Specificity and evolution of bacterial two-component signal transduction systems

by

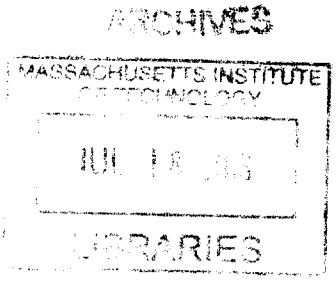
Emily Jordan Capra

A.B. Molecular Biology
Princeton University, Princeton, New Jersey, 2008

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2013



Signature of Author: _____

Emily Jordan Capra
Department of Biology
April 29, 2013

Certified by: _____

Michael T. Laub
Associate Professor of Biology
Thesis supervisor

Accepted by: _____

Amy E. Keating
Associate Professor of Biology

Specificity and evolution of bacterial two-component signal transduction systems

by

Emily J. Capra

Submitted to the Department of Biology
on April 29, 2013 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biology at the Massachusetts Institute of Technology

ABSTRACT

Cells possess a remarkable capacity to sense and process a diverse range of signals. Duplication and divergence of a relatively small number of gene families has provided the raw material enabling cells to quickly increase their signaling capacity. After duplication, however, all pathway components are identical in sequence and function. To evolve a new role, the pathways must become insulated at the level of signal transduction. Two-component signal transduction systems, consisting of a sensor histidine kinase and a cognate response regulator, are the main means by which bacteria sense and respond to their environment. These systems have undergone extensive duplication and lateral gene transfer such that most species encode dozens to hundreds of these pathways, yet there is little evidence of cross-talk at the level of signal transduction. Previous work has shown that interaction specificity is dictated by molecular recognition and determined by a small set of specificity residues.

I begin by studying the evolutionary trajectories of specificity residues in a duplicated two-component system that lead to insulation of pathways while at the same time maintaining interaction between cognate kinases and regulators. I then examine specificity residues in orthologs of a single two-component system and show that specificity residues are typically under purifying selection, but, as a result of additions to the two-component signaling network, can undergo bursts of diversification followed by extended stasis. By reversing these mutations I demonstrate that avoidance of cross-talk is a major selective pressure. Finally, I show that covalent attachment of the response regulator to a kinase represents an alternative mechanism for enforcing specificity. In these cases, no changes are needed to accommodate a duplication; the high effective concentration of the covalently attached response regulator prevents cross-talk with other two component proteins in the cell. This may allow hybrid kinases to be duplicated or transferred between genomes more easily. This work sheds light on the apparent ease with which two-component systems have expanded to become the dominant signaling system in bacterial genomes and, more generally, how a small number of gene families can be responsible for signal transduction in all organisms.

Thesis Supervisor: Michael T. Laub
Title: Associate Professor of Biology

ACKNOWLEDGEMENTS

This work would not have been possible without the help and support of a large number of people. I'd like to thank the following people in particular:

I'd first like to thank my advisor Mike Laub for all of his scientific guidance. It's been a great run and a great graduate experience. You've really made the lab a fantastic place to do science.

My thesis committee, Amy Keating and Aviv Regev for all of their insightful comments on the project, and their willingness and timeliness in writing me so many letters of recommendation.

Mike Springer for agreeing to be the outside member on my committee.

The lab as a whole—the amount of knowledge and scientific interest in the lab is impressive. I'd like to thank everyone for their willingness to help and for making my graduate experience what it was. Within the lab there are certain people who I'd like to single out. I need to especially thank Barrett. The immense number of profiles wouldn't have been the same without the company in the hot room. Your knowledge and instincts were invaluable in helping me to decide what path to take. All of my papers would have taken a lot longer if not for the numerous protein purifications that you were willing to help with. You have been my sounding board scientifically a great friend. I couldn't imagine the Laub lab without you. I'd also like to thank Erin, Kasia, and Christos for welcoming me into the lab and for paving the way. Christos and Diane for being awesome baymates and ensuring that I spent very little time in lab alone. Anna for being a great addition to the specificity side of the lab and always being willing to bounce ideas around with me and to distract me with coffee. And finally, the pilates group for convincing me to leave lab at a reasonable hour once a week for cannolis and exercise.

My classmates for going on this journey with me and for all of the random Boston adventures. I'm so thankful to have met you. I'd especially like to thank Lori for the coffee chats and for being my fifth floor buddy. Also the trivia group-Jen, MK, Josh, Jason, Jenny, and Lori for trying make sure that we see each other outside of lab.

My roommates, Ashley, Sarah, and Julia, and my honorary roommate Max. I can't believe it's been five years. I don't know what I would have done without you guys. From family dinners to scientific conversations to random adventures around Boston, I'll miss you guys and I'm so happy that we decided to embark on this journey together.

Finally my family, for their love and support in all that I do.

TABLE OF CONTENTS

TABLE OF FIGURES AND TABLES

Chapter 1: Introduction

Chapter 2: Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that insures specificity of two-component signaling pathways

Chapter 3: Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families

Chapter 4: Spatial tethering of kinases to their substrates relaxes evolutionary constraints on specificity

Chapter 5: Conclusions and future directions

Chapter 1

Introduction:
Evolution of two-component signal transduction systems

This chapter is adapted from work originally published as Emily J. Capra and Michael T. Laub. 2012. Annu Rev Microbiol. 66:325-47.

EJC and MTL wrote the manuscript and designed the figures. EJC made all of the changes from the original manuscript.

Overview

Two-component signal transduction systems are a predominant means by which bacteria sense and respond to their environments. These systems are generally comprised of a receptor histidine kinase that senses a specific signal and translates that input into a desired output through the phosphorylation of its cognate response regulator. The success of two-component signaling systems as a strategy for coupling changes in the environment to changes in cellular physiology is underscored by their prevalence throughout the bacterial kingdom. These signaling proteins have been found in the genomes of nearly all sequenced bacteria, with the majority of species encoding dozens, and sometimes hundreds, of two-component proteins. They have been uncovered in countless genetic screens and shown to respond to an enormous range of signals and stressors (for reviews, see (Laub, 2011; Stock et al., 2000).

Although tremendous progress has been made in understanding the structure and function of some individual systems, additional aspects of these pathways have recently garnered significant interest. How does a single cell coordinate so many highly related signaling pathways? The kinases and regulators encoded by a given organism are often highly similar at the sequence and structural levels, yet cells are able to match specific inputs to the desired output. How is unwanted cross-talk avoided? Do cells leverage the similarity of these proteins to integrate signals or diversify responses?

Histidine kinases and response regulators have an intrinsic modularity that separates signal input, phosphotransfer, and output response; this modularity has allowed bacteria to dramatically expand and diversify their signaling capabilities. Gene duplication and lateral (horizontal) gene transfer (LGT) provide the raw materials for producing new

pathways and, in either case, the introduction of new signaling proteins requires a flurry of changes if the new proteins are to be maintained over the course of evolution. The new pathway must gain a new function to provide a selective advantage and to warrant maintenance in the genome. Domain shuffling likely plays a critical role and recent work has begun to reveal how, at a mechanistic level, this process occurs. New pathways must also avoid cross-talk with other pathways, and *vice versa*, leading to changes in the specificity determinants of these pathways at multiple levels, including receptor dimerization and kinase-substrate partnering. Recent work has begun to reveal the molecular basis by which two-component proteins evolve. How and why do orthologous signaling proteins diverge? How do cells gain new pathways and recognize new signals? What changes are needed to insulate a new pathway from existing pathways? What constraints are there on gene duplication and lateral gene transfer?

The two-component signal transduction paradigm

The eponymous two-component signaling pathway contains a sensor histidine kinase and a cognate response regulator (Figure 1.1A). Upon receipt of a stimulus, the histidine kinase catalyzes an autophosphorylation reaction on a conserved histidine residue. This phosphoryl group is then transferred to a conserved aspartate on a cognate response regulator. Phosphorylation of the regulator usually drives a conformational change that activates its output response, often leading to changes in gene expression (Gao et al., 2007; Gao and Stock, 2009; Gao and Stock, 2010; Stock et al., 2000). These systems thus represent versatile, powerful ways to couple changes in external or environmental conditions to corresponding changes in cellular physiology and gene expression. In most cases, histidine kinases are bifunctional such that, when not stimulated to

autophosphorylate, they act as phosphatases for their cognate response regulators; thus it is ultimately the ratio of kinase to phosphatase activity that is responsible for modulating the output response (Huynh and Stewart, 2011; Jin and Inouye, 1993; Yang and Inouye, 1993). In some cases, input signals may promote the phosphatase state rather than stimulating autophosphorylation (Raivio and Silhavy, 1997).

All histidine kinases contain two highly conserved domains, the dimerization and histidine phosphotransfer (DHp) domain, which harbors the conserved histidine that is the site of both the autophosphorylation and phosphotransfer reactions, and the catalytic and ATP-binding (CA) domain. Histidine kinases also usually contain at least one (and often several) additional domain N-terminal to the DHp domain (Figure 1.1B). For the vast majority of kinases this includes 1-13 transmembrane domains (Galperin, 2005) with signal recognition occurring primarily in the periplasmic or extracellular portion of the protein. Although some common domains have been noted, signal recognition domains tend to be more variable than the other domains. Most kinases also have at least one domain between the transmembrane and DHp domains, with PAS, HAMP, and GAF domains by far the most common (Galperin et al., 2001). These domains can either relay signals from the periplasmic sensory domains to the DHp and CA domains or, in some cases, directly recognize cytoplasmic signals (Moglich et al., 2009b; Parkinson, 2010).

Response regulators share a common, well-conserved receiver domain (RD) that catalyzes phosphotransfer from its cognate histidine kinase. Phosphorylation then promotes a conformational change on one face of the receiver domain, which in turn effects an output (Gao et al., 2007). In single-domain response regulators, the

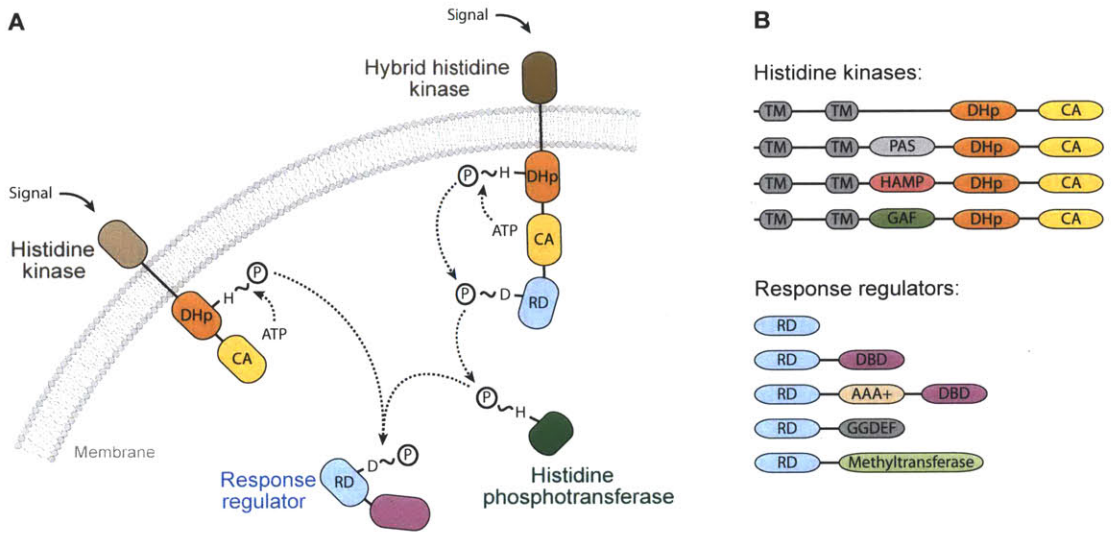


Figure 1.1 Overview of two-component signal transduction.

(A) In the canonical two component pathway (*left*), the CA domain of a histidine kianse binds ATP and autophosphorylates a conserved histidine in the DHp domain. The phosphoryl group is then transferred to an aspartate in the RD of the cognate response regulator, activating its output domain to effect cellular changes, often through changes in transcription. In a phosphorelay system (*right*), a hybrid histidine kinase autophosphorylates and transfers its phosphoryl group intramolecularly to a RD. A histidine phosphotransferase (HPT) then shuttles the phosphoryl group to a soluble response regulator that effects a pathway output. (B) Common domain organizations of histidine kinases and response regulators are shown. For histidine kinases, the DHp and CAdomains are shown with comnon intracellular domains: Per-Arnt-Sim (PAS), histidine kinase and methyl-accepting proteins (HAMP), and cGMP-specific phosphodiesterase adenyl cyclase and FhlA (GAF). Note that some kinases have multiple copies of such domains. Two TM domains are shown on the kinases, but kinases can harbor from 0-13 TM domains. A wide range of sensory domains (not shown) are often found in the periplasmic portions of membrane-bound histidine kinases. For response regulators, the conserved receiver domain is shown alone or with common output domains including a DNA-binding domain (DBD), a AAA+ and DNA-binding domain, a GGDEF domain involved in cyclic-di-GMP synthesis, or a CheB-like methyltransferase domain.

conformational change in the receiver domain allows the protein to directly produce an output response. Most response regulators, however, contain a DNA binding output domain (Galperin, 2006) (Figure 1.1B). For these regulators, phosphorylation induces homodimerization of the receiver domain, stimulating DNA binding and leading to

transcriptional changes. Other common output domains include diguanylate cyclases and methyltransferases.

A common variant of the two-component paradigm is the so-called phosphorelay (Burbulys et al., 1991) (Figure 1.1A). These extended pathways typically initiate with a hybrid kinase, which is a histidine kinase with a receiver domain fused to its C-terminus. After autophosphorylation and an intramolecular phosphotransfer to the receiver domain, the phosphoryl group is shuttled to a histidine phosphotransferase (HPT), and from there to a terminal response regulator that effects an output. Nearly 25% of all histidine kinases are hybrids (Cock and Whitworth, 2007), suggesting that phosphorelays are common.

Evolution of genome content and gene number

Two-component signaling proteins are among the most prevalent bacterial genes, and histidine kinases and response regulators constitute two of the largest paralogous gene families in bacteria (Galperin, 2005). Both kinases and regulators are easily identified by sequence homology, in contrast to many eukaryotic signaling systems in which protein kinases are easily identified but their substrates are not. Many histidine kinases are encoded in the same operon as their cognate regulators, allowing for cognate pairs to be identified through sequence analysis. Census-taking is thus straightforward and easily applied to fully sequenced bacterial genomes (Figure 1.2A). Such analyses have revealed that the total number of two-component genes per genome typically grows as a square of the genome size (Galperin, 2005) (Figure 1.2B). In addition, the number of two-component genes appears to correlate strongly with ecological and environmental niche (Alm et al., 2006; Galperin, 2005; Galperin et al., 2001; Koretke et al., 2000). Bacteria

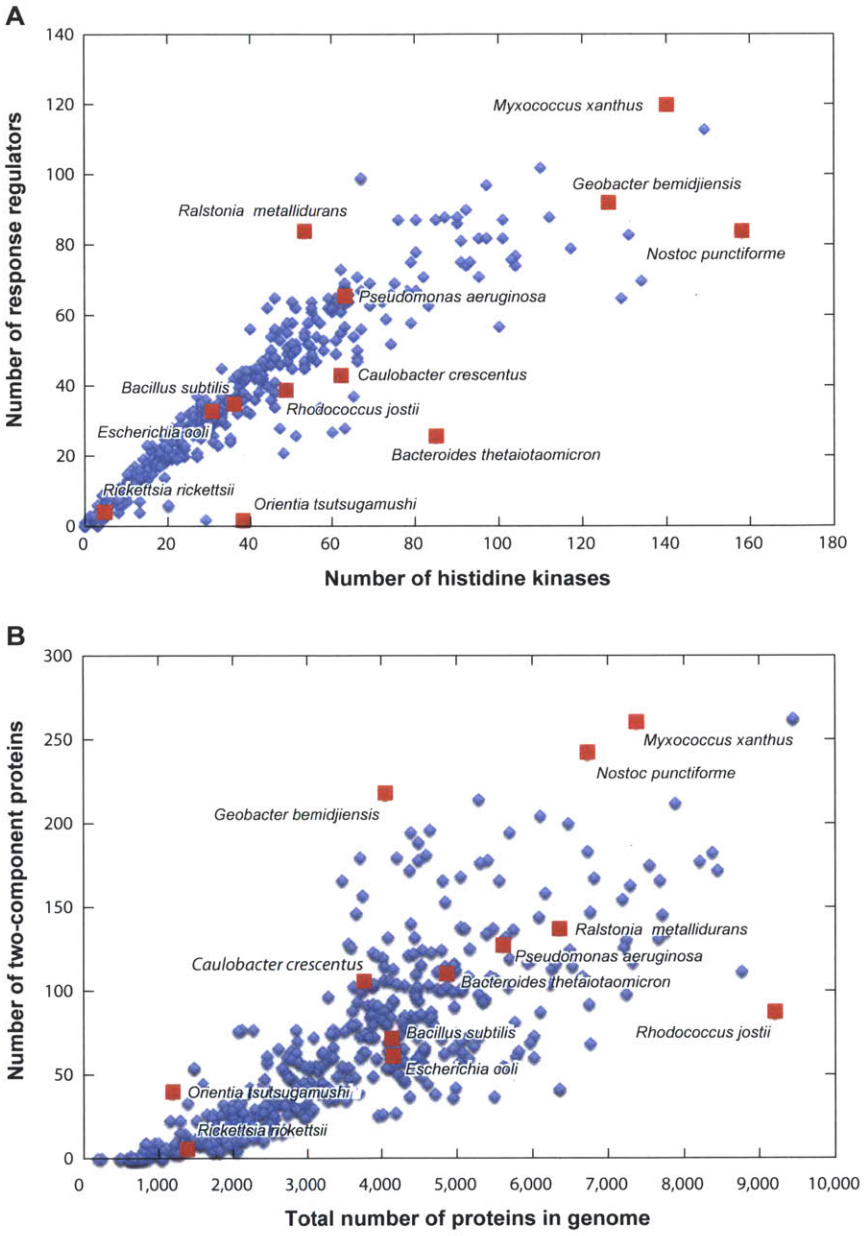


Figure 1.2 Diversity of two-component signaling gene content in bacterial genomes.

(A) Plot showing the number of histidine kinases and response regulators in set of bacterial genomes. Generally, most genomes contain equal numbers of kinases and regulators, as pathways typically comprise one kinase and one cognate regulator. When the ratio is not 1:1, there are usually more kinases than regulators, suggesting response regulators may sometimes integrate signals from multiple kinases. (B) Plot showing the number of two-component proteins as a function of genome size for the same organisms as in panel A. Each plot is based on 504 bacterial genomes with data taken from (Galperin et al., 2010). A handful of well-studied and notable species are marked with red squares.

that live primarily in constant environments typically encode relatively few two-component signaling genes, even taking into account their smaller genome sizes and characteristic reductive genome evolution. In the extreme, many obligate intracellular parasites and endosymbionts harbor only a few pathways or sometimes none at all, as with *Mycoplasma* and *Amoebophilus*. By contrast, bacteria that inhabit rapidly changing or diverse environments typically encode large numbers of these signaling proteins. Extreme cases include *Myxococcus xanthus* with 136 histidine kinases and 127 response regulators and *Nostoc punctiforme* with 160 kinases and 98 regulators (Ulrich and Zhulin, 2010) (Figure 1.2). In some species, nearly 3% of the genome encodes for histidine kinases alone (Galperin, 2005). These patterns of gene content strongly suggest that organisms expand their set of two-component signaling genes to help adapt to fluctuations in their environment.

Although most abundant in the genomes of gram-negative bacteria and cyanobacteria, two-component signaling genes are found in all three domains of life (Koretke et al., 2000; Schaller et al., 2011). However, they are considerably less abundant in archaea and eukaryotes. The majority of systems found in eukaryotes involve hybrid kinases and phosphorelays; whether there is selective pressure against canonical systems is unknown. Many of the archaeal and eukaryotic systems likely originated through multiple, independent lateral gene transfers from bacteria (Kim and Forst, 2001; Koretke et al., 2000); plants likely gained two-component pathways through the integration of chloroplast genes into the nuclear genome (Martin et al., 2002). In plants, the two-component genes obtained through lateral transfer likely expanded through duplication

and diversification and now play integral roles in diverse developmental pathways (Ren et al., 2009).

Whereas two-component genes are found in yeasts, filamentous fungi, slime molds, and plants, they are conspicuously absent from higher eukaryotes and metazoans. The absence of two-component signaling proteins from humans, combined with their well-documented role in bacterial pathogenesis (Gooderham and Hancock, 2009; Miller et al., 1989), has made these proteins attractive new targets for antibiotic development (Gotoh et al., 2010). Indeed, a recent study sequenced individual isolates over the course of a *Burkoholderia dolosa* outbreak of patients with cystic fibrosis and discovered that a two-component system, FixL/FixJ was under the strongest positive selection of any gene over the course of infection (Lieberman et al., 2011). Evolutionarily, the absence of two-component systems in metazoans begs the question of why they were supplanted as the primary means of signaling by pathways employing serine, threonine, and tyrosine phosphorylation. Although a definitive answer is lacking, the intrinsic lability of phosphoryl groups on aspartates may have contributed. In eukaryotes, a need for longer, more stable outputs may have been desirable, and perhaps necessary, for transmitting signals from the cell membrane to the nucleus without signal loss *en route* in the form of phosphoryl group hydrolysis. Consistent with this idea, many of the two-component pathways in eukaryotes do not regulate transcription and instead target other cytoplasmic proteins. For example, in *Saccharomyces cerevisiae*, the Sln1-Ypd1-Ssk1 phosphorelay modulates the activity of a MAP kinase pathway that is also located in the cell membrane (Posas et al., 1996). Nevertheless, there are cases of eukaryotic response regulators that directly affect transcription, particularly in plants. In these cases a histidine

phosphotransferase typically shuttles phosphoryl groups from a cytoplasmic hybrid histidine kinase to a response regulator in the nucleus that is constitutively associated with the DNA (Grefen and Harter, 2004; Imamura et al., 2001). Signal transmission may be successful in these cases because a histidyl-phosphate moiety is considerably more stable than an aspartyl-phosphate moiety.

Where did two-component signaling pathways, in any organism, evolve from in the first place? Given their ancient origin, an unequivocal answer to this question may not be attainable. However, one clue is that histidine kinases share distant homology in their ATP-binding domains with Hsp90, the mismatch repair protein MutL, and type II topoisomerases (Dutta and Inouye, 2000; Dutta et al., 1999). These proteins, members of the so-called GHKL superfamily, are thought to bind ATP in similar ways and share significant structural similarities; in some cases this domain is used to drive ATP hydrolysis, and in the case of histidine kinases the γ-phosphoryl group is transferred to a histidine in the DHp domain. It is thus plausible that histidine kinases emerged from one of these ATPases. In contrast to histidine kinases, there are no such weak homologies for response regulators and their origin remains a mystery.

There are likely two sources of histidine phosphotransferases. Some, particularly those that are monomeric (Ulrich et al., 2005; Xu et al., 2009), may have evolved *de novo* from a range of other proteins, as there are few structural and sequence requirements to function as a histidine phosphotransferase beyond a phosphorylatable histidine within an α-helical bundle. Others are dimeric and may have evolved through the degeneration of histidine kinases. For example, *Bacillus subtilis* Spo0B has two domains with significant similarity to those in histidine kinases (Varughese et al., 1998; Zapf et al., 2000). The

domain that contains the crucial histidine is structurally similar to the DHp domain of histidine kinases and the other is topologically and structurally similar to a CA domain but lacks key residues usually involved in ATP binding. A similar scenario of recruitment and degeneration of a histidine kinase may hold for the phosphotransferase ChpT in *Caulobacter crescentus* (Biondi et al., 2006). In general, however, evolutionary analysis of histidine phosphotransferases has been limited by the difficulty of identifying these proteins from sequence alone, in contrast to histidine kinases and response regulators.

Mechanisms for evolving changes in two-component signaling gene content

Given the prevalence of two-component signaling pathways in bacterial genomes, it is natural to ask how new proteins and pathways arise. The possibilities fall into two broad categories: gene duplication and divergence, sometimes also referred to as lineage-specific expansion (LSE), and lateral gene transfer (LGT). To assess the contributions made by these two mechanisms, one study systematically examined the origins of histidine kinases from 207 genomes, using BLAST to identify the closest homologs of each kinase (Alm et al., 2006). For those most closely related to a kinase within the same genome, gene duplication, or lineage-specific expansion, was inferred as the source. If the closest homolog was from a closely related species, and if a gene tree built from all homologs matched a species tree, the kinase was classified as ancient and vertically transmitted. If, however, the closest homolog for a given kinase was from a distantly related species, lateral gene transfer was invoked. This interpretation assumes that multiple gene losses are less parsimonious and hence less likely to have occurred. However, gene loss occurs at very high rates in bacteria. In addition, inferences of lateral

transfer can be confounded by the inaccuracy of sequence-based distances and heterotachy, the notion that substitution rates in different lineages often vary significantly (Kurland et al., 2003).

Nevertheless, lateral gene transfer of two-component pathways undoubtedly has occurred and these systematic studies provide a general sense of the frequency, both across all species and within individual genomes (Alm et al., 2006). Overall, lineage-specific expansion, or gene duplication, appears to explain the origin of the vast majority of kinases. However, the relative balance of duplication and lateral transfer varies substantially from species to species. For example, in *Streptomyces coelicolor*, essentially all of its 140 histidine kinases appear to be ancient or derived from lineage-specific expansions. By contrast, in *Pseudomonas syringae* and *Ralstonia solanacearum*, many of the recently derived kinases probably came from lateral transfer events.

The lateral transfer of genes in bacteria can occur in several ways, including through phage and plasmids, by direct conjugation, or by competence and the direct uptake of extracellular DNA. There are examples of two-component signaling genes encoded on plasmids, such as the VanR-VanS system found in enterococci that senses and responds to vancomycin (Arthur et al., 1992; Wright et al., 1993). In *R. solanacearum*, many of the laterally derived histidine kinases are encoded on a megaplasmid that may have moved laterally (Salanoubat et al., 2002). There are also cases of two-component signaling proteins encoded on pathogenicity islands, such as the SpiR-SsrB system in *Salmonella*, which frequently move through conjugation (Deiwick et al., 1999). However, for many chromosomally encoded two-component genes derived by lateral transfer, the mechanism of transfer remains difficult to infer.

Both gene duplication and lateral transfer events have occurred more frequently than suggested by phylogenetic analyses. However, in most cases the newly introduced genes were likely eliminated from the genome, and thus are no longer present in extant species. Bacteria typically have high rates of gene loss through mutation and deletion. Indeed, histidine kinases and response regulators are among the most common pseudogenes present in bacterial genomes (Liu et al., 2004); these pseudogenes likely arose through relatively recent duplications or lateral transfers, and were then inactivated, but have not yet been removed from the genome. To be fixed in a population, duplicated or laterally transferred genes must provide a substantial selective advantage within a relatively short period, as gene loss and pseudogeneization occur rapidly in bacteria (Hooper and Berg, 2003; Kuo and Ochman, 2010).

The function of a particular two-component system can also influence its evolutionary history. For example, an analysis of six species of *Xanthomonas* compared the complement of signaling genes present in each genome and found extensive gene loss (Qian et al., 2008). Notably, those pathways involved in *Xanthomonas* pathogenesis were never lost or duplicated, whereas other, presumably less critical, pathways experienced more flux. Similarly, in *C. crescentus*, where two-component signaling proteins play important roles in cell cycle progression and development pathways, those that are essential for viability are highly-conserved in other *Alphaproteobacteria*, whereas those that are non-essential in *C. crescentus* are less well-conserved (Skerker et al., 2005). In most species there is probably a core set of two-component proteins that is maintained and relatively fixed, and an additional set that can be lost, or modified, more easily.

This notion of fixed core signaling genes and malleable auxiliary factors has been well-characterized in the context of bacterial chemotaxis, which centers on a two-component pathway, CheA-CheY. In *Escherichia coli*, where chemotaxis has been best studied, signal recognition requires a methyl-accepting chemoreceptor protein (MCP) and an adaptor protein CheW. Virtually all chemotactic bacteria encode orthologs of these core components: MCP, CheW, CheA, and CheY (Wuichet and Zhulin, 2010). In contrast, many of the auxiliary components, including the methyltransferase CheR and the methylesterase CheB that influence signal adaptation, are not universally conserved and are often missing or replaced by other types of regulators (Wuichet and Zhulin, 2010).

Gene fusions, rearrangements, and duplications

Many two-component genes are encoded in operons as cognate kinase-regulator pairs, allowing for the duplication or lateral transfer of an intact signaling pathway. It is rare to see operon shuffling and the mixing and matching of genes encoded in operons. Hence, for a given kinase-regulator pair, the orthologs are also usually found together in an operon and in the same relative order (Koretke et al., 2000; Whitworth and Cock, 2009). Fusions of kinases and regulators to create hybrid kinases also seem to be rare, but there are some examples. For instance, analysis of six species of *Xanthomonas* found that the individual domains of a hybrid histidine kinase in one species were most similar to, and likely derived from, an operonic kinase-regulator pair encoded as separate open reading frames in a closely related species (Qian et al., 2008). Such fusions probably occur through the mutation of stop codons in operons where the histidine kinase is upstream of the response regulator, although hybrid kinases may also form through the fusion of previously separated genes (Qian et al., 2008; Whitworth and Cock, 2009; Zhang and

Shi, 2005). As might be expected, fusion events that create hybrid kinases are rare for response regulators that contain DNA-binding output domains (Cock and Whitworth, 2007; Zhang and Shi, 2005). There are, however, examples of such hybrid kinases (Sonnenburg et al., 2006), but the mechanism by which these systems regulate transcription remains unclear.

Although *E. coli* encodes 55 of its 62 two-component genes in operons, many organisms encode a substantial fraction of their two-component genes as orphans. Frequently only one gene from an operon is duplicated (or both are duplicated and one is lost) resulting in the production of orphan two-component signaling genes. An orphan kinase, however, may retain the ability to phosphorylate the regulator in the operon from which it was derived. Such duplication events, coupled with a change in kinase input domain, may be a primary mechanism for generating cross-regulated systems in which multiple, independent signals can trigger the same response. A classic example is in *B. subtilis*, in which each of the five orphan kinases KinA/B/C/D/E, which probably evolved through duplication, can each phosphorylate Spo0F and initiate the sporulation phosphorelay (Stephenson and Hoch, 2002). Similarly, duplication of only the response regulator from a given kinase-regulator pair can lead to a scenario in which a single sensor kinase can drive multiple outputs. For example, in cyanobacteria NblS-RpaB forms an essential two-component system. During divergence of the cyanobacteria in the clade including *Synechococcus* species, a duplication of RpaB produced a second response regulator SrrA. This regulator retained the ability to be phosphorylated by NblS, but appears to affect transcription manner different than that by RpaB (Lopez-Redondo et al., 2010).

Evolution of signaling protein structure and function

Gene duplication and lateral gene transfer ultimately provide the raw material for generating new two-component signaling pathways. But what happens immediately after new signaling genes are introduced? Owing to large population sizes and selective pressure to minimize genome size (Mira et al., 2001), new signaling proteins presumably must quickly gain new functions to be retained. There are undoubtedly many mutations that must occur to produce a pathway that can respond to a new input or effect a new output. These mutations presumably include single amino acid substitutions, although rapid changes in function may rely heavily on larger-scale rearrangements such as domain shuffling. Below I summarize the current understanding of how cells generate new signaling functions from duplicated genes, focusing on (i) changes in kinase sensory domains and pathway inputs, (ii) changes in response regulators and pathway outputs, and (iii) changes required to insulate new pathways from existing pathways, before describing the work that I have done, particularly regarding the question of interaction specificity between kinase and regulator.

Histidine kinase sensory domain evolution

After the duplication of a histidine kinase, whether alone or with a cognate response regulator, the duplicate histidine kinases must differentiate themselves and find new roles within the signaling network of a cell. One mechanism to accomplish this is through changes in the sensory domains of one or both kinases (Cheung and Hendrickson, 2010; Krell et al., 2010). For most orthologous kinases, the sensory domains are less well-conserved than their catalytic domains. The ability to sense a new signal often arises via domain shuffling, which may occur coincident with, or shortly after, a duplication. Over

70% of recently duplicated histidine kinases show an input domain structure different from that of their closest paralog (Alm et al., 2006) (Figure 1.3A). Domain shuffling can also occur between histidine kinases and other proteins. Sequence analyses indicate that the sensory domains of some histidine kinases are closely related to domains found on other types of proteins, including serine/threonine kinases (Zhulin et al., 2003), chemotaxis proteins, and diguanylate cyclases (Zhang and Hendrickson, 2010).

The domain shuffling observed in histidine kinases suggests that these proteins are intrinsically modular and, consequently, that the rational design of new kinases may be possible. Indeed, several groups have successfully fused the conserved phosphotransfer and catalytic domains from a histidine kinase to the sensory domain of another kinase, or even the sensory domain of completely unrelated proteins. The first such example, dubbed Taz, is a chimeric protein that fused the sensory domain of the aspartate chemoreceptor Tar with the DHp and CA domains of the model histidine kinase EnvZ, producing an aspartate-responsive kinase (Utsumi et al., 1989). In addition to demonstrating the fundamental modularity of histidine kinases, Taz has been used to dissect the functions and activities of EnvZ *in vivo* (Dutta et al., 2000; Jin and Inouye, 1993; Zhu and Inouye, 2003). Other functional chemoreceptor-EnvZ constructs have also been made (Baumgartner et al., 1994; Rampersaud et al., 1991).

How does domain shuffling, either during evolution or during rational construction of chimeric proteins, produce successful, signal-responsive proteins? Is there a particular way in which sensory domains must be fused to the catalytic domains to function? These questions were recently examined in the context of a chimeric protein that fused a light-sensing PAS domain, taken from the *B. subtilis* protein YtvA (which is not a kinase),

with the DHp and CA domains of the histidine kinase FixL from *Bradyrhizobium japonicum*. Successful fusions of the PAS domain to FixL led to light-responsive changes in FixL signaling and FixL-FixJ-dependent gene expression (Moglich et al., 2009a). Successful fusions had linkers, which form coiled coils, separating the PAS and DHp domains that differed in length by exactly seven amino acids. Inspection of other histidine kinases containing PAS domains further revealed that the linkers are of variable lengths, but often differ by multiples of seven. Together, these results suggest that maintaining the heptad periodicity of the coiled-coil linker may be critical to the construction of functional chimeras, either during evolution or for rational engineering purposes. Further work demonstrated that, by following similar rules, multiple PAS domains could be engineered into the same kinase, allowing it to integrate multiple signals (Moglich et al., 2010). Naturally occurring histidine kinases also often have multiple input domains, suggesting that partial gene duplications, in which only a single input domain is duplicated, may be a common mechanism for generating input diversity. In sum, these efforts to engineer novel proteins are not only producing valuable tools, but are also providing important new insights into how domain shuffling occurs and how it contributes to the origin of new two-component signaling pathways in nature.

An additional mechanism for acquiring new input signals is through accumulated substitutions in a sensory domain rather than its complete replacement. A prime example comes from the NarX and NarQ sensor kinases in *E. coli* (Figure 1.3B). A gene duplication event led to the emergence of these two related kinases, although which is more ancestral is unclear. Nevertheless, studies of signal recognition have demonstrated that NarQ responds to both nitrate and nitrite whereas NarX responds preferentially to

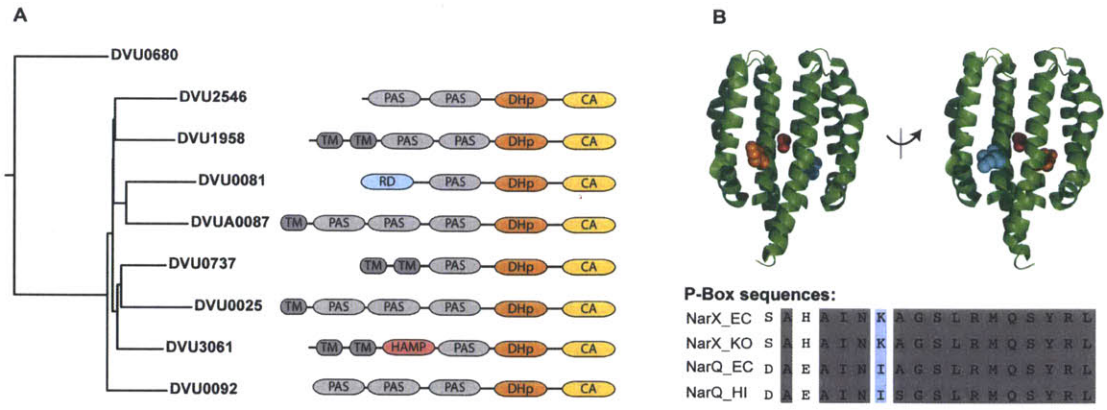


Figure 1.3 Evolution of sensory domains.

(A) A tree of a recent lineage specific expansion in *Desulfovibrio vulgaris* shows the extent of domain shuffling that can occur after duplication. These paralogs show differences in the number and types of signaling domains, as well as in the presence and number of transmembrane domains. The lineage specific expansion was identified from (Alm et al., 2006). A neighbor-joining tree was constructed using the PHYLIP software package (Felsenstein, 1989) with another *D. vibrio* kinase, DVU0680, as the outgroup. Only the DHp and CA domains of the kinases were used to build the tree. Domains were identified using the Pfam database (Punta et al., 2012) and colored according to the same scheme as used in Figure 1.1. Transmembrane domains were predicted by TMHMM (Krogh et al., 2001). The lineage specific expansion was rapid, as shown by the difficult to resolve branches between members of the expansion. The diversity in the number of PAS domains could represent partial duplications of the histidine kinase. (B) The crystal structure of the ligand binding domain of NarX shown in complex with NO_3^- (Cheung and Hendrickson, 2009). NarX autophosphorylates preferentially in the presence of NO_3^- when compared to NO_2^-. NarQ, which is a paralog of NarX, autophosphorylates in response to both NO_2^- and NO_3^-. A mutation of a lysine (shown in orange) to an isoleucine (shown in blue), causes NarX to behave more like NarQ in that it responds equally to both NO_2^- and NO_3^- (Williams and Stewart, 1997). The larger and more hydophobic isoleucine may cause a kink in the helices that affect how they transduce the signal in response to ligand binding. Shown below is an alignment of the P-boxes of NarX and NarQ orthologs from *Escherichia coli* (EC), *Klebsiella oxytoca* (KO), and *Haemophilus influenze* (HI). All residues that are conserved throughout the alignment are highlighted in gray, while the residue that determines NarX-like vs. NarQ-like ligand discrimination is highlighted in blue.

nitrate (Rabin and Stewart, 1993). Although the periplasmic domains of NarQ and NarX are significantly diverged, they do share substantial similarity, particularly in a region critical to ligand binding (Cheung and Hendrickson, 2009). Notably, a single point mutation in this region of NarX that substitutes a lysine with an isoleucine, as found at the equivalent position in NarQ, reduced the ability of NarX to discriminate between nitrate and nitrite, rendering a more NarQ-like response pattern (Williams and Stewart,

1997) (Figure 1.3B). This study highlights how the accumulation of single point mutations is a plausible means of rapidly generating new and different inputs to two-component signaling pathways.

Divergence and evolution of pathway outputs

Within a two-component signaling pathway, the response regulator is the ultimate arbiter of physiological change. How does the output of a response regulator evolve, and how are new output responses generated by response regulators after they emerge through duplication or following lateral transfer? As the majority of response regulators direct changes in gene expression, the evolution of pathway outputs can be easily studied by following changes in target genes.

One of the best-studied examples is the PhoQ-PhoP system found in the *Enterobacteriaceae*. In response to low extracellular concentrations of Mg^{2+}, the histidine kinase PhoQ drives phosphorylation of PhoP, which then regulates gene expression. The direct regulon of PhoP has been mapped in both *Salmonella enterica* serovar Typhimurium and *Yersinia pestis* (Perez et al., 2009), which probably shared a common ancestor ~200 million years ago. Strikingly, only three genes were directly regulated by PhoP in both species: the autoregulated *phoQ* and *phoP* genes and *slyB*, which encodes a lipoprotein thought to be a critical regulator of PhoQ activity (Figure 1.4A). There were also some genes, such as *pbgP* and *ugd*, that were directly regulated in one species, but indirectly regulated in the other; the overall regulatory logic for these genes was thus conserved, but the precise mechanism has changed. Despite these examples, the vast

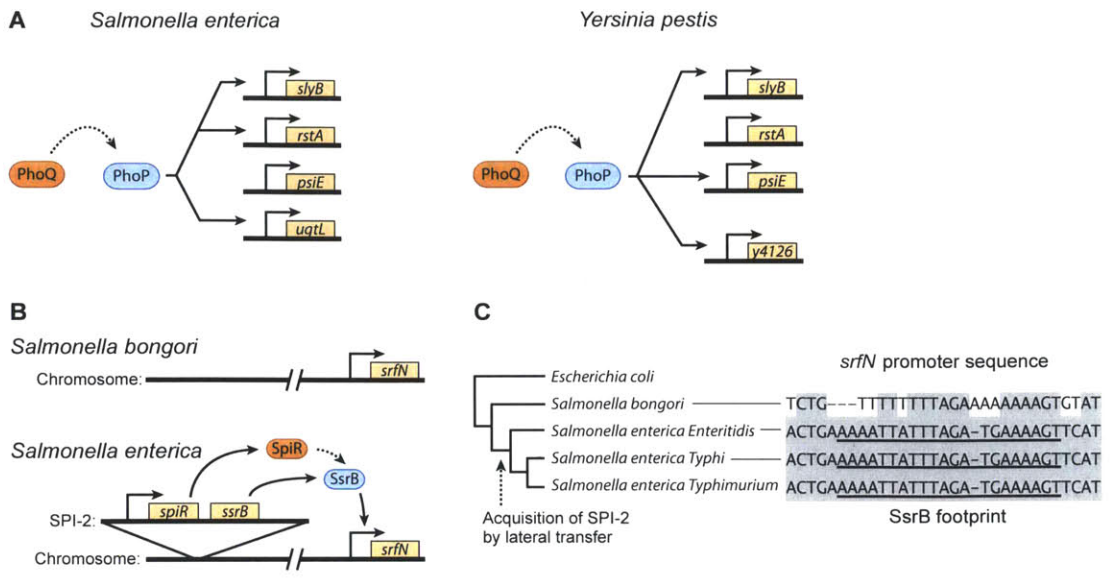


Figure 1.4 Evolution of transcriptional circuits controlled by two-component pathways.

(A) Examples of genes directly regulated by the two-component pathway PhoQ-PhoP in *Salmonella enterica* and *Yersinia pestis*. *slyB* is conserved and directly regulated by PhoP in both species. *rstA* and *psiE* are conserved but directly regulated by PhoP in only one of the two species. *ugtL* and *y4126* are directly regulated and are unique to *S. enterica* and *Y. pesis*, respectively. (B) Schematic of the *Salmonella bongori* and *S. enterica* chromosomes, each harboring a *srfN* ortholog. The horizontally acquired SpiR-SsrB system, encoded on *Salmonella* pathogenicity island 2 (SPI-2) in *S. enterica* but not *S. bongori,* evolved to transcriptionally activate *srfN*. (C) De novo evolution of a response regulator-binding site. SPI-2 encodes the two-component pathway SpiR-SsrB, which was acquired after the divergence of *S. enterica* from *S. bongori*. The gene *srfN*, ancestral to the *Salmonella* lineage, accumulated promoter mutations that enabled activation by SsrB, a transcriptional link that contributes to *Salmonella* virulence. The relevant portion of the *srfN* prmoter is shown with conserved poisitions shaded gray and the region bound by SsrB in *S. enterica* underlined.

majority of genes directly regulated by PhoP in each organism were not conserved. Instead, transcriptional rewiring appears to have been prevalent since the divergence of *Salmonella* and *Yersinia*, leading to the gain and loss of PhoP-regulated genes in each species (Figure 1.4A). It is tempting to speculate that these changes have tailored the response of each species to magnesium limitation.

Notably, the change in PhoP regulons between *Salmonella* and *Yersinia* may not always result from a simple gain or loss of PhoP binding sites. In some cases, regulon differences may reflect changes in (i) the orientation of, and distance between, a PhoP binding site and the transcriptional start site and (ii) concomitant changes in how PhoP recruits RNA polymerase. For instance, the PhoP binding site in the promoter of *mgtC* in *Yersinia* is located in a position and orientation that enables gene activation by *Yersinia* PhoP, but not by *Salmonella* PhoP, even though *Salmonella* PhoP can bind the *mgtC* promoter (Perez and Groisman, 2009b). The ability to change the targets of a response regulator without necessarily changing DNA-binding sites is also seen in *Desulfovibrio*, in which two recently duplicated response regulators share DNA-binding motifs but regulate non-overlapping target genes (Rajeev et al., 2011). Point mutations in OmpR have also been identified that allow it to activate the *kdpABC* operon, usually activated by KdpE, not by changing DNA binding but by changing the ability to interact with RNA polymerase while bound to the promoter (Ohashi et al., 2005). Thus with a single point mutation, and without any changes needed in the promoters of target genes, a duplicated response regulator can regulate a new set of target genes. Collectively, these studies demonstrate that two-component pathway outputs can evolve through changes in the DNA-binding sites of response regulators or through changes in how response regulators interact with RNA polymerase. They also highlight the critical need to couple computational analyses of binding sites with experimental studies to reveal the functional and evolutionary consequences of binding site conservation or loss.

Changes in response regulator outputs may also frequently occur after duplication or lateral transfer events. After gene duplication, a change in the output response of one or

both regulators is likely a critical step in the establishment of new functions and, consequently, the maintenance of the duplicated proteins. For instance, in *E. coli*, a duplication event likely gave rise to the paralogous systems NarX-NarL and NarQ-NarP which respond to nitrate and nitrite in anaerobic conditions (Rabin and Stewart, 1993). While the regulators NarP and NarL share significant similarity and even recognize highly similar consensus binding sites, divergent evolution has enabled each response regulator to recognize different promoter architectures and to activate different genes (Price et al., 2008). The duplication of the Nar two-component system has thus led to an increase in complexity of the transcriptional control of genes necessary for growth in anaerobic conditions.

The evolution of response regulator outputs in response to lateral gene transfer has also been recently explored. A particularly illuminating example comes from studies of *Salmonella* pathogenicity island-2 (SPI-2), which encodes a two-component signaling system called SpiR-SsrB (Figure 1.4B-C). In addition to regulating the expression of other SPI-2-encoded genes, the response regulator SsrB directly regulates the expression of genes outside SPI-2 (Worley et al., 2000), indicating that SsrB-binding sites probably evolved *de novo* within the promoters of these genes. This hypothesis was tested by examining the evolution of a *Salmonella* gene, *srfN* (Osborne et al., 2009). This gene is ancestral to the *Salmonella* lineage and present in both *S. enterica* and *S. bongori*. By contrast, SPI-2 and SsrB are found in *S. enterica* but not *S. bongori* (Figure 1.4B). A comparison of the cis-regulatory regions of *srfN* indicated that the binding site for SsrB was not present in *S. bongori* meaning it likely arose in the lineage leading to *S. enterica* (Figure 1.4C). Importantly, this recruitment of an ancestral gene into the regulon of a

horizontally-acquired response regulator provided *S. enterica* with an adaptive advantage as a pathogen. When the promoter of *S. enterica srfN* was replaced with that found in *S. bongori*, cells were rendered significantly less virulent compared to the wild-type.

Conversely, the genes encoded on SPI-2 have evolved to be regulated by ancestral two-component pathways. A case in point is the expression of *ssrB* and *spiR*, which are themselves regulated by OmpR and PhoP, two response regulators found throughout the *Gammaproteobacteria* (Bijlsma and Groisman, 2005; Lee et al., 2000). By controlling *spiR* and *ssrB*, these ancestral regulators likely help to ensure that virulence genes are maximally expressed when *Salmonella* enters host cells. For instance, the PhoQ-PhoP system is activated by the low-magnesium conditions that *Salmonella* experiences inside host macrophages; the consequent activation of *ssrB* and *spiR* would then drive the expression of virulence genes.

Although this introduction includes only a few cases, it is clear that response regulator outputs can, and do, change rapidly. The observed changes to transcriptional circuitry observed suggest that bacteria are resilient to, and capable of, transcriptional rewiring (Perez and Groisman, 2009a). This notion was tested systematically by artificially rewiring transcriptional connections; promoters for 26 different sigma and transcription factors (including some response regulators) were combined with the open reading frames of 23 of these transcriptional regulators and introduced into *E. coli* cells on a high-copy plasmid (Isalan et al., 2008). Strikingly, over 95% of these constructs, many of which led to substantial transcriptional rewiring, were tolerated, with little to no growth defect under standard laboratory conditions. One implication of this study is that after a new DNA-binding response regulator is introduced by gene duplication or lateral

transfer, there is time to "scan" different regulatory possibilities. A new combination that yields even a slight benefit could then be selected and rapidly fixed in a population. Finally, the evolvability of response regulators and their outputs may also benefit from the fact that most prokaryotic transcription factors regulate only a few genes, either directly or indirectly (Madan Babu and Teichmann, 2003), decreasing the number of binding sites that would need to co-evolve with the DNA-binding domain of a response regulator, thereby increasing the likelihood that they can change (Rajewsky et al., 2002).

Dimerization specificity

After duplication, the generation of new, functional, and insulated pathways requires changes to the residues that mediate homodimerization of histidine kinases and response regulators. To establish new and insulated pathways, substitutions are needed that eliminate heterodimerization of the diverging paralogous proteins while maintaining the ability to homodimerize.

Most, if not all, histidine kinases form homodimers in order to autophosphorylate. There is almost no evidence of physiologically-relevant heterodimerization, with one exception in *Pseudomonas aeruginosa* (Goodman et al., 2009), indicating that histidine kinases must harbor a set of amino acids that enforce homodimerization. Many of these residues are likely to reside in the DHp domain, although upstream domains, such as PAS and HAMP domains, could also contribute to dimerization specificity and stability. To better pinpoint the residues mediating specificity, one recent study looked for coevolving residues in a set of more than 15,000 histidine kinase sequences (Ashenberg et al., 2011). This approach revealed a small set of strongly coevolving residues that mapped primarily

to the DHp domain and mostly within the lower half of the four-helix bundle (Figure 1.5). Homodimerization specificity could be changed through directed mutagenesis of these residues (Ashenberg et al., 2011).

Nearly 50% of response regulators, including all members of the OmpR family (Gao and Stock, 2010), also form homodimers upon phosphorylation. Homodimerization is often crucial for producing an output response as many response regulators have DNA-binding domains and recognize tandem or inverted repeat elements within target promoters. A systematic study of the 17 OmpR-family response regulators from *E. coli* demonstrated that essentially all of them specifically homodimerize (Gao et al., 2008). Although intermolecular interactions on the dimer interface involves highly conserved residues within the receiver domain, some interfacial residues do vary, perhaps providing a mechanism for ensuring homodimerization and excluding heterodimerization (Toro-Roman et al., 2005). As with kinase dimerization amino acid coevolution studies have identified a subset of interfacial residues that may help enforce homodimerization and prevent heterodimerization (Weigt et al., 2009). These residues are likely to change following gene duplication as a means of insulating paralogous response regulators from one another, thereby enabling distinct outputs to result from the phosphorylation of each regulator.

Evolution of phosphotransfer specificity and the insulation of pathways

The flow of information through two-component signaling pathways depends critically on the transfer of phosphoryl groups from a histidine kinase to its cognate response regulator. Despite early suggestions of rampant cross-talk, there is little evidence for such

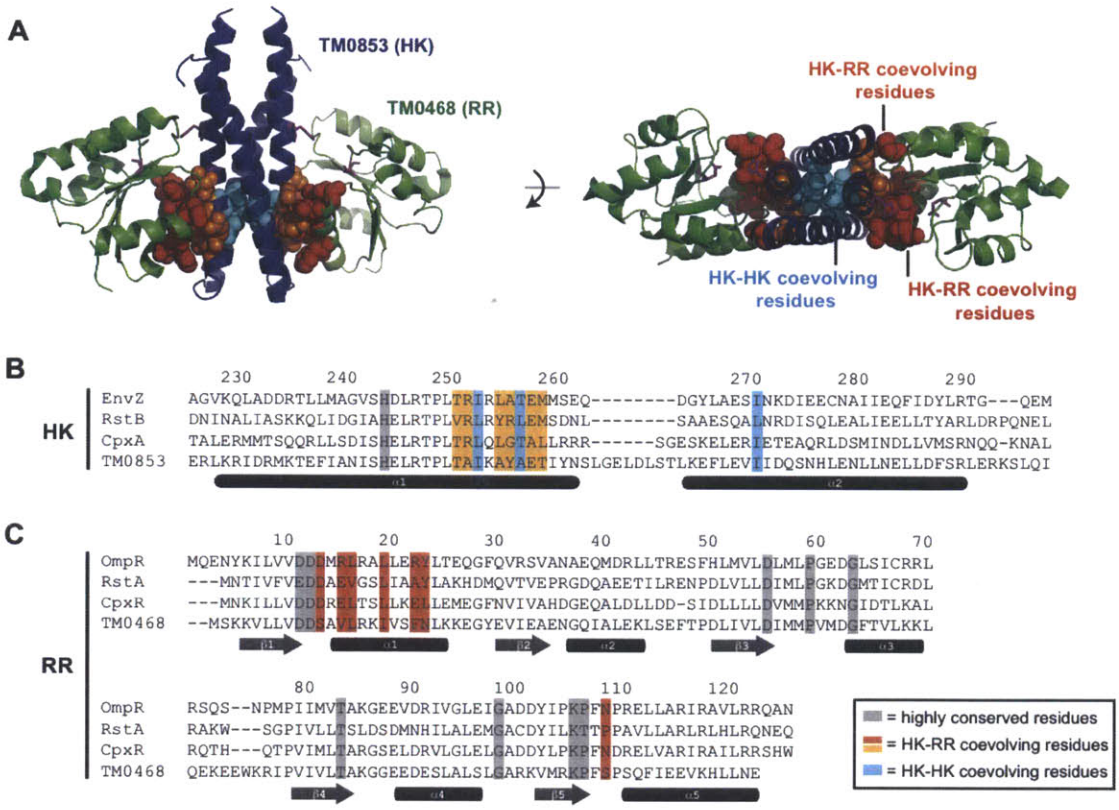


Figure 1.5 Amino acid coevolution in two-component signaling proteins.

(A) Residues that coevolve in cognate pairs of histidine kinases (HKs) and response regulators (RRs) are shown with space-filling on the crystal structure of the *Thermotoga maritima* kinase TM0853 bound to its cognate response regulator TM0468. Only the DHp domain of the kinase and the receiver domain of the response regulator are shown. The histidine and the asparate that are involved in phosphotransfer are shown as sticks in purple. Residues in histidine kinases that coevolve strongly with other kinase residues are shown in cyan, while residues on the kinase that coevolve with those on the response regulator are shown in orange and red respectively. (B-C) coevolving residues from panel A are shown on (B) a sequence alignment of TM053 with three *Escherichia coli* kinases, EnvZ, RstB, and CpxA, and (C) an alignment of TM0468 with three *E. coli* regulators, OmpR, RstA, and CpxR. Secondary structure elements are indicated beneath the primary sequence.

promiscuity *in vivo* with most kinases having one response regulator substrate or, on occasion, two or three (Laub and Goulian, 2007; Skerker et al., 2005; Yamamoto et al., 2005). This *in vivo* preference is mirrored *in vitro,* with histidine kinases harboring a strong kinetic preference for phosphotransfer to their *in vivo* partner. For example, a systematic, global study of phosphotransfer from *E. coli* EnvZ to each of the 32 response

regulators in *E. coli* demonstrated that OmpR was the preferred substrate. EnvZ transferred to other substrates only after extended incubation times (Skerker et al., 2005). These *in vitro* studies demonstrate that the specificity of two-component signaling pathways is based primarily on molecular recognition rather than reliance on scaffolds or other cellular strategies. This observation further suggests that the information necessary for promoting the "correct", or desired, interaction and preventing "incorrect" interactions is encoded at the sequence level (Skerker et al., 2008).

A consequence of relying on molecular recognition for specificity is that, during the course of evolution, any mutation in a residue contributing to a kinase-regulator interaction may disrupt signaling and place cells at a strong fitness disadvantage. Survival would then depend on reversion of the mutation or a compensatory mutation in the partner protein. Consistently, computational analyses of large sets of cognate kinase-regulator pairs have revealed extensive amino acid coevolution (Burger and van Nimwegen, 2008; Skerker et al., 2008; Weigt et al., 2009). Conspicuously, the most significantly coevolving pairs of residues map to the molecular interface formed during phosphotransfer (Casino et al., 2009) suggesting they mediate the specificity of this protein-protein interaction (Figure 1.5). These residues, which have been called specificity residues, map to the same region of the DHp domain that is also responsible for homodimerization specificity. These specificity residues that dictate partnering specificity are on the solvent exposed region of the α-helix, while the dimerization specificity residues are buried within the four-helix bundle. Using *E. coli* EnvZ as a model kinase, a subset of these residues was shown to be sufficient, when mutated, to reprogram substrate specificity both *in vitro* and *in vivo* (Skerker et al., 2008). For

example, mutating as few as three residues in EnvZ to match those found at equivalent positions in RstB led EnvZ to preferentially phosphorylate RstA instead of OmpR (Figure 1.5A-B). Similarly, the response regulator CheY from *Rhodobacter sphaeroides*, has been rationally rewired to interact with non-cognate kinases by mutating the coevolving, specificity-determining residues (Bell et al., 2010). Directed evolution has also been used to rewire two-component specificity. For example, mutants of the *E. coli* kinase CpxA that phosphorylate and dephosphorylate OmpR were selected; many of the mutated residues were also identified in the studies of kinase-regulator coevolution (Siryaporn et al., 2010).

Another mechanism that may be used to enforce the specificity between cognate kinases and response regulators is the use of alternative phosphotransfer mechanisms. Originally, histidine kinases were thought to autophosphorylate in *trans*—i.e. the CA domain of one homodimer phosphorylates the histidine in the DHp domain of the alternate homodimer (Yang and Inouye, 1991). Recent crystallization studies have shown, however, that some histidine kinases autophosphorylate in *cis* (Casino et al., 2009). This *cis* vs. *trans* autophosphorylation mechanism has been attributed to the loop between α-helix-1 and α-helix-2 of the DHp domain (Ashenberg et al., 2013), the same region that was also found to be important for switching the phosphotransfer specificity of some kinases (Skerker et al., 2008). Intriguingly, the histidine kinases for which both the specificity residues and the loop need to be introduced into EnvZ in order to switch phosphotransfer specificity include those kinases that are known to phosphorylate *in cis*. The slightly different contacts made between the response regulator and the DHp domain of a kinase that

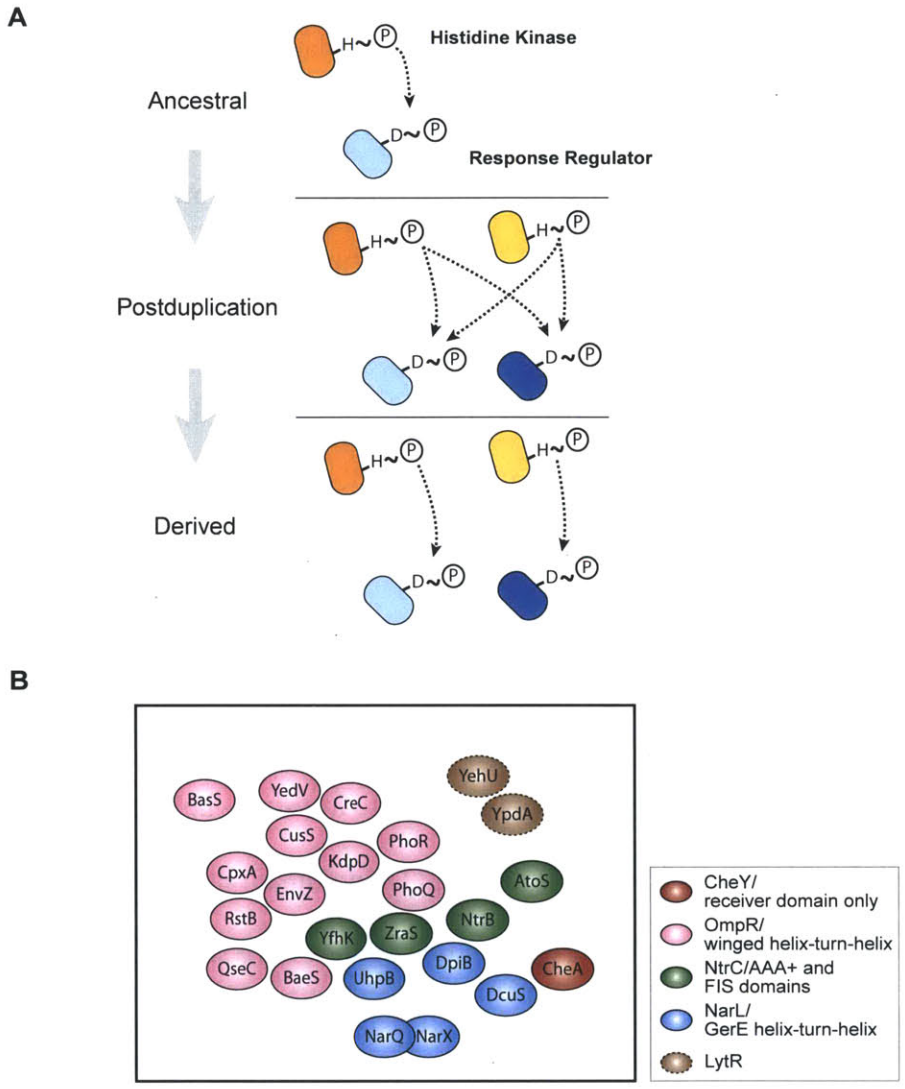


Figure 1.6 Insulation of two-component pathways following gene duplication.

(A) Schematic of major steps in the insulation of two pathways following a duplication event. The duplication of an ancestral pathway initially produces two identical pathways that cross-talk at the level of phosphotransfer. Through the accumulation of mutations in specificity-determining residues, the two pathways can become insulated. A similar process must occur, but is not shown, at the levels of kinase and regulator homodimerization. (B) Schematic summarizing the distribution of histidine kinase sequence space defined by their specificity-determining residues. Each sphere represents the set of response regulators that a given kinase phosphorylates. This set will include, but is not limited to, the cognate response regulator. With the exception of NarQ and NarX, these spheres are presented as nonoverlapping to reflect the minimal cross-talk between pathways. The relative positions of the spheres are based on the ability of individual kinases to phosphorylate non-cognate response regulators after extended times *in vitro* (Skerker et al., 2005; Yamamoto et al., 2005, unpublished data). Positions are approximate, and a two-dimensional representation of a multidimensional sequence space. The diagram is intended to convey a general sense of how kinases are distributed in sequence space. Spheres are colored according to the subfamily of each kinase's cognate response regulator. Spheres with dashed outlines indicate kinases for which no data exist to infer relative positions. Hybrid histidine

kinases, which are under different selective pressures, are excluded.

autophosphorylates in *cis* or in *trans*, may help to enforce phosphotransfer specificity, particularly after a duplication event.

Although specificity-determining residues do coevolve, these correlated changes appear to be rare events as the specificity residues of many kinase-regulator systems are nearly invariant over relatively long timescales. So when and why do specificity residues change and coevolve? One strong possibility is that substitutions occur following gene duplication, helping to insulate the duplicate kinase-regulator pairs from each other (Figure 6A). That is, a series of mutations presumably must occur to prevent cross-talk between two duplicated pathways, while maintaining the interaction within each pair. Such an accumulation of changes in specificity residues, however, is inherently risky business for a bacterium. Due to large population sizes, even slightly deleterious mutations are likely to be quickly removed from the population. Hence, for a new kinase-regulator pair to be maintained in the genome, the mutational intermediates between its initial state and its final, insulated state must be neutral, or nearly neutral. In other words, cognate kinase-regulator pairs must retain their ability to interact as the specificity residues coevolve and find a region of sequence space in which they are insulated from other two-component proteins within the cell. This may be one reason that hybrid kinases are overrepresented among recent gene duplications (Alm et al., 2006). Similarly, after a lateral gene transfer event involving two-component signaling genes, the newly introduced kinase-regulator pair may need to accumulate substitutions in phosphotransfer

specificity residues to avoid cross-talk with existing systems, thereby maintaining the fidelity of information flow within the cell.

The notion of insulation, or orthogonality, in sequence space can be extended from individual, recently duplicated pairs of signaling proteins to the entire complement of two-component signaling proteins in a given organism. For example, all 29 histidine kinases in *E. coli* ultimately arose through some combination of gene duplication and lateral transfer. The net result is a system of signaling pathways that are, with a few exceptions, insulated from one another in sequence space and with respect to phosphotransfer, as observed by global phosphotransfer profiling (Figure 1.6B). This system-wide insulation suggests that negative selection and the avoidance of cross-talk are powerful forces influencing the evolution of two-component signaling proteins. Negative selection has been suggested to influence other paralogous signaling protein families, such as SH3-domain-containing proteins found in eukaryotes (Zarrinpar et al., 2003). Two notable exceptions to the orthogonality of phosphotransfer specificity in *E. coli* are the kinases NarQ and NarX, which share significant similarity in terms of phosphotransfer specificity residues and, consistently, both phosphorylate the response regulators NarL and NarP, although with different kinetic preferences (Noriega et al., 2010). Other exceptions to the orthogonality of specificity residues include hybrid histidine kinases, which phosphotransfer intramolecularly to an attached receiver domain. This spatial arrangement may enforce the specificity of phosphotransfer in hybrid kinases and, consequently, their specificity-determining residues may not be under the same pressure to avoid cross-talk with other two-component pathways.

Research approach

I sought to investigate the evolutionary pressures that act on the interaction between histidine kinases and response regulators. In chapter two, I begin by asking what mutational trajectories can two-component proteins follow after a duplication event in order to insulate the two new pathways? And, in particular, how do duplicated proteins move through the sequence space defined by the specificity-determining residues of histidine kinases and response regulators? Answering these questions through sequence analysis is problematic because transient intermediates may not be captured in extant sequences and the behavior of ancestral or intermediate states is difficult to infer from sequence alone. To circumvent these issues, I experimentally examined all possible specificity intermediates between the two *E. coli* kinases EnvZ and RstB and demonstrated that a cognate kinase-regulator pair can, in fact, move in sequence space from the region occupied by EnvZ-OmpR to that occupied by RstB-RstA while (i) introducing only one mutation at a time, (ii) maintaining the interaction between the kinase and the regulator and (iii) avoiding to introduction of cross-talk to other closely related pathways such as CpxA-CpxR. Notably though, only a small fraction of all possible mutational trajectories satisfy these criteria, indicating that the evolution of signaling proteins post-duplication may be fundamentally constrained.

In chapter three I focus on identifying the evolutionary forces driving the evolution of specificity residues. I showed that in the *Alphaproteobacteria*, NtrB/NtrC likely duplicated to produce NtrB/NtrC and NtrX/NtrY. Subsequent divergence of NtrX/NtrY to insulate the duplicated pathways likely lead to the overlap of NtrX/NtrY sequence space with that already occupied by PhoR/PhoB. In order to resolve the cross-talk, the

specificity residues of PhoR/PhoB evolved specifically in the *Alphaproteobacteria.* I demonstrated that in the absence of duplication, specificity residues remain remarkably conserved between orthologous histidine kinases. Duplication of a two-component system can introduce cross-talk with pre-existing two-component systems within the cell. This cross-talk provides a strong selective disadvantage *in vivo*, resulting in rapid diversification of specificity residues until insulation of two-component systems is achieved.

In chapter four, I shifted my focus from canonical kinases to hybrid kinases. As described above, hybrid kinases are comprised of a histidine kinase that is covalently attached to its cognate receiver domain. Although they comprise nearly 25% of all histidine kinases, they remain less well studied and, unlike canonical two-component systems, the interactions between the kinase and the receiver domain remain incompletely understood. I employed the same approach of covariation to identify specificity residues that are responsible for determining interactions between the histidine kinase and the receiver domain of hybrid kinases and demonstrated that, as expected, their specificity residues are not under the same evolutionary pressure to diverge post-duplication. I showed that after duplication, the pressure for specificity residues to diverge is weaker; the high effective concentration of the covalently attached receiver domain acts to prevent cross-talk with other two component systems within the cell. Surprisingly, however, the specificity residues of a hybrid kinase are still important in order to allow a histidine kinase to interact with its cognate, covalently attached, receiver domain. Only kinases and receiver domains that are able to interact *in vitro* when separated are able to interact when covalently attached. Thus covalent attachment of a

receiver domain serves primarily to prevent cross-talk, not to increase cognate interactions.

In chapter five, I review the conclusions and implications of the work. I also outline the remaining questions, including understanding the size, shape, and distribution of sequence space and the evolution of input and output responses after duplication. In addition, one large unanswered question involves the ubiquity of certain DNA binding domains. Similarly to how I have explored the mechanisms by which cells can coordinate a large number of two-component pathways even given their high sequence and structural homology, most organisms also encode a large number of highly similar response regulators containing homologous DNA-binding domains. I outline several important questions and approaches that can be used in order to understand specificity in the response regulator-DNA-binding interaction, and approaches to elucidate the mechanisms by which two-component systems can be insulated on the transcriptional, or output, level.

Acknowledgements

Support was provided by the National Institutes of Health and the National Science Foundation. M.T.L is an Early Career Scientist at the Howard Hughes Medical Institute.

References

Alm, E., Huang, K., and Arkin, A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol *2*, e143.

Arthur, M., Molinas, C., and Courvalin, P. (1992). The VanS-VanR two-component regulatory system controls synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. J Bacteriol *174*, 2582-2591.

Ashenberg, O., Keating, A.E., and Laub, M.T. (2013). Helix Bundle Loops Determine Whether Histidine Kinases Autophosphorylate in cis or in trans. J Mol Biol.

Ashenberg, O., Rozen-Gagnon, K., Laub, M.T., and Keating, A.E. (2011). Determinants of homodimerization specificity in histidine kinases. J Mol Biol *413*, 222-235.

Baumgartner, J.W., Kim, C., Brissette, R.E., Inouye, M., Park, C., and Hazelbauer, G.L. (1994). Transmembrane signalling by a hybrid protein: communication from the domain of chemoreceptor Trg that recognizes sugar-binding proteins to the kinase/phosphatase domain of osmosensor EnvZ. J Bacteriol *176*, 1157-1163.

Bell, C.H., Porter, S.L., Strawson, A., Stuart, D.I., and Armitage, J.P. (2010). Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex. PLoS Biol *8*, e1000306.

Bijlsma, J.J., and Groisman, E.A. (2005). The PhoP/PhoQ system controls the intramacrophage type three secretion system of Salmonella enterica. Mol Microbiol *57*, 85-96.

Biondi, E.G., Reisinger, S.J., Skerker, J.M., Arif, M., Perchuk, B.S., Ryan, K.R., and Laub, M.T. (2006). Regulation of the bacterial cell cycle by an integrated genetic circuit. Nature *444*, 899-904.

Burbulys, D., Trach, K.A., and Hoch, J.A. (1991). Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay. Cell *64*, 545-552.

Burger, L., and van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol *4*, 165.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Cheung, J., and Hendrickson, W.A. (2009). Structural analysis of ligand stimulation of the histidine kinase NarX. Structure *17*, 190-201.

Cheung, J., and Hendrickson, W.A. (2010). Sensor domains of two-component regulatory systems. Current Opinion in Microbiology *13*, 116-123.

Cock, P.J., and Whitworth, D.E. (2007). Evolution of prokaryotic two-component system signaling pathways: gene fusions and fissions. Mol Biol Evol *24*, 2355-2357.

Deiwick, J., Nikolaus, T., Erdogan, S., and Hensel, M. (1999). Environmental regulation of Salmonella pathogenicity island 2 gene expression. Mol Microbiol *31*, 1759-1773.

Dutta, R., and Inouye, M. (2000). GHKL, an emergent ATPase/kinase superfamily. Trends Biochem Sci *25*, 24-28.

Dutta, R., Qin, L., and Inouye, M. (1999). Histidine kinases: diversity of domain organization. Mol Microbiol *34*, 633-640.

Dutta, R., Yoshida, T., and Inouye, M. (2000). The critical role of the conserved Thr247 residue in the functioning of the osmosensor EnvZ, a histidine Kinase/Phosphatase, in Escherichia coli. J Biol Chem *275*, 38645-38653.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164-166.

Galperin, M.Y. (2005). A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol *5*, 35.

Galperin, M.Y. (2006). Structural classification of bacterial response regulators: diversity of output domains and domain combinations. J Bacteriol *188*, 4169-4182.

Galperin, M.Y., Higdon, R., and Kolker, E. (2010). Interplay of heritage and habitat in the distribution of bacterial signal transduction systems. Mol Biosyst *6*, 721-728.

Galperin, M.Y., Nikolskaya, A.N., and Koonin, E.V. (2001). Novel domains of the prokaryotic two-component signal transduction systems. FEMS Microbiol Lett *203*, 11-21.

Gao, R., Mack, T.R., and Stock, A.M. (2007). Bacterial response regulators: versatile regulatory strategies from common domains. Trends Biochem Sci *32*, 225-234.

Gao, R., and Stock, A. (2009). Biological insights from structures of two-component proteins. Annu Rev Microbiol *63*, 133-154.

Gao, R., and Stock, A.M. (2010). Molecular strategies for phosphorylation-mediated regulation of response regulator activity. Curr Opin Microbiol *13*, 160-167.

Gao, R., Tao, Y., and Stock, A.M. (2008). System-level mapping of Escherichia coli response regulator dimerization with FRET hybrids. Mol Microbiol *69*, 1358-1372.

Gooderham, W.J., and Hancock, R.E. (2009). Regulation of virulence and antibiotic resistance by two-component regulatory systems in Pseudomonas aeruginosa. FEMS Microbiol Rev *33*, 279-294.

Goodman, A.L., Merighi, M., Hyodo, M., Ventre, I., Filloux, A., and Lory, S. (2009). Direct interaction between sensor kinase proteins mediates acute and chronic disease phenotypes in a bacterial pathogen. Genes Dev *23*, 249-259.

Gotoh, Y., Eguchi, Y., Watanabe, T., Okamoto, S., Doi, A., and Utsumi, R. (2010). Two-component signal transduction as potential drug targets in pathogenic bacteria. Curr Opin Microbiol *13*, 232-239.

Grefen, C., and Harter, K. (2004). Plant two-component systems: principles, functions, complexity and cross talk. Planta *219*, 733-742.

Hooper, S.D., and Berg, O.G. (2003). On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol *20*, 945-954.

Huynh, T.N., and Stewart, V. (2011). Negative control in two-component signal transduction by transmitter phosphatase activity. Mol Microbiol *82*, 275-286.

Imamura, A., Yoshino, Y., and Mizuno, T. (2001). Cellular localization of the signaling components of Arabidopsis His-to-Asp phosphorelay. Biosci Biotechnol Biochem *65*, 2113-2117.

Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., and Serrano, L. (2008). Evolvability and hierarchy in rewired bacterial gene networks. Nature *452*, 840-845.

Jin, T., and Inouye, M. (1993). Ligand binding to the receptor domain regulates the ratio of kinase to phosphatase activities of the signaling domain of the hybrid Escherichia coli transmembrane receptor, Taz1. J Mol Biol *232*, 484-492.

Kim, D., and Forst, S. (2001). Genomic analysis of the histidine kinase family in bacteria and archaea. Microbiology *147*, 1197-1212.

Koretke, K.K., Lupas, A.N., Warren, P.V., Rosenberg, M., and Brown, J.R. (2000). Evolution of two-component signal transduction. Mol Biol Evol *17*, 1956-1970.

Krell, T., Lacal, J., Busch, A., Silva-Jimenez, H., Guazzaroni, M.E., and Ramos, J.L. (2010). Bacterial sensor kinases: diversity in the recognition of environmental signals. Annual Review of Microbiology *64*, 539-559.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.

Kuo, C.H., and Ochman, H. (2010). The extinction dynamics of bacterial pseudogenes. PLoS Genet *6*.

Kurland, C.G., Canback, B., and Berg, O.G. (2003). Horizontal gene transfer: a critical view. Proc Natl Acad Sci U S A *100*, 9658-9662.

Laub, M.T. (2011). The role of two-component signal transduction systems in bacterial stress responses. In Bacterial Stress Responses, G. Storz, Hengge, R., ed. (Washington, D.C.: ASM Press).

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annu Rev Genet *41*, 121-145.

Lee, A.K., Detweiler, C.S., and Falkow, S. (2000). OmpR regulates the two-component system SsrA-ssrB in Salmonella pathogenicity island 2. J Bacteriol *182*, 771-781.

Lieberman, T.D., Michel, J.B., Aingaran, M., Potter-Bynoe, G., Roux, D., Davis, M.R., Skurnik, D., Leiby, N., LiPuma, J.J., Goldberg, J.B., *et al.* (2011). Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. Nat Genet *43*, 1275-1280.

Liu, Y., Harrison, P.M., Kunin, V., and Gerstein, M. (2004). Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol *5*, R64.

Lopez-Redondo, M.L., Moronta, F., Salinas, P., Espinosa, J., Cantos, R., Dixon, R., Marina, A., and Contreras, A. (2010). Environmental control of phosphorylation pathways in a branched two-component system. Mol Microbiol *78*, 475-489.

Madan Babu, M., and Teichmann, S.A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. Nucleic Acids Res *31*, 1234-1244.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A *99*, 12246-12251.

Miller, S.I., Kukral, A.M., and Mekalanos, J.J. (1989). A two-component regulatory system (phoP phoQ) controls Salmonella typhimurium virulence. Proc Natl Acad Sci U S A *86*, 5054-5058.

Mira, A., Ochman, H., and Moran, N.A. (2001). Deletional bias and the evolution of bacterial genomes. Trends Genet *17*, 589-596.

Moglich, A., Ayers, R.A., and Moffat, K. (2009a). Design and signaling mechanism of light-regulated histidine kinases. J Mol Biol *385*, 1433-1444.

Moglich, A., Ayers, R.A., and Moffat, K. (2009b). Structure and signaling mechanism of Per-ARNT-Sim domains. Structure *17*, 1282-1294.

Moglich, A., Ayers, R.A., and Moffat, K. (2010). Addition at the molecular level: signal integration in designed Per-ARNT-Sim receptor proteins. J Mol Biol *400*, 477-486.

Noriega, C.E., Lin, H.Y., Chen, L.L., Williams, S.B., and Stewart, V. (2010). Asymmetric cross-regulation between the nitrate-responsive NarX-NarL and NarQ-NarP two-component regulatory systems from Escherichia coli K-12. Mol Microbiol *75*, 394-412.

Ohashi, K., Yamashino, T., and Mizuno, T. (2005). Molecular basis for promoter selectivity of the transcriptional activator OmpR of Escherichia coli: isolation of mutants that can activate the non-cognate kdpABC promoter. J Biochem *137*, 51-59.

Osborne, S.E., Walthers, D., Tomljenovic, A.M., Mulder, D.T., Silphaduang, U., Duong, N., Lowden, M.J., Wickham, M.E., Waller, R.F., Kenney, L.J., *et al.* (2009). Pathogenic adaptation of intracellular bacteria by rewiring a cis-regulatory input function. Proc Natl Acad Sci U S A *106*, 3982-3987.

Parkinson, J.S. (2010). Signaling mechanisms of HAMP domains in chemoreceptors and sensor kinases. Annu Rev Microbiol *64*, 101-122.

Perez, J.C., and Groisman, E.A. (2009a). Evolution of transcriptional regulatory circuits in bacteria. Cell *138*, 233-244.

Perez, J.C., and Groisman, E.A. (2009b). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc Natl Acad Sci U S A *106*, 4319-4324.

Perez, J.C., Shin, D., Zwir, I., Latifi, T., Hadley, T.J., and Groisman, E.A. (2009). Evolution of a bacterial regulon controlling virulence and Mg(2+) homeostasis. PLoS Genet *5*, e1000428.

Posas, F., Wurgler-Murphy, S.M., Maeda, T., Witten, E.A., Thai, T.C., and Saito, H. (1996). Yeast HOG1 MAP kinase cascade is regulated by a multistep phosphorelay mechanism in the SLN1-YPD1-SSK1 "two-component" osmosensor. Cell *86*, 865-875.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2008). Horizontal gene transfer and the evolution of transcriptional regulation in Escherichia coli. Genome Biol *9*, R4.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Qian, W., Han, Z.J., and He, C. (2008). Two-component signal transduction systems of Xanthomonas spp.: a lesson from genomics. Mol Plant Microbe Interact *21*, 151-161.

Rabin, R.S., and Stewart, V. (1993). Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in Escherichia coli K-12. J Bacteriol *175*, 3259-3268.

Raivio, T.L., and Silhavy, T.J. (1997). Transduction of envelope stress in Escherichia coli by the Cpx two-component system. J Bacteriol *179*, 7724-7733.

Rajeev, L., Luning, E.G., Dehal, P.S., Price, M.N., Arkin, A.P., and Mukhopadhyay, A. (2011). Systematic mapping of two component response regulators to gene targets in a model sulfate reducing bacterium. Genome Biol *12*, R99.

Rajewsky, N., Socci, N.D., Zapotocky, M., and Siggia, E.D. (2002). The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res *12*, 298-308.

Rampersaud, A., Utsumi, R., Delgado, J., Forst, S.A., and Inouye, M. (1991). Ca2(+)-enhanced phosphorylation of a chimeric protein kinase involved with bacterial signal transduction. J Biol Chem *266*, 7633-7637.

Ren, B., Liang, Y., Deng, Y., Chen, Q., Zhang, J., Yang, X., and Zuo, J. (2009). Genome-wide comparative analysis of type-A Arabidopsis response regulator genes by overexpression studies reveals their diverse roles and regulatory mechanisms in cytokinin signaling. Cell Res *19*, 1178-1190.

Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J.C., Cattolico, L., *et al.* (2002). Genome sequence of the plant pathogen Ralstonia solanacearum. Nature *415*, 497-502.

Schaller, G.E., Shiu, S.H., and Armitage, J.P. (2011). Two-component systems and their co-option for eukaryotic signal transduction. Curr Biol *21*, R320-330.

Siryaporn, A., Perchuk, B.S., Laub, M.T., and Goulian, M. (2010). Evolving a robust signal transduction pathway from weak cross-talk. Mol Syst Biol *6*, 452.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Sonnenburg, E.D., Sonnenburg, J.L., Manchester, J.K., Hansen, E.E., Chiang, H.C., and Gordon, J.I. (2006). A hybrid two-component system protein of a prominent human gut symbiont couples glycan sensing in vivo to carbohydrate metabolism. Proc Natl Acad Sci U S A *103*, 8834-8839.

Stephenson, K., and Hoch, J.A. (2002). Evolution of signalling in the sporulation phosphorelay. Mol Microbiol *46*, 297-304.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

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

Ulrich, D.L., Kojetin, D., Bassler, B.L., Cavanagh, J., and Loria, J.P. (2005). Solution structure and dynamics of LuxU from Vibrio harveyi, a phosphotransferase protein involved in bacterial quorum sensing. J Mol Biol *347*, 297-307.

Ulrich, L.E., and Zhulin, I.B. (2010). The MiST2 database: a comprehensive genomics resource on microbial signal transduction. Nucleic Acids Res *38*, D401-407.

Utsumi, R., Brissette, R.E., Rampersaud, A., Forst, S.A., Oosawa, K., and Inouye, M. (1989). Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. Science *245*, 1246-1249.

Varughese, K.I., Madhusudan, Zhou, X.Z., Whiteley, J.M., and Hoch, J.A. (1998). Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase. Mol Cell *2*, 485-493.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A *106*, 67-72.

Whitworth, D.E., and Cock, P.J. (2009). Evolution of prokaryotic two-component systems: insights from comparative genomics. Amino Acids *37*, 459-466.

Williams, S.B., and Stewart, V. (1997). Discrimination between structurally related ligands nitrate and nitrite controls autokinase activity of the NarX transmembrane signal transducer of Escherichia coli K-12. Mol Microbiol *26*, 911-925.

Worley, M.J., Ching, K.H., and Heffron, F. (2000). Salmonella SsrB activates a global regulon of horizontally acquired genes. Mol Microbiol *36*, 749-761.

Wright, G.D., Holman, T.R., and Walsh, C.T. (1993). Purification and characterization of VanR and the cytosolic domain of VanS: a two-component regulatory system required for vancomycin resistance in Enterococcus faecium BM4147. Biochemistry *32*, 5057-5063.

Wuichet, K., and Zhulin, I.B. (2010). Origins and diversification of a complex signal transduction system in prokaryotes. Sci Signal *3*, ra50.

Xu, Q., Carlton, D., Miller, M.D., Elsliger, M.A., Krishna, S.S., Abdubek, P., Astakhova, T., Burra, P., Chiu, H.J., Clayton, T., *et al.* (2009). Crystal structure of histidine phosphotransfer protein ShpA, an essential regulator of stalk biogenesis in Caulobacter crescentus. J Mol Biol *390*, 686-698.

Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R., and Ishihama, A. (2005). Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. J Biol Chem *280*, 1448-1456.

Yang, Y., and Inouye, M. (1991). Intermolecular complementation between two defective mutant signal-transducing receptors of Escherichia coli. Proc Natl Acad Sci U S A *88*, 11057-11061.

Yang, Y., and Inouye, M. (1993). Requirement of both kinase and phosphatase activities of an Escherichia coli receptor (Taz1) for ligand-dependent signal transduction. J Mol Biol *231*, 335-342.

Zapf, J., Sen, U., Madhusudan, Hoch, J.A., and Varughese, K.I. (2000). A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. Structure *8*, 851-862.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

Zhang, W., and Shi, L. (2005). Distribution and evolution of multiple-step phosphorelay in prokaryotes: lateral domain recruitment involved in the formation of hybrid-type histidine kinases. Microbiology *151*, 2159-2173.

Zhang, Z., and Hendrickson, W.A. (2010). Structural characterization of the predominant family of histidine kinase sensor domains. J Mol Biol *400*, 335-353.

Zhu, Y., and Inouye, M. (2003). Analysis of the role of the EnvZ linker region in signal transduction using a chimeric Tar/EnvZ receptor protein, Tez1. J Biol Chem *278*, 22812-22819.

Zhulin, I.B., Nikolskaya, A.N., and Galperin, M.Y. (2003). Common extracellular sensory domains in transmembrane receptors for diverse signal transduction pathways in bacteria and archaea. J Bacteriol *185*, 285-294.

Chapter 2

Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways

This work was published as Emily J. Capra*, Barrett S. Perchuk*, Emma A. Lubin, Orr Ashenberg, Jeffrey M. Skerker, and Michael T. Laub. 2010. PLoS Genet. Nov 24;6(11): e1001220

EJC, BSP, JMS, and MTL conceived and designed the experiments. EJC and BSP performed the experiments. EAL contributed reagents. OA performed the computational analysis. EJC and MTL wrote the paper. EJC and BSP contributed equally to the work.

Abstract

Two-component signal transduction systems enable bacteria to sense and respond to a wide range of environmental stimuli. Sensor histidine kinases transmit signals to their cognate response regulators via phosphorylation. The faithful transmission of information through two-component pathways and the avoidance of unwanted cross-talk requires exquisite specificity of histidine kinase-response regulator interactions to ensure that cells mount the appropriate response to external signals. To identify putative specificity-determining residues we have analyzed amino acid coevolution in two-component proteins and identified a set of residues that can be used to rationally rewire a model signaling pathway, EnvZ-OmpR. To explore how a relatively small set of residues can dictate partner selectivity, we combined alanine-scanning mutagenesis with an approach we call trajectory-scanning mutagenesis, in which all mutational intermediates between the specificity residues of EnvZ and another kinase, RstB, were systematically examined for phosphotransfer specificity. The same approach was used for the response regulators OmpR and RstA. Collectively, the results begin to reveal the molecular mechanism by which a small set of amino acids enables an individual kinase to discriminate amongst a large set of highly-related response regulators and *vice versa*. Our results also suggest that the mutational trajectories taken by two-component signaling proteins following gene or pathway duplication may be constrained and subject to differential selective pressures. Only some trajectories allow both the maintenance of phosphotransfer and the avoidance of unwanted cross-talk.

Author Summary

Maintaining the specificity of signal transduction pathways is critical to the ability of cells to process information, make decisions, and regulate their behavior. Preventing cross-talk often relies predominantly on molecular recognition and a set of specificity-determining residues in cognate proteins. Identifying these residues and understanding how they dictate specificity is still a major challenge. Additionally, we have a rudimentary understanding of how specificity evolves, particularly after gene duplication events. We tackled these questions using two-component signaling proteins, the largest family of bacterial signaling proteins. Using analyses of amino acid coevolution, we pinpointed a set of specificity residues in histidine kinases and their cognate substrates. Then, using systematic mutagenesis we characterized the complete set of intermediates between two different signaling systems, EnvZ/OmpR and RstA/RstB. The results demonstrate that specificity residues contribute unequally and, importantly, that some residues depend substantially on the identity of neighboring residues. We also demonstrate how the specificity of EnvZ/OmpR can be reprogrammed to match that of RstB/RstA through a series of individual substitutions without disrupting the kinase/regulator interaction. Notably, this property is not shared by all trajectories from EnvZ/OmpR to RstA/RstB, suggesting that the duplication/divergence process that likely produced these two pathways may have been fundamentally constrained.

Introduction

Protein-protein interactions are crucial to virtually every cellular process. Within the crowded confines of the cell, proteins must distinguish between their cognate partners and non-cognate partners, in order to avoid unproductive and potentially deleterious interactions. The problem of interaction specificity is particularly acute for paralogous protein families where proteins with diverse cellular functions share significant structural and sequence similarity. Cells have evolved many mechanisms to cope with potential cross-talk and to ensure the specificity of protein-protein interactions (Schwartz and Madhani, 2004; Ubersax and Ferrell, 2007). In multicellular organisms, spatial mechanisms that prevent related, but distinct, proteins from coming in contact with one another are often used to create specificity. For example, scaffold proteins, the localization of proteins to different subcellular compartments, and tissue-specific expression can all insulate distinct pathways. Temporal mechanisms, such as the differential timing of expression, are also used to insulate pathways. Although cells employ each of these strategies, in many cases the primary means of preventing unwanted interactions is molecular recognition. However, our understanding of precisely how proteins discriminate between cognate and non-cognate partners at the molecular level is surprisingly rudimentary. Identifying the amino acids responsible, elucidating the precise roles played by each residue, and understanding their complex interdependencies remain major challenges for most protein-protein interactions.

Two component signal transduction pathways provide a tractable system for addressing these questions. These signaling pathways, which are the dominant form of signaling in bacteria, typically consist of a sensor histidine kinase (HK) and a cognate response regulator (RR) (Stock et al., 2000). Upon activation of the pathway, a histidine kinase dimer will autophosphorylate on a conserved histidine that then serves as the phosphodonor for a cognate response regulator. Phosphorylation of the response

regulator typically activates an output domain which can effect changes in cellular physiology, often by modulating gene expression (Gao et al., 2007). Many histidine kinases are bifunctional and when not active for autophosphorylation, will drive the dephosphorylation of their cognate response regulators.

Two-component signaling systems are used for sensing and adapting to a wide range of environmental and intracellular stimuli (Stock et al., 2000) and most bacterial species encode dozens, if not hundreds of kinase-regulator pairs. Most histidine kinases have only one or two cognate response regulators, and there is minimal cross-talk between different pathways at the level of phosphotransfer (Laub and Goulian, 2007; Skerker et al., 2005). The specificity of phosphotransfer is dictated, on a system-wide level, at the level of molecular recognition (Skerker et al., 2005). That is, histidine kinases exhibit a large kinetic preference *in vitro* for their *in vivo* cognate regulator(s) relative to all other response regulators (Fisher et al., 1996; Grimshaw et al., 1998; Skerker et al., 2005). Hence, cellular context is not essential and the basis of *in vivo* phosphotransfer specificity can be dissected *in vitro*.

To identify the amino acids that govern the specificity of phosphotransfer in two-component pathways, several groups have examined patterns of amino acid coevolution in cognate pairs of histidine kinases and response regulators (Burger and van Nimwegen, 2008; Skerker et al., 2008; Weigt et al., 2009; White et al., 2007). The rationale behind this approach is that if a residue critical to molecular recognition mutates, it must either revert or be compensated for by a mutation in the cognate protein. Many of the residues identified in these computational approaches are at the molecular interface formed in a co-crystal structure of a histidine kinase-response regulator complex (Casino et al., 2009). However, residues in direct contact do not necessarily dictate specificity (Skerker et al., 2008) and computational approaches alone cannot reveal how a histidine kinase discriminates between cognate and non-cognate substrates.

Using the *E. coli* histidine kinase EnvZ as a model, we mapped a subset of coevolving residues that are critical to the specificity of phosphotransfer (Skerker et al., 2008). Mutating as few as three residues within the DHp (Dimerization and Histidine phosphotransfer) domain of EnvZ was sufficient to reprogram its phosphotransfer specificity from OmpR to the non-cognate substrate RstA. Although a set of residues that could switch the phosphotransfer specificity of EnvZ was identified, several fundamental questions remain unanswered. Can phosphotransfer specificity also be rewired by making mutations in a response regulator? Do individual specificity residues function as positive elements to promote cognate interactions, as negative elements to prevent non-cognate interactions, or both? Do individual residues contribute equally and independently or are there "hot spots" and dependencies at the amino acid level?

Here, we couple analysis of amino acid coevolution with alanine-scanning mutagenesis and an approach we call trajectory-scanning mutagenesis to systematically dissect the basis of phosphotransfer specificity in two-component signaling pathways. The results provide new insights into how histidine kinases use a set of amino acids to "choose" their cognate substrates, and *vice versa*. The results have important implications for understanding the evolution of two-component signaling pathways and the mechanisms that cells can use to insulate pathways following gene duplication.

Results

Identification of coevolving residues in cognate kinase-regulator pairs

To identify the amino acids responsible for determining the specificity of phosphotransfer in two-component signaling pathways, we searched for residues that covary in cognate HK-RR pairs. Histidine kinases and response regulators that are encoded in the same operon typically form exclusive one-to-one pairings, exhibiting a highly specific interaction both *in vivo* and *in vitro*. We identified ~4500 operonic pairs of histidine kinases and response regulators from a phylogenetically diverse set of 400 sequenced bacterial genomes. To identify coevolving residues, we concatenated cognate HK-RR pairs, performed a large multiple sequence alignment, and then measured mutual information between columns of the sequence alignment. We noted that some columns tended to have high mutual information scores with many other columns in the alignment, an observation also made in other analyses of mutual information (Gloor et al., 2005). For example, positions 8 and 270 have relatively broad score distributions with long tails, while positions 18 and 202 have narrower distributions centered closer to the origin (Figure 2.1A-B). Consequently, the pairs 8-270 and 18-202, which possess identical mutual information scores of 0.35, cannot be treated identically. We used a relatively simple correction in which raw MI scores were normalized by each column's average raw MI score with all 310 positions in the sequence alignment (Figure 2.1C).

At an adjusted score threshold of 3.5, we found 12 coevolving pairs, comprising 9 residues in the histidine kinases and 7 in the response regulators (Figure 2.2A-C). These residues form a single, densely-interconnected cluster of coevolving residues. The

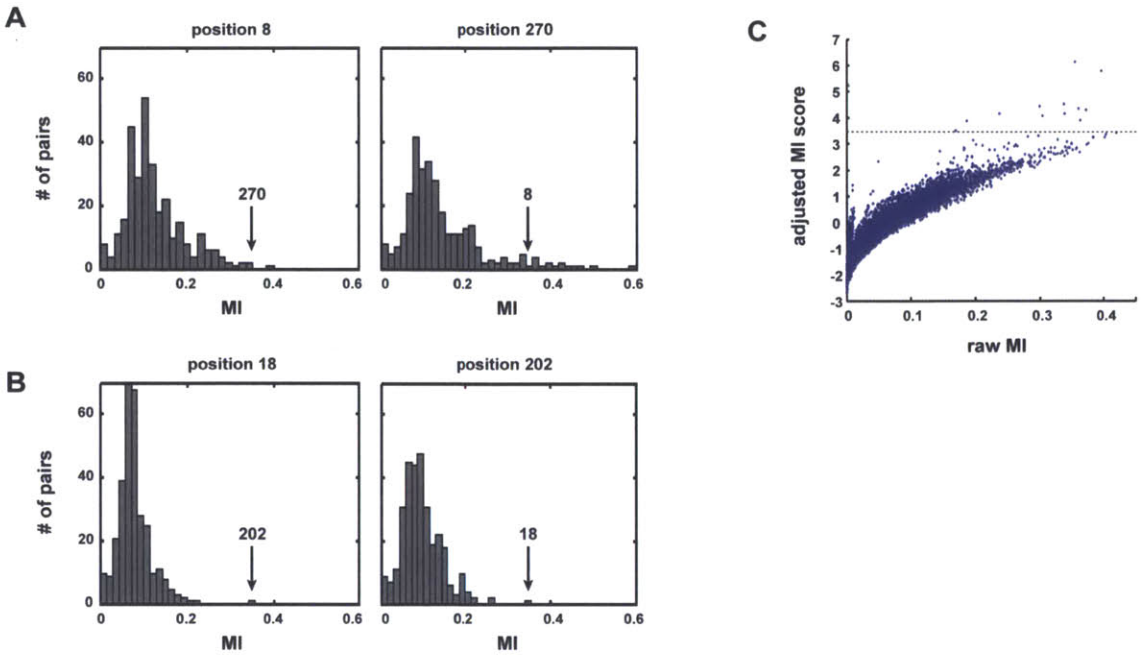


Figure 2.1 Adjusted mutual information analysis of amino acid covariation in two-component proteins.

(A) Histograms summarizing the raw mutual information scores for columns 8 and 270 in the kinase-regulator alignment. The arrow indicates the location of the score for the column pair 8-270. (B) Same as panel A, but for positions 18 and 202 in the alignment. (C) Scatterplot of raw mutual information scores against adjusted mutual information scores, as described in the main text and in Materials and Methods. Dashed line indicates the score cutoff of 3.5 used in Figure 2.2.

Residues are all solvent-exposed in the individual molecules, but buried within the molecular interface formed in a co-crystal structure of *T. maritima* HK853 and RR468 (Figure 2.2D) (Casino et al., 2009). The residues identified here overlap substantially with, but are not identical to, those we identified previously (Skerker et al., 2008). Of the coevolving residues in the kinase, all are in the DHp domain, consistent with this domain being the primary site of interaction with the response regulator. Within the DHp domain, the coevolving residues are found on both alpha helices and are located below the histidine phosphorylation site (Figure 2.2D). The covarying residues in the response regulator are spatially near the conserved aspartic acid phosphorylation site (Figure

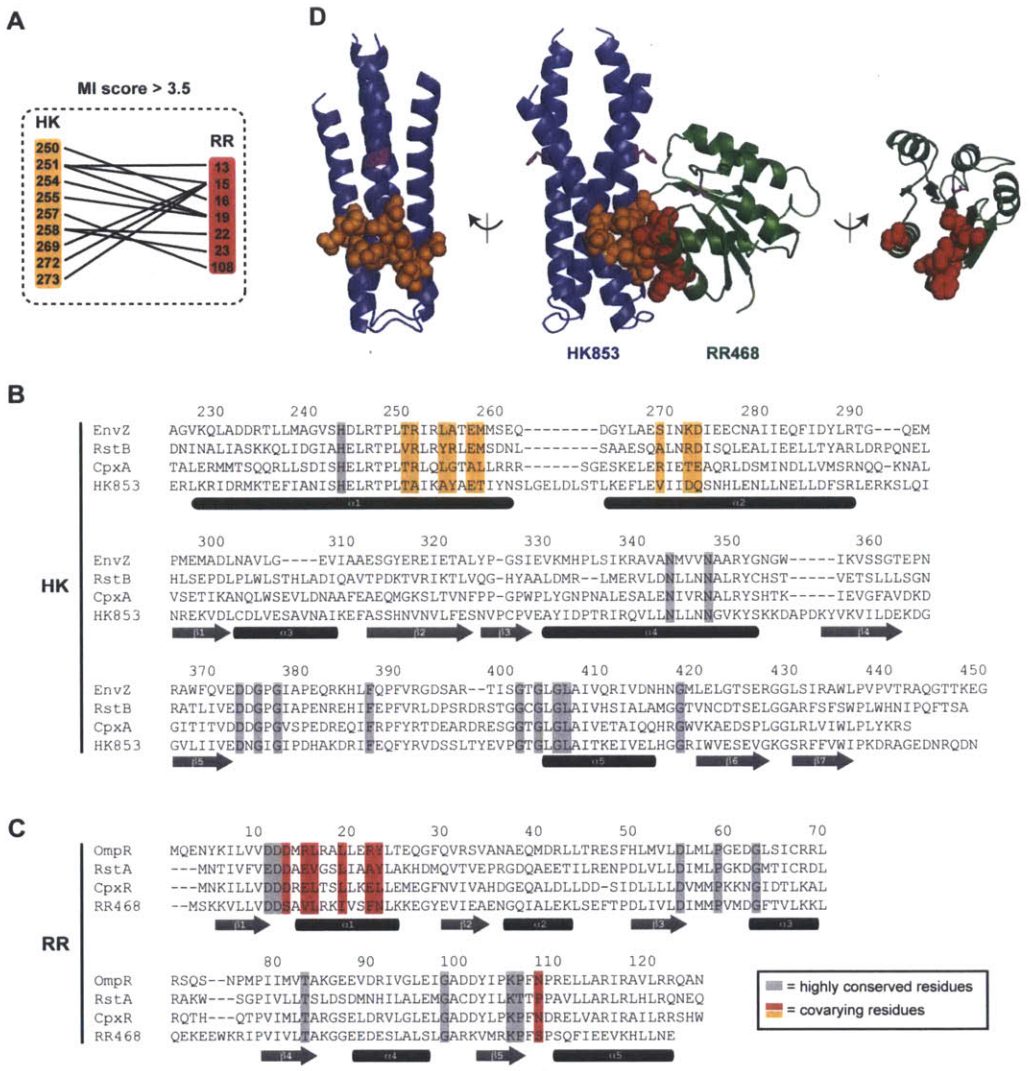


Figure 2.2 Identification of coevolving amino acids in cognate pairs of histidine kinases and response regulators.

(A) Residues in histidine kinases and response regulators that strongly coevolve (adjusted MI score > 3.5) are listed with lines connecting covarying pairs. Residues are numbered according to their position in *E. coli* EnvZ and OmpR. (B-C) Residues in histidine kinases that coevolve with residues in response regulators are shown on a primary sequence alignment of HK853 from *T. maritima* and EnvZ, RstB, and CpxA from *E. coli*. Residues in response regulators that strongly coevolve with residues in histidine kinases are shown on a primary sequence alignment of RR468 from *T. maritima* and OmpR, RstA, and CpxR from *E. coli*. Residues highly conserved across all two-component signaling proteins are shaded in grey. Coevolving residues are shown in orange and red for the kinase and regulator, respectively. Secondary structure elements, based on the co-crystal structure of HK853 and RR468 from *T. maritima* (Casino et al., 2009), are shown beneath the sequences. (D) Coevolving residues mapped onto the HK853-RR468 structure. Coevolving residues are shown by space-filling and colored as in panels A-C. The side chains of the conserved phosphorylatable histidines and aspartate are shown as magenta sticks. The HK853-RR468 complex is shown in the center with each individual molecule rotated 90° and shown separately.

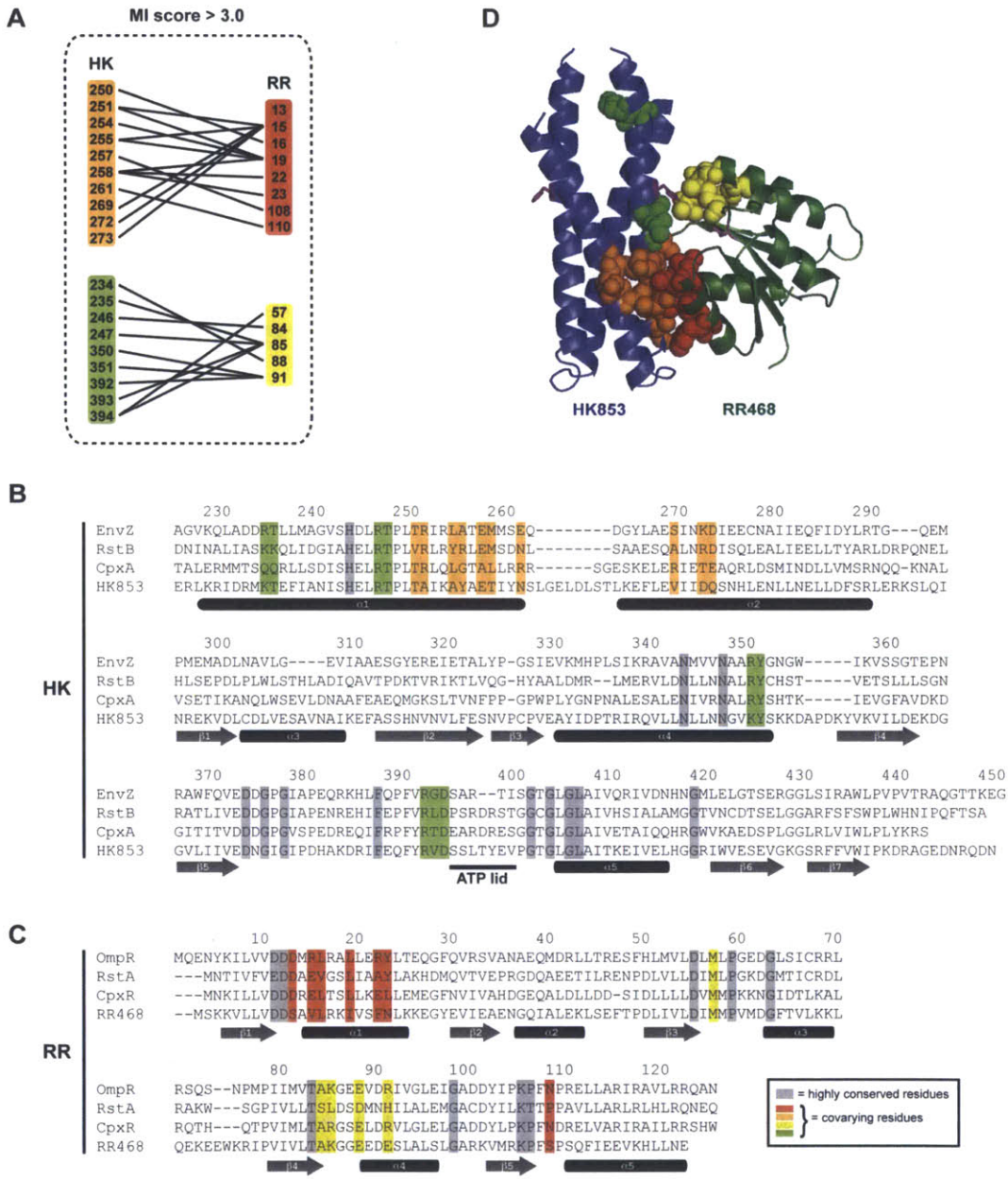


Figure 2.3 Identification of coevolving amino acids in cognate pairs of histidine kinases and response regulators.

Same as Figure 2.2, except at a score threshold of 3.0. (A) residues in histidine kinases and response regulators that strongly coevolve (adjusted MI score > 3.0) are listed with lines connecting covarying pairs. Residues are numbers according to their position in *E. coli* EnvZ and OmpR and colored as in panels B-D. (B-C) Residues in histidine kinases that coevolve with

residues in response regulators are shown on a primary sequence alignment of HK853 from *T. maritima* and EnvZ, RstB, and CpxA from *E. coli*. Residues highly conserved across all two-component signaling proteins are shaded in grey. Coevolving residues above and below the phosphorylation site in the kinase are shown in green and orange, respectively. These two sets of residues coevolve with residues in the response regulator shaded in yellow and red, respectively. Secondary structure elements, based on the co-crystal structure of HK853 and RR468 from *T. maritima* (Casino et al., 2009), are shown beneath the sequences. (D) Coevolving residues mapped onto the HK853-RR458 structure. Coevolving residues are shown by space-filling and colored as in panels A-C. The side chains of the conserved phosphorylatable histidines and aspartate are shown as magenta sticks.

2.2D), predominantly on a single face of alpha helix-1 in the receiver domain with one additional residue within the β5-α5 loop. At lower score thresholds, an additional cluster of coevolving residues are found (Figure 2.3), but we focus here on the set of 16 residues identified at a threshold of 3.5.

Rewiring response regulator specificity

Our previous studies demonstrated that many of the coevolving residues in the kinase (Figure 2.2) are critical to the phosphotransfer specificity of EnvZ and when mutated can reprogram its substrate selectivity (Skerker et al., 2008). To test whether we could also rewire the specificity of a response regulator, we again coupled our analyses of coevolution with site-directed mutagenesis. We aimed to mutate the response regulator OmpR such that it was no longer phosphorylated by its cognate kinase EnvZ and instead was phosphorylated by the non-cognate kinase CpxA or RstB. Each kinase was autophosphorylated, purified away from unincorporated nucleotide, and tested for phosphotransfer. In our reaction conditions at a 1 minute time point, EnvZ phosphotransfers exclusively to OmpR, whereas CpxA and RstB phosphotransfer exclusively to CpxR and RstA, respectively (Figure 2.4).

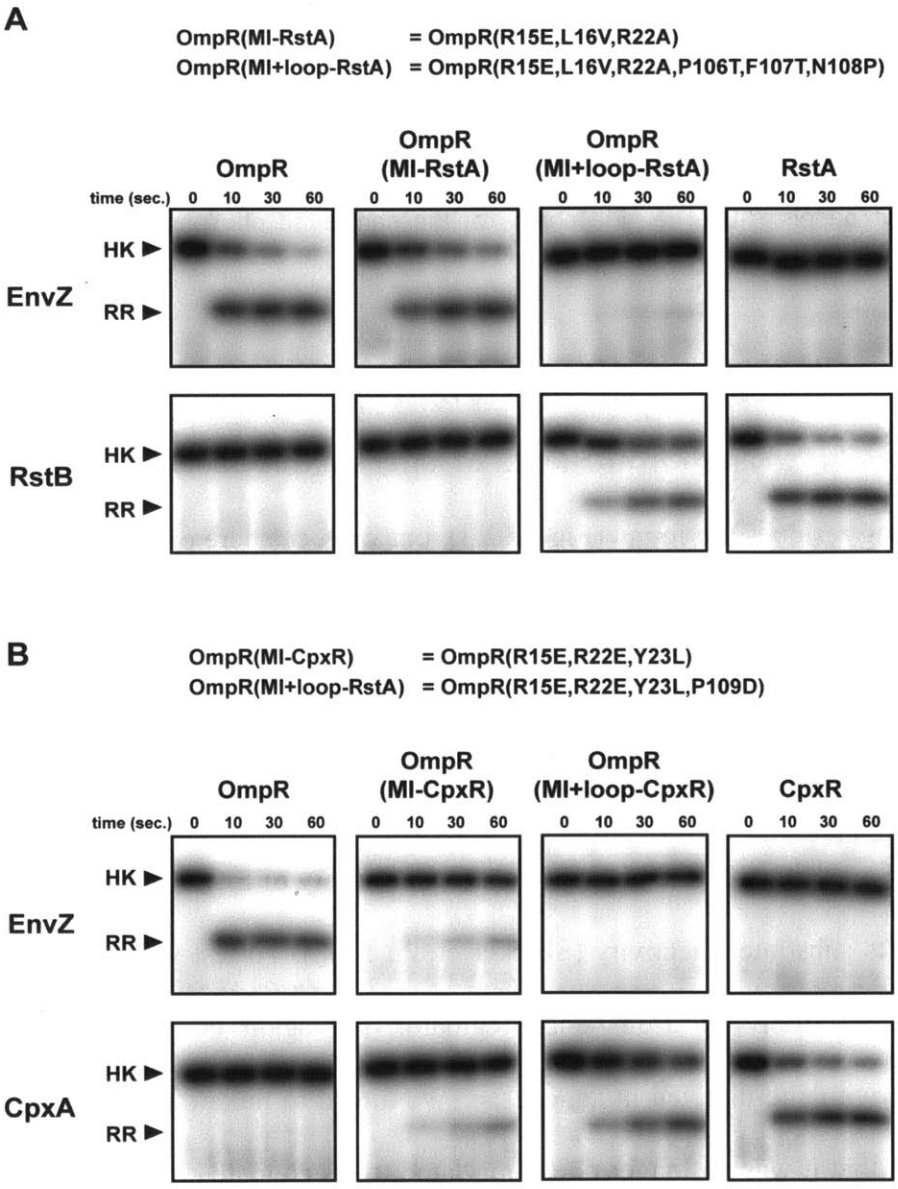


Figure 2.4 Rewiring the specificity of response regulators.

(A) The histidine kinases EnvZ and RstB were autophosphorylated and examined for phosphotransfer to the response regulators indicated. The mutations in OmpR(MI-RstA) and OmpR(MI+loop-RstA) are listed at the top. (B) The histidine kinases EnvZ and CpxA were autophosphorylated and examined for phosphotransfer to the response regulators indicated. The mutations in OmpR(MI-CpxR) and OmpR(MI+loop-CpxR) are listed at the top. Each gel image shows phosphotransfer after 0, 10, 30, and 60 seconds. Bands corresponding to autophosphorylated kinases are labeled on the left. If phosphotransfer occurred, bands corresponding to the phosphorylated regulator appear below the kinase band.

We first substituted residues in OmpR at the positions within alpha helix-1 identified by mutual information analysis with the corresponding residues from CpxR and RstA to create OmpR(MI-CpxR) and OmpR(MI-RstA); in each case three amino acid substitutions were made in OmpR. The mutant OmpR(MI-RstA) was not phosphorylated to a significant extent by RstB and was still a robust target of EnvZ (Figure 2.4A). The mutant OmpR(MI-CpxR) showed diminished phosphotransfer from EnvZ and was now phosphorylated by CpxA, although less efficiently than wild type CpxR (Figure 2.4B). The residues in alpha helix-1 are thus important for phosphotransfer specificity, but other residues must contribute. We hypothesized that residues within the β5-α5 loop may also affect specificity of the regulator. One of these residues covaried strongly with residues in the histidine kinase (Figure 2.2) and other loop residues covaried at a slightly lower score threshold of 2.8. We thus swapped the residues in the OmpR loop with those from CpxR and RstA to create OmpR(MI+loop-RstA) and OmpR(MI+loop-CpxR), respectively, and examined phosphotransfer to each of these constructs; the former required three amino acid substitutions and the latter just one. Both constructs exhibited a nearly complete switch in phosphotransfer specificity. EnvZ was unable to phosphotransfer to either OmpR(MI+loop-RstA) or OmpR(MI+loop-CpxR), whereas phosphotransfer from RstB or CpxA to the respective rewired OmpR mutants was efficient and at near wild-type rates (Figure 2.4). Thus, the top coevolving residues appear sufficient, when mutated along with the β5-α5 loop, to rewire the phosphotransfer specificity of OmpR.

We note that the residues mutated to change the specificity of OmpR constitute a subset of the molecular interface formed by a cognate kinase and regulator (Figure 2.2D). For instance, the residues in the β4-α4 loop of the response regulator contact the histidine

kinase, are in close proximity to the top coevolving residues, and coevolve with sites in the kinase at lower score thresholds (Figure 2.3), but mutating them was not required to change phosphotransfer specificity (Figure 2.4). We conclude that the strongest coevolving residues are necessary and sufficient to change the phosphotransfer partnering specificity of OmpR. Other residues may fine-tune the interaction, but do not make major contributions.

Alanine-scanning mutagenesis and the role of individual residues

Our results indicate that kinase-substrate interaction specificity in two-component pathways is determined by a relatively small set of residues. But does each residue contribute equally to specificity or are there "hotspots" that contribute disproportionately? Do individual residues help bind the cognate substrate or help prevent interaction with non-cognate substrates? To address these questions, we performed alanine-scanning mutagenesis on the DHp domain of EnvZ. Surprisingly, despite being one of the best-characterized histidine kinases, EnvZ has never been explored through alanine-scanning mutagenesis. One study described a series of cysteine mutants (Qin et al., 2003), but the set of residues examined was limited and the interpretation of cysteine mutations can be ambiguous. We created a series of 33 EnvZ mutants to probe the role of most of the solvent-exposed residues in the DHp domain, generating alanine mutations for all residues except for A255, which was substituted with a threonine (Figure 2.5A).

We first examined the autophosphorylation activity of each EnvZ mutant (Figure 2.5B, 2.6A). As expected, mutating the conserved phosphorylation site H243 (data not shown),

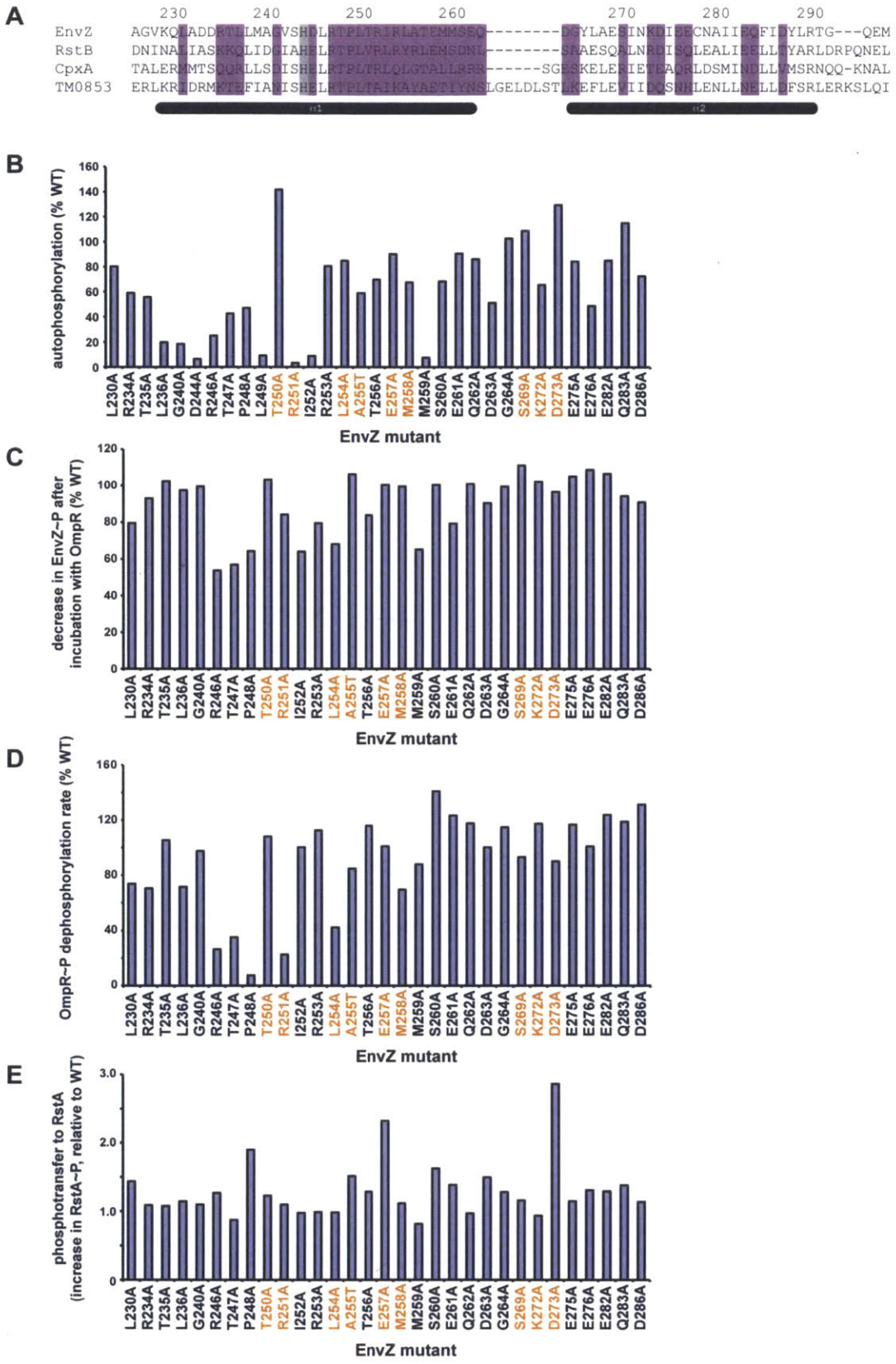

A
230
240
250
260
270
280
290
EnvZ AGVKQLADDATLLMAGVSHDLRTPLTRIRLATEMMSEQ--------DGYLAESINKDIEECNAIIEQFIDYLRTG---QEM
RstB DNINALIASKKQLIDQIAHELRTPLVRLRYRLEMSDNL--------SAAESQALNRDISQLEALIEELLTYARLDRPQNEL
CpxA TALERAMTSQQRLLSTISHELRTPLTRLQLGTALLRSR------SGESKELERIETEAQRLDSMINDLLVMSRNQQ-KNAL
TM0853 ERLKRIDRMKTEFIAMISHELRTPLTAIKAVAETIYNSLGELDLSTLKEFLEVIIDQSMHLENLLNELLDFSRLERKSLQI
n1
n2
B
autophosphorylation (% WT)
EnvZ mutant
C
decrease in EnvZ~P after incubation with OmpR (% WT)
EnvZ mutant
D
OmpR~P dephosphorylation rate (% WT)
EnvZ mutant
E
phosphotransfer to RstA (increase in RstA~P, relative to WT)
EnvZ mutant
L230A R234A T235A L236A G240A D244A R246A T247A P248A L249A T250A R251A I252A R253A L254A A255T T256A E257A M258A M259A S260A E261A Q262A D263A G264A S269A K272A D273A E275A E276A E282A Q283A D286A

Figure 2.5 Alanine-scanning mutagenesis of EnvZ.

(A) Sequence of the DHp domain of EnvZ showing the residues substituted with alanine in purple. The conserved histidine phosphorylation site is shaded in grey. Numbering and secondary structure elements indicated as in Figure 2.2C. (B) Autophosphorylation levels of each EnvZ alanine mutant after a 1 minute incubation, expressed as a percentage of that measured for wild-type EnvZ. For gel images, see Figure 2.6A. (C) Decrease in EnvZ~P band after incubation with OmpR. Each value was expressed as a percentage of the decrease measured for wild-type EnvZ. Mutants that do not show a decrease in EnvZ~P could be defective either in phosphotransfer or in dephosphorylation of OmpR~P (see text for details). (D) Phosphatase activity of EnvZ alanine mutants. Each alanine mutant was tested for dephosphorylation of OmpR~P and the rate expressed as a percentage of that measured for wild-type EnvZ. (E) Phosphotransfer from EnvZ alanine mutants to RstA. Phosphotransfer was assessed by measuring the increase in labeled RstA after a 10 second incubation. For each mutant, the increase in RstA was normalized to the autophosphorylation level for that kinase and then reported as a fold-change relative to the phosphotransfer for wild-type EnvZ to RstA. In panels B-E, the specificity residues are listed in orange, as in Figure 2.2C. For panels C and E, the mutant kinases were autophosphorylated for 60 minutes prior to assessing phosphotransfer. Mutants D244A and L249A did not autophosphorylate significantly enough to examine phosphotransfer. For gel images for panels C-D, see Figure 2.6B. For panel D, the mutant kinases were tested for dephosphorylation of OmpR~P at 0.5, 1, and 2 minutes (Figure 2.7).

or the highly conserved aspartate that follows, D244, completely abolished autophosphorylation. Other residues strongly affecting autophosphorylation flank H243, including L236, G240, R246, T247, P248, L249, R251, and I252. Many of these residues are highly conserved among all histidine kinases suggesting they are critical for catalyzing phosphoryl transfer from ATP to histidine. Alternatively, they may impact folding or stability of the kinase; however, these residues are mostly solvent-exposed and none of the mutants significantly affected purification of soluble protein (data not shown). Of the top coevolving residues (Figure 2.2), only R251A showed substantially lower autophosphorylation than wild type, suggesting that residues required for docking to a response regulator are distinct from those required for docking to the kinase's CA (catalytic ATP-binding) domain.

For each EnvZ mutant that was able to autophosphorylate to reasonably high levels after an extended incubation, we tested phosphotransfer to OmpR, CpxR, and RstA (Figure 2.5

C-E, 2.6A). For an assessment of significance, see Figure 2.6C and Materials and Methods. For wild-type EnvZ, phosphotransfer to OmpR manifests as a decrease in the EnvZ~P band and a weak or absent OmpR~P band, resulting from high rates of phosphotransfer and subsequent dephosphorylation of OmpR~P by EnvZ. Several alanine mutants did not show the same decrease in EnvZ~P as the wild-type protein. However, for most of these mutants, such as R246A, T247A, and P248A, a more intense OmpR~P band was also seen, suggesting that phosphotransfer had occurred but that the mutant could no longer dephosphorylate OmpR~P. We confirmed the loss of phosphatase activity by measuring the dephosphorylation of purified OmpR~P by each EnvZ mutant (Figure 2.5D, 2.7). Only one mutant, I252A, showed a significant defect in phosphotransfer with no effect on phosphatase activity. Strikingly, mutating most of the coevolving specificity residues, including T250, R251, A255, E257, M258, S269, K272, and D273 had no major effect on phosphotransfer to OmpR. This finding suggests that there is no single "hot spot" and, instead, that specificity and molecular recognition are distributed over a number of residues. There may also be non-additive or synergistic effects between residues such that single point mutations do not significantly affect phosphotransfer in isolation, a possibility probed in more detail below.

Finally, we examined the EnvZ alanine mutants for phosphotransfer to the non-cognate regulators RstA and CpxR (Figure 2.5D, 2.6B). For these reactions, in contrast to those shown in Figure 2, EnvZ constructs were autophosphorylated and tested for phosphotransfer without purifying them away from ATP. Under these conditions, EnvZ phosphotransfers weakly to RstA, permitting us to assess whether the alanine mutations affected this non-cognate interaction. Most mutants phosphorylated RstA at a level

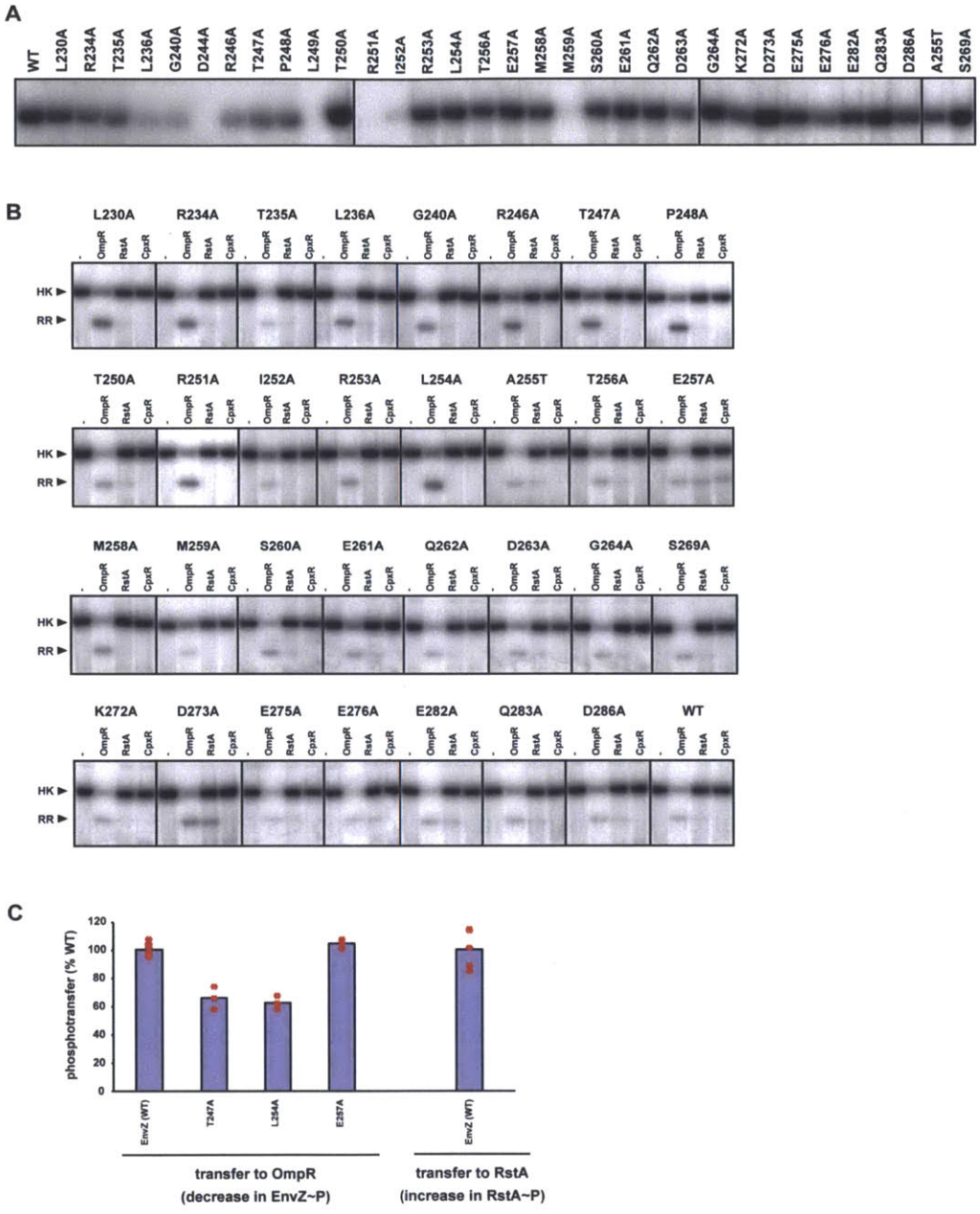


Figure 2.6 Alanine scanning mutagenesis of EnvZ.

(A) Each EnvZ mutant was autophosphorylated for 1 minute before reactions were stopped by the addition of loading buffer. Kinases were then examined by SDS-PAGE and phosphorimaging using four separate protein gels that were handled identically. Scanned images were concatenated; vertical bars separate lanes from different gels. For quantification see Figure 2.5B. (B) Each EnvZ mutant was autophosphorylated for 60 minutes and then examined for phosphotransfer to OmpR, RstA, and CpxR. Phosphotransfer was assessed by measuring the decrease in labeled EnvZ after a 10 second incubation with OmpR. For quantification, see Figure 2.5 C and E. (C) Reproducibility of phosphotransfer assays. Wild-type EnvZ was examined for phosphotransfer to OmpR and RstA six times while mutants T247A, L254A, and E257A were examined three times. The graph shows the mean and the individual values in red.

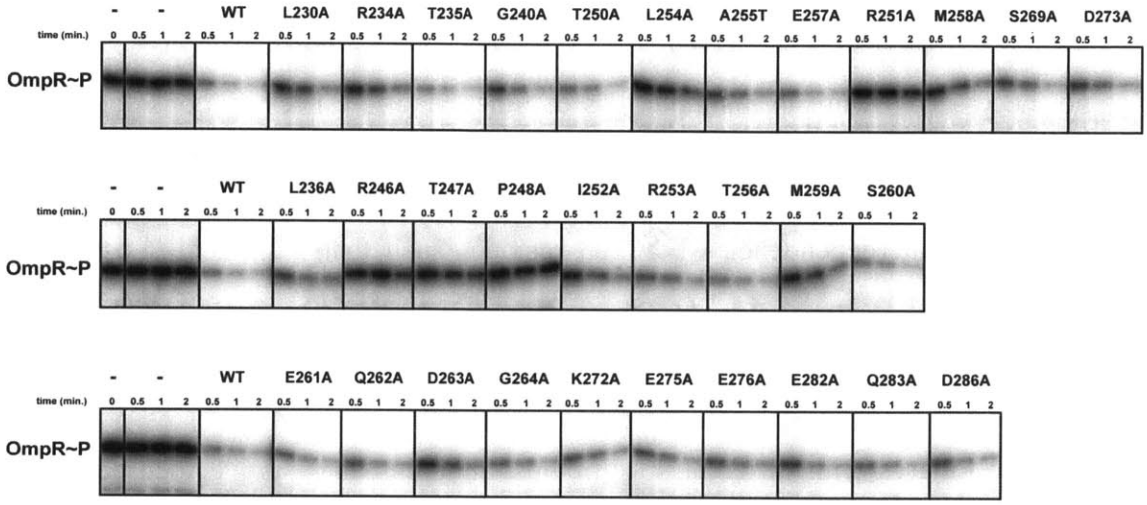


Figure 2.7 Dephosphorylation of OmpR~P by EnvZ alanine mutants.

Phosphorylated OmpR was purified and incubated with each EnvZ mutant for 0.5, 1, and 2 minutes. For a quantification of rates relative to wild-type EnvZ see Figure 2.5D.

equivalent to or less than the wild type EnvZ. However, four mutants, P248A, A255T, E257A, and D273A, each showed increases in RstA phosphorylation; E257A also showed detectable phosphorylation of CpxR. Notably, three of the four residues were identified as specificity residues (Figure 2.2) in our coevolution analysis. The increase in cross-talk seen with these mutants suggests that these residues function, at least in part, as negative elements that prevent phosphotransfer to non-cognate substrates without significantly affecting transfer to the cognate substrate.

Characterization of all intermediates along the mutational trajectories separating EnvZ and RstB

Although alanine-scanning provides some insight into specificity, an alanine substitution does not necessarily result in a simple loss of functionality, especially considering that EnvZ has a specificity residue that is already an alanine. In addition, as noted, there may be non-additive interdependencies between residues such that individual substitutions have minimal effect. We therefore sought to characterize the role of specificity-determining residues by examining the complete set of mutational intermediates between two histidine kinases with different specificities. For this analysis we focused on the paralogous systems EnvZ/OmpR and RstB/RstA, and term the approach trajectory-scanning. We constructed each possible specificity intermediate between EnvZ and RstB. This was feasible as the conversion of EnvZ phosphotransfer specificity to match that of RstB required only three substitutions, T250V, L254Y, and A255R (Skerker et al., 2008); the other major specificity residues identified by coevolution analysis are identical between EnvZ and RstB. In addition, we were able to rewire the specificity of RstB to match that of EnvZ by mutating the same three sites (Figure 2.8). The triple mutant RstB(V228T, Y232L, and R233A) no longer phosphorylated RstA and, instead, efficiently phosphorylated OmpR. These three residues thus play the dominant roles in dictating the specificity of both EnvZ and RstB. Other residues may make minor contributions.

We constructed each possible single and double mutant intermediate between EnvZ and RstB, in the context of each protein for a total of 12 mutants. To simplify nomenclature we have named mutants based on the protein mutated and the identity of the three

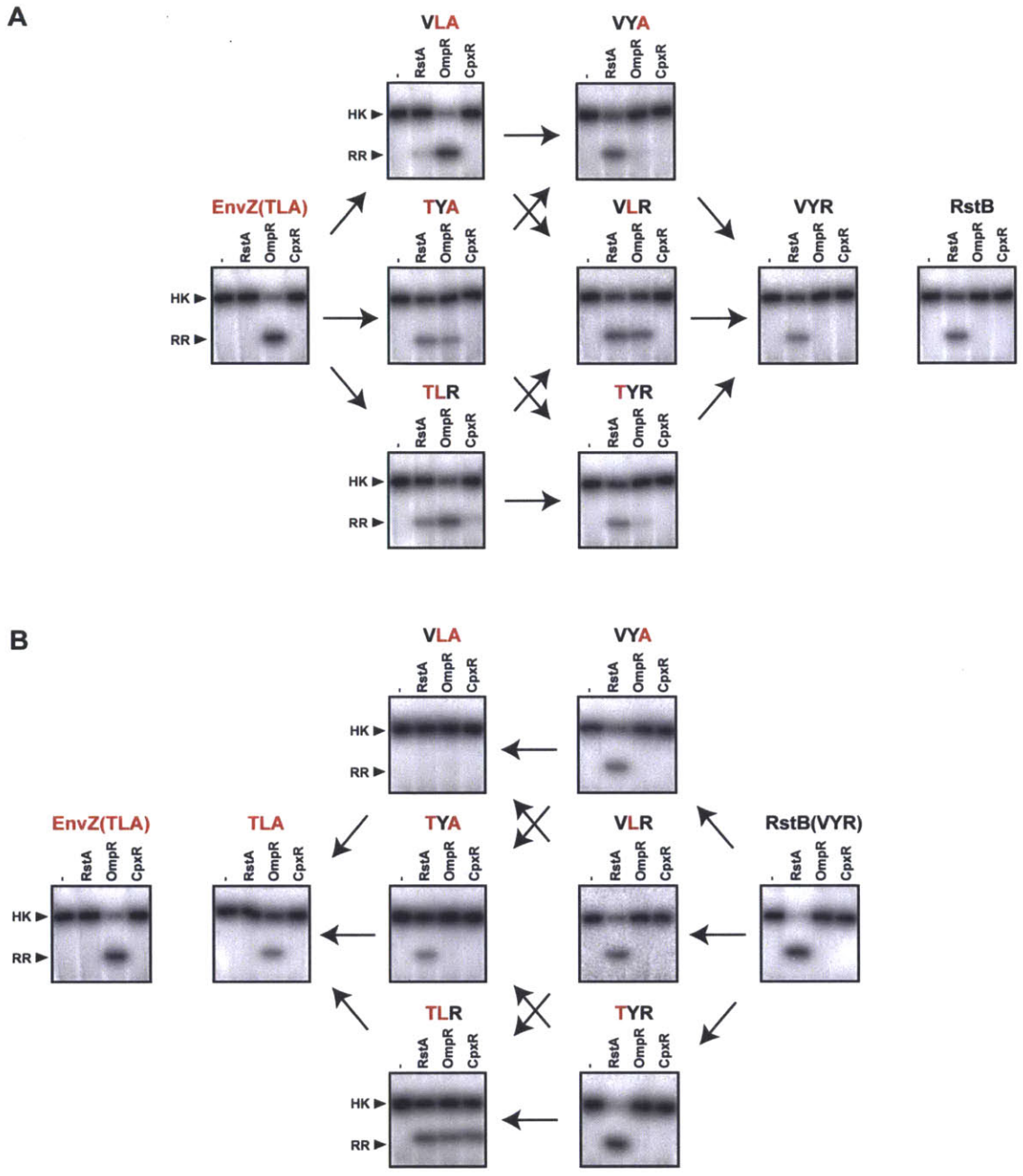


Figure 2.8 Converting the phosphotransfer specificity of EnvZ to match RstB and *vice versa*.

(A) Converting the phosphotransfer specificity of EnvZ to that of RstB. Wild-type EnvZ and each single, double, and triple mutant on the trajectory from EnvZ to RstB were autophosphorylated and then incubated alone or with one of three response regulators, as indicated, for 10 seconds. Wild-type RstB (far right) is shown for comparison to EnvZ(VYR). (B) Converting the phosphotransfer specificity of RstB to that of EnvZ. Wild-type RstB and each single, double, and triple mutant on the trajectory from RstB to EnvZ was autophosphorylated and then incubated alone or with one of three response regulators, as indicated, for 60 seconds. Wild-type EnvZ (far left) is shown for comparison to RstB(TLA). Arrows connect profiles of mutants differing by a

single amino acid substitution.

specificity residues being considered. For example, wild-type EnvZ is EnvZ(TLA) and the single point mutant EnvZ(T250V) is EnvZ(VLA). Each mutant was tested for phosphotransfer to the regulators OmpR, RstA, and CpxR (Figure 2.8). Under the conditions used, the wild type EnvZ and RstB are specific for, and only phosphorylate, their cognate substrates, OmpR and RstA, respectively.

In the context of EnvZ, each single mutant continued to phosphorylate OmpR (Figure 2.8A). The single mutants EnvZ(TYA) and EnvZ(TLR) also showed weak phosphorylation of RstA. Of the double mutants, EnvZ(VYA) and EnvZ(TYR) both preferentially phosphorylated RstA, with the former not detectably phosphorylating OmpR and the latter only weakly phosphorylating OmpR. The other double mutant, EnvZ(VLR) appeared to have an approximately equal preference for phosphotransfer to RstA and OmpR. In the context of RstB, none of the three single mutants had a major effect on specificity and each continued to phosphotransfer only to RstA (Figure 2.8B). By contrast, the double mutants each behaved differently; the mutant RstB(TYA) phosphorylated only RstA, the mutant RstB(TLR) was promiscuous and phosphorylated RstA, OmpR, and CpxR, while the mutant RstB(VLA) did not phosphorylate any of the response regulators under these reaction conditions.

The systematic mapping of the mutational trajectories from EnvZ to RstB and *vice versa* led to several interesting observations (Figure 2.8). First, the behaviors of intermediates along individual trajectories are often quite different. The most dramatic example is the double mutants of RstB, with RstB(TLR) phosphorylating all three substrates examined,

RstB(TYA) phosphorylating only RstA, and RstB(VLA) not phosphorylating any of the substrates. Second, we found that the individual specificity residues strongly influence each other. For example, the substitution V228T in the wild type RstB had very little effect on substrate preference, while the same substitution into RstB(VLA) converted a kinase that phosphorylated none of the regulators into a kinase that specifically phosphorylates OmpR (Figure 2.8B). The effect of the V228T substitution thus depends critically on the identity of other residues. As another example, the substitution Y230L in wild type RstA had little effect on specificity, but when introduced into RstA already harboring the V228T substitution produced a kinase that phosphorylated OmpR, RstA, and CpxR (Figure 2.8B). Similar observations were made for each of the other residues. Collectively, these data indicate that each specificity residue does not contribute independently or additively to the overall substrate specificity of a kinase. Rather, their contributions are frequently epistatic to one another and display context-dependence.

A complete specificity map of the mutational trajectories separating EnvZ/OmpR and RstB/RstA

The mutational trajectory scanning done for both EnvZ and RstB was extended to the response regulator OmpR. Converting OmpR to have the phosphotransfer specificity of RstA required 3 mutations in alpha helix-1 and 3 mutations in the β5-α5 loop (Figure 2.4A). We treated the loop as a single entity and made the 15 possible OmpR-RstA intermediates: 4 single, 6 double, 4 triple, and 1 quadruple mutant. We then examined phosphotransfer from each of the 7 EnvZ-RstB mutants (Figure 2.8A), as well as wild type EnvZ, RstB, and CpxA, to each of the 15 OmpR mutants and to wild-type OmpR, RstA, and CpxR, for a total of 180 pairwise combinations. The complete data are shown

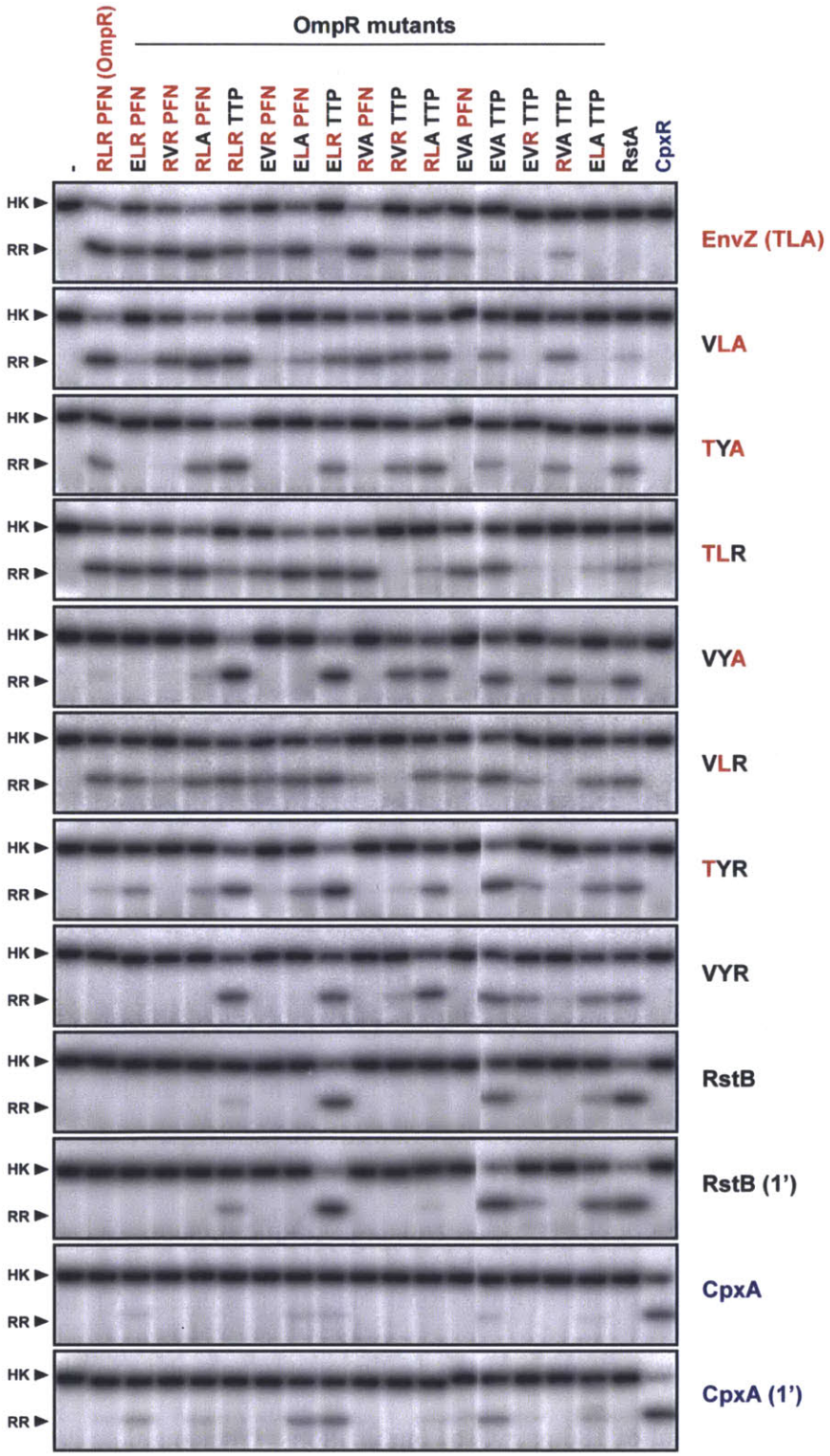

OmpR mutants
-
RLR PFN (OmpR)
ELR PFN
RVR PFN
RLA PFN
RLR TTP
EVR PFN
ELA PFN
ELR TTP
RVA PFN
RVR TTP
RLA TTP
EVA PFN
EVA TTP
EVR TTP
RVA TTP
ELA TTP
RstA
CpxR
HK
RR
EnvZ (TLA)
VLA
TYA
TLR
VYA
VLR
TYR
VYR
RstB
RstB (1')
CpxA
CpxA (1')

Figure 2.9 Complete trajectory-scanning mutagenesis of EnvZ and OmpR.

Each histidine kinase, indicated on the far right, was autophosphorylated and tested for phosphotransfer to each of the response regulators listed across the top. Mutants of EnvZ are named according to the identity of the three specificity residues being examined; for instance, wild-type EnvZ is 'TLA' whereas the mutant T250V is 'VLA'. Mutants of OmpR are named similarly. All phosphotransfer reactions were incubated for 10 seconds with the exception of RstB and CpxA, which were examined at both 10 seconds and 1 minute. Each kinase profile was composed of two separate gels that were run, exposed to phosphor screens, and scanned in parallel. The resulting two gel images were treated identically and then stitched together between OmpR(EVAPFN) and OmpR(EVATTP).

in Figures 2.9 and 2.10. All phosphotransfer reactions were run for 10 seconds, except for RstB and CpxA, which were run for 10 seconds and for 1 minute. To evaluate phosphotransfer, we quantified the relative intensity of each response regulator band for a given histidine kinase, yielding a profile of phosphotransfer activity for each kinase. From the comprehensive profiles, several observations and trends emerged (Figure 2.9, 2.10).

First, the triple mutant EnvZ(VYR) robustly phosphorylated wild type RstA as well as the quadruple mutant of OmpR in which all major specificity residues have been mutated to match those found in RstA. EnvZ(VYR) no longer phosphorylated OmpR, consistent with a complete change in specificity. However, it still phosphorylated two other OmpR mutational intermediates that the wild type RstB kinase did not, at least at the time point examined. This comparison supports the notion that the three residues we mutated in EnvZ are the dominant determinants of partner specificity, but that other residues play minor, fine-tuning roles, particularly in preventing non-cognate interactions.

Second, the data demonstrated that EnvZ and OmpR can tolerate some mutations in the specificity residues of their partner and still retain the ability to readily phosphotransfer. Wild-type EnvZ phosphorylated each of the single mutants of OmpR and three of the six

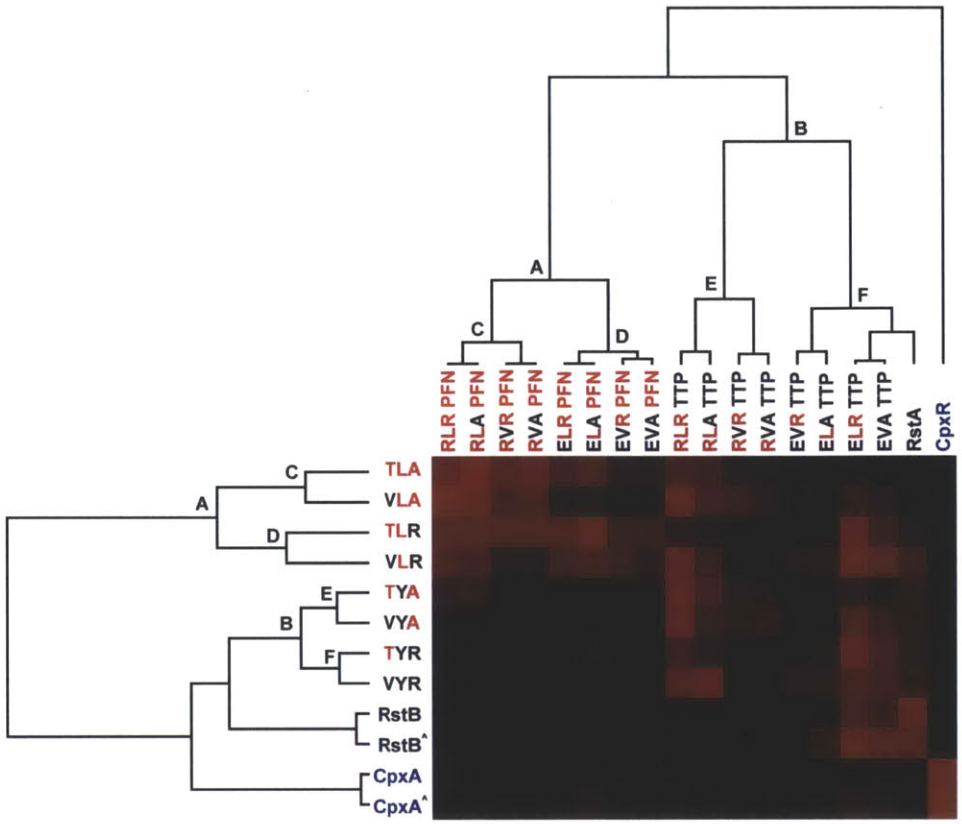


Figure 2.10 Hierarchical clustering of trajectory-scanning mutagenesis of EnvZ and OmpR.

Phosphotransfer profiles for each EnvZ construct examined in Figure 5 were quantified. The intensity of each response regulator band within a given kinase profile was expressed as a percentage of the maximally phosphorylated response regulator in that profile. Profiles were then clustered in two-dimensions, with the resulting tree shown for the response regulators (top) and histidine kinases (left). For each tree, the major clusters of EnvZ and OmpR mutants are designated by letters. The 1 minute time point profiles for RstB and CpxA are indicated by '^'.

double mutants nearly as well as it phosphorylated wild-type OmpR; however, it did not significantly phosphorylate the triple mutants or the quadruple mutant. Wild-type OmpR was efficiently phosphorylated by each of the EnvZ single mutants and one of the double mutants, but not by the triple mutant.

Third, these profiles reveal mutational paths from the specificity of the EnvZ/OmpR pair to that of RstB/RstA in which phosphotransfer is maintained. In other words, there is an

ordered series of single mutations that can be made in EnvZ and OmpR that convert them to the specificity of RstB and RstA, respectively, without disrupting their ability to phosphotransfer to one another along the way. For example, wild-type EnvZ phosphorylates OmpR and the single mutant OmpR(RLAPFN) to similar levels, and conversely the single mutant EnvZ(TLA) phosphorylates both OmpR and OmpR(RLAPFN). In Figure 2.11 we extend this example to show how EnvZ and OmpR could, in principle, change its specificity to that of the RstB/RstA system by a series of alternating mutations in the two molecules without ever severely disrupting their interaction. There are several such paths, although each path is not necessarily equivalent because CpxA phosphorylates some mutational intermediates of OmpR and some EnvZ mutants phosphorylate CpxR. For instance, EnvZ(TLR) phosphorylated CpxR, and OmpR(ELRPFN) was phosphorylated by CpxA (Figure 2.9, also see Figure 2.8). The avoidance of cross-talk may limit the possible evolutionary pathways between EnvZ/OmpR and RstA/RstB, or at least favor some relative to others (Figure 2.11).

We also quantified the phosphotransfer profiles for each EnvZ mutant and the wild type kinases (Figure 2.9) and performed hierarchical clustering in two dimensions, *i.e.* both the kinase and regulator dimensions (Figure 2.10). As expected, clustering the kinases places RstB close to the EnvZ(VYR) while CpxA is separated from EnvZ, the EnvZ mutants, and RstB. Similarly, clustering the regulators placed RstA close to the quadruple mutant OmpR(EVATTP) while CpxR formed a clear outgroup on its own.

The hierarchical clustering analysis provides insight into the relative importance of individual specificity residues. The profiles were clustered based on phosphorylation levels, but show a clear correspondence to sequence features. For instance, the two

primary clusters of OmpR mutants (labeled A and B in Figure 10) differ in the identity of their β5-α5 loops; that is, each OmpR mutant in cluster A has the residues 'PFN' whereas each mutant in cluster B has the residues 'TTP'. The branch lengths separating these clusters are long relative to the total length of the tree, indicating that the identity of the loop strongly splits the phosphotransfer profiles of the regulators. Within both cluster A and B, the next split in the tree correlates with the identity of position 1; that is, each OmpR mutant in cluster C (or cluster E) has an arginine at position 1 while each OmpR mutant in cluster D (or cluster F) has a glutamate at position 1. Again, the branch lengths are relatively long indicating a clear correlation between phosphotransfer behavior and sequence. The next split is based on identity at the second position, either a leucine or valine. The final split is based on the identity at the third position. In each case, this final split has extremely short branch lengths, reflecting the near identity of each profile pair that follows the split. In sum, the clustering analysis suggests a hierarchy to the contribution made by individual specificity residues within the regulators. The loop, which includes three residues, made the strongest contribution, followed by, in order, positions 1 > 2 > 3. A similar analysis was applied to the EnvZ mutants revealing that position 2 (Y or L) drives the initial clustering of EnvZ mutants, followed by position 3 (R or A), and finally position 1 (V or T).

Discussion

Determinants of specificity in paralogous protein families

Maintaining specificity and preventing unwanted cross-talk between highly similar proteins is a fundamental challenge for cells, and one that remains poorly understood. In many cases molecular recognition plays a critical role, but the ability to pinpoint the amino acids responsible and to determine the contributions of each residue to specificity has been elusive. Here, we tackled this problem in the context of bacterial two-component signal transduction systems where specificity is dictated by molecular recognition (Skerker et al., 2005). We note, however, that two-component signaling pathways are not insulated at all levels – for instance, distinct signaling pathways sometimes converge transcriptionally by regulating overlapping sets of genes (Laub and Goulian, 2007). However, the focus here is on the specificity of phosphotransfer for which there is little evidence of significant, physiologically-relevant cross-talk (Laub and Goulian, 2007).

To identify the amino acids that enforce the specificity of phosphotransfer, we examined patterns of amino acid coevolution in cognate kinase-regulator pairs. However, computational approaches alone do not unequivocally establish which residues are critical for specificity or reveal how each contributes to substrate selection. We therefore focused on experimentally rewiring the specificity of the model two-component proteins, EnvZ and OmpR. Previously we reported that EnvZ could be rewired to exhibit the substrate specificity of RstB by mutating as few as three of the coevolving residues (Skerker et al., 2008). Here we extended these results by rewiring OmpR to partner specifically with the histidine kinase RstB instead of EnvZ.

The residues mutated to rewire the partnering specificity of EnvZ and OmpR are predicted to be in close physical proximity during phosphotransfer. While no structure of EnvZ bound to OmpR exists, a co-crystal structure of a histidine kinase from *Thermotoga maritima* in complex with its cognate response regulator was recently solved (Casino et al., 2009) and can be used to infer physically proximal residues for EnvZ and OmpR. However, the spatial proximity of residues does not reveal how they govern specificity and whether individual residues promote the binding of a cognate protein or prevent interactions with non-cognate proteins. Moreover, the relative contribution made by each residue is difficult to discern from structural or spatial considerations alone.

To better dissect the role played by individual residues, we used alanine-scanning mutagenesis of EnvZ. However, of the nine major specificity residues in EnvZ (Figure 2.2), only one disrupted phosphotransfer to OmpR when mutated to alanine. These data suggest that no major hot spot exists for the EnvZ-OmpR interaction and that specificity is distributed across the interface. However, single alanine mutants do not always reveal the role of a particular residue. For example, EnvZ(L254A) showed very little change in substrate specificity, whereas EnvZ(L254Y) (Figure 2.8A) showed a significant level of cross-talk to RstA. Alanine-scanning mutagenesis also ignores any potential interdependencies that may exist between residues. Such relationships and non-additive effects on specificity were revealed in our comprehensive characterization of the mutational intermediates separating EnvZ and RstB. In several cases, the effect of a given substitution on phosphotransfer specificity depended significantly on what other substitutions had already been made; for example the mutation A255R in EnvZ had very little effect in the context of EnvZ(VYA) but led to significant promiscuity in the context

of EnvZ(TLA). These sorts of contextual and epistatic effects have been seen in other studies of molecular interaction specificity including corticosteroid receptor-ligand interaction (Ortlund et al., 2007) and transcription factor-DNA binding (Carlson et al., 2010). In principle, the context dependence of amino acids could lead to 'negative' epistasis in which one mutation on its own is detrimental until a second mutation is introduced. For example, the protein β-lactamase has evolved resistance to cefotaxime by accumulating five different mutations (Weinreich et al., 2006). While each mutation contributes to resistance, certain mutations actually decrease resistance unless, or until, one of the other mutations also occurs. We did not see any obvious case of negative epistasis when converting EnvZ to RstB or converting OmpR to RstA, as each mutation either increased interaction with the target molecule or had no effect. However, negative epistasis could exist when converting the specificity of other two-component signaling proteins.

Evolutionary implications

Our trajectory-scanning analysis provides a glimpse into the possible evolutionary history of two-component signaling proteins. The EnvZ/OmpR and RstB/RstA systems are relatively closely related and likely evolved by duplication of a common progenitor followed by sequence divergence, including at specificity sites. Mutations in specificity residues following duplication presumably required corresponding changes in their cognate regulators in order to maintain operation of each pathway as they diverged from one another to avoid pathway cross-talk. Our results demonstrate that an ordered series of mutations could occur in EnvZ and OmpR such that the two proteins would maintain significant levels of phosphotransfer while transiting through sequence space to the

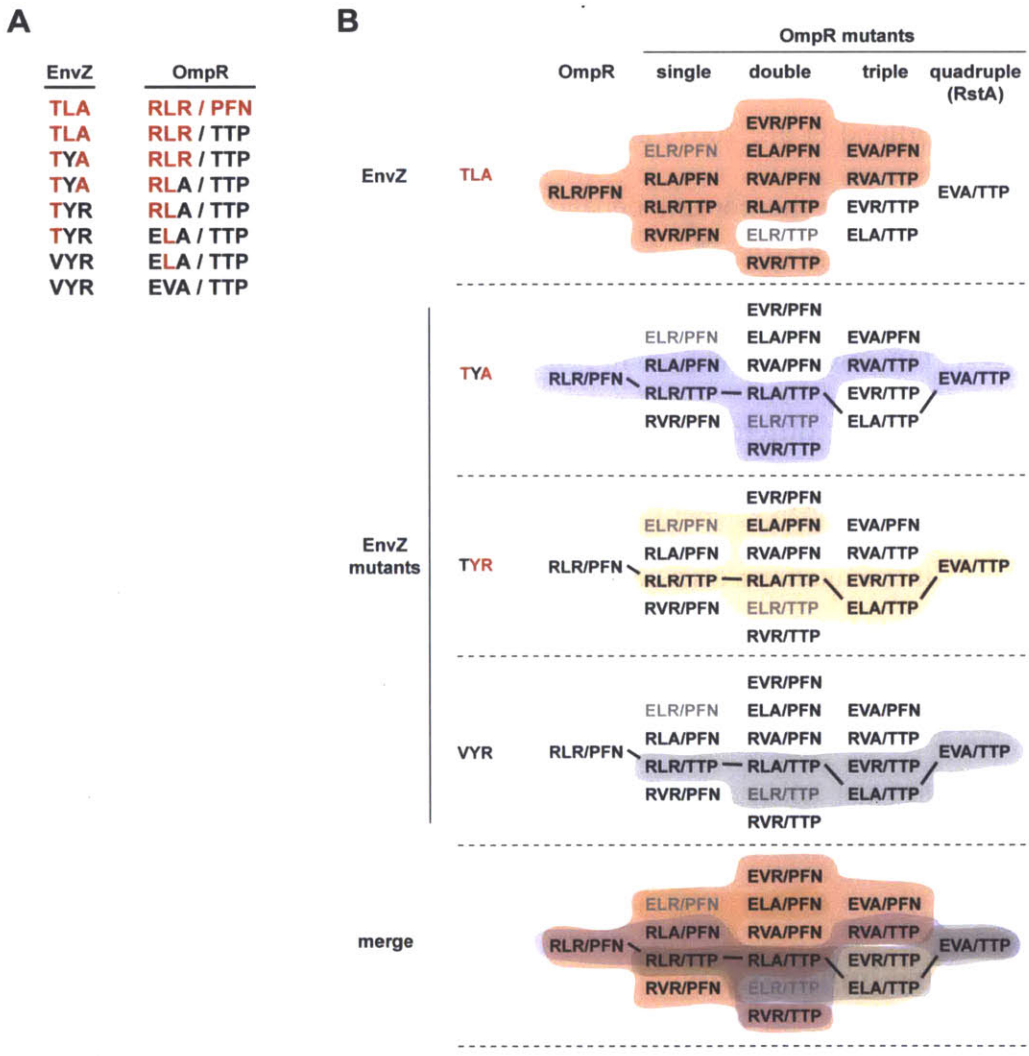


Figure 2.11 Mutational trajectories from EnvZ/OmpR to RstB/RstA.

EnvZ and OmpR can be converted by a series of single mutations to harbor the specificity residues found in RstB and RstA, respectively, without disrupting phosphotransfer in intermediate stages. (A) A series of single mutations can convert the specificity of EnvZ to match that of RstB and OmpR to match RstA. Starting with the wild type specificity residues in red text at the top, each subsequent line introduces a single mutation (shown in black text) until both sets of specificity residues have been completely changed. As noted in the text, we treated the loop as a single mutation. As shown in panel B, each kinase-regulator pair listed is capable of phosphotransfer and does not include a regulator that is phosphorylated by CpxA. (B) The complete set of intermediates between wild type OmpR (RLR/PFN) and the quadruple mutant (EVA/TTP) are listed. For wild type EnvZ (TLA), the single mutant EnvZ(TYA), the double mutant EnvZ(TYR), and the triple mutant EnvZ(VYR), the set of OmpR mutants recognized by each kinase are shaded, with a merge of all four at the bottom. Mutants that are phosphorylated by CpxA are listed in grey text, all others in black text. Bold lines connect the mutant series shown in panel A.

specificity residues of RstB/RstA (Figure 2.11), or *vice versa*. In addition, this series of mutations can occur without ever entering the sequence space occupied by another closely related (in sequence) pair, CpxA/CpxR thereby preventing cross-talk. Interestingly though, not all mutational trajectories have these characteristics of maintaining phosphotransfer and avoiding cross-talk, raising the possibility that sequence evolution following duplication is constrained or that natural selection may have favored certain trajectories over others. Analysis of other proteins, including β-lactamase, lambdoid phage integrases, hormone receptors, and the metabolic enzyme isopropylmalate dehydrogenase (Bridgham et al., 2009; Dorgai et al., 1995; Lunzer et al., 2005; Weinreich et al., 2006), have led to similar suggestions about the constraints on protein evolution.

Our trajectory scanning approach is related to other systematic studies of protein-protein interaction specificity, including homolog-scanning (Cunningham et al., 1989) and site-saturation mutagenesis (Miyazaki and Arnold, 1999). In many cases, however, such approaches involve single substitutions rather than an exploration of the entire mutational landscape separating two different proteins. Because the major specificity-determining residues of two-component signaling proteins have been previously mapped and are relatively limited in number, we were able to systematically generate all intermediates between EnvZ/OmpR and RstB/RstA. We note, however, that for the three major specificity residues in EnvZ, T250, L254, and A255, conversion to the corresponding residue in RstB requires two nucleotide substitutions. There are thus a great number of additional mutational intermediates that will be important to characterize in the future when considering the evolutionary history of EnvZ and RstB.

Intriguingly, our clustering analysis of the trajectory-scanning data also reveals an underlying hierarchy of the specificity-determining residues in EnvZ and OmpR. The clusters mapped based on phosphotransfer relationships were strongly correlated with the sequence of specificity residues. For example, the first branch point in the histidine kinase clusters separated those with a leucine at position 254 in EnvZ from those with a tyrosine at that position. These observations demonstrate that different residues contribute unequally to specificity. So although our alanine-scanning mutagenesis did not reveal any major hot spots and suggested that specificity is distributed, the trajectory-scanning study indicates that certain residues play more important roles than others. It will be interesting to see whether the hierarchies revealed here have influenced or constrained evolutionary trajectories of two-component signaling proteins, and if the relative importance of positions is similar in other two-component pairs.

Rational rewiring of two-component signaling pathways

The rational rewiring of two-component signaling proteins represents a stringent test of how well specificity is understood. Additionally, it opens the door to improved construction of synthetic signaling pathways in bacteria. Here, we used analyses of amino acid coevolution to guide the rational rewiring of the response regulator OmpR, a prototypical DNA-binding response regulator. With only a handful of mutations, the phosphotransfer specificity of OmpR was rewired to match that of RstA or CpxR. A recent study of *Rhodobacter* used structural data to guide the rewiring of chemotaxis response regulators to partner with the non-cognate kinase $CheA_3$ (Bell et al., 2010). The residues mutated in that study were in alpha helix 1 of the response regulator and most were identified here as coevolving residues. A genetic screen for altered partnering

specificity of the regulator PhoB also identified residues in alpha helix 1 (Haldimann et al., 1996). The successful rewiring of CheY and PhoB along with EnvZ and OmpR suggests that two-component proteins will be generally amenable to synthetic biology. However, it is not yet clear whether any histidine kinase (or response regulator) can be reprogrammed to behave like any other histidine kinase (or response regulator). For example, response regulators have been categorized into eight subfamilies, with the majority falling into just three (Grebe and Stock, 1999). OmpR, RstA, and CpxR all fall within one subfamily perhaps facilitating the interconversion of their specificities. Another important challenge for the future is to create novel kinase-regulator pairs with specificity residues that are orthogonal to those used in naturally occurring pairs. The functional hierarchies and interdependencies identified here will be important guides in engineering new, specific interactions. Similarly, these functional relationships should help in designing better algorithms for predicting kinase-regulator pairs in genomes of interest.

Final perspective

The life of a cell depends critically on the specificity of protein-protein interactions. Yet we still have a relatively primitive understanding of how such specificity is encoded within proteins and how a set of amino acids can allow binding of a cognate partner while excluding all other non-cognate partners. Two-component signal transduction systems represent an ideal model for addressing these fundamental issues as specificity is determined predominantly by a small set of residues. The consequent reduction in scope and scale enabled the systematic and comprehensive analyses presented here. More generally, the approaches used, including analyses of amino acid coevolution and

trajectory-scanning mutagenesis, will be widely applicable to the study of specificity and molecular recognition in many other protein-protein interactions.

Materials and Methods

Sequence analysis

The software HMMER (http://hmmer.org) was used, with an E-value cutoff of 0.01, to identify and align histidine kinase and response regulator sequences from fully sequenced bacterial genomes in GenBank. For histidine kinases, the models HisKA, HisKA_2, HisKA_3, and HWE_HK from the PFAM database were used. For response regulators, the model Response_reg was used. Histidine kinases and response regulators with GenBank genome identifier numbers differing by one, indicating adjacent genes, were identified, concatenated, and treated as cognate pairs. Sequences were filtered to ensure that no two sequences were more than 90% identical. The final set contained 4375 concatenated pairs of histidine kinase and response regulators. Columns in the multiple sequence alignment (MSA) containing greater than 10% gaps were eliminated.

Mutual information (MI) between columns was measured as described previously (Skerker et al., 2008). MI scores were adjusted to account for differences in the average MI of each column. For columns i and j in a multiple sequence alignment, we defined $MI(i,j)_{adj} = MI(i,j)_{raw} / (MI(i)_{avg} + MI(j)_{avg})/2$ where $MI(i)_{avg}$ and $MI(j)_{avg}$ are the average MI scores for column i and j paired with every other column in the alignment.

Clustering

Phosphorylation profiles in Figure 2.10 were constructed by quantifying response regulator bands in each profile (Figure 2.9) using ImageQuant (GE Healthcare) and then normalizing such that each regulator's value was represented as a percentage of the maximally phosphorylated regulator for a given kinase. Profiles were then subjected to

hierarchical clustering in two dimensions, with response regulators clustered using uncentered correlation and histidine kinases using Euclidean distance. Profiles were clustered using Cluster 3.0 (de Hoon et al., 2004) and visualized using Java Treeview (Saldanha, 2004).

Protein purification

All cloning and site-directed mutagenesis was done with Gateway pENTR vectors (Invitrogen) following procedures described previously (Skerker et al., 2008). Mutagenesis primers are listed in Table 2.1. Clones in pENTR vectors were mobilized into destination vectors for expression and purification using Gateway LR reactions according to the manufacturer's protocol (Invitrogen). Histidine kinases were moved into pDEST-His_6-MBP and response regulators into pDEST-TRX-His_6. Expression and purification was carried out exactly as described previously (Skerker et al., 2005).

Table 2.1 –
Primers

Primer Name	Sequence*
OmpR(R15E)	GTCGATGACGACATGGAGCTGCGTGCGCTGCTG
OmpR(L16V)	GACGACATGCGCGTGCGTGCGCTGCTG
OmpR(R22A)	GCGCTGCTGGAAGCTTATCTCACCGAA
OmpR(R15E;L16V)	GATGACGACATGGAGGTGCGTGCGCTGCTG
OmpR(R22E;Y23L)	CGTGCGCTGCTGGAAGAACTGCTCACCGAACAAGGC
OmpR(P106T,F107T,N108P)	GACTACATTCCAAAAACGACGCCGCCGCGTGAACTGCTG
OmpR(P109D)	AAACCGTTTAACGACCGTGAACTGCTG
EnvZ(T250V)	ACGCCGCTGGTGCGTATTCGC
RstB(V229T)	CGAACACCGTTAACGCGCCTGCGTTAT
RstB(Y233L)	GTGCGCCTGCGTCCTCGACTGGAGATG
RstB(R234A)	CGCCTGCGTTATGCACTGGAGATGAGC
RstB(Y233L;R234A)	TTAGTGCGCCTGCGTCTTGCACTGGAGATGAGCGAT
RstB(Y233L)_onV229T	ACGCGCCTGCGTCTTCGACTGGAGATG
RstB(V229T;Y233L)_onR234A	CGAACACCGTTAACGCGCCTGCGTCTT
EnvZ(L230A)	GGTGTTAAGCAAGCGGCGGATGACCGC
EnvZ(R234A)	CTGGCGGATGACGCCACGCTGCTGATG
EnvZ(T235A)	TGGCGGATGACCGCGCGCTGCTGATGGCGGG

EnvZ(L236A)	CGGATGACCGCACGGCGCTGATGGCGGGGGT
EnvZ(G240A)	CGCTGCTGATGGCGGCGGTAAGTCACGACTT
EnvZ(D244A)	CGGGGGTAAGTCACGCGTTGCGCACGCCGCT
EnvZ(R246A)	TAAGTCACGACTTGGCGACGCCGCTGACGCG
EnvZ(T247A)	GTCACGACTTGCGCGCGCCGCTGACGCGTAT
EnvZ(P248A)	GACTTGCGCACGGCGCTGACGCGTATT
EnvZ(L249A)	TTGCGCACGCCGGCGACGCGTATTCGC
EnvZ(T250A)	TGCGCACGCCGCTGGCGCGTATTCGCCTGGC
EnvZ(R251A)	GCACGCCGCTGACGGCGATTCGCCTGGCGAC
EnvZ(I252A)	CCGCTGACGCGTGCTCGCCTGGCGACT
EnvZ(R253A)	CGCTGACGCGTATTGCGCTGGCGACTGAGAT
EnvZ(L254A)	ACGCGTATTCGCGCGGCGACTGAGATG
EnvZ(A255T)	CGTATTCGCCTGACGACTGAGATGATG
EnvZ(T256A)	ATTCGCCTGGCGGCTGAGATGATGAGC
EnvZ(E257A)	CGCCTGGCGACTGCGATGATGAGCGAG
EnvZ(M258A)	GCCTGGCGACTGAGGCGATGAGCGAGCAGGA
EnvZ(M259A)	TGGCGACTGAGATGGCGAGCGAGCAGGATGG
EnvZ(S260A)	CGACTGAGATGATGGCGGAGCAGGATGGCTA
EnvZ(E261A)	CTGAGATGATGAGCGCGCAGGATGGCTATCT
EnvZ(Q262A)	AGATGATGAGCGAGGCGGATGGCTATCTGGC
EnvZ(D263A)	TGATGAGCGAGCAGGCGGGCTATCTGGCAGA
EnvZ(G264A)	TGAGCGAGCAGGATGCGTATCTGGCAGAATC
EnvZ(S269A)	TATCTGGCAGAAGCGATCAATAAAGAT
EnvZ(K272A)	CAGAATCGATCAATGCGGATATCGAAGAGTG
EnvZ(D273A)	AATCGATCAATAAAGCGATCGAAGAGTGCAA
EnvZ(E275A)	TCAATAAAGATATCGCGGAGTGCAACGCCAT
EnvZ(E276A)	ATAAAGATATCGAAGCGTGCAACGCCATCAT
EnvZ(E282A)	GCAACGCCATCATTGCGCAGTTTATCGACTA
EnvZ(Q283A)	ACGCCATCATTGAGGCGTTTATCGACTACCT
EnvZ(D286A)	TTGAGCAGTTTATCGCGTACCTGCGCACCGG

*Site-directed mutagenesis was done using the primer listed as well as its reverse complement.

Autophosphorylation and phosphotransfer reactions

For autophosphorylation analysis of alanine mutants, histidine kinases were at a final concentration of 5 μM in HKEDG buffer (10 mM HEPES-KOH pH 8.0, 50 mM KCl, 10% glycerol, 0.1 mM EDTA, 2 mM DTT) supplemented with 5 mM $MgCl_2$, 500 μM ATP, and 0.5 μCi [γ^{32}P]-ATP from a stock at ~6000 C_i/mmol (Perkin Elmer). Reactions were incubated at room temperature for 1 minute, stopped by the addition of 4X loading

buffer (500 mM Tris-HCl pH 6.8, 8% SDS, 40% glycerol, 400 mM β-mercaptoethanol), and analyzed by SDS-PAGE and phosphorimaging.

For phosphotransfer analysis, histidine kinases were autophosphorylated as above, but were incubated for 60 minutes at 30°C. Phosphotransfer was assessed by incubating autophosphorylated kinases with response regulators, each at a final concentration of 2.5 μM, at room temperature for the indicated time (either 10 seconds or 1 minute). Reactions were stopped by the addition of loading buffer, and analyzed by SDS-PAGE and phosphorimaging. For the experiments in Figures 2.4, 2.8 and 2.9, autophosphorylated kinases were purified away from unincorporated nucleotides by diluting them 1:10 in HKEDG and then washing eight times in Nanosep 30K Omega columns (Pall Life Sciences) to minimize the effect of any phosphatase activity. The final eluate was diluted back to the original volume and $MgCl_2$ added to 5 mM before assessing phosphotransfer.

For alanine-scanning mutagenesis, to gauge reproducibility and assess significance in the changes observed, we repeated the phosphotransfer reactions for wild type EnvZ six times and a subset of the mutants three times. Standard deviations in each case were ~5-10% of the mean.

Acknowledgements

We thank A. Keating for helpful comments on the manuscript.

References

Bell, C.H., Porter, S.L., Strawson, A., Stuart, D.I., and Armitage, J.P. (2010). Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex. PLoS Biol *8*, e1000306.

Bridgham, J.T., Ortlund, E.A., and Thornton, J.W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. Nature *461*, 515-519.

Burger, L., and van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol *4*, 165.

Carlson, C.D., Warren, C.L., Hauschild, K.E., Ozers, M.S., Qadir, N., Bhimsaria, D., Lee, Y., Cerrina, F., and Ansari, A.Z. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. Proc Natl Acad Sci U S A *107*, 4544-4549.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Cunningham, B.C., Jhurani, P., Ng, P., and Wells, J.A. (1989). Receptor and antibody epitopes in human growth hormone identified by homolog-scanning mutagenesis. Science *243*, 1330-1336.

de Hoon, M.J., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453-1454.

Dorgai, L., Yagil, E., and Weisberg, R.A. (1995). Identifying determinants of recombination specificity: construction and characterization of mutant bacteriophage integrases. J Mol Biol *252*, 178-188.

Fisher, S.L., Kim, S.K., Wanner, B.L., and Walsh, C.T. (1996). Kinetic comparison of the specificity of the vancomycin resistance VanSfor two response regulators, VanR and PhoB. Biochemistry *35*, 4732-4740.

Gao, R., Mack, T.R., and Stock, A.M. (2007). Bacterial response regulators: versatile regulatory strategies from common domains. Trends in biochemical sciences *32*, 225-234.

Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*, 7156-7165.

Grebe, T.W., and Stock, J.B. (1999). The histidine protein kinase superfamily. Adv Microb Physiol *41*, 139-227.

Grimshaw, C.E., Huang, S., Hanstein, C.G., Strauch, M.A., Burbulys, D., Wang, L., Hoch, J.A., and Whiteley, J.M. (1998). Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in Bacillus subtilis. Biochemistry *37*, 1365-1375.

Haldimann, A., Prahalad, M.K., Fisher, S.L., Kim, S.K., Walsh, C.T., and Wanner, B.L. (1996). Altered recognition mutants of the response regulator PhoB: a new genetic strategy for studying protein-protein interactions. Proc Natl Acad Sci U S A *93*, 14361-14366.

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annual review of genetics *41*, 121-145.

Lunzer, M., Miller, S.P., Felsheim, R., and Dean, A.M. (2005). The biochemical architecture of an ancient adaptive landscape. Science *310*, 499-501.

Miyazaki, K., and Arnold, F.H. (1999). Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. J Mol Evol *49*, 716-720.

Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., and Thornton, J.W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. Science *317*, 1544-1548.

Qin, L., Cai, S., Zhu, Y., and Inouye, M. (2003). Cysteine-scanning analysis of the dimerization domain of EnvZ, an osmosensing histidine kinase. J Bacteriol *185*, 3429-3435.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246-3248.

Schwartz, M.A., and Madhani, H.D. (2004). Principles of MAP kinase signaling specificity in Saccharomyces cerevisiae. Annu Rev Genet *38*, 725-748.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

Ubersax, J.A., and Ferrell, J.E., Jr. (2007). Mechanisms of specificity in protein phosphorylation. Nat Rev Mol Cell Biol *8*, 530-541.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A *106*, 67-72.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science *312*, 111-114.

White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2007). Features of protein-protein interactions in two-component signaling deduced from genomic libraries. Methods Enzymol *422*, 75-101.

Chapter 3

Adaptive mutations that prevent cross-talk enable the expansion of paralagous signaling protein families

This work was published as Emily J. Capra, Barrett S. Perchuk, Jeffrey M. Skerker, and Michael T. Laub. 2012. Cell.

EJC and MTL conceived and designed the experiments. EJC performed all of the experiments. BSP helped with the protein purifications and the profiles shown in Figure 3.3 JMS contributed constructs. EJC and MTL wrote the paper.

Abstract

Orthologous proteins often harbor numerous substitutions, but whether these differences result from neutral or adaptive processes is usually unclear. To tackle this challenge, we examined the divergent evolution of a model bacterial signaling pathway comprising the kinase PhoR and its cognate substrate PhoB. We show that the specificity-determining residues of these proteins are typically under purifying selection, but have, in α-proteobacteria, undergone a burst of diversification followed by extended stasis. By reversing mutations that accumulated in an α-proteobacterial PhoR, we demonstrate that these substitutions were adaptive, enabling PhoR to avoid cross-talk with a paralogous pathway that arose specifically in α-proteobacteria. Our findings demonstrate that duplication and the subsequent need to avoid cross-talk strongly influence signaling protein evolution. These results provide a concrete example of how system-wide insulation can be achieved post-duplication through a surprisingly limited number of mutations. Our work may help explain the apparent ease with which paralogous protein families expanded in all organisms.

Introduction

The evolutionary forces and selective pressures that influence protein sequences remain poorly understood at a detailed, molecular level. A comparison of orthologs often reveals tens to hundreds of amino acid differences. How and why do functionally equivalent proteins diverge in different organisms? Many of the accumulated substitutions may be functionally neutral and result from processes such as genetic drift. However, some mutations may have been adaptive and provided a fitness advantage. Identifying these beneficial mutations and pinpointing the advantage that they provide are difficult problems. Comparative sequence analyses, such as measures of codon substitution patterns or dN/dS ratios (Yang and Bielawski, 2000), can help to identify residues that are potentially adaptive, but such approaches are frequently insufficient and difficult to validate. Additionally, elucidating why certain mutations are beneficial requires a genetically manipulatable organism and an ability to probe the effects of individual mutations *in vivo*.

In many cases where protein evolution has been studied experimentally (reviewed in (Dean and Thornton, 2007)), the relevant proteins were examined *in vitro* or in heterologous hosts, and thus outside their native cellular context, possibly eliminating or obscuring important evolutionary constraints. For example, signal transduction proteins are often part of large paralogous families that expand through duplication and divergence. The duplication-divergence process thus runs an inherent risk of introducing cross-talk with existing pathways. A study of SH3 domains from *S. cerevisiae* and humans suggested that the avoidance of cross-talk may represent an important selective pressure in the evolution of paralogous protein families (Zarrinpar et al., 2003). However,

a direct demonstration that cross-talk influences the evolution of signaling proteins and, more importantly, an understanding of how this occurs at the amino-acid level are lacking.

To tackle these challenges we examined the evolution of two-component signal transduction proteins in bacteria. These pathways, a primary means of signal transduction in prokaryotes, typically involve a sensor histidine kinase that, upon receipt of an input stimulus, autophosphorylates and then transfers its phosphoryl group to a cognate response regulator, which in turn modulates gene expression (Stock et al., 2000). Most histidine kinases are bifunctional and can, in the absence of an input signal, stimulate the dephosphorylation of their cognate response regulators, effectively acting as phosphatases (Huynh and Stewart, 2011).

Although most bacteria encode between 20 and 200 two-component signaling pathways (Alm et al., 2006), very little cross-talk occurs at the level of phosphotransfer *in vivo* (Grimshaw et al., 1998; Laub and Goulian, 2007; Siryaporn and Goulian, 2008; Skerker et al., 2005). Two-component pathways are highly specific, typically with one-to-one relationships between cognate kinases and regulators. When not stimulated to autophosphorylate, the phosphatase activity of a given histidine kinase can help to eliminate cross-talk and the errant phosphorylation of its cognate regulator (Siryaporn and Goulian, 2008). However, when stimulated as a kinase, molecular recognition is the dominant mechanism for preventing phosphotransfer cross-talk and thereby maintaining the fidelity of distinct signaling pathways. Systematic analyses of phosphotransfer have demonstrated that histidine kinases are endowed with an intrinsic ability to discriminate their *in vivo* cognate substrate from all other non-cognate substrates (Skerker et al.,

2005). Analyses of amino acid coevolution in cognate signaling proteins identified the key specificity-determining residues in histidine kinases and response regulators (Bell et al., 2010; Capra et al., 2010; Casino et al., 2009; Skerker et al., 2008; Weigt et al., 2009). Rational mutagenesis of these residues can reprogram the partnering specificity of a histidine kinase or response regulator (Bell et al., 2010; Capra et al., 2010; Skerker et al., 2008).

For a given two-component pathway there is likely strong purifying selective pressure on its key specificity-determining residues to preserve the kinase-substrate interaction. Even single amino acid changes in specificity residues can drastically change the interaction capabilities and preferences of a histidine kinase or response regulator (Capra et al., 2010). Nevertheless, an inspection of orthologous kinases or response regulators often reveals divergent evolution and variability in the specificity residues of certain subsets of orthologs, raising the question of whether these changes resulted from neutral or adaptive processes. We favored the latter, hypothesizing that specificity residues must change in order to avoid cross-talk between pathways following gene duplication events. Two-component signaling pathways often expand through duplication (Alm et al., 2006), and following such events, bacteria presumably must accumulate mutations that insulate the new pathways from each other and maintain their isolation from other, existing two-component pathways. Here, we provide direct experimental evidence that the avoidance of cross-talk is indeed a major selective force in the evolution of two-component signaling pathways. Through *in vitro* studies and fitness competition assays, we identify specific substitutions in a model two-component pathway, PhoR-PhoB, that represent an adaptation to the duplication of another two-component signaling system. Similar

adaptations likely accompanied each of the duplication and divergence events underlying the massive expansion of two-component signaling protein families in bacteria. Accordingly, global analyses of specificity-determining residues in extant bacterial genomes reveal a pervasive trend toward orthogonality in these signaling proteins.

Results

To identify vertical inheritance of PhoR and PhoB

To examine the divergent evolution of two-component signaling pathways, we focused on the PhoR-PhoB signaling pathway (Wanner and Chang, 1987), which is found throughout the bacterial kingdom and helps a wide range of organisms respond to phosphate starvation. To systematically identify orthologs of *E. coli phoR* and *phoB*, we used a modified version of reciprocal best hits in BLAST analysis that allows for the identification of putative duplications. Most proteobacteria, except a small number of δ-proteobacteria, were found to encode a single ortholog of each *phoR* and *phoB*, suggesting that these genes are rarely duplicated, particularly in the α-, β-, and γ-proteobacteria (Figure 3.1A). Additionally, gene trees for *phoR* and *phoB* closely matched a species tree (Figures 3.1B, 3.2A-B), indicating that this signaling system has likely been vertically inherited in these clades.

Given that *phoR* and *phoB* genes were rarely duplicated during the evolution of proteobacteria, it might be expected that the residues dictating phosphotransfer specificity would be relatively constant in order to preserve the interaction between PhoR and PhoB. We thus examined the six residues in PhoR and seven residues in PhoB previously identified as critical determinants of specificity in two-component signaling proteins (Capra et al., 2010). We extracted these residues, hereafter referred to simply as specificity residues, from 149 PhoR orthologs and 92 PhoB orthologs, and built sequence logos representing the relative frequency of amino acids at each specificity position (Figure 3.1B). The difference in number of PhoR and PhoB orthologs results from the

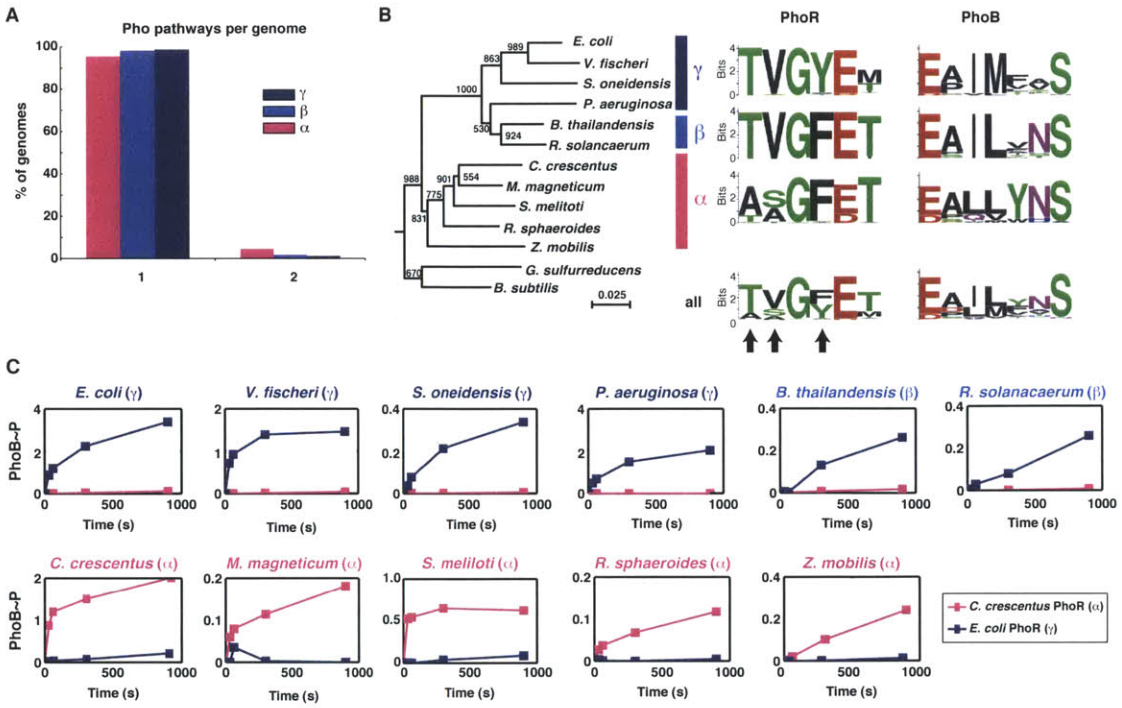


Figure 3.1 Phosphotransfer specificity of PhoR is different in α– and γ-proteobacteria.

(A) Percentage of genomes harboring one or two Pho pathways. (B) Neighbor-joining tree for a subset of PhoR orthologs and sequence logos for specificity residues in PhoR and PhoB orthologs. Logos are shown for orthologs in each subdivision and as a combined set. Bootstrap values are out of 1000. A neighbor-joining tree for PhoB orthologs and a species tree for the species used are in Figure 3.1A-B. (C) Time-courses of phosphotransfer from *C. crescentus* PhoR (α) and *E. coli* PhoR (γ) to each of 11 PhoB orthologs from representative α-, γ-, and β-proteobacteria as noted above each graph. Band intensities for each PhoB were normalized in each experiment to the initial amount of autophosphorylated kinase. Values for PhoB phosphorylation can be greater than one as ATP was in excess in the reaction, allowing for re-autophosphorylation of PhoR and subsequent transfer to PhoB. For original gel images, see Figure 3.2D.

independent identification of kinase and regulator orthologs; most organisms encode both PhoR and PhoB.

The specificity residues of both PhoR and PhoB are generally well-conserved (Figures 3.1B, 3.2C), although several positions showed substantial variability. We then split the

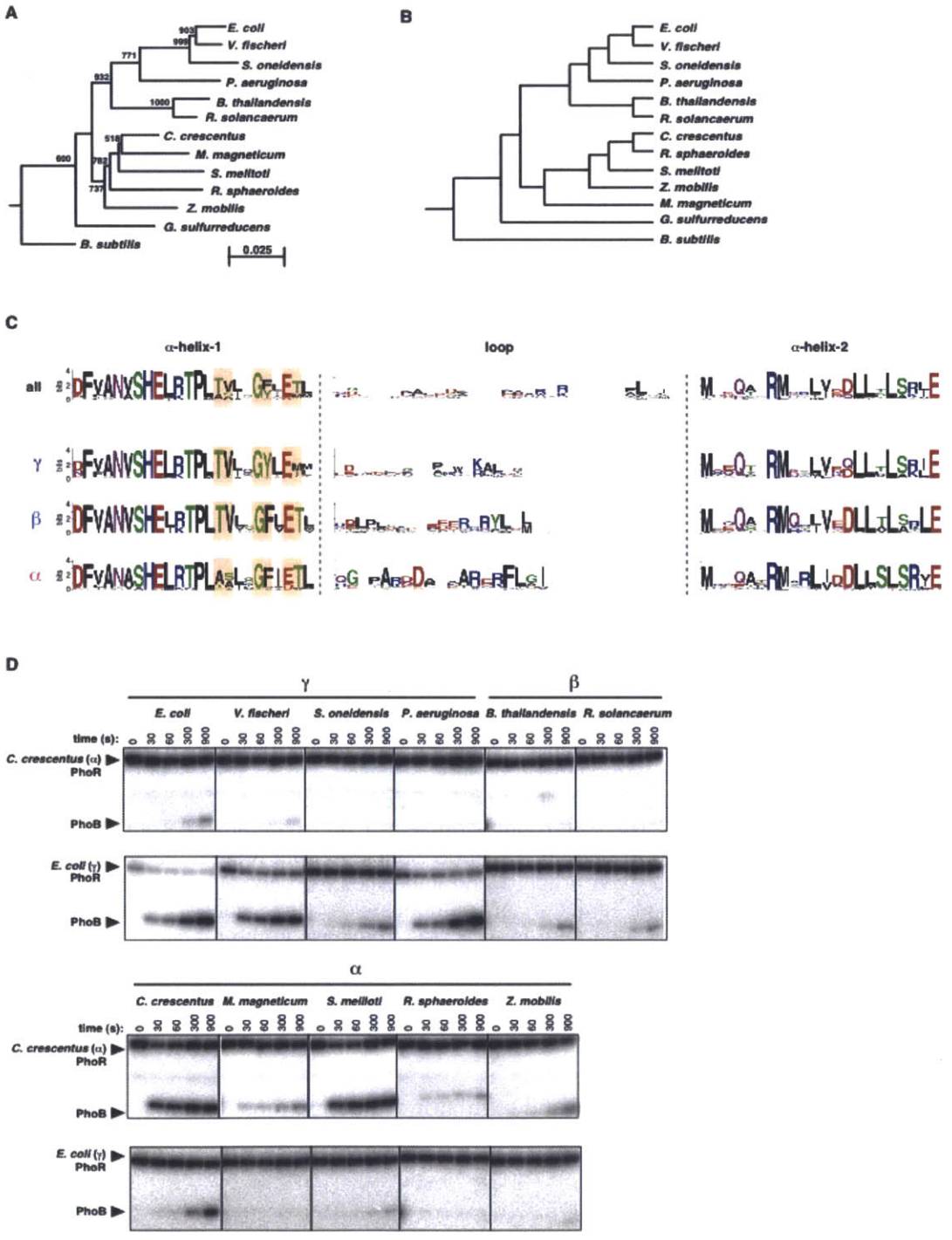


Figure 3.2 Phylogenetic analyses of PhoR and PhoB.

(A) Neighbor-joining tree built using receiver domains of PhoB orthologs from the organisms indicated. Bootstrap values are out of 1000. (B) Species tree for the species represented in the gene trees of PhoR and PhoB. The species tree was obtained from microbesonline.org and built using highly conserved genes that were likely vertically inherited. (C) Sequence logos for the DHp domain of PhoR orthologs, split by clade or as a complete set. The DHp domain was divided into

α-helix 1, the loop region, and α-helix 2 as the loop region is of variable length and aligns poorly. Specificity residues in helix 1 are shaded. (D) Time courses of phosphotransfer from *C. crescentus* and *E. coli* PhoR to a set of 11 PhoB orthologs. *C. crescentus* and *E. coli* PhoR constructs were autophosphorylated and tested for phosphotransfer to each PhoB at the times indicated. The species and proteobacterial subdivision from which each PhoB was taken are indicated at the top. Quantifications are in Figure 2.1C.

PhoR and PhoB sequences into groups corresponding to the three major proteobacterial subdivisions, α, β, and γ. Sequence logos built for each phylogenetic group revealed that differences between subdivisions can account for nearly all of the variability in the combined sequence logos (Figure 3.1B). For instance, in γ- and β-proteobacteria the first two positions are almost always threonine and valine, whereas in α-proteobacteria these positions are usually alanine and serine or two alanines. Similar observations were made for the specificity residues of PhoB orthologs grouped according to phylogenetic subdivision. Importantly, each PhoR and PhoB sequence logo was built using species that are highly diverged. The strong conservation within each clade thus suggests that specificity residues are usually subject to strong purifying selection. Why, though, have specificity residues diverged between clades?

Identification of adaptive mutations that prevent cross-talk *in vitro*

The clade-specific differences in PhoR and PhoB specificity residues may simply reflect degeneracy in the residues that enable PhoR and PhoB to interact. Alternatively, the differences may have produced functional changes such that a PhoR from one clade is less efficient at interacting with a PhoB from a different clade. To distinguish between these possibilities, we purified PhoR kinases from representative γ– and α–proteobacteria, *E. coli* and *C. crescentus*, and examined their ability to phosphorylate a panel of 11 PhoB orthologs from α, β, and γ-proteobacteria (Figure 3.1C, 3.2D). For each

PhoB from a γ-proteobacterium, phosphotransfer from the *E. coli* (γ) PhoR was significantly faster than from *C. crescentus* (α) PhoR. Similarly, each PhoB from an α-proteobacterium was preferentially phosphorylated by the α–PhoR. For the two chosen β–PhoB orthologs, we observed more rapid phosphorylation by the γ–PhoR than the α–PhoR, consistent with the specificity residues of the β–PhoR and β–PhoB orthologs being more similar to those found in γ–proteobacteria than those in α–proteobacteria. We conclude that within each proteobacterial subdivision, the phosphotransfer specificity of PhoR and PhoB orthologs is relatively static. However, substitutions in the specificity residues of α-PhoR and α-PhoB orthologs have led to significant differences in phosphotransfer specificity between clades.

The changes in PhoR-PhoB specificity residues, and consequent alteration of interaction specificity, could have resulted from neutral drift. However, the strong conservation of specificity residues within each clade, which includes species that are widely divergent, suggests that such drift is extremely rare. Instead, the alternative PhoR-PhoB specificity residues in α-proteobacteria may be adaptive and provide an important selective advantage. We hypothesized that the substitutions in α-PhoR and α-PhoB specificity residues prevent unwanted cross-talk with another pathway that is specific to the α-proteobacteria, *i.e.* negative selection led to changes in α-PhoR and α-PhoB. This model predicts that PhoR orthologs from γ-proteobacteria may phosphorylate response regulators found exclusively in α-proteobacteria, which the α-PhoR orthologs have adapted to avoid phosphorylating.

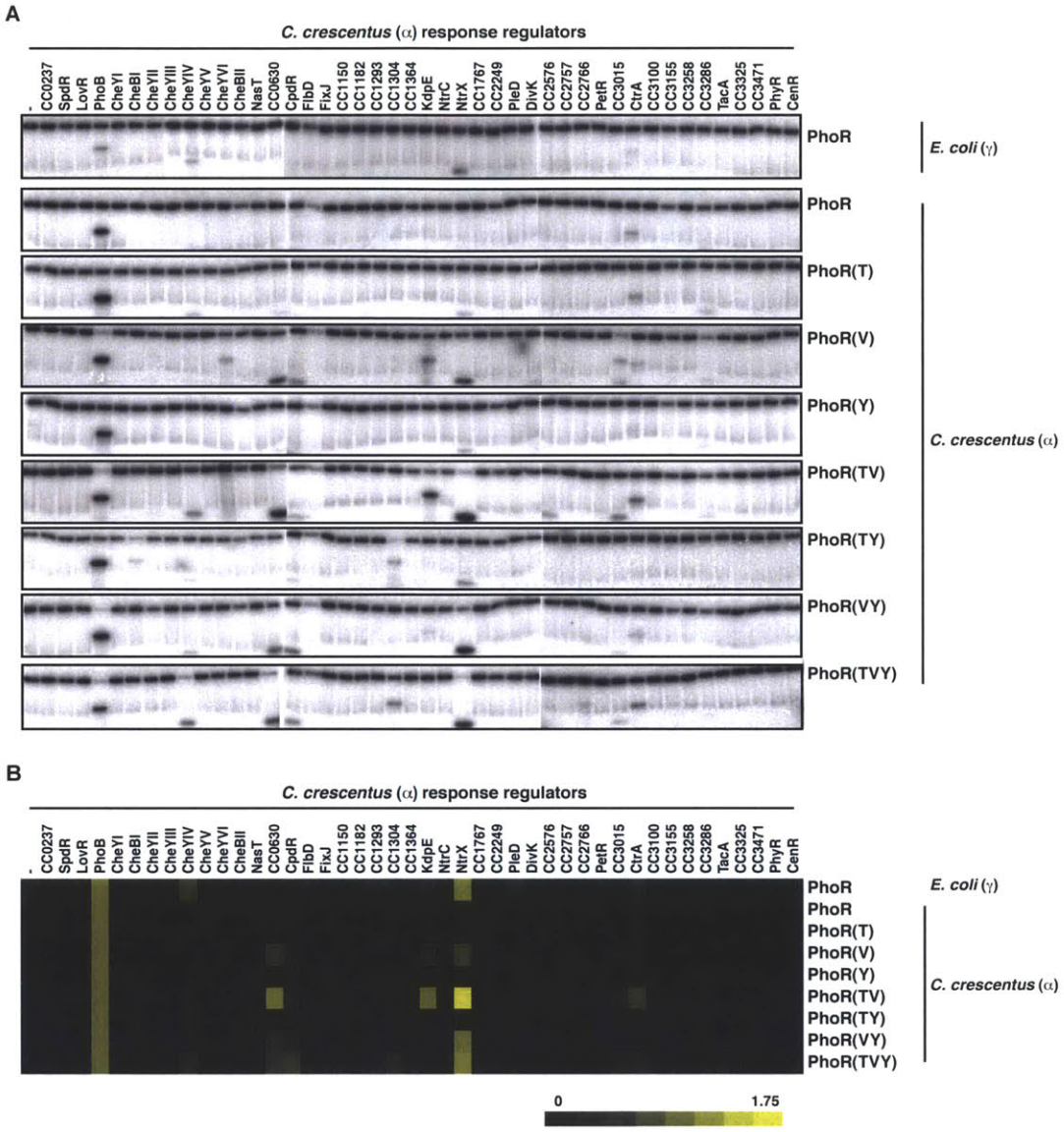


Figure 3.3 Substituting γ-like specificity residues into α-PhoR increases phosphorylation of NtrX.

(A) Phosphotransfer profiling of *E. coli* PhoR, *C. crescentus* PhoR, and *C. crescentus* PhoR mutants containing γ-proteobacterial specificity residues. Each autophosphorylated kinase, indicated on the right, was incubated with each of 44 *C. crescentus* response regulators, indicated across the top, for 15 minutes. (B) Quantification of profiles in panel A. Band intensities for each response regulator were normalized to the level of autophosphorylated kinase and then plotted relative to PhoB.

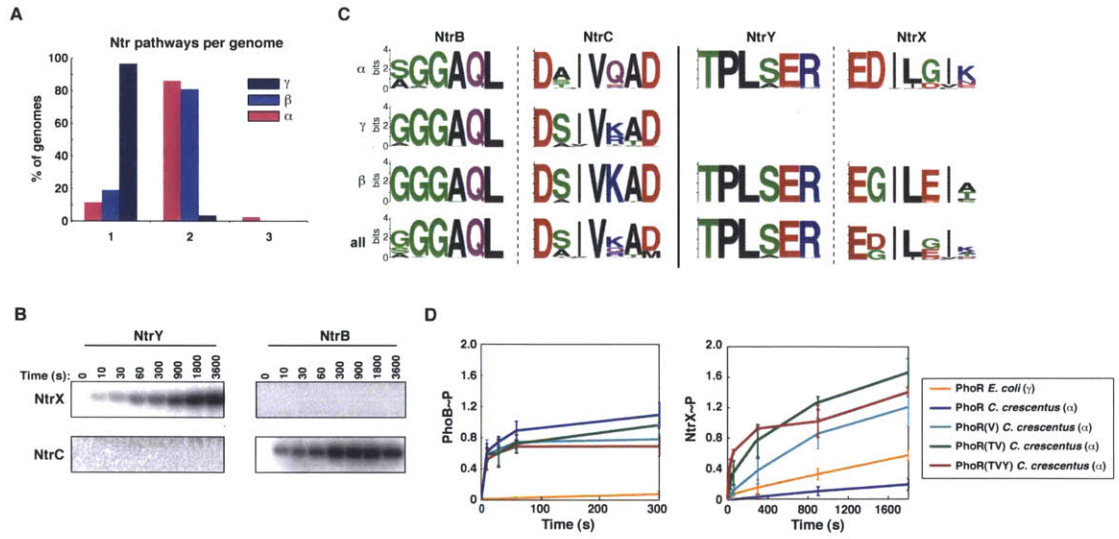


Figure 3.4 The divergent evolution of NtrX after duplication led initially to cross-talk with PhoR in α-proteobacteria.

(A) Percentage of genomes harboring one, two, or three Ntr pathways. (B) Time course of phosphotransfer from *C. crescentus* kinases NtrY and NtrB to *C. crescentus* response regulators NtrX and NtrC. The NtrY and NtrB constructs were autophosphorylated and examined for phosphotransfer at the time points indicated. Error bars represent standard deviations, n=3. (C) Sequence logos for specificity residues in NtrB-NtrC and NtrY-NtrX orthologs. Logos are shown for orthologs in each subdivision and as a combined set. (D) Time course of phosphotransfer from *E. coli* PhoR, *C. crescentus* PhoR, and *C. crescentus* PhoR mutants, listed in the legend, to *C. crescentus* PhoB and NtrX. Error bars represent standard deviations, n=3.

To test this possibility, we performed comprehensive phosphotransfer profiling of *E. coli* (γ) and *C. crescentus* (α) PhoR. Both PhoR constructs were autophosphorylated *in vitro* and then examined, in parallel, for phosphotransfer to the 44 response regulators encoded by *C. crescentus* (Figure 3.3). Both PhoR constructs phosphorylated the *C. crescentus* PhoB, consistent with their orthologous relationship; although as noted above, phosphotransfer from the α-PhoR is more robust. Interestingly, the γ-PhoR showed significant phosphotransfer to NtrX, whereas the α-PhoR construct did not. Notably, most α-proteobacteria encode two paralogous Ntr systems, NtrB-NtrC and NtrX-NtrY, while the γ–proteobacteria typically encode only one, NtrB-NtrC (Figure 3.4A). The two

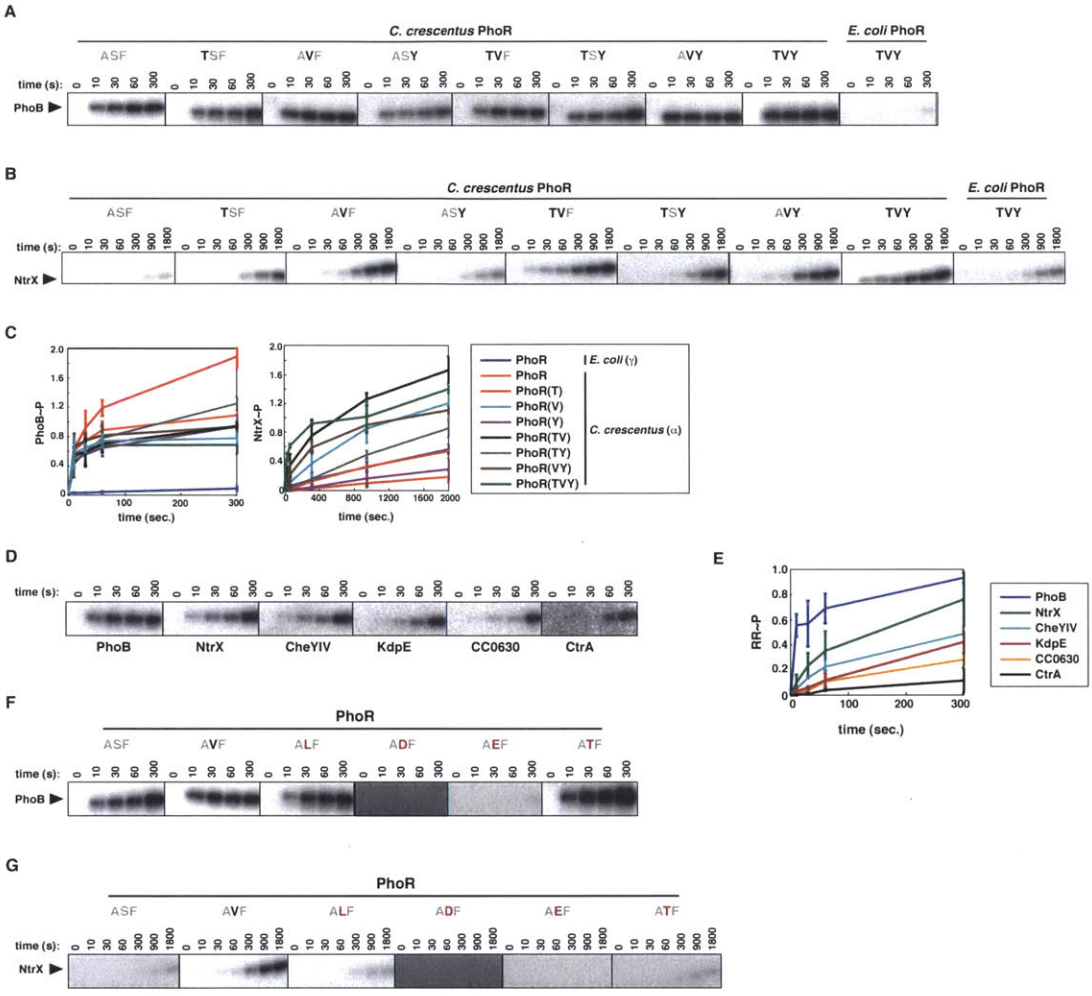


Figure 3.5 Time courses of phosphotransfer from *C. crescentus* PhoR specificity mutants.

(A-B) Time courses of phosphotransfer from *E. coli* PhoR, *C. crescentus* PhoR, and *C. crescentus* PhoR mutants to either *C. crescentus* PhoB (A) or *C. crescentus* NtrX (B). For each kinase the identities of specificity positions 1, 2, and 4 (see Figure 3.1B) are listed. Wild-type *C. crescentus* PhoR has A, S, and F and wild-type *E. coli* PhoR has T, V, and Y. Kinase constructs were autophosphorylated and then examined for phosphotransfer at the time points indicated. Representative gels from three independent replicates are shown. Only the response regulator band is shown. (C) Quantifications of phosphotransfers from panels A-B and replicates. Error bars indicate standard deviations, n=3. (D) Time courses of phosphotransfer from PhoR(TV) to the *C. crescentus* response regulators that were phosphorylated in the profile shown in Figure 2. Only the response regulator bands are shown. Representative gels from three independent experiments are shown. (E) Quantification of the phosphotransfers from panel D and replicates. Error bars indicate standard deviations, n=3. (F-G) Time courses of phosphotransfer from *C. crescentus* PhoR harboring various mutations at specificity position 2 to either *C. crescentus* PhoB (F) or *C. crescentus* NtrX (G). Wild-type specificity residues for *C. crescentus* PhoR are A, S, and F. Only the response regulator band is shown.

α–Ntr systems, which likely arose through duplication and divergence, do not cross-talk with each other *in vitro* (Figure 3.4B) and, consistently, have different specificity residues (Figure 3.4C). Collectively, our observations suggest that the different PhoR specificity residues seen in α–proteobacteria may have evolved to accommodate the presence of a second, lineage-specific pathway, NtrX-NtrY. Such a change in PhoR was presumably accompanied by changes in the PhoB specificity residues (see Figure 3.1B) to maintain phosphotransfer from PhoR.

Thus, we hypothesized that the alanine, serine, and phenylalanine found at specificity positions 1, 2, and 4 of α–PhoR proteins represent adaptive mutations that prevent cross-talk to NtrX. To test this hypothesis, we created a series of *C. crescentus* PhoR mutants in which specificity residues were replaced with the corresponding residues from γ–proteobacterial PhoR. We made each single mutant, three double mutants, and the triple mutant. Each mutant kinase was then profiled against the complete set of *C. crescentus* response regulators to examine what effect, if any, these residues have on phosphotransfer specificity. Strikingly, each mutant led to a significant increase in NtrX phosphorylation (Figure 3.3). We also examined detailed time courses of phosphotransfer from each mutant PhoR, as well as the wild-type kinases, to the *C. crescentus* regulators PhoB and NtrX. Each mutant kinase exhibited an increase in cross-talk with NtrX compared to the wild-type *C. crescentus* (α) PhoR, but retained the ability to phosphorylate *C. crescentus* PhoB at rates comparable to the wild-type PhoR (Figures 3.4D, 3.5A-C). Although some mutant PhoR kinases phosphorylated several substrates (see Figure 3.3), we focused on PhoB and NtrX as time-courses of phosphotransfer indicated these as the two preferred targets of mutant PhoR constructs (Figure 3.5D-E).

The most significant cross-talk to NtrX occurred for PhoR mutants with a valine substituted for serine at specificity position 2. Importantly, substantial cross-talk was not observed when this serine was substituted with other residues including leucine, aspartate, glutamate, and threonine. Only valine, corresponding to that found in γ-proteobacterial PhoR orthologs, produced significant cross-talk (Figure 3.5F-G).

Taken together, our *in vitro* studies support the notion that alanine, serine, and phenylalanine at specificity positions 1, 2, and 4 represent adaptive mutations that prevent cross-talk to NtrX in α-proteobacteria.

Avoidance of cross-talk is a significant selective pressure

To test whether these mutations also prevent cross-talk *in vivo*, we engineered the chromosomal copy of *phoR* in the α-proteobacterium *C. crescentus* to produce a mutant PhoR in which specificity positions 1 and 2 are threonine and valine, respectively, as they are in γ-proteobacteria; hereafter this mutant strain is referred to as PhoR(TV). Based on our *in vitro* experiments, we expected that cross-talk from PhoR(TV) to NtrX would be induced during growth in phosphate-limited media (Figure 3.6A). During growth in such conditions, wild-type PhoR is stimulated to autophosphorylate and phosphotransfer to PhoB, which then activates genes involved in responding to phosphate limitation. Thus, any effects of increased cross-talk to NtrX by the PhoR(TV) kinase should be manifest specifically during growth in phosphate-limited media. We grew cells to mid-logarithmic phase in phosphate-limited media and measured the rate of growth by monitoring the accumulation of optical density at 600 nm. In minimal media containing either 50 μM phosphate or 5 μM phosphate, the PhoR(TV) mutant grew significantly more slowly than

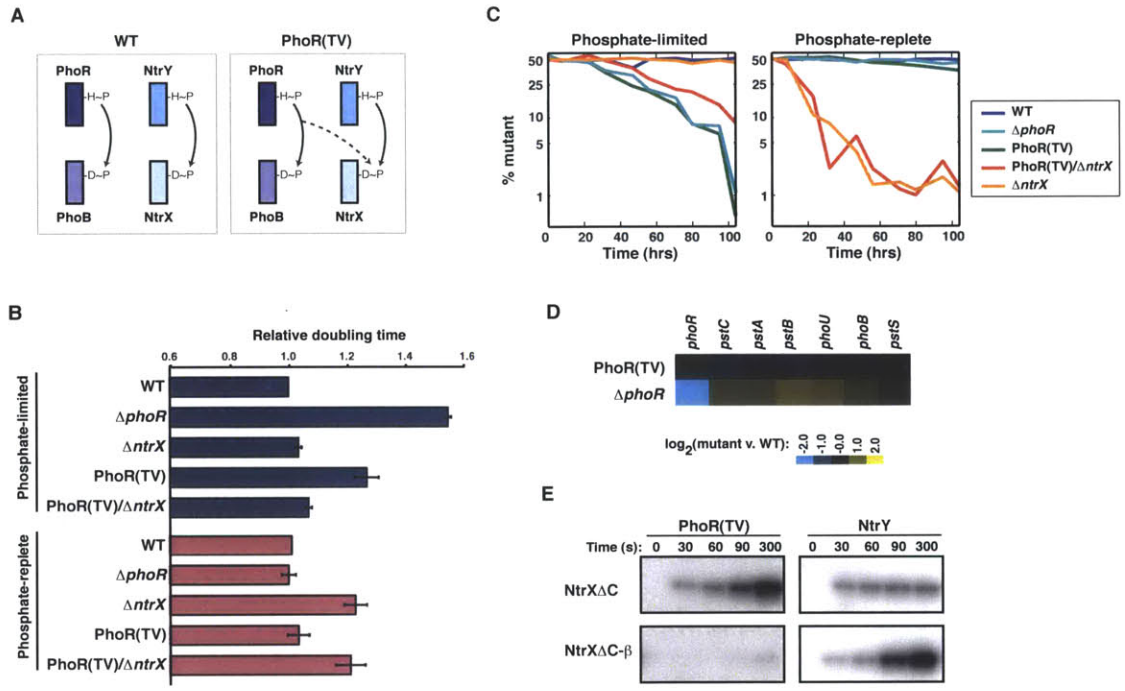


Figure 3.6 Cross-talk between PhoR(TV) and NtrX leads to a growth defect and fitness disadvantage in phosphate-limited media.

(A) Schematic of strains examined. In the wild type, PhoR-PhoB and NtrY-NtrX are insulated, whereas the PhoR(TV) mutant leads to cross-talk between PhoR and NtrX. (B) Doubling times of *C. crescentus* strains Δ*phoR*, Δ*ntrX*, PhoR(TV), and PhoR(TV)/Δ*ntrX* relative to wild type in M5G (phosphate-limited) and M2G (phosphate-replete) media. Error bars indicate standard errors, n=3. For growth curves in M8G medium (5 μM), see Figure 3.7A-B. (C) Wild type, Δ*phoR*, Δ*ntrX*, PhoR(TV), and PhoR(TV)/Δ*ntrX* were each competed against the wild type in M2GX and M5GX. The percentage of mutant cells in the population was measured periodically for 104 hours. Curves represent the average of two independent competitions with swapped fluorophores. Also see Figure 3.7C-D. (D) Expression data for known members of the *pho* regulon in PhoR(TV) and Δ*phoR* in M2G (phosphate-replete) media. Data are expressed as $\log_2$ values of the ratio between a given mutant and wild-type *C. crescentus*, and are color-coded according to the legend. (E) Time courses of phosphotransfer from kinases PhoR(TV) and NtrY to the regulators NtrXΔC and an NtrXΔC harboring β-like specificity substitutions. Only the response regulator band is shown.

wild type, with a doubling time ~30% longer than wild type in each case (Figures 3.6B, 3.7A-B). This growth defect was almost as severe as that observed for a Δ*phoR* strain which cannot mount a proper transcriptional response to phosphate-limitation. To assess whether cross-talk from PhoR(TV) to NtrX contributed to the slow growth phenotype

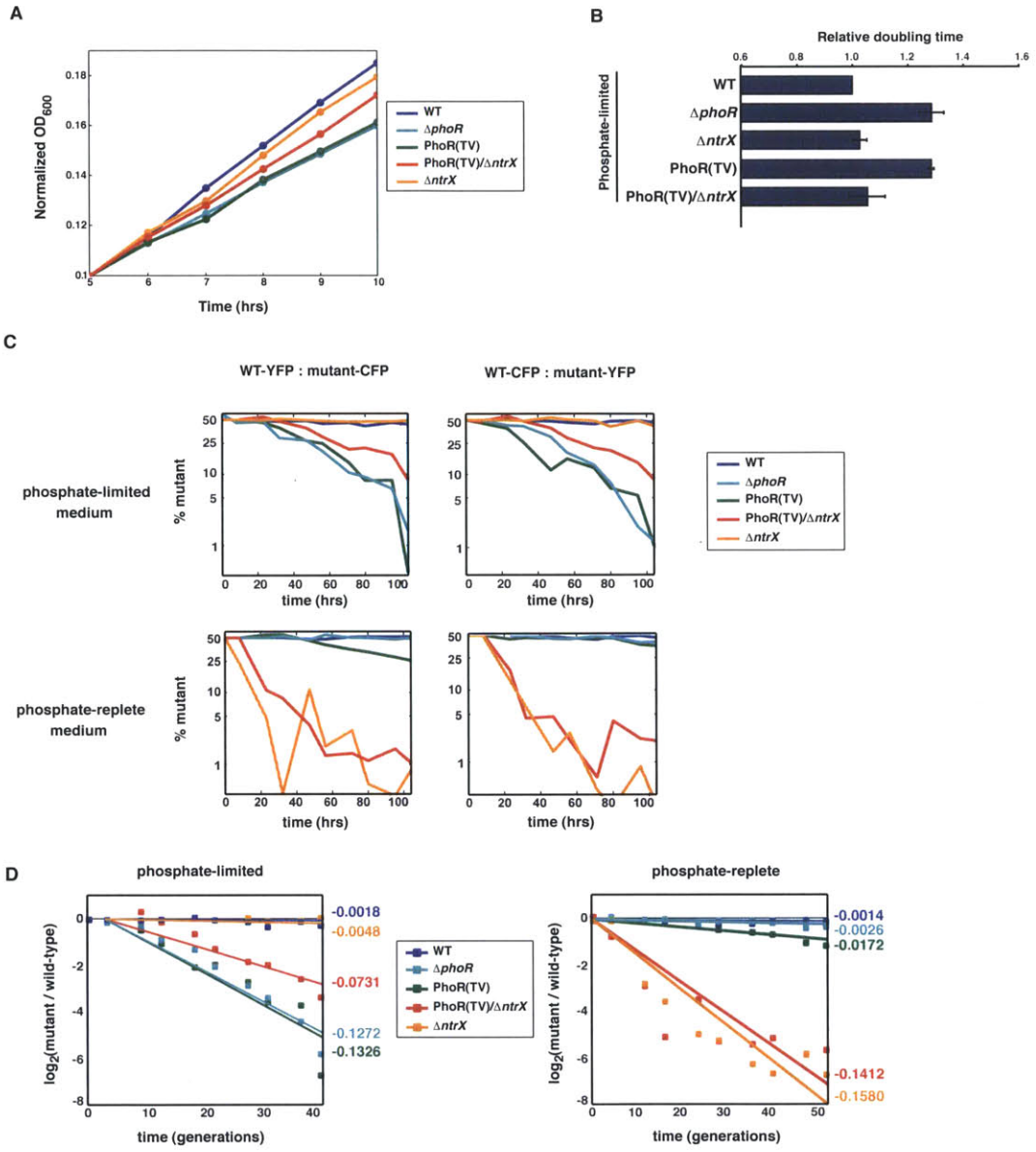


Figure 3.7 The specificity substitutions AS→TV in *C. crescentus* PhoR lead to a selective disadvantage in phosphate-limited media.

(A) Growth curves for wild type, Δ*phoR*, Δ*ntrX*, PhoR(TV), and PhoR(TV)/Δ*ntrX* in M8G, which contains 5 μM phosphate. Data points represent the average of 3 replicates. (B) Relative doubling times calculated for the growth curves in panel A. Error bars represent standard error, n=3. (C) Wild type, Δ*phoR*, Δ*ntrX*, PhoR(TV), and PhoR(TV)/Δ*ntrX* were each competed against the wild type in M2GX (phosphate-replete) and M5GX (phosphate-limited) for 104 hours. One set of competitions used YFP-labeled wild type and CFP-labeled mutant cells, whereas the other used CFP-labeled wild type and YFP-labeled mutant cells, as indicated at the top. Growth medium is listed to the left. The identity of the mutant cells for each competition is shown in the legend. (D) The average values of the competitions shown in panel C, plotted as $\log_2$(mutant/wild-type) against the number of wild-type generations in each medium. Best-fit lines for each competition are shown. The selective coefficients, per generation, are listed on the right

side of each graph.

observed, we deleted *ntrX* in the PhoR(TV) strain. Indeed, the deletion of *ntrX* significantly reduced the growth deficiency of the PhoR(TV) mutant (Figures 3.6B, 3.7A-B) suggesting that cross-talk with NtrX contributes significantly to the slow growth phenotype of a PhoR(TV) strain. The suppression observed was not a non-specific acceleration of growth as the *ntrX* deletion alone had no effect on growth in phosphate-limited medium. In phosphate-replete medium, the PhoR(TV) mutant strains grew at a rate nearly identical to the wild type (Figure 3.6B), indicating that, as expected, cross-talk to NtrX requires PhoR to be activated as a kinase. The *ntrX* deletion and PhoR(TV)/*ΔntrX* strains grew more slowly in phosphate-replete medium, as the NtrY-NtrX pathway is likely necessary for responding to a signal or metabolite produced in M2G medium.

To corroborate our growth rate measurements, we performed competitive fitness assays in which each mutant strain was mixed with the wild type at a ratio of 1:1 and grown in the same flask for 104 hours, or approximately 40 wild-type generations. The mutant and wild-type strains were engineered to constitutively produce CFP or YFP, allowing for a rapid assessment of relative strain abundance using fluorescence microscopy. In phosphate-limited conditions, the PhoR(TV) strain showed a significant growth disadvantage, being almost completely eliminated from the population after 104 hours (Figures 3.6C, 3.7C). The fitness disadvantage of the PhoR(TV) mutant was comparable to that of *ΔphoR* competed against wild type in the same phosphate-limited medium. Consistent with our growth measurements, deleting *ntrX* in the PhoR(TV) background

improved competitive fitness (Figures 3.6C, 3.7C-D). In phosphate-replete conditions, the PhoR(TV) and *ΔphoR* mutants retained a ratio with wild type close to 1:1, demonstrating that the selective disadvantage of introducing ancestral specificity residues into PhoR likely occurs only in conditions in which PhoR is a kinase. Collectively, these data further support a model in which the α-specific substitutions in PhoR specificity residues (T→A and V→S at specificity positions 1 and 2) are selectively advantageous because they help prevent phosphotransfer cross-talk to NtrX, and perhaps other response regulators.

The growth and competitive fitness defects of PhoR(TV) in phosphate-limited media were comparable to that seen for Δ*phoR*. This similarity suggested that the detrimental effect of cross-talk in the PhoR(TV) strain stems from an inability to phosphorylate PhoB and activate PhoB-dependent genes in phosphate-limited conditions. To test this hypothesis directly, we examined global gene expression patterns in the PhoR(TV) and Δ*phoR* strains during growth in phosphate-limited conditions. These expression profiles exhibited strong similarity with a Pearson correlation coefficient of ~0.9, supporting a model in which phosphorylation cross-talk from PhoR(TV) to NtrX comes at the expense of phosphorylating PhoB. The inappropriate phosphorylation of NtrX could also contribute to the growth defect of the PhoR(TV) mutant. However, NtrX-dependent genes (see Materials and Methods) were not significantly affected in the PhoR(TV) strain during growth in phosphate-limited conditions; NtrX-dependent genes behaved similarly in the PhoR(TV) and Δ*phoR* strains in phosphate-limited conditions. This may result from NtrY, the cognate kinase for NtrX, functioning as a phosphatase to prevent the accumulation of phosphorylated NtrX in phosphate-limited media. Consistent with this

notion, *ntrX* and *ntrY* are not required for growth in phosphate-limited media, suggesting that in this condition NtrY is likely in a phosphatase state.

Importantly, and in contrast to NtrY, PhoR functions as a kinase, not a phosphatase, in phosphate-limited media. Thus, our results indicate that the α-specific substitutions in PhoR specificity residues (T→A and V→S) impact fitness by affecting cross-talk at the level of phosphotransfer. Consistently, in phosphate-replete media, when PhoR is primarily active as a phosphatase, these substitutions had little to no effect on competitive fitness (Figure 3.6B-C, 3.7C-D). To further confirm that these substitutions do not significantly impact the phosphatase activity of PhoR, we examined global patterns of gene expression in the PhoR(TV) mutant grown in a phosphate-replete medium. Under these conditions, PhoR likely acts as a phosphatase to eliminate any errant phosphorylation of PhoB. Accordingly, the expression levels of known PhoB-dependent genes, such as *pstC*, *pstA*, and *pstB*, were modestly elevated in a Δ*phoR* strain grown in phosphate-replete medium (Figure 3.6D). By contrast, these genes were not affected, or were slightly downregulated, in the PhoR(TV) strain grown in the same phosphate-replete conditions, indicating that PhoR(TV) retains phosphatase activity *in vivo*. Collectively, our data demonstrate that the growth and fitness defect of the PhoR(TV) mutant stems from inappropriate phosphotransfer to NtrX, and perhaps other non-cognate substrates.

Different adaptive mutations prevent cross-talk in other proteobacterial clades

Our results suggest that α-proteobacteria have accumulated substitutions in PhoR that prevent unwanted cross-talk with the non-cognate substrate NtrX. There could, however, be other ways to avoid cross-talk between these systems in other clades. Like the α-proteobacteria, most β-proteobacteria encode NtrY-NtrX orthologs (Figure 3.4C). However, the β-PhoR orthologs have specificity residues at positions 1 and 2, similar to those found in γ-PhoR orthologs. This observation suggests that either the β-proteobacteria can tolerate cross-talk between PhoR and NtrX, or other mutations have emerged to prevent PhoR from phosphorylating NtrX. We favored the latter possibility as a comparison of sequence logos for the NtrX orthologs from α- and β-proteobacteria revealed differences at two critical positions (Figure 3.4C). Whereas most α-NtrX orthologs have aspartate, glycine, and lysine at specificity positions 2, 5, and 7, respectively, the β-NtrX orthologs typically have glycine, glutamate, and alanine at these same three respective positions. We speculated that the different specificity residues in a given β-NtrX may eliminate cross-talk from a β-PhoR; that is, β-proteobacteria may have evolved to avoid cross-talk by accumulating substitutions in NtrX rather than PhoR and PhoB.

To test this hypothesis, we asked whether introducing the β-NtrX specificity residues into an α-NtrX would eliminate cross-talk from α-PhoR(TV) which, as shown above, phosphotransfers to α-NtrX *in vitro* and *in vivo*. Indeed, whereas *C. crescentus* NtrX was robustly phosphorylated by PhoR(TV), a mutant NtrX harboring the β-like substitutions

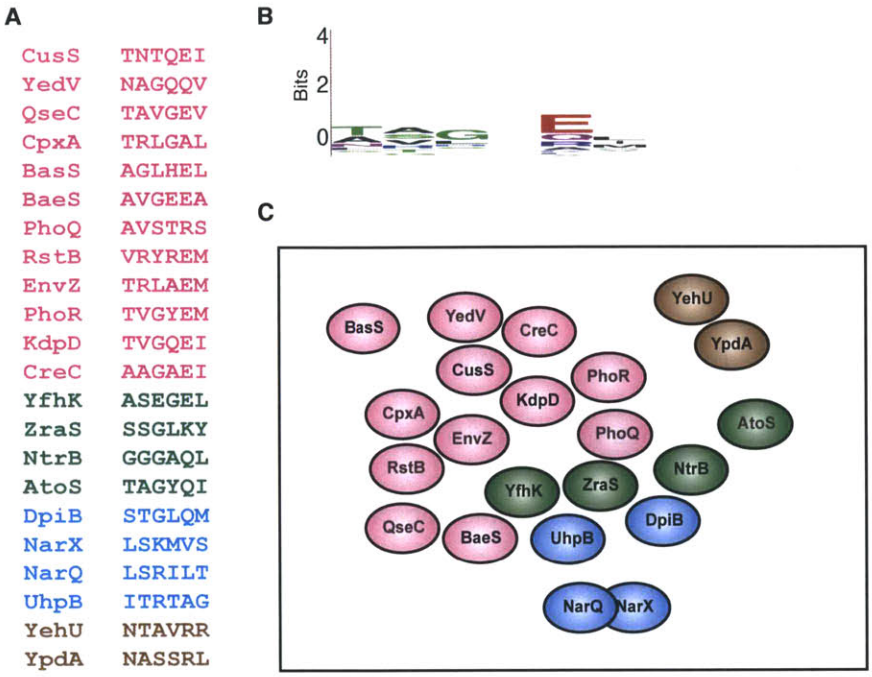


Figure 3.8 Extant two-component signaling pathways are insulated from each other at the level of phosphotransfer.

(A) The six primary specificity residues are shown for each of the 22 canonical *E. coli* histidine kinases. Hybrid histidine kinases and the non-canonical kinases DcuS and CheA were omitted. The histidine kinases are separated into groups by color based on the family of their cognate response regulator: pink, OmpR/winged helix-turn helix; green, NtrC/AAA+ and Fis domains; blue, NarL/GerE helix-turn-helix; brown, LytR. For specificity residues from *E. coli* response regulators and *C. crescentus* histidine kinases and response regulators, see Figure 3.9. (B) Sequence logo for the specificity residues in panel A. (C) A qualitative two-dimensional representation of the distribution of *E. coli* histidine kinases in the sequence space defined by the six primary specificity-determining residues. Each oval represents the set of response regulators recognized by a histidine kinase given its specificity residues. Spheres are colored using the same scheme as in panel A. With the exception of NarQ and NarX (see text), the spheres are non-overlapping, indicating a lack of cross-talk *in vivo* and *in vitro*. Kinases were placed relative to one another based roughly on their ability to phosphorylate the cognate regulators of other histidine kinases after extended incubation times *in vitro* (Skerker et al., 2005; Yamamoto et al., 2005). For example, CpxA shows a strong preference for phosphotransfer to its cognate regulator CpxR, but will phosphorylate the cognate regulators of EnvZ and RstB after extended periods of time.

D13G, G20E, F107I, and K108A was not detectably phosphorylated (Figure 3.6E). This mutant NtrX was not simply unfolded or unphosphorylatable as it was still robustly phosphorylated by α-NtrY. Hence, the substitutions introduced specifically eliminated

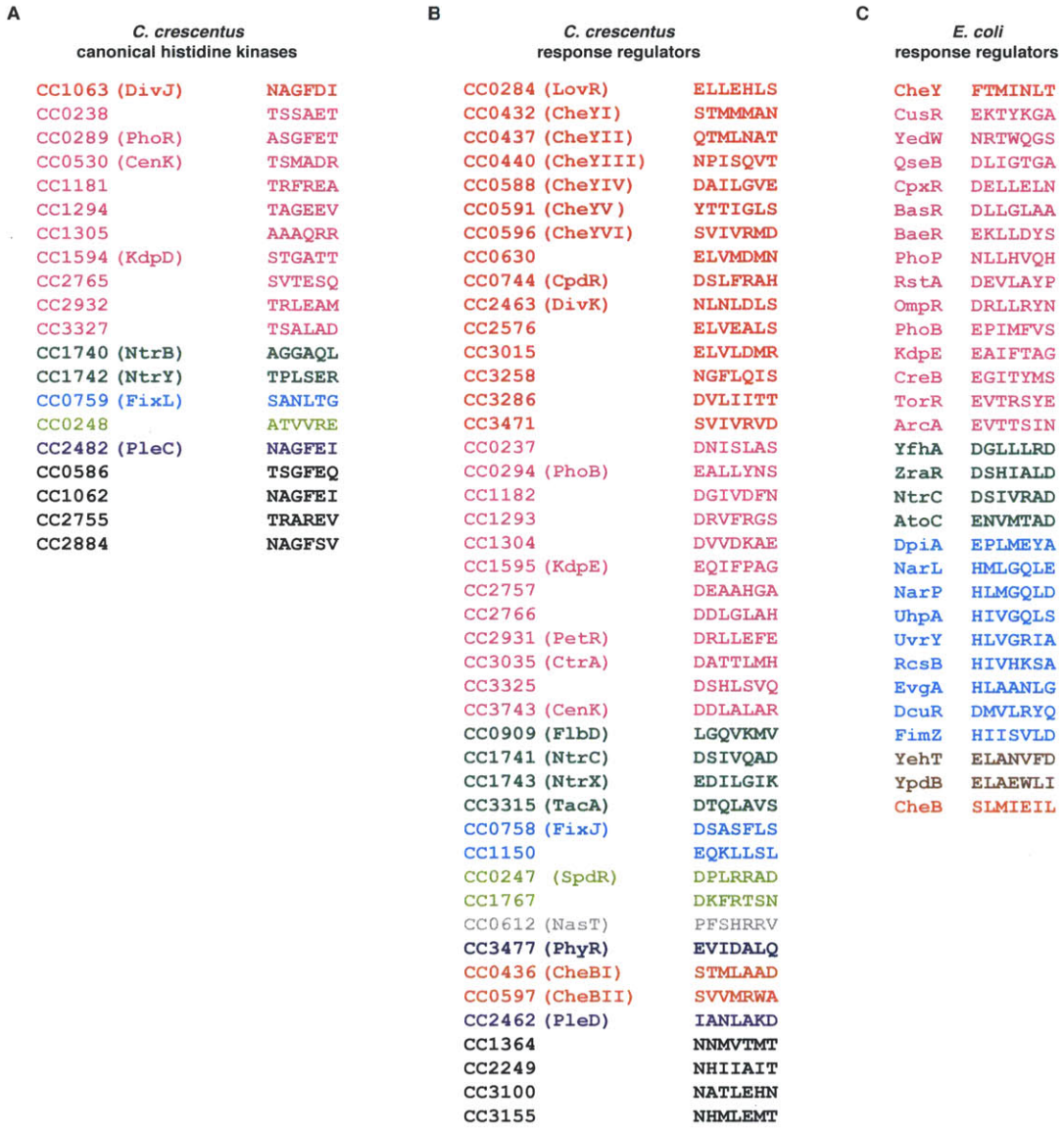


Figure 3.9 Orthogonality of specificity residues in *E. coli* and *C. crescentus* two-component signaling proteins.

(A-C) The specificity residues of (A) all canonical histidine kinases from *C. crescentus*, (B) *C. crescentus* response regulators, or (C) *E. coli* response regulators. The text for residues are colored based on the sub-family of response regulators or, in the case of kinases, by the sub-family of its cognate response regulator, if known: red, receiver domain only; pink, OmpR/winged helix-turn helix; green, NtrC/AAA+ and Fis domains; blue, NarL/GerE helix-turn-helix; light green, ActR; grey, AmiR; navy, PhyR; orange, CheB/methyltransferase; purple, GGDEF; brown, LytR; black, no known cognate for the histidine kinase or no identified sub-family for the response regulator.

cross-talk from PhoR(TV), while still allowing for interaction with the cognate kinase NtrY. Taken together, these results suggest that in β-proteobacteria, substitutions in NtrX alleviated cross-talk with PhoR while in α-proteobacteria substitutions in PhoR prevented cross-talk with NtrX. Although the substitutions are different, the net result in both cases was an insulation of the Ntr and Pho systems.

Global optimization of signaling fidelity

Our results with the Pho and Ntr signaling pathways indicate that the avoidance of cross-talk following gene duplication is a major selective pressure that drives the accumulation of adaptive substitutions in the specificity-determining residues of two-component signaling proteins. More generally, this model predicts that the specificity residues of two-component signaling proteins in extant organisms should be sufficiently different from, or orthogonal to, one another to prevent cross-talk. To test this prediction, we extracted the six major specificity residues from each of the 22 canonical histidine kinases encoded in the *E. coli* K12 genome (Figure 3.8A). Pairwise comparisons indicated that kinases typically had no more than three identities with every other kinase at these six specificity sites, often with non-conservative differences at the remaining sites. One notable exception is NarX and NarQ, which contain two identities and four conservative differences. However, these kinases, which likely arose through gene duplication, each phosphorylate the response regulators NarL and NarP *in vitro* and likely *in vivo*, and hence represent a case of physiologically beneficial cross-regulation (Noriega et al., 2010). Aside from these two kinases, there is a general pattern of orthogonality between specificity residues in the system-wide set of *E. coli* histidine kinases. This orthogonality is further reflected by a lack of information in a sequence logo built from

the specificity residues of the 22 *E. coli* histidine kinases (Figure 3.8B), particularly in comparison to the sequence logos built from orthologous histidine kinases (Figures 3.1B, 3.4C). A similar pattern of orthogonality was evident in the specificity residues of the 20 canonical histidine kinases in *C. crescentus*, as well as the specificity residues of the response regulators from both *E. coli* and *C. crescentus* (Figure 3.9). These observations, in combination with our detailed investigation of the Ntr and Pho proteins across phylogenies, suggest that the avoidance of cross-talk is a pervasive and significant selective pressure driving the system-wide insulation of two-component signaling pathways, and consequently, that in extant organisms, two-component systems are largely insulated from one another (Figure 3.8C).

Discussion

Signaling protein families, in both prokaryotes and eukaryotes, frequently expand through gene duplication (Ohno, 1970; Pires-daSilva and Sommer, 2003). The retention of the duplicated genes often requires mutations that insulate them from one another, allowing each to transmit signals without inducing cross-talk. This divergence process may be additionally constrained by a need to avoid cross-talk with other, existing members of the same protein family. For two-component pathways, the duplication-divergence process can be conceptually framed by considering the sequence space defined by the specificity-determining residues of histidine kinases (Figures 3.9-3.10). For each kinase, these residues dictate the substrates it can phosphorylate, with different kinases recognizing largely non-overlapping, or orthogonal, sets of substrates. Gene duplication leads initially to a complete overlap and requires that one or both of the duplicates accumulate changes in its specificity residues, thus separating them in sequence space (Figure 3.10).

The mutational path taken by a given kinase may cause it to infringe on the sequence space occupied by another kinase, as was likely the case with the Ntr and Pho systems in α-proteobacteria. Such overlap then necessitates additional mutations to achieve a system-wide optimization of specificity. Our results indicate that such optimization and the avoidance of cross-talk are important selective pressures influencing two-component systems and they can drive the divergent evolution of orthologous proteins.

How did the NtrY-NtrX pathway arise if cross-talk is detrimental? Although we cannot infer the order of events with complete certainty, a plausible scenario is that the NtrY-

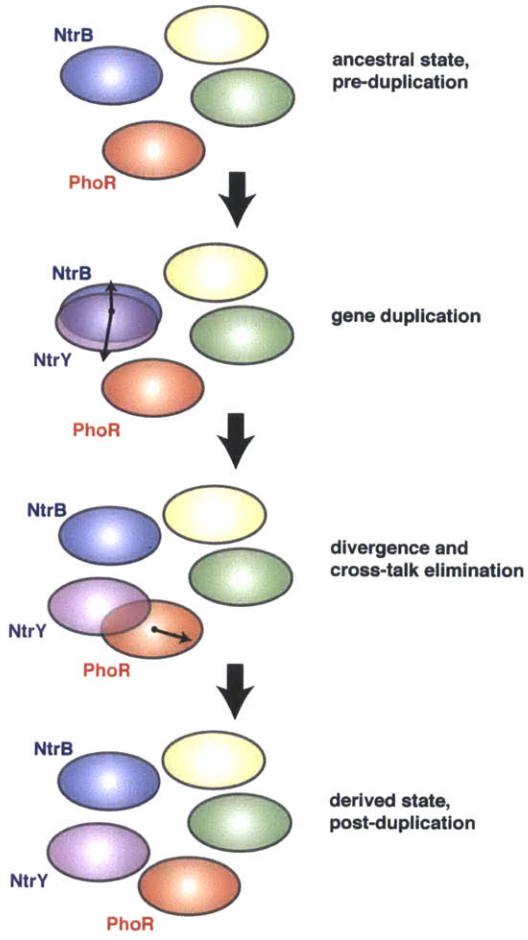


Figure 3.10 Adaptive divergence of duplicated signaling pathways involves the elimination of cross-talk.

Ovals represent the set of response regulators recognized by a histidine kinase, as determined by its specificity residues (see also Figure 3.8). The NtrB-NtrC pathway is shown duplicating to produce the paralogous system NtrY-NtrX. As these pathways diverged, the specificity of NtrY overlapped that of PhoR, necessitating a change in PhoR specificity to yield the derived state with insulated pathways.

NtrX pathway arose during growth in phosphate-replete conditions where it provided a selective advantage, as suggested by the slow growth of a Δ*ntrX* strain in these conditions. Subsequent growth in phosphate-limited conditions would then select for strains that have accumulated mutations eliminating cross-talk between PhoR and NtrX. We showed that such insulation can occur with only one or two point mutations,

supporting the plausibility of this scenario in an ancestral α-proteobacterium. Interestingly, the β-proteobacteria likely followed a different mutational path to avoiding cross-talk, accumulating substitutions in NtrX rather than PhoR. The difference between the two clades of proteobacteria may reflect the inherent stochasticity of mutations and selection. Alternatively, the growth conditions or genomic context of the ancestral organisms in which gene duplication occurred may have deterministically influenced selection.

In sum, we propose that the evolution of two-component signaling genes is characterized by long periods of stasis with specificity-determining residues subject to strong purifying selection to ensure robust phosphotransfer from kinase to regulator. Gene duplication, or lateral transfer events, can disrupt this stasis, requiring a global re-optimization of existing signaling proteins to accommodate the new pathway. The specificity residues are thus likely subject to bursts of diversifying selection; however, these residues would not necessarily exhibit commonly used signatures of diversifying selection such as large dN/dS values. Instead, our work emphasizes that a molecular-level understanding of protein evolution and the identification of adaptive mutations ultimately demands an integration of sequence analysis with focused biochemical and genetic characterizations.

Our approach and findings are relevant beyond two-component signaling as paralogous signaling protein families are found throughout biology. In fact, most organisms use a remarkably small number of types of signaling protein to carry out their diverse information-processing tasks. In all cases, duplication and divergence is a primary means by which new pathways are created and, consequently, issues of specificity and the fidelity of information transfer are critical. While eukaryotes sometimes rely on tissue-

specific expression of paralogous genes or spatial mechanisms like scaffolds to enforce specificity, many common signaling proteins and domains, such as PDZ, SH3, SH2, and bZIP proteins (Hou et al., 2009; Liu et al., 2011; Newman and Keating, 2003; Stiffler et al., 2007; Tonikian et al., 2008; Zarrinpar et al., 2003), rely on molecular recognition and a relatively small set of specificity-determining residues. Hence, our observation that pathway insulation in bacteria can be achieved with a limited number of mutations may help to explain how organisms in all domains of life have exploited gene duplication to expand and diversify their signaling repertoires.

Materials and Methods

Identification of orthologs and construction of gene trees

A modified version of reciprocal best blast hits was used to identify orthologous proteins. For *E. coli* PhoR, the DHp domain was used as a query in BLAST searches against fully sequenced bacterial genomes in GenBank (September 2009). The top ten hits from each genome were then subjected to reciprocal BLAST searches against the *E. coli* MG1655 genome. If only the top hit identified *E. coli* PhoR as the best match, it was called as a PhoR ortholog. If multiple hits identified *E. coli* PhoR as the best match, the top hit was called as an ortholog and additional hits were evaluated as follows. If an additional hit had an E-value within 10^3 of the top hit and was closer to the top hit than to the fifth hit (which generally had an E-value reflecting the overall paralogous relationship of histidine kinases), we also called it as an ortholog of *E. coli* PhoR and examined the next hit similarly. For genomes with more than one hit called as an ortholog, duplications were inferred and each hit deemed a member of the PhoR orthogroup. A similar procedure was followed to identify orthologs of PhoB, NtrB, NtrC, NtrX, and NtrY using as query sequences the *E. coli* PhoB receiver domain, the *C. crescentus* NtrB and NtrY DHp domains, or the *C. crescentus* NtrC and NtrX receiver domains.

Orthologous sequences were aligned using ClustalX (Chenna et al., 2003). Sequence logos for the specificity residues, extracted from the aligned sequences, were built using WebLogo (Crooks et al., 2004). To help correct for phylogenetic biases in genome sequencing efforts, sequences were filtered to ensure that no two sequences were more than 95% identical.

PhoR DHp domains and PhoB receiver domains were extracted using HMMER with models for a His_KA domain or REC domain, respectively (Wistrand and Sonnhammer, 2005), and used to build gene trees through the PHYLIP package (Felsenstein, 1989) using the neighbor-joining algorithm provided. The tree was rooted using *B. subtilis* PhoR as the outgroup. Reported bootstrap values are out of 1000.

For genome-wide analyses of specificity residues, only canonical histidine kinases were included. Canonical kinases were defined as those containing the PFAM HisKA domain and no REC domain.

Growth conditions and strain construction

C. crescentus cells were grown at 30°C in PYE, M2G (10 mM phosphate), M5G (50 μM phosphate), or M8G (same as M5G but with 5 μM phosphate), supplemented when necessary with oxytetracycline (1 μg/ml), kanamycin (25 μg/ml), 0.2% glucose or 0.3% xylose. *E. coli* strains were grown at 37°C in LB supplemented with carbenicillin (100 μg/ml) or kanamycin (50 μg/ml). Transductions were performed using ΦCr30 (Ely, 1991).

Table 3.1 –
Strains and plasmids

Name	Description	Source
E. coli		
BL21-tuner	*E. coli* strain for protein expression and purification	Novagen
DH5α	*E. coli* general cloning strain	Invitrogen
C. crescentus		
CB15N	synchronizable derivative of wild-type CB15	(Evinger and Agabian, 1977)

ΔCC0289	Δ*phoR*	(Skerker et al., 2005)
ΔCC1743	Δ*ntrX*	(Skerker et al., 2005)
ML1934	PhoR(TV)	this study
ML1935	PhoR(TV) Δ*ntrX*	this study
ML1936	P_{xyl}-*yfp-xylX*	this study
ML1937	P_{xyl}-*cfp-xylX*	this study
ML1938	PhoR(TV) P_{xyl}-*cfp-xylX*	this study
ML1939	PhoR(TV), P_{xyl}-*yfp-xylX*	this study
ML1940	Δ*phoR* P_{xyl}-*cfp-xylX*	this study
ML1941	Δ*phoR* P_{xyl}-*yfp-xylX*	this study
ML1942	Δ*ntrX* P_{xyl}-*cfp-xylX*	this study
ML1943	Δ*ntrX* P_{xyl}-*yfp-xylX*	this study
ML1944	PhoR(TV) Δ*ntrX* P_{xyl}-*cfp-xylX*	this study
ML1945	PhoR(TV) Δ*ntrX* P_{xyl}-*yfp-xylX*	this study
General purpose vectors		
pENTR/D-TOPO	ENTRY vector for Gateway cloning system (kan^R)	Invitrogen
pML310	Destination vector, TRX-HIS_6 (kan^R)	(Skerker et al., 2005)
pML333	Destination vector, HIS_6-MBP (kan^R)	(Skerker et al., 2005)
pNPTS138	integration vector ($sacB^S$, kan^R)	(Skerker et al., 2005)
pXCFPN-4	P_{xyl}-*cfp*	(Thanbichler et al., 2007)
pXYFPN-4	P_{xyl}-*yfp*	(Thanbichler et al., 2007)
Plasmids		
pXCFPN-4:P_{xyl}-*cfp-xylX*	intermediate cloning plasmid	this study
pXYFPN-4:P_{xyl}-*yfp-xylX*	intermediate cloning plasmid	this study
pNPTS138:P_{xyl}-*yfp-xylX*	integration of YFP behind the xylose promoter	this study
pNPTS138:P_{xyl}-*cfp-xylX*	integration of CFP behind the xylose promoter	this study
pNPTS138:PhoR(TV)	allelic replacement of *C. crescentus phoR* with *phoR*(TV)	this study
Protein expression plasmids		
pML333:PhoR_*Cc*	*C. crescentus* HIS_6-MBP-PhoR	this study
pML333:PhoR(T)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(T)	this study
pML333:PhoR(V)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(V)	this study
pML333:PhoR(Y))_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(Y)	this study
pML333:PhoR(TV)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(TV)	this study
pML333:PhoR(VY)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(VY)	this study
pML333:PhoR(TY)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(TY)	this study
pML333:PhoR(TVY)_*Cc*	*C. crescentus* HIS_6-MBP-PhoR(TVY)	this study
pML333:PhoR_*Ec*	*E. coli* HIS_6-MBP-PhoR	this study
pML333:NtrB	*C. crescentus* HIS_6-MBP-NtrB	this study
pML333:NtrY	*C. crescentus* HIS_6-MBP-NtrY	this study
pML310-NtrXΔC-β	*C. crescentus* TRX-HIS_6-NtrXΔC(D13G, G20E, F107I, K108A)	this study

pML310:CC1741ΔC	*C. crescentus* TRX-HIS$_6$-NtrCΔC	this study
pML310:CC1743ΔC	*C. crescentus* TRX-HIS$_6$-NtrXΔC	this study
pML310:CC3743ΔC	*C. crescentus* TRX-HIS$_6$-CenRΔC	this study
pML310:b0399	*E. coli MG1655* TRX-HIS$_6$-PhoBΔC	this study
pML310:VF1988	*V. fischeri ES114* TRX-HIS$_6$-PhoBΔC	this study
pML310:SO1558	*S. oneidensis MR-1* TRX-HIS$_6$-PhoBΔC	this study
pML310:PA5360	*P. aeruginosa PAO1* TRX-HIS$_6$-PhoBΔC	this study
pML310:BTHI2768	*B. thailandensis E264* TRX-HIS$_6$-PhoBΔC	this study
pML310:RSc1534	*R. solanacearum GMI1000* TRX-HIS$_6$-PhoBΔC	this study
pML310:CC0294	*C. crescentus CB15* TRX-HIS$_6$-PhoBΔC	this study
pML310:amb1370	*M. magneticum sp AMB-1* TRX-HIS$_6$-PhoBΔC	this study
pML310:SMc02140	*S. meliloti 1021* TRX-HIS$_6$-PhoBΔC	this study
pML310:RSP2599	*R. sphaeroides 2.4.1* TRX-HIS$_6$-PhoBΔC	this study
pML310:ZMO1164	*Z. mobilis ZM4* TRX-HIS$_6$-PhoBΔC	this study

Strains used are listed in Table 3.1. To construct ML1934, PhoR(TV), a region from nucleotide position -30 (relative to the PhoR start codon) to position 168 was amplified using CB15N genomic DNA as template and the primers PhoR_int_upstream_for and PhoR_int_upstream_rev, and a region from 147 to 1416 was amplified using pENTR-CC_PhoR(TV) as template and the primers PhoR_int_for and PhoR_int_rev. Primer sequences are listed in Table 3.2. The two amplicons were then fused using SOE-PCR and ligated into pNPTS138, which had been cut with EcoRV and phosphatased using SAP, to create pNPTS-CC_PhoR(TV). This plasmid was used for allelic replacement in CB15N following procedures described previously (Skerker et al., 2005). Integrants were tested for kanamycin sensitivity, sucrose resistance, and sequence-verified using primers listed in table S8. To create ML1935, PhoR(TV)/Δ*ntrX*, a tetracycline-marked *ntrX* deletion was transduced into ML1934; transductants were verified by PCR.

Strains used for competition assays (ML1936-ML1945) contained either the coding region for CFP or YFP driven by the *xylX* promoter. *xylX* was amplified from CB15N genomic DNA using primers xylX_for and xylX_rev, digested with KpnI and AgeI and

ligated into pXCFPN-4 and pXYFPN-4 using the same restriction sites, producing pXCFPN-4:P_{xyl}-*cfp-xylX* and pXYFPN-4:P_{xyl}-*yfp-xylX*. The inserts were then amplified using primers xyl_xfp_for and xyl_xfp_rev, digested with HindIII and EcoRI, and ligated into pNPTS138 digested with the same enzymes. These vectors, pNPTS138:P_{xyl}-*cfp-xylX* and pNPTS138:P_{xyl}-*yfp-xylX*, were then integrated into the chromosomes of CB15N, Δ*phoR*, Δ*ntrX*, PhoR(TV), and PhoR(TV)/Δ*ntrX* through transformation and selection on kanamycin followed by counterselection on sucrose, leading to markerless integrations of CFP or YFP at the native *xylX* locus of each strain.

Table 3.2 –
Primers

Genome	Accession number	Forward primer	Reverse primer
E. coli MG1655	NP_414933.1	CACCTTGGCGAGACGTATTCTGGTC	TTACGCCATTGGCGAAATACG
V. fischeri ES114	YP_205371.1	CACCTTGGCTAGAAGGATCCTTGTTGT AGAAGATGAAG	TTATGATGTTGGAGAGACACGACGAATT AC
S. oneidensis MR-1	NP_717171.1	CACCTTGTTGACATTCACTGACAGAAAA TC	CTATTTGATGCGGGCAACTAGCTC
P. aeruginosa PAO1	NP_254047.1	CACCTTGGTTGGCAAGACAATCCTCAT CGTTGATG	TTAGTCGCCAGGCCCGGTGCG
B. thailandensis E264	YP_443280.1	CACCTTGCCCAGCAACATTCTC	TTACTTGATCCGCGCCATCAG
R. solanacearum GMI1000	NP_519655.1	CACCTTGCCGAGCATATTCTG	TTACTTGATGCGGGCAAGCAAC
C. crescentus CB15	NP419113.1	CACCTTGACTCCCTACGTTTTGGTGGT CGAAGAC	TTACAGACCCGGGCGGATGCG
M. magneticum sp AMB-1	YP_420733.1	CACCTTGACCGCCCGCGAGACCGCC	CTAGCGGACGCGGGCCATCAGTTC
S. meliloti 1021	NP_384621.1	CACCTTGTTGCCGAAGATTGCCGTAGT CG	TTAAACCTCGGGCTTGGCGCG
R. sphaeroides 2.4.1	YP_352657.1	CACCTTGTCGCCCGCCGATCAGCCCG	CTAGCGCACCCGCGCCATCAGTTC
Z. mobilis ZM4	YP_162899.1	CACCTTGGCCGCTTTACGGCTACTACT C	CTAAAGGACTCTTGCAACCAGCTC
C. crescentus	CC_0289, *phoR*	CACCCTGAACCGGCGAAAGGCCGT	TCAGGCGCTTCCCGCTTCCGCC
C. crescentus	CC_1742, *ntrY*	CACCTTCGGCGTGCTGGTCAACCG	TCATATCATCTCCTCAACGCCA
C. crescentus	CC_1740, *ntrB*	CACCTTGGCCACCGAAGCTCTGAAA	TCATGCTCGGACGTCTCCGGAA
C. crescentus	CC_1741, *ntrCΔC*	CACCTTGAACGCCGCGAGCAAGAAA	TCAAGTGTCCGCCGGCCGCGACAA
C. crescentus	CC_3743, *cenRΔC*	CACCTTGTTCCGGCTTATGGCGCAACG C	TCATGAAGCCTCGTAGCTGCGCAGCTG C
E. coli	b_3868, *ntrCΔC*	CACCTTGCAACGAGGGATAGTCTGG	TCACTGGTAATGACTGATAGCGCG
E. coli	b_0400, *phoR*	CACCAATCTGGTGCTCAACACCGG	TGGCTAATCATGCGAACAAA
Primers for site directed mutagenesis			

Name	Template	Forward primer	Reverse primer
CC_PhoR(TV)	pENTR-CC_PhoR	CGCACGCCGCTCACCGTGTTGTCCGGCTTC	GAAGCCGGACAACACGGTGAGCGGCGTGCG
CC_PhoR(T)	pENTR-CC_PhoR	CTGCGCACGCCGCTCACCTCGTTGTCCGGC	GCCGGACAACGAGGTGAGCGGCGTGCGCAG
CC_PhoR(V)	pENTR-CC_PhoR	ACGCCGCTCGCCGTGTTGTCCGGCTTC	GAAGCCGGACAACACGGCGAGCGGCGT
CC_PhoR(Y)	pENTR-CC_PhoR, PhoR(TV), PhoR(T)	TTGTCCGGCTACATCGAGACC	GGTCTCGATGTAGCCGGACAA
CC_PhoR(VY)	pENTR-CC_PhoR	GTGTTGTCCGGCTACATCGAGACCCTG	CAGGGTCTCGATGTAGCCGGACAACAC
CC_NtrXΔC	pENTR-CC_NtrX	CACCTTGAACGCCGCGAGCAAGAAA	TCAAGTGTCCGCCGGCCGCGACAA
CC_NtrX_D13G	pENTR-CC_NtrXΔC	GTGGATGACGAGGCCGGCATTCGGGATCTCGTC	GACGAGATCCCGAATGCCGGCCTCGTCATCCAC
CC_NtrX_G20E	pENTR-CC_NtrXΔC(D13G)	CGGGATCTCGTCGCCGAAATCCTGGAGGATGAA	TTCATCCTCCAGGATTTCGGCGACGAGATCCCG
CC_NtrX_F107I_K108A	pENTR-CC_NtrXΔC(D13G,G20E)	GAGTTCCTCGAAAAGCCGATCGCATCGGACCGGCTTTTGCTG	CAGCAAAAGCCGGTCCGATGCGATCGGCTTTTCGAGGAACTC

Primers for strain construction

Name	Template	Forward primer	Reverse primer
PhoR_int_upstream	CB15N genomic	GCCTCTCGCTTGAATCGGTGAAGCTC	GAAAACGCCTGCGCCGCCCAC
PhoR_int	pENTR-CC_PhoR(TV)	GTGGGCGGCGCAGGCGTTTTCCTGAACCGGCGAAAGGCCGT	TCAGGCGCTTCCCGCTTCCGC
xylX	CB15N genomic	CAGCAGGGTACCTAAGTGGGCGTGAGTGAATTCCT	CAGCAGACCGGTTTAGAGGAGGCCGCGGCCGG
xylX_xfp	pXCFPN-4:P_{xyl}-*cfp-xylX*, pXYFPN-4:P_{xyl}-*yfp-xylX*	CAGCAGAAGCTTCTTGGCCGGCGGCTTGACCT	CAGCAGGAATTCTTAGAGGAGGCCGCGGCCGG
NtrX_tet_conf	Δ*ntrX*, PhoR(TV)/Δ*ntrX* genomic	GCACGTGTGGAAGTCGAGC	CGTTGAAGGACCGAGAAAGG
PhoR_int_sequence	PhoR(TV) genomic	CTCGCGACACGCACTAAGGC	GGCGCTGTGCTTATGCGAC

Expression vectors were built by moving pENTR clones into destination vectors using the Gateway LR reaction (Invitrogen), and then transformed into BL21 *E. coli* for expression and purification. All site-directed mutagenesis was done on pENTR clones using primers listed in Table 3.2 and sequence-verified.

Protein purification and phosphotransfer assays

Expression, protein purification, and phosphotransfer experiments were carried out as described previously (Skerker et al., 2005). Phosphotransfer profiles against all *C. crescentus* regulators comprise three gels, which were run in parallel and exposed to the same phosphorscreen. Gel images were then stitched together for presentation. Profiles used full-length response regulators except for CC1741 (*ntrC*), CC1743 (*ntrX*), and CC3743 (*cenR*) for which only receiver domains were used. For time courses of phosphotransfer in Figure 3.1C, each PhoB construct contained only the receiver domain.

Growth and competitive fitness assays

Cultures were grown overnight in M2G. Cultures were then diluted to $OD_{600} \sim 0.025$ and resuspended in either M2G or M5G. Samples were taken every hour and growth rates calculated 8-14 hours post-dilution to ensure phosphate-limitation of cells grown in M5G. For more severe phosphate limitation, growth curves were repeated in M8G, which contains 10-fold less phosphate. For these experiments, cultures were grown overnight in M5G and resuspended at an $OD_{600} \sim 0.07$ in M8G.

For competitive fitness assays, cultures were grown overnight in M2GX, resuspended in either M2GX or M5GX to an OD_{600} of 0.05, and mixed 1:1 with a competitor strain in 10 mL of media in a 150 mL flask. After 9 hours, a sample was taken and fixed using paraformaldehyde. Cells were then diluted to $OD_{600} \sim 0.01$ in 10 mL of M2GX or M5GX. After 15 hours, another sample was taken and cells diluted to $OD_{600} \sim 0.05$. After 9 hours, another sample was taken and cells diluted to $OD_{600} \sim 0.01$. This growth and dilution process was repeated until 104 total hours had elapsed. Cultures typically

remained below OD_{600} ~ 0.85 at all times. Cells from each sample collected were immobilized on 1.5% agarose pads made with PBS, and imaged using a Zeiss Axiovert 200 microscope with a 100x objective. Multiple fields of CFP, YFP, and phase images were taken for each sample. Roughly 500 cells were counted for each time point using a custom MATLAB script with counts checked manually using ImageJ. Competition experiments were done once with wild type expressing *cfp* and mutant expressing *yfp*, repeated with fluorescent proteins swapped, and results averaged.

Microarray analysis

Cultures were grown to mid-log phase in M2G and either RNA was harvested or cells were washed, resuspended in M5G, and grown for 11 hours in phosphate-limited conditions before RNA was harvested. RNA was extracted, labeled, and hybridized to custom-designed 8x15K Agilent expression arrays as described previously (Gora et al., 2010). NtrX-dependent genes were defined as those genes exhibiting at least a 4-fold decrease in expression in the *ΔntrX* strain compared to wild-type *C. crescentus* in M2G. Complete array data are deposited in GEO.

Acknowledgements

We thank O. Ashenberg for help with bioinformatics, Y.E. Chen for help with strain construction, and A. Podgornaia, A. Keating, O. Ashenberg, and K. Foster for helpful comments on the manuscript. Sequence analyses were performed on a computer cluster supported by NSF grant 0821391. M.T.L. is an Early Career Scientist of the Howard Hughes Medical Institute. This work was supported by an NSF graduate fellowship to E.J.C and an NSF CAREER award to M.T.L.

References

Alm, E., Huang, K., and Arkin, A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol *2*, e143.

Bell, C.H., Porter, S.L., Strawson, A., Stuart, D.I., and Armitage, J.P. (2010). Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex. PLoS Biol *8*, e1000306.

Capra, E.J., Perchuk, B.S., Lubin, E.A., Ashenberg, O., Skerker, J.M., and Laub, M.T. (2010). Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet *6*, e1001220.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res *31*, 3497-3500.

Crooks, G., Hon, G., Chandonia, J., and Brenner, S. (2004). WebLogo: a sequence logo generator. Genome Research *14*, 1188-1190.

Dean, A.M., and Thornton, J.W. (2007). Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet *8*, 675-688.

Ely, B. (1991). Genetics of Caulobacter crescentus. Methods Enzymol *204*, 372-384.

Evinger, M., and Agabian, N. (1977). Envelope-associated nucleoid from Caulobacter crescentus stalked and swarmer cells. J Bacteriol *132*, 294-301.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics *5*, 164-166.

Gora, K.G., Tsokos, C.G., Chen, Y.E., Srinivasan, B.S., Perchuk, B.S., and Laub, M.T. (2010). A cell-type-specific protein-protein interaction modulates transcriptional activity of a master regulator in Caulobacter crescentus. Mol Cell *39*, 455-467.

Grimshaw, C.E., Huang, S., Hanstein, C.G., Strauch, M.A., Burbulys, D., Wang, L., Hoch, J.A., and Whiteley, J.M. (1998). Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in Bacillus subtilis. Biochemistry *37*, 1365-1375.

Hou, T., Xu, Z., Zhang, W., McLaughlin, W.A., Case, D.A., Xu, Y., and Wang, W. (2009). Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Molecular & Cellular Proteomics *8*, 639-649.

Huynh, T.N., and Stewart, V. (2011). Negative control in two-component signal transduction by transmitter phosphatase activity. Mol Microbiol *82*, 275-286.

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annu Rev Genet *41*, 121-145.

Liu, B.A., Shah, E., Jablonowski, K., Stergachis, A., Engelmann, B., and Nash, P.D. (2011). The SH2 domain-containing proteins in 21 species establish the provenance and scope of phosphotyrosine signaling in eukaryotes. Sci Signal *4*, ra83.

Newman, J.R., and Keating, A.E. (2003). Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science *300*, 2097-2101.

Noriega, C.E., Lin, H.Y., Chen, L.L., Williams, S.B., and Stewart, V. (2010). Asymmetric cross-regulation between the nitrate-responsive NarX-NarL and NarQ-NarP two-component regulatory systems from Escherichia coli K-12. Mol Microbiol *75*, 394-412.

Ohno, S. (1970). Evolution by Gene Duplication (New York: Springer).

Pires-daSilva, A., and Sommer, R.J. (2003). The evolution of signalling pathways in animal development. Nat Rev Genet *4*, 39-49.

Siryaporn, A., and Goulian, M. (2008). Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in E. coli. Mol Microbiol *70*, 494-506.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Stiffler, M.A., Chen, J.R., Grantcharova, V.P., Lei, Y., Fuchs, D., Allen, J.E., Zaslavskaia, L.A., and MacBeath, G. (2007). PDZ domain binding selectivity is optimized across the mouse proteome. Science *317*, 364-369.

Stock, A., Robinson, V., and Goudreau, P. (2000). Two-component signal transduction. Annual Review of Biochemistry *69*, 183-215.

Thanbichler, M., Iniesta, A.A., and Shapiro, L. (2007). A comprehensive set of plasmids for vanillate- and xylose-inducible gene expression in Caulobacter crescentus. Nucleic Acids Res *35*, e137.

Tonikian, R., Zhang, Y., Sazinsky, S.L., Currell, B., Yeh, J.H., Reva, B., Held, H.A., Appleton, B.A., Evangelista, M., Wu, Y., *et al.* (2008). A specificity map for the PDZ domain family. PLoS Biol *6*, e239.

Wanner, B.L., and Chang, B.D. (1987). The *phoBR* operon in *Escherichiacoli* K-12. J Bacteriol *169*, 5569-5574.

Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., and Hwa, T. (2009). Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A *106*, 67-72.

Wistrand, M., and Sonnhammer, E. (2005). Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. BMC Bioinformatics *6*, 99.

Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R., and Ishihama, A. (2005). Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. J Biol Chem *280*, 1448-1456.

Yang, Z., and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. Trends Ecol Evol *15*, 496-503.

Zarrinpar, A., Park, S.H., and Lim, W.A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature *426*, 676-680.

Chapter 4

Spatial tethering of kinases to their substrates relaxes evolutionary constraints on specificity

This work was published as Emily J. Capra, Barrett S. Perchuk, Orr Ashenberg, Charlotte A. Seid, Hana R. Snow, Jeffrey M. Skerker, Michael T. Laub. 2012. Mol Microbiol. Dec;86(6):1393-403.

EJC and MTL conceived and designed the experiments. EJC performed most of the experiments. BSP helped with the protein purifications and the profiles shown in figures 4.3, 4.4 and 4.5. CAS, HRS, and JMS contributed reagents and helped with preliminary experiments. OA and MTL performed the computational analysis. EJC and MTL wrote the paper.

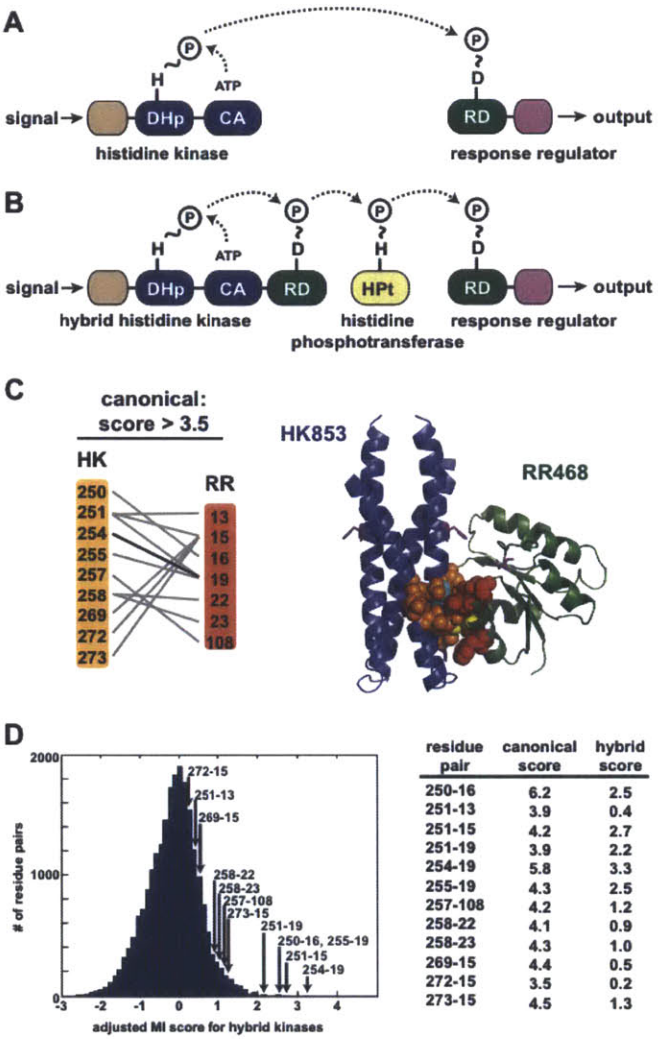
Abstract

Signal transduction proteins are often multidomain proteins that arose through the fusion of previously independent proteins. How such a change in the spatial arrangement of proteins impacts their evolution and the selective pressures acting on individual residues is largely unknown. We explored this problem in the context of bacterial two-component signaling pathways, which typically involve a sensor histidine kinase that specifically phosphorylates a single cognate response regulator. Although usually found as separate proteins, these proteins are sometimes fused into a so-called hybrid histidine kinase. Here, we demonstrate that the isolated kinase domains of hybrid kinases exhibit a dramatic reduction in phosphotransfer specificity *in vitro* relative to canonical histidine kinases. However, hybrid kinases phosphotransfer almost exclusively to their covalently attached response regulator domain, whose effective concentration exceeds that of all soluble response regulators. These findings indicate that the fused response regulator in a hybrid kinase normally prevents detrimental cross-talk between pathways. More generally, our results shed light on how the spatial properties of signaling pathways can significantly affect their evolution, with additional implications for the design of synthetic signaling systems.

Introduction

Cells can sense and respond to a remarkable diversity of signals and stimuli. This sensory capability typically involves a limited number of signal transduction protein families that have expanded through gene duplication. Although the relative ease of duplication and divergence has enabled cells to dramatically expand their signaling repertoires, the use of highly related signaling proteins has a significant cost, or risk. Cells must avoid detrimental cross-talk and ensure the fidelity of information flow through different signaling pathways. How the specificity of each signaling pathway is determined and how it evolves following gene duplication events are important problems that remain incompletely understood.

In bacteria, the dominant form of signal transduction is known as two-component signaling and typically involves a sensor histidine kinase that can autophosphorylate and then transfer its phosphoryl group to a cognate response regulator, which effects changes in cellular physiology or behavior (Stock et al., 2000) (Figure 4.1A). Two-component signaling genes have undergone extensive duplication and horizontal transfer, such that most species possess tens or hundreds of these pathways (Galperin, 2005). Previous work has shown that the interaction between a histidine kinase and its cognate response regulator is highly specific with limited cross-talk between pathways *in vivo* (Capra et al., 2012; Fisher et al., 1996; Grimshaw et al., 1998; Laub and Goulian, 2007; Skerker et al., 2005). This specificity is determined predominantly at the level of molecular recognition rather than relying on cellular factors such as scaffolds. Consequently, a histidine kinase preferentially phosphorylates its cognate response regulator *in vitro*, relative to all other response regulators (Skerker et al., 2005).



residue pair	canonical score	hybrid score
250-16	6.2	2.5
251-13	3.9	0.4
251-15	4.2	2.7
251-19	3.9	2.2
254-19	5.8	3.3
255-19	4.3	2.5
257-108	4.2	1.2
258-22	4.1	0.9
258-23	4.3	1.0
269-15	4.4	0.5
272-15	3.5	0.2
273-15	4.5	1.3

Figure 4.1 Amino acid coevolution analysis of hybrid histidine kinases.

(A) Diagram of canonical two-component signaling pathways and (B) phosphorelays, indicating the conserved domains in each protein. (C) Coevolving residues in cognate pairs of canonical histidine kinases and response regulators. Residue pairs with adjusted mutual information scores greater than 3.5 are listed, connected by lines (left), and shown in spacefilling on a structure of the *T. maritima* HK853-RR468 complex (right). The only pair in the hybrid kinase alignment with a score greater than 3.0 is highlighted. For clarity, only the DHp domain of HK853 is shown. Residue numbers correspond to positions within EnvZ and OmpR (see Figure 4.2A-B). (D) Histogram of adjusted mutual information scores for all residue pairs in the hybrid histidine kinase alignment. Arrows indicate the residue pairs scoring higher than 3.5 in the analysis of canonical two-component proteins, with scores for these pairs in each alignment listed in the table.

Canonical histidine kinases harbor two highly-conserved domains, a dimerization and histidine phosphotransfer (DHp) domain and a catalytic and ATP binding (CA) domain.

The DHp domain promotes homodimerization and harbors the histidine that is autophosphorylated by the CA domain. Response regulators also typically have two domains, a receiver domain and an output domain. The receiver domain contains a conserved aspartate that receives a phosphoryl group from the autophosphorylated kinase while the output domains are variable, but are often DNA-binding domains.

Phosphotransfer relies primarily on an interaction between the DHp domain of the kinase and the receiver domain of the regulator (Casino et al., 2009). The residues that determine the specificity of this interaction were identified through analyses of amino acid coevolution in large sets of cognate kinase-regulator pairs (Capra et al., 2010; Skerker et al., 2008). These studies pinpointed a small set of strongly coevolving residues that determine the specificity of two-component signaling proteins and that enable the rational rewiring of both the kinase and the regulator (Bell et al., 2010; Capra et al., 2010; Skerker et al., 2008).

The coevolution of specificity-determining residues in two-component signaling proteins is driven by negative selection against pathway cross-talk following gene duplication (Capra et al., 2012). The insulation of recently duplicated two-component proteins requires changes in the residues that govern molecular recognition, such that each cognate pair of signaling proteins continues interacting while avoiding cross-talk with the other pathway. In some cases, changes in the specificity residues of other two-component signaling proteins, that were not recently duplicated, are also necessary to achieve a system-wide insulation of all pathways in a given cell (Capra et al., 2012).

A common variant of two-component signaling involves hybrid histidine kinases, in which a conventional histidine kinase is fused to a receiver domain similar to those found in soluble response regulators (Figure 4.1B). Hybrid kinases autophosphorylate and are thought to transfer the phosphoryl group intramolecularly to their receiver domains. The phosphoryl group can then be transferred to a histidine phosphotransferase and finally to a soluble response regulator, completing a phosphorelay. Hybrid histidine kinases are found in over 50% of all bacterial genomes and nearly 25% of all bacterial histidine kinases are hybrids (Wuichet et al., 2010). These hybrid kinases likely arise through the fusion of canonical, co-operonic histidine kinases and response regulators, and may further expand through gene duplication (Whitworth and Cock, 2009; Zhang and Shi, 2005).

Despite their prevalence, the phosphotransfer properties and specificity of hybrid kinases are poorly characterized relative to canonical histidine kinases. Here, we investigated the global phosphotransfer specificity of hybrid histidine kinases. We find that these hybrid kinases exhibit significantly reduced phosphotransfer specificity when liberated from their receiver domains. The covalently attached receiver domain thus normally serves as an intramolecular phosphoacceptor and helps prevent unwanted cross-talk inside cells. Our data further indicate that, following the duplication of a hybrid kinase, there is reduced selective pressure to diversify the residues responsible for binding its attached response regulator domain, in stark contrast to canonical histidine kinases. In sum, we propose that the spatial arrangement of domains in hybrid histidine kinases strongly influences the evolution of these proteins with implications for understanding the

evolution of multi-domain signaling proteins throughout biology and for designing synthetic circuits.

Results

Hybrid kinases show reduced amino acid coevolution between kinase and receiver domains

Analyses of amino acid coevolution using mutual information as a metric have helped pinpoint the residues that govern protein-protein interaction specificity in two-component signal transduction systems (Capra et al., 2010; Skerker et al., 2008). These analyses identified a small set of residues that map to the molecular interface formed during phosphotransfer (Casino et al., 2009), and were used to guide the rational rewiring of substrate specificity for the model histidine kinase EnvZ, validating their role in dictating specificity (Skerker et al., 2008). To assess whether the same residues coevolve in hybrid histidine kinases, we examined amino acid coevolution in a large set of hybrid kinases. This analysis was performed on a multiple sequence alignment containing 2681 hybrid histidine kinases, drawn from a wide phylogenetic range of organisms. This sequence alignment contained the DHp and CA domains of each hybrid kinase as well as its receiver domain, but omitted sensory domains. To measure coevolution we used a mutual information-based algorithm that helps adjust for phylogenetic and sampling biases in sequence alignments (Martin et al., 2005). Adjusted MI values were calculated for all possible pairs of positions within the sequence alignment (Figure 4.1C, 4.2A-D). A similar analysis for canonical kinase-regulator pairs was used for comparison (Capra et al., 2010). The two alignments have similar entropy at each position, facilitating a comparison of mutual information scores (Figure 4.2E-F).

We focused primarily on residue pairs in which one position corresponds to a site within the DHp or CA domains and the other to a site within the receiver domain. The overall

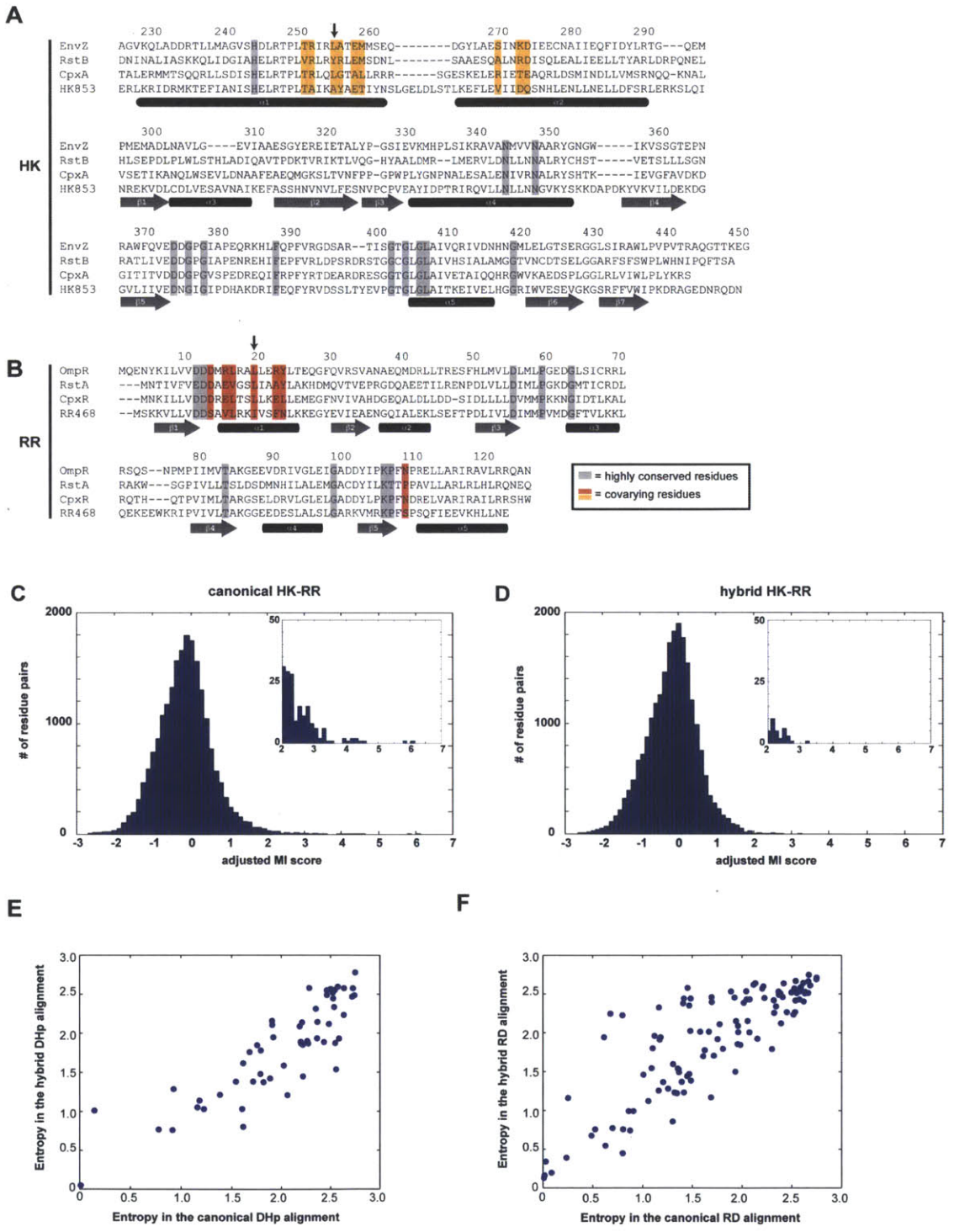


Figure 4.2 Amino acid coevolution analysis of hybrid histidine kinases.

(A) Sequence alignment of three canonical histidine kinases from *E. coli* (EnvZ, RstB, CpxA) and one canonical kinase from *T. maritima* (HK853). Alignment numbering corresponds to HK853. The alignment only shows the DHp and CA domains. (B) Sequence alignment of the cognate response regulators for the kinases in panel A. For panels A and B, the residues that strongly

coevolve and that dictate specificity in canonical two-component signaling proteins are shaded orange and red. The pair of residues that also strongly coevolved in the hybrid kinases is marked by arrows above each alignment. Highly conserved residues is all kinases and receiver domains are shaded grey. (C-D) Histograms of adjusted mutual information scores for residue pairs in the multiple sequence alignment of (C) 4,375 canonical, cognate kinase-regulator pairs or (D) 2,681 hybrid histidine kinases (see text for details). Insets show tails of each distribution. (E) Scatter plot of entropy values for each position in the multiple sequence alignment of DHp domains in canonical and hybrid kinases. (F) Scatter plot of entropy values for each position in the alignment of receiver domains in canonical and hybrid kinases.

shape of the distribution of adjusted MI values was similar for the canonical kinase-regulator pairs and the hybrid kinase-receiver domain pairs (Figure 4.2C-D). However, the hybrid kinase distribution did not contain the same long tail seen in the canonical distribution. There are 12 pairs of amino acids in the canonical kinase-regulator alignment that have adjusted MI values greater than 3.5, which indicates significant coevolution. In contrast, in the hybrid kinase-receiver domain alignment, no residue pair had an MI value greater than 3.5, and only one pair had a value greater than 3.0 (Figure 4.1C).

The scores for residue pairs in the hybrid kinase alignment were not simply reduced relative to those from the canonical alignment. Of the 12 top-scoring residue pairs from the canonical kinase-regulator alignment, only 5 were included in the top 12 scoring pairs from the hybrid kinase alignment. The other 7 had substantially reduced scores, falling throughout the distribution, although each had a positive score (Figure 4.1D). This analysis suggests that hybrid kinases do not exhibit the same extensive amino acid coevolution between DHp and receiver domains as canonical kinase-regulator pairs.

Hybrid kinases exhibit limited phosphotransfer specificity

To determine whether the reduced coevolution in hybrid kinases translates into a difference in kinase specificity, we performed phosphotransfer profiling (Skerker et al.,

2005). In this approach, a histidine kinase is autophosphorylated using [γ-^{32}P]ATP and then systematically tested for phosphotransfer to a large panel of full-length response regulators or receiver domains, using SDS-PAGE and phosphorimaging. Robust phosphotransfer typically manifests both with a band corresponding to a phosphorylated response regulator and, sometimes, with depletion of the radiolabeled kinase band.

We profiled 10 different hybrid kinases from the α-proteobacterium *C. crescentus*. In each case we purified an epitope-tagged construct harboring the DHp and CA domains, but not the receiver domain. We first profiled each kinase against the entire set of receiver domains from the 27 annotated *C. crescentus* hybrid kinases, using incubation times of 15 minutes (Figure 4.3A-B, 4.4). Strikingly, most of the kinases phosphorylated several of the hybrid kinase receiver domains. In fact, some kinases phosphorylated the majority of the receiver domains. These profiles stand in sharp contrast to our results with canonical histidine kinases in which the phosphotransfer profiles were typically extremely sparse, with kinases phosphorylating a single cognate response regulator (Skerker et al., 2008; Skerker et al., 2005).

Interestingly, not all of the hybrid histidine kinases phosphorylated their own receiver domains. For example, the kinase CC0723 phosphorylated the receiver domains of CC3075 and CC2670, but not its own, even though other hybrid kinases were able to phosphorylate the CC0723 receiver domain. There were also several cases in which a hybrid kinase phosphorylated its own receiver domain, but did so more weakly than other receiver domains. For example, CC3191 phosphorylated the CC0921 receiver domain to a greater extent than its own (Figure 4.3A, 4.6B). Thus, unlike canonical kinases for

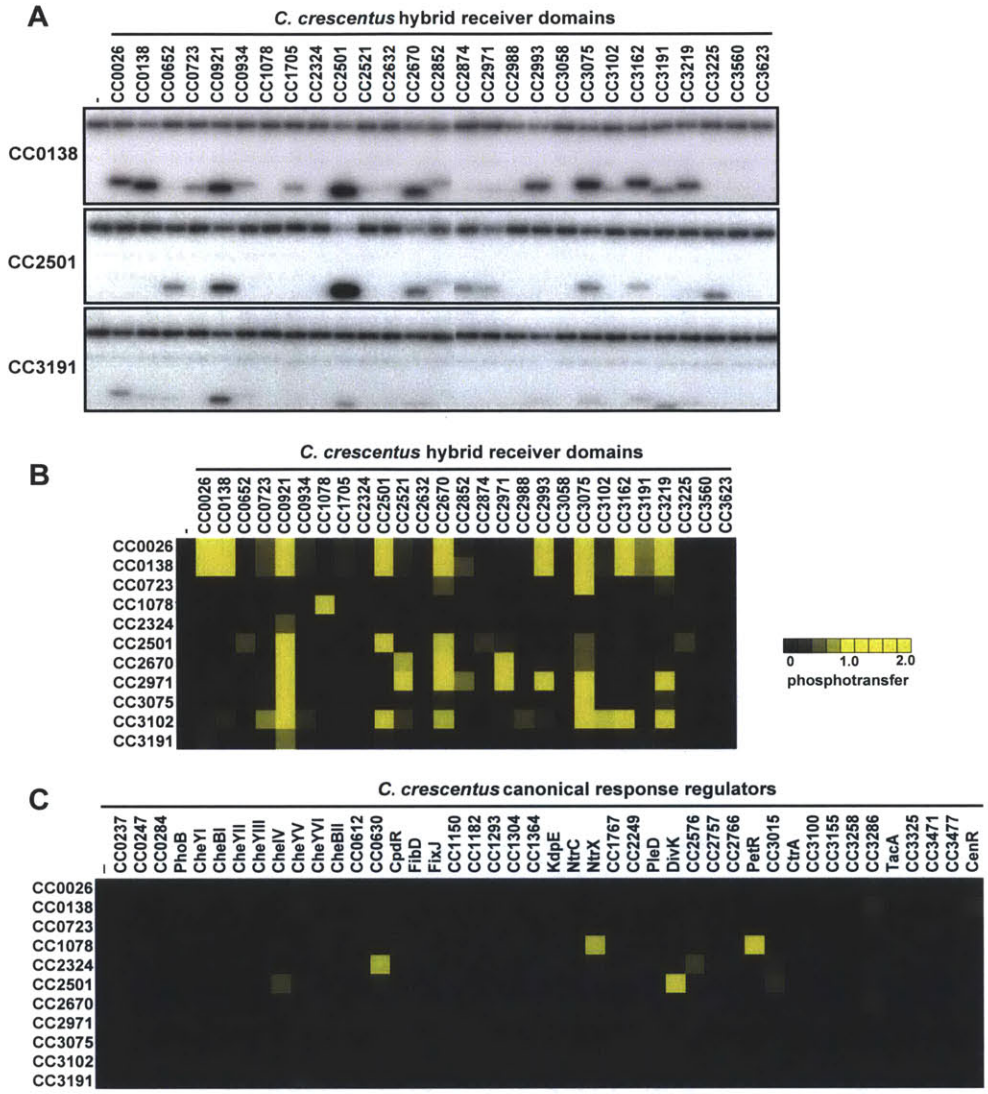


Figure 4.3 Hybrid histidine kinases show reduced phosphotransfer specificity *in vitro*.

(A) Phosphotransfer profiles for kinase domains from three *C. crescentus* hybrid histidine kinases against all 27 receiver domains from hybrid kinases. (B) Quantification of phosphotransfer profiles for 10 hybrid kinases against the 27 hybrid kinase receiver domains; for raw profile data, see Figure 4.4. (C) Quantification of phosphotransfer profiles for 10 hybrid kinases against the 44 soluble *C. crescentus* response regulators; for raw profile data, see Figure 4.5. For panels B-C, the ratio of receiver domain or response regulator band intensity to the autophosphorylated kinase band intensity was calculated and converted to color based on the legend shown. All phosphotransfer reactions were incubated 15 minutes.

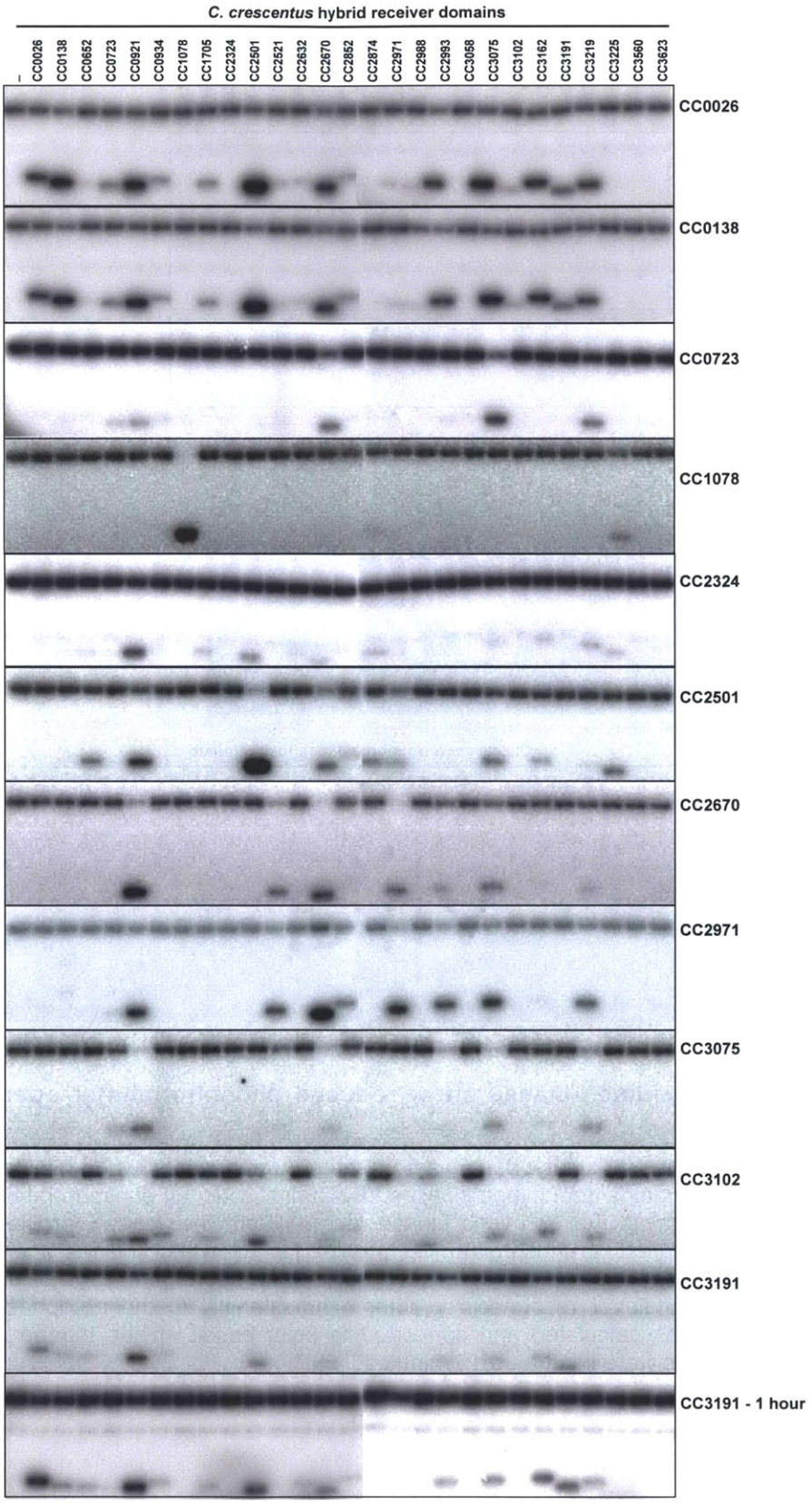

C. crescentus hybrid receiver domains
–
CC0026
CC0138
CC0652
CC0723
CC0921
CC0934
CC1078
CC1705
CC2324
CC2501
CC2521
CC2632
CC2670
CC2852
CC2874
CC2971
CC2988
CC2993
CC3058
CC3075
CC3102
CC3162
CC3191
CC3219
CC3225
CC3560
CC3623
CC0026
CC0138
CC0723
CC1078
CC2324
CC2501
CC2670
CC2971
CC3075
CC3102
CC3191
CC3191 - 1 hour

Figure 4.4 Phosphotransfer profiles against receiver domains.

The kinase domains from 10 different *C. crescentus* hybrid histidine kinases were each profiled against the receiver domains from the 27 hybrid kinases in *C. crescentus*. Each profile involved 15 minute phosphotransfer reaction times except CC3191 which was profiled at 15 minutes and 1 hour.

which the cognate response regulator is usually the kinetically preferred target, hybrid kinases display a variety of behaviors, and often harbor substantially less specificity.

Next, we profiled each of the 10 hybrid kinases against the entire set of 44 canonical, soluble response regulators encoded in the *C. crescentus* genome (Figure 4.3C, 4.5). Although these profiles were sparser than those performed against the hybrid kinase receiver domains, there were significant interactions observed with several of response regulators. For instance, the kinase domain of CC2501 showed significant phosphotransfer to the regulators CheYIV, DivK, and CC3015. There were also several response regulators that were phosphorylated by multiple hybrid kinases, including CC0630, CC2576, CC3015, and CC3286. Finally, we noted that two hybrid kinases, CC0723 and CC2324, showed stronger phosphotransfer to CC0630 than to any of the hybrid kinase receiver domains, including their own. These profiles reinforce the conclusion that hybrid kinases exhibit relaxed phosphotransfer specificity and are fundamentally different in this respect from canonical histidine kinases.

Physical attachment of a receiver domain reduces signaling cross-talk

Although our data demonstrated a reduced specificity of hybrid kinases, these profiles were performed using kinases that had been physically separated from their receiver domains. The kinetic preference and phosphotransfer behavior of these liberated kinase

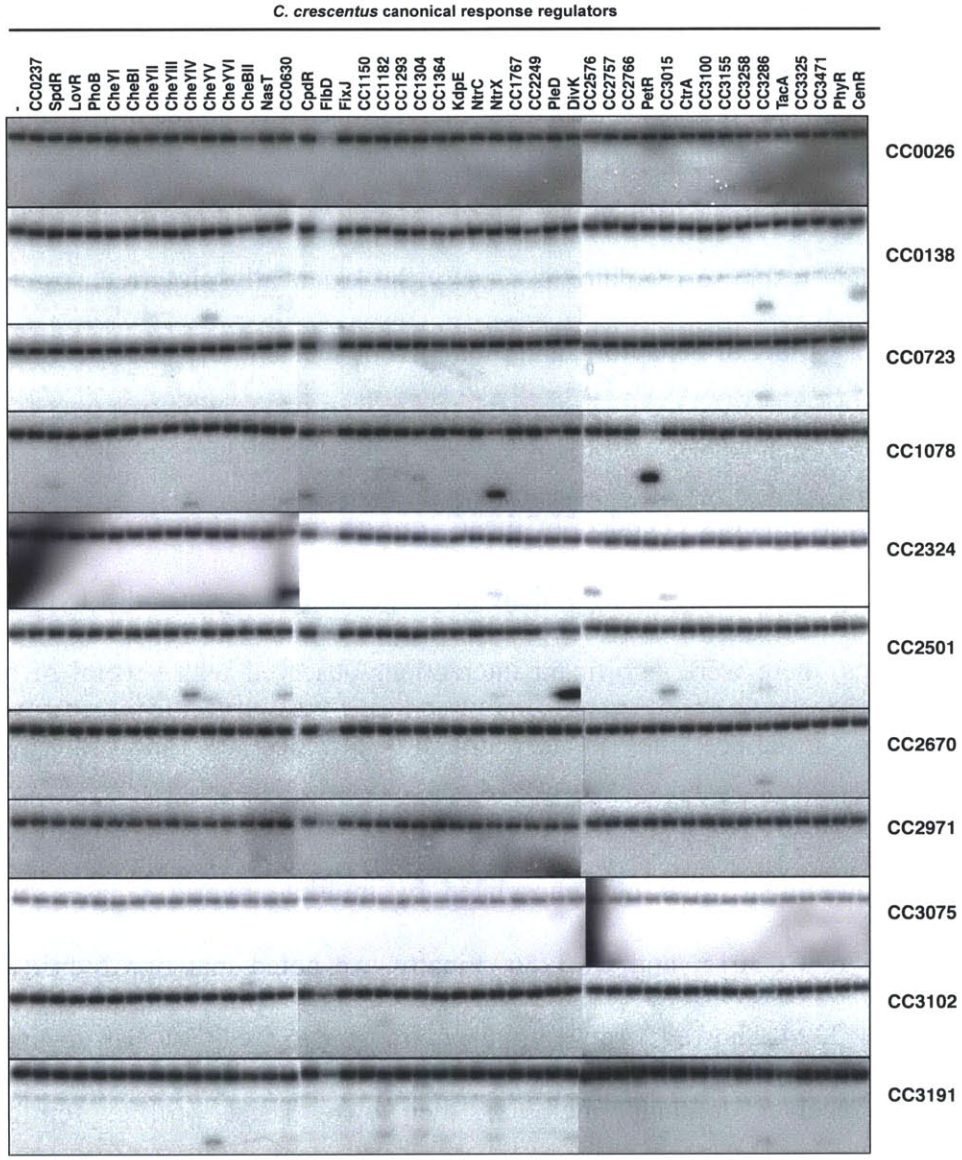


Figure 4.5 Phosphotransfer profiles against response regulators.

The kinase domains from 10 different *C. crescentus* hybrid histidine kinases were each profiled at 15 minutes against the 44 soluble response regulators in *C. crescentus*.

domains likely differ substantially from those of full-length hybrid kinases. For example, although the kinase domain for CC0138 (ShkA) phosphorylated 16 receiver domains and 3 full-length response regulators, previous studies have indicated that ShkA exclusively phosphorylates its own receiver domain *in vivo* (Biondi et al., 2006b). Similarly, although

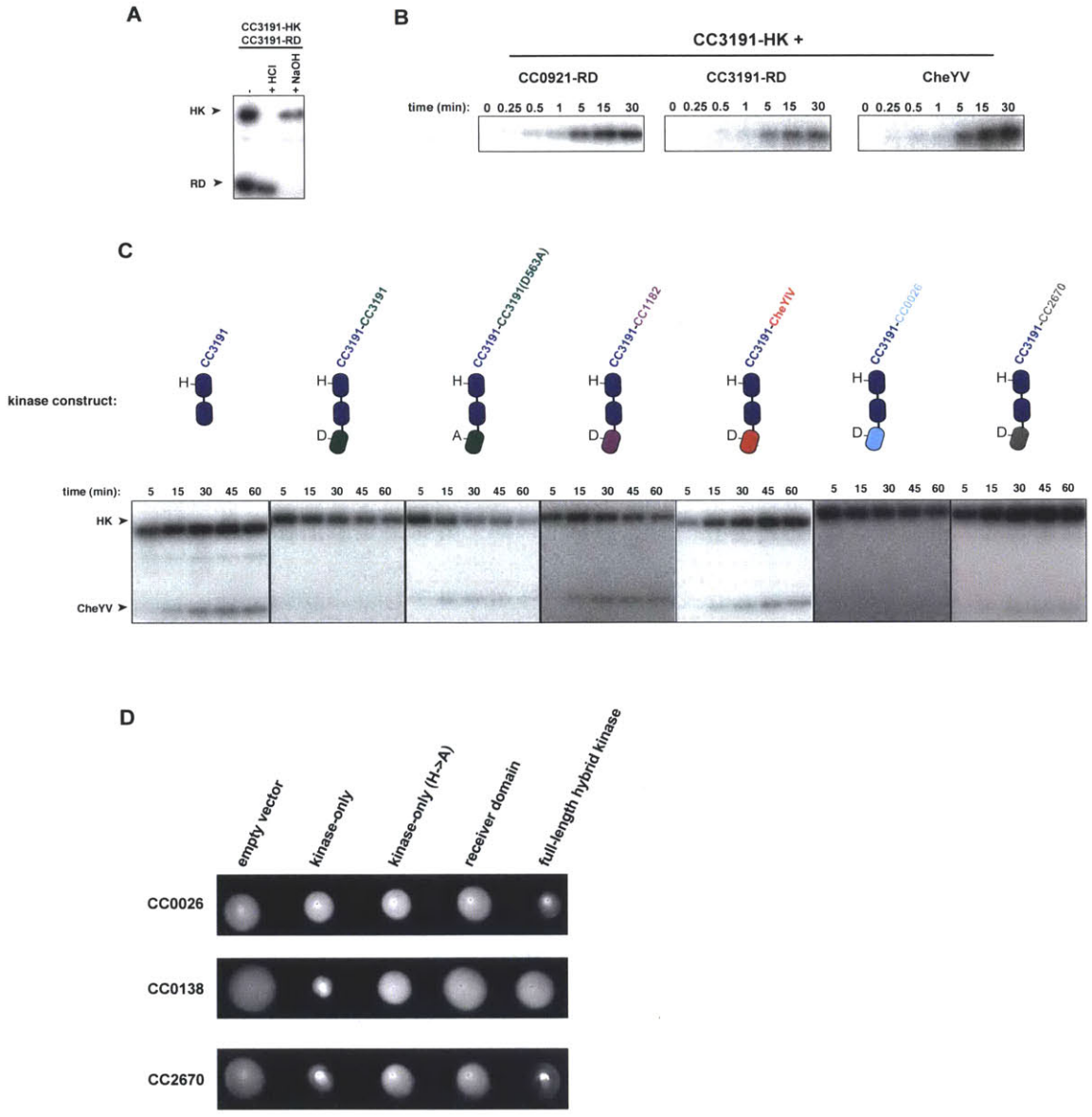


Figure 4.6 Hybrid kinases lacking their receiver domains exhibit cross-talk.

(A) The kinase-only CC3191 was incubated with CC3191 receiver domain in the presence of [γ-32]ATP and either buffer, HCl, or NaOH. (B) The kinase-only portion of CC3191 was autophosphorylated and examined for phosphotransfer to the receiver domains of CC0921 and CC3191, and to the response regulator CheYV, at the time points indicated. (C) Representative gels showing the time-course of phosphotransfer to CheYV from each kinase construct in shown in Figure 4.7B. (D) Representative swarm plates for strains expressing various domains of the three kinases indicated. Quantifications are shown in Figure 4.7D.

the kinase domain of CC1078 (CckA) showed apparent promiscuity *in vitro* and phosphorylated the response regulator PetR, there is no evidence of cross-talk to this regulator *in vivo* and CckA does not activate PetR-dependent genes *in vivo* (Biondi et al., 2006a). Thus, we propose that the high local concentration of a covalently attached receiver domain normally allows this domain to outcompete other response regulators for access to an autophosphorylated kinase domain.

To further probe the effect of covalently attaching a receiver domain to a histidine kinase, we focused on the hybrid kinase CC3191. We first compared the phosphotransfer behavior of the CC3191 construct used in Figure 4.3 that harbors the DHp and CA domains to a construct that also contains the C-terminal receiver domain of CC3191. The kinase-only construct for CC3191 phosphorylated its own receiver domain *in vitro*, although it also phosphorylated the soluble response regulator CheYV at a similar rate (Figure 4.3A, 4.6B). In contrast, the longer construct containing the C-terminal receiver domain no longer detectably phosphotransferred to CheYV (Figure 4.6C, 4.7A). This result demonstrates that the receiver domain in a hybrid kinase normally prevents cross-talk between the kinase domain and other, soluble response regulators.

The suppression of cross-talk provided by a receiver domain could arise through steric hindrance or because the kinase domain is engaged in intramolecular phosphotransfer. To determine whether productive phosphotransfer contributes, we first generated a full-length CC3191 construct in which the phosphoaccepting aspartate (D563) in the receiver domain was mutated to alanine. This construct exhibited significantly more phosphotransfer to soluble CheYV than the wild-type CC3191 construct, indicating that engagement of the kinase domain in intramolecular phosphotransfer contributes to the

suppression of cross-talk (Figure 4.7B), although the receiver domain may also prevent cross-talk, in part, by occluding the binding of other regulators.

To further understand the contribution of a receiver domain to the prevention of cross-talk, we created chimeric hybrid kinases, fusing the kinase domain of CC3191 to a receiver domain from CheYIV or CC1182 (soluble response regulators) or from CC0026 or CC2670 (hybrid kinases). In our profiling studies, the liberated kinase domain of CC3191 had not detectably phosphorylated CheYIV, and had only weakly phosphorylated CC1182 and the receiver domain of CC2670, but it had strongly phosphorylated the receiver domain of CC0026 (Figure 4.3C). To test whether these four chimeras could phosphotransfer intramolecularly from the CC3191 kinase domain to the heterologous receiver domain attached, we autophosphorylated each in buffer, acid, or base (Figure 4.7A). Histidyl-phosphate bonds are sensitive to acid and aspartyl-phosphate bonds are sensitive to base (Figure 4.6A). The phosphorylation of CC3191 was decreased in the presence of either acid or base, indicating that it was phosphorylated on both the histidine and aspartate. In contrast, the phosphorylation of CC3191(D563A) was primarily acid sensitive. Together, these patterns of acid/base sensitivity indicate that CC3191 normally autophosphorylates and transfers its phosphoryl group intramolecularly to its receiver domain.

We observed a similar pattern, consistent with intramolecular phosphotransfer, for the chimera CC3191-CC0026 and, to a lesser extent, CC3191-CC2670, but not CC3191-CheYIV or CC3191-1182. These findings are consistent with our results indicating that the CC3191 kinase domain alone can phosphorylate its own receiver domain and the receiver domains of CC0026 and CC2670, but not CC1182 or CheYIV (Figure 4.3).

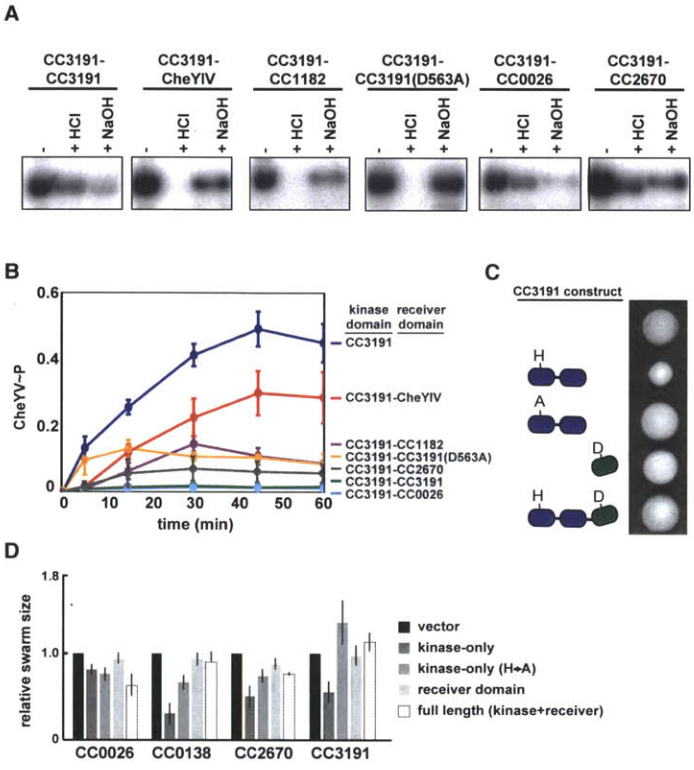


Figure 4.7 Hybrid kinases lacking their receiver domains exhibit cross-talk.

(A) Chimeric hybrid kinases were autophosphorylated in the presence of buffer, HCl, or NaOH to assess whether phosphoryl groups resided on the conserved histidine, aspartate, or both. (B) Chimeric hybrid kinases were autophosphorylated and then tested for phosphotransfer to soluble CheYV at the time points indicated. Error bars represent standard deviation from three independent replicates. Raw gel images are shown in Figure 4.6C. The identity of domains in each chimeric kinase are listed. (C) Swarm plate assay for strains expressing each of the CC3191 constructs listed or vector alone. (D) Quantification of swarm sizes for strains expressing various constructs for each of the four hybrid histidine kinases indicated. Swarm areas were measured and plotted relative to the empty vector control. Error bars represent standard deviations from three replicates. Swarm plate images are shown in Figure 4.6D.

These results also indicate that tethering non-cognate receiver domains to a histidine kinase is not always sufficient to promote phosphotransfer.

Next, we tested whether the four chimeras would phosphorylate, or cross-talk to, soluble CheYV. All four chimeras showed reduced phosphotransfer to CheYV compared to the CC3191 kinase-only construct (Figure 4.6C, 4.7B), with the strongest suppression of

cross-talk occurring with CC3191-CC2670 and CC3191-CC0026, the two chimeras that also demonstrated the most significant intramolecular phosphotransfer. Only the CC3191-CC0026 chimera, whose kinase and receiver domains displayed an interaction similar to that of CC3191-CC3191, both in isolation and when fused, completely prevented cross talk. Taken together, our results indicate that the receiver domain of a hybrid histidine kinase plays an important role in reducing, or eliminating, cross-talk with other response regulators by interacting with, and receiving phosphoryl groups from, the linked kinase domain.

Hybrid kinases lacking their receiver domains likely cross-talk to other response regulators in vivo

Previous work has shown that, with only a few exceptions, canonical histidine kinase-response regulator pairs are insulated from each other *in vivo* (Laub and Goulian, 2007; Skerker et al., 2005) and, importantly, that cross-talk between non-cognate pairs can be severely detrimental to an organism's fitness (Capra et al., 2012). We have shown here that many of the hybrid kinases, when separated from their receiver domains, interact readily with noncognate response regulators *in vitro*. Thus, we hypothesized that expressing only the kinase domain of a hybrid histidine kinase might induce cross-talk *in vivo* and affect the growth or fitness of cells.

We tested this hypothesis by inducing expression of CC3191 lacking its C-terminal receiver domain in *C. crescentus* and assessing cellular growth in swarm plates. Wild-type *C. crescentus* cells can swim through low-percentage agar, creating a large circular colony, or swarm; defects in motility, chemotaxis, cell growth, or cell division can affect swarm size, making this a convenient assay for assessing gross cellular phenotype

(Skerker et al., 2005). We found that cells producing the kinase-only portion of CC3191 produced a small swarm relative to the wild type without affecting growth or morphology. This observation is consistent with the notion that a kinase-only version of CC3191 inappropriately phosphotransfers to CheYV *in vivo*, as it does *in vitro* (Figure 4.3C). In contrast, cells synthesizing either a full-length construct that contains the receiver domain or the receiver domain alone did not exhibit significant swarm phenotypes (Figure 4.7C-D). The phenotype seen with cells expressing the kinase portion of CC3191 was dependent on autophosphorylation, as cells overexpressing a construct in which the conserved histidine was mutated to an alanine no longer exhibited a severe swarm phenotype.

We then tested the effects of overexpressing three other hybrid histidine kinases that we profiled above: CC0026, CC0138, and CC2670. Like CC3191, these kinases do not contain transmembrane domains. As with CC3191, overproducing the N-terminal and kinase domains of CC0138 and CC2670 led to a small swarm phenotype, whereas constructs containing both the kinase and receiver domains, or the receiver domain alone, did not (Figure 4.6D, 4.7D). For the kinase-only constructs of CC0138 and CC2670, the phenotype was suppressed by substituting the phosphorylatable histidine with an alanine suggesting that autokinase activity is required for the small swarm phenotype. Unlike CC0138 and CC2670, cells synthesizing the kinase-only version of CC0026 did not exhibit a significant swarm phenotype. Notably, however, the kinase domain of CC0026 had not significantly phosphorylated any non-hybrid receiver domains *in vitro* (Figure 4.3C). Taken together, these data are consistent with the idea that some hybrid kinases

are promiscuous, but that their attached receiver domains normally help to prevent cross-talk with other response regulators *in vivo*.

Hybrid histidine kinases are under reduced selective pressure to diversify

Collectively, our results indicate that hybrid histidine kinases are subject to different selective pressures than canonical histidine kinases. We previously found that canonical histidine kinases and response regulators are under strong selective pressure to diversify their specificity residues following gene duplication, but are otherwise relatively static (Capra et al., 2012). This diversification of specificity residues post-duplication is critical to preventing cross-talk and ultimately ensures the system-wide optimization of phosphotransfer specificity (Capra and Laub, 2012; Capra et al., 2012). Consistently, inspection of the six key specificity residues (those from α-helix 1 in the DHp domain) in genome-wide sets of canonical histidine kinases indicates fewer than three identities at these six positions in most pairwise comparisons (Figure 4.8).

We extracted the corresponding six residues from each of 24 hybrid histidine kinases in *C. crescentus* (Figure 4.8). Although there are 27 annotated hybrid kinases that contain CA and receiver domains, 3 did not have intact DHp domains. Strikingly, many of the 24 hybrid kinases share four, five, or even six identities at these positions with other hybrid kinases. This similarity does not arise simply because the hybrid kinases duplicated recently, as pairwise comparisons of the entire DHp and CA domains demonstrated extensive variability at other sites (Figure 4.2E-F), resulting in significant separation in a neighbor-joining tree built from those domains (Figure 4.9A).

C. crescentus hybrid histidine kinases

		kinase domain	receiver domain
CC0026		NGGVHV	NTNVRIS
CC0138	(ShkA)	NGGMRL	NINLTLT
CC0652		TAGFAL	NLNMAIT
CC0723		NGGLQA	YVNVMME
CC0921		NGGMHA	NINVTMK
CC0934		NGGMQI	NTNVTLE
CC1078	(CckA)	TALRDE	EAVVRLD
CC1705		NANGGR	HINTGIS
CC2324		TGHVAA	DLNMAVS
CC2501		TAGFEV	NVNLAIS
CC2521		NGAIDR	HTNVIVH
CC2632		NSGFQL	HINIALQ
CC2670		NGALAA	HINVLIT
CC2852		NGGMEV	HTNVLMS
CC2874		TVGADV	DQVLAMS
CC2971		NGALSA	NNNVLLT
CC2988		NGAMDA	HVNVAIA
CC2993		NAGLEV	NANVLLA
CC3075		NGGLQA	HVNVLFD
CC3102		NGGLHA	NVNVTIQ
CC3191		NGGVHL	NTNIRMS
CC3219		NGAMDV	NTNVAVE
CC3225		AGGSEM	ETVLDTA
CC3623		SAAGGR	HVNALVD

C. crescentus canonical histidine kinases

CC1063	(DivJ)	NAGFDI
CC0238		TSSAET
CC0289	(PhoR)	ASGFET
CC0530	(CenK)	TSMADR
CC1181		TRFREA
CC1294		TAGEEV
CC1305		AAAQRR
CC1594	(KdpD)	STGATT
CC2765		SVTESQ
CC2932		TRLEAM
CC3327		TSALAD
CC1740	(NtrB)	AGGAQL
CC1742	(NtrY)	TPLSER
CC0759	(FixL)	SANLTG
CC0248		ATVVRE
CC2482	(PleC)	NAGFEI
CC0586		TSGFEQ
CC1062		NAGFEI
CC2755		TRAREV
CC2884		NAGFSV

C. crescentus response regulators

CC0284	(LovR)	ELLEHLS
CC0432	(CheYI)	STMMMAN
CC0437	(CheYII)	QTMLNAT
CC0440	(CheYIII)	NPISQVT
CC0588	(CheYIV)	DAILGVE
CC0591	(CheYV)	YTTIGLS
CC0596	(CheYVI)	SVIVRMD
CC0630		ELVMDMN
CC0744	(CpdR)	DSLFRAH
CC2463	(DivK)	NLNLDLS
CC2576		ELVEALS
CC3015		ELVLDMR
CC3258		NGFLQIS
CC3286		DVLIITT
CC3471		SVIVRVD
CC0237		DNISLAS
CC0294	(PhoB)	EALLYNS
CC1182		DGIVDFN
CC1293		DRVFRGS
CC1304		DVVDKAE
CC1595	(KdpE)	EQIFPAG
CC2757		DEAAHGA
CC2766		DDLGLAH
CC2931	(PetR)	DRLLEFE
CC3035	(CtrA)	DATTLMH
CC3325		DSHLSVQ
CC3743	(CenK)	DDLALAR
CC0909	(FlbD)	LGQVKMV
CC1741	(NtrC)	DSIVQAD
CC1743	(NtrX)	EDILGIK
CC3315	(TacA)	DTQLAVS
CC0758	(FixJ)	DSASFLS
CC1150		EQKLLSL
CC0247	(SpdR)	DPLRRAD
CC1767		DKFRTSN
CC0612	(NasT)	PFSHRRV
CC3477	(PhyR)	EVIDALQ
CC0436	(CheBI)	STMLAAD
CC0597	(CheBII)	SVVMRWA
CC2462	(PleD)	IANLAKD
CC1364		NNMVTMT
CC2249		NHIIAIT
CC3100		NATLEHN
CC3155		NHMLEMT

Figure 4.8 Genome-wide sets of specificity residues from two-component signaling proteins.

The key specificity determining residues, as defined through coevolution analysis of canonical two-component signaling proteins (Figure 4.1), were extracted from each of the histidine kinases, response regulators, and hybrid histidine kinases in *C. crescentus*.

The lack of variability at the sites corresponding to the six key specificity residues in canonical kinases was also evident in sequence logos for the 24 hybrid and 21 canonical kinases from *C. crescentus* (Figure 4.9B). The logo for canonical kinases indicated relatively low conservation at each specificity position except the first, which may be constrained due to involvement in autophosphorylation (Capra et al., 2010; Casino et al., 2010). In contrast, the logo for hybrid kinases indicated higher conservation at each site.

The kinase domains of hybrid histidine kinases are likely under less selective pressure than canonical kinases to diversify following gene duplication. The effective concentration of the attached receiver domain is high enough to ensure that a hybrid kinase will transfer its phosphoryl group intramolecularly and not to another regulator or receiver domain. Hence, after duplication of a hybrid kinase, the residues that bind to the receiver domain do not need to change to insulate the new proteins from one another, as occurs in canonical kinases (Figure 4.10). Consistent with this hypothesis, many of the hybrid histidine kinases in *C. crescentus,* which were likely derived from a common ancestral gene through duplication and divergence, had similar specificity residues and exhibited similar phosphotransfer profiles when liberated from their receiver domains (Figure 4.3B). One exception to this trend was CC1078 (CckA), which had a distinct set of specificity residues relative to the other hybrid kinases and, consequently, had a

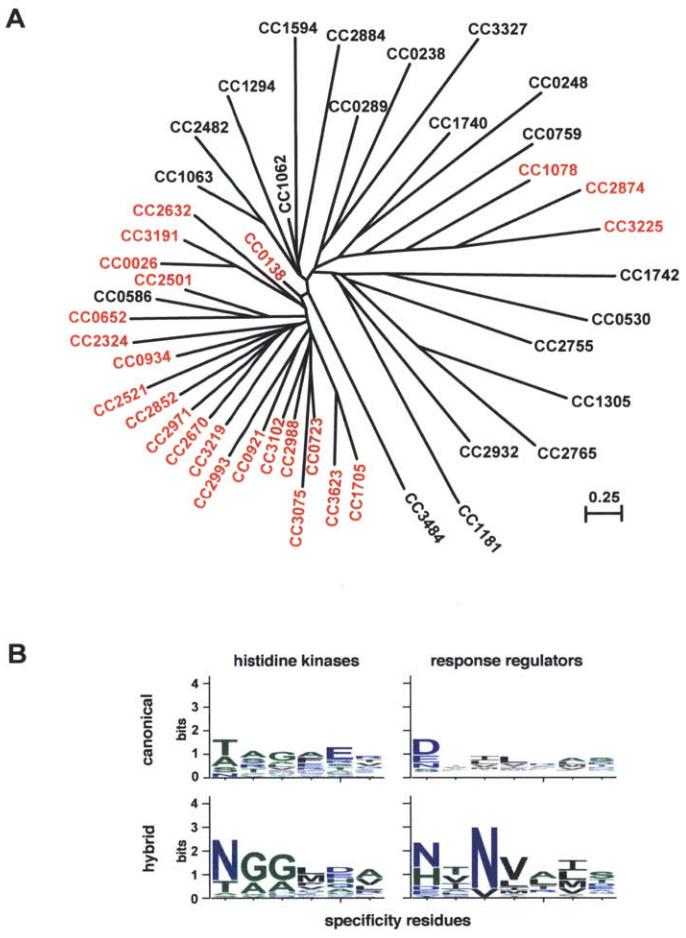


Figure 4.9 Specificity residues are conserved among hybrid histidine kinases.

(A) An unrooted neighbor-joining tree of the *C. crescentus* kinases was built from an alignment of the DHp domains of all 24 hybrid and 21 canonical histidine kinases from *C. crescentus*. Hybrid kinases are labeled in red. (B) Sequence logos for the residues that dictate phosphotransfer specificity in canonical kinase-regulator pairs. Logos were built from an alignment of the 21 canonical histidine kinases and 44 soluble response regulators (top), and from an alignment of the 24 hybrid histidine kinases in *C. crescentus* (bottom).

significantly different phosphotransfer profile. Notably, CckA did not group with the other hybrid kinases in a tree of *Caulobacter* kinases (Figure 4.9A) suggesting that CckA may be relatively ancient and not derived from a recent duplication.

Discussion

The expansion of existing signaling protein families has enabled cells to rapidly evolve the ability to sense and response to a wide range of stimuli. In bacteria, two-component signaling proteins have expanded dramatically, such that most species encode dozens, and sometimes hundreds, of these proteins. For canonical pathways involving a single histidine kinase and response regulator, these pathways are exquisitely specific and a cognate response regulator can outcompete all other non-cognate regulators to receive phosphoryl groups from a given histidine kinase. Consequently, phosphotransfer profiles of canonical kinases have demonstrated that each possesses a strong kinetic preference for its cognate substrate (Skerker et al., 2005). This preference is determined by a small number of specificity-determining residues in both the kinase and regulator. These residues must coevolve to maintain a tight, specific interaction between cognate partners, particularly after a gene duplication event as a means of insulating the new pathways from one another (Figure 4.10) (Capra et al., 2012).

In contrast to the canonical systems, we demonstrated here that kinase domains of hybrid kinases typically exhibit relaxed substrate specificity, often phosphorylating soluble response regulators or other receiver domains as well or better than they phosphorylate their own receiver domains. A similar observation was made previously in *Myxococcus xanthus* with a limited set of response regulators. In that case, the kinase domain of RodK was shown to preferentially phosphorylate the soluble regulator RokA relative to its own receiver domain, RodK-R3 even though the latter is the *in vivo* target of RodK (Wegener-Feldbrugge and Sogaard-Andersen, 2009).

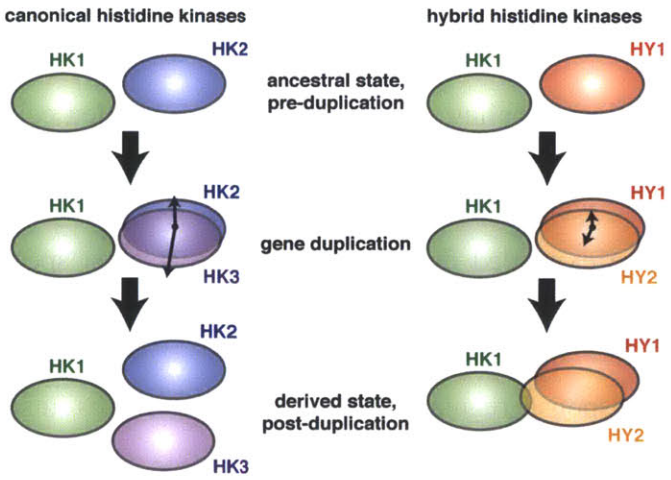


Figure 4.10 Model for changes in specificity residues following duplication of canonical and hybrid histidine kinases.

Ovals represent niches within sequence space, or the set of response regulators recognized by a given histidine kinase as determined by its specificity residues. Post-duplication, canonical kinases separate in sequence space to insulate the two pathways and prevent cross-talk. In contrast, hybrid kinases do not separate, as the tethered receiver domain effectively insulates the duplicated kinases against cross-talk.

Although hybrid kinases are more promiscuous on their own, our data indicate that the covalently attached receiver domain helps to prevent cross-talk with other cytoplasmic response regulators. The local concentration of an attached receiver domain likely exceeds the concentration of all soluble response regulators quite significantly. Consequently, intramolecular phosphotransfer from the kinase domain to the attached receiver domain will be strongly favored, thereby ensuring minimal cross-talk to other pathways.

The enforcement of intramolecular phosphotransfer specificity through spatial tethering of domains likely eliminates selective pressure to diversify the residues in a hybrid kinase that mediate docking to the receiver domain. Hence, after a hybrid kinase duplicates, these residues either will not change or will change more rarely through processes such as

genetic drift (Figure 4.9B). The net result of the reduced rate of change is that for hybrid kinases in extant organisms, the interfacial residues show substantially reduced variability compared to the same set of residues in canonical histidine kinases.

The enforcement of phosphotransfer within hybrid kinases has also likely reduced the need for their kinase and receiver domains to coevolve (Figure 4.1). Mutations that reduce or weaken the interaction of these domains are probably more easily tolerated because the domains are spatially tethered. By contrast, with canonical two-component pathways, the cognate proteins are under strong pressure to coevolve, as a means of maintaining their interaction and preventing interaction with non-cognate proteins. However, merely increasing the effective concentration of a receiver domain was not always sufficient to induce phosphotransfer from a kinase domain (Figure 4.7A) indicating some requirement for molecular recognition and a proper pairing of interfacial residues. It may be that the fusion of domains in a hybrid kinase serves primarily to prevent cross talk, rather than driving phosphotransfer.

Why some two-component pathways involve hybrid histidine kinases instead of canonical kinases is not clear. Hybrid kinases are often involved in phosphorelays, and the additional number of components in a phosphorelay may create additional points for integrating signals (Burbulys et al., 1991). However, not all hybrid kinases necessarily participate in phosphorelays. Recent work with the hybrid kinase VirA from *Agrobacterium tumefaciens* suggests that the receiver domain binds the response regulator VirG, somehow stimulating its activity as a transcriptional activator (Wise et al., 2010). There are also hybrid kinases in some Gram-positive bacteria, such as *Bacteroides thetaiotaomicron*, that have DNA-binding domains C-terminal to their

receiver domains, suggesting that these kinases may directly regulate transcription (Raghavan and Groisman, 2010). In short, although nearly a quarter of all kinases are of the hybrid variety, our understanding of their functions, properties, and advantages remains limited.

The notion that spatial proximity can overcome relaxed specificity of signaling proteins is relevant in all cells. Multi-domain signaling proteins are quite common, particularly in eukaryotes. Additionally, some signal transduction proteins are spatially constrained through the action of scaffolds. For example, in the *S. cerevisiae* pheromone pathway, the scaffold Ste5 enforces the proximity of three separate MAP kinases, helping to prevent them from inappropriately phosphorylating other substrates (Choi et al., 1994). This spatial colocalization may, in turn, have relaxed evolutionary constraints on these MAP kinases.

Finally, our results suggest that information flow through two-component pathways could be rationally engineered by fusing together non-cognate kinases and regulators. Such an arrangement can also prevent unwanted cross-talk with other pathways. Indeed, we showed here that fusing heterologous receiver domains to a hybrid kinase was, in some cases, sufficient to allow phosphotransfer and prevent cross-talk with a soluble regulator. Synthetic scaffolds that bring non-cognate two-component signaling proteins in close proximity may also be used to promote phosphotransfer or prevent cross-talk. A similar approach of artificially colocalizing proteins has been applied in metabolic engineering studies, where enzymes have been tethered together to enhance the synthesis and yield of desired compounds (Dueber et al., 2009).

In sum, our work has revealed new aspects of signaling protein evolution in bacteria that will likely inform similar evolutionary studies in other organisms and help guide efforts to construct synthetic signaling circuits.

Materials and Methods

Sequence analyses

Histidine kinase and response regulator receiver domains were identified, aligned, and filtered as described previously (Capra et al., 2010). Hybrid kinases were defined as those proteins that had a single match to each of the three Pfam models: HisKA, HATPase_C, and Response_reg. The final alignment included 2681 hybrid kinases. Shannon entropy values were calculated for each position in the alignment. Mutual information for every pair of columns in the sequence alignment was calculated as previously reported (Capra et al., 2010). Sequence logos were built using WebLogo (weblogo.berkeley.edu). Neighbor-joining trees were built using the PHYLIP package and multiple sequence alignments built from the DHp domain of each canonical and hybrid histidine kinase in the *C. crescentus* genome.

Strain construction and growth conditions

E. coli and *C. crescentus* strains were grown as described previously (Skerker et al., 2005). Primers used are listed in Table 4.1. Full-length hybrid kinases and the kinase domains of hybrid kinases were amplified from genomic CB15N DNA and ligated into the Gateway pENTR vector (Invitrogen). Chimeric hybrid kinases were cloned by separately amplifying the kinase domain from CC3191 and the specified receiver domain, amplifying the chimeric sequence using splicing with overlap extension PCR and ligating the resulting product into pENTR. pENTR clones were moved into pDEST-His_6-MBP or pDEST-TRX-His_6 vectors for purification, or the pDEST-P_{xyl}-M2 vector derived from

pJS71 for overexpression studies. Overexpression vectors were introduced into wild-type CB15N via electroporation.

Table 4.1 –
Primers

Histidine Kinases		
Kinase-only	**Forward**	**Reverse**
CC0026	CACCTTGTCCCAGGCGTCGACCCCC	TCAGCTGGTCAGCAGCTCCGAGGTC
CC0723	CACCAAGGCCGCCCAGACCAACGG	TCAGGGGTGTTCGCCGCTGAGCGGT
CC2324	CACCCGCATGTTCCGCCGGCAGCA	TCAGGCGCGGGGCGCCCCGGAGGCC
CC2501	CACCGGCCAGCGCTTGCGCCTCGA	TCAGCCCGCGTCGGCGTCAAGGGGC
CC2670	CACCTTGAGAAGAACGACCGCCCAC	TCAGCGCGCCATCACCTGGCCG
CC2971	CACCTTGGCGCGCTACCAAGGGGTG	TCAGGCCGCCGCCGTCAGCGCG
CC3075	CACCCACCAGCGCGGGGCCTCGCG	TCAGCCCTGCAGGGCGGCGGCGCCG
CC3102	CACCCGCACCATGCGCGCCTCGGC	TCAGCCGTCGAGTTGGGTCACGGCG
CC3191	CACCTTGGGCAAGCGCTTGGATACA	TCACCCGTCGAACAGAGGGCCGTCG
Full-length	**Forward**	**Reverse**
CC0026	CACCTTGTCCCAGGCGTCGACCCCC	TTAGGCGACGGCGCAGCGGGGG
CC0138	CACCTTGAGCGACAGCAGATCCGAC	TCAGCCGGCGACCTTGGCTCGC
CC2670	CACCTTGAGAAGAACGACCGCCCAC	TCAGCGCGCCATCACCTGGCCG
CC3191	CACCTTGGGCAAGCGCTTGGATACA	TTAGGAGGCCTTGCGCTGGCGC
Chimeras		
	Forward	**Reverse**
CC3191-HK	CACCTTGGGCAAGCGCTTGG	CCCGTCGAACAGAGGGCCGT
CheYIV	ACGGCCCTCTGTTCGACGGG TTGGGCGCGATGCGAATC	CTAGGACGCCGAACGCACCATC
CC1182	ACGGCCCTCTGTTCGACGGG TTGGAAAACGTCCAAAACGCCGC	TTACTTGGGCAGATAGTCGTCGGCGC
CC0026	ACGGCCCTCTGTTCGACGGG TTGAAAGTCCTTGTCGTGGAGGACAATCCAC	TTAGGCGACGGCGCAGCGGG
CC2670	ACGGCCCTCTGTTCGACGGG TTCAAGGTGCTGCTGGCCGAGGATCAC	TCAGCGCGCCATCACCTGGCC

Site-directed mutagenesis

	Forward
CC3191 (D563A)	GACCTGATCCTCATGGCCATCCAGATGCCGGTC
CC0026 (H537A)	TTCCTGGCCAATATGAGTGCCGAAATCCGCACCCCCATG
CC0138 (H23A)	CAGCTGGCGACCCTGAGCGCCGAGTTCCGCACGCCCCTG
CC2670 (H330A)	TTCCTGGCCAATATGAGCGCCGAGATCCGCACGCCTTTG
CC3191 (H275A)	TTCCTGGCCAATATGAGCGCCGAGATCCGGACGCCCATG

Protein purification and phosphotransfer assays

Expression, protein purification, and phosphotransfer profiling experiments were carried out as described previously (Biondi et al., 2006a; Capra et al., 2012; Skerker et al., 2008; Skerker et al., 2005). All reactions used 500 μM ATP, and 0.5 μCi/μL [γ–^{32}P]ATP. For phosphotransfer experiments in Figure 4.7A, CC3191-HK was autophosphorylated under the same conditions as the phosphotransfer profiles and then incubated with the given receiver domain in a 1:1 ratio for the time indicated. For phosphotransfer experiments in Figure 4.7C, 2.5 μM of the specified kinase was mixed with 2.5 μM CheYV before ATP was added the reaction allowed to proceed for the indicated time before being stopped with the addition of 4X loading buffer. To test acid or base stability of phosphoryl groups, 5 μM of kinase was autophosphorylated at room temperature for 15 minutes. The reaction was then stopped by the addition of 4X loading buffer, and then buffer, 1 M HCl or 0.5 M NaOH was added. After 20 minutes, reactions were neutralized. All phosphotransfer experiments were analyzed by SDS-PAGE and phosphorimaging.

Acknowledgements

We thank Anna Podgornaia for helpful comments on the experiments and on the manuscript. This work was supported by an NSF CAREER award to MTL and an NSF GRFP award to EJC. MTL is an Early Career Investigator at the Howard Hughes Medical Institute.

References

Bell, C.H., Porter, S.L., Strawson, A., Stuart, D.I., and Armitage, J.P. (2010). Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex. PLoS Biol *8*, e1000306.

Biondi, E.G., Reisinger, S.J., Skerker, J.M., Arif, M., Perchuk, B.S., Ryan, K.R., and Laub, M.T. (2006a). Regulation of the bacterial cell cycle by an integrated genetic circuit. Nature *444*, 899-904.

Biondi, E.G., Skerker, J.M., Arif, M., Prasol, M.S., Perchuk, B.S., and Laub, M.T. (2006b). A phosphorelay system controls stalk biogenesis during cell cycle progression in Caulobacter crescentus. Mol Microbiol *59*, 386-401.

Burbulys, D., Trach, K.A., and Hoch, J.A. (1991). Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay. Cell *64*, 545-552.

Capra, E.J., and Laub, M.T. (2012). Evolution of Two-Component Signal Transduction Systems. Annu Rev Microbiol.

Capra, E.J., Perchuk, B.S., Lubin, E.A., Ashenberg, O., Skerker, J.M., and Laub, M.T. (2010). Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet *6*, e1001220.

Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive Mutations that Prevent Crosstalk Enable the Expansion of Paralogous Signaling Protein Families. Cell *150*, 222-232.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Casino, P., Rubio, V., and Marina, A. (2010). The mechanism of signal transduction by two-component systems. Curr Opin Struct Biol *20*, 763-771.

Choi, K.Y., Satterberg, B., Lyons, D.M., and Elion, E.A. (1994). Ste5 tethers multiple protein kinases in the MAP kinase cascade required for mating in S. cerevisiae. Cell *78*, 499-512.

Dueber, J.E., Wu, G.C., Malmirchegini, G.R., Moon, T.S., Petzold, C.J., Ullal, A.V., Prather, K.L., and Keasling, J.D. (2009). Synthetic protein scaffolds provide modular control over metabolic flux. Nat Biotechnol *27*, 753-759.

Fisher, S.L., Kim, S.K., Wanner, B.L., and Walsh, C.T. (1996). Kinetic comparison of the specificity of the vancomycin resistance VanSfor two response regulators, VanR and PhoB. Biochemistry *35*, 4732-4740.

Galperin, M.Y. (2005). A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol *5*, 35.

Grimshaw, C.E., Huang, S., Hanstein, C.G., Strauch, M.A., Burbulys, D., Wang, L., Hoch, J.A., and Whiteley, J.M. (1998). Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in Bacillus subtilis. Biochemistry *37*, 1365-1375.

Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways. Annu Rev Genet *41*, 121-145.

Martin, L.C., Gloor, G.B., Dunn, S.D., and Wahl, L.M. (2005). Using information theory to search for co-evolving residues in proteins. Bioinformatics *21*, 4116-4124.

Raghavan, V., and Groisman, E.A. (2010). Orphan and hybrid two-component system proteins in health and disease. Curr Opin Microbiol *13*, 226-231.

Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub, M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J.M., Prasol, M.S., Perchuk, B.S., Biondi, E.G., and Laub, M.T. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Stock, A.M., Robinson, V.L., and Goudreau, P.N. (2000). Two-component signal transduction. Annu Rev Biochem *69*, 183-215.

Wegener-Feldbrugge, S., and Sogaard-Andersen, L. (2009). The atypical hybrid histidine protein kinase RodK in Myxococcus xanthus: spatial proximity supersedes kinetic preference in phosphotransfer reactions. J Bacteriol *191*, 1765-1776.

Whitworth, D.E., and Cock, P.J. (2009). Evolution of prokaryotic two-component systems: insights from comparative genomics. Amino Acids *37*, 459-466.

Wise, A.A., Fang, F., Lin, Y.H., He, F., Lynn, D.G., and Binns, A.N. (2010). The receiver domain of hybrid histidine kinase VirA: an enhancing factor for vir gene expression in Agrobacterium tumefaciens. J Bacteriol *192*, 1534-1542.

Wuichet, K., Cantwell, B.J., and Zhulin, I.B. (2010). Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol *13*, 219-225.

Zhang, W., and Shi, L. (2005). Distribution and evolution of multiple-step phosphorelay in prokaryotes: lateral domain recruitment involved in the formation of hybrid-type histidine kinases. Microbiology *151*, 2159-2173.

Chapter 5

Conclusions and future directions

Conclusions

In this work I use bacterial two-component signal transduction pathways to investigate the evolution of protein-protein interactions between signaling proteins and explore the mechanisms by which cells can expand their signaling repertoires via duplication and divergence. More specifically I have investigated the molecular mechanisms by which two-component pathways can become insulated at the level of signal transduction and identified selective pressures that act on the kinase/regulator interaction.

Previous work has shown that the interaction between kinase and regulator is incredibly specific, with most histidine kinases phosphorylating only a single, cognate, response regulator (Skerker et al., 2008; Skerker et al., 2005). This specificity is mediated through molecular recognition and determined by a limited set of residues on the kinase and regulator (Bell et al., 2010; Casino et al., 2009; Skerker et al., 2008). Focusing on this set of residues that dictate partnering specificity, I showed that the evolutionary trajectory towards insulation of duplicated two-component systems may be constrained by the need to maintain interaction between cognate kinases and response regulators while, at the same time, preventing unwanted interactions with the other two-component proteins in the genome (Chapter 2). Thus, after a duplication, there may only be a few accessible paths that result in insulation and the ability to successfully create two insulated pathways may be determined by the other two-component pathways already in the cell. Furthermore, I demonstrated that the evolution of these specificity residues is driven by selection against pathway cross-talk following gene duplication and that cross-talk between pathways leads to a selective disadvantage *in vivo* (Chapter 3). I determined that is possible to insulate pathways using only a limited number of mutations and that after a

duplication event, changes in the specificity residues of other two-component signaling proteins already in the genome may be necessary insulate all pathways within a cell (Chapter 3). This work may help to explain why some two-component pathways appear to be more easily duplicated or transferred than others.

In Chapter 4, I investigated covalent attachment as an alternative means of enforcing interaction specificity between a kinase and its cognate receiver domain. Hybrid kinases, in which the kinase and the receiver domains are covalently attached, represent almost 25% of all sequenced histidine kinases and nearly all eukaryotic histidine kinases, yet they remain less well understood than canonical pathways. I employed the same coevolution approach that had been used in canonical two-component pathways to identify specificity residues between kinase and receiver domain in hybrid histidine kinases and showed that there was significantly less coevolution in hybrid kinases. In hybrid histidine kinases, no changes to the specificity residues are necessary to accommodate a duplication; the high effective concentration of the covalently attached response regulator prevents cross-talk with the other cytoplasmic response regulators in the cell. Thus the covalent attachment of kinase to substrate may allow a two-component system to more easily integrate itself into the genome via duplication or horizontal gene transfer by preventing cross-talk with other two-component proteins already in the genome.

In short, the question of specificity in two-component signal transduction systems can be reduced to only a few residues that are sufficient to dictate the partnering specificity. The impressive fidelity in the kinase-substrate interaction observed in these systems is due to the significant selective disadvantage against cross-talk and the relatively limited number

of mutations necessary to change the specificity of the kinase/regulator interaction. This work sheds light on the apparent ease with which two-component signal transduction systems have expanded to become the dominant signaling system in bacterial genomes and, more generally, how a small number of gene families, through duplication and divergence, can be responsible for signal transduction in all organisms.

Future Directions

HPT specificity and expansion

In this work, I investigated the specificity between histidine kinase and response regulator, and shown that only a few mutations are sufficient to insulate pathways post-duplication. Furthermore, in Chapter 4, I demonstrated that covalent attachment of a receiver domain to the kinase is another mechanism to enforce specificity and that the kinase specificity residues of hybrid kinases are not under selective pressure to change after duplication. I proposed that this may allow hybrid histidine kinases to be more easily duplicated or transferred between genomes. While this is true, it ignores the downstream components of the phosphorelay pathway. After transferring intramolecularly from kinase to receiver domain, the hybrid histidine kinase then transfers to a histidine phosphotransferase (HPT) and finally to a response regulator which effects an output (Figure 5.1A). Although there are no crystal structures of hybrid histidine kinases, much less ones in complex with a HPT, the crystal structure of the response regulator Spo0F with the HPT Spo0B (Zapf et al., 2000) shows that the kinase/regulator specificity residues also map to the interface between response regulator and HPT (Skerker et al., 2008). If the interaction surface is conserved and the same

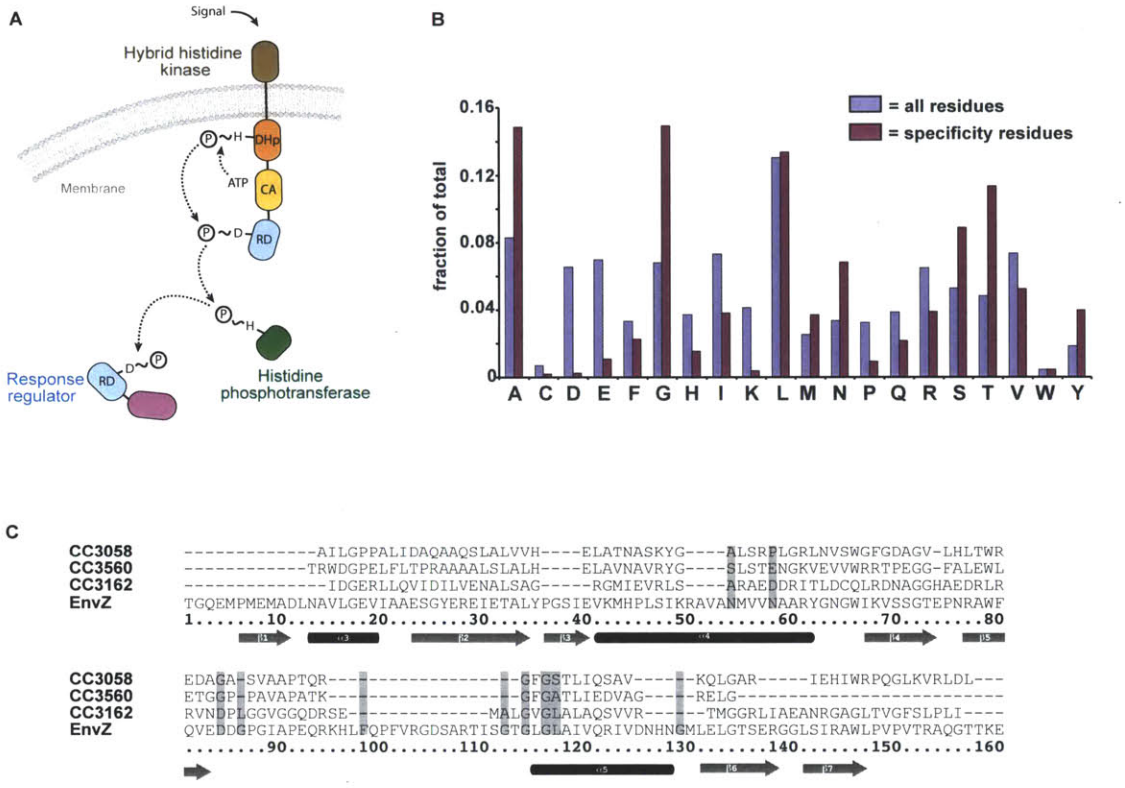


Figure 5.1 Evolution and specificity of HPT domains.

(A) Outline of a traditional phosphorelay pathway. The kinase/receiver domain interaction in the hybrid kinase was described in chapter 4, however after transferring intramolecularly, in order to reach the output of the pathway, two additional phosphotransfer steps are necessary: from the receiver domain of the hybrid kinase to a histidine phosphotransferase (HPT) and then from the HPT to the final receiver domain. (B) Graph showing the distribution of amino acids in the specificity residues of canonical histidine kinases compared to the amino acids that comprise the DHp domain. Within the specificity residues, there is an overrepresentation of small and nonpolar amino acids and an underrepresentation of large and charged amino acids. Whether this distributions represents constraints that are due to the interaction between kinase and receiver domain or constraints that are due to the need for the kinase to autophosphorylate is unknown. Data are taken from (Podgornaia and Laub, 2013) (C) Sequence alignment of the CA domains of 3 annotated hybrid kinases within the *C. crescentus* genome that appear to have degenerated with the *E. coli* kinase EnvZ. All three have homology to the HATPase_c_2 (CA domain) family according to PFAM, as well as to the response_reg family (receiver domain). Although the histidine is conserved, key residues in the CA domain that are important for autophosphorylation (shaded in gray), and that are conserved in all or most functional kinases are changed or missing. These degenerated hybrid kinases represent possible HPTs that may have been used to increase the signaling complexity in response to a large scale lineage specific expansion of hybrid kinases in *Caulobacter*.

specificity residues are used, what benefit is there to increasing the number of components in the pathway?

One possibility is that the amino acid composition of specificity residues in canonical kinases may be constrained. Indeed, certain amino acids are over and under represented in the specificity residues of canonical histidine kinases (Figure 5.1B) As I showed in chapter 2, certain substitutions in the specificity residues affect the autophosphorylation abilities of the kinase. An HPT, however, acts only as a shuttle for phosphate and has no catalytic abilities of its own. Perhaps using an HPT leads to new and unoccupied sequence space, defined by specificity residues that prevent autophosphorylation in a kinase, becoming accessible. Thus the hybrid kinase, whose specificity residues may be constrained by catalytic function, does not need to evolve new specificity residues in response to duplication. The HPT, whose specificity residues are unconstrained, can then find a new, orthogonal and previously unavailable, sequence space.

The second question that the proliferation of hybrid kinases in certain genomes brings up is where do HPTs come from? There is a defined HPT domain in the PFAM database based on the structure of the HPT domain from the *E. coli* hybrid kinase ArcB. However, almost any protein can act as an HPT as long as there is a histidine within a four helix bundle. This HPT model fails to identify several characterized HPTs, such as ChpT in *Caulobacter crescentus* (Biondi et al., 2006) and HPTs remain difficult to find. In nearly all cases there are many fewer identifiable HPTs in a genome than hybrid kinases. One possibility is that duplication of hybrid kinases could be used to create more complex signaling pathways—*i.e.* multiple hybrid kinases feed into a single HPT. While this may

be true in some systems, particularly the *Pseudomonas aeruginosa* virulence pathway, it is unlikely to be true in all cases.

One intriguing possibility for the generation of HPTs is that duplicated kinases or hybrid kinases can degenerate into HPTs. This is probably the origin of the *Bacillus* sporulation HPT Spo0B. In *C. crescentus*, which as I showed in chapter 4 has a large set of recently duplicated hybrid kinases, 3 of the annotated hybrid kinases are in missing key residues in the CA domain that are necessary for autophosphorylation (Figure 5.1C). Each is cytoplasmic and retains a histidine within an α-helix. Intriguingly, at least one of these "hybrid kinases" retains what appears to be an intact receiver domain, indicating that perhaps it transfers intramolecularly from the HPT to the receiver domain. Further biochemical assays are necessary to determine whether or not these degenerated hybrid histidine kinases represent actual HPTs, and whether the receiver domains are indeed functional.

Explorations of sequence space

In Chapters 2 and 3, I often reference the idea of sequence space in the kinase/regulator interaction. Experiments from both chapters demonstrate that single mutations in the specificity residues of either the kinase or the regulator often have little effect on interaction specificity. However, as also shown in both chapters, certain mutations can have profound effects. Predicting which mutations will fall into each category *a priori* is currently impossible. Even predicting which kinases will interact with which regulators purely by sequence has proved challenging, as most methods rely on phylogeny rather than biochemical principles (Burger and van Nimwegen, 2008; Procaccini et al., 2011).

There are two approaches that can be taken to begin to answer this question of sequence space. The first approach is structure based. Currently there is only one structure of a kinase in complex with its cognate response regulator. Although based on the coevolution studies and the co-crystal structure of Spo0F and Spo0B (Zapf et al., 2000), the interface appears to be conserved between kinase/regulator pairs, how a point mutant might affect the interaction surface remains unclear. In addition, the specificity residues used by extant two-component systems include a wide range of amino acids. How are different charges or sizes accommodated along the same interface? What are the biochemical principles that determine whether or not a kinase will interact with a given response regulator? By swapping the specificity residues of the *Thermotoga maritima* two-component pair that comprises the co-crystal structure (Casino et al., 2009) with those of alternative two-component systems, it may be possible to achieve alternative co-crystal structures that may elucidate the roles of individual point mutations on specificity and ability to interact, as well as determine how alternate specificity residues might be accommodated along the same interface.

The second approach is library based. In chapters 2 and 3 of this thesis, I made a large number of directed mutations and determined whether or not these mutations change specificity. This approach can be extended to a library based approach, where instead of introducing each mutation individually, all possible combinations of specificity residues are introduced into via a single library. An *in vivo* based screen for kinase/regulator interaction, such as a fluorescent reporter behind a promoter that is activated by a given response regulator (Figure 5.2), can then be used to identify the kinases from the library that preserve the interaction with the cognate response regulator. Illumina sequencing can

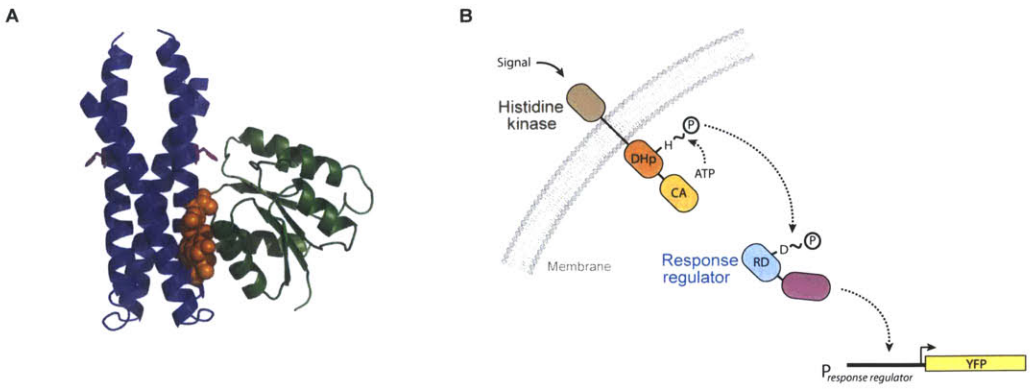


Figure 5.2 Library screen to determine sequence space.

(A) The specificity residues of a given kinase (shown in orange) can be replaced with random amino acids. As the specificity residues are close together in the primary amino acid sequence, all six important specificity residues can be changed at once. This library of histidine kinases can be introduced via plasmid and then screened for function. (B) By creating a fluorescent reporter, the activity of the two-component system can be monitored via microscopy or FACs. Most two-component systems autoregulate their own expression, and thus for two-component systems for which the regulon is unknown, the promotor of the operon containing the histidine kinase and response regulator can be used for the reporter fusion. Most kinases are bifunctional, and thus in order to be classified as wild-type, the kinase must be able to activate the promoter in response to a signal and to turn off the reporter in the absence of the signal. Cells that possess a kinase with these functions can be sorted and the specificity residues of the kinase sequenced in order to determine the sequence space of the kinase/regulator pair.

then be used to identify all combinations of specificity residues on the kinase that enable the interaction between kinase and regulator. This high-throughput screen will allow for the comprehensive definition of sequence space for a kinase/regulator pair. By repeating this screen using different two-component pairs within a given genome, the questions of how sequence space is divided between different two-component systems can begin to be addressed.

Furthermore, by employing the library approach and gaining a comprehensive understanding of the distribution of specificity space between two-component pairs within a genome, it may be possible to gain a deeper understanding as to the mechanisms

by which duplications of two-component systems are accommodated in the genome. Gene duplication and divergence has been well-studied, and several models have been proposed to explain the fate of genes post-duplication: (i) nonfunctionalization, where one of the duplicated copies becomes non-functional due to mutational accumulation, (ii) neofunctionalization, where one copy is now free to gain a novel function while the other retains the ancestral function, and (iii) subfunctionalization, where each copy retains a subset of the ancestral function (Force et al., 1999; He and Zhang, 2005; Kimura and Ota, 1974; Lynch, 2002; Lynch and Conery, 2000; Lynch and Force, 2000).

Two types of subfunctionalization have been defined. The first type, and the original model, is known as duplication-degeneration-complementation (DDC). In DDC, after a duplication event, the ancestral function is divided between the two extant copies. This means, that post-duplication, both copies are required to carry out the same function as the ancestral copy. The second type of subfunctionalization, which could also be thought of as a hybrid neo- and subfunctionalization model, is escape from adaptive conflict (EAC). In EAC, the ancestral gene has two or more functions. However, when the functions are encoded in the same protein, the protein cannot be optimized for either function (Hittinger and Carroll, 2007). Duplication and divergence allows the paralogs to become specialized and to optimize each of the duplicates to become optimized for the two ancestral functions. In this case, the functions of the two duplicated copies represent a gain of function over the ancestral copy. While subfunctionalization has been shown to be more prevalent, the most commonly studied "function" is expression patterns (Force et al., 1999; Huminiecki and Wolfe, 2004; Lynch and Force, 2000). In multicellular organisms subfunctionalization is often accomplished through the degeneration of

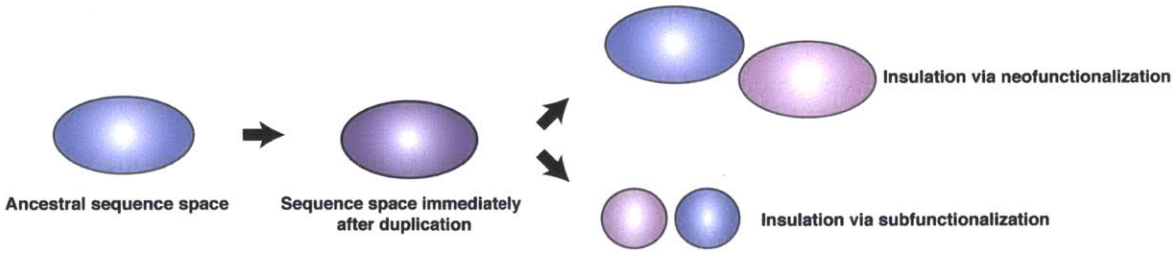


Figure 5.3 Two models for insulation of pathways post-duplication.

The ancestral two-component pair occupies some defined sequence space. Immediately after duplication, the specificity residues of, and thus the sequence spaces occupied by, the duplicated kinases are identical. Over time, in order to insulate pathways, the specificity residues diverge and the sequence spaces become insulated. In the case of neofunctionalization, one two-component pair occupies the ancestral sequence space, while the other two-component pair finds a new, unoccupied, sequence space. In the case of subfunctionalization, the two-duplicated pairs now occupy a subset of the ancestral sequence space.

promoter elements, leading to differential tissue-specific expression patterns (Force et al., 1999; McClintock et al., 2002; Prince and Pickett, 2002; Tümpel et al., 2006).

Groups have found evidence for both subfunctionalization (Conant and Wolfe, 2007; Tuch et al., 2008; Wapinski et al., 2007b; Wapinski et al., 2010) and neofunctionalization (He and Zhang, 2005; Kellis et al., 2004; Tirosh and Barkai, 2007; Wapinski et al., 2007a, b) in regulatory networks of single-celled organisms post-duplication. However, many of the studies of protein-protein interactions of duplicates were done in *Saccharomyces cerevisiae*, where the occurrence of the whole-genome duplication could change the evolutionary landscape post-duplication (Guan et al., 2007; Tirosh and Barkai, 2007). In addition, the "function" of a gene is not always easy to define. The evolutionary history of duplicated pathways, where the "function" is the ability to faithfully transmit a signal, has not yet been studied.

In the context of kinase-regulator interaction, where function is determined by sequence space occupied by a given two-component system, neofunctionalization vs.

subfunctionalization can be assessed by determining the distribution of ancestral sequence space post-duplication (Figure 5.3). Based upon the models, if subfunctionalization is more prevalent, it would be expected that the older and more often duplication two-component pairs would occupy more constrained sequence spaces, as each duplication would result in a narrowing of the sequence space. Subfunctionalization may be easier, as the duplicated two-component systems wouldn't have to traverse new sequence space that may be occupied by other two-component pairs in the genome. However, the work in chapter 3 showed that it is possible to insulate pathways, even if insulation involves interacting with other two-component pairs in the genome. It is unclear, and will remain unknown until sequence space of a genome is mapped, how narrowly defined a sequence space can be. Is there a limit to the number of times that a two-component system can duplicate and diverge via subfunctionalization?

Although high-throughput library based approaches are necessary in order to define the size and shape of sequence space, to accurately access which model of duplication and divergence occurred in a given duplication event, detailed ancestral reconstructions are necessary. Ancestral reconstructions rely on maximum likelihood methods to determine the most likely sequence of the ancestral proteins. As the function investigated is specificity, reconstructions can be limited to the amino acid content of the specificity residues. By performing ancestral reconstructions on relatively recent duplications and testing the phosphotransfer specificity of these proteins via phosphorylation experiments, it will be possible to determine which model of duplication and divergence has governed that particular duplication event. Additional questions, including the number and type of mutations that are necessary to insulate the pathways and how long cross-talk persists in a

duplicated system can also be answered. However, the speed of change in gene content, the long branch lengths, and hard to resolve phylogenies between species may make true ancestral reconstruction difficult.

Sequence space in the response regulator/DNA interaction

In this work, I have outlined the mechanisms by which a two-component system can become insulated at the level of signal transduction from other two-component systems also in the genome. However, the ultimate purpose of duplication or horizontal gene transfer of signaling proteins is to expand the signaling capacity of a cell. In order for this to be achieved, the output response of the signaling pathway has to change after entry into the genome and specificity, although in this case the specificity refers to the response regulator/DNA interaction, has to be ensured. Changes in the output response of one or both response regulators is likely a critical step in the establishment of new functions and the maintenance of the duplicated copies within the genome. Yet, there have been few global studies of the DNA binding specificity of response regulators. Very little understanding of molecular basis by which response regulators gain new targets after duplication, or even the distribution of DNA binding sites and the specificity of the response regulator–target gene interaction on a global level exists.

Understanding the specificity of the response regulator-target gene interaction is an important one, as response regulators control a large variety of bacterial cellular processes. Although certain response regulators and their regulons are well studied, for example CtrA in *Caulobacter crescentus*, many more remain unstudied at a genome-wide level. Systematic determination of the binding sites for all response regulators in a given

genome will allow for the exploration of several important questions. First, how many genes do most response regulators regulate? How much cross-regulation is there at the output level? For genes with multiple response regulators serving as regulators, do the response regulators bind to the same place in the promoter, thus constraining the DNA-binding specificities of these response regulators, or do they have separate and independent binding sites? The second question is the distribution of response regulator/DNA-binding specificity within a genome. Like in the kinase/regulator interaction, paralagous response regulators often share a high sequence and structural homology. How are so many related response regulators able to coexist in the same genome and activate genes in a specific way? How much information is encoded in the DNA binding motif of a given response regulator and how similar or different are the DNA binding motifs of related response regulators?

The difficulty in identifying specific signals for a given two-component system has prevented the systematic study of the response regulator/DNA interaction. Even in *E. coli*, which has perhaps the best-studied set of two-component systems, the signals responsible for activating many of these systems remains unclear. As most response regulators are only active after phosphorylation-induced dimerization, it is necessary to activate the pathway enough to significantly phosphorylate the response regulator in order to determine DNA-binding preferences. One way to get around this problem is to look at *in vitro* rather than *in vivo* binding (Rajeev et al., 2011). By examining DNA-binding *in vitro*, the response regulators can all be phosphorylated to similar levels using

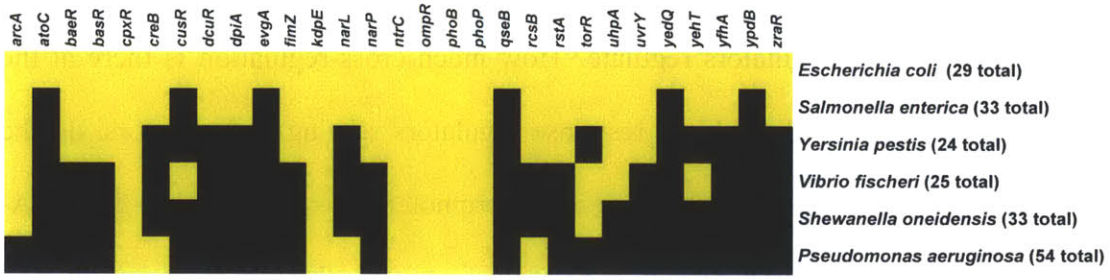


Figure 5.4 Distribution of *E. coli* response regulators in a set of well-studied γ-proteobacteria.

A chart showing presence (yellow) or absence (black) of a given *E. coli* response regulator in the genomes shown at right. The total number of DNA-binding response regulators in each genome are also listed. These genomes represent a wide distribution of number types of response regulators within the γ-proteobacteria. Response regulators whose orthologs are absent in *E. coli* are not shown. Duplications in species other than *E. coli* are also not shown on this diagram.

small molecules or the cognate kinases. Several methods exist to interrogate the binding properties of a DNA-binding protein, including HiTS-FLIP (Nutiu et al., 2011), MITOMI (Fordyce et al., 2010) and *in vitro* CHIP-seq. Results from these *in vitro* methods can be compared to *in vivo* CHIP-seq for those two-component systems whose signals are known.

One way to extend this work is to look at DNA-binding specificities of orthologs and paralogs, instead of focusing on a single genome. How do the DNA-binding specificities differ between orthologs? And how do they change after a duplication event? Previous work into the evolution of the PhoP response regulator regulon has demonstrated that while the direct targets of PhoP are highly variable across species, the consensus binding site is remarkably conserved (Perez et al., 2009). Likewise, although members of the PhoB regulon can, and do, vary, PhoB boxes are conserved across bacteria. However, Both PhoP and PhoB have been vertically inherited, and neither has undergone duplication in a lineage in which they have been studied. However, the histories of most

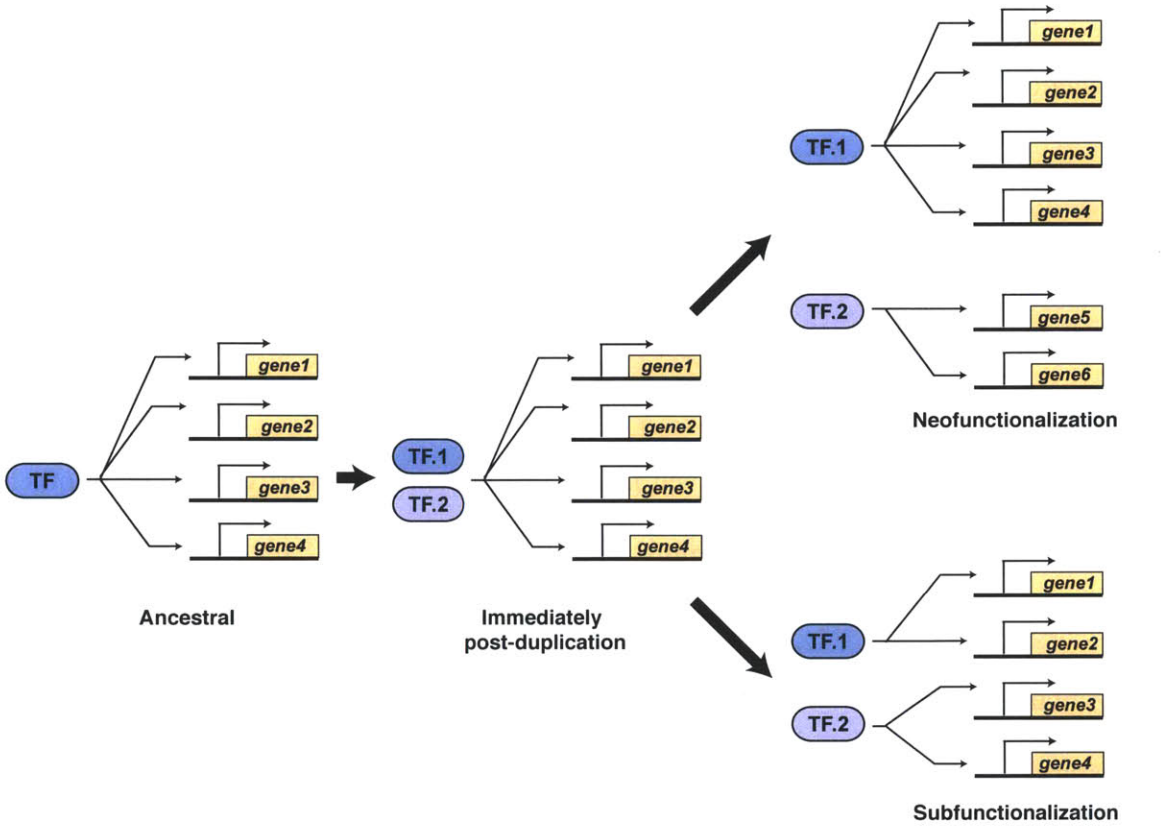


Figure 5.5 Evolution of transcriptional networks post-duplication.

The evolution of transcriptional networks after duplication of a transcription factor can also be described using the models of neo and subfunctionalization. In this case, the function refers to the identity of the target genes. Due to the speed of transcriptional rewiring, it is unlikely that any response regulator duplication would fall neatly into either category. By focusing on the core set of genes, however, it may be possible to divide duplications into one category or the other.

DNA-binding response regulators are not so easy to discern, and there is a great deal of gain and loss even within the γ-proteobacteria (Figure 5.4). How do the transcriptional networks controlled by orthologous response regulators change between species? One recent study looked at the regulons of OmpR in two different *Salmonella* strains, *S. typhi* and *S. typhimurium* and found only a small subset of conserved targets between the two (Perkins et al., 2013). From previous work, it seems as if transcriptional rewiring is rapid, although a core set of genes are often conserved. By looking at this core set of genes in

response regulators that have duplicated, it will be possible to determine whether neo- or subfunctionalization of target genes is more frequent (Figure 5.5).

Extending the analysis of DNA binding specificities of response regulators to multiple genomes would allow for the following questions to be answered: **(1)** are the DNA binding motifs of most orthologous response regulators conserved, **(2)** are the trends observed in the *E. coli* response regulator set in terms of distribution of binding specificity the same as those observed in other organisms with vastly different numbers of DNA binding response regulators and **(3)** what happens to the DNA binding specificity of a response regulator after a duplication or after entering a genome via horizontal gene transfer. As only purified response regulators and genomic DNA are necessary for *in vitro* methods to ascertain DNA-binding specificities, it should be relatively simple to obtain a comprehensive data set from multiple species that represent a wide variety of ecological and phylogenetic niches.

Concluding remarks

Cells have developed a remarkable capacity to sense and interact with their environment. This impressive complexity is mediated through only a small number of different types of signaling pathways that have expanded via duplication and divergence. Understanding how multiple paralagous pathways can coexist in a genome while at the same time maintaining signaling fidelity is an important question in understanding the evolution of complexity. How are input/output relationships maintained? How is a new pathway incorporated into the signaling network of the cell? Is there a limit to the number of copies of a given type of signaling pathway that can be encoded in a single cell? Bacterial

two-component systems represent a tractable model for studying the expansion of signaling proteins and the mechanisms of duplication and divergence employed in order to generate complexity. In this work I have shown, in at least one case, the evolutionary pressures that dictate how, and if, a duplicated two-component pair is integrated into the signaling network of the cell.

In addition, much of the recent work in synthetic biology has focused on building new signaling pathways. Two-component pathways make appealing targets for synthetic biology approaches due to the wide range of signals that are sensed by histidine kinases and the modularity of these systems. However, the interplay between the synthetic pathways and the naturally occurring signaling pathways has been ignored. I have shown that the introduction of a new two-component system into a genome can affect the ability of pre-existing two-component systems to function. In order for a new, synthetic, pathway to behave as expected it needs to occupy an orthogonal space from the native pathways. Understanding of the distribution of sequence space will help with the success of synthetic approaches to biology using two-component systems.

Understanding the mechanisms by which a new signaling pathways can integrate itself into the signaling network of the cell and allow the cell to respond to novel signals will help to shed light on the processes by which cells are able to gain complexity and also help with efforts to engineer new signaling pathways.

References:

Bell, C., Porter, S., Strawson, A., Stuart, D., and Armitage, J. (2010). Using structural information to change the phosphotransfer specificity of a two-component chemotaxis signalling complex. PLoS Biol *8*, e1000306.

Biondi, E.G., Reisinger, S.J., Skerker, J.M., Arif, M., Perchuk, B.S., Ryan, K.R., and Laub, M.T. (2006). Regulation of the bacterial cell cycle by an integrated genetic circuit. Nature *444*, 899-904.

Burger, L., and van Nimwegen, E. (2008). Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Mol Syst Biol *4*, 165.

Casino, P., Rubio, V., and Marina, A. (2009). Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell *139*, 325-336.

Conant, G., and Wolfe, K. (2007). Increased glycolytic flux as an outcome of whole-genome duplication in yeast. Mol Syst Biol *3*, 129.

Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. Genetics *151*, 1531-1545.

Fordyce, P.M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J.L., and Quake, S.R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nat Biotechnol *28*, 970-975.

Guan, Y., Dunham, M., and Troyanskaya, O. (2007). Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics *175*, 933-943.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics *169*, 1157-1164.

Hittinger, C.T., and Carroll, S.B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. Nature *449*, 677-681.

Huminiecki, L., and Wolfe, K. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res *14*, 1870-1879.

Kellis, M., Birren, B., and Lander, E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature *428*, 617-624.

Kimura, M., and Ota, T. (1974). On some principles governing molecular evolution. Proc Natl Acad Sci U S A *71*, 2848-2852.

Lynch, M. (2002). Genomics. Gene duplication and evolution. Science *297*, 945-947.

Lynch, M., and Conery, J. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151-1155.

Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. Genetics *154*, 459-473.

McClintock, J., Kheirbek, M., and Prince, V. (2002). Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development *129*, 2339-2354.

Nutiu, R., Friedman, R.C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G.P., and Burge, C.B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat Biotechnol *29*, 659-664.

Perez, J.C., Shin, D., Zwir, I., Latifi, T., Hadley, T.J., and Groisman, E.A. (2009). Evolution of a bacterial regulon controlling virulence and Mg(2+) homeostasis. PLoS Genet *5*, e1000428.

Perkins, T.T., Davies, M.R., Klemm, E.J., Rowley, G., Wileman, T., James, K., Keane, T., Maskell, D., Hinton, J.C., Dougan, G., *et al.* (2013). ChIP-seq and transcriptome analysis of the OmpR regulon of Salmonella enterica serovars Typhi and Typhimurium reveals accessory genes implicated in host colonization. Mol Microbiol *87*, 526-538.

Podgornaia, A.I., and Laub, M.T. (2013). Determinants of specificity in two-component signal transduction. Curr Opin Microbiol.

Prince, V., and Pickett, F. (2002). Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet *3*, 827-837.

Procaccini, A., Lunt, B., Szurmant, H., Hwa, T., and Weigt, M. (2011). Dissecting the specificity of protein-protein interaction in bacterial two-component signaling: orphans and crosstalks. PLoS One *6*, e19729.

Rajeev, L., Luning, E.G., Dehal, P.S., Price, M.N., Arkin, A.P., and Mukhopadhyay, A. (2011). Systematic mapping of two component response regulators to gene targets in a model sulfate reducing bacterium. Genome Biol *12*, R99.

Skerker, J., Perchuk, B., Siryaporn, A., Lubin, E., Ashenberg, O., Goulian, M., and Laub, M. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*, 1043-1054.

Skerker, J., Prasol, M., Perchuk, B., Biondi, E., and Laub, M. (2005). Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol *3*, e334.

Tirosh, I., and Barkai, N. (2007). Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. Genome Biol *8*, R50.

Tuch, B., Galgoczy, D., Hernday, A., Li, H., and Johnson, A. (2008). The evolution of combinatorial gene regulation in fungi. PLoS Biol *6*, e38.

Tümpel, S., Cambronero, F., Wiedemann, L., and Krumlauf, R. (2006). Evolution of cis elements in the differential expression of two Hoxa2 coparalogous genes in pufferfish (Takifugu rubripes). Proc Natl Acad Sci U S A *103*, 5419-5424.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007a). Automatic genome-wide reconstruction of phylogenetic gene trees. Bioinformatics *23*, i549-558.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007b). Natural history and evolutionary principles of gene duplication in fungi. Nature *449*, 54-61.

Wapinski, I., Pfiffner, J., French, C., Socha, A., Thompson, D., and Regev, A. (2010). Gene duplication and the evolution of ribosomal protein gene regulation in yeast. Proc Natl Acad Sci U S A *107*, 5505-5510.

Zapf, J., Sen, U., Madhusudan, Hoch, J., and Varughese, K. (2000). A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction. Structure *8*, 851-862.