UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

DETERMINANTS OF GENE TARGETING AND REGULATION BY RESPONSE REGULATORS IN *CLOSTRIDIOIDES DIFFICILE*

A

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

SKYLER D. HEBDON
Norman, Oklahoma
2018

DETERMINANTS OF GENE TARGETING AND REGULATION BY RESPONSE REGULATORS IN *CLOSTRIDIOIDES DIFFICILE*

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Dr. Ann West, Chair

Dr. Elizabeth Karr

Dr. Christina Bourne

Dr. George Richter-Addo

Dr. Wai Tak Yip

Dedicated to Terry Wayne Hebdon

1957-2014

Acknowledgements

I would like to thank the University of Oklahoma Department of Chemistry and Biochemistry, in particular Dr. West and members of my graduate committee, for their continuous support and encouragement that helped me arrive at this culminating work. I have benefited from—and thoroughly enjoyed—the camaraderie among the researchers in Dr. West's lab, for which I owe thanks to Dr. Ann West, Dr. Smita Menon, Rachel Kim, Jamie Sykes, Dr. Emily Kennedy, Dr. Katie Branscum, Dr. Clay Foster, Dr. Fabiola Janiak-Spens, and the host of undergraduate researchers who have frequented the laboratory. In fact, I must extend that gratitude to everyone on the second floor, East wing of the Stephenson Life Sciences building—the open lab concept works well when you have good neighbors. I especially thank Rachel Kim for her efforts to cure the figures in this dissertation of their colorblindness inherited from me, a role that Jamie Sykes also played for many of my seminar talks.

Finally, and most importantly, I thank my family: parents, siblings, wife, and children. I am blessed to leave the lab behind and go home to Lisa, Claire, Todd and Julie every day. Their love and support has helped me keep my sanity mostly intact.

Direct scientific contributions are acknowledged at the end of each chapter.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

3-AT	3-amino-1,2,4,-triazole
5-IAF	5-iodoacetamido fluorescein
ABC	Adenosine triphosphate-binding cassette
ATP	Adenosine triphosphate
B1H	Bacterial one-hybrid
BLAST	Basic local alignment search tool
BLASTP	Basic local alignment search tool for proteins
CFU	Colony forming units
COBRE	Center of Biomedical Research Excellence
DBD	DNA-binding domain
DOOR2	Database of prokaryotic operons version 2
EMSA	Electrophoretic mobility shift assay
FASTA	Fast alignment sequence format
FTIR	Fourier-transform infrared
GFP	Green fluorescent protein
GO	Gene ontology
GOMo	Gene ontology for motifs
HK	Histidine kinase
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K_d	Equilibrium dissociation constant
kDa	kiloDaltons
LB	Luria broth
MALS	Multiple angle light scattering
MEME	Multiple em for motif elicitation
OD	Optical density
PA	Phosphoramidate
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
REC	Receiver domain
RNAP	Ribonucleic acid polymerase
RR	Response regulator
RSAT	Regulatory sequence analysis tools
SDS	Sodium dodecyl sulfate
SEC	Size exclusion chromatography
UV	Ultraviolet
ωRNAP	Omega (ω) subunit of RNAP

Abstract

Bacteria use two-component signal transduction systems to sense the conditions of the environment and adapt their behavior to ensure survival. The separate roles of signal perception and response output are carried out by histidine kinase and response regulator proteins, respectively. Most response regulators alter gene transcription, but it is not yet possible to predict which genes they regulate. The precise prediction of these gene regulatory outputs could enable researchers to predict, and doctors and patients to mitigate, various bacterial behaviors.

This dissertation presents the elucidation of the genes and biological functions regulated by two response regulators, RR_1586 and RR_1677, from the hypervirulent human pathogen *Clostridioides difficile* R20291. The data presented herein supports the conclusion that RR_1586 regulates genes involved in phosphate transport, and further characterization of its activity suggests several mechanisms it uses to regulate those genes. The bacterial one-hybrid assay used in this study to elucidate the gene regulatory targets of RR_1586 could potentially be used to find a generalizable solution to predicting gene targets of all response regulators from genome sequences alone. In the process of optimizing the bacterial one-hybrid assay for such a broadly impactful endeavor, it was found that RR_1677 appears to regulate the processes of protein synthesis and cell wall synthesis. A bioinformatics pipeline was also constructed by combining several existing utilities to analyze and interpret the experimental results.

Two additional lines of research are also presented. The first is the adaptation of Fourier-transform infrared spectroscopy to observe response regulator phosphorylation. This method offers an alternative from the more labor-intensive methods currently available. The second is an analysis of the biophysical interactions between a response

regulator from *Saccharomyces cerevisiae* and its binding partner, a histidine-containing phosphotransfer protein. Characterization of this interaction elucidates the physical impetus behind the evolutionary conservation of a specific glycine residue near the active site whose role was previously unknown.

Chapter 1: General introduction: two-component signal transduction

Two-component signal transduction systems comprised of histidine kinases (HKs) and response regulators (RRs) are the primary means that bacteria use to sense and respond to their surroundings. HKs are a class of proteins that allow bacteria to perceive external stimuli such as nutrient concentrations, antibiotics, environmental stresses, or quorum signals. These signals stabilize active or inactive conformations of the HKs, as described elsewhere (1). A kinase-active HK autophosphorylates a conserved histidine residue, which serves as a phosphoryl reservoir for the cognate RR. An RR is capable of phosphoryl transfer from the phospho-histidine residue of an HK—and to a lesser degree from small molecule phosphoryl donors—to a conserved aspartate in the RR active site (2). Accommodation of the negatively charged phosphoryl group triggers conformational changes in the RR, resulting in altered biological activity. Thus, information about the environment is translated into the chemical form of a phosphoryl group by the HK and transferred to the RR, which in turn, interprets the chemical information into a biological response.

The mechanism by which RRs transduce the chemical energy of the phosphoryl group into a biological output is mediated by conserved features of a receiver or REC domain. A REC domain is the identifying characteristic of all RRs. It consists of five βα secondary structure motifs organized into a globular domain (Figure 1.1A) (2). The five β strands form a parallel sheet at the core of the REC domain surrounded by a perimeter of the five α helices. Highly conserved residues at the C-terminus of the β strands comprise the active site. A cluster of acidic residues in immediate proximity of the phosphorylatable aspartate at the C terminus of β3 is largely unchanged by

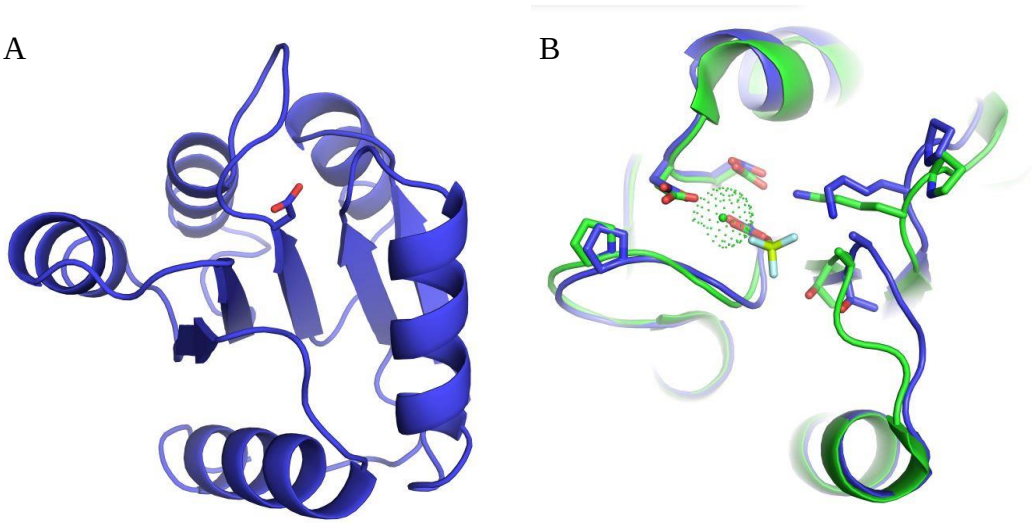

Figure 1.1 Structure of a REC domain
A) The $(\beta\alpha)_5$ fold and phosphorylatable aspartate residue (sticks) of the RR CheY (PDB: 2che). B) The loops in the active site of CheY in apo (blue, PDB 2che) and activated (green, 1fqw) forms. Conserved active site residues and beryllium fluoride phosphoryl mimic are shown in sticks. The position of the metal cation is shown as stippled dots. Figures prepared using PyMOL v1.3 (3).

phosphorylation, but the entire α4-β5-α5 region shifts as a conserved lysine residue on the β5-α5 loop and a serine or threonine residue on the β4-α4 loop shift several angstroms to form a new salt bridge or hydrogen bond interaction with the phosphoryl group (Figure 1.1B). These conformational changes often alter quaternary structures mediated by the REC domain and/or the activity of the C-terminal effector domain.

The effector domain of an RR determines the type of response it can produce, which must correlate to the perceived stimulus to be advantageous. RRs are categorized into families by the identity or absence of an effector domain. These families broadly represent domains with catalytic function, often synthesizing or degrading second messengers, protein-binding domains, and RNA- or DNA-binding domains. The most abundant is the OmpR family (4), which regulates gene expression through a winged helix-turn-helix DNA-binding effector domain. Phosphorylation of OmpR-family RRs often results in formation of a dimer with the DNA-binding domains (DBDs) oriented

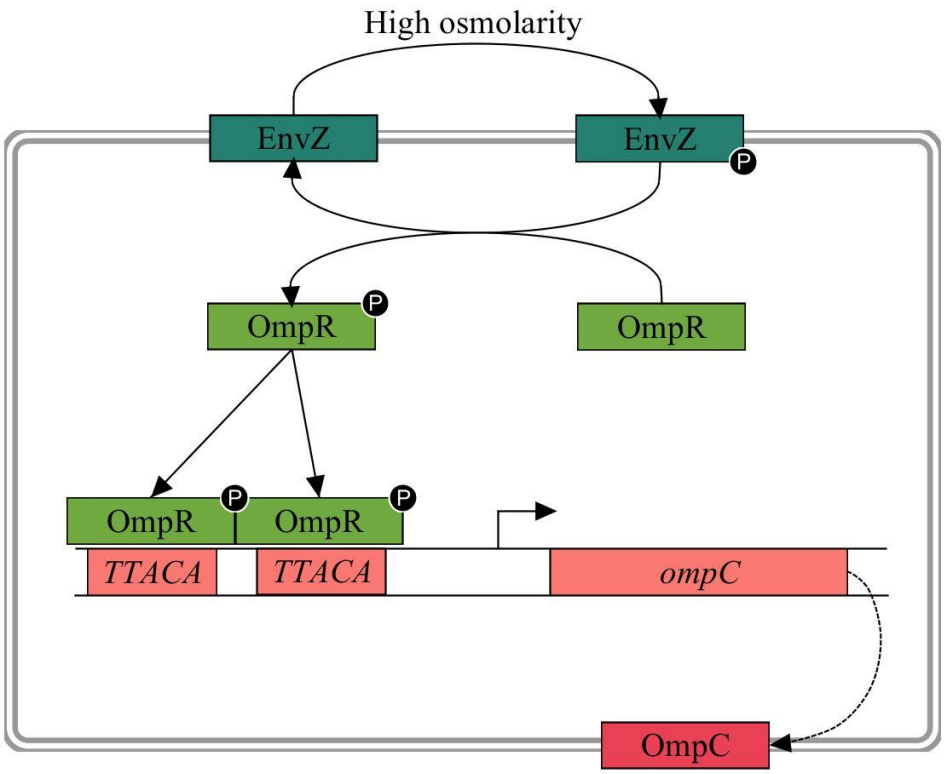

Figure 1.2 Example of RR gene transcription
The model RR protein OmpR activates transcription of the *ompC* gene encoding an outer membrane porin protein in response to high osmolarity sensed by EnvZ. Phosphorylation induces dimerization and increases binding to direct repeats of the TTACA motif (6). Expression of the OmpC porin is necessary for *E. coli* to transition to the animal gut environment. This figure was created using PathVisio 3.3.0 (7).

head-to-tail to recognize and bind to direct repeats of specific DNA motifs (Figure 1.2) (5-7). In the case of OmpR and other DNA-binding RRs, the biological response produced depends on DNA-binding specificity and the genes affected by binding.

Bacterial genomes generally encode multiple RRs, each regulating a distinct biological process. The link between an RR and the process it regulates can sometimes be inferred through homology or its genomic context. For example, the genome of the hypervirulent human pathogen *Clostridioides difficile* R20291 (CDR20291) encodes homologues of VanR, KdpE, Spo0A and EutV, which regulate vancomycin resistance, potassium starvation response, sporulation, and ethanolamine metabolism, respectively

(8). These are important aspects of *C. difficile* biology and pathogenicity. *C. difficile* is an obligate anaerobe that can colonize the lower intestinal tract and causes symptoms ranging from diarrhea to fatal pseudomembranous colitis (9). Spores, formed after transduction of a signal through the master regulator Spo0A (10), are the primary means of transmission between human hosts through the hostile aerobic environment. Inside the lower intestines, KdpE and EutV likely help *C. difficile* to compete with other microorganisms for efficient uptake and utilization of nutrients (11, 12). The roles of other important RRs in *C. difficile* have been identified experimentally (13-15). AgrA2 and CdtR are most notable because they regulate production of toxins A and B (14)—the primary symptom-causing virulence factors—or *C. difficile* binary toxin (15), respectively. These, however, comprise only a fraction of the 57 total RRs in the CDR20291 genome (8). Several factors, including the intractability of modifying the *C. difficile* genome, have hampered the study of the remaining RRs.

A more complete study of gene regulation in *C. difficile* by RRs could uncover pathways that can be exploited for therapeutic advantage. Two-component systems are absent in animals, but their ubiquitous presence among bacteria makes them attractive therapeutic targets. However, no such drugs had yet reached clinical trials as of late 2017 (16). The true benefit of understanding two-component systems lies in being able to predict and therefore modify bacterial behaviors. For example, a recent study found that the acquisition of genes involved in metabolism of the low calorie sweetener trehalose may have given a competitive advantage to, and facilitated the outbreak of, a hyper-virulent strain of *C. difficile* (17). Patients may therefore wish to avoid trehalose as a simple step to lessen their symptoms or chances of repeated infection. A more

complete understanding of how RRs respond to their environment is likely to reveal similar strategies that could be used, for example, to avoid stimulating RRs that induce sporulation or toxin production, or to purposefully stimulate RRs that block these processes.

Despite the wealth of structural and functional information available concerning HK and RR proteins, we cannot reliably predict their biological role from sequence alone. Even sequence similarity to a well-characterized RR is not reliable enough to predict the function of an uncharacterized RR (4). Conversely, unrelated proteins may serve identical functions in unrelated species. For example, a minimum diversity of HK sensory domains is expected in order to perceive the diverse classes of relevant stimuli, but many independent evolutionary pathways can lead to unrelated protein structures used to sense the same stimulus (18). The sequence-to-function relationship is disrupted even further by the persistence of nonfunctional sensory domains that have diverged from a functional ancestor protein (19). Predicting the function of RRs is also difficult, especially because the function does not depend wholly on the RR but also on the downstream signaling partner. Hypothetically speaking, two RRs could recognize the same DNA sequence, but the biological role of the RRs could be completely different if that DNA sequence is associated with virulence genes in one organism and genes encoding a particular metabolic pathway in another.

Several screens have been developed, and new strategies are continually being sought, to elucidate the ligands or other inputs perceived by HKs (20-22). Several strategies have also been developed to detect the specificities of DNA-binding RRs, which account for the majority of RRs. The classical demonstration of gene regulation

is to detect changes in transcription of gene targets in response to the loss of function of the RR being analyzed, usually by gene knockouts. Transposon mutagenesis and next-generation sequencing techniques have been developed to detect gene-phenotype relationships in massively parallel experiments (23), but at the expense of losing molecular resolution of the physical, causal interactions. A systematic application of *in vitro* methods was used in a genome-wide mapping of RRs to downstream DNA binding sites and putative gene targets (24), but not all proteins are amenable to *in vitro* studies due to insolubility, instability or other problems that arise during overexpression and purification. Another approach—using a bacterial one-hybrid (B1H) assay to determine the DNA binding specificity of proteins expressed to low levels in a heterologous host—seems to circumvent the limitations of *in vitro* studies while still achieving similar results (25). B1H analysis of the homeodomains that regulate gene transcription in *Drosophila* has disentangled protein-family-wide patterns correlating a protein sequence to its preferred DNA recognition sequence (26). Similar correlations between RR amino acid sequences and their respective DNA-binding specificities could be also be drawn from high-throughput B1H analysis. Using a sequence-derived binding motifs to predict downstream gene targets could enable annotation of RR functions based solely on an organism's genome sequence. This would bridge one of the largest gaps remaining in the field of two-component signal transduction and would be of great value to medical researchers.

Research Scope

The work presented in this dissertation demonstrates the conceptual and technical framework that has the potential to facilitate *de novo* prediction of genes

directly regulated by an RR. Chapter 2 describes the application of a bacterial one-hybrid assay to determine the DNA-binding specificity of RR_1586 from *C. difficile* R20291. This specificity, in light of bioinformatics analysis and experimental evidences, suggests that RR_1586 regulates phosphate ion homeostasis. The effects of phosphorylation on RR_1586 also suggest regulatory mechanisms described in Chapter 2. This work was published with the intention that it could serve as a template for other RRs (27), but significant effort was required to optimize application of the assay to other RRs. Chapter 3 provides an account of the advancements made since the study presented in Chapter 2, including the determination of the DNA-binding specificity of an RR_1677 and the development of a streamlined bioinformatics application. Chapter 4 describes the novel use of vibrational spectroscopy to monitor RR phosphorylation by small molecule phosphodonors in real time and independent of radioisotope tracers. This method is anticipated to extend the utility of real-time, optical methods to all RRs, as opposed to current fluorescence-based methods that only apply to a small subset of RRs. Finally, Chapter 5 presents my contribution toward a collaborative study of a multi-step histidine-to-aspartate phosphorelay signal transduction system found in yeast. A manuscript describing this work will likely be accepted for publication (28).

Chapter 2: Regulatory targets and mechanisms of the response regulator RR_1586 from *Clostridioides difficile* R20291

The work described in this chapter has been published (Copyright © 2018 Hebdon et al.) (27) and is reproduced here in accordance with the publishing agreements.

I employed a bacterial one-hybrid (B1H) assay to characterize the DNA-binding specificity, and bioinformatics to predict the genomic binding sites, of the OmpR-family RR encoded by *CDR20291_1586* (RR_1586). Along with other RRs, RR_1586 appears to be involved in processes important to sporulation according to transposon mutagenesis assays (29). In the B1H assay (25), a chimera of the RNA polymerase ω subunit (ωRNAP) and the transcription factor of interest binds to a randomized DNA sequence upstream of *his3* and *ura3* genes. The weak *his3-ura3* promoter is not recognized by RNAP unless guided there by an interaction between the chimera (bait) protein and the upstream random DNA sequence (prey) as depicted in Figure 2.1. The

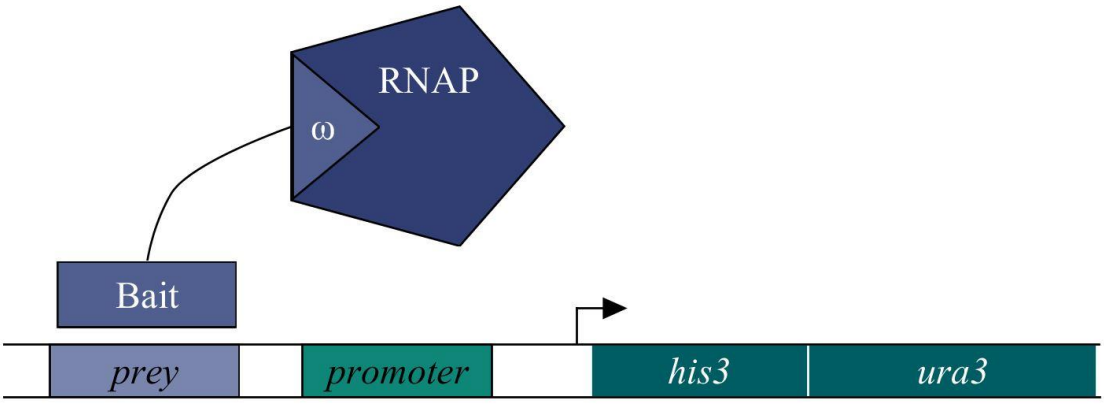

Figure 2.1 Diagram of bacterial one-hybrid assay
Transcription of the *his3* and *ura3* genes is induced by the interaction between the bait protein and a specific subset among millions of random prey sequences. The weak promoter is not recognized by RNAP unless guided there by the bait-prey interaction. Thus survival of auxotrophic *E. coli* on minimal medium agar plates indicates an interaction between the bait protein and prey DNA. Patterns of bait specificity can be found among surviving colonies. This figure was created using PathVisio 3.3.0 (7).

prey sequences from colonies surviving on histidine- and uracil-free medium contain the binding motif of the transcription factor, in this case RR_1586.

The putative regulon of RR_1586 was identified by searching for genes in the *C. difficile* R20291 genome with evolutionarily conserved binding sites. These findings are supported by an *E. coli*-based green fluorescent protein (GFP)-fusion reporter assay. We also report *in vitro* characterization of the effects of phosphorylation on oligomerization and DNA binding, and propose a working model for gene regulation by RR_1586. We anticipate that similar analyses of other RRs and transcription factors could lead to a global understanding of gene regulation by RRs in *C. difficile* and other pathogenic bacteria.

Results

RR_1586-DNA interaction specificity

The DNA motif recognized by RR_1586 was identified using an *E. coli*-based B1H assay (25). In this assay, transcription of *his3* and *ura3* genes is made possible if the bait protein binds to a randomized 28 base pair fragment inserted immediately upstream of the promoter. Binding of the bait chimera protein to both the prey DNA and the RNA polymerase enzyme (through the ωRNAP subunit) induces gene expression and cell survival in the absence of histidine and uracil. By plating millions of cells harboring a diverse library of randomized prey sequences on selective medium, a subset of prey sequences compatible with the bait chimera will survive. In a successful selection, this subset of sequences contains an overrepresentation of the DNA recognition motif to which the transcription factor binds.

Although ωRNAP fusions of full-length and three constructs of various lengths of the RR_1586 C-terminal DBD were tested (beginning at Arg124, Ser131, Gln151),

only the fusion at Ser131 yielded significant levels of selection. This position includes the predicted β platform (except for the first strand) and the winged helix-turn-helix domain of RR_1586 (see Figure 3.2 below). It does not include any of the unstructured linker leading to the receiver domain. The number of colonies that survived selection was only ~7-fold higher than background compared to the zinc finger positive control bait protein that exhibits at least 100-fold higher than background (data not shown). This is probably because zinc finger proteins have much higher affinity to DNA than RRs belonging to the OmpR family. This low yield could indicate that the DNA-RR_1586 interaction affinity is near the lower limit of detection for the B1H assay.

Overrepresented motifs were identified in the randomized fragments from colonies that survived selection (Appendix A). Low stringency selection with 10 mM 3-aminotriazole (3-AT) failed to produce a concise motif, whereas analysis of sequences

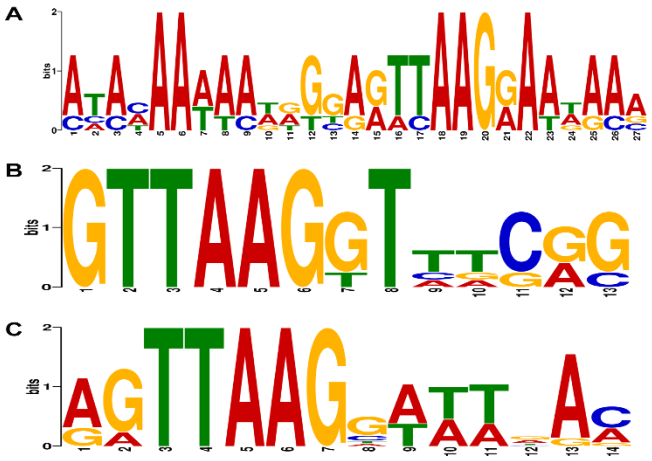

Figure 2.2 DNA-binding specificity of RR_1586
The motifs were overrepresented in colonies isolated from A) low stringency (10 mM 3-AT) or B) high stringency (20 mM 3-AT) selection or in C) both data sets. Statistical confidence in these motifs increases with stringency and sample sizes. Associated E-values are 3.1×10^{-4}, 2.6×10^{-15}, and 7.7×10^{-24}, respectively.

from high stringency screening or combined data sets produced highly significant motifs (Figure 2.2). The statistical significance of these motifs increased with the stringency of selection as expected (Figure 2.2) (25). More details of data processing are available in Appendix B.

Binding of full-length RR_1586 to the observed motif was confirmed by electrophoretic mobility shift assays (EMSAs) as shown in Figure 2.3. Given that OmpR-family response regulators typically bind direct repeats and the observation that purified RR_1586 is a dimer (see below), we presumed that the biological motif recognized by RR_1586 is a direct repeat of the B1H-derived motif. This was affirmed by EMSAs, where we observed a gel-shifted band pattern in the presence of RR_1586 for a direct repeat of the motif, and to a much lesser extent for a single motif. A

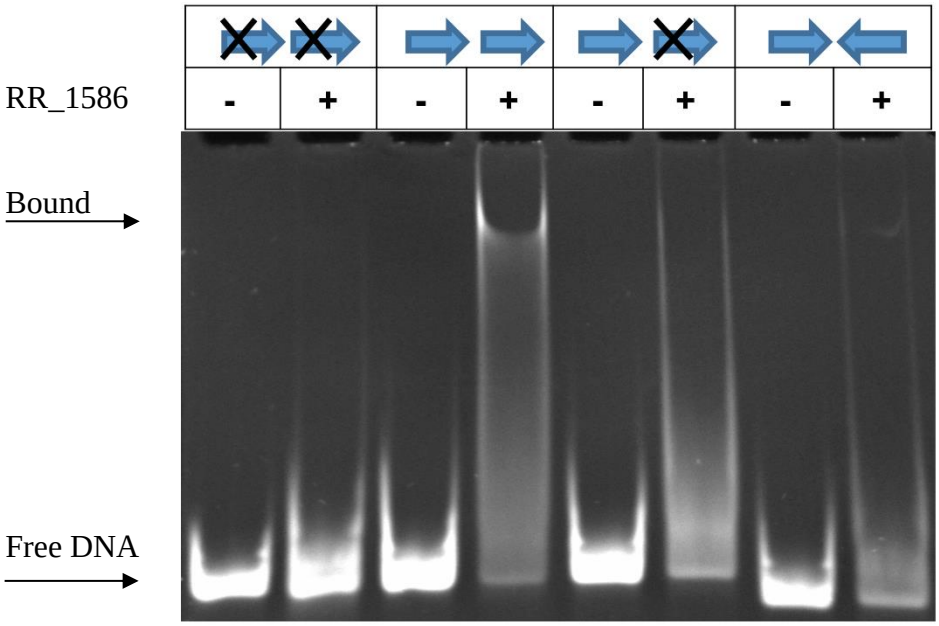

Figure 2.3 RR_1586 binds direct repeats *in vitro*
The presence of RR_1586 shifted direct repeats of the B1H-derived motif (blue arrow) in EMSAs. Inverting or substituting (black cross) one or both of the repeats diminished the shift suggesting a weakened interaction. Oligo sequences listed in Table S2. Presence (+) or absence (-) of RR_1586 protein is indicated above each lane.

substituted repeat or an inverted repeat showed little to no gel shift in the presence of RR_1586 suggesting a weakened or no interaction (Figure 2.3). Hereafter, we refer to a direct repeat of the first 11 nucleotides of the B1H-derived motif (Figure 2.2C) as the RR_1586 binding site.

Putative regulon of RR_1586

Having characterized the DNA-RR_1586 interaction specificity, we used bioinformatics to identify a putative regulon and the biological functions associated with the RR_1586 binding site. The assumption that the biological role of RR_1586 is evolutionarily conserved suggests that the RR, its downstream target genes, and the associated RR_1586 binding sites will also be conserved. We used scripts from the Regulatory Sequence Analysis Tools suite (RSAT) (30) to evaluate the co-conservation of these three elements relative to RR_1586 in the genomes of 26 species in the *Peptostreptococcaceae* family (Appendix C), to which *C. difficile* belongs. Orthologues of RR_1586, defined as bidirectional best BLASTP hits (31), were found in 17 genomes, which were further analyzed for conservation of RR_1586 binding sites (Appendix C). Fourteen putative gene targets were identified. These genes represent the operons that comprise the conserved core of the putative RR_1586 regulon, including seven operons that encode ion or ABC-type transporters (Table 2.1).

Single-genome scanning revealed additional, non-conserved putative gene targets. The matrix-scan script in RSAT identified several hundred potential target operons with statistically significant matches ($p < 0.0005$) to the RR_1586 binding site. Negative controls using permuted motifs (matrix-quality script) suggested a high false positive rate, probably because the AT-rich motif and scrambled derivatives are similar

Table 2.1 Locus tags of operon leaders and the upstream RR_1586 binding site.

Operon leader[a]	Annotation	Predicted RR_1586 site
CDR20291_2142[R,E]	hypothetical protein	AA**TTAAG**GTATAA**TTAAG**TTTT
CDR20291_3145[R.E]	Protease	AG**TTAAG**GTTTAA**TTAAG**ATTA
CDR20291_0818[R]	speADEB	TT**TTGAG**TTTTAG**TAAGC**TTTT
CDR20291_0879[R]	potABCD	AG**TAAAC**AAAATG**TTTAG**TAAA
CDR20291_1470[R]	transcriptional regulator	AA**TCGAG**GGAAAG**TTAAC**AAAA
CDR20291_1527[R]	hypothetical protein	AG**TTAAG**GTATAA**TTATT**TTAT
CDR20291_1565[R]	hypothetical protein	AT**TTAAG**CTTTAT**TTAAG**GTTA
CDR20291_1626[R]	Na(+)/phosphate cotransporter	TA**TTAAT**GTTTTG**TTAAG**TATA
CDR20291_1855[R]	tyrosine recombinase	AT**TTAGG**GAATAG**TTAGT**GATA
CDR20291_2009[R]	Na(+)/H(+) antiporter	GG**ATATA**GAATAG**ATAAG**AAAA
CDR20291_2188[R]	two-component system	TC**TTAAG**AAATAT**TTAAG**AATT
CDR20291_2890[R]	ABC transporter	AT**GTAAT**ATTTAC**TTAAG**GATT
CDR20291_3121[R]	phosphate transport (pst)	TA**TTAGG**ATTAAG**TTAAG**CAAG
CDR20291_3239[R]	ABC transporter	TG**TAAAG**GATATA**TTAAG**ACAA
CDR20291_2468[E]	neutral Zn metallopeptidase	AG**TTAAG**TGAATA**TTAAG**AGGA
CDR20291_0571[E]	peptidase	GA**TTAAG**TATGAA**TTAAG**CATG
CDR20291_0578[E]	chloride ion channel protein	TA**TTAAG**AATGGG**TTAAG**AGTA
CDR20291_0610[E]	ATP-dependent peptidase	GA**TTAAG**TATTTA**TTAAG**TATT
CDR20291_0884[E]	signaling protein	TA**TTAAG**TATTTA**TTAAG**TAAA
CDR20291_2143[E]	signaling protein	AA**TTAAG**GTATAA**TTAAG**TTTT
CDR20291_0477[P]	sleB	AA**ATAAG**CTAAAA**ATAAG**TAGA
CDR20291_0523[P]	cotJC1	TA**TTAAA**TATATA**TTAAG**GAGG
CDR20291_1583[P]	hypothetical protein	AA**TTAAG**GAGCAA**TTAAA**TGAT
CDR20291_3401[P]	spoIIR	TA**TTATG**AATAAA**TTAAA**TTTA
Consensus	-	DR**TTAAG**nwwwDR**TTAAG**nwww

[a] Superscripts in this column indicate the justification for including each gene as part of the regulon.

[R] conserved in RSAT footprint-scan searches

[E] exact match to the consensus search motif

[P] high similarity to the consensus and is involved in a phenotype relevant to RR_1586 including sporulation, germination, or self-activation.

a majority fraction of the AT-rich genome. We therefore manually selected several potential binding sites for further testing: ideal binding sites and sites upstream of sporulation/germination-associated genes or upstream of the *CDR20291_1583* operon (which includes the *CDR20291_1586* gene) according to the DOOR2 operon database (32). These genes, along with the 14 genes mentioned above, are listed in Table 2.1.

The above bioinformatic analysis identified associations between the RR_1586 binding site and downstream genes. Further analysis using GOMo (33) identified significant associations between the presence of the RR_1586 binding site and the biological functions of downstream targets. An enrichment of terms associated with ABC transport, ion transport, and phosphate transport was observed. These parallel

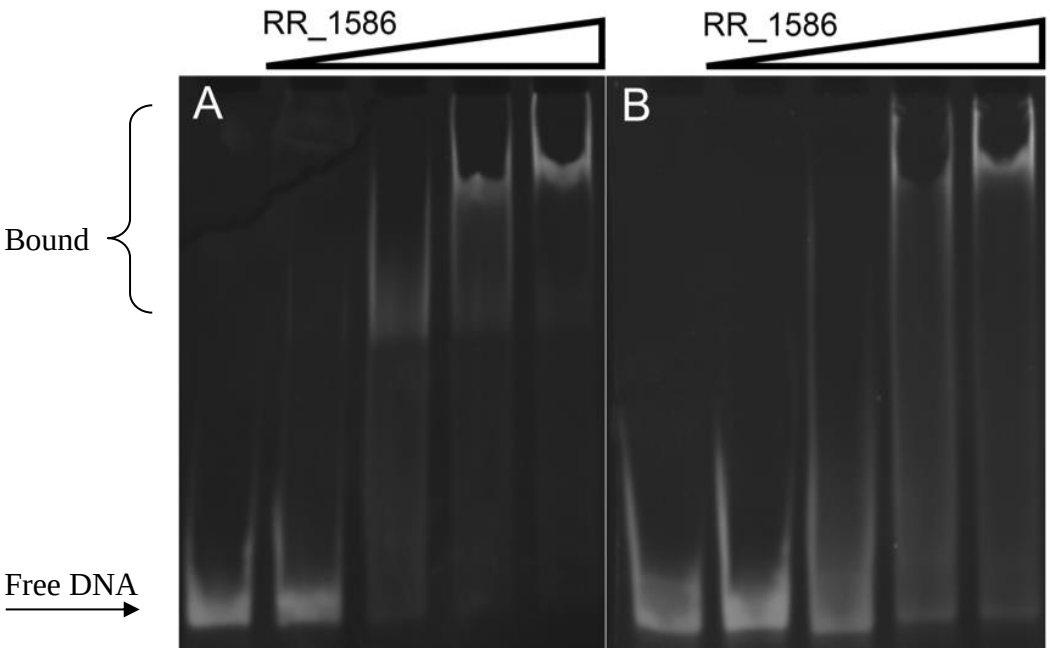

Figure 2.4 *In vitro* validation of RR_1586 binding sites
Titrations of RR_1586 against predicted genomic binding sites upstream of A) CDR20219_3145 and B) CDR20291_3121, with zero or two nucleotides, respectively, mismatching the search model. Each gel shows 500 nM DNA alone and in the presence of 1X, 5X, 10X and 20X molar equivalents of RR_1586. Oligo sequences used in these EMSAs are listed in Appendix E.

bioinformatic approaches both led to the same conclusion: RR_1586 appears to regulate genes involved in ion transport, particularly phosphate ion transport.

Experimental evaluation of putative regulon

In vitro binding of RR_1586 to all the promoter regions of genes in the proposed RR_1586 regulon (Table 2.1) was confirmed using EMSAs. Figure 2.4 shows the titration of RR_1586 against an ideal binding site upstream of *CDR20291_3145* (panel A) and a binding site with two mismatches upstream of *CDR20291_3121* (panel B), both identified as part of the conserved RR_1586 regulon.

To evaluate the potential for regulatory interactions at these binding sites, several putative target promoters were tested in a GFP reporter assay in *E. coli*. The promoter regions from *C. difficile* R20291 genes, including at least one and up to 10 codons of the open reading frame, were cloned in frame with super-fold GFP (34). *E. coli* expressed GFP from the tested promoters (Figure 2.5), while the induction of

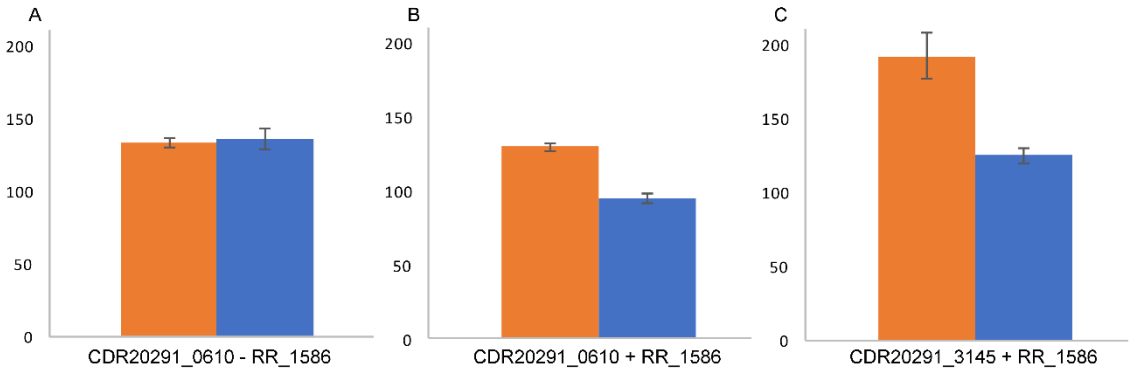

Figure 2.5 Expression of GFP from CDR20291 promoters in response to RR_1586 Cell density-normalized fluorescence (F/OD_{600}) of GFP protein was observed in *E. coli* Rosetta cells transformed with a reporter plasmid and/or RR_1586 expression vector (indicated below each graph). Samples were recorded in the absence (orange) and presence (blue) of 40 μM IPTG used to induce production of RR_1586. IPTG had no effect on fluorescence in the absence of RR_1586-encoding vector (panel A), but a decrease in fluorescence was observed for vectors reporting transcription from *CDR20291_0610* (panel B) and *CDR20291_3145* (panel C) promoters.

RR_1586 repressed expression of this GFP reporter gene. These results support the hypothesis of transcriptional regulation by RR_1586 at these identified sites.

Effects of phosphorylation on oligomerization and DNA binding

The results presented thus far define the components of a putative RR_1586 regulon but provide little evidence for the mechanisms governing regulation. OmpR-family proteins are often monomeric and form dimers upon phosphorylation to promote binding to their genomic targets (35, 36). We observed, however, that RR_1586 is purified as a dimer and shifts to an apparent tetrameric species in the presence of a small-molecule phosphoryl donor phosphoramidate (PA) (Figure 2.6) as judged by multi-angle light scattering in line with size-exclusion chromatography (SEC-MALS). RR_1586 with the phosphorylatable aspartate, Asp50, mutated to a glycine (D50G) was not affected by PA, indicating that the dimer-to-tetramer shift was dependent on phosphorylation of the active site aspartate. Secondary structure analysis by protein

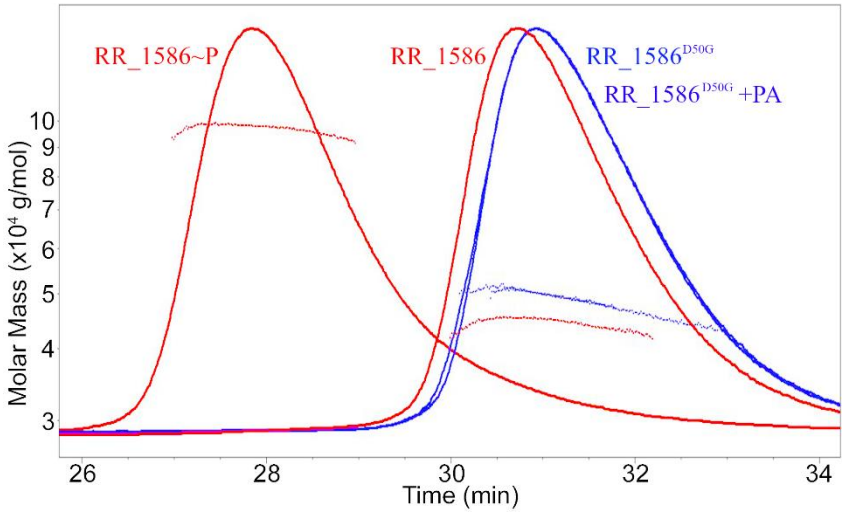

Figure 2.6 Phosphorylation-dependent RR_1586 oligomerization
SEC elution (curves) and light scattering (dots) profiles are shown. Addition of small-molecule phosphodonor phosphoramidate shifts the molecular weight of wild-type RR_1586 (red) from 57.5 to 119 kDa. In contrast, the apparent molecular weight of non-phosphorylatable RR_1586^{D50G} (blue) only shifts from 59.4 to 52.6 kDa. Monomeric RR_1586 is expected to be 28 kDa.

Fourier-transform infrared spectroscopy (37) showed no significant difference between the wild-type and D50G proteins, validating RR_1586^{D50G} as a well-folded, phosphorylation-negative control (Appendix D).

We also tested the effects of phosphorylation on DNA binding by the addition of PA to the EMSA reaction buffer. The most striking effect is that phosphorylation diminishes binding to sites that do not perfectly match the RR_1586 binding site, such as the one found upstream of *CDR20291_1583* (Figure 2.7B). PA could disrupt binding through ionic interaction with the RR_1586 DBD or DNA. However, all effects of PA were reversed by using RR_1586^{D50G} (Figure 2.7A and 2.7B), demonstrating the importance of phosphorylation of the active site aspartate. Binding to ideal sites, such as the one upstream of *CDR20291_2142*, was not disrupted by phosphorylation (Figure 2.7A). The amplitude of the electrophoretic shift was altered in some cases, although

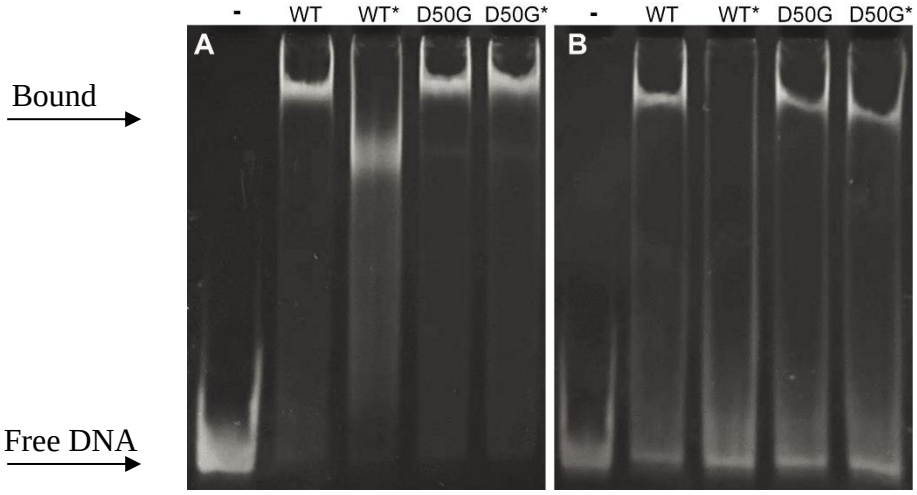

Figure 2.7 Phosphorylation-dependent DNA binding
The presence (*) of phosphoramidate A) has only minor effects on RR_1586 binding to a site upstream of CDR20291_3145, but B) disrupts binding to imperfect matches to the B1H-derived motif, such as the one upstream of CDR20291_1583. In both cases, use of the phosphorylation-deficient mutant RR_1586^{D50G} reverses these effects. Oligo sequences are listed in Appendix E.

the convoluted effects on shape and charge make it impossible to reliably interpret the significance of such a change (38). These results show that phosphorylation-dependent changes in binding occur in a sequence-dependent manner.

Discussion

One of the goals of this work was to accelerate the study of gene regulation by two-component systems in *Clostridioides difficile*, particularly the hypervirulent R20291 strain. Methods in synthetic biology and bioinformatics were used as a framework to predict the direct regulon of RR_1586. RR_1586 is encoded in a five open reading frame operon including *CDR20291_1583* to *CDR20291_1587*, annotated as a hypothetical protein, a putative DNA-binding protein, a putative lipoprotein, an RR, and an HK, respectively. Of these genes, *CDR20291_1586* (encoding RR_1586) is reported as essential for sporulation, and the orthologue of *CDR20291_1583* was differentially regulated to a detectable degree during germination of the CD630 strain (29, 39). No information was reported regarding the other genes in this operon, even though the assays tested the whole genome at apparently single-gene resolution. This potential connection to a biologically and medically important phenotype, and the lack of any other relevant information, made RR_1586 a suitable target for this study.

Extension of the B1H findings to genomic context

The hypothesis of self-regulation can be a very useful starting point for identifying downstream targets of transcription-regulating RRs (24), but initial attempts to observe *in vitro* binding of RR_1586 to regions upstream of the *CDR20291_1583* gene and neighboring operons failed to detect binding (data not shown). The B1H assay employed here screens for binding to a large library of randomized DNA sequences in parallel (25). It depends primarily on the design of a suitable DBD fusion construct, not

on the accuracy of an initial hypothesis. Several versions of a B1H assay have been applied to RRs (40-43). However, to our knowledge only two RRs have been characterized using the improved, ωRNAP fusion-based assay; both reports used full-length RRs (41, 42). B1H selection was unsuccessful when using the full-length RR_1586, but one DBD construct in our series of three met the conditions for successful selection and identification of a specific motif. With the aid of an empirically-derived DNA-binding specificity motif, we found that RR_1586 does indeed bind upstream of its own operon *in vitro* (Figure 2.7). The confirmed binding site partially extends into the coding region, which was not included in initial tests. This anecdote exemplifies the utility of a B1H screen for precisely defining potential genomic binding sites. This approach—predicting downstream targets from an observed specificity motif—circumvents the need for genetic manipulation required for approaches where binding sites are inferred from differentially expressed genes.

The DNA motif derived in the B1H assay is not a direct representation of the genetic regulatory element recognized by RR_1586 in *C. difficile*. The 5-7 base-specific and 3 AT-rich positions and overall length of the RR_1586 specificity motif are consistent with binding of a monomeric RR, but OmpR-family response regulator proteins often dimerize and bind direct repeats (35). Furthermore, the B1H assay utilizes a synthetic library coupled to a synthetic signaling pathway, and we were only able to identify a motif under conditions that exclude low-activity binding sites (44). This is in contrast to consensus motifs derived from transcriptomics data, which represent coevolving interactions between protein and DNA elements tuned to the needs

of the cell. These considerations led us to rely on comparative genomics strategies to identify putative gene targets.

Evolutionary conservation of regulatory function implies that the response regulator, its downstream target genes, and their respective binding sites will likely be conserved across related species. We evaluated every gene in the CDR20291 genome for the possibility that it fits these conditions of conservation among a set of *Peptostreptococcaceae* genomes, the family to which *C. difficile* belongs. This search identified operons including the *speADEB* and *potABCD* operons encoding spermidine biosynthesis and transport pathways, respectively. Many ABC transporter systems, and particularly *potABCD*, have been reported to be important for sporulation and/or germination (29, 39). RR_1586 binding sites are statistically correlated ($q < 0.05$) to gene ontology terms referencing ion transport and ABC-type transporter systems, suggesting that a possible role of RR_1586 is to regulate transport of ions. These conclusions may explain why *CDR20291_1586* was found to be essential for sporulation in a high-throughput screen (29), given that inorganic phosphate induces sporulation in *Clostridium perfringens* (45).

Interpretation of phosphorylation-dependent regulation

The main driver of two-component signal transduction is phosphoryl transfer between an HK and an RR. Phosphorylation of RR_1586 results in changes to the oligomeric state and DNA binding, implying possible regulatory mechanisms. Binding to an RR_1586 binding site positioned from -17 to +4 relative to the annotated translational start site of *CDR20291_1583* is likely to inhibit the advancement of transcriptional machinery and repress expression of downstream genes, including

CDR20291_1586. Phosphorylation of RR_1586 disrupted binding to this position *in vitro,* suggesting a potential feedback loop wherein RR_1586 represses self-expression until it becomes phosphorylated. This theme of phosphorylation-driven release of binding is repeated across most of the tested binding sites.

Reporter assays showed that expression of RR_1586 repressed expression of GFP from *CDR20291_0610* and *CDR20291_3145* promoters, which encode perfect matches to the RR_1586 consensus binding sites. Although the magnitude of repression is relatively low in *E. coli*, we anticipate a greater effect in *C. difficile* in the presence of native transcriptional machinery. *In vitro* binding to these sites, and to the other ideal RR_1586 binding sites, was not affected by phosphorylation. Regulation of these genes would be subject to changes in oligomeric state and in expression levels of RR_1586, with the latter potentially being mediated by phosphorylation-dependent self-regulation described above. Because the phosphorylation state is likely to change more rapidly

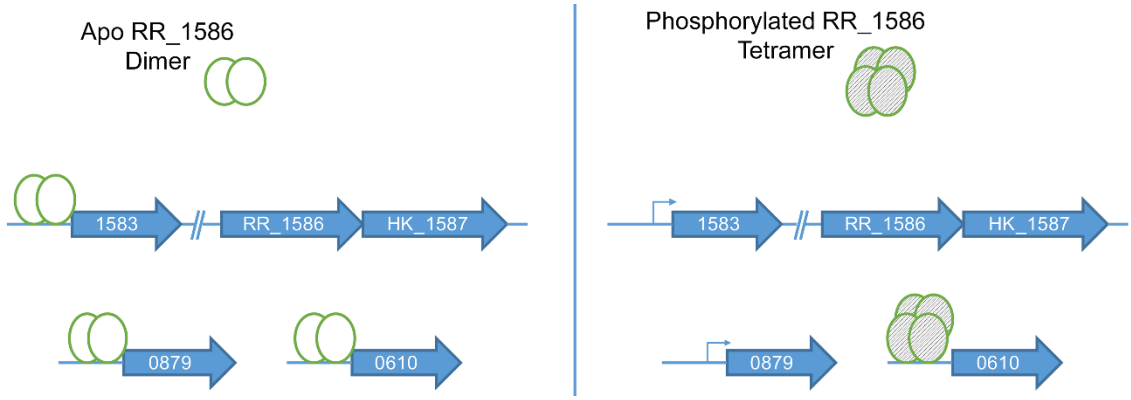

Figure 2.8 Working model of gene regulation by RR_1586
RR_1586 binds a larger set of predicted sites as dimeric apo protein (open circles, left) than it does as a phosphorylated tetramer (filled circles, right). Both forms of RR_1586 bind to ideal sites such as one upstream of *CDR20291_0610*. Binding sites such as those upstream of *CDR20291_0879* (leader of operon encoding potABDE system) and *CDR20291_1583* (leader of operon encoding RR_1586 and partner HK_1587) are bound by apo but not phosphorylated RR_1586. The position of the *CDR20291_1583* binding site (from -17 to +4 relative to the translational start site) suggests that binding of RR_1586 would disrupt transcription thus creating a phosphorylation-dependent transcription feedback loop.

than protein concentration, our results may also point to temporal differences in regulation of gene targets. The themes of regulatory mechanisms of the RR_1586 regulon are depicted in Figure 2.8. We propose that differential binding between dimeric apo-RR_1586 and tetrameric phospho-RR_1586 suggest both phosphorylation-dependent and -independent regulatory mechanisms for different gene promoter targets. These two mechanisms appear to correlate to binding sites with low and high identity to the B1H-derived binding site, respectively.

Phosphorylation-dependent changes in binding may reflect the conformational states accessible to apo and phosphorylated forms of the protein. One model of RR activation proposes that phosphorylation restricts receiver domain conformation from a mixed population to a nearly homogenous population of an activated conformational state (46). As phosphorylation stabilizes one conformation, absence of other conformations may preclude binding to certain DNA sequences. On the other hand, rather than constricting sequence space, phosphorylation may shift the center of sequence space recognized by RR_1586. This would manifest as changes in the DNA-binding specificity motif upon phosphorylation. We are actively studying this possibility and look forward to *in vivo* verification of these conclusions.

Finally, we emphasize the reliability of this experimentally-informed bioinformatics approach. For the samples we tested, *in vitro* binding was accurately indicated if the site surpassed statistical thresholds set in the bioinformatic searches. This is not surprising considering that the B1H assay selects for preferential binding to a 28 base pair sequence in competition with the entire *E. coli* genome—a simulation of the selectivity required for regulation in the native host. Similarly, using bioinformatic

constraints to identify conserved binding sites simulates evolutionary conservation of function, identifying targets most likely to be functionally conserved. Both the B1H and GFP reporter assays are performed without purified protein, meaning that a putative regulon could potentially be identified and initially validated even for proteins that are not amenable to overexpression and purification. We anticipate that this study will serve as a model for analysis of two-component gene regulation in *C. difficile* and other pathogenic bacteria.

Methods

Sources of strains and plasmids are listed by their associated method in Appendix E. Custom DNA oligonucleotides and primers (Sigma) used in this study are also listed in Appendix E. EmeraldAmp GT master mix (Clontech) was used for all PCR reactions unless explicitly stated. ZymoPURE midiprep, DNA Clean & Concentrator 5 (DNA C&C), and Oligo Clean & Concentrator (Oligo C&C) kits were purchased from Zymo Research. Restriction enzymes, ligases and RecA were purchased from NEB. Sequencing was performed by the Oklahoma Medical Research Foundation DNA Sequencing Core facility.

Preparation of RR_1586 and derivatives

Hexahistidine-tagged RR_1586 was purified for *in vitro* analysis using nickel affinity and size exclusion column chromatography. Proteins were expressed in BL21(DE3) Rosetta cells from pSGC plasmids constructed in the laboratory of Dr. Steve Almo at Albert Einstein College of Medicine. Cells were lysed by sonication in 20 mM HEPES pH 7.5, 300 mM NaCl, 20 mM imidazole, and 5% glycerol. The lysate was loaded on to a 5 mL hand-poured MCLAB Ni-NTA column and washed with lysis buffer. RR_1586 was eluted in lysis buffer with increasing imidazole concentrations in

100 mM steps (100-500 mM imidazole). Size exclusion chromatography was performed in the OU COBRE Protein Production Core facility using a 24 mL Superdex 200 Increase column (GE Healthcare) equilibrated with 20 mM HEPES pH 7.5 and 150 mM NaCl. Fractions were pooled following SDS-PAGE analysis. Protein concentrations were estimated using the BioRad Protein Assay reagent standardized against BSA. Phosphorylated RR_1586 was obtained by incubating 50 μM pure protein with 50 mM phosphoramidate in 20 mM Tris pH 8.0, 50 mM NaCl, and 10 mM $MgCl_2$ for 10 minutes at room temperature. Phosphoramidate was synthesized following published procedures (47).

Primer-directed mutagenesis was used to substitute the RR_1586 phosphorylatable aspartate with glycine (D50G). Primers are listed in Appendix E. LongAmp Taq 2x master mix (NEB) was used as recommended by the manufacturer to incorporate the mutation during whole-plasmid PCR amplification. The mutation was confirmed by sequencing. The RR_1586^{D50G} protein was expressed and purified as described above for wild-type RR_1586.

Multi-angle light scattering and protein Fourier-transform infrared spectroscopy

A MiniDawn Treos (Wyatt) multi-angle light scattering instrument in line with a Superdex 200 Increase SEC column was used to measure the molar masses of purified RR_1586 and its derivatives. We also measured infrared absorbance spectra of RR_1586 and RR_1586^{D50G} to detect changes in the amide I and amide II bands associated with protein secondary structure. We used a Bruker Confocheck Tensor II instrument fitted with an AquaSpec I sample cell (Bruker) and temperature regulated by a Ministat 125 (Ruber) water bath set to 23 °C. Protein was equilibrated into 20 mM

HEPES pH 7.5 and 150 mM NaCl by SEC before analysis. Opus 7.5 software was used to evaluate secondary structure features using manufacturer's protocols.

Construction of a "prey" plasmid library and "bait" plasmids for bacterial one-hybrid assay

A library of plasmids encoding *his3* and *ura3* expressed under control of randomized 28-mers was constructed based on published methods (48). The complementary strand to a commercial 71-mer oligo was synthesized by PCR. *Not*I-generated restriction fragments were separated on a 20% polyacrylamide gel. The larger band was excised and digested with *Eco*RI. Final purification by Oligo Clean and Concentrator kit retained the desired sticky-ended 28-mer library but not the 6-nucleotide byproduct. The insert was ligated into pH3U3-mcs and the resulting library was transformed directly into the counter selection strain. We performed counter selection three times using liquid gel medium instead of solid agar (49). The final plasmid library was tested using positive and negative control plasmids (pB1H2w2-mutOdd and pB1H2w2-Zif268).

The omega subunit of RNAP was fused to full-length RR_1586 and three constructs of its DNA-binding domain (at positions Arg124, Ser131 and Gln151) using sequence and ligation independent cloning (50). PCR amplification of vector pB1H2w2-prd (NEB, Long-Amp Taq 2X Master Mix) and insert introduced complementary overhangs to be recombined *in vitro* by RecA. Plasmid construction was confirmed by Sanger sequencing. Plasmid DNA from overnight cultures in 50 mL of LB was isolated and concentrated by ethanol precipitation in preparation for selection (48).

Bacterial one-hybrid selection and data analysis

Selection proceeded as described previously (25). The RR_1586 bait and prey library were cotransformed by electroporation into USO cells and plated onto minimal medium lacking histidine and uracil at a density between 5×10^4 and 5×10^5 CFU/cm^2. Plates were wrapped with Parafilm M and incubated at 37 °C until colonies were large enough to be counted and picked for colony PCR. The PCR product was purified by phenol-chloroform extraction and sequenced. MEME (v4.12.0) (51) was used to identify over-represented motifs in RR_1586-selected sequences. Zero or one instance of the motif were allowed per sequence (using the "-zoops" option) on the given or complementary strand ("-revcomp") with a minimum width of 3 nucleotides ("-minw 3"). An E-value threshold of 0.005 was set. All other parameters were left at default.

Genome scanning and comparative genomics

The accession numbers for genome assemblies analyzed in this study are listed in Appendix C. The pattern-search and footprint-scan scripts from the Regulator Sequence Analysis Tools (RSAT) suite were used for single genome scans and comparative genomics approaches (30). Search models for the direct repeat were constructed by duplicating the first eleven positions in the search strings or matrices.

Multispecies GOMo (33), part of MEME-Suite, was used to identify statistically significant correlations between promoters with RR_1586 binding sites and gene ontology terms associated with downstream genes from 13 *Peptostreptococcaceae* genomes with RRs highly similar to RR_1586. The GO terms assigned to CDR20291 proteins, by BLAST2GO (52), were also to their respective orthologues. A union of terms from members of an operon was assigned to the leading gene to account for

species-specific operon structure. Orthology and operon structure were inferred using RSAT (30), except CDR20291 operon predictions are from the DOOR2 database (32). GO maps for all genomes were combined into a single input file for GOMo analysis (33).

Electrophoretic mobility shift assays

Binding of full-length RR_1586 to DNA was observed *in vitro* using EMSAs. Pairs of synthetic single-stranded oligonucleotides (Appendix E) in 10 mM Tris pH 8.0 and 50 mM NaCl were annealed at 95°C for 5 minutes and passively cooled to room temperature. Titrations of protein and 5 pmol DNA in 10 μL of 10 mM Tris pH 8.0, 50 mM NaCl, and 10 mM $MgCl_2$ were incubated at room temperature for 10 minutes, then 5 μL of 50% glycerol were added to aid in loading. Samples were loaded onto pre-run 10% native polyacrylamide gels with 0.5X TBE as the running buffer. Gels were run at 120 V for one hour with the gel box submerged in ice. DNA was stained by rocking the gel in running buffer spiked with ethidium bromide for 5 minutes and briefly washed with fresh buffer before imaging. Images were captured using a Gel Logic 100 system with a UV transilluminator.

Recombinant reporter assay

The tetracycline biosensor plasmid pJKR-L-tetR (34) was repurposed as a GFP reporter of transcription in *E. coli*. Restriction sites were introduced to replace the existing ribosomal binding site with inserts spanning the upstream region and first few codons of *C. difficile* R20291 genes. TetR promoters were kept intact to serve as an anhydrotetracycline-inducible control of GFP expression and to screen for properly integrated inserts. All vectors were confirmed by sequencing.

RR_1586-dependent expression of GFP was tested in *E. coli* BL21(DE3) Rosetta cells. Saturated overnight cultures were diluted 100-fold in fresh LB and shaken at 37 °C for 3 hours. RR_1586 expression was then induced with 40 μM IPTG. Cell growth and GFP fluorescence were monitored as described (34). Cell density-normalized fluorescence at the final 15 hour time point is reported.

Acknowledgements

The experiments in this chapter that used purified protein were performed by or in collaboration with Dr. Smita Menon who was a co-author on the published version of this work (27). We were both fully involved in the design, analysis, and interpretation of these assays.

This work was funded by the Price Family Foundation (AHW, GBRA and EAK), the Oklahoma Center for the Advancement of Science and Technology (HR18-110, AHW), and Grayce B. Kerr Endowment funds (AHW). The OU Protein Production Core was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103640. The funding sources had no role in study design, data collection and interpretation, or the decision to submit the work for publication. We thank Dr. Fares Najar (OU Chemistry and Biochemistry Bioinformatics Core) for his contributions to genomic analyses early in this project. We are grateful to Dr. Ann Stock for a gifted sample of phosphoramidate. Phosphoramidate synthesis was also performed in house by Robert Fogle and Dr. Erwin Abucayon. We also thank Dr. Steve Almo for the plasmid encoding RR_1586 and Dr. Jimmy Ballard for *C. difficile* genomic DNA.

Chapter 3: Toward high-throughput bacterial one-hybrid-bioinformatics analysis of OmpR response regulators.

Successful B1H selection relies on an appropriate balance of several factors: the size and diversity of the prey library, the stringency of selection, and the activity and expression levels of the bait fusion protein. A sufficiently large and diverse library was made and was validated using RR_1585 (as described in Chapter 2), and is universally applicable to all active bait proteins. Selection stringency is determined by the amount of 3-AT in the selection medium, which inhibits the activity of the *his3* gene product. Survival, therefore, requires more copies of this product, which are produced more efficiently by tightly interacting bait-prey pairs. The activity and expression levels of the bait protein are both controlled by the cloning process (Figure 3.1). The bait protein can be cloned into one of three vectors so it can be expressed from a dual promoter

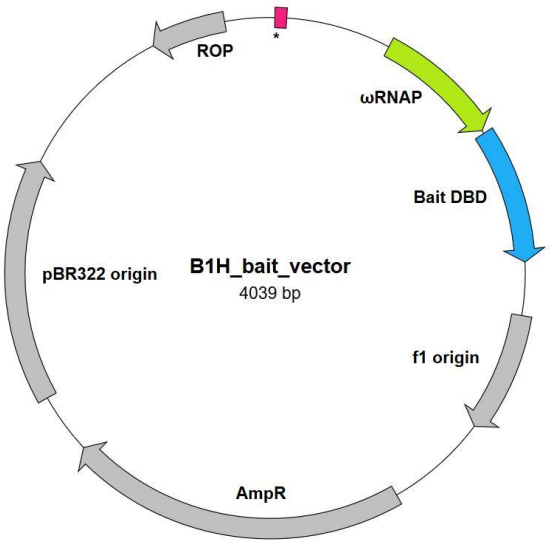

Figure 3.1 Diagram of bait plasmids
Bait fusion proteins can be expressed to varying degrees using the pB1H2w series of plasmids, each having a different promoter of varying strength (*). Control of expression by altering chemical inducer concentrations can cause changes to metabolism that interfere with the selection process. While increased expression can increase the chances of binding between the bait and prey molecules, it can also be fatal as the bait protein binds to and alters expression of genes in the *E. coli* genome. This figure was prepared using a Savvy application server (53).

(pB1H2wL), a single native promoter (pB1H2w5), or a weakened, mutant promoter (pB1H2w2) (25). This is preferred over varying concentrations of chemical inducer, which can affect the growth rates of the cells independent of protein expression. The activity of the bait protein seems to depend on the length of the linker between the DNA-binding domain and the ωRNAP domain, as described for RR_1586 in Chapter 2.

This chapter summarizes progress toward a more generalized B1H selection strategy for OmpR-family RRs. It also describes custom Perl scripts designed to perform the data analysis and bioinformatic processes described in Chapter 2 in an automated and customizable pipeline.

Results

The amino acid sequences of the OmpR-family RRs encoded by *Clostridioides difficile* R20291 (8) were aligned using PROMALS, a secondary structure-informed multiple sequence alignment server (54). A sample of this alignment is shown in Figure 3.2. This specific alignment strategy ensures that points of fusion can be designed for individual proteins in context of the secondary and tertiary structural features common to the entire family of proteins.

Constructs of RR_1522 and RR_1677 were prepared to be equivalent to the RR_1586 Ser131 fusion junction, but no selection was detected in the B1H assay. A longer construct of RR_1586 (beginning at Arg124) was cloned into pB1H2ω2 with a shortened linker. This construct showed similar selection levels and produced a motif equivalent to the motif discussed in Chapter 2 (Figure 3.3A). Several RR_1677 constructs were tested with the shortened vector. The RR_1677 Ser144 construct showed positive selection, but to a lesser extent than either of the RR_1586 constructs. A DNA-binding specificity motif for RR_1677 was identified among selected sequences (Appendix A), but the probability of achieving a similar result by random chance (E-value), though acceptable, was not as low as expected (Figure 3.3B)

```
CDR20291_1677  121 KELLVRVSALLRRVAKDDS-----SVKSSE 145
CDR20291_0860  106 NELISRIKALLRRYNVAS--------NVNE 127
CDR20291_1522  107 LEVVARVKTQLRRYMRYNNSYEQQSIIVNE 136
CDR20291_1586  105 EELVARVYAILRTNGKIK--------ERNG 126
Consensus_ss:      hhhhhhhhhhhhh               e

CDR20291_1677  146 IVSPPFILDIDKRKLFKNGKEIELTPTEFS 175
CDR20291_0860  128 LSSNNITIKLLENRVFKGEFEVELTTAEYK 157
CDR20291_1522  137 YDIKGLIINKETHKCSLFGKEVALTPIEFS 166
CDR20291_1586  127 LEFKSLYLDTLEKRVYIEKEEIKLQNKQFN 156
Consensus_ss:      eee  eeee   eeeee  eeee  hhhhh

CDR20291_1677  176 IVKYLISNAKQSLSRDQILDEVWG-TNYLY 204
CDR20291_0860  158 LLCLFMKNKNIVLTRKNILDKLWDGNGSFI 187
CDR20291_1522  167 ILWYLCEHQGKVVPSEELFEAVWG-EKYLD 195
CDR20291_1586  157 LLEYFVLNKGSILLKEQIYDRIWG-IDSDA 185
Consensus_ss:      hhhhhhh       hhhhhhhh
```

Figure 3.2 PROMALS alignment of OmpR-family RRs
This sample of the alignment spans from the α5 helix of the REC domain into a portion of the winged helix-turn-helix domain (wing and helix 3 not shown). Points of fusion for some of the tested constructs are **bolded** or underlined if they were tested in the **full-length** or shortened vector, respectively.

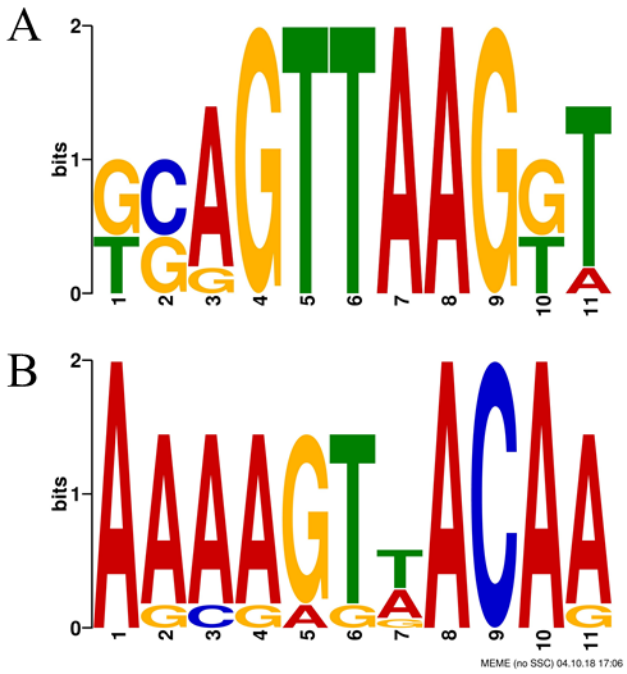

Figure 3.3 DNA-binding specificity motifs
Motifs selected by A) RR_1586 Ser131 and B) RR_1677 Arg144 constructs in a B1H assay. E-values for these motifs are , respectively.

Two Perl scripts (B1H_analysis.pl and prep_go_tables.pl) were created to streamline data processing. The strategy used for analysis of RR_1586 in Chapter 2 uses three major functions: 1) meme (MEME-Suite), to identify a common motif among input sequences, 2) footprint-scan (RSAT), to identify conserved gene targets of a putative regulon, and 3) GOMO (MEME-Suite), to identify biological functions associated with a given motif. The prep_go_tables.pl script maintains a directory of input files required for GOMO analysis, including nucleotide sequences and operon predictions—both extracted from RSAT—and gene ontology annotations prepared using the InterProScan method in BLAST2GO (52). B1H_analysis.pl identifies sequencing result (.seq) files in a given directory and performs, according to options

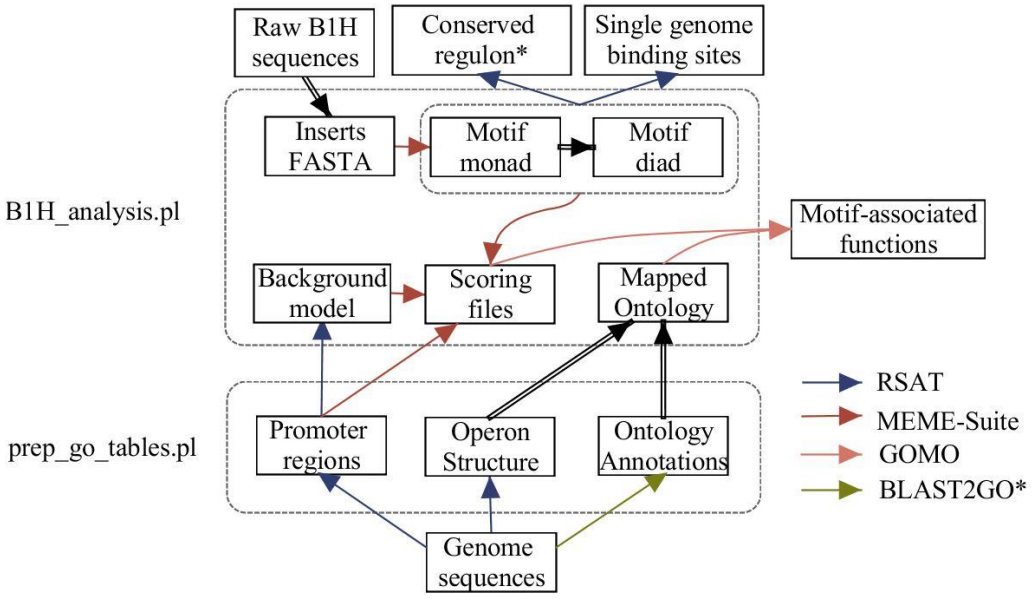

Figure 3.4 Diagram of custom bioinformatics pipeline scripts
Two Perl scripts were written to coordinate several publicly available bioinformatics utilities (colored arrows) to streamline the process of B1H data analysis and interpretation. The 28 bp prey inserts selected during the B1H assay were previously retrieved manually from the 400 bp sequencing results. New functions were written (black double arrows) to parse these and other files to be compatible with subsequent analysis steps. Solid boxes represent data files used and/or produced throughout the process. This figure was created using PathVisio 3.3.0 (7).

defined by the user, all aspects of analysis described in Chapter 2. A flowchart of the scripted pipeline is summarized in Figure 3.4. A more complete description is found in the Methods section of this chapter, and the full scripts are included in Appendices F and G.

These scripts were validated by analyzing sequences selected by the Ser131 and Arg124 constructs of RR_1586. This analysis confirmed previous results, but with greater resolution. Fewer GO terms were associated with the RR_1586 Ser131 motif than in the previous analysis (3 versus 11), but the statistical confidence in the results improved (p- and q-values decreased) by an order of magnitude. This could be the result of using genome-wide GO annotations for all genomes, rather than extending the *C.*

difficile R20291 annotations to predicted orthologues as was done in Chapter 2. Regardless of these differences, all results point to RR_1586 regulating phosphate transport.

The RR_1677 motif was also analyzed using the new custom scripts. Orthologues of RR_1677 were found among 24 of the 26 tested genomes, compared to 17 for RR_1586. The high conservation of this RR among related species is consistent with the observation that it is essential to cell viability (29), and therefore any changes to, or loss of, the encoding gene would require costly compensations to avoid cell death. Genomes with orthologues of RR_1677 are listed in Appendix C.

From these analyses, RR_1677 appears to regulate a broad range of functions. The gene ontology terms associated with the RR_1677 motif are listed in Table 3.1. One of the striking features of these results is that the top hits all have identical q-values, which are calculated by correcting the p-values to account for repeated testing. The p-values are derived by randomly shuffling the motif-to-gene similarities and calculating new scores for the scrambled motif-to-gene-to-function associations. The

Table 3.1 Gene ontology terms associated with the RR_1677 motif

GO term	Score	q-value[a]	Definition
GO:0019867	4.4E-02	4.9E-06	outer membrane
GO:0009254	4.9E-02	4.9E-06	peptidoglycan turnover
GO:0003735	7.6E-02	4.9E-06	structural constituent of ribosome
GO:0042026	8.0E-02	4.9E-06	protein refolding
GO:0006412	1.0E-01	1.3E-04	translation
GO:0005840	1.3E-01	4.0E-03	ribosome
GO:0045892	1.3E-01	4.9E-03	negative regulation of transcription
GO:0051539	1.4E-01	5.6E-03	4 iron, 4 sulfur cluster binding
GO:0016021	1.4E-01	9.1E-03	integral component of membrane
GO:0003690	1.5E-01	1.4E-02	double-stranded DNA binding
GO:0008966	1.5E-01	2.2E-02	phosphoglucosamine mutase activity
GO:0005737	1.6E-01	3.6E-02	cytoplasm
GO:0005515	1.6E-01	4.2E-02	protein binding

[a] q-values were derived from 10^5 negative control calculations.

empirical frequency of finding an association with a score greater than or equal to the non-shuffled scores is reported as the p-value and adjusted for multiple testing as the q-value. By default 10^3 negative controls were calculated, but gave identical p-values for the five top hits. Rerunning the analysis with 10^4 and 10^5 control sets gave matching p-values for the top four and two hits, respectively. Furthermore, among these three analyses, the p-values of the top hits changed by exact factors of 10 while other p-values changed as little as 1-2%. This trend shows that the shared highest p- and q-values are the limits of the empirical analysis. That is to say, out of 10^5 randomized data sets, no associations scored higher than the association to the outer membrane and peptidoglycan turnover. A two-fold difference in p-values of the first four hits is lost during multiple testing correction.

Noteworthy trends are also seen among the definitions of the gene ontology terms listed in Table 3.1. RR_1677 was associated with several related processes of transcription, translation, and protein folding. These included genes encoding structural components of the ribosome. It is also worth noting that phosphoglucosamine mutase is a key enzyme in synthesizing the peptidoglycan component of the cell wall (55). Thus, the majority of the gene ontology terms associated with RR_1677 contribute toward one of the two themes of cell wall synthesis and protein synthesis.

Conclusions

A general solution to determining RR-DNA binding specificity would be a significant advancement to the field of two-component signal transduction and to all fields involving prokaryotes. The B1H assay has the potential to produce the data required for such a solution if a sufficiently large and diverse population of RRs can be tested efficiently. To this end, B1H selection by the RR_1586 Ser131 construct was

mimicked using other RRs from *C. difficile* R20291. These DBD constructs begin at a loop connecting the first and second strands of an antiparallel β sheet, known as the β platform of OmpR proteins. Although RR_1586 appears to tolerate this disruption to the β platform, the other RRs may fail B1H selection because their DBD structure is destabilized. To achieve a similar DBD-ωRNAP distance as the RR_1586 S131 version while also including the entire β platform, the longer RR_1586 Arg124 construct was cloned into a vector with a shortened linker (pB1H2w2Δ). Although the RR_1586 Arg124 construct had previously failed B1H selection, this new design with the shortened vector produced a similar number of surviving colonies as RR_1586 Ser131.

The elongation of the DBD construct and compensatory shortening of the vector appear to be a step toward an OmpR protein family-wide strategy for successful B1H selection. With this strategy, RR_1677 selected a detectable, though lesser quality, motif. The differences in statistical parameters between RR_1586 and RR_1677 motifs can be attributed to two factors. First, although a similar number of colonies were sequenced for both proteins, many of the RR_1677 colonies were clonal copies containing the same prey plasmid. In fact, one prey vector was identified during multiple independent RR_1677 selections, but never identified while testing other RRs. The presence of many clonal copies indicates a survival or reproductive advantage associated with a highly active bait-prey pair. This hypothesis is supported by the observation that the calculated motif was found in the prey sequence. The second factor that affects E-values is the calculation of the negative control. The input nucleotide sequences are scrambled and the probability of finding a similar motif from these scrambled sequences is calculated. The RR_1677 motif is very “A-rich” in the given

strand. This simple binding motif will have low variability even after random rearrangement, reducing the impact of the negative control calculations. Selection by RR_1677 may yet be improved by using the pB1H2ω5 vector, which expresses bait proteins from a stronger promoter. Stable, soluble full-length RR_1586 expresses exceptionally well in *E. coli*, much more so than full-length RR_1677, as observed during purification. If this difference holds for their bait fusion constructs, then RR_1677 may require stronger expression to achieve similar concentrations of effective bait protein.

The development of scripts to automate the data processing and analysis will also aid in the advancement toward a general solution to DNA specificity. The complete processing from raw sequencing data to associated functions (excluding prediction of genes in the regulon) through the B1H_analysis.pl script now takes only a few minutes. These scripts also ensure that analysis is consistent across data sets. For example, analysis of the RR_1677 motif using footprint-scan identified around 1,000 operons in the conserved, putative regulon. This could indicate that RR_1677 is a master regulator or there is an anomaly in the analysis process. However, because identical analysis of the RR_1586 motif identified a regulon of only 14 operons, we can more readily believe the result that RR_1677 appears to have at least a large regulon. A large putative regulon and an association to genes involved in translation and cell wall synthesis (Table 3.1) are consistent with experimental results indicating that RR_1677 is essential for *C. difficile* viability (29).

In conclusion, significant progress has been made in preparing tools to elucidate the structure-function relationship between RR primary structure and DNA binding

specificity and biological function. In addition to progress in technical aspects of the B1H assay and data analysis, the DNA-binding characteristics and putative regulons and functions of two OmpR proteins from *C. difficile* R20291 have been determined. These results pave the way for future studies involving broader application of the B1H-bioinformatics approach to potentially elucidate organism- or protein family-wide models of DNA-binding specificity.

Methods

All cloning and B1H selections were performed as described in Chapter 2.

Description of custom scripts

A Perl script was written to enable consistent application of bioinformatics programs to the analysis of B1H raw data. The script, entitled B1H_analysis.pl, simplifies a series of about one dozen steps into a single script. It recognizes the prey vector sequences flanking the randomized insert and extracts and stores them in two separate FASTA files as either a complete set or a set purged of redundant clonal copies. The purged data set is submitted to MEME for motif identification. The identified motif—or an automatically-generated direct repeated motif of a customizable size—is then used to search for matches in the CDR20291 genome and/or to search for association to a conserved function (GOMO). This script uses RSAT to identify genomes with orthologues of the bait RR and builds a table mapping the operon leader genes to gene ontology terms describing all constituents of the operon. Gene ontology annotations were derived from BLAST2GO's InterProScan search for all genes. A second script (prep_go_table.pl) was written to parse BLAST2GO output files and information from the RSAT database, such as operon predictions and upstream nucleotide sequences, into a format compatible with GOMO. In short, B1H_analysis.pl

performs all the analysis described in Chapter 2 except for RSAT's footprint-scan. It does return a command that can be copied into the terminal to run footprint-scan, a process that takes more than six hours using all the resources available on a MacPro.

B1H_analysis.pl and prep_go_table.pl have built-in help functions which can be accessed from the command line interface using the –h or –help flags. These scripts can be found in Appendices F & G.

Acknowledgements

I thank Dr. Fares Najar (OU Chemistry and Biochemistry Bioinformatics Core) for his prescient diagnosis of a bug in a previous draft of my scripts. This work was funded by the Price Family Foundation (AHW, GBRA and EAK), the Oklahoma Center for the Advancement of Science and Technology (HR18-110, AHW), and Grayce B. Kerr Endowment funds (AHW).

Chapter 4: Protein FTIR spectroscopic indicators of response regulator phosphorylation by small molecules.

Phosphoryl transfer from an HK drives the transition between the "on" and "off" states of an RR and its regulatory output. Although two-component signaling transduction pathways don't always fit two discrete states, the study of apo and phosphorylated forms of an RR can define the two extremes of a spectrum of regulatory outcomes. Phosphorylated RRs can be prepared *in vitro* by using a purified cognate HK protein and ATP or by using small-molecule phosphoryl donors such as phosphoramidate and acetylphosphate (56). Obtaining a pure and active HK construct is non-trivial, and not all RRs can accept phosphoryl groups directly from small molecules donors. Therefore, when testing the effects of phosphorylation on an RR, it is important to distinguish between the lack of phosphorylation versus no phosphorylation-dependent change in the tested property of the RR.

Most existing methods to detect RR phosphorylation involve radioactive tracers and/or gel electrophoresis (57, 58). One notable exception is the use of intrinsic fluorescence spectroscopy to detect the phosphorylation of the model RR CheY using small molecule donors (56, 59). CheY has a single tryptophan residue adjacent to the phosphorylatable aspartate residue. The intrinsic fluorescence signal from this residue is sensitive to the active site phosphorylation state (56). This allows the reaction to be probed in real time under reaction conditions rather than using a separate set of conditions for analysis. This method has other advantages (60), but it applies only to the subset of RRs with tryptophan residues near the active site, which limits its utility.

FTIR spectroscopy uses wavelengths sensitive to vibrational modes of molecules instead of the electronic transitions measured by fluorescence in the UV and

visible spectrums. Protein FTIR usually emphasizes the normal vibrational modes of the amide bonds of the main chain. The sensitivity of these vibrational modes to secondary structure is used to determine the distribution of α helices and β strands of a protein in solution (37). Additionally, phosphorylation of serine, threonine, and tyrosine groups have been detected directly by the formation of new absorbance bands around 1060 and 1080 cm^{-1} corresponding to the vibrational modes of the phosphoryl moiety (61). Likewise, phosphorylation of an aspartate residue in a sarcoplasmic reticulum Ca^{2+} ATPase enzyme is thought to contribute to an increase in absorbance at 1131 cm^{-1} (62). The identification of an FTIR band indicative of response regulator phosphorylation could extend the advantages of optical methods to all response regulators even in the absence of fluorescent side chains. This chapter presents the identification of phosphorylation-specific changes in the RR_1586 FTIR spectrum. Follow-up experiments are proposed to validate this technique for RRs in general.

Results

The FTIR vibrational spectra of apo and phosphorylated RR_1586 were measured to identify spectral features that could potentially indicate RR phosphorylation. The most prominent features of a protein FTIR spectrum are the amide I and amide II bands which indicate the relative contributions of α helices or β strands to the overall protein structure (Figure 4.1A). An increase in α-helices and a slight decrease in β strands was detected upon phosphorylation of RR_1586 (Table 4.1). The magnitude of these changes in the secondary structure are larger than variations between samples across a relevant range of RR_1586 concentrations (Table 4.1).

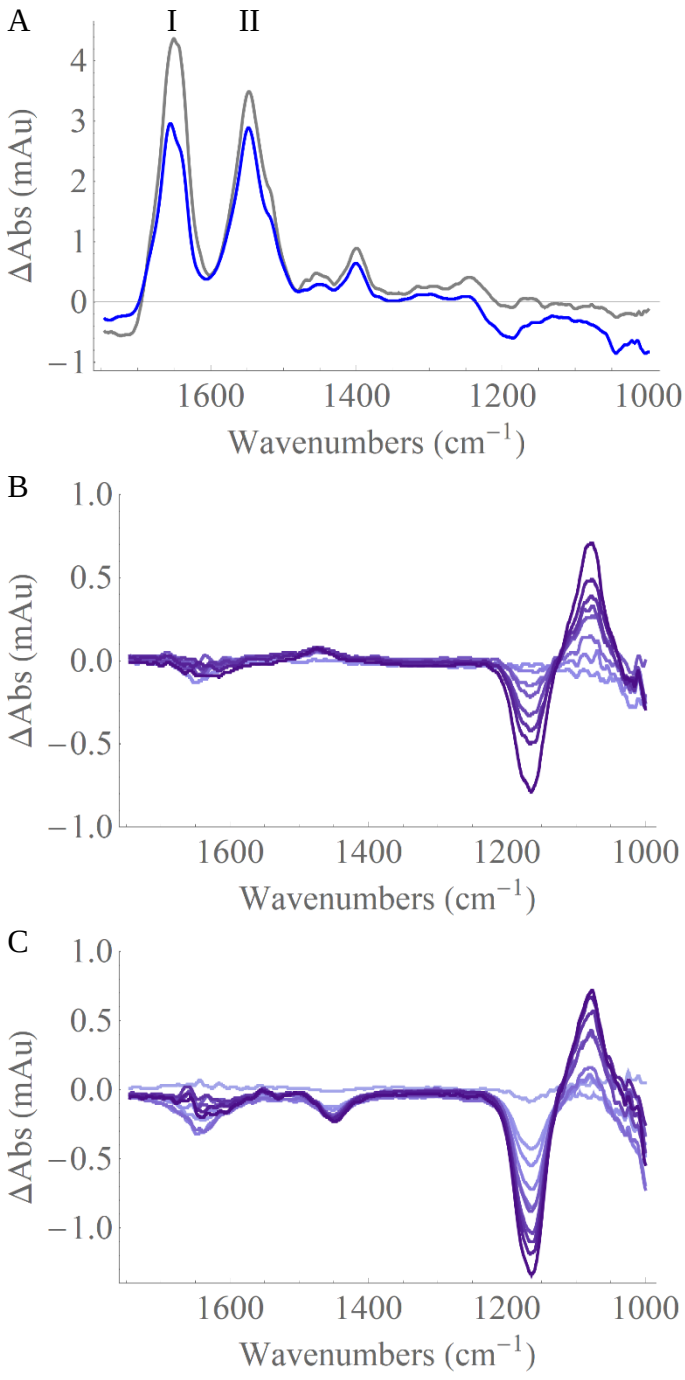

Figure 4.1 FTIR spectra of apo and phosphorylated RR_1586
A) The FTIR spectra of apo RR_1586 (gray) and phosphorylated RR_1586 (blue) are overlaid. The most prominent peaks, amide I (~1650 cm^{-1}) and amide II (~1550 cm^{-1}), indicate changes in secondary structure upon phosphorylation. Incubation of RR_1586 (B) and RR_1586^{D50G} (C) with PA was observed over time using FTIR. Spectra were collected every 3 min. Darker lines indicate later time points. A decrease in absorbance of the wild-type sample around 1450 cm^{-1} contrasts with the increase in absorbance in this region for the non-phosphorylatable mutant.

Table 4.1 Effects of phosphorylation on secondary structure of RR_1586

Protein	α helix (%)	β strand (%)	Concentration range (mg/mL)
Apo RR_1586	40 ± 1	15.9 ± 0.3	1.3-6.3
Phospho RR_1586	46 ± 2	13 ± 1.8	2.5-3.5

A new peak corresponding to the phosphorylated moiety could not be assigned based on these spectra because of the noise between samples and the low mass of the phosphoryl group relative to the protein. However, reaction-induced differential FTIR spectroscopic methods can be used to increase sensitivity by observing changes in a single sample over the course of the reaction rather than comparing two samples (37). To this end, two series of difference spectra, for RR_1586 and the non-phosphorylatable RR_1586^{D50G} protein, were measured in the presence of PA (Figures 4.1B and 4.1C). Peaks in these spectra show only changes in absorbance relative to the time-zero reference sample. Two negative peaks (1645 and 1165 cm^{-1}) and one positive peak (1078 cm^{-1}) were seen for both wild-type RR_1586 (Figure 4.1B) and non-phosphorylatable RR_1586^{D50G} proteins (Figure 4.1C). The low wavenumber peaks

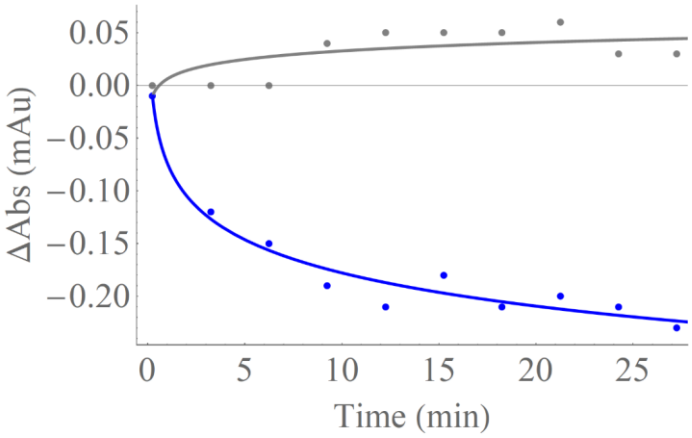

Figure 4.2 Changes in absorbance at 1450 cm^{-1} during phosphorylation
The changes over time in absorbance at 1450 cm^{-1} for wild-type RR_1586 (blue) and RR_1586^{D50G} (gray) in the presence of PA. These are preliminary data and will be refined as more replicates are performed.

increased linearly through time, but the 1645 cm^{-1} peak did not show a discernable trend. Finally, contrasting positive and negative peaks in the mutant and wild-type proteins, respectively, appeared around 1450 cm^{-1}. These changes in absorbance are plotted versus time in Figure 4.2.

The reaction-induced difference spectra detect changes in all chemical species, including reagents, products, and buffer components. Changes in the spectra can occur from protein phosphorylation and from protein-independent hydrolysis of PA. The spectra of fresh and hydrolyzed PA samples were compared to identify their contributions to the reaction spectra (Figure 4.3). Fresh PA exhibits strong bands at 1453 and 1164 cm^{-1}. Upon extensive hydrolysis of the sample by heating, the 1164 cm^{-1} band nearly completely disappeared, and an intense new band at 1078 cm^{-1} appeared (Figure 4.3). The 1453 cm^{-1} band also shifts upward three wavenumbers and increased intensity by about 20% as measured by both maximum absorbance and integrated peak

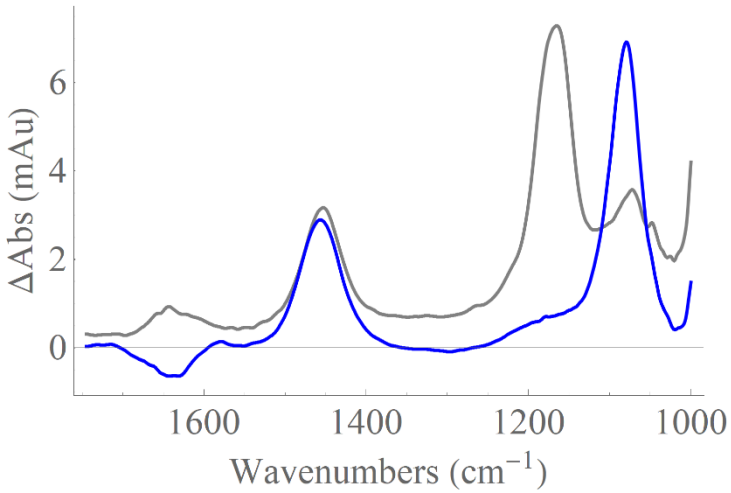

Figure 4.3 Spectra of pre- and post-hydrolysis phosphoramidate
Buffer-subtracted spectra of phosphoramidate before (gray) and after (blue) hydrolysis in the absence of protein. Hydrolysis was induced by incubation at 65 °C for 1.75 hr. The sample was cooled on ice and centrifuged to collect the condensate.

area. A negative peak also appeared near 1650 cm^{-1} and may be the result of changes in the buffer which contributed to a peak in this region before subtraction.

Discussion

The most prominent methods to detect RR phosphorylation require either a precisely positioned fluorescent residue or separation of products from reactants by SDS-PAGE and detection by radiography, western-blot or protein staining. Protein FTIR has been used previously to characterize the phosphorylation of other classes of proteins, but to the best of our knowledge, not RRs. The use of FTIR to detect phosphorylation would potentially extend the advantages of existing fluorescent methods to all RRs regardless of the placement of any fluorescent residues. RR_1586 was found to be phosphorylated by PA as previously confirmed by observing changes in oligomeric state and DNA-binding activity. In this study, the phosphorylation of RR_1586 by PA was observed using FTIR spectroscopic methods.

Comparison of phosphorylated and apo RR_1586 proteins revealed small shifts in secondary structure. While the absolute intensities of the amide I and amide II peaks indicate protein concentration, their widths, peak wavenumbers and relative heights can indicate secondary structure (37). In the apo protein spectrum, the amide I peak was more intense than the amide II peak. In the phosphorylated protein spectrum, these peaks were of equal intensity. These changes are interpreted as increase of α-helical character consistent with phosphorylation-driven reorganization of the unstructured REC-DBD linker into an α helix, as has been observed in another OmpR protein (63). This structural change may indicate the mechanism of phosphorylation-induced changes in the oligomeric state, but are not necessarily applicable to all RRs and not a universal indicator of RR phosphorylation.

Considering that phosphoaspartate contributed to a peak at 1131 cm^{-1} in another enzyme (62), similar peaks would be expected for phosphorylated RR_1586. However, any phosphorylation-related changes in this area of the spectrum were masked by sample-to-sample variations. From these observations, it is apparent that the instrument is too sensitive to allow comparison of apo and phosphorylated RR samples for routine detection of phosphorylation.

Rather than comparing the spectrum of pure protein to that of the purified phosphorylated product, the conversion of reactants to products was observed in real time. Repeated measurement of the same sample drastically reduces sample-to-sample noise, and even very small changes in the spectrum can be detected. Because phosphorylation of RRs by small molecules is slower than phosphorylation by a cognate HK protein, high donor concentrations or long incubations times are required. At these scales, the hydrolysis of phosphoramidate to phosphate and ammonium becomes relevant, although aqueous ammonium is not usually detected by FTIR at low millimolar concentrations (64). The decrease in absorbance at 1165 cm^{-1} and the increase at 1078 cm^{-1} are seen in all reactions involving PA, whether reacting with RR_1586 or hydrolyzing in the presence or absence of RR_1586^{D50G}, and are consistent with the cleavage of the higher energy P-N bond and the formation of a lower energy P-O bond, respectively. These changes appear to be linear over the observed time courses, consistent with pseudo-zero order kinetics during the early stages of the hydrolysis reaction.

Two other regions of the spectra also showed changes during the reaction. Sporadic changes appeared in the amide I band at ~1650 cm^{-1} without any changes in

the amide II region at ~1550 cm^{-1}. A decrease in absorbance at the amide I relative to the amide II region was anticipated based on the spectra of purified apo and phosphorylated RR_1586 (Figure 4.1A). Furthermore, the spectrum of PA also decreased in this region upon hydrolysis (Figure 4.3). It is possible that the sporadic nature of the shifts in the amide I band arise from competing effects on secondary structure due to active site phosphorylation and bulk changes in the solution driven by hydrolysis. The reaction with RR_1586^{D50G} also showed some shifts in the amide I band even though it is incapable of phosphorylation. Further investigation would be required to elucidate the source(s) of the shift seen in the amide I band during the phosphorylation reaction.

The second set of changes in absorbance was observed around 1450 cm^{-1}. The wild-type RR_1586 reaction with PA decreased in absorbance and tended toward a saturating effect. Saturation suggests that this is not due to protein-independent hydrolysis of PA, which is in excess, but due to a limiting reagent, such as the protein. However, it cannot be attributed directly to forming the phosphoaspartate bond, because a decrease in FTIR absorbance indicates a loss of vibrational modes by either bonds breaking or by stabilizing and/or hindering existing modes. Several amino acid side chains can contribute to this region including phenylalanine, tryptophan, proline, and acidic residues (37). The spectrum of the proline side chain is sensitive to the conformation of the main chain because of its unique structure (65). The contribution of acidic residues to absorbance in this region is possible when they chelate a metal cation (66). The two prolines, three acidic residues, and magnesium cation that are present in the active site of RR_1586—and highly conserved among all RRs—would not only

contribute to absorbance near 1450 cm^{-1}, but would also be subject to changes, or losses, of vibrational modes as the active site stabilizes the phosphoaspartate residue.

These observations have identified a shift in FTIR absorbance present during the phosphorylation of wild-type RR_1586, but absent in the negative control reactions. These changes could arise from amino acid side chains stabilizing the active conformation. On the other hand, upon phosphorylation wild-type RR_1586 transitions from a dimer to a tetrameric form. RR_1586^{D50G} is incapable of phosphorylation or tetramer formation, so the effects could also be attributed to side chain interactions at the new tetramer interface. The difference between these two interpretations is critical to the goals of this chapter, but could not be further explored due to an untimely instrument failure. A spectral signature indicative of universally conserved active site residues adjusting to RR phosphorylation is of much greater value to the field than a signature of RR_1586 oligomerization. Analysis of other RRs, especially the single domain RR CheY—which does not change oligomeric state upon phosphorylation—is likely to provide the final evidence to support or refute the role of a decrease in absorbance at 1450 cm^{-1} as a universal indicator of RR phosphorylation.

Methods

RR_1586 and RR_1586^{D50G} were purified by nickel affinity and size exclusion column chromatography as described in Chapter 2. All samples were prepared in 20 mM HEPES pH 7.5, 150 mM NaCl, with or without 10 mM $MgCl_2$ as indicated. FTIR measurements (30 scans) were taken using a Bruker Tensor II instrument mounted with an AquaSpec I sample cell and temperature controlled to 25.0 °C by a Ruber Ministat 125 water bath. Except for the timed series of reaction-induced difference spectra, all samples were collected with water as the background and buffer was subtracted during

analysis. The backgrounds for the time series were collected immediately after adding PA to the protein, and the first spectrum was immediately collected. Ten total measurements were made at three-minute intervals. Previous experience with preparing samples SEC-MALS indicated that this timeframe was sufficient to achieve 100% phosphorylation of RR_1586.

Phosphorylated RR_1586 was prepared by adding to the protein small volumes of PA and $MgCl_2$, both to final concentrations of 10 mM. SEC-MALS was used to confirm that RR_1586 had completely shifted to the phosphorylated, tetrameric form. PA was synthesized as described elsewhere (47).

Acknowledgements

Phosphoramidate used in this chapter was synthesized by Rob Fogle and Dr. Erwin Abucayon. I thank Dr. Erwin Abucayon for his insightful discussions concerning FTIR analysis and data interpretation.

Chapter 5: Evidence for the functional role of an evolutionarily conserved glycine in histidine-containing phosphotransfer proteins.

The HK and RR modules of two-component signaling pathways can also be organized into phosphorelay signaling pathways where a histidine-containing phosphotransfer (HPt) protein mediates transfer to a second RR in the pathway. The structure of HPt proteins resembles the four helix bundle corresponding to the dimerization and histidine phosphorylation domain of HKs (67), however HPt proteins are phosphorylated by upstream RRs rather than by ATP-dependent kinase activity. The addition of a third protein module increases the diversity of possible signaling domain organizations, which can be emphasized by bioinformatic analysis (68). Some bacterial phosphorelays are encoded as a single protein with an HK-RR-HPt-RR domain structure that undergoes internal His-Asp-His-Asp phosphotransfer. In yeasts, most HPt proteins are single-domain proteins and are the foci of converging upstream and/or diverging downstream branched pathways (69).

The Sln1 phosphorelay pathway in *Saccharomyces cerevisiae* has been studied as a model for two-component signal transduction in eukaryotes. Under non-stress conditions, the hybrid HK-RR sensory kinase Sln1 initiates phosphoryl transfer to the HPt Ypd1, which phosphorylates either Skn7 or Ssk1 RR proteins (70). Sln1 becomes inactive under hyperosmotic stress and unphosphorylated Skn7 and Ssk1 activate transcription of cell wall synthesis genes and the Hog1 mitogen-activated protein kinase cascade, respectively (71).

The phosphotransfer kinetics and binding between Ypd1 and each of the three RRs in this pathway have been characterized. Yeast two-hybrid assays were used to examine the effects that alanine substitutions at several positions on Ypd1 had on

binding to each of the three RRs (72). All three RRs were found to share a common docking site on Ypd1. In an analysis of the kinetics of phosphotransfer in this pathway, the equilibrium dissociation constants (K_d) of the Ypd1-RR interactions were all within the single digit μM range (73). Some of the Ypd1 mutants were tested in both the kinetic and yeast two-hybrid assays with good correlation between relative and quantitative results, except for one case. The $Ypd1^{G68Q}$ protein was found to drastically diminish binding in the yeast two-hybrid assay (72), but estimates from the kinetic data suggested only a slight decrease in binding affinity to the Sln1 RR domain (Sln1-R1) (73). This minor discrepancy was outside the scope of both of these publications and could be explained given the difference between *in vitro* and *in vivo* assays or the imprecisions associated with estimating binding from kinetics of a minimally active mutant.

This chapter describes the development and application of a fluorescence-based binding assay to quantify the effects that substitutions at position G68 of Ypd1 have on binding to Sln1-R1. This began as a project during my rotation in Dr. West's lab, and was included in a manuscript by Kennedy et al. that has been accepted for publication (28). The manuscript represents a collaborative body of work combining data gathered by several past and present researchers in Dr. West's lab.

Results

A fluorescence-based *in vitro* binding assay was developed to gain insight into binding between Ypd1 and Sln1-R1. A substitution was made on Ypd1 (T12) adjacent to the common RR docking site to introduce a unique, solvent exposed cysteine. This cysteine was used for thiol-specific labeling with a fluorescent probe, 5-iodoacetamidofluorescein (5-IAF). $Ypd1^{T12C}$ functions similarly to wild-type Ypd1 in *in*

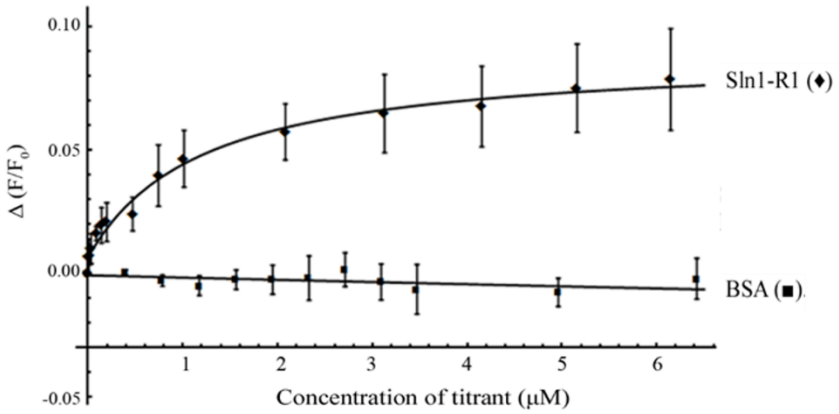

Figure 5.1 Ypd1-Sln1-R1 binding assay
Change in fluorescence signal indicative of binding between Ypd1 and Sln1-R1. Binding of Sln1-R1 to fluorescein-labeled Ypd1 increased the fluorescence signal with an apparent K_d of 0.9 ± 0.4 μM. Titration with bovine serum albumin had no effect on the fluorescence signal.

vitro phosphotransfer assays (data not shown). The titration of 5-IAF-labeled Ypd1 with Sln1-R1 increased fluorescence intensity above baseline buffer dilutions, resulting in binding curves that appear to reach saturation (Figure 5.1). Observed K_d values calculated from these binding curves for a panel of mutant Ypd1 proteins ranged from 0.5 to 3 μM as shown in Table 5.1. The observed K_d for the Sln1-R1 and Ypd1^{T12C}-F interaction, 0.9 ± 0.4 μM, is in agreement with the observed K_d between Sln1-R1 and wild-type Ypd1 calculated from published kinetics data (73, 74). Although the estimated K_d values show only minor changes in binding, substitutions with hydrophobic side chains showed a slightly increased K_d while hydrophilic side chains showed the opposite trend (Table 5.1).

Table 5.1 K_d values of Ypd1^{T12C}-F interacting with Sln1-R1

Ypd1^{T12C} variant	K_d (μM)[a]
WT	0.9 ± 0.4
G68V	2.9 ± 0.1
G68L	1.3 ± 0.1
G68E	0.5 ± 0.1
G68Q	0.6 ± 0.3

[a]The average and standard deviation of triplicates are reported

Discussion

The K_d values for wild-type Ypd1 and Sln1-R1 are in good agreement between the fluorescence- and kinetics-based assays (0.9 ± 0.4 and 1.4 ± 0.3 μM, respectively). The G68Q substitution did not have any measurable effects on binding (K_d = 0.6 ± 0.3 μM). These observations support the conclusions of the kinetics assays while improving the precision of the estimated parameters for G68Q. Attempts were made to also quantify binding using biolayer interferometry, but significant non-specific binding between the sensor and the analyte made this technique unusable for this study.

In the context of the manuscript by Kennedy et al. (28) that describes both structural and biochemical characterization of a set of G68X mutants, these fluorescence-based binding data provide evidence for the biophysical constraints imposing strict evolutionary conservation of small residues at this position of all HPt proteins. Glycine is generally known for its flexible nature, but crystal structures of free Ypd1 and Ypd1 in complex with Sln1-R1 show no relevant changes in the αC helix, which contains G68 (75, 76). Kennedy et al. found that substitutions at G68 with residues larger than serine severely disrupted phosphotransfer activity (28). It was hypothesized that the disruption in activity was due to bulky side chains hindering binding between Ypd1 and Sln1-R1. In light of the results presented in this chapter, it is apparent that binding was not affected by these substitutions. Analysis of the $Ypd1^{G68Q}$ crystal structure and its comparison to the published structures of the Ypd1-Sln1-R1 protein complex showed clashes between the G68Q residue and the catalytic lysine residue from Sln1-R1. These clashes would keep Sln1-R1 from stabilizing the intermediate and/or final phosphorylation states. Thus, the purpose of small residues at position 68 of Ypd1 and of HPt proteins in general seems to be to allow residues from

the cognate RR to access the phosphorylated histidine side chain and catalyze histidine to aspartate phosphotransfer. It appears that the void created by G68 is a conserved feature of HPt proteins as an important part of the cognate RR's function.

Methods

Preparation of Ypd1-fluorescein probes

$Ypd1^{T12C}$ and $Ypd1^{T12C\text{-}G68X}$ mutants were purified as described elsewhere (77) and buffer exchanged into 50 mM potassium phosphate, pH 9.0 and 1mM β-mercaptoethanol. Proteins were incubated in darkness for 2 hours at room temperature with a 7-fold molar excess of 5-IAF for covalent labeling of $Ypd1^{T12C}$. Unincorporated 5-IAF was removed by exchanging labeled Ypd1 proteins into 20 mM Tris, pH 8.0, 50 mM NaCl and 10 mM $MgCl_2$ using a GE HiTrap desalting column. The concentration of bound fluorescein was estimated by absorption at 492 nm, and protein concentration was estimated by BioRad protein assay revealing that 70-90% of the Ypd1 molecules were labeled. Fluorescently-labeled proteins were aliquoted and stored in the presence of 10% glycerol at -20 ˚C.

The Sln1-R1 receiver domain was purified as described previously, however, fluorescence reaction buffer (20 mM Tris, pH 8.0, 50 mM NaCl and 10 mM $MgCl_2$) was substituted during size exclusion chromatography. Chelex® resin (BioRad) was used to strip contaminating cations from the Tris-salt solution before magnesium chloride was added to the buffer.

Fluorescence intensity measurements

Binding of Sln1-R1 to fluorescein-labeled $Ypd1^{T12C}$ induces a change in the fluorescein moiety resulting in altered fluorescence intensity. A Fluoromax 4 Spectrofluorometer from Horiba Scientific, temperature controlled to 23.0 °C, was used

to observe changes in fluorescence caused by binding. IAF-labeled Ypd1^{T12C} (30 pmol) was diluted to 1.9 mL in fluorescence reaction buffer, and Sln1-R1 was titrated into the reaction such that the concentration in the cuvette ranged from 10 nM to 6 μM. Upon addition of Sln1-R1, the solution was mixed with a magnetic stir bar for 30 seconds and allowed to rest for an additional 20 seconds before reading fluorescence intensity with absorbance at 492 nm and emission at 515 nm. Intensity after each addition (F) as a fraction of intensity from labeled Ypd1 alone (F_0) was calculated for titration with Sln1-R1 and buffer alone. The difference between these two normalized intensities (F/F_0, Sln1-R1 – F/F_0, buffer) indicates binding of Sln1-R1 to labeled Ypd1. Plotting change in fluorescence intensity caused by Sln1-R1 versus concentration of Sln1-R1 shows a binding curve with saturation at high concentrations. These curves were fitted using Mathematica (78) to an expanded quadratic equation with three variable parameters accounting for 1) fluorescence from bound Ypd1, 2) fluorescence from unbound Ypd1, and 3) the dissociation constant for the Ypd1:Sln1-R1 complex. The average dissociation constant and standard deviation of the mean are reported here.

Acknowledgements

The work described in this chapter was funded by the National Science Foundation (MCB 1158319). I thank Dr. Christina Bourne for the use of her fluorimeter and Dr. Fabiola Janiak-Spens for the construction of Ypd1^{T12C}.

References

1. Zschiedrich CP, Keidel V, Szurmant H. 2016. Molecular mechanisms of two-component signal transduction. J Mol Biol 428:3752-75. http://doi.org/10.1016/j.jmb.2016.08.003
2. Bourret RB. 2010. Receiver domain structure and function in response regulator proteins. Curr Opin Microbiol 13:142-9. http://doi.org/10.1016/j.mib.2010.01.015
3. The PyMOL Molecular Graphics System, v1.3. Schrodinger, LLC,
4. Galperin MY. 2010. Diversity of structure and function of response regulator output domains. Curr Opin Microbiol 13:150-9. http://doi.org/10.1016/j.mib.2010.01.005
5. Bachhawat P, Swapna GV, Montelione GT, Stock AM. 2005. Mechanism of activation for transcription factor PhoB suggested by different modes of dimerization in the inactive and active states. Structure 13:1353-63. http://doi.org/10.1016/j.str.2005.06.006
6. Yoshida T, Qin L, Egger LA, Inouye M. 2006. Transcription regulation of ompF and ompC by a single transcription factor, OmpR. J Biol Chem 281:17114-23. https://doi.org/10.1074/jbc.M602112200
7. van Iersel MP, Kelder T, Pico AR, Hanspers K, Coort S, Conklin BR, Evelo C. 2008. Presenting and exploring biological pathways with PathVisio. BMC Bioinformatics 9:399. https://doi.org/10.1186/1471-2105-9-399
8. Ortet P, Whitworth DE, Santaella C, Achouak W, Barakat M. 2015. P2CS: Updates of the prokaryotic two-component systems database. Nucleic Acids Res 43:D536-41. http://doi.org/10.1093/nar/gku968
9. George RH, Symonds JM, Dimock F, Brown JD, Arabi Y, Shinagawa N, Keighley MRB, Alexanderwilliams J, Burdon DW. 1978. Identification of *Clostridium difficile* as a cause of pseudomembranous colitis. British Medical Journal 1:695-695. https://doi.org/10.1136/bmj.1.6114.695
10. Pettit LJ, Browne HP, Yu L, Smits WK, Fagan RP, Barquist L, Martin MJ, Goulding D, Duncan SH, Flint HJ, Dougan G, Choudhary JS, Lawley TD. 2014. Functional genomics reveals that *Clostridium difficile* Spo0A coordinates sporulation, virulence and metabolism. BMC Genomics 15:160. http://doi.org/10.1186/1471-2164-15-160
11. Freeman ZN, Dorus S, Waterfield NR. 2013. The KdpD/KdpE two-component system: Integrating K(+) homeostasis and virulence. PLoS Pathog 9:e1003201. https://doi.org/10.1371/journal.ppat.1003201
12. Garsin DA. 2010. Ethanolamine utilization in bacterial pathogens: Roles and regulation. Nat Rev Microbiol 8:290-5. http://doi.org/10.1038/nrmicro2334
13. Suarez JM, Edwards AN, McBride SM. 2013. The *Clostridium difficile* cpr locus is regulated by a noncontiguous two-component system in response to

type A and B lantibiotics. J Bacteriol 195:2621-31. http://doi.org/10.1128/JB.00166-13

14. Darkoh C, DuPont HL, Norris SJ, Kaplan HB. 2015. Toxin synthesis by *Clostridium difficile* is regulated through quorum signaling. MBio 6:e02569. http://doi.org/10.1128/mBio.02569-14

15. Carter GP, Lyras D, Allen DL, Mackin KE, Howarth PM, O'Connor JR, Rood JI. 2007. Binary toxin production in *Clostridium difficile* is regulated by CdtR, a LytTR family response regulator. J Bacteriol 189:7290-301. http://doi.org/10.1128/JB.00731-07

16. Cardona ST, Choy M, Hogan AM. 2018. Essential two-component systems regulating cell envelope functions: Opportunities for novel antibiotic therapies. J Membr Biol 251:75-89. https://doi.org/10.1007/s00232-017-9995-5

17. Collins J, Robinson C, Danhof H, Knetsch CW, van Leeuwen HC, Lawley TD, Auchtung JM, Britton RA. 2018. Dietary trehalose enhances virulence of epidemic Clostridium difficile. Nature 553:291-294. https:/doi.org/10.1038/nature25178

18. Mascher T, Helmann JD, Unden G. 2006. Stimulus perception in bacterial signal-transducing histidine kinases. Microbiol Mol Biol Rev 70:910-38. https://doi.org/10.1128/MMBR.00020-06

19. Hoch JA. 2000. Two-component and phosphorelay signal transduction. Current Opinion in Microbiology 3:165-170. https://doi.org/10.1016/S1369-5274(00)00070-9

20. van Rensburg JJ, Fortney KR, Chen L, Krieger AJ, Lima BP, Wolfe AJ, Katz BP, Zhang ZY, Spinola SM. 2015. Development and validation of a high-throughput cell-based screen to identify activators of a bacterial two-component signal transduction system. Antimicrob Agents Chemother 59:3789-99. https://doi.org/10.1128/AAC.00236-15

21. McKellar JL, Minnell JJ, Gerth ML. 2015. A high-throughput screen for ligand binding reveals the specificities of three amino acid chemoreceptors from *Pseudomonas syringae* pv. actinidiae. Mol Microbiol 96:694-707. https://doi.org/10.1111/mmi.12964

22. Ng WL, Wei Y, Perez LJ, Cong J, Long T, Koch M, Semmelhack MF, Wingreen NS, Bassler BL. 2010. Probing bacterial transmembrane histidine kinase receptor-ligand interactions with natural and synthetic molecules. Proc Natl Acad Sci U S A 107:5575-80. https://doi.org/10.1073/pnas.1001392107

23. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, Deutschbauer A. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio 6:e00306-15. https://doi.org/10.1128/mBio.00306-15

24. Rajeev L, Luning EG, Dehal PS, Price MN, Arkin AP, Mukhopadhyay A. 2011. Systematic mapping of two component response regulators to gene targets in a

model sulfate reducing bacterium. Genome Biol 12:R99. http://doi.org/10.1186/gb-2011-12-10-r99

25. Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, Wolfe SA. 2008. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. Nucleic Acids Res 36:2547-60. http://doi.org/10.1093/nar/gkn048

26. 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 133:1277-89. https://doi.org/10.1016/j.cell.2008.05.023

27. Hebdon SD, Menon SK, Richter-Addo GB, Karr EA, West AH. 2018. Regulatory targets of the response regulator RR_1586 from *Clostridioides difficile* identified using a bacterial one-hybrid screen. Journal of Bacteriology 200:e00351-18. https://doi.org/10.1128/jb.00351-18

28. Kennedy EN, Hebdon SD, Menon SM, Foster CA, Copeland DM, Xu Q, Janiak-Spens F, West AH. Role of the highly conserved G68 residue in the yeast phosphorelay protein Ypd1: Implications for interactions between histidine phosphotransfer (HPt) and response regulator proteins. BMC Biochemistry, revised manuscript submitted.

29. Dembek M, Barquist L, Boinett CJ, Cain AK, Mayho M, Lawley TD, Fairweather NF, Fagan RP. 2015. High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. MBio 6:e02383. http://doi.org/10.1128/mBio.02383-14

30. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J. 2015. RSAT 2015: Regulatory sequence analysis tools. Nucleic Acids Res 43:W50-6. http://doi.org/10.1093/nar/gkv362

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421

32. Mao F, Dam P, Chou J, Olman V, Xu Y. 2009. DOOR: A database for prokaryotic operons. Nucleic Acids Res 37:D459-63. https://doi.org/10.1093/nar/gkn757

33. Buske FA, Boden M, Bauer DC, Bailey TL. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. Bioinformatics 26:860-6. http://doi.org/10.1093/bioinformatics/btq049

34. Rogers JK, Guzman CD, Taylor ND, Raman S, Anderson K, Church GM. 2015. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. Nucleic Acids Res 43:7648-60. http://doi.org/10.1093/nar/gkv616

35. 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. http://doi.org/10.1016/S0969-2126(02)00761-X

36. He H, Zahrt TC. 2005. Identification and characterization of a regulatory sequence recognized by *Mycobacterium tuberculosis* persistence regulator MprA. J Bacteriol 187:202-12. 10.1128/JB.187.1.202-212.2005

37. Barth A. 2007. Infrared spectroscopy of proteins. Biochim Biophys Acta 1767:1073-101. http://doi.org/10.1016/j.bbabio.2007.06.004

38. Fried MG. 1989. Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. Electrophoresis 10:366-76. http://doi.org/epdf/10.1002/elps.1150100515

39. Dembek M, Stabler RA, Witney AA, Wren BW, Fairweather NF. 2013. Transcriptional analysis of temporal gene expression in germinating *Clostridium difficile* 630 endospores. PLOS One 8:e64011. http://doi.org/10.1371/journal.pone.0064011

40. Guo M, Feng H, Zhang J, Wang W, Wang Y, Li Y, Gao C, Chen H, Feng Y, He ZG. 2009. Dissecting transcription regulatory pathways through a new bacterial one-hybrid reporter system. Genome Res 19:1301-8. http://doi.org/10.1101/gr.086595.108

41. An H, Douillard FP, Wang G, Zhai Z, Yang J, Song S, Cui J, Ren F, Luo Y, Zhang B, Hao Y. 2014. Integrated transcriptomic and proteomic analysis of the bile stress response in a centenarian-originated probiotic *Bifidobacterium longum* BBMN68. Mol Cell Proteomics 13:2558-72. http://doi.org/10.1074/mcp.M114.039156

42. Zhai Z, Douillard FP, An H, Wang G, Guo X, Luo Y, Hao Y. 2014. Proteomic characterization of the acid tolerance response in *Lactobacillus delbrueckii subsp. bulgaricus* CAUH1 and functional identification of a novel acid stress-related transcriptional regulator Ldb0677. Environ Microbiol 16:1524-37. http://doi.org/10.1111/1462-2920.12280

43. Svensson SL, Huynh S, Parker CT, Gaynor EC. 2015. The *Campylobacter jejuni* CprRS two-component regulatory system regulates aspects of the cell envelope. Mol Microbiol 96:189-209. http://doi.org/10.1111/mmi.12927

44. Noyes MB. 2012. Anaylsis of specific protein-DNA interactions by bacterial one-hybrid assay. *In* Deplancke B, Gheldof N (ed), Gene Regulator Networks, vol 786.

45. Philippe VA, Mendez MB, Huang IH, Orsaria LM, Sarker MR, Grau RR. 2006. Inorganic phosphate induces spore morphogenesis and enterotoxin production in the intestinal pathogen *Clostridium perfringens*. Infect Immun 74:3651-6. http://doi.org/10.1128/IAI.02090-05

46. Volkman BF, Lipson D, Wemmer DE, Kern D. 2001. Two-state allosteric behavior in a single-domain signaling protein. Science 291:2429-33. http://doi.org/10.1126/science.291.5512.2429

47. Sheridan RC, McCullough JF, Wakefield ZT, Allcock HR, Walsh EJ. 2007. Phosphoramidic acid and its salts. doi:http://doi.org/10.1002/9780470132449.ch6:23-26. http://doi.org/10.1002/9780470132449.ch6

48. Meng X, Wolfe SA. 2006. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nat Protoc 1. http://dx.doi.org/10.1038/nprot.2006.6

49. Elsaesser R, Paysan J. 2004. Liquid gel amplification of complex plasmid libraries. Biotechniques 37:200, 202.

50. Scholz J, Besir H, Strasser C, Suppmann S. 2013. A new method to customize protein expression vectors for fast, efficient and background free parallel cloning. BMC Biotechnol 13:12. http://doi.org/10.1186/1472-6750-13-12

51. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: Tools for motif discovery and searching. Nucleic Acids Res 37:W202-8. http://doi.org/10.1093/nar/gkp335

52. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420-35. http://doi.org/10.1093/nar/gkn176

53. Bond SR, Naus CC. 2012. RF-Cloning.org: An online tool for the design of restriction-free cloning projects. Nucleic Acids Res 40:W209-13. https://doi.org/10.1093/nar/gks396

54. Pei J, Grishin NV. 2007. PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23:802-8. https://doi.org/10.1093/bioinformatics/btm017

55. Jolly L, Ferrari P, Blanot D, van Heijenoort J, Fassy F, Mengin-Lecreulx D. 1999. Reaction mechanism of phosphoglucosamine mutase from *Escherichia coli*. European Journal of Biochemistry 262:202-210. https://doi.org/10.1046/j.1432-1327.1999.00373.x

56. Lukat GS, McCleary WR, Stock AM, Stock JB. 1992. Phosphorylation of bacterial response regulator proteins by low molecular weight phospho-donors. Proc Natl Acad Sci U S A 89:718-22. http://doi.org/10.1073/pnas.89.2.718

57. Barbieri CM, Stock AM. 2008. Universally applicable methods for monitoring response regulator aspartate phosphorylation both in vitro and in vivo using Phos-tag-based reagents. Anal Biochem 376:73-82. https://doi.org/10.1016/j.ab.2008.02.004

58. Buckler DR, Stock AM. 2000. Synthesis of [^{32}P]phosphoramidate for use as a low molecular weight phosphodonor reagent. Anal Biochem 283:222-7. https://doi.org/10.1006/abio.2000.4639

59. Silversmith RE, Appleby JL, Bourret RB. 1997. Catalytic mechanism of phosphorylation and dephosphorylation of CheY: Kinetic characterization of

imidazole phosphates as phosphodonors and the role of acid catalysis. Biochemistry 36:14965-74. https://doi.org/10.1021/bi9715573

60. Bourret RB, Thomas SA, Page SC, Creager-Allen RL, Moore AM, Silversmith RE. 2010. Measurement of response regulator autodephosphorylation rates spanning six orders of magnitude. 471:89-114. https://doi.org/10.1016/S0076-6879(10)71006-5

61. Faupel MD, Ishii K, Meyrueis P, Yoshihashi SS, Chihara K, Awazu K. 2004. FT-IR analysis of phosphorylated protein. 5461:17. https://doi.org/s10.1117/12.541092

62. Barth A, Mäntele W. 1998. ATP-induced phosphorylation of the sarcoplasmic reticulum Ca^{2+} ATPase: Molecular interpretation of infrared difference spectra. Biophysical Journal 75:538-544. https://doi.org/10.1016/S0006-3495(98)77543-5

63. Ames SK, Frankema N, Kenney LJ. 1999. C-terminal DNA binding stimulates N-terminal phosphorylation of the outer membrane protein regulator OmpR from *Escherichia coli*. Proceedings of the National Academy of Sciences 96:11792-11797. https://doi.org/10.1073/pnas.96.21.11792

64. Max J-J, Chapados C. 2013. Aqueous ammonia and ammonium chloride hydrates: Principal infrared spectra. Journal of Molecular Structure 1046:124-135. https://doi.org/10.1016/j.molstruc.2013.04.045

65. Johnston N, Krimm S. 1971. An infrared study of unordered poly-L-proline in $CaCl_2$ solutions. Biopolymers 10:2597-605. https://doi.org/10.1002/bip.360101219

66. Tackett JE. 2016. FT-IR characterization of metal acetates in aqueous solution. Applied Spectroscopy 43:483-489. https://doi.org/10.1366/0003702894202931

67. Jacob-Dubuisson F, Mechaly A, Betton JM, Antoine R. 2018. Structural insights into the signalling mechanisms of two-component systems. Nat Rev Microbiol doi:https://doi.org/10.1038/s41579-018-0055-7. https://doi.org/10.1038/s41579-018-0055-7

68. Salvado B, Vilaprinyo E, Sorribas A, Alves R. 2015. A survey of HK, HPt, and RR domains and their organization in two-component systems and phosphorelay proteins of organisms with fully sequenced genomes. PeerJ 3:e1183. https://doi.org/10.7717/peerj.1183

69. Fassler JS, West AH. 2013. Histidine phosphotransfer proteins in fungal two-component signal transduction pathways. Eukaryot Cell 12:1052-60. https://dx.doi.org/10.1128%2FEC.00083-13

70. Saito H. 2003. The Sln1-Ypd1-Ssk1 multistep phosphorelay system that regulates an osmosensing MAP kinase cascade in yeast, p 397-419, Histidine Kinases in Signal Transduction doi:https://doi.org/10.1016/B978-012372484-7/50020-5. Academic Press.

71. Catlett NL, Yoder OC, Turgeon BG. 2003. Whole-genome analysis of two-component signal transduction genes in fungal pathogens. Eukaryotic Cell 2:1151-1161. https://doi.org/10.1128/EC.2.6.1151-1161.2003

72. Porter SW, West AH. 2005. A common docking site for response regulators on the yeast phosphorelay protein YPD1. Biochim Biophys Acta 1748:138-45. https://doi.org/10.1016/j.bbapap.2004.12.009

73. Janiak-Spens F, Cook PF, West AH. 2005. Kinetic analysis of YPD1-dependent phosphotransfer reactions in the yeast osmoregulatory phosphorelay system. Biochemistry 44:377-386.

74. Stojanovski K, Ferrar T, Benisty H, Uschner F, Delgado J, Jimenez J, Solé C, de Nadal E, Klipp E, Posas F, Serrano L, Kiel C. 2017. Interaction dynamics determine signaling and output pathway responses. Cell Rep 19:136-149. http://dx.doi.org/10.1016/j.celrep.2017.03.029

75. Xu Q, West AH. 1999. Conservation of structure and function among histidine-containing phosphotransfer (HPt) domains as revealed by the crystal structure of YPD1. J Mol Biol 292:1039-50. https://doi.org/10.1006/jmbi.1999.3143

76. Zhao X, Copeland DM, Soares AS, West AH. 2008. Crystal structure of a complex between the phosphorelay protein YPD1 and the response regulator domain of SLN1 bound to a phosphoryl analog. J Mol Biol 375:1141-51. https://doi.org/10.1016/j.jmb.2007.11.045

77. Xu Q, Nguyen V, West AH. 1999. Purification, crystallization and preliminary X-ray diffraction analysis of the yeast phosphorelay protein YPD1. Acta Crystallogr D Biol Crystallogr 55:291-3. https://doi.org/10.1107/S090744499800866X

78. Wolfram S. 2013. Wolfram research. Inc, Mathematica, Version 8:23.

Appendix A: Sequences derived from bacterial one-hybrid selections

Table A.1 Sequences selected by RR_1586 Ser131.

Insert Sequence	Stringency[a]
GAGGACAACAGCTTTGGGTACTTGAAAA	High
TACCTTGCTCCCGGGTTAAGCTTACGCA	High
GAAGATCCCAAAACTTAACTGCCACTAA	High
GCCACCCTGTTACCTTAACTGTCGCCAC	High
ACTCAAGTGAAACCTTAACTATACACCT	High
GTATTTTTGTGCGTTTGATTTTTTTTTG	High
cggccgcGACCTTAACTGATGGTTGTTAAAACACG[b]	High
TGGGTTCCGAAACCTTAACACGACCTAC	High
TTTTTTTGTGTGTTCTTGTTTTTTTTTT	High
TAAAAAAAAGTTTAAGTTAAGCTTTGAC	High
ATCTACGAAGATTAGTTAAGTTTAAACG	High
TAAAAAACAAAAGGAGTTAAGGAATGAC	High
TTTTTGTCTATGAGTTAAGGATTGAG	High
TCTGTTTGGTCGTTGGTTAAGGTTTTAC	Low
CCTACAAAAAAGTGAGTTAAGAAAAAAC	Low
AGGGTGGGTGTGGGTGTATTGTGGGTGG	Low
TTTGACGGTCAACAGTTAAGGATTTAAT	Low
AAAAAAATACAAGTAGACAAGAATGACG	Low
TGTTAAGGTAATGCTCCTCACTTGTGCC	Low
TCGGTCATTGTTGGGGTTGGGGGGGGGG	Low
CTCCTGGATGGCTGTTTGACTTTGCGTT	Low
TACACAATTAGGGCAATTAAGGAAAAAA	Low
GAGCTTTTTGGGGGGTCGGACTGTTTGC	Low
TATATAAAAATTGGAATTAAGGAATAAA	Low
AACCAGACGCCCGCGAACGACAAATAAA	Low
AATCAAAAAATAGGGGTTAAGGAATGAA	Low
TGAAACGATCGTCAATTGACTTCACGCC	Low
ATATGAGAGGAGAAGGGCAGAAATCAAT	Low
TAGTGTATTTGTTGATGTTTGTTTG	Low
AGCCATCTCGAAAACTTAACTGAAACAA	Low

[a] High stringency is defined as 20 mM 3AT, and low stringency is defined as 10 mM 3AT in the selection medium.

[b] Bolded text indicates sequence outside the randomized regions. See Appendix B for more details.

Table A.2 Sequences selected by RR_1586 Arg124

Insert Sequence	Count
ATCGGGTGTGTGTTGCGTAGGGGGGGT	2
GACCTTAACTGATGGTTGTTAAAACACG	1
AGCCATCTCGAAAACTTAACTGAAACAA	7
GAAGATCCCAAAACTTAACTGCCACTAA	1
AGGAAACTTAACTCCATAGACAATCCAC	1
CAAGAACATCAACCTTAACCGAACGGACG	1
AGAACTAACCAACCTTAACTCCTCCTGA	2
TAAAAGAGAAGGGAGTTAAGGAATGAAT	1

Table A.3 Sequences selected by RR_1677

Insert Sequence	Count
TCCAAATAACGAGATTGCAAAGCATAAC	1
GCCTCCTGCCCCCGAACCTAAGCACCCT	1
CTACCTCCCCCCCCCACCGCACCCTCAAC	1
AACAAAAAAAAAAAAGGGACAAAACAAAA	1
GAAACCGCTAAGAAGTTACAGGCGGAAA	1
CCGCCTCAGACCATTCCCCCCTATACCA	1
TATAAGTCAAAAAGTTACAAGAATGAAA	1
AACGACGTAAAAAAGTAACAAAAAAGAA	6
AGTGCAGAGATTTGGTAACAAAAGTGCA	2
ATTCTCCTCTCACCCCTTCTTTCCCAGT	1
GGTTTGACATTTTTATTGTTATTGTTTC	1
GTCTAAGAAGCAAAAGTTACAAAAGAAAC	2
AAGACGCAGAAGACTAGACAAAGAAACG	2
TAGGACCAAAAAGTAACAACGGATGAAG	1
GTACAGATCAAAGGTTACAAGCAGGATA	1
CTTCAAACAAATACTTGTCAGAGAAATA	2
AACCACGCATATNAGCATCCNNNNAGAA	1
AATCTTAGCATCGCCTTGACAACTGCAT	2

Appendix B: Data analysis for motif discovery from RR_1586 S131 bacterial one-hybrid selections

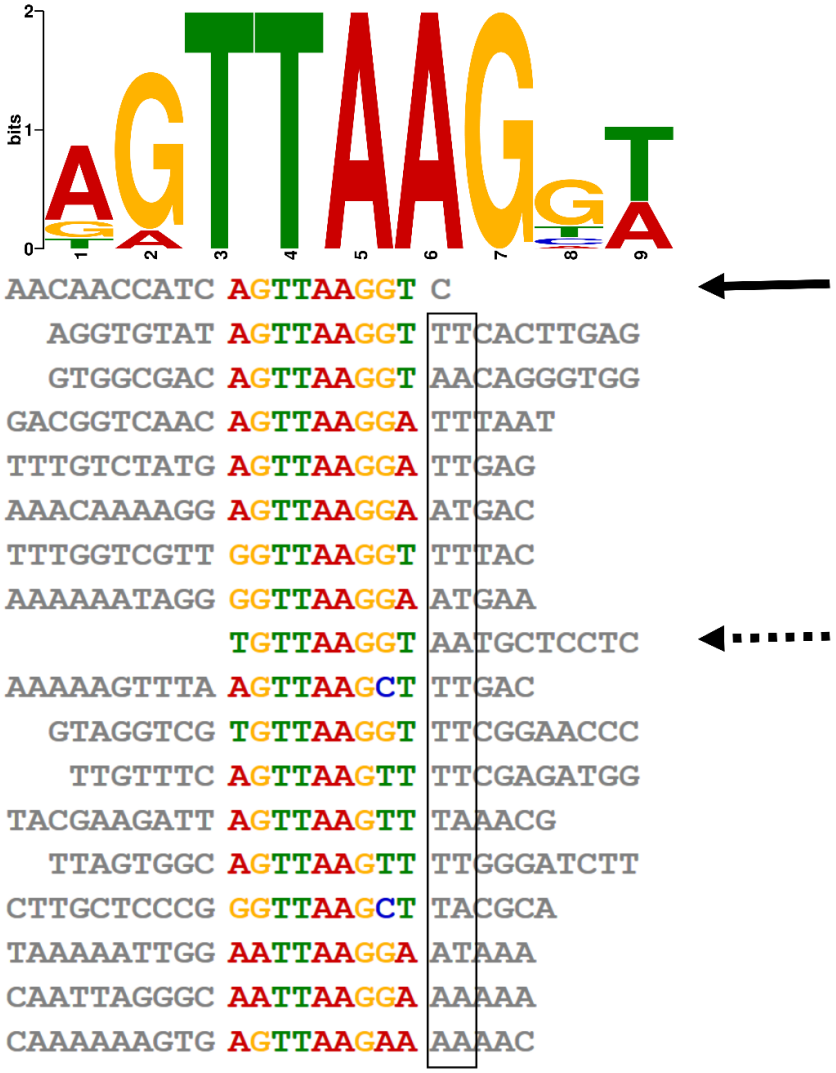

Figure B.1 Data analysis for motif discovery
A nine nucleotide motif was identified by MEME in the raw selected sequences. The conserved nucleotides contributing to the motif are colored, but two more nucleotide positions (boxed) have conserved A or T nucleotides. These nucleotides were excluded from the motif because of a penalty applied for reaching the end of the sequence indicated by a solid arrow. It is possible that RR_1586 bound a region straddling the border between the vector and randomized insert. We therefore included six nucleotides of the adjacent vector sequence in our final analysis to achieve the logo presented in Figure 2.2. Full sequences are shown in Table A.1. Adding vector nucleotides to the sequence marked with a dashed arrow did not significantly change the conserved motif and was therefore left unchanged.

Appendix C: Genome assemblies used for bioinformatics searches

Table C.1 Summary of orthologue searches for RR_1586 and RR_1677

Organism	RefSeq Accession	RR_1586[a]	RR_1677[a]
Acetoanaerobium noterae	GCF_900168025.1	-	+
Acetoanaerobium sticklandii	GCF_000196455.1	-	+
Asaccharospora irregularis	GCF_900129815.1	+	+
Clostridioides difficile QCD-66c26	GCF_000003215.1	+	+
Clostridioides mangenotii	GCF_000498755.1	+	+
Clostridioides difficile R20291	GCF_000027105.1	+	+
Criibacterium bergeronii	GCF_001693775.1	+	+
Filifactor alocis	GCF_000163895.2	+	+
Intestinibacter bartlettii	GCF_000154445.1	+	+
Paeniclostridium sordellii	GCF_000444095.1	+	+
Paraclostridium benzoelyticum	GCF_001006285.1	+	+
Paraclostridium bifermentans	GCF_000452225.2	+	+
Peptoanaerobacter stomatis	GCF_000238095.2	-	+
Peptoclostridium acidaminophilum	GCF_000597865.1	-	-
Peptoclostridium litorale	GCF_000699585.1	-	+
Peptostreptococcaceae bacterium AS15	GCF_000287695.1	-	+
Peptostreptococcaceae bacterium VA2	GCF_000686145.1	+	+
Peptostreptococcaceae bacterium oral taxon 113	GCF_000467935.1	+	+
Peptostreptococcus anaerobius	GCF_000178095.1	+	+
Peptostreptococcus stomatis	GCF_000147675.1	+	-
Proteocatella sphenisci	GCF_000423525.1	-	+
Romboutsia timonensis	GCF_900106845.1	+	+
Tepidibacter formicigenes	GCF_900142235.1	-	+
Tepidibacter thalassicus	GCF_900129915.1	-	+
Terrisporobacter glycolicus	GCF_000373865.1	+	+
Terrisporobacter othiniensis	GCF_000808015.1	+	+

[a] Columns indicate if a bidirectional best BLAST hit orthologue of the RR was (+) or was not (-) found in each of the listed genomes.

Appendix D: FTIR analysis of wild-type and mutant proteins

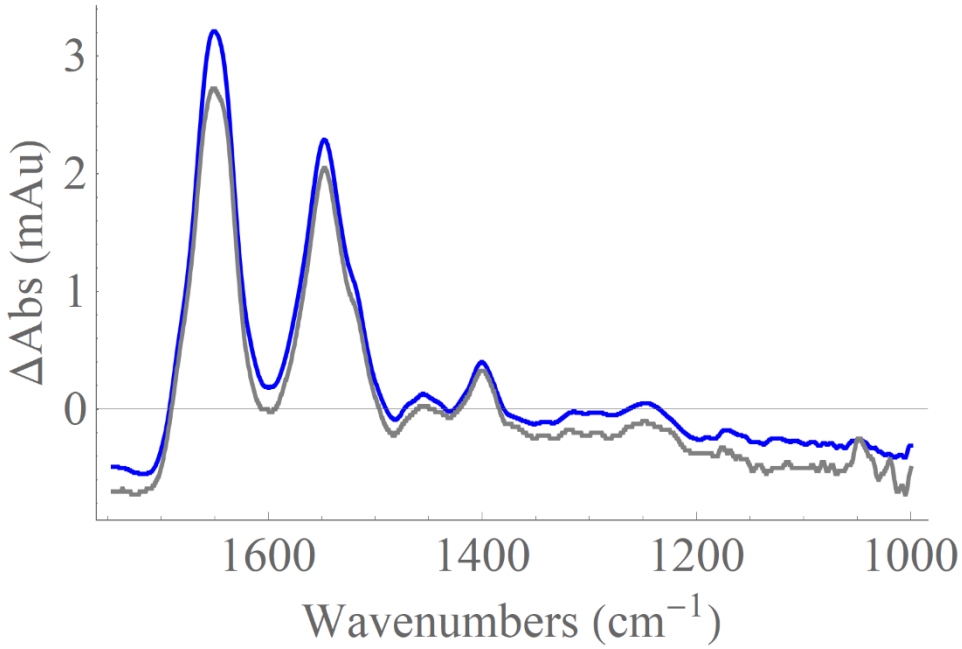

Figure D.1 FTIR spectra of RR_1586 and RR_1586D50G
The FTIR absorbance spectra and secondary structures of RR_1586 (blue) and RR_1586^{D50G} (gray) proteins are nearly identical. The RR_1586^{D50G} protein was at a lower concentration than RR_1586 and was stretched and shifted vertically to compensate for differences due to concentration. This better demonstrates the similarities in the peak positions and contours that indicate no significant changes in secondary structure (Table D.1).

Table D.1 Parameters calculated from FTIR the spectra in Figure D.1

Protein	α (%)	β (%)	Concentration (mg/mL)
RR_1586	41.7 ± 0.8	19.6 ± 0.6	1.04 ± 0.03
RR_1586^{D50G}	40.0 ± 1.0	19.6 ± 0.6	2.7 ± 0.2

Appendix E: Plasmids, strains, and primers used in this dissertation

Table E.1 Plasmids and strains

Strain or Plasmid	Use	Source (Addgene #)
pSGC-RR_1586	Protein Expression	Steve Almo
pSGC-RR_1586^{D50G}	Protein Expression	This work
USO hisB- pyrF- rpoZ- (*E. coli*)	B1H (selection)	Scot Wolfe (18049)
USO hisB- pyrF- (*E. coli*)	B1H (counter selection)	Scot Wolfe (12614)
pH3U3-mcs	B1H (prey)	Scot Wolfe (12609)
pB1H2w2-Zif268	B1H (positive control)	Scot Wolfe (18045)
pB1H2w2-Prd	B1H (cloning template)	Scot Wolfe (18038)
pB1H2w2-mutOdd	B1H (negative control)	Scot Wolfe (18044)
pB1H2w2-1586_FL	B1H (selection)	This work
pB1H2w2-1586_R124	B1H (selection)	This work
pB1H2w2-1586_S131	B1H (selection)	This work
pB1H2w2-1586_Q151	B1H (selection)	This work
pB1H2w2Δ-1586_R124	B1H (selection)	This work
pB1H2w2-1677_S144	B1H (selection)	This work
pB1H2w2Δ-1677_S144	B1H (selection)	This work
pB1H2w2Δ-1677_V141	B1H (selection)	This work
pJKR-L-tetR	GFP reporter	George Church (62562)
p0610-GFP	GFP reporter	This work
P3145-GFP	GFP reporter	This work

Table E.2 Oligonucleotides/Primers

ID	Sequence[a]	Experiment[b]
pr117_F	gttattttagmgttaatgcttcctaatataaatg	RR_1586^{D50G}
pr118_R	agcattaacmctaaaataactaaatcatat	RR_1586^{D50G}
pr21_F	ggcgcgaattcgnnnnnnnnnnnnnnnnnnnnnnnnnnnnngcggccgcaaggtagctgattccgttctcgc	B1H library synthesis
pr22_R	gcgagaacggaatcagctacctt	B1H library synthesis
pr23_R	gccgctgttggtaccatg	pB1H2w2 amplification
pr24_F	aagcttggagccacccgcggagcagctgcctcagtg	pB1H2w2 amplification
pr125_R	gccgctgttggtaccatgcttgtcgtcgtcatccttgtagtc	pB1H2w2Δ amplification
pr25_R	gcgggtggctccaagcttccgtatccacctttactg	B1H: RR_1586 insert
pr122_F	catggtaccaacagcggctctttatatttagatacactag	B1H: RR_1586 DBD (Ser131) insert
pr123_F	catggtaccaacagcggccaaaacaaacagtttaacttg	B1H: RR_1586 DBD (Gln151) insert
pr28_F	catggtaccaacagcggcagaaacggattagaatttaaat	B1H: RR_1586 DBD (Arg124) insert
pr201_F	catggtaccaacagcggcagtgaaatagtttctccacc	B1H: RR_1677 DBD (S144) insert

pr108_F	catggtaccaacagcggcgttaagtcgagtgaaatagttt	B1H: RR_1766 DBD (Val141) insert
pr31_F	caaatatgtatccgctcatgac	B1H colony PCR and sequencing primer
pr176_R	ccagagcatgtatcatatggtccagaaaccc	B1H colony PCR
pr156_F	attttacacgaaatgggcacgaagtatac	*CDR20291_0578*, double substituted
pr157_R	gtatacttcgtgcccatttcgtgtaaaat	*CDR20291_0578*, double substituted
pr154_F	attttattaagaatgggttaagagtatac	*CDR20291_0578*, native
pr155_R	gtatactcttaacccattcttaataaaat	*CDR20291_0578*, native
pr158_F	attttattaagaatgggcacgaagtatac	*CDR20291_0578*, single substituted
pr159_R	gtatacttcgtgcccattcttaataaaat	*CDR20291_0578*, single substituted
pr191_F	ttgcttgtttaagacatacttaaattttagg	*CDR20291_1833* inverted repeat
pr192_R	cctaaaatttaagtatgtcttaaacaagcaa	*CDR20291_1833* inverted repeat
pr188_F	atagttaaggtttaattaagattaaa	*CDR20291_3145*
pr189_R	tttaatcttaattaaaccttaactat	*CDR20291_3145*
pr268_F	ctatattaggattaagttaagcaagtgt	*CDR20291_3121*
pr269_R	acacttgcttaacttaatcctaatatag	*CDR20291_3121*
pr225_F	aggaattaaggagcaattaaatgatg	*CDR20291_1583*
pr226_R	catcatttaattgctccttaattcct	*CDR20291_1583*
pr291_F	tcggtaccgaaaaggaagagctagaaaaaagac	GFP: *CDR20291_3145*
pr292_R	agtaagcttcatctcataccactccctatc	GFP: *CDR20291_3145*
pr239_F	catggtaccaacagcggcatcattttcatttcaactaaatttg	GFP: *CDR20291_0610*
pr240_R	gcgggtggctccaagcttcatattcacacctcagg	GFP: *CDR20291_0610*
pr241_F	tcggtaccagtaagcttatgcgtaaaggtgaagaactg	Adapt pJKR-L-tetR for GFP reporter
pr242_R	actggtaccgaattcggtcatgcgtcc	Adapt pJKR-L-tetR for GFP reporter
pr243_F	gacggcacgtacaaaacccgtg	Adapt pJKR-L-tetR for GFP reporter
pr244_R	tgtacgtgccgtcgtctttgaaagag	Adapt pJKR-L-tetR for GFP reporter
pr245_F	gagatactgagcacatcagcagg	GFP reporter sequencing primer

[a] The conserved TTAAG, or substituted nucleotides, are underlined for oligos used in Figure 2.

[b] CDR20291 locus tags indicate nearest downstream gene from RR_1586 binding site.

Appendix F: B1H_analysis.pl script

This script is not intended to be read here, but is included as a source for others to copy and paste into an appropriate editor such as Atom, developed by GitHub. It was designed to run on the MacPro in Dr. West's lab and may require modifications to run on other computers.

```
#!/usr/bin/env perl

##############################################################################
#
# Title: B1H_analysis.pl
# Authors: Skyler Hebdon
#
##############################################################################
use strict;
use warnings;
use Getopt::Long;
use Term::ANSIColor;
use File::Basename;
use Pod::Usage;

=pod

=head1 Authors

Skyler Hebdon shebdon at gmail dot com

=head1 Description

Parses directories of SEQ files into a FASTA (one per directory) of
prey from the bacterial one-hybrid assay. Prey sequences are
identified using restriction sites and sequences flanking the N(28) or
N(10) libraries. FASTA files are automatically analyzed by MEME (MEME-
Suite). Options to run matrix-scan (RSAT) or GOMO (MEME-Suite) are
encoded. A command to run footprint-scan (RSAT) can be formatted and
printed but will not be run automatically because it can take over six
hours on an eight-core MacPro.

=head1 USAGE
```

B<You must define the RSAT environmental variables> before using this script.

source /applications/rsat/RSAT_config.bashrc

B1H_analysis.pl -m -g -f -tf locus_tag --eval number --pval number --lib [10 or 28] <directories>

=head2 ARGUMENTS

=over

=item <directories>

Enter or drag-and-drop one or more directories into the terminal. Directories must contain '.seq' files from colonies selected by the same transcription factor. Multiple directories can be handled in series. See below for Parallelization.

B<Compatability issue:> You must identify the locus tag of the bait transcription factor by either using for example '-tf CDR20291_1586' or by naming the folder as the locus tag. Only the second option is compatable with analyzing multiple folders.

=back

=head2 Optional Arguments

=over

=item -m -g -f

Run matrix-scan, GOMo, and/or footrpint-scan. Note: the footrpint_scan command is printed to be run by the user when ready to commit the computer to hours of processing.

=item -tf locus_tag

Identify the locus tag of the transcription factor if the folder name doesn't already do so.

Default behavior: -tf <directory>

=item --lib [28 or 10]

Set to 28 or 10 depending on cloning strategy for pH3U3 library.

Default behavior: --lib 28

=item --eval number

Sets upper threshold of statistical significance in MEME motif search. Accepts scientific notation.

Default behavior: --eval 0.05 or 5e-2.

=item --pval number

Sets upper threshold of statistical significance in matrix-scan search. Accepts scientific notation.

Default behavior: --pval 0.005 or 5e-3.

=item --diad (number)

Creates and processes a direct repeat of the discovered motif. Optional (number) describes period of repeat in nucleotides.

Default behavior: --diad 11

=item --org organism

Organism must match supported-organism in RSAT_config

Default behavior: --org Clostridium_difficile_R20291

=item --taxon taxon

Case sensitive, capitalize first letter. Passed to RSAT, must be compatible.

Default behavior: --taxon Peptostreptococcaceae

=item --bgfile path_to_bgfile

Drag and drop the background file for GOMo and footrpint_scan

Default behavior: --bgfile /Applications/Bioinformatics_Tools/CDR20291_1st_order_bg.txt

```
=item --help or -h

Stop everything and show this help file.

=back

=head1 Parallelization

The following example works if you first set working directory to a
parent directory containing multiple folders of SEQ files organized by
TF. Files must be named by the locus tag of the TF to run GOMO.

ls | parallel "perl B1H_analysis.pl -m --lib 28 --eval 1e-5 {}"

=cut

#-------------------------------
#configure preferences and paths to meme, rsat and directory tree made
by prep_go_table.pl
my $uid = $ENV{LOGNAME} || $ENV{USERNAME} || $ENV{USER};
my $memebin = "/Users/$uid/meme/bin";                 #path to meme
my $go_data = "/Applications/Bioinformatics_Tools";   #path to
prep_go_table.pl output
my $rsat = "/Applications/RSAT";                      #path to RSAT
my $opt_org = "Clostridium_difficile_R20291";         #organism to
study
my $opt_tax = "Peptostreptococcaceae";                #taxon to study
my $opt_bgfile =
"/Applications/Bioinformatics_Tools/CDR20291_1st_order_bg.txt";

#declare global variables, set defaults
my $opt_meme_eval = 0.05;          #default MEME cutoff E-value
my $opt_scan_pval = 0.005;         #defaults matrix-scan cutoff
my $opt_gomo = 0;                  #default skip gomo
my $opt_m_scan = 0;               #default skip matrix-scan
my $opt_f_scan = 0;       #default skip footprint-scan
my $opt_diad = 1;                  #1 defaults to not run
my $opt_library = 28;              #default to 28 bp lib
my $maxw = 28;                     #meme motif maxw
my $opt_tf = 1;                    #default uses folder name as tf
my $opt_help;

#----------------------- process options ----------------------------
GetOptions ('g' => \$opt_gomo, 'm' =>\$opt_m_scan, 'f' =>
\$opt_f_scan, 'lib=i' => \$opt_library, 'eval=s' => \$opt_meme_eval,
```

```
'pval=s' => \$opt_scan_pval, 'diad:i' => \$opt_diad, 'tf=s' =>
\$opt_tf , 'help!' => \$opt_help , 'h!' => \$opt_help, 'org=s' =>
\$opt_org, 'taxon=s' => \$opt_tax, 'bgfile=s' => \$opt_bgfile) or
pod2usage(-verbose => 1) && exit;
pod2usage(-verbose => 2) && exit if defined $opt_help;

if ((scalar @ARGV) < 1) {
  warn_print ("Give at least one directory.");
  exit;
}

unless ( -f $opt_bgfile ) { #stop if file is incorrect.
  warn_print ("Background file not found.");
  exit;
}

#if called but not set, set diad length to 11 nucleotides
if ($opt_diad == 0) {
   $opt_diad = 11;
   $maxw = 11;
 } elsif ($opt_diad == 1) {
   $maxw = 27;
 } else {
   $maxw = $opt_diad;
 }

#set bait flanking sequences for 28 or 10 bp (zf12-directed)
libraries.
my $before;
my $after;
if ($opt_library == 28) {
  $before = "GCGGCCGC";
  $after = "CGAATTC";
} elsif ($opt_library == 10) {
  $before = "ATGGATCC";
  $after = "TGGGCGGCT";
} else {
  warn_print ("Values of 10 or 28 expected for -lib option");
  exit;
}

#------------------------------ MAIN WORKFLOW -----------------------
foreach my $cwd (@ARGV) {
  my $motif_path = "$cwd/meme_out/monad"; #default to monad
  if ($opt_tf eq 1) {#assume dir is named after TF locus tag ID.
```

```
    $opt_tf = basename($cwd);
  }
  raw2meme ($cwd);
  if ("$opt_diad" >=2) {
    command ("convert-matrix -i",
    "$cwd/meme_out/monad.meme.txt -o $cwd/meme_out/motif.tab -from
meme",
    "-to tab");
    monad2diad ($cwd, $opt_diad);
    $motif_path = "$cwd/meme_out/diad";
  }

  if ($opt_gomo == 1) { run_gomo ($cwd, $motif_path) };
  if ($opt_m_scan == 1 ) {
    print "\nRunning matrix-scan on $opt_org.\n";
    command ("matrix-scan -v 2 -m $motif_path.transfac -matrix_format
transfac -bgfile $opt_bgfile ",
    "-2str -uth pval $opt_scan_pval -origin end -offset -50 ",
    "-i $go_data/sequences/$opt_org.leader_up_nt.fasta -o
$cwd/matrix_scan_out.txt");
  }
  if ( $opt_f_scan == 1 ) {
    warn_print ("\nCopy/paste the following text into the terminal to
run footrpint-scan in parallel");
    print "gene-info -descr -q \"\" -org $opt_org | cut -f1 -
d\$\'\\t\' | parallel --skip-first-line ",
    "\"footprint-scan -m $motif_path.transfac -matrix_format transfac
-bgfile $opt_bgfile ",
    "-tf $opt_tf -org $opt_org -sep_genes -infer_operons -taxon
$opt_tax -q \{\} -task ",

"orthologs,operons,ortho_seq,query_seq,purge,occ_sig,filter_scan,map,o
cc_sig_graph ",
    "-o $cwd\"";
  }
}
print "\n\nFinished\n";


#----------------------------- DEFINE SUBS -------------------------
#finds sequences, creates fasta, purges prey, finds motif in purged
set.
sub raw2meme {
  #capture options and declare variables
  my $folder = $_[0];
```

```
  my @files;        #array of SEQ files names in directory
  my $file;
  my $file_path;    #complete path to SEQ file

  #----------------------- PARSE Raw Sequences into FASTA -----------
  #collect files
  opendir (my $DIR, $folder) or die warn_print ("Can't open folder:
$folder");
  @files = grep {/\.seq/ } readdir($DIR);
  closedir $DIR;

  #exit if no files found in directory
  warn_print ("No .seq files found in $folder") && exit if (scalar
@files == 0);

  #read files, extract & clean prey.
  my @no_seq;
  my @fasta = ();   #array to hold prey
  foreach $file (@files) {
    $file_path = "$folder/$file";
    open my $raw_seq, '<' , $file_path or die warn_print ("Could not
open file $file_path\n");
    $/ = undef;          #slurping files
    if ( <$raw_seq> =~ /$before(.+?)$after/is ) {
    my $catch = $1;
    $catch =~ s/\s+//g;
      push (@fasta, ">$file", "$catch"); #list of caught prey
    } else {
      push (@no_seq, "$file_path");   #list files without prey
    }
    close $raw_seq;
  }

  #report files missing prey
  warn_print ("WARNING: Some files didn't have readable prey",
@no_seq) if (scalar @no_seq > 0);

  #save prey to file & purge duplicates
  if (scalar @fasta/2 >= 1) {
    open($file, '>' , "$folder/prey.fasta") or die warn_print ("Can't
write to $folder/prey.fasta");
    print $file "$_\n" for @fasta;
    close $file;
    command ("$memebin/purge ", "$folder/prey.fasta",
      " 100 -n -q -o >", "$folder/purged_prey.fasta");
```

```
  } else {
      warn_print ("Couldn't build fasta from $folder") && exit;
  }

  #-------------------- MOTIF SEARCH (MEME) -------------------------
  #add environmental variable and run meme, still slurping from ~30
lines above
  $ENV{PATH} = join ":", $ENV{PATH},"$memebin";
  command ("meme ", "$folder/purged_prey.fasta",
    " -dna -mod zoops -nmotifs 1 -minw 3 -maxw $maxw -revcomp -evt ",
"$opt_meme_eval", " -oc ",
    "$folder/meme_out/", " -nostatus");
  rename "$folder/meme_out/meme.txt",
"$folder/meme_out/monad.meme.txt";
  open ($file, '<' ,"$folder/meme_out/monad.meme.txt") or die
warn_print (
    "Can't open $folder/meme_out/monad.meme.txt");
  if ( not <$file> =~ /MEME-1/i) {
    warn_print ("No motif found in sequences.");
    close $file;
    exit;
  }
  close $file;
  print "\nMonad motif written to $folder/meme_out/monad.meme.txt\n";
  command ("convert-matrix -i",
  "$folder/meme_out/monad.meme.txt -o $folder/meme_out/monad.transfac
-from meme",
  "-to transfac");
  return;
}

#motif.tab file in $_[0] converted to direct repeat of length $_[1]
sub monad2diad {
  open my $filein, '<' , "$_[0]/meme_out/motif.tab" or
    die warn_print ("File not found: $_[0]/meme_out/motif.tab");
  $/ = "\n";            #stop slurping
  my @diad;
  while ( my $line = <$filein> ) {
    chomp $line;
    my @monad = split (/\t/, $line);
    until ($#monad >= $_[1]) {
      push @monad, 1;
    }
    push my @diad_line, [@monad[0..$_[1]], @monad[1..$_[1]]];
    push @diad, @diad_line;
```

```
  }
  close $filein;
  open my $fh, '>', "$_[0]/meme_out/diad.tab" or die $!;
  print $fh (join("\t", @$_), "\n") for @diad[0..3];
  print $fh '\\';
  close $fh;
  command ("fasta-get-markov $_[0]/purged_prey.fasta
>$_[0]/meme_out/zomm.txt ",
    "&& convert-matrix -i",
    "$_[0]/meme_out/diad.tab -o $_[0]/meme_out/diad.transfac -from tab
-to transfac ",
    "&& transfac2meme -bg $_[0]/meme_out/zomm.txt
$_[0]/meme_out/diad.transfac ",
    "> $_[0]/meme_out/diad.meme.txt");

  #open original motif file and extract e value
  open my $fhin, '<' , "$_[0]/meme_out/monad.meme.txt" or die $1;
  my $evalue;
  while ( my $line = <$fhin> ) {
    if ($line =~ /(^letter-probability.+)/) {     #find row starts
w/letter prob
    $evalue = substr($1, -12);                    #keep last 12 digits
    }
  }
  close $fhin;

  #read file, add e value change motif name and print
  open my $fhout, '<', "$_[0]/meme_out/diad.meme.txt" or die $1;
  my @lines = <$fhout>;
  close $fhout;
  my @new_lines;
  foreach (@lines) {
  $_ =~ s/E= 0/$evalue/g;
  $_ =~ s/diad\.tab/$opt_tf/g;
  push(@new_lines,$_);            #store modified lines
  }
  open my $fhout2, '>', "$_[0]/meme_out/diad.meme.txt" or die $1;
  print $fhout2 @new_lines;        #print midified lines with e value.
  close $fhout2;
  print "Diad motif written to $_[0]/meme_out/diad.meme.txt \n";
  return;
}

sub run_gomo {#given directory, motif file,
  $/ = "\n";            #stop slurping
```

```
  my $motif = $_[1];
  my $cwd = $_[0];
  my %tf_go_table;
  my @score_files;
  my @organisms;
  warn_print ("Getting genomes with orthologs of $opt_tf\n");
  system "footprint-scan -org $opt_org -m $motif.transfac -
matrix_format transfac -bgfile $opt_bgfile -tf $opt_tf -sep_genes -
taxon $opt_tax -q $opt_tf -task orthologs_tf -o $rsat/footprints";
  my $ortho_tf_file = join ("_",
$opt_tf,$opt_org,$opt_tax,"ortho_bbh_tf.tab" );
  open my $fhin, '<' ,
"$rsat/footprints/$opt_tax/$opt_org/$opt_tf/$opt_tf/$ortho_tf_file" or
die "Can't open orthologs file";
  while (my $line = <$fhin> ) {
    push @organisms, $line;
  }
  close $fhin;
  chomp ( @organisms );
  foreach my $org (@organisms) {
    open my $fhin, '<' ,
"$go_data/GO_tables/$org.mapping.go_table.txt" or die "\nRun
prep_go_table.pl for and restart\n";
    while ( my $line = <$fhin> ) {
      chomp $line;
      my @entry = split (/\t/, $line);
      my $go_term = shift @entry;
      foreach my $leader (@entry) {
        push( @{ $tf_go_table { $go_term } }, $leader );
        }
    }
    close $fhin;
  }
  print "Making GO tables \n";
  open my $fhout, '>' , "$cwd/$opt_tf.go_table.txt";
  foreach my $key (keys %tf_go_table) {
    my @out = @{ $tf_go_table { $key } };
    print $fhout join("\t", $key, @out, "\n");
  }
  close $fhout;

  mkdir "$cwd/ama_dump" unless (-f "$cwd/ama_dump" );

  print "Running MEME's ama function\n";
  foreach my $org (@organisms) {
```

```
    if (-f "$go_data/sequences/$org.leader_up_nt.fasta") {
      command ("ama -verbosity 1 -pvalues -oc $cwd/ama_dump/$org
$motif.meme.txt ",
      "$go_data/sequences/$org.leader_up_nt.fasta $opt_bgfile");
      push @score_files, "$cwd/ama_dump/$org/ama.xml";
    } else {
      warn_print ("Can't find nt.fasta for GOMo. Need to run
prep_go_table.pl");
    }
  }
  my $go_data = join ( ' ', "@score_files");
  warn_print ("Running MEME's gomo function");
  command ("gomo --verbosity 1 --oc $cwd/gomo_out --shuffle_scores
2000 --motifs $motif.meme.txt $cwd/$opt_tf.go_table.txt $go_data ");
}

#print array items in new lines, first line bold.
sub warn_print {
  print color 'bold';
  print "\n$_[0]\n";
  print color 'reset';
  print "$_\n" for @_[1..$#_];
  print "\n";
  return;
}

#Variables evaluated before sending to bash.
sub command {
  system "@_";
  return;
}
```

Appendix G: Prep_go_table.pl script

As with B1H_analysis.pl in Appendix F, this script is not intended to be read here, but is included as a source for others to copy and paste into an appropriate editor such as Atom, developed by GitHub. It was designed to run on the MacPro in Dr. West's lab and may require modifications to run on other computers.

```
#!/usr/bin/env perl
##############################################################################
#
# Title: prep_go_table.pl
# Authors: Skyler Hebdon
#
##############################################################################
use strict;
use warnings;
use List::MoreUtils qw(uniq firstidx);
use File::Copy qw (copy);
use lib '/Applications/rsat/perl-scripts/lib/';
use Getopt::Long;
use Pod::Usage;
=POD

=head1 Authors

Skyler Hebdon shebdon at gmail dot com

=head1 Description

This script can be used to maintain the information necessary to run
GOMo and matrix-scan as part of the B1H_analysis.pl pipeline. It is
not required for motif identification or for footprint-scan analysis.

=head1 Usage

B<You must define the RSAT environmental variables> before using this
script.

source /applications/rsat/RSAT_config.bashrc

prep_go_table.pl
```

```
The user will be prompted to identify BLAST2GO export tables for each
of the species in the given taxon (default: Peptostreptococcaceae).

=head2 Optional ARGUMENTS

=over

=item -taxon taxon

Must use RSAT-supported taxon spelling and capitalization.

=item -h or -help

Print this help page.

=cut
my $path_base = "/Applications/Bioinformatics_Tools";
my $from = -450;
my $to = 50;
my %operons;
my $opt_help;
my $opt_tax = "Peptostreptococcaceae";
GetOptions ( 'help!' => \$opt_help , 'h!' => \$opt_help, 'taxon=s' =>
\$opt_tax)
 or pod2usage(-verbose => 1) && exit;
 pod2usage(-verbose => 2) && exit if defined $opt_help;
my @organisms = `supported-organisms -taxon $opt_tax`;
chomp ( @organisms );
#take care of database
foreach my $org (@organisms) {#unless you can Find the files, make
them
  unless (-f "$path_base/GO_tables/$org.mapping.go_table.txt")
{#find/make GO Tables
    unless (-f "$path_base/GO_tables/$org.operons.tab") {#find/make
operons file
      my $params = "-org $org -all -return operon,q_info -o
$path_base/GO_tables/$org.operons.tab";
      system "infer-operons $params";
      PredictOperons ( "$path_base/GO_tables/$org.operons.tab")
    }
    unless (-f "$path_base/GO_tables/$org.mapping.txt") {#find or
identify/copy BLAST2GO file
      print "Can't find mapping file from BLAST2GO.\n";
      print "Leave blank to skip.\nDrag and drop a file to have it
added for $org.\n";
```

```
      my $mapping_file = <STDIN>;
      chomp $mapping_file;
      $mapping_file =~ s/^\s+|\s+$//g;  #remove white spaces
      if (-f $mapping_file) {#if file exists, move it to the right
place/name
        copy("$mapping_file",
"$path_base/GO_tables/$org.mapping.txt");
      } else {
        print "No file found.\n";
        next;
      }
    }
    MakeGoTable ( "$path_base/GO_tables/$org.mapping.txt",
"$path_base/GO_tables/$org.operons.tab");
  }
  unless (-f "$path_base/sequences/$org.leader_up_nt.fasta")
{#find/make FASTA files
    PredictOperons ( "$path_base/GO_tables/$org.operons.tab" );
    my $out_params = "-type upstream -from $from -to $to -label ID -
format FASTA -ids_only";
    my @leaders = values %operons;
    my @query = join (" -q ", @leaders);
    system "retrieve-seq -org $org $out_params -q @query -noorf -o
$path_base/sequences/$org.leader_up_nt.fasta";
  }
}
sub PredictOperons {#input path to file to be changed
  my $outfile = "$_[0]";
  $outfile =~ s{\.[^.]+$}{};   # to remove file extention
  {#change operon naming scheme to use only ID's not names
    my %renamed_operons;
    {#fix single-gene operons in line. accumulate changes to multigene
operons
      open my $fhout, '>' , "$outfile.temp" or die $!;
      open my $fhin, '<' , "$outfile.tab" or die $!;
        while ( my $line = <$fhin> ) {
          my @entry = split (/\t/, $line);
          my $bad_operon = $entry[0];
          my $operon_length = scalar ( split (/;/, $entry[0]) );
          if ( $operon_length > 1) {
            if ( exists $renamed_operons{"$entry[0]"} ) {#update or
write new operon.
              my $fixed = $renamed_operons{"$entry[0]"};
              $fixed =~ s/$entry[3]/$entry[2]/g;
              $renamed_operons{"$entry[0]"} = $fixed;
```

```
          } else {
            $bad_operon =~ s/$entry[3]/$entry[2]/g;
            $renamed_operons{"$entry[0]"} = $bad_operon;
          }
        } else {
          $line =~ s/$entry[3]/$entry[2]/g;
        }
        print $fhout $line;
      }
    close $fhin;
    close $fhout;
  }
  {#apply changes to multigene operons
    my $data;
    open my $fhin, '<' , "$outfile.temp" or die $!;
    $/ = undef;                  # slurp file
    $data = <$fhin>;
    while (my($k, $v) = each %renamed_operons) {
      $data =~ s/$k/$v/g;
    }
    close $fhin;
    $/ = "\n";             #stop slurping
    open my $fhout, '>' , "$outfile.tab" or die $!;
      print $fhout $data;
    close $fhout;
    unlink "$outfile.temp";
  }
}
{#interpret operon leaders, save as hash: %operons{gene}=leader.
  # ^because infer-operons sometimes returns <NULL> leader.
  %operons = ();
  open my $fhin, '<' , "$outfile.tab" or die $!; #open operons file
  while (my $line = <$fhin> ) {
    my @entry = split (/\t/, $line);
    my $gene = $entry[1];
    my @operon = split (/;/, $entry[0]);
    my $leader ="";
    if (scalar @operon == 1) {
      $leader = $operon[0];
    } elsif ( $entry[4] eq "R") {
      $leader = $operon[0];
    } elsif ( $entry[4] eq "D") {
      $leader = $operon[-1];
    } else  {
```

```
        print "error: Skyler didn't know what he was doing when he
wrote this";
      }
      $gene =~ s/^\s+|\s+$//g;    #remove white spaces
      $leader =~ s/^\s+|\s+$//g;  #remove white spaces
      $operons{$gene} = "$leader";
      delete $operons{"query"};
    }
    close $fhin;
  }
}
sub MakeGoTable {#input mapping file,operon.tab file just in case its
needed
  my %go_table;
  {#push leaders from %operons onto hash of arrays
%go_table{term}=@leaders
    open my $file, '<' , "$_[0]" or die $!; #open $mapping_file
    my @firstLine = split (/\t/, <$file> );
    my $names_col = firstidx{ $_ eq 'SeqName' } @firstLine;
    my $gos_col = firstidx{ $_ eq 'GO IDs' } @firstLine;
    close $file;
    open my $fhin, '<' , "$_[0]" or die $!; #open $mapping_file
    while ( my $line = <$fhin> ) {
      my @entry = split (/\t/, $line);
      my $gene = $entry[$names_col];
      $gene =~ s/^\s+|\s+$//g;  #remove white spaces
      my @goterms = split (/;/, $entry[$gos_col] );
      foreach my $term (@goterms) {
        $term =~ s/^\s+|\s+$//g;   #remove white spaces
        push( @{ $go_table { $term } }, ${operons{$gene}} );
      }
    }
    close $fhin;
  }
  {#write file from hash of arrays with key, elements (tab delimited).
    my $outfile = "$_[0]";
    $outfile =~ s{\.[^.]+$}{};   # to remove file extention
    open my $fhout, '>' , "$outfile.go_table.txt" or die $!;
    foreach my $key (keys %go_table) {
      my @out = uniq @{${go_table{$key}}};
      print $fhout join("\t", $key, @out, "\n");
    }
    close $fhout;
  }
}
```