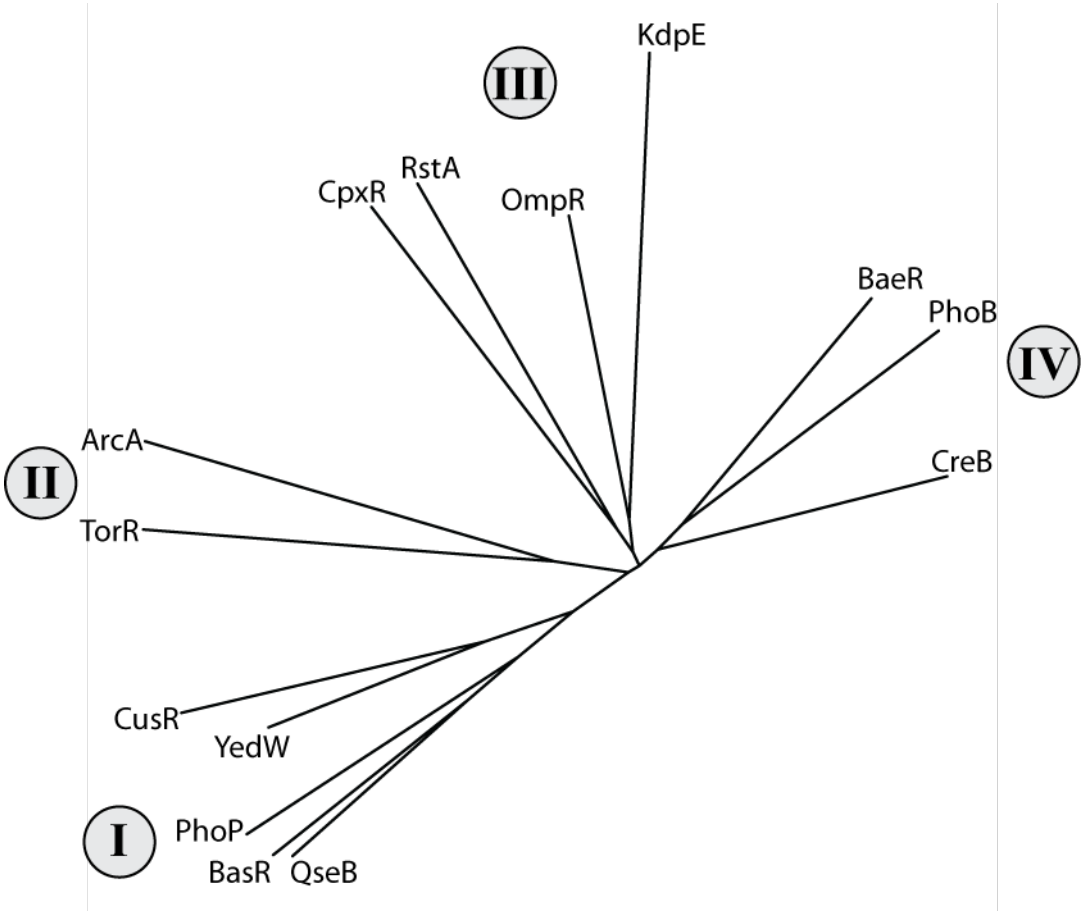


**Figure S2.1: Phylogenetic tree of OmpR family represented in *E. coli*.** A phylogenetic tree was constructed based on consensus representations of eRR ortholog groups. Categories I-IV correspond to Figure 2.1c.

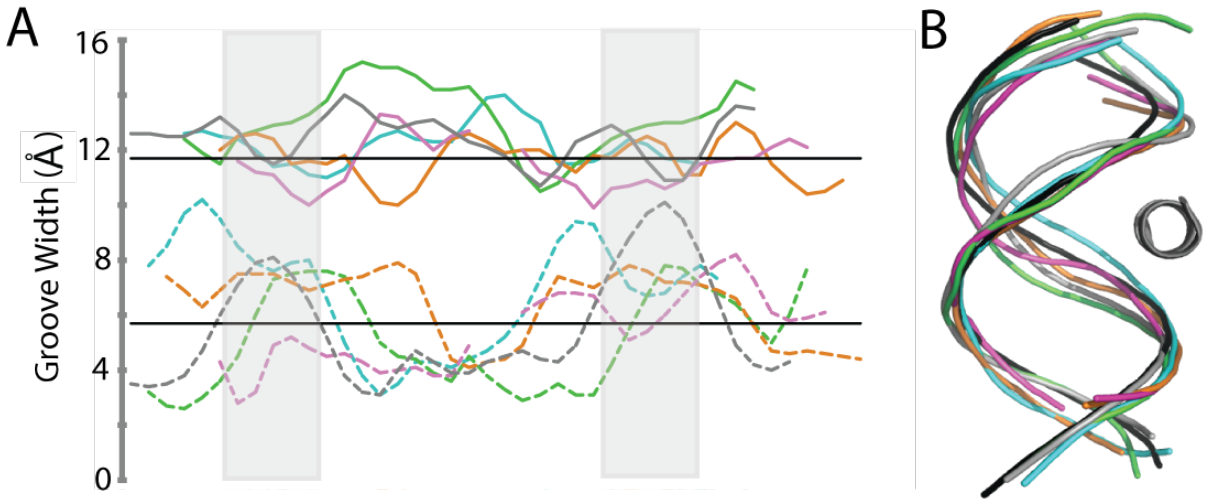


**Figure S2.2: DNA shape varies between OmpR family RRs** **A.** Groove widths for DNA major (solid lines) and minor (dashed lines) are displayed for DNA sequences in complex with dimeric RR: RstA (PDB code: 4NHJ, magenta). KdpE (PDB code: 4KNY, green), PhoB (PDB code: 1GXP, cyan), PhoP (PDB code: 5ED4, orange), PmrA (PDB code: 4S05, gray). DNA half-site sequences were aligned based on a structural overlay of bound RR monomers (a gap was introduced in the linker region for RstA to account for its reduced half-site spacing). Horizontal, black lines display the average B-form width parameters for the major (upper, 11.7Å) and minor (lower, 5.7Å) grooves. Gray boxes indicate the central 3 bases of each half-site. **B.** Recognition helices for 'upstream' RR monomers were aligned in Pymol (solely by the recognition helix) to normalize angle of groove entry, and the resultant phosphate backbone trajectories are represented according to the color scheme described in (**A**).

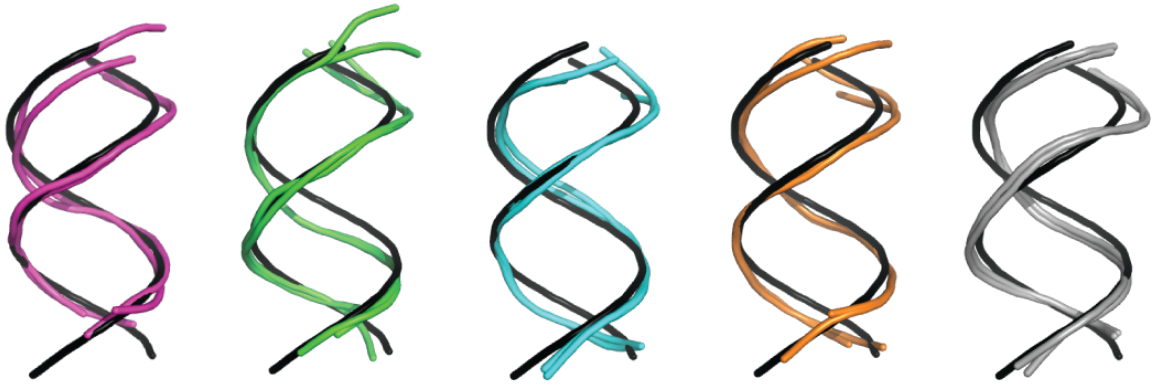

**Figure S2.3: Deviation of phosphate backbone induced by RR binding.** Each structure presents an overlay of bound half-sites (colored) and a DNA structure corresponding to ideal B-form parameters (black). Half-sites were aligned using the RH, as in Figure 1D. DNA structures were taken from: RstA (PDB code: 4NHJ, magenta). KdpE (PDB code: 4KNY, green), PhoB (PDB code: 1GXP, cyan), PhoP (PDB code: 5ED4, orange), PmrA (PDB code: 4S05, gray).

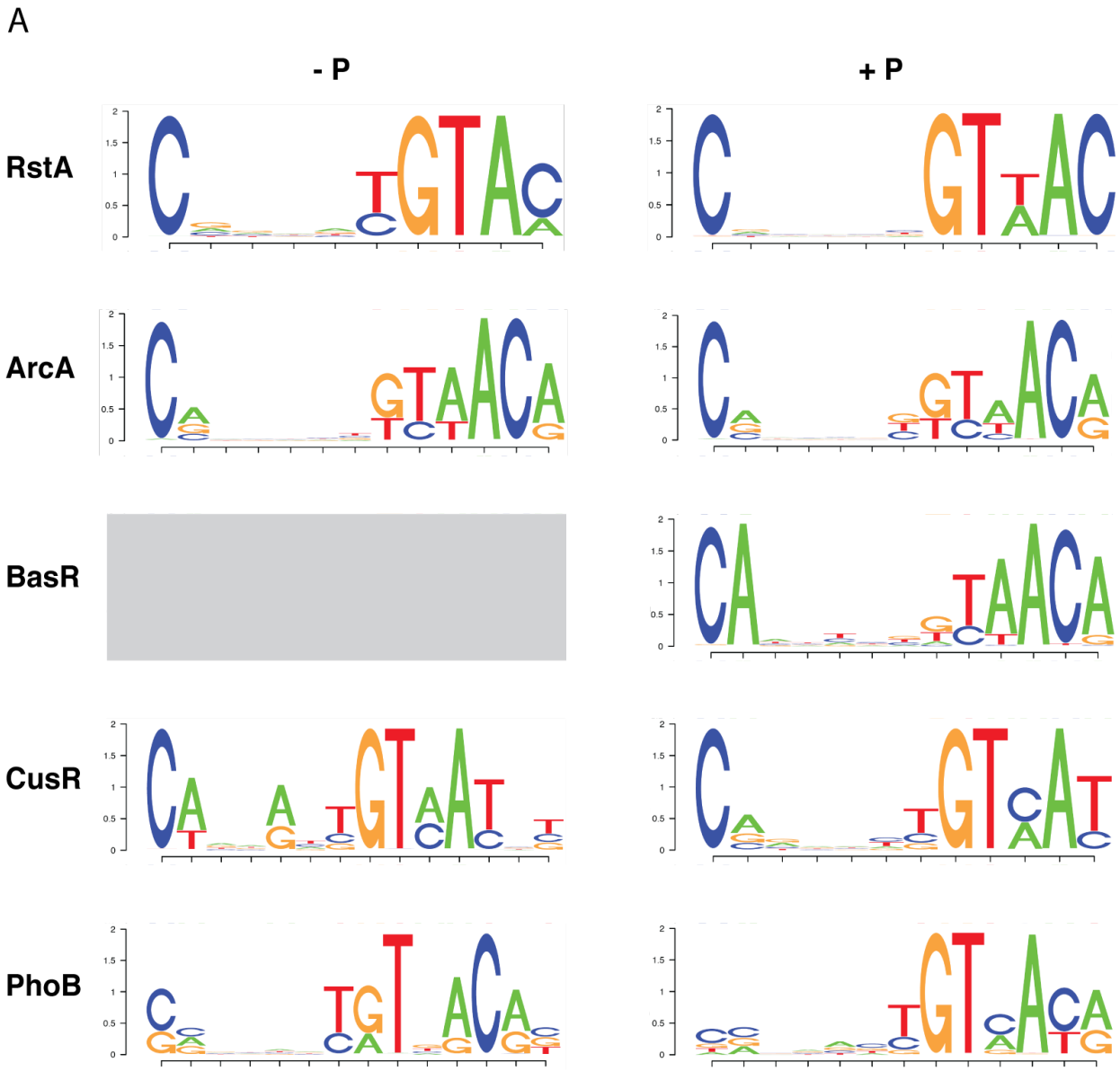


**Figure S2.4: Binding orientation revealed by base preferences adjacent to fixed half-site sequence.** Selected logos are the raw results of the *de novo* motif finding tool Bioprospector, oriented to place the synthetic anchor sequence ('AGGTAA') to the left of the sequence logo.

## 1.7 References


1. Magli MC, Largman C, Lawrence HJ (1997) Effects of HOX homeobox genes in blood cell differentiation. J Cell Physiol, 173, 168–177

2. Pineault KM, Wellik DM (2014) Hox genes and limb musculoskeletal development. Curr Osteoporos Rep, 12, 420–427

3. Mallo M, Wellik DM, Deschamps J (2010) Hox genes and regional patterning of the vertebrate body plan. Dev Biol, 344, 7–15

4. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell, 133, 1277–1289

5. Berger MF, Badis G, Gehrke AR et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell, 133, 1266–1276

6. Chu SW, Noyes MB, Christensen RG, Pierce BG, Zhu LJ, Weng Z, Stormo GD, Wolfe SA (2012) Exploring the DNA-recognition potential of homeodomains. Genome Res, 22, 1889–1898

7. Barrera LA, Vedenko A, Kurland JV et al. (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science, 351, 1450–1454

8. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol, 24, 1429–1435

9. Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol, 23, 988–994

10. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature, 527, 384–388

11. Chang YK, Srivastava Y, Hu C, Joyce A, Yang X, Zuo Z, Havranek JJ, Stormo GD, Jauch R (2017) Quantitative profiling of selective Sox/POU pairing on hundreds of sequences in parallel by Coop-seq. Nucleic Acids Res, 45, 832–845

12. Zuo Z, Chang Y, Stormo GD (2015) A quantitative understanding of lac repressor's binding specificity and flexibility. Quant Biol, 3, 69–80

13. Ishihama A, Shimada T, Yamazaki Y (2016) Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. Nucleic Acids Res, 44, 2058–2074

14. Stanton BC, Nielsen AA, Tamsir A, Clancy K, Peterson T, Voigt CA (2014) Genomic mining of prokaryotic repressors for orthogonal logic gates. Nat Chem Biol, 10, 99–105

15. Carrera J, Covert MW (2015) Why Build Whole-Cell Models. Trends Cell Biol, 25, 719–722

16. Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, Luisi BF, Sigler PB (1988) Crystal structure of trp repressor/operator complex at atomic resolution. Nature, 335, 321–329

17. Aeling KA, Steffen NR, Johnson M, Hatfield GW, Lathrop RH, Senear DF (2007) DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions. IEEE/ACM Trans Comput Biol Bioinform, 4, 117–125

18. Tong H, Mrázek J (2014) Investigating the interplay between nucleoid-associated proteins, DNA curvature, and CRISPR elements using comparative genomics. PLoS One, 9, e90940

19. Harlocker SL, Bergstrom L, Inouye M (1995) Tandem binding of six OmpR proteins to the ompF upstream regulatory sequence of Escherichia coli. J Biol Chem, 270, 26849–26856

20. Laub MT, Goulian M (2007) Specificity in two-component signal transduction pathways. Annu Rev Genet, 41, 121–145

21. Wang X, Gao H, Shen Y, Weinstock GM, Zhou J, Palzkill T (2008) A high-throughput percentage-of-binding strategy to measure binding energies in DNA-protein interactions: application to genome-scale site discovery. Nucleic Acids Res, 36, 4863–4871

22. Weirauch MT, Cote A, Norel R et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity. Nat Biotechnol, 31, 126–134

23. Inukai S, Kock KH, Bulyk ML (2017) Transcription factor-DNA binding: beyond binding site motifs. Curr Opin Genet Dev, 43, 110–119

24. Schneider TD (2002) Consensus sequence Zen. Appl Bioinformatics, 1, 111–119

25. Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci, 23, 109–113

26. Zhao Y, Ruan S, Pandey M, Stormo GD (2012) Improved models for transcription factor binding site identification using nonindependent interactions. Genetics, 191, 781–790

27. Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW (2016) DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. Cell Syst, 3, 278–286.e4

28. Yang L, Dror I, Zhou T, Mathelier A, Wasserman WW, Gordân R, Rohs R (2015) 15 TFBSshape: a motif database for DNA shape features of transcription factor binding sites. J Biomol Struct Dyn, 33 Suppl 1, 9

29. Bareket-Samish A, Cohen I, Haran TE (1997) Repressor assembly at trp binding sites is dependent on the identity of the intervening dinucleotide between the binding half sites. J Mol Biol, 267, 103–117

30. Mathelier A, Wasserman WW (2013) The next generation of transcription factor binding site prediction. PLoS Comput Biol, 9, e1003214

31. Qin Q, Feng J (2017) Imputation for transcription factor binding predictions based on deep learning. PLoS Comput Biol, 13, e1005403

32. Chen X, Hughes TR, Morris Q (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics, 23, i72–9

33. Nishino K, Honda T, Yamaguchi A (2005) Genome-wide analyses of Escherichia coli gene expression responsive to the BaeSR two-component regulatory system. J Bacteriol, 187, 1763–1772

34. Park DM, Akhtar MS, Ansari AZ, Landick R, Kiley PJ (2013) The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. PLoS Genet, 9, e1003839

35. Batchelor E, Walthers D, Kenney LJ, Goulian M (2005) The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. J Bacteriol, 187, 5723–5731

36. Feldheim YS, Zusman T, Speiser Y, Segal G (2016) The Legionella pneumophila CpxRA two-component regulatory system: new insights into CpxR's function as a dual regulator and its connection to the effectors regulatory network. Mol Microbiol, 99, 1059–1079

37. Flores-Valdez MA, Fernández-Mora M, Ares MÁ, Girón JA, Calva E, De la Cruz MÁ (2014) OmpR phosphorylation regulates ompS1 expression by differentially controlling the use of promoters. Microbiology, 160, 733–741

38. Shimada T, Takada H, Yamamoto K, Ishihama A (2015) Expanded roles of two-component response regulator OmpR in Escherichia coli: genomic SELEX search for novel regulation targets. Genes Cells, 20, 915–931

39. Harari O, Park SY, Huang H, Groisman EA, Zwir I (2010) Defining the plasticity of transcription factor binding sites by Deconstructing DNA consensus sequences: the PhoP-binding sites among gamma/enterobacteria. PLoS Comput Biol, 6, e1000862

40. He X, Wang S (2014) DNA consensus sequence motif for binding response regulator PhoP, a virulence regulator of Mycobacterium tuberculosis. Biochemistry, 53, 8008–8020

41. Perreault AA, Venters BJ (2016) The ChIP-exo Method: Identifying Protein-DNA Interactions with Near Base Pair Precision. J Vis Exp,

42. Yamamoto K, Ishihama A (2006) Characterization of copper-inducible promoters regulated by CpxA/CpxR in Escherichia coli. Biosci Biotechnol Biochem, 70, 1688–1695

43. Urano H, Umezawa Y, Yamamoto K, Ishihama A, Ogasawara H (2015) Cooperative regulation of the common target genes between $H_2O_2$-sensing YedVW and $Cu^{2+}$-sensing CusSR in Escherichia coli. Microbiology, 161, 729–738

44. Finn RD, Coggill P, Eberhardt RY et al. (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res, 44, D279–85

45. Narayanan A, Kumar S, Evrard AN, Paul LN, Yernool DA (2014) An asymmetric heterodomain interface stabilizes a response regulator-DNA complex. Nat Commun, 5, 3282

46. Itou H, Tanaka I (2001) The OmpR-family of proteins: insight into the tertiary structure and functions of two-component regulator proteins. Journal of biochemistry, 129, 343–350

47. Sandelin A, Wasserman WW (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. J Mol Biol, 338, 207–215

48. Blanco AG, Sola M, Gomis-Rüth FX, Coll M (2002) Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. Structure, 10, 701–713

49. Galperin MY (2010) Diversity of structure and function of response regulator output domains. Curr Opin Microbiol, 13, 150–159

50. De Wulf P, McGuire AM, Liu X, Lin EC (2002) Genome-wide profiling of promoter recognition by the two-component response regulator CpxR-P in Escherichia coli. J Biol Chem, 277, 26652–26661

51. Ogasawara H, Hasegawa A, Kanda E, Miki T, Yamamoto K, Ishihama A (2007) Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. J Bacteriol, 189, 4791–4799

52. Kato A, Tanabe H, Utsumi R (1999) Molecular characterization of the PhoP-PhoQ two-component system in Escherichia coli K-12: identification of extracellular Mg2+-responsive promoters. J Bacteriol, 181, 5516–5520

53. Yang C, Huang TW, Wen SY, Chang CY, Tsai SF, Wu WF, Chang CH (2012) Genome-wide PhoB binding and gene expression profiles reveal the hierarchical gene regulatory network of phosphate starvation in Escherichia coli. PLoS One, 7, e47314

54. Cariss SJ, Tayler AE, Avison MB (2008) Defining the growth conditions and promoter-proximal DNA sequences required for activation of gene expression by

CreBC in Escherichia coli. J Bacteriol, 190, 3930–3939

55. Motlhabi LM, Stormo GD (2011) Assessing the effects of symmetry on motif discovery and modeling. PLoS One, 6, e24908

56. Narayanan A, Paul LN, Tomar S, Patil DN, Kumar P, Yernool DA (2012) Structure-function studies of DNA binding domain of response regulator KdpE reveals equal affinity interactions at DNA half-sites. PLoS One, 7, e30102

57. Ogasawara H, Yamada K, Kori A, Yamamoto K, Ishihama A (2010) Regulation of the Escherichia coli csgD promoter: interplay between five transcription factors. Microbiology, 156, 2470–2483

58. Stormo GD, Zuo Z, Chang YK (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. Brief Funct Genomics, 14, 30–38

59. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat Biotechnol, 29, 480–483

60. Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc Natl Acad Sci U S A, 110, 12349–12354

61. Lou YC, Weng TH, Li YC, Kao YF, Lin WF, Peng HL, Chou SH, Hsiao CD, Chen C (2015) Structure and dynamics of polymyxin-resistance-associated response regulator PmrA in complex with promoter DNA. Nat Commun, 6, 8838

62. He X, Wang L, Wang S (2016) Structural basis of DNA sequence recognition by the response regulator PhoP in Mycobacterium tuberculosis. Scientific reports, 6, 24442

63. Ogasawara H, Shinohara S, Yamamoto K, Ishihama A (2012) Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. Microbiology, 158, 1482–1492

64. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif, 41, 207–234

# Chapter 3. Specificity-determining residues in DNA-binding by OmpR family TFs[2]

[2] *Ibid*

# 3.1 Context and motivation for research[3]

### 3.1.1 Structure-based modeling of protein:DNA interactions

Sequence-specific protein:DNA interactions are critical for proper cellular - functioning; consequently, there is substantial interest in predicting and/or reengineering their specificity. Amino acid changes in DNA-binding proteins can act as driving alterations that lead to disease [1–3] or evolutionary adaptation [4]. Changes in the affinities of transcription factors for mutated binding sites can also alter the occupancy and identity of bound proteins in gene regulatory regions, resulting in phenotypic consequences that may fuel evolutionary change [5–8]. Scientists have applied tools from structural biology to achieve an atomic-level understanding of binding mechanisms for a number of protein:DNA complexes [9]. The structures of these complexes have shed considerable light on the determinants of DNA sequence readout [10], effectively refuting the idea of a simple and general "code" for protein:DNA recognition [11] (**Figure 3.1.1a**), while at the same time enabling rational structure-guided engineering of DNA interaction specificity for certain families [12].

Structure-based computational approaches to binding prediction seek to rationalize observed specificity patterns and predict new interactions; this approach contrasts with more widely applied probabilistic models, which instead seek to model the downstream

[3] This section is adapted from the following published manuscript: Joyce, A.P., Zhang, C., Bradley, P., Havranek J.J. (2015) Structure-based Modeling or Protein:DNA Specificity. *Brief Funct Genomics*. 14(1):39-49.

effect of sequence recognition. Broadly speaking, structure-based approaches proceed by constructing three-dimensional models of protein:DNA complexes (**Figure 3.1.1b**) and deriving estimates of binding affinity and/or specificity from them. Structure-based approaches vary in their degree of computational and physical rigor, ranging from relatively low-resolution, statistically-based potentials to all-atom molecular dynamics simulation. In comparison, non-structural approaches are often far less computationally intensive, require little or no knowledge of physical interactions, and frequently yield models of equal or greater quality than state-of-the-art structure-based calculations when provided with sufficient experimental binding data for training.

Structural modeling of the protein:DNA interface can provide substantial information beyond predictions of binding specificity. First, the physical forces that govern protein:DNA interactions are generalizable to any protein:DNA complex; therefore advances in structure-based modeling can have immediate and significant impact on our ability to model thousands of individual genomic interactions. Second, structural models of protein:DNA complexes are highly useful to model secondary binding events, such as interactions with a protein cofactor or allosteric regulator. Third, structural models facilitate the *in silico* exploration of mutations or covalent modifications to protein or DNA (e.g. CpG methylation, DNA damage, or protein phosphorylation). Lastly, energy functions and sampling methods developed for binding site prediction have the potential to drive innovation in the engineering and design of genomic tools, such as synthetic transcription factors and site-specific nucleases [13–15] .

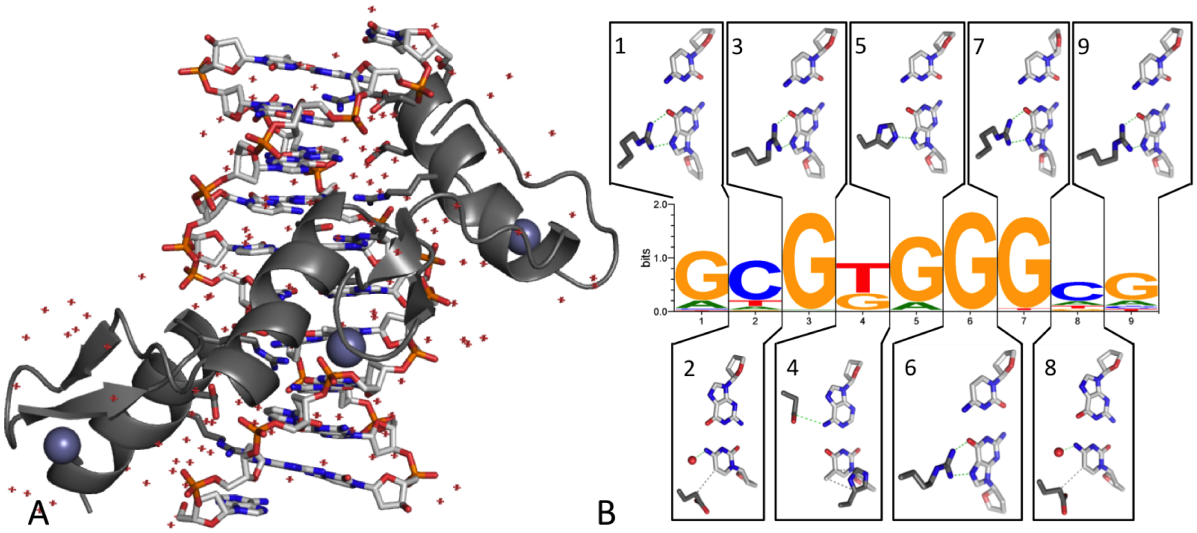


**Figure 3.1.1.** Atomically detailed structures of protein:DNA complexes illuminate the molecular mechanisms underlying sequence specific binding: the overall structure (with protein shown in cartoon representation, the DNA in sticks, zinc ions as spheres, and crystal waters as red crosses) (A) and per-position specificity-determining interactions (B) seen in the high-resolution crystal structure of the C2H2 zinc finger Zif268 bound to a high-affinity target site (PDB ID 1aay [16]; PWM data downloaded from the Uniprobe database [17]; structure figures generated in PyMOL [18]).

### 3.1.2 Recent improvements in statistical potentials

The accuracy of structure-based binding predictions depends critically on the quality of the potential energy functions used to estimate binding affinities from modeled complexes. The potential energy functions that have been used for this purpose can be roughly classified as being either physics- or knowledge-based. The functional form of the energy terms in physics-based potentials is derived from a physicochemical model of the underlying interactions, and as a result these potentials can be quite sensitive to the

atomic coordinates: small changes in atomic position can lead to large changes in computed energy due to steric or electrostatic clashes. In knowledge-based statistical potentials, on the other hand, the interaction potentials are derived from experimentally determined protein:DNA structural information. The probabilities of observing different kinds of interactions in crystal structures are calculated and converted into potential energies, for example by using the inverse Boltzmann approach. Statistical potentials can model any previously observed behavior even if the underlying physical phenomena are poorly understood. However, they cannot predict atomic interaction patterns absent from the training set of available protein:DNA structures [19; 20]. The resolution of statistical potentials can vary from atom-level to residue-level; in general they do not have the sensitivity of molecular mechanics potentials.

The moderate spatial resolution of statistical potentials makes them a good match for scoring the approximate structural models generated by homology modeling or by the docking of unbound structures (**Figure 3.1.2**). In contrast, molecular mechanics potentials may be less forgiving in these cases, due to the steric clashes often present in these complexes. Chen et al. used structural alignment to generate synthetic protein:DNA complexes from structures of unbound proteins, and applied a statistical potential to predict position weight matrices (PWMs) for these proteins [21]. Although PWMs generated using this approach were less accurate than those generated from native complexes, results were comparable to those obtained from complexes generated by docking, and were better than those obtained from homologous complexes generated by

bound structural templates from the same protein family. Their analysis demonstrated the utility of statistical potentials for predicting PWMs given approximate models, and also indicated that correctly capturing the conformational changes of proteins upon binding DNA will be important for future improvements. A number of alternate approaches exist for generating synthetic complexes, and it remains to be seen whether they yield improved models for predicting protein:DNA specificity [22].

Atomic resolution is usually preferred when using statistical potentials to predict protein:DNA binding specificity, yet atomistically detailed statistical potentials have very large numbers of parameters which can make them challenging to train robustly (for example, a pairwise atomic potential with 30 atom types and 10 distance bins has 4,650 free parameters). Recently, improvements have been made to train the potentials more efficiently. Xu et al. developed an energy function that was trained to include the target structure templates themselves in recognizing transcription factor binding sites [23]. This development led to increased prediction accuracy and robustness compared to their previous potential, vcFIRE [24]. Their method also out-performed sequence-based approaches in prediction accuracy in cases for which limited experimental data was available. In another approach, the training incorporated experimentally determined PWMs. Traditionally, statistical potentials count the number of times a given interaction is observed across protein:DNA complexes and assume that each complex is equally likely. However, the occurrence frequencies can also be weighted proportionally to the binding affinity of the protein for different DNA sequences. AlQuraishi & McAdams

trained their potentials by weighting DNA sequences differently according to their experimental probability of occurrence specified by their corresponding PWMs [25]. Although this approach did not significantly improve PWM predictions, it was a novel step in the long-term goal of combining structural data with biochemical data for protein:DNA binding site prediction.

In contrast to atomistic potentials, coarse-grained residue-level potentials do not generally have sufficient resolution to make predictions for PWMs. However, they are well-suited for generating protein:DNA complexes by docking unbound structures. Although residue-level potentials have far fewer parameters then atom-level potentials and require less computing power, docking with large decoy sets can still be computationally intensive. Parisien et al. applied machine-learning techniques to reduce the number of parameters required in their residue-level potential function to fifteen [26]. Their rigid body docking protocol performed well at rebuilding native protein:DNA contacts for both bound and unbound structures, although it was still a challenge to achieve RMSDs below 5 Å when using unbound structures as the starting point. Besides reducing parameters, efforts have been made to make statistical potentials more accurate. Most statistical potentials are distance-based and thus may benefit from including an angular term. Takeda *et al.* derived a novel orientation-dependent residue-level potential for protein:DNA docking [27]. Their potential performed significantly better than their previous multi-body potential in docking accuracy. Its binding affinity prediction was also greatly improved and was on par with some

atom-level statistical potentials, though it was still less accurate than others (e.g, vcFIRE). [24]

Finally, because statistical potentials usually require much less computational power than physics-based potentials, they can easily be adapted to run on web servers. Three web servers for predicting PWMs using protein:DNA complexes have been constructed in the past few years, making these statistical potentials easily accessible to researchers without a computational background: 3D-footprint [28], 3DTF [29], and PiDNA [30].

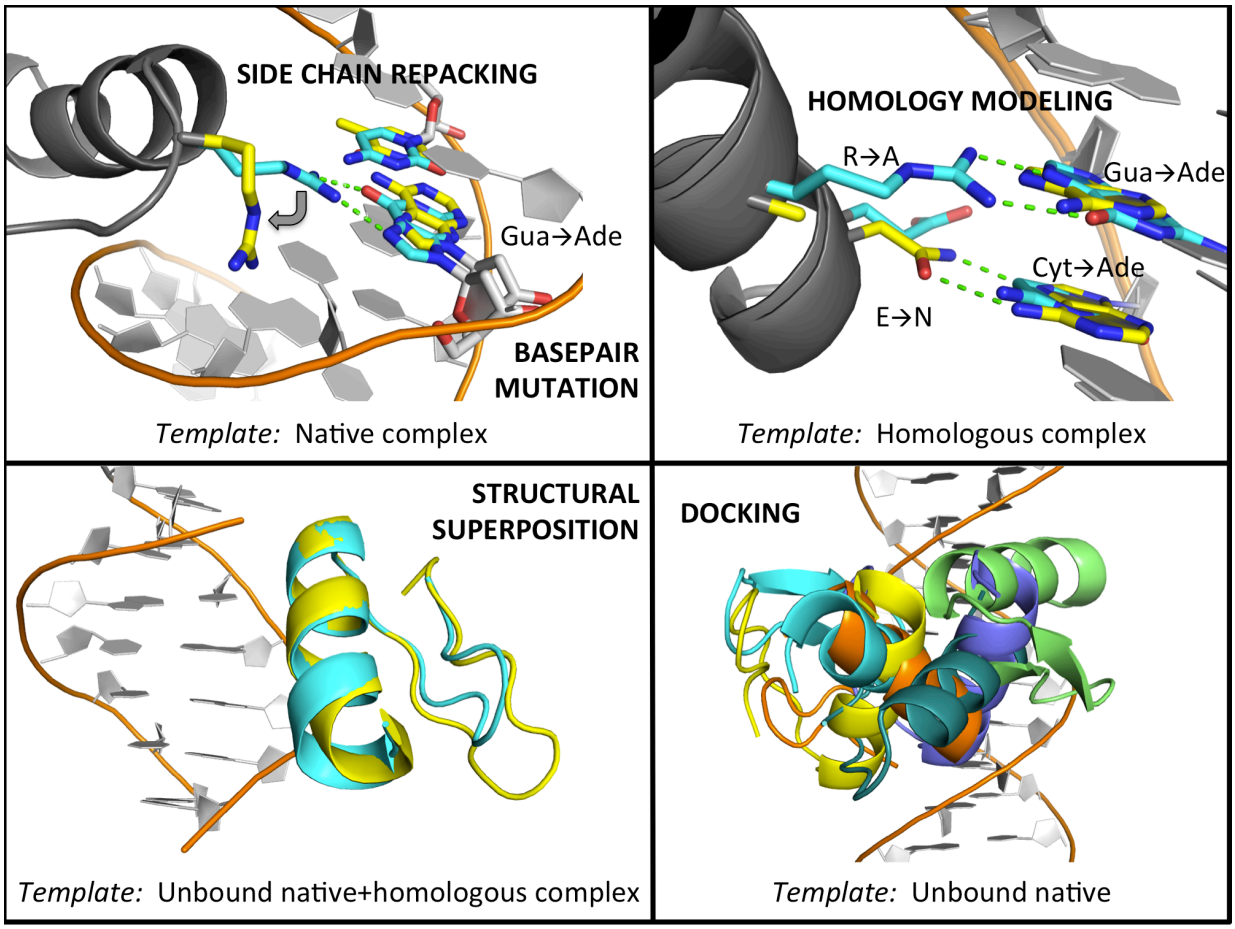


**Figure 3.1.2.** Modeling protein:DNA complexes. The choice of protocol depends on the structural 'template' available for constructing the model. If a bound structure is available for the protein of interest ('Native complex', top left), the modeling needed for binding predictions involves primarily base pair mutations ('Gua→Ade': template in cyan and model in yellow) and side chain rearrangements (grey arrow). Building a model using a homologous complex as a template will require protein ('R→A', 'E→N') as well as base pair mutations, and may require protein and DNA backbone relaxation. If the unbound structure of the native protein is known, a DNA-bound model can be constructed by superimposing this unbound structure onto the structure of a homologous factor in a bound structure (bottom left), or by de novo 'docking' onto DNA (bottom right, multiple candidate docked conformations shown).

### 2.1.3 Modeling water in protein-DNA interfaces

Modeling the role of water is likely to be more important for protein-DNA interfaces than for other macromolecular calculations. Biochemical and structural data indicate that water-mediated interactions play a key role in protein-DNA recognition (**Figure 3.1.3a**) [31; 32]. This is in contrast to the modeling tasks of protein folding and docking, which have achieved notable successes without incorporating explicit water molecules [33; 34]. In addition, the polyanionic nature of nucleic acids suggests that electrostatics, also commonly omitted from protein modeling, will figure prominently in any energetic description of protein:DNA complexes. Water plays an important role in quantitative models for electrostatic phenomena by virtue of its high dielectric constant. Finally, protein:DNA interfaces possess many polar and charged amino acids that are sequestered from bulk solvent, yet must still satisfy their hydrogen bonding potential. Water can serve this role by filling voids in the interface and providing hydrogen bond donors or acceptors for polar groups in both the protein and DNA.

The effects of water upon the energetics of a protein-DNA complex can be treated at several levels of detail. At one extreme is the complete neglect of explicit water molecules, perhaps partially compensated by the inclusion of an implicit solvation potential [35]. In one study, the ability of water to attenuate hydrogen bonds, but not to participate in them, was considered [36]. At the other end of the spectrum is the explicit treatment of water molecules that fully solvate a macromolecule or complex using molecular mechanics [37–39]. Computational protocols also differ in where and how

explicit water molecules are introduced into a model. For instance, water networks have been constructed en masse, with the goal of optimizing hydrogen bonding across an entire interface [40]. Water molecules have also been attached to polar groups in amino or nucleic acids at optimal geometries for hydrogen bonding, giving rise to the 'solvated rotamer' strategy [41]. In some approaches, the locations of water molecules are determined simultaneously along with the conformational sampling that optimizes the protein:DNA interface [42; 43]. The specific choices of how water molecules are modeled, and where and when they enter into the calculation, are based on trade-offs between the accuracy of the physical potential used, the scale of conformational sampling that is to be considered, and the computational resources that are available.

Unsurprisingly, given the extra computational requirements and significant uncertainty in the optimal approach for modeling water in protein:DNA interfaces, few studies include water in the calculation of protein-DNA binding specificity. Nevertheless, some conclusions can be drawn regarding the impact of including explicit water. Van Dijk et al. incorporated explicit water into the protein:DNA docking capabilities of HADDOCK[43]. While they were not explicitly calculating DNA-binding preferences, the methods they describe are readily transferable to the protein:DNA homology modeling problem. Water molecules were first placed on unbound models for both protein and DNA based on the results of molecular dynamics simulations. During a subsequent docking step, water was removed or added from the developing complex using a Monte Carlo approach. The inclusion of water molecules led to modest but

significant improvements in the docked complex geometries. In particular, they were able to recover specific water-mediated hydrogen bonds in the Engrailed homeodomain:DNA interface [44]. They found most consistent success in those cases where the bound and unbound conformations of the protein were very similar, which is expected for the homology modeling calculations required to estimate PWMs.

Li and Bradley directly studied the effect of explicit water molecules on predicting protein:DNA recognition specificity [42]. Their method considered water molecules only at the consensus minor and major groove locations that have been determined from crystallographic studies [45]. Water occupancy at these locations was allowed to vary during the course of the structural optimization. Similar to Van Dijk et.al., they observed limited but significant improvement over a large test set of protein:DNA complexes. Notably, the inclusion of explicit water led to improvements in the description of water-mediated hydrogen bonds that are known to be important for the specificity of the EcoRI restriction enzyme (**Figure 3.1.3b**). Interestingly, neglecting explicit water molecules yielded a specificity profile consistent with EcoRI 'star activity'. Star activity has been linked experimentally to the release of bound interfacial waters thought to participate in the formation of the cognate protein:DNA complex [46]. Of particular interest for the calculation of PWMs, their method was able to predict correctly that in the case of one experimentally determined protein:DNA complex, a higher affinity DNA sequence than the one in the crystal structure could be found. This demonstrates that it is possible for structure-based calculations to use an experimental structure as a

homology modeling template to accurately describe water-mediated protein:DNA interactions not found in the original complex.

In summary, the consideration of explicit water molecules can lead to a more faithful description of protein:DNA recognition specificity. The improvements have been found to be clear, if modest in effect [42; 43]. However, in certain cases key water-mediated interactions appear to be crucial for describing specificity, and approaches that neglect explicit water may not generate useful PWMs. In the near future, we are likely to witness improvements in the placement and scoring of water molecules and their interactions, as well as in the computational efficiency of calculating these effects.

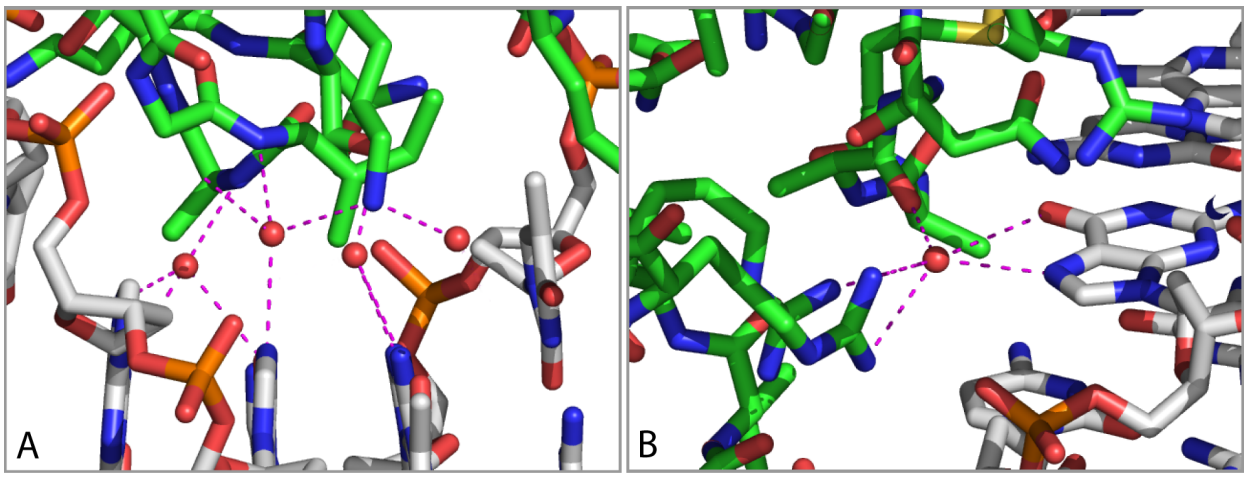


**Figure 3.1.3.** Water molecules at the protein:DNA interface participate in hydrogen bonding networks. **A.** The trp repressor protein achieves recognition of its operator sequence through multiple water mediated contacts, involving both protein sidechain and mainchain atoms. **B.** The EcoRI restriction enzyme interacts with its cognate cleavage site with both and water-mediated contacts. Failure to model water molecules explicitly leads to a relaxed DNA specificity profile reminiscent of 'star activity', which has been attributed to the loss of bound interfacial water.

### 3.1.4 Flexibility at the protein:DNA interface

The Protein Data Bank contains representative structures for the majority of known DNA-binding protein families in complex with DNA (~3000 total structures, with substantial redundancy) [47], and predictions based on these homologous "template" structures have the potential to expand our knowledge of sequence-specificity to thousands of uncharacterized proteins. However, homologous template complexes present a single, static conformation that is unique to the crystallized protein and DNA

molecules, and sequence changes to either partner often result in steric clashes or, conversely, novel low-energy states. In these cases, it is necessary to sample and evaluate any deviations from template coordinates within a set of allowable conformations reflecting the total "flexibility" of the protein backbone, amino acid side chains, bases or base pairs, and the sugar-phosphate backbone. Physically, flexibility is integral to the process of protein:DNA recognition. Within a single protein:DNA complex, both inter- and intra-molecular contacts vary according to DNA sequence, and individual side-chains freely adopt alternative conformations in specific and non-specific binding modes [48]. Comparison of protein:DNA interfaces in the free and DNA-bound states has revealed greater intrinsic structural variation in protein:DNA interfaces than other surface areas [49–51]. Additionally, crystallographic studies have shown that extensive contact with proteins can induce significant deviation from the canonical B-form DNA backbone and standard base pair geometry [52]. Collectively, these findings demonstrate that both protein and DNA can exhibit conformational changes relative to their unbound structures.

The incorporation and conformational sampling of new side chains is essential for the prediction of sequence specificity using homologous proteins or unbound structures as templates. Typically, this search is discretized using libraries of torsionally rotamerized side chains [53; 54]. Using Monte Carlo optimization of rotamer selection, Havranek et al. demonstrated recovery of both identity and native conformation for DNA-contacting residues in the presence of DNA, with accuracy comparable to modeling of monomeric proteins [54]. This model was further extended to include a simplified representation of

DNA strain; however, compared to full conformational relaxation of both protein side chains and DNA in a single native complex, a “static model” allowing neither side chain nor DNA motion reproduced experimental PWMs more accurately in most cases [35]. In this study, conformational sampling was least accurate when water molecules were omitted from the structural templates. Parisien and colleagues also found that side chain reorganization in unbound structures significantly reduced the recovery of native protein:DNA contacts in 47 protein:DNA structures using the rigid-body docking tool FTDock [26]. Together, these studies illustrate that additional degrees of freedom in interfacial side chains, in the absence of appropriate constraints can reduce the accuracy of structural and specificity prediction.

Currently, most homology-based predictions of protein:DNA specificity rely on the assumption that the target and template structures possesses sufficiently similar, if not identical, backbone coordinates. Violations of this assumption can have dramatic functional consequences [55], and, given that increased backbone flexibility has been commonly observed in protein:DNA interfaces [49; 50], this assumption is likely to be inappropriate for modeling many DNA-binding proteins (**Figure 3.1.4a**). Moreover, polar amino acids with long side chains, which are enriched at protein:DNA interfaces, commonly form distance- and orientation-constrained contacts with specific DNA bases, and will experience large deviations in torsional sampling space following subtle backbone movements [56]. Correct backbone placement is therefore essential for an accurate depiction of protein:DNA contacts. Using a novel fragment insertion protocol to

improve backbone torsional sampling, Yanover and Bradley generated homology-models of C2H2 zinc fingers that recapitulated near-native docking conformations, base-specific contacts, and experimentally-generated models of sequence specificity [57]. Havranek and Baker introduced structure-guided backbone flexibility using a motif library of observed side chain:base contacts, termed "inverse rotamers" [56]. In this approach, after incorporating a motif into the DNA template, the adjacent protein backbone was allowed to sample nearby positions; changes were accepted if the backbone could accommodate the motif in an energetically favorable conformation.

The protein-bound DNA backbone frequently displays both local and global deformation from the standard B-form helix, and resultant changes in the positions of phosphate atoms and base parameters can substantially impact binding conformation and sequence recognition (**Fig. 3.1.4b**). Siggers and Honig developed a torsional sampling approach in which mutated base pairs were introduced with co-planarity to the template bases and subsequently conformationally diversified by means of small, compensating rotations about four DNA backbone torsion angles [22; 58]. This increase in DNA flexibility substantially improved specificity prediction, especially for templates with low similarity to the target structure. Yanover and Bradley introduced conformational diversity into both protein and DNA backbones simultaneously by insertion of fragments from multiple template structures of protein:DNA complexes [57]; the DNA backbone sampling procedure of Siggers and Honig was then applied to minimize the local impact of fragment insertion [22]. Full-atom simulation of bound and unbound DNA using

molecular mechanics force fields is another powerful, though computationally intensive approach to modeling DNA deformation [59]. Steady gains in computing power and optimization of nucleic acid force field parameters have improved the speed and accuracy of MD-based methods [60; 61], but the combinatorial challenge of minimizing all possible DNA sequences that a protein may bind has been a major barrier to the application of MD to specificity prediction. Using the ADAPT methodology, Deremble and colleagues developed a technique to sub-divide the DNA interface into overlapping pentanucleotide segments, which are independently evaluated and summed to yield the total, sequence-dependent energy of the protein:DNA complex [62]. The substantial reduction in computing time permitted the simultaneous conformational relaxation of both protein and DNA, and achieved accurate structural predictions for proteins bound to highly deformed DNA [63].

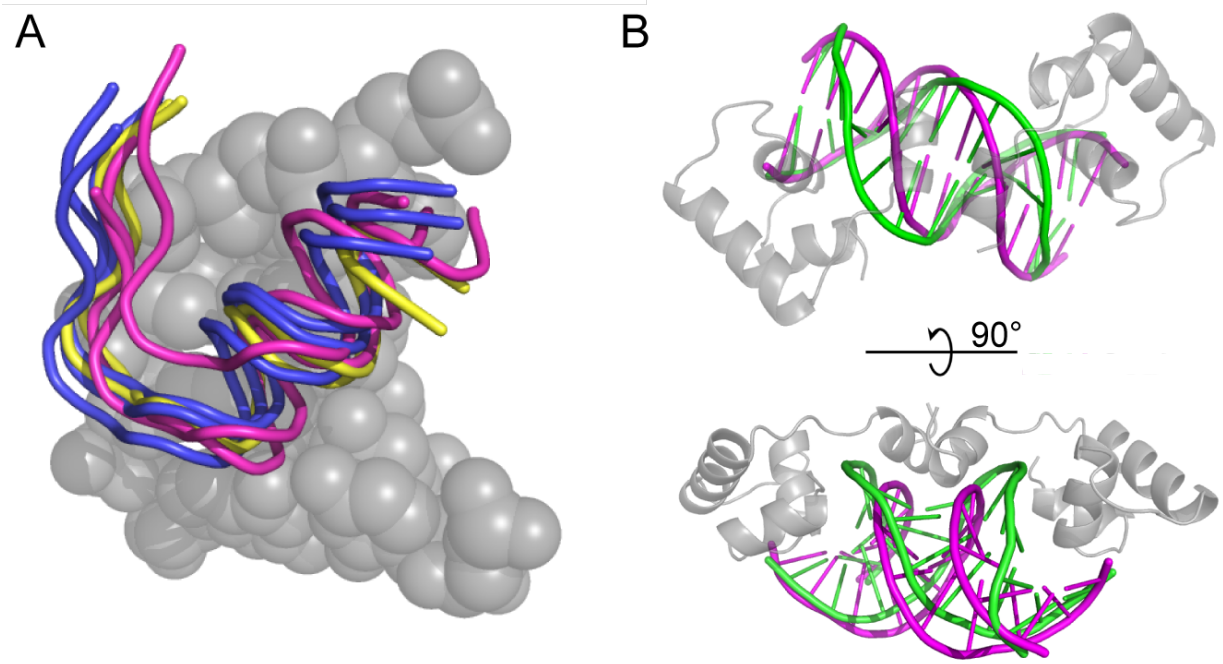


**Figure 3.1.4.** Protein and DNA adopt diverse backbone conformations and orientations in complex. **A.** Variation in triplet-docking orientation of the protein backbone for eight zinc finger domains from Zif268 (1AAY, blue), Tramtrack (2DRP, yellow) and TFIIIA (1TF6, magenta) **B.** The recognition element of PurR undergoes substantial deformation from the unbound state (1HQ7, magenta) upon protein binding in the minor groove (1QPZ, gray/green) (Left panel: top view. Right panel: side view)

### 3.1.5 Evaluating improvements in protein:DNA modeling

A wealth of experimental data on protein:DNA interactions is now available for training and testing structure-based approaches. High-throughput *in vitro* [64–68] and *in vivo* [69] experimental methods have been developed that can produce rich binding affinity profiles for multiple DNA binding proteins relatively rapidly. These methods enable the mapping of affinity landscapes for individual DNA binding proteins with

unprecedented depth and resolution, facilitating the detection of subtle binding features such as secondary motifs [70], correlations between target site positions [68], higher-order binding interactions [65], and DNA-shape mediated readout [71]. In addition, these methods have been applied to survey large families of homologous factors, providing valuable data on the mapping between protein sequence and DNA binding specificity within families [72; 73].

The standard approach to benchmarking a structure-based algorithm has been to reduce the reference experimental dataset to a position weight matrix (PWM), to similarly condense the output of the prediction algorithm, and then to assess the agreement between the two PWMs by aligning them and scoring the strength of the alignment using one of a number of established PWM comparison metrics [74; 75]. This approach ignores the richness of deep binding affinity datasets , and it also overlooks the potential of structure-based approaches to rationalize exactly those higher-order effects that are neglected by the PWM representation. Historically, it has been a challenge to recapitulate even the first order, position-independent binding profile, and this remains a valuable assessment for benchmarking, particularly in template-based approaches. We anticipate, however, that as structural modeling methods continue to improve it will be increasingly informative to directly compare predicted and experimentally measured relative affinities for large sets of full-length target site sequences (rather than PWM columns or consensus sequences), particularly for target proteins with a bound, high-resolution crystal structure. This comparison should be particularly enlightening when applied across families for

which multiple experimental binding profiles and co-crystal structures are available, giving insight into the origins of binding specificity divergence among related proteins.

### 3.1.6 Prospects for the future of structure-based modeling

The structure-based prediction of protein:DNA specificity will be affected by several ongoing trends. First, we can expect that high-throughput experimental techniques will continue to provide a wealth of protein:DNA affinities useful in both training and testing the robustness of structure-based prediction algorithms. Second, the number of experimentally determined crystal structures of protein:DNA complexes will continue to grow. The availability of examples of additional structural families will expand the number of DNA-binding proteins that are amenable to structural modeling of specificity, and increased DNA sequence coverage for similar or identical proteins will provide additional examples of sequence-specific molecular contacts. The availability of complexes with different DNA sequence specificities, altered binding modes, and diversified backbone conformations will provide more appropriate starting templates for homology modeling, lessening the need to incorporate protein or DNA flexibility in modeling calculations. Examples of novel protein:DNA complexes will also add to the set of training data for statistical potentials. Finally, the steady increase in computing power will facilitate improvements in scoring potentials and conformational sampling previously described. Furthermore, the nature of specificity calculations (involving evaluations of a protein bound to multiple DNA sequences) make them an ideal fit for the

parallel architectures increasingly available to individual researchers at reasonable costs.

### 3.1.7 Definition and detection of specificity-determining residues

While the quantitative prediction of protein:DNA specificity from structure-based calculations remains challenging, numerous statistical methods have been developed to identify positions in DNA-binding proteins with a role in specificity determination. For this purpose, 'specificity-determining residues' (SDRs) occur at positions that vary within protein families to diversify specific biochemical properties, such as ligand recognition or protein interactions; this is in contrast to highly conserved residues with an invariant role in these processes, such as (for TFs) establishing a global mode of DNA sequence recognition. A practical evolutionary framework for SDR identification was adapted by Mirny and Gelfand [76], relying on the assumption that SDRs would be conserved among orthologs, or proteins with similar function generated through a speciation event, and variable within paralogs, distinct proteins that stem from an duplication event and subsequently evolved divergent function through a period of relaxed selection. Presumably, over evolutionary scales, SDRs will leave a trace detectable through the analysis of residue covariation in large, well-sampled protein alignments. Applying a statistical metric based on mutual information, Mirny and Gelfand successfully identified SDRs within the LacI/PurR family, revealing two distinct clusters localized to the DNA-contacting surface as well as a region responsible both for binding chemical ligands and mediating homo-dimerization (a necessary step in its DNA-binding mechanism). In a

more recent analysis, Sloutsky and Naegle [77] developed a method to identify 'partial SDRs' (by their terminology, SDPs), which are residue positions that are 'heterogeneously conserved' within distinct ortholog sets known as specificity groupings. This approach identified numerous putative SDRs localized to protein and DNA interfaces with greater sensitivity than existing tools, but experimental validation is necessary to explore the biochemical role of these residues. This and related techniques do potentially advance a concept highly relevant to the design of novel DNA-binding proteins: that the ultimate effect of SDRs can be highly scaffold-dependent.

SDR identification using evolutionary criteria is reliant on the selection and multiple alignment of protein sequence across many species, and great care must be taken to minimize faulty assumptions and data artefacts. Algorithms for the multiple alignment of protein sequences can be highly sensitive to various features of the dataset, including the number of sequences [78] as well as the upper and lower sequence similarity thresholds for inclusion of a particular protein in a set of homologous sequences [79; 80]. For example, phylogenetically distantly proteins may have diverged so substantially in sequence and biochemical activity as to contribute little information to a sequence-based alignment. This is especially problematic in the case of proteins with regions under low purifying selection, such as variable-length domain linkers and loops connecting elements of secondary structure. Phylogeny-aware alignment algorithms are specifically designed to incorporate the evolutionary history of protein sequences, potentially reducing the impact of substitution and gaps in poorly conserved regions, but these come at an

increased computational cost [81]. Incorporating large groups of highly similar sequences (i.e., low sampling depth) unavoidably introduces systematic bias toward the structural and functional properties of sub-groups [82]. In regard to the latter case, biological sequence repositories contain well-known imbalances due to the attention paid to certain clinically- or experimentally-common species and genera, but attempts have been made to reduce data redundancy [83]. Overall, a set of 'best practices' to guide sequence selection and increase alignment accuracy may include, but are certainly not limited to: incorporation of alternate sources of information (e.g., a well-curated structures [84]), applying multiple algorithms [85], conducting hierarchical subalignments [77], or purging sequence outliers [86].

In addition to technical artefacts, the success of different approaches to SDR identification depends critically upon a number of biological and evolutionary assumptions that are especially relevant to two-component signaling systems. First, two-component systems vary widely between species in both number - known OmpR homologues per bacterial genome ranges between 0 and 41[79] - and pathway combinatorics, likely facilitated by their highly modular construction. Because a single pathway minimally includes a co-operonic response regulator and histidine kinase, one duplication event can generate a fully functionally redundant pathway. Paralogous systems are thus, with significant frequency, able to explore many novel evolutionary trajectories with relaxed selective constraint. Transcriptional regulatory networks under the control of two-component systems also evolve rapidly within and between lineages,

as different species are challenged by unique environmental conditions [87; 88]. Importantly, these characteristics reduce confidence in the assumption that sequence-similarity between homologous response regulators is a strong indicator of functional conservation, as, in any given species, the gain or loss of a single system can dramatically alter the evolutionary constraint on orthogonal pathways and their subordinate gene regulatory networks. In a telling example, the *E. coli* PhoB and OmpR share 37% sequence identity and govern completely distinct transcriptional programs; similarly, the evolutionarily distant, functionally divergent *Bacillus subtilis* Spo0F and *Caulobacter crescentus* DivK are more similar than known functionally conserved orthologs [79]. As previously observed, purely sequence-based annotation and ortholog assignment of response regulators must be undertaken with caution.

## 3.2 Introduction

Precise determination of protein:DNA specificity and crucial SDRs remains a challenging, unsolved problem in regulatory biology, hampering our ability to design proteins, predict disease-causing mutations, and decode phenotypic variation. Structure-based modeling, although entering a promising time, currently lacks the ability to describe the physicochemical complexity of the protein:DNA interface [89]. Phylogenetic and sequence-based prediction algorithms require careful sequence curation and accurate, large-scale alignments that may be challenging for certain systems; in contrast, a biophysical 'protein-DNA code' is theoretically applicable to any solved or modeled structure. Although technically possible, the complete experimental characterization of any putative SDRs is resource-intensive, and would require the development of novel high-throughput analytical techniques. Integrative solutions combining the strength of these three approaches have the potential to greatly accelerate SDR detection and validation.

Several DNA-binding attributes of OmpR family TFs complicate the identification of residues important for sequence recognition. First, there are ~20 residue positions (per monomer) oriented toward the DNA helix via the phosphate backbone and major or minor grooves. Many exhibit preferences *in vivo* for multi-meric binding [90] and/or recognition of curved DNA [91], so a role for large-scale changes in shape or geometry may play an outsized role in target recognition. As demonstrated in the previous chapter, half-site specificity may also be asymmetric, suggesting that residues play multiple roles

depending on their position in the larger context of the DNA-bound homodimer (or higher-order complexes). In this chapter, I present a simple approach to predict, refine, and characterize putative DNA-binding SDRs in the OmpR family. Specific attention was paid to minimize the application of arbitrary statistical thresholds and complex evolutionary models; rather, a small amount of biological and structural data very effectively prioritized SDRs with strong indicators of function. Validated SDRs provide evidence for a generalized structural model of canonical wHTH:DNA interactions, and the overall prediction-validation strategy is potentially applicable to other families of DNA-binding response regulators.

## 3.3 Results

### 3.3.1 Specificity-centric MI subnetwork distinguishes the RH as a SDR hub

Capra and coworkers [92] previously conducted an analysis of mutual information (MI) between cognate HK:RR pairings to predict interfacial SDR pairs responsible for selective phosphotransfer, based on the assumption that such pairs would necessarily co-evolve. We hypothesized that this subset of receiver-SDRs would additionally co-vary with determinants of DNA-binding specificity due to the convergent evolutionary pressure to maintain overall TCSP-specificity (**Figure 3.3.1a**). As a first step toward building this broader 'specificity-centric' network, we generated a MI network based on our alignment of ~2000 OmpR family protein; in this approach, the cumulative MI (cMI) for each position represents MI contributions summed over all possible pairings over the full protein length, and is thus expected to increase at positions with many significant pairings (as might be expected at the DNA interface) [51]. Residues with high cMI localized to the HK:RR and DBD:DNA interfaces; however, high cMI values were also observed for a cluster of poorly-conserved, core-facing residues in the β1-4 region of the DBD (**Figure 3.3.1b**). We next isolated the first-order, cross-domain contact network centered on previously validated receiver-SDRs [46], and observed a strong enrichment for residues in the RH, trans-activation loop, and α2 regions (**Figure 3.3.1c**). Notably, β1-4 residues are absent from this sub-network, casting doubt on an active role in the maintenance of TCSP-specificity. Further, because receiver-SDRs are themselves highly

correlated, we expected each DBD-SDR to exhibit high MI with multiple receiver-SDRs, which was true for the RH, α2, and TA loop (**Figure 3.3.1d**).

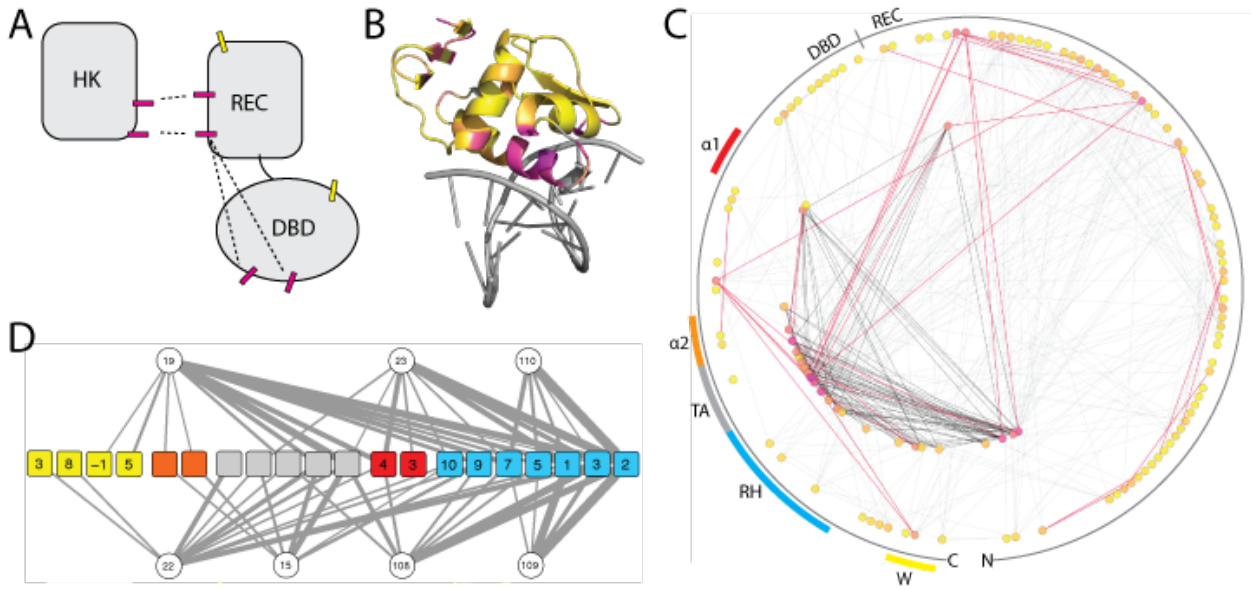


**Figure 3.3.1: Sequence- and structure-informed identification of specificity-determining residues at the protein-DNA interface.** **A.** cMI scores were projected onto the structure of PhoP (PDB code: 5ED4) on a scale from minimum cMI (yellow) to maximum cMI (magenta). **B.** Non-interacting SDRs in distant regions of the protein may exhibit covariance due to their shared selective pressure against TCSP crosstalk. **C.** Network representation of MI between positions in full-length RR. Residues are ordered counter-clockwise by primary sequence in a circular layout with the start (N) and end (C) positions at bottom, and colored according to cMI [93]. Edges between positions separated by <10 residues were removed, and nodes with <3 edges were subsequently filtered. Edges in the top 10% of MI scores are shown as solid red lines. Nodes that constitute the first-order network of receiver SDRs have been shifted to an inner, concentric ring and are connected by solid black lines. **D.** First-order contact MI involving the wing (yellow), α2 (orange), transactivation loop (gray), recognition helix (blue), and α1 (red) is represented by gray lines, with MI magnitude proportional to thickness. Numbered nodes reflect the reference coordinates previously established for residues in the wing, recognition helix, and α1 regions.

### 3.3.2 SDRs alter different aspects of eRR-DNA specificity

To functionally validate predicted SDRs, I performed an exchange of $RH_{7,10}$ between OmpR and CpxR (**Figure 3.3.2a**), and analyzed the interface variants (**Figure 3.3.2b,c**) using Spec-seq. Strikingly, CpxR-$RH_{7\text{-}10}$[QISR] underwent a complete conversion to an OmpR-like binding motif (**Figure 3.3.2d**); however, the reciprocal mutant OmpR-$RH_{7\text{-}10}$[HISN] retained its original specificity with greatly reduced apparent affinity. I further examined the effect of the $RH_{7\text{-}10}$[AISR] variants (RstA-like) in both backbone contexts, but found no substantial difference between RHs containing '**A**ISR' and '**Q**ISR'. This result shows that the $RH_{(7)}$[Q] side chain fulfills no unique biochemical role in sequence recognition, and is in good agreement with the near-identical sequence preferences for wild-type OmpR and RstA (**Figure 2.3.4c-e**).

From previous conservation analysis, OmpR-$RH_7$[Q] and OmpR-$RH_7$[H] are known to occur naturally at similar frequencies, while OmpR-$RH_{10}$[R] is highly conserved (**Figure 2.3.1c**). Interestingly, the single OmpR-$RH_7$[Q→H] variant lost canonical 'GT-A' preference; instead, OmpR- $RH_7$[H] recognized an A/T-rich tract most similar to the model generated by phospho-CpxR at its highest concentration. Interestingly, although this variant lost the canonical *preference* for '$G_{+2}T$' it retained specificity *against* '$C_{+2}G$' at the same positions, highlighting that base preferences are the net reflection of affinity gains and losses due to distinct molecular interactions. Further, CpxR-$RH_7$[Q], recognized a 'hybrid' motif, consisting of canonical binding in the 5' portion of the half-site ('$tG_{+2}T$') and CpxR-like base preferences in the 3' segment ('$AAG_{+7}AA$'). In the

context of previous mutants, this suggests that $RH_{10}$ independently governs base preferences in the 3' segment of each half-site: $RH_{10}$[R] ~> $C_{+6}$ / $RH_{10}$[N] ~> 'AA$G_{+7}$AA'. $RH_7$ appears to play a role in specifying the canonical $G_{+2}$. Overall, it appears that CpxR and OmpR are capable of specificity inter-conversion, although context does play a broader role.

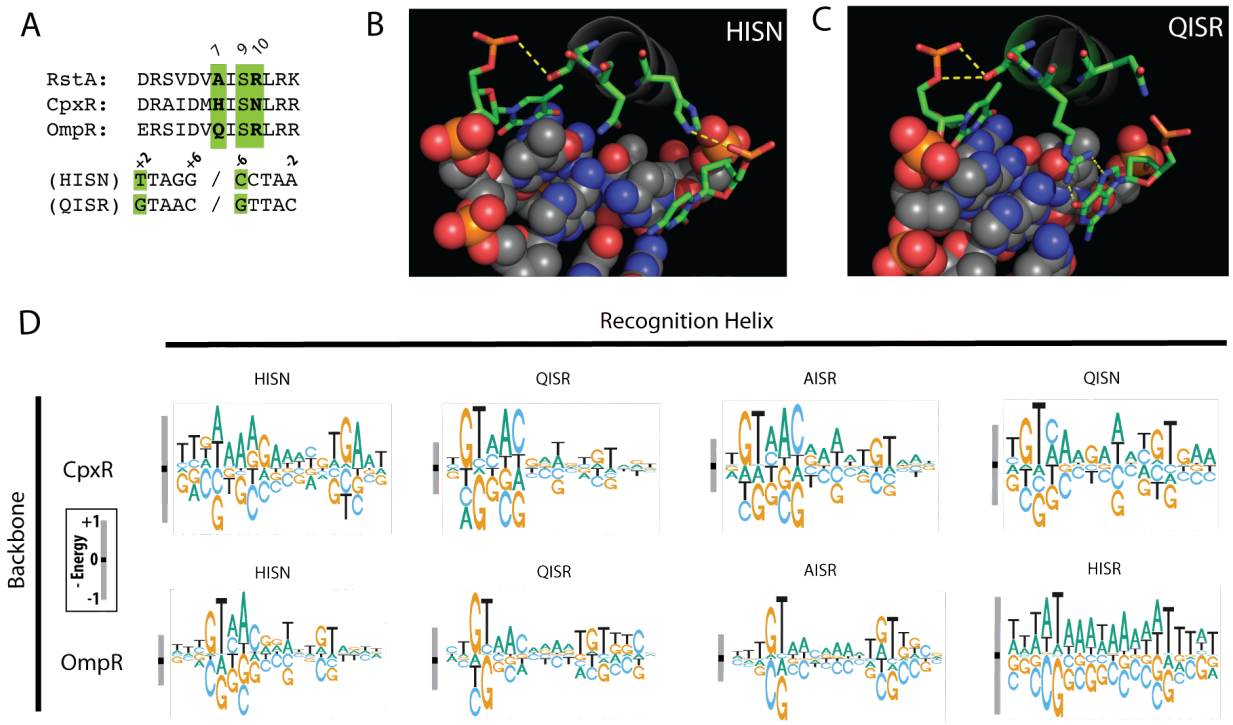


**Figure 3.3.2: Quantitative determination of binding specificity for recognition helix (RH) variants in the CpxR and OmpR proteins.** **A.** An alignment is given for the primary RH sequences of OmpR, CpxR and RstA (corresponding to the blue shaded region in panel 1A). DNA sequences refer to the bases in the structural models of variants depicted in (B) and (C). Bold positions in the amino acid alignment will be subject to mutation in (D). **B.** Structural model of a 'HISN'-containing RH modeled using the crystal structure of the *Klebsiella pneumoniae* PmrA protein (native sequence: HIHN) (PDB code: 4S05 [94]) in complex with a TTAGG half-site sequence as a template. **C.** Structural model of 'QISR'-containing RH, using the *Klebsiella pneumoniae* RstA protein in complex with a GTAAC half-site sequence as a template (PDB code: 4NHJ [95]). **D.** Energy logos for RH variants were generated from Spec-seq experiments using phospho-proteins are displayed. The background protein context (CpxR or OmpR) for the indicated mutations are shown in the left of the panel, and the RH residues are shown above each energy logo. The Y axis is in units of –kT, and the grey bar on the side of each logo reflects the range from -1 to +1.

## 3.4 Discussion

In this chapter, I demonstrate that residue covariation analysis - typically applied across protein and protein:DNA interfaces – identifies SDRs effectively when combined with limited biological information. This approach yielded a subset of putative DNA-contacting SDRs with which to prioritize experimental validation. While structural information was never directly incorporated in SDR prediction, putative SDRs localized almost exclusively to the DNA interface and transactivation loop, hypothetical hubs of paralog-specific biochemical activities. Modest residue changes at two putative SDRs successfully converted the specificity of two paralogous OmpR family TFs, providing the first high-resolution, quantitative characterization of 'portable' SDRs in the OmpR family.

We initially set out to expand a DNA-contacting network through a common application of mutual information, and were surprised to learn that it robustly identified a network of SDRs in the receiver domain involved in TCSP-specific phospho-transmission. It was novel, to our knowledge, to consider that SDRs important for many different, TCSP-specific functions might co-vary in a 'specificity-centered' network through a process of convergent evolution; in other analyses, in fact, such relationships may even be considered noise or random correlations arising due to shared phylogeny. One of the major benefits of this knowledge-based approach was that it required signal between completely non-interacting residue positions, implicitly controlling for statistical

correlations between residues adjacent in sequence and/or three-dimensional structure. Additionally, the approach required no structural input outside the fact that the two domain interfaces were non-interacting; as we used no explicit structural information, it was significant that the all high-ranking SDRs localized to regions of the protein in immediate contact with variable regions of the DNA half-site.

Structural analyses of OmpR family TFs have yielded substantial insight into the assembly of the protein:DNA complex, such as the orientation and potential contacts between bound monomers, providing context in which to interpret the role of individual side-chains on sequence recognition. A low-throughput, semi-quantitative analysis of RH residues in the *Bacillus anthracis* WalR protein (highly similar to ArcA) demonstrated that non-conservative mutations to $RH_2$, $RH_5$, $RH_6$, $RH_9$, and $RH_{10}$ near-uniformly reduced binding affinity toward a native operator sequence ('TGTAACATAACTGTAAC'); the single exception was $RH_5$[D→R], which increased non-specific binding affinity [96]. In contrast, my Spec-seq experiment demonstrated that $RH_{10}$ is definitively involved in sequence-specific recognition at $G_{-5}$, a conclusion with strong structural support. Additionally, $RH_{10}$ [R] appeared to increase the stability of the dimeric complex in both native and mutant contexts, with the slight exception of OmpR-$RH_7$[H]. However, neither the stability (i.e., percent correctly-folded protein) nor the activity (i.e., percent phosphorylated protein) were quantified in these assays, so it is not formally possible to distinguish effects on DNA-binding affinity from cumulative protein 'activity'.

Previously, the *E. coli* OmpR variant $RH_6$[V→M] was shown to selectively enhance binding affinity for the *ompF* enhancer sequences, while causing a reduction in affinity toward binding sites in the *ompC* promoter [97]. Conversion of a single base position in the *ompC* target to its *ompF* counterpart ('GTT$\mathbf{T}_{+5}$C' → 'GTT$\mathbf{A}_{+5}$C') selectively restored wild-type binding levels, indicating a direct interaction between $RH_6$ and the fifth base pair in the canonical half-site. Importantly, this residue was not identified as an SDR, indicating that it has not been a significant target of natural evolution to modify the specificity of sequence recognition by OmpR homologues. In the aforementioned study, it was also shown that $RH_6$[V→M] affected OmpR phosphorylation via an undetermined mechanism; it follows that a complex, multi-functional role precludes modification of this residue. In the majority of OmpR family paralog groups (including OmpR itself), $RH_6$[V] is moderately conserved; in CpxR, however, both Val and Met are equally represented at this position. This suggests that this residue may play a context-specific role in specificity determination, a concept that is supported both in this work and suggested in previous work using the LacI family [77].

Overall, this work demonstrates that residues in the recognition helix exert significant (if not primary) control over sequence recognition, complex architecture, and, more speculatively, DNA-binding affinity. Although not examined in this work, DNA-contacting residues outside this region are predicted to play a role in specificity determination, and provide a resource for residues for future DNA-binding analysis, target prediction for newly discovered OmpR homologues, and the design of new

regulatory tools for synthetic biology. Further, this work provides a basis for the continued study of two-component system evolution, which will help to decipher the regulation of complex homeostatic, pathogenic, and industrially relevant bacterial processes.

## 3.5 Materials and Methods

**Residue covariation analysis**

OmpR family orthologs were identified using a reciprocal best-BLAST criterion from a previously curated list of OmpR family members spanning 896 bacterial genomes [79]. To qualify as orthologous, a conservative cutoff of 40% sequence identity was also imposed along with a 90% redundancy threshold within each group using CD-HIT [82], and the remaining sequences were aligned using M-Coffee [85]. A family-wide alignment was constructed using step-wise, progressive profile alignments in Clustal Omega [98] guided by a preliminary tree based on pairwse amino acid similarity. A mutual information network was constructed for this multiple alignment using the MISTIC web interface with default parameters (projected onto the PhoP:DNA complex structure, PDB Code: 5ED4) [93].

**Construction of mutants**

All mutants were generated by site-directed mutagenesis of the appropriate wild-type plasmid construct by Gibson assembly method using Gibson Assembly Master Mix (New England Biolabs) [99]. Expression and purification were carried out as described for wild-type proteins.

**Expression, and purification**

Coding sequences of 14 response regulators (RRs) of the OmpR sub-family (ArcA, BaeR, BasR, CpxR, CreB, CusR, KdpE, OmpR, PhoB, PhoP, QseB, RstA, TorR, YedW) were amplified directly from *E. coli* MG1655 genomic DNA. Coding sequence for the StrepTagII affinity tag (WSHPQFEK) was added by PCR amplification along with upstream and downstream restriction sites for MfeI and XhoI, respectively. Strep-RR fusion protein sequences were sub-cloned into the pET-42a(+) expression vector in-frame with N-terminal GST and 6xHis purification tags and a thrombin protease cleavage site, generating triple-tagged constructs. Stock plasmids were stored, purified and handled using standard laboratory techniques.

ArcticExpress (DE3) competent cells (Agilent) were chemically transformed with expression plasmids, and single colonies from selective (Kan) LB-agar plates were used to inoculate 5ml LB-Kan starter cultures. After 6-8 hours growing at 37°C, starter cultures were scaled up to 400ml expression cultures in triple-baffled 4L flasks prepared with auto-induction media containing Kanamycin according to the Studier method [100]. Cultures were expanded at 37°C for 3-6 hours, then grown several hours past saturation (24-36 hours total growth time) to achieve maximum protein yield. Bacterial pellets were harvested by centrifugation, sonicated, re-pelleted at high speed to remove cellular debris, and lysate (diluted with 1X PBS to reduce viscosity) was passed through a 0.45μm syringe-tip filter for clarification. Lysate was passed over a HiTrap GST affinity column (1 ml capacity, GE Healthcare) and eluted under manufacturer-specified buffer conditions.

Fusion protein was cleaved with 5U thrombin protease, and GST-6xHis was removed with two rounds of treatment with Ni-NTA resin (Thermo Scientific). Protein samples were cleared completely of resin by passage through 0.22 μm syringe-tip filters (MANUFAC). Purity was assessed by both SDS-PAGE and size-exclusion chromatography, and protein concentration was determined by NanoDrop (Thermo Scientific).

**Spec-seq**

Binding reactions were prepared on ice in 12μl volumes containing 20ng FAM-labeled dsDNA library in 1X EMSA Buffer (25 mM Tris-Cl, 60 mM KCl, 140 mM NaCl, 1.5 mM $MgCl_2$, 0.2 mg/ml BSA, 5% glycerol, 10 ng/μl salmon sperm DNA, pH 8.3 at 8°C). Reactions were incubated 2 hours at 32°C with 25mM ammonium phosphoramidate or ammonium chloride for binding of phosphorylated and non-phosphorylated response regulators, respectively. Bound and unbound DNA pools were separated by native PAGE (8% polyacrylamide, 0.8x TBE [72 mM Tris-borate, 1 mM EDTA]) at 8°C. Gels were visualized on a Typhoon FLA 9500 (GE Healthcare) biomolecular imager. Bands containing bound and unbound DNA were excised, and DNA was extracted by the crush and soak method in gel diffusion buffer (.3 M sodium acetate, 1 mM EDTA). Eluted DNA was concentrated using the Qiaex II gel extraction kit (Qiagen).

## 3.6 References


1. Alibes A, Nadra AD, De Masi F, Bulyk ML, Serrano L, Stricher F (2010) Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. Nucleic Acids Res, 38, 7422–7431

2. Muller PAJ, Vousden KH (2013) p53 mutations in cancer. Nat. Cell Biol, 15, 2–8

3. D'Elia AV, Tell G, Paron I, Pellizzari L, Lonigro R, Damante G (2001) Missense mutations of human homeoboxes: a review. Human Mutation, 18, 361–374

4. Luscombe NM, Thornton JM (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol, 320, 991–1009

5. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007) Divergence of transcription factor binding sites across related yeast species. Science, 317, 815–819

6. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A, 104 Suppl 1, 8605–8612

7. Schmidt D, Wilson MD, Ballester B et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science, 328, 1036–1040

8. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet, 8, 206–216

9. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol, 1, REVIEWS001

10. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS (2010) Origins of specificity in protein-DNA recognition. Annu Rev Biochem, 79, 233–269

11. Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol, 301, 597–624

12. Wolfe SA, Ramm EI, Pabo CO (2000) Combining structure-based design with phage display to create new Cys(2)His(2) zinc finger dimers. Structure, 8, 739–750

13. Thyme S, Baker D (2014) Redesigning the specificity of protein-DNA interactions with Rosetta. Methods Mol Biol, 1123, 265–282

14. Thyme SB, Baker D, Bradley P (2012) Improved modeling of side-chain--base interactions and plasticity in protein--DNA interface design. J Mol Biol, 419, 255–274

15. Thyme SB, Boissel SJ, Arshiya Quadri S, Nolan T, Baker DA, Park RU, Kusak L, Ashworth J, Baker D (2014) Reprogramming homing endonuclease specificity through computational design and directed evolution. Nucleic Acids Res, 42, 2564–2576

16. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO (1996) Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure, 4, 1171–1180

17. Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res, 37, D77–82

18. Schrodinger LLC (2010) The PyMOL Molecular Graphics System, Version 0.99.

19. Boas FE, Harbury PB (2007) Potential energy functions for protein design. Curr Opin Struct Biol, 17, 199–204

20. Fornes O, Garcia-Garcia J, Bonet J, Oliva B (2014) On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein-Protein, Protein-DNA, and Protein-RNA Interactions. Adv Protein Chem Struct Biol, 94, 77–120

21. Chen CY, Chien TY, Lin CK, Lin CW, Weng YZ, Chang DT (2012) Predicting target DNA sequences of DNA-binding proteins based on unbound structures. PLoS One, 7, e30446

22. Siggers TW, Honig B (2007) Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Res, 35, 1085–1097

23. Xu B, Schones DE, Wang Y, Liang H, Li G (2013) A structural-based strategy for recognition of transcription factor binding sites. PLoS One, 8, e52460

24. Xu B, Yang Y, Liang H, Zhou Y (2009) An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Proteins, 76, 718–730

25. AlQuraishi M, McAdams HH (2013) Three enhancements to the inference of statistical protein-DNA potentials. Proteins, 81, 426–442

26. Parisien M, Freed KF, Sosnick TR (2012) On docking, scoring and assessing protein-DNA complexes in a rigid-body framework. PLoS One, 7, e32647

27. Takeda T, Corona RI, Guo JT (2013) A knowledge-based orientation potential for transcription factor-DNA docking. Bioinformatics, 29, 322–330

28. Contreras-Moreira B (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucleic Acids Res, 38, D91–7

29. Gabdoulline R, Eckweiler D, Kel A, Stegmaier P (2012) 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. Nucleic Acids Res, 40, W180–5

30. Lin CK, Chen CY (2013) PiDNA: Predicting protein-DNA interactions with structural models. Nucleic Acids Res, 41, W523–30

31. Li Z, Lazaridis T (2007) Water at biomolecular binding interfaces. Phys Chem Chem Phys, 9, 573–581

32. Schwabe JW (1997) The role of water in protein-DNA interactions. Curr Opin Struct Biol, 7, 126–134

33. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science, 309, 1868–1871

34. Wang C, Bradley P, Baker D (2007) Protein-protein docking with backbone flexibility. J Mol Biol, 373, 503–519

35. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res, 33, 5781–5798

36. Temiz NA, Camacho CJ (2009) Experimentally based contact energies decode interactions responsible for protein-DNA affinity and the role of molecular waters at the binding interface. Nucleic Acids Res, 37, 4076–4088

37. Beierlein FR, Kneale GG, Clark T (2011) Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. Biophys J, 101, 1130–1138

38. Liu LA, Bader JS (2009) Structure-based ab initio prediction of transcription factor-binding sites. Methods Mol Biol, 541, 23–41

39. Seeliger D, Buelens FP, Goette M, de Groot BL, Grubmuller H (2011) Towards computational specificity screening of DNA-binding proteins. Nucleic Acids Res, 39, 8281–8290

40. Li Y, Sutch BT, Bui HH, Gallaher TK, Haworth IS (2011) Modeling of the water network at protein-RNA interfaces. J Chem Inf Model, 51, 1347–1352

41. Jiang L, Kuhlman B, Kortemme T, Baker D (2005) A "solvated rotamer" approach to modeling water-mediated hydrogen bonds at protein-protein interfaces. Proteins, 58, 893–904

42. Li S, Bradley P (2013) Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model. Proteins, 81, 1318–1329

43. van Dijk M, Visscher KM, Kastritis PL, Bonvin AM (2013) Solvated protein-DNA docking using HADDOCK. J Biomol NMR, 56, 51–63

44. Tucker-Kellogg L, Rould MA, Chambers KA, Ades SE, Sauer RT, Pabo CO (1997) Engrailed (Gln50→ Lys) homeodomain–DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. Structure, 5, 1047–1054

45. Schneider B, Cohen D, Berman HM (1992) Hydration of DNA bases: analysis of crystallographic data. Biopolymers, 32, 725–750

46. Robinson CR, Sligar SG (1994) Hydrostatic pressure reverses osmotic pressure effects on the specificity of EcoRI-DNA interactions. Biochemistry, 33, 3787–3793

47. Rose PW, Bi C, Bluhm WF et al. (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res, 41, D475–82

48. Kalodimos CG, Bonvin AMJJ, Salinas RK, Wechselberger R, Boelens R, Kaptein R

(2002) Plasticity in protein–DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA‐binding domain. EMBO J, 21, 2866–2876

49. Gunther S, Rother K, Frommel C (2006) Molecular flexibility in protein-DNA interactions. Biosystems, 85, 126–136

50. Sunami T, Kono H (2013) Local conformational changes in the DNA interfaces of proteins. PLoS One, 8, e56080

51. Weikl TR, von Deuster C (2009) Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. Proteins, 75, 104–110

52. Varnai P, Djuranovic D, Lavery R, Hartmann B (2002) Alpha/gamma transitions in the B-DNA backbone. Nucleic Acids Res, 30, 5398–5406

53. Endres RG, Schulthess TC, Wingreen NS (2004) Toward an atomistic model for predicting transcription-factor binding sites. Proteins, 57, 262–268

54. Havranek JJ, Duarte CM, Baker D (2004) A simple physical model for the prediction and design of protein-DNA interactions. J Mol Biol, 344, 59–70

55. Ashworth J, Havranek JJ, Duarte CM, Sussman D, Monnat RJJ, Stoddard BL, Baker D (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. Nature, 441, 656–659

56. Havranek JJ, Baker D (2009) Motif-directed flexible backbone design of functional interactions. Protein Sci, 18, 1293–1305

57. Yanover C, Bradley P (2011) Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. Nucleic Acids Res, 39, 4564–4576

58. Cahill M, Cahill S, Cahill K (2002) Proteins wriggle. Biophys J, 82, 2665–2670

59. Pérez A, Luque FJ, Orozco M (2011) Frontiers in molecular dynamics simulations of DNA. Accounts of chemical research, 45, 196–205

60. Dans PD, Perez A, Faustino I, Lavery R, Orozco M (2012) Exploring polymorphisms in B-DNA helical conformations. Nucleic Acids Res, 40,

10668–10678

61. Perez A, Lankas F, Luque FJ, Orozco M (2008) Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Res, 36, 2379–2394

62. Deremble C, Lavery R, Zakrzewska K (2008) Protein–DNA recognition: Breaking the combinatorial barrier. Comput Phys Commun, 179, 112–119

63. Zakrzewska K, Bouvier B, Michon A, Blanchet C, Lavery R (2009) Protein–DNA binding specificity: a grid-enabled computational approach applied to single and multiple protein assemblies. Phys Chem Chem Phys, 11, 10712–10721

64. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, Bulyk ML (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol, 24, 1429–1435

65. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327–339

66. Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. Science, 315, 233–237

67. Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, Dervan PB, Phillips GNJ, Ansari AZ (2006) Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A, 103, 867–872

68. Zhao Y, Granas D, Stormo GD (2009) Inferring binding energies from selected binding sites. PLoS computational biology, 5, e1000590

69. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science, 316, 1497–1502

70. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, Chan ET, Metzler G, Vedenko A, Chen X (2009) Diversity and complexity in DNA recognition by transcription factors. Science, 324, 1720–1723

71. Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep, 3, 1093–1104

72. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell, 133, 1266–1276

73. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, Wolfe SA (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell, 133, 1277–1289

74. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol, 8, R24

75. Persikov AV, Singh M (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res, 42, 97–108

76. Mirny LA, Gelfand MS (2002) Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol, 321, 7–20

77. Sloutsky R, Naegle KM (2016) High-Resolution Identification of Specificity Determining Positions in the LacI Protein Family Using Ensembles of Sub-Sampled Alignments. PLoS One, 11, e0162579

78. Liu K, Linder CR, Warnow T (2010) Multiple sequence alignment: a major challenge to large-scale phylogenetics. PLoS Curr, 2, RRN1198

79. Galperin MY (2010) Diversity of structure and function of response regulator output domains. Curr Opin Microbiol, 13, 150–159

80. Buslje CM, Santos J, Delfino JM, Nielsen M (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics, 25, 1125–1131

81. Löytynoja A (2014) Phylogeny-aware alignment with PRANK. Methods Mol Biol, 1079, 155–170

82. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28, 3150–3152

83. Abriata LA (2016) Structural database resources for biological macromolecules.

Brief Bioinform,

84. Poleksic A (2009) Algorithms for optimal protein structure alignment. Bioinformatics, 25, 2751–2756

85. Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C (2007) The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. Nucleic Acids Res, 35, W645–8

86. Jehl P, Sievers F, Higgins DG (2015) OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics, 16, 269

87. Perez JC, Groisman EA (2009) Evolution of transcriptional regulatory circuits in bacteria. Cell, 138, 233–244

88. Perez JC, Groisman EA (2009) Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc Natl Acad Sci U S A, 106, 4319–4324

89. Joyce AP, Zhang C, Bradley P, Havranek JJ (2015) Structure-based modeling of protein: DNA specificity. Brief Funct Genomics, 14, 39–49

90. Park DM, Kiley PJ (2014) The influence of repressor DNA binding site architecture on transcriptional control. MBio, 5, e01684–14

91. Takayanagi K, Mizuno T (1992) Activation of the osmoregulated ompF and ompC genes by the OmpR protein in Escherichia coli: a study involving chimeric promoters. J Biochem, 112, 1–6

92. Capra EJ, Perchuk BS, Lubin EA, Ashenberg O, Skerker JM, Laub MT (2010) Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet, 6, e1001220

93. Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C (2013) MISTIC: Mutual information server to infer coevolution. Nucleic Acids Res, 41, W8–14

94. Lou YC, Weng TH, Li YC, Kao YF, Lin WF, Peng HL, Chou SH, Hsiao CD, Chen C (2015) Structure and dynamics of polymyxin-resistance-associated response

regulator PmrA in complex with promoter DNA. Nat Commun, 6, 8838

95. Li YC, Chang CK, Chang CF, Cheng YH, Fang PJ, Yu T, Chen SC, Li YC, Hsiao CD, Huang TH (2014) Structural dynamics of the two-component response regulator RstA in recognition of promoter DNA element. Nucleic Acids Res, 42, 8777–8788

96. Dhiman A, Rahi A, Gopalani M, Bajpai S, Bhatnagar S, Bhatnagar R (2017) Role of the recognition helix of response regulator WalR from Bacillus anthracis in DNA binding and specificity. Int J Biol Macromol, 96, 257–264

97. Tran VK, Oropeza R, Kenney LJ (2000) A single amino acid substitution in the C terminus of OmpR alters DNA recognition and phosphorylation. J Mol Biol, 299, 1257–1270

98. Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. Methods Mol Biol, 1079, 105–116

99. Gibson DG (2011) Enzymatic assembly of overlapping DNA fragments. Methods Enzymol, 498, 349–361

100. Studier FW (2005) Protein production by auto-induction in high density shaking cultures. Protein Expr Purif, 41, 207–234

# Chapter 4. Conclusions and future directions

This thesis explores the sequence-specific binding of DNA by *E. coli* paralogs belonging to the OmpR sub-family of two-component response regulators, an essential sub-family of bacterial transcription factors (TFs) possessing a winged helix-turn-helix (wHTH) DNA-binding domain (DBD). Despite the prevalence of the wHTH domain in bacterial TFs, prior to this work neither the DNA sequence recognition potential nor structural basis of sequence-specific binding had been fully characterized on a large scale *in vitro*. In this investigation, I discovered both canonical and non-canonical forms of binding utilized by different members of this TF family, examined the conservation and co-variation of amino acid residues, and both predicted and validated specificity-determining residues (SDRs) with the potential to alter specificity via both DNA base and backbone contacts. These data greatly expand our understanding of the basic structural mechanisms by which OmpR homologues select specific genomic targets, and can support future lines of investigation:

### 4.1 Validation and characterization of non-canonical binding mode

Using *in vitro* SELEX, it was determined that a representative majority of OmpR homologues recognized a related binding motif, which I termed the 'canonical' mode. This mechanism of binding was especially prevalent when proteins were presented with a 'seed' half-site held constant in SELEX libraries; however, in the absence of a partial site, an alternative binding motif. Strikingly, both the canonical and alternative motifs bear resemblance to some previously observed for a eukaryotic wHTH (Forkhead) TF family

[1]. Anecdotally, this motif also appears frequently in regions enriched for binding by QseB and BasR in a previous study using genomic SELEX [2]. Moreover, it appears repetitively in verified enhancers for the closely related BasR, as well as in the sequence selected for structural analysis [3; 4]. There are relatively few *bona fide* examples of TF multi-specificty; as such, this result warrants further study.

Although not fully reported in this work, I attempted to replicate the interaction between QseB and both canonical and non-canonical motifs by EMSA; under identical conditions, only oligonucleotides containing the canonical motif were bound. These attempts to demonstrate non-canonical binding by an orthogonal method were undertaken in a gel format - formally, a kinetic technique. SELEX, however, was conducted in solution under equilibrium conditions. If the alternative complex is dynamic or very large, a gel format could lower its stability. Steps could also be taken to generate higher-quality protein, and/or to quantify the degree of phosphorylation.

On the balance of circumstantial evidence, I offer three suggestions for continued study of this potential mode of binding. First, Future analyses should take these important assay properties into account; perhaps solution-state measurements or *in vivo* activity assays can shed light on this mode of DNA interaction. Second, genomic SELEX analysis has provided a variety of potential alternative target sites for QseB, but only a single one was selected for validation. An attempt should be made to screen binding to a broader set of genomic regions, guided by genomic SELEX enrichment and known QseB target genes. Third, KdpE and BasR were also observed to enrich for alternative motifs, and

could also be the subject of these analyses.

### 4.2 Characterization of native targets by Spec-seq

A major conclusion of this work was the relative uniformity of sequence recognition by distantly-related, functionally-distinct OmpR paralogs. Motif discrimination is unlikely to produce the level of genomic specificity inferred from analysis of native sequences. Therefore, there is likely a great deal of interesting 'specificity mechanisms' encoded directly in cis-regulatory sequences (e.g., multimeric binding, cooperativity, low-affinity / transient interactions, etc.) that warrants further study.

Importantly, the significant contrast between motifs identified through SELEX and known genomic binding sites suggests far greater complexity than synthetic libraries can feasibly represent. In a hypothetical follow-up experiment, regions enriched by genomic SELEX can be randomly diversified (e.g., through error-prone PCR) then analyzed using Spec-seq. Spec-seq is an ideal approach, given its ability to resolve complexes of different molecular weights, a necessity given the diverse size range of native DNA targets and potential for multi-protein complex formation. Results can shed light on how affinity is 'tuned' at specific target sites, and, much like CpxR, reveal alternative sequence-specificities driven by multimeric binding.

### 4.3 Expanding SDR prediction to other TF families

There is no shortage of techniques to predict DNA-, ligand-, and protein-binding SDRs; although algorithmically distinct, many produce very similar results and require

arbitrary statistical thresholds. In this work, I applied a simple biological principle: that SDRs engaged in disparate biochemical processes and spatial domains, if linked through a selective pressure toward functional specificity, evolve convergently as a single specificity-centric network. This assumption enabled the application of information theoretic methods typically applied to identify interacting residues across interfaces; importantly, because SDR groupings do not physically interact (in this work, SDRs occur in distinct domains), this approach implicitly eliminates covariance signal due to structure- or sequence-proximity. Although the success of this approach was not rigorously tested (e.g., through large-scale mutagenesis), it did enrich strongly and specifically for DNA-contacting residues, and two moderately-ranked predictions were functionally validated and characterized.

Beyond the OmpR sub-family, bacteria contain large numbers of TFs that bind DNA in response to chemical and protein ligands, and, for many such families, binding pockets and interfaces have been predicted and described. Given the broad availability of protein sequences in public database repositories, it should be possible to perform analyses similar to the one proposed here to identify DNA-binding SDRs on a large scale. For certain ligand-binding families, such as LacI, extensive mutational studies have already been conducted, and may provide necessary functional benchmarks with which to evaluate success.

**1) Training structure-based models**

Large-scale functional characterization of naturally-occurring TFs has, in some ways,

reduced the need to develop biophysically accurate models of protein:DNA interactions. However, it would be unwise to de-emphasize the development of tools for structure-based design. First, many of the challenges in modeling the protein:DNA interface are generic to nucleic acids, so the potential impacts will extend far beyond the prediction of sequence recognition potential. Second, given recent progress in genomic sequencing, there will always be more disease- and phenotypically-relevant variants than can be characterized experimentally [5]. Due to their high sequence variability and functional diversity, for example, two-component response regulators are especially appropriate targets for modeling. Thirdly, it is currently possible to calculate the structure of multi-protein / DNA complexes only at relatively low resolution, and scoring functions or algorithms intended for protein design have been shown to aid in refinement. Thusly, protein:DNA modeling can, even in its current, imperfect state, make a significant impact on structural and regulatory biology [6]. Finally, and more generally, advances in modeling will continue to benefit from and drive computing and algorithmic advances.

# References


1. Nakagawa S, Gisselbrecht SS, Rogers JM, Hartl DL, Bulyk ML (2013) DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc Natl Acad Sci U S A, 110, 12349–12354

2. Ishihama A, Shimada T, Yamazaki Y (2016) Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. Nucleic Acids Res, 44, 2058–2074

3. He X, Wang S (2014) DNA consensus sequence motif for binding response regulator PhoP, a virulence regulator of Mycobacterium tuberculosis. Biochemistry, 53, 8008–8020

4. Ogasawara H, Shinohara S, Yamamoto K, Ishihama A (2012) Novel regulation targets of the metal-response BasS-BasR two-component system of Escherichia coli. Microbiology, 158, 1482–1492

5. Barrera LA, Vedenko A, Kurland JV et al. (2016) Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science, 351, 1450–1454

6. Mao B, Tejero R, Baker D, Montelione GT (2014) Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding X-ray crystal structures. J Am Chem Soc, 136, 1893–1906