Pervasive degeneracy and epistasis in a protein-protein interface

by

Anna Igorevna Podgornaia

B.A. Molecular and Cell Biology
University of California, Berkeley (2009)

SUBMITTED TO THE GRADUATE PROGRAM IN COMPUTATIONAL AND SYSTEMS BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2014

Signature of Author:__ Signature redacted ____________________
Anna Igorevna Podgornaia
Graduate Program in Computational and Systems Biology
July 21, 2014

Certified by: _______ Signature redacted ____________________
Michael T. Laub
Associate Professor of Biology
Thesis supervisor

Accepted by: _______ Signature redacted ____________________
Christopher B. Burge
Professor of Biology and Biological Engineering
Director of Computational and Systems Biology PhD program

Pervasive degeneracy and epistasis in a protein-protein interface

by

Anna Igorevna Podgornaia

Submitted to the Graduate Program in Computational and Systems Biology
on July 21, 2014 in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at the Massachusetts Institute of Technology

ABSTRACT

Signal transduction pathways rely on transient yet specific protein-protein interactions. How a limited set of amino acids can enforce cognate protein interactions while excluding undesired pairings remains poorly understood, even in cases where the contacting residues have been identified on both protein partners. To tackle this challenge, I performed structure-guided and library-based mutagenesis studies of bacterial two-component signaling pathways. These pathways, typically consisting of a histidine kinase and a response regulator, are an ideal model system for studying protein-protein interactions as they rely almost exclusively on molecular recognition for specificity. The kinase uses a limited set of residues to recognize the regulator in both phosphorylation and dephosphorylation reactions, and to prevent docking with all non-cognate regulators.

In this thesis I characterized the extent to which interface residues in two-component signaling proteins can be modified without changing the overall behavior of the pathway. In collaboration with another research group I have performed a mutagenesis study of a two-component system from *Thermotoga maritima* that has proven amenable to structural analysis. By solving the co-crystal structure of a histidine kinase and response regulator containing interface residues from a different interacting pair, we learned the biophysical basis for accommodating these new residues. To understand how many different residue combinations can support a functional interaction, I comprehensively mapped the sequence space of the interface formed by *Escherichia coli* histidine kinase PhoQ and its partner PhoP. I used a robust high-throughput assay to screen a library of 20^4 (160,000) PhoQ variants in which I had completely randomized the four key specificity-determining residues. Using deep sequencing, I identified ~1,600 (1 %) variants that can phosphorylate and dephosphorylate PhoP as well as the wild-type PhoQ. Strikingly, PhoQ can interact with PhoP via many sets of interfacial residues that are completely different from the wild type. This combinatorial approach to mapping sequence space revealed interdependencies between individual amino acids, illustrating its power relative to screens that only examine substitutions at individual sites. This thesis provides a framework for mapping the sequence space of histidine kinases and has broad implications for understanding protein-protein interaction specificity and the evolution of bacterial signaling pathways.

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

ACKNOWLEDGEMENTS

"Одна голова – хорошо, а две – лучше" (Russian proverb)

Thank you Michael Laub for your support at every stage of my PhD project. I really appreciate your optimism and enthusiasm for my work. Also, Sadie's artwork is very nice.

Thank you to my committee members, Amy Keating and Aviv Regev, for asking intimidating questions and pushing me to think harder about various aspects of my project. Thank you Dane Wittrup for serving on the defense committee. Angela Brooks and Ali Mortazavi, I really appreciate you as mentors, for teaching me about bioinformatics as an undergraduate student and always being so excited about my progress.

Members of the Laub lab have made it a great place to work over four years. Thank you Emily Capra for teaching me how to do science and always reminding me to step back and look at the big picture. Thank you Barrett Perchuk – it was valuable to learn experimental techniques from you, but even more valuable to learn how to be a reasonable adult. The early members of the lab – Kasia, Erin, Christos, Orr, Kristina – inspired me to join them and work really hard. Thank you Joshua Modell for the lab cheer. "Remember, labs are temporary, but bay mates are for life". Thank you Chris and Diane for going on this journey with me – you are some of the smartest and nicest people that I know. Thank you Tung, Anjana, Kyle, and Leonor for being wonderful lab mates. Thank you Katie, Salazar, and Conor for being great rotation students and continuing the good fight (studying specificity). Thanks to my numerous rotation students – you do not appreciate just how little you know until you need to train others. Finally, thank you to Peter for making me look fashionable by comparison with your ripped jeans. All the members of the Laub lab have contributed to my scientific growth in different ways and your hard work inspires me.

Thank you Chris Burge, the CSB program, and especially my CSB2009 classmates. It's been a lot of fun going through grad school with you guys! MIT Building 68 is a great place for basic research and I am especially grateful for the 5th floor labs as the best intersection of microbiology and protein biochemistry.

My PhD has progressed swimmingly thanks to administrative help and emotional support from Bonnielee Whang, Jacquie Carota, and Sally MacGillivray. My experiments benefitted greatly from technical help at the Koch flow cytometry facility and the BioMicro center.

Outside of lab I thank my new East Coast family, Linda and Phil, for nice dinners and great company. Also thank you to Marjon Moulai for dragging me out of the lab to do fun things.

Thank you to my parents, Natasha and Igor, for your continuous support and concern for my overall well being. Thank you Brian Fiske for everything.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction: The sequence space of two-component signaling pathways

A subset of this chapter was published as Podgornaia, A.I., Laub, M.T. *Current Opinion in Microbiology 16*, 156-162, April 2013.

The advent of high-throughput DNA synthesis and sequencing technologies has revolutionized biological research. Despite the abundance of -omic datasets appearing every day, the relationship between protein sequence and protein function remains poorly understood. How robust is a particular protein to mutations? How many different sequences can encode a given biochemical activity? Has nature sampled all of these possible combinations over the course of evolution? Answers to these questions can be found by mapping protein sequence space (Smith, 1970), or the complete set of amino acid sequences that can generate a given function. However, making such a map for just one protein requires assessment of 20^N sequences for every N positions of interest. Because this can be experimentally challenging and sometimes even impossible to achieve, fundamental questions about protein sequence-function relationships remain largely unanswered.

Protein sequence-function relationships

One function that is often encoded by a protein sequence is the ability to participate in protein-protein interactions, which are crucial for many cellular activities. Permanent protein associations can lead to the formation of higher order oligomers used for catalytic or structural purposes, as in the case of microfilaments formed by actin subunits. Protein-protein interactions can also be used for the creation or sequestration of a binding site (Phizicky and Fields, 1995), such as when an inhibitor covers an active site and prevents its use. Transient protein-protein interactions are prevalent in phosphorylation-based regulatory pathways, including the signal transduction activities of both eukaryotic and prokaryotic cells. The binding strength and orientation necessary for proper interactions between two proteins, as well as deterrents to non-specific binding in the crowded cellular environment, are ultimately encoded by the primary protein sequence.

To understand the relationship between protein sequence and function, one can introduce a mutation into the protein and analyze how the activity changes relative to wild type. The first functional studies of proteins mutagenized in the lab came in the late 1960's with the advent of solid-phase protein synthesis, which allowed specific amino acids to be incorporated at a precise position in a protein chain. In an early application of the method, the point mutations F120A and F120Y in ribonuclease confirmed that residues that are too small or too big will both alter the protein's activity (Hodges and Merrifield, 1974). Subsequent advances in molecular biology have facilitated amplification and mutagenesis of DNA, which in turn allows rapid and easy creation of protein variants with any number of amino acid substitutions. Introducing mutations in pairs allows researchers to construct double mutant cycles and understand whether the effects of the mutations are independent or coupled. An early instance of this analysis was applied to the barnase-barstar complex, which had been crystallized and has a very tight K_d of 10^{-14} M (Schreiber and Fersht, 1995). The study showed that residues found within 7Å of each other at the interface exhibit significant coupling energies.

Although glycine is the smallest amino acid, alanine substitutions are most often used to query the importance of a particular residue at its position in the protein. Unlike glycine, an alanine substitution removes all side-chain atoms past the beta carbon without compromising backbone structure (Morrison and Weiss, 2001). The technique of alanine scanning, or assaying the effect of an alanine substitution on a set of functional residues, has become a widespread assay. In one exemplary study using this method, the authors introduced single alanine substitutions at 62 positions in human growth hormone (hGH) and identified the precise residues responsible for its specific binding to the hGH receptor (Cunningham and Wells, 1989). In a high throughput extension of this assay, eleven residues in λ repressor were substituted with alanine in various combinations to produce 2,048 proteins (Gregoret and Sauer, 1993). By comparing the prevalence of individual versus paired mutations, the investigators found that seven out of fifty five residue pairings were found to have significant effects. Their data indicate that substitutions to a protein-protein interface change the interaction strength in a non-additive way. These findings imply that the sequence space for a given protein is described not just by the biochemical identity of a particular mutation, but also by higher-order dependencies with other residues in the protein.

The effects of mutations on protein-protein interactions can be assayed using a variety of different *in vitro* biochemical methods. Although there are too many to describe in depth here, these methods can be grouped into direct and indirect approaches (Kastritis and Bonvin, 2013). As the name would suggest, "direct" methods directly measure the concentrations of bound and unbound proteins of interest. Examples of these methods include analytical ultracentrifugation and separation by electrophoresis. "Indirect" methods measure a signal associated with the binding reaction, such as surface plasmon resonance or optical methods that monitor fluorescence. A number of methods have also been developed to study protein-protein interactions *in vivo*, such as Förster resonance energy transfer (FRET) between two fluorescently labeled interaction partners or the yeast two-hybrid system that provides a transcriptional read-out for an interaction (Phizicky and Fields, 1995). The application of these methods has resulted in the finding that the strengths of physiological protein-protein interactions span many orders of magnitude, from a K_d of 10^{-15} M for biotin-avidin to non-specific interactions with a K_d in the high millimolar range (Harding and Rowe, 2010; Livnah et al., 1993).

High-throughput protein mutagenesis studies

Although *in vitro* biochemical methods for assaying protein-protein interactions have been very important for biological research, they are laborious and often not amenable to higher throughput. Improvements in DNA synthesis and DNA sequencing have paved the way for library-based approaches that allow for faster screening of a large number of mutagenized proteins. The basic approach of these high-throughput methods is summarized in Figure 1.1. The three key steps common to these approaches are generation of library diversity, screening for the desired function, and analysis of the selected variants (Jäckel et al., 2008).

I. Generation of library diversity

The first step in a high-throughput mutagenesis study is the generation of a diverse protein library (Fig. 1.1), usually by changing the underlying nucleotide sequence. Mutations can be introduced randomly throughout the coding sequence using a variety of approaches, including chemical mutagens or UV radiation, error-prone polymerases, and bacterial mutator strains (Neylon, 2004). Changes to a specific position or region in a protein can likewise be achieved using synthetic DNA. The latter is used as degenerate oligos for PCR or to replace a region of a protein together with traditional molecular cloning techniques. Another targeted approach is homologous recombination, which creates new proteins variants by making chimeric molecules from segments of homologous proteins. This approach produces novel characteristics not found in the original versions and has been successfully used by the Arnold lab to generate proteins with improved stability for thermal or mechanical fluctuations (Trudeau et al., 2013).

Although it has been argued that the genetic code has been optimized to favor amino acid substitutions most likely to retain function, it makes coverage of the entire possible protein

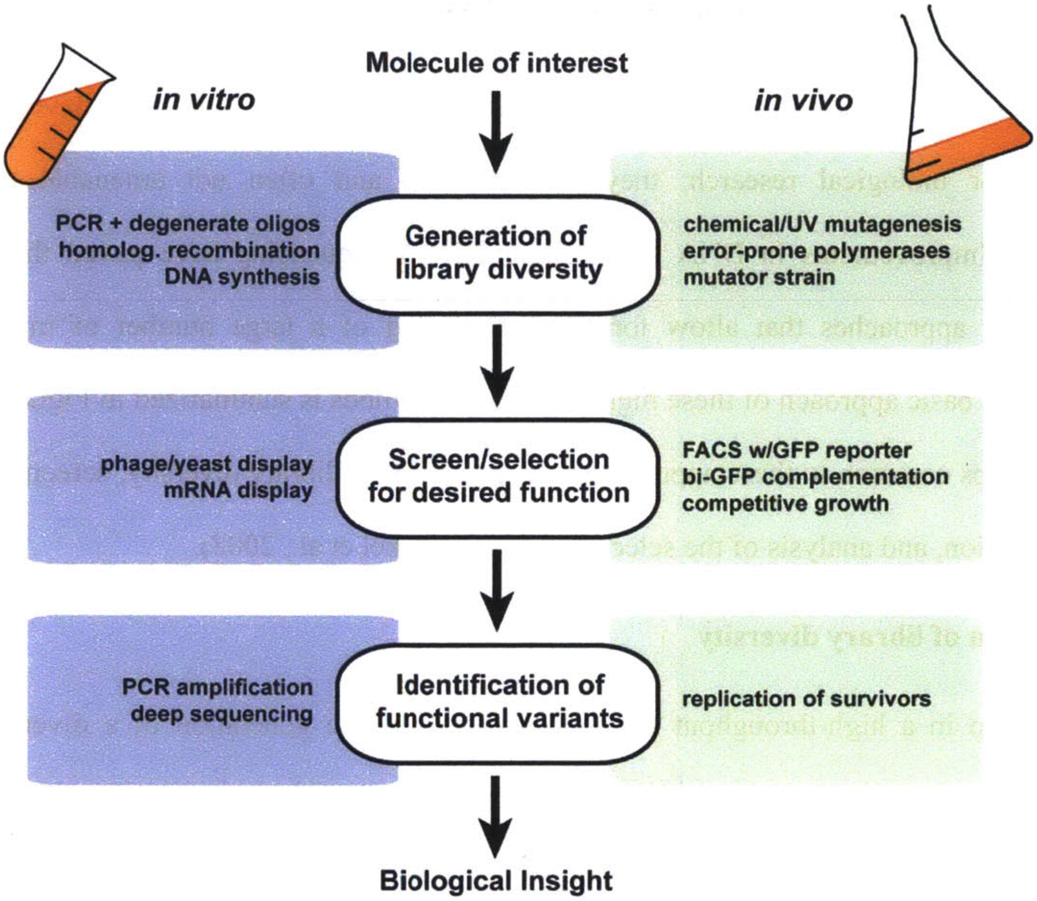


Figure 1.1 Experimental pipeline for high-throughput protein mutagenesis studies.

Schematic illustrating the three main stages in high-throughput studies of protein function: generation of library diversity, screening for the desired function, and analysis of the selected variants protein design. A number of *in vivo* and *in vitro* approaches are available for each stage and these can be mixed and matched in accordance with the overall experimental design. PCR: polymerase chain reaction. FACS: fluorescence-activated cell sorting. GFP: green fluorescent protein. Adapted from (Jäckel et al., 2008).

sequence space more difficult (Neylon, 2004). One approach to reduce this complexity is to substitute the 64 'NNN' codons with 32 'NNS' (S = C/G), which encode all 20 amino acids using fewer DNA sequences. Another way to circumvent the problem is to use mixtures of tri-nucleotides rather than single nucleotides for gene synthesis, thus reducing each amino acid to a single DNA codon. Although the relevant chemical protocol (TRIM) (Virnekäs et al., 1994) has been available for twenty years, its cost and complicated methodology has precluded its widespread use. A similar approach, called the MAX method of randomization, employs 20 different primers to incorporate the different codons into a location (Hughes et al., 2003). However, this method requires a large number of PCR steps with the potential to introduce mutations at other sites in the protein and complicate downstream analysis.

User-defined DNA sequences can be synthesized using the technology developed for printing oligonucleotide-based microarrays, but the total number is limited to ~50,000 probes (Baker, 2011). This approach does not currently have widespread use because it suffers from high error rates and the potential for cross-hybridization between adjacent clusters on a microchip (Kosuri et al., 2010). This method was used in 2012 to synthesize 55,000 oligos in order to write a "book" using DNA (Church et al., 2012). George Church, the senior author on this project, referred to the parallel advances in DNA synthesis and DNA sequencing as a Catch-22: "'If you do high-throughput synthesis on chips, you have a bottleneck in sequencing. If you use high-throughput sequencing [to verify low-throughput synthesis], you have a bottleneck in making the oligos'" (Baker, 2011). As these technologies continue to improve, experimental screening of the entire library, rather than generation of library diversity, will become the bottleneck in protein mutagenesis projects.

II. Screens and selections for library function

Arguably the most important and experimentally challenging step in a high-throughput study is screening or selection for variants that possess the desired trait (Fig. 1.1). In a screen each individual variant is tested for activity, whereas under selection all variants are subjected to an experimental condition simultaneously and the survivors retained. *In vivo* studies have the advantage of measuring protein-protein interactions in their native cellular environment. If the protein interaction of interest leads to transcriptional changes, then using a fluorescent reporter with fluorescence-activated cell sorting (FACS) can enable a rapid screening protocol. FACS can also be used to rapidly screen cells engineered with tags for bi-fluorescence complementation. In this method, one half of a green fluorescent protein (GFP) molecule is tagged to each interaction partner, and a sufficiently strong interaction results in the reconstitution of a functional, fluorescent protein molecule (Hu et al., 2002).

In vivo systems can also be adapted for fitness selections on large populations of cells. In a recent study, a yeast population harboring 180 unique point mutations in the essential chaperone Hsp90 was grown at an elevated non-permissive temperature (Hietpas et al., 2011). The mutations were introduced at nine positions, including two exposed residues predicted to bind to hydrophobic regions on unfolded proteins. The authors used this medium-throughput approach to calculate fitness coefficients for the different amino acid substitutions and found that most positions could accommodate varied amino acids. For example, the substitution G584F results in fitness comparable to wild type but could not have been predicted on the basis of amino acid side chain similarity. Furthermore, the diversity of allowed amino acid substitutions is much greater than the diversity predicted by the phylogenetic record for Hsp90.

If the library screen is conducted *in vitro*, it is important to maintain a connection between the protein being assayed and the underlying DNA sequence, which is used to decode the relevant protein sequence. Phage display is a powerful technique that fulfills this requirement by fusing DNA sequences of interest to a coat protein gene of bacteriophage (Smith, 1985). The expressed protein is delivered to the surface during infection of *E. coli* and selected for binding to an immobilized target. Phage display has been used extensively in antibody engineering, as well as for querying cDNA libraries for binding partners and mapping the specificity residues responsible for protein-protein interactions (Sidhu, 2000). Although phage-displayed libraries have been used successfully to make human antibodies for biomedical applications, some eukaryotic proteins are not properly expressed in a bacterial system. Yeast surface display was developed to address this issue as *Saccharomyces cerevisiae* has the proper machinery to correctly translate, modify, and secrete most mammalian proteins (Boder and Wittrup, 1997). Using this approach, a library of protein fragments is fused to the yeast cell-surface protein Aga2p mating adhesion receptor and assayed for binding.

Another innovative way to preserve the relationship between genotype and phenotype is through mRNA display, a method developed in the Szostak lab. This approach uses *in vitro* translation of synthetic mRNAs, which is modified to result in covalent fusion between the information-carrying mRNA and the function-carrying protein portion (Roberts and Szostak, 1997). Using this approach, the authors of one study constructed a library of $6x10^{12}$ unique mRNA-displayed protein sequences comprised of 80 random amino acids. They enriched the library for variants that bind ATP and identified four novel ATP-binding proteins in this vast sequence space (Keefe and Szostak, 2001). The authors concluded that functional proteins occur at a high enough frequency in sequence space that they can be sampled by chance during evolution.

Although high-throughput screens and selections allow more variants to be sampled, this comes with an associated error rate and a substitution of a measurable activity for the bona fide phenotype. It may be necessary to constrain the experimental conditions by localizing the mutations to a given area, for example at an active site or binding interface (Romero and Arnold, 2009). When available, computational pre-screening of library variants can help to reduce the complexity of the project. This approach has been successfully used to design selective bZIP-binding peptides, improve the stability of a small protein, and facilitate an unnatural protein-

protein interaction (Allen et al., 2010; Grigoryan et al., 2009; Guntas et al., 2010). One limitation of these methods is the requirement of NMR or X-ray crystal structures for accurately modeling the protein of interest.

III. Identification of functional variants

Most high-throughput studies of protein function to date have focused on finding a handful of improved variants for a downstream application rather than comprehensive mapping of protein sequence space. The advent of automated DNA sequencing in the late 1980s allowed straightforward determination of the coding sequences for these variants, but this approach remained too expensive and time-consuming to truly map out sequence space (Smith et al., 1986). Efforts to sequence the large eukaryotic, metazoan, and eventually human genomes sped up the development of next-generation sequencing technologies and commercial pyrosequencing (sequencing-by-synthesis) methods have been available since 2005 from several vendors (Mardis, 2013). These instruments permit the sequencing of millions of variants in a given library experiment, both before and after selection.

The application of next-generation sequencing to protein studies was demonstrated in two recent publications from the University of Washington. In the first study the authors used phage display to assay 600,000 variants of a human WW domain containing 1-3 amino acid substitutions distributed at random throughout the protein (Fowler et al., 2010). The clones were sequenced both before and after six rounds of selection to reveal that 97% of the variants in the input library were deleterious relative to the starting wild-type sequence. The authors of the second study constructed two libraries by introducing all possible point mutations at ~50 amino acid positions of two different influenza hemagglutinin proteins for a total of 2 x 20 x 50 = 2,000 unique variants (Whitehead et al., 2012). The libraries were screened using a combination of yeast surface display and FACS to quantify the effect of each substitution, and the resulting data set used to improve computational energy functions and to produce high-affinity binders.

The use of next-generation sequencing in these and similar studies allows the investigators to confirm the expected diversity of the starting library population and to rapidly quantify large-scale experiments. However, one limitation of the studies described above is that single point mutations do not permit a comprehensive sampling of sequence space. The authors do not explore context dependence of the introduced mutations and make the assumption that mutations always behave in simple additive ways. Finally, these studies focus only on protein binding and therefore may not capture aspects of protein function that involve allosteric changes, interactions with other cellular components, or effects on organismal fitness.

The concept of protein sequence space

In 1970, long before next-generation sequencing was a widespread technique, John Maynard Smith put forward the concept of protein sequence space as a model for the evolution of information-coding sequences in biological organisms. At the time there appeared to be a contradiction between "the idea of natural selection and the idea of the uniqueness of the gene" (Salisbury, 1969). In other words, how can a DNA sequence coding for a functional enzyme arise out of a seemingly infinite number of possible sequences produced by random mutagenesis? Smith suggested that functional sequences are only a small part of all possible sequences and that they can be reached via unit steps through a continuous network (Smith, 1970) (Fig. 1.2). His framework simplified the problem, but fundamental questions remained regarding the size and connectivity of the sequence space, as well as the fraction that has been

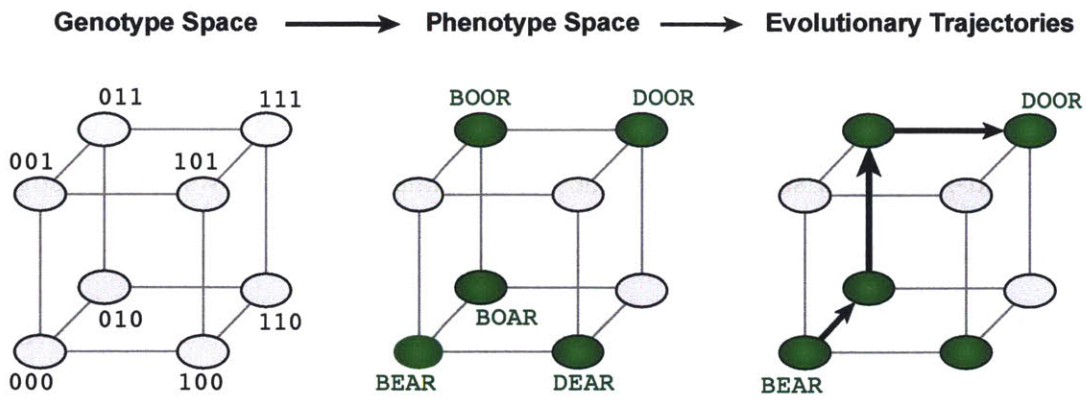


Figure 1.2 Mapping protein sequence-function relationships.

The protein genotype space represents all DNA sequences that encode a protein or a subset of its residues. At this level the genotypes are all equivalent and connected to each other by mutational steps. In the simplest case (left), 0 represents the ancestral state at a position and 1 represents the acquisition of a mutation. At the level of phenotypes, only some of the proteins are functional (center). In this example, the functional nodes correspond to words in the English language and are colored green. Using this sequence space as a map, we can find evolutionary trajectories between the starting node 'BEAR' and the end node 'DOOR' that use only functional intermediates (right). Adapted from [1,112].

explored to date. Although a vast literature on protein science has built up since 1970, fundamental questions remain unanswered.

The contradiction that inspired Smith to write about sequence space arises due to seemingly disparate results from forward and reverse approaches. Borrowing terminology from genetics, "forward" approaches identify functional variants from a pool of randomly mutagenized sequences, whereas "reverse" approaches introduce mutations into a specific protein residue. Forward approaches have estimated the prevalence of functional proteins to be anywhere between 1 in 10^{11} for ATP-binding activity to 1 in 10^{63} for the probability of a 93-residue sequence adopting the λ repressor fold (Keefe and Szostak, 2001; Reidhaar-Olson and Sauer, 1990). However, the work on λ repressor also demonstrated that the wild-type protein is quite degenerate in sequence, with half of the positions examined in the paper accepting nine or more different amino acid substitutions. Indeed, it has been estimated that 50-70% of point mutations are neutral or only slightly deleterious (Romero and Arnold, 2009). These numbers vary depending on the location in the protein, but the overall picture is that a given function can be carried out by more than one particular string of residues.

The finding that functional proteins are robust to mutations yet rarely observed in a pool of completely random sequences suggests that these variants are not uniformly distributed throughout sequence space. Computational studies have shown that sequence space is organized like a funnel around a "prototype sequence" that is optimized for both stability and a large number of neighboring sequences (Bornberg-Bauer and Chan, 1999). Work from the Arnold lab suggests that increased stability allows proteins to tolerate additional destabilizing but beneficial mutations (Bloom et al., 2006). The resulting pattern of alternating stabilizing mutations and functional changes has been documented for clinically relevant cases, including the evolution of influenza and antibiotic resistance (Gong et al., 2013; Wang et al., 2002).

Other factors contributing to the shape of sequence space include epistasis between residues and the layout of the genetic code. In a seminal study of β-lactamase, the authors showed that only 8 out of a possible 120 trajectories can transform the wild-type protein into a version that contains five mutations and greatly increases bacterial antibiotic resistance (Weinreich et al., 2006). Epistasis between these five positions limits the number of accessible paths through sequence space, because the mutations have to appear in a particular order. Results from a subsequent mutagenesis study of β-lactamase suggested that the layout of the genetic code places a further constraint on these mutational trajectories (Firnberg and Ostermeier, 2013). Out of the 19 possible amino acid substitutions, on average only 6 are accessible via a single nucleotide change. Although the authors postulate that the genetic code is laid out to enrich for adaptive mutations, these restrictions limit the number of sequences that can be explored. By reconstructing the precursors of vertebrate glucocorticoid and mineralocorticoid receptors, members of the Thornton lab learned that neutral or "permissive" mutations came before mutations that actually affected function (Ortlund et al., 2007). These findings shed light on epistasis among protein residues, however such small-scale studies cannot tell us exactly how many other permissive mutations exist for this protein.

Given that the estimates of the number of functional protein sequences span many orders of magnitude, there is currently no answer regarding the percentage of sequence space that has been explored in the course of natural evolution. Recent literature on the subject does not provide a consensus view yet. In a publication from 2008, Dryden and colleagues challenge the assumptions that the number of functional proteins is large and only a small proportion has been accessed (Dryden et al., 2008). Using rough estimates from published data the authors conclude that protein sequence space has been fully explored by extant life on Earth, and possibly even before the appearance of prokaryotes. Directly contradicting this, a study from 2010 likens protein sequence to the expanding universe and suggests that evolution has not sampled all functional sequences (Povolotskaya and Kondrashov, 2010). The authors base their claims on rates of divergence away from and towards ancestral reconstructions of proteins. Neither of these studies provides a satisfactory answer, and solid experimental evidence for either of the hypotheses is not available.

In summary, protein sequence space is a useful theoretical framework, but biological data to flesh out the model is lacking. Experimental techniques are just beginning to achieve the throughput necessary to query all 20^N sequences populating a given sequence space. In the case of protein-protein interactions, the problem is magnified because both sides of the interface must be taken into consideration. Rational choice of experimental system, thought-out experimental design, and application of cutting-edge techniques are all essential for any investigation of protein sequence space.

Two-component signal transduction pathways

The experimentally tractable bacterium *Escherichia coli* is a perfect model organism for the types of large-scale questions posed in the previous section, because it has a long history of use in the lab, a wide range of molecular tools, and a short doubling time. Like the majority of bacteria, *E. coli* uses two-component signaling (TCS) pathways to sense the environment and to respond appropriately (Stock et al., 2000). These pathways typically consist of a histidine kinase (HK) that autophosphorylates using ATP, and a response regulator that gets phosphorylated by its partner kinase in response to a signal (Fig. 1.3). Once phosphorylated, the regulator enacts an appropriate cellular response, typically by modulating transcription. In many cases the histidine kinase is bi-functional and dephosphorylates its partner regulator after the stimulus subsides.

The best-studied canonical two-component system EnvZ/OmpR was discovered using genetic complementation analysis over thirty years ago (Sarma and Reeves, 1977). Biochemical studies subsequently revealed that EnvZ is a membrane-bound protein that autophosphorylates on the conserved residue His-243, transfers the phosphoryl group to its partner OmpR, and thereby regulates the expression of outer-membrane proteins encoded by *ompC* and *ompF*. The phospho-His bond is unstable and acid-labile, so it cannot be assayed using conventional phosphoamino acid chromatography at acidic pH (Gao and Stock, 2009). Instead, His-Asp phosphotransfer is monitored by incubation of recombinant proteins with [γ-^{32}P]-ATP. Today, the traditional tools

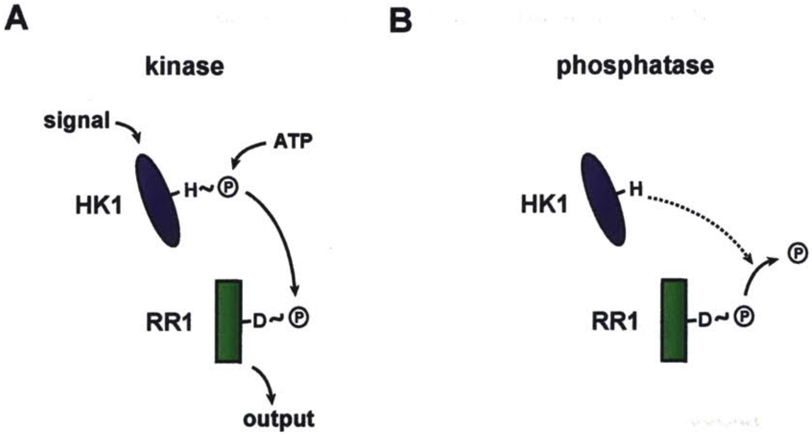


Figure 1.3 Canonical two-component signal transduction system.

(A) When activated by an input signal, canonical histidine kinases use ATP to autophosphorylate on a conserved histidine. The phosphoryl group is transferred to a conserved aspartate on the cognate response regulator, which can then enact a response by changing cellular physiology or gene expression. (B) Most histidine kinases are bi-functional such that, in the absence of an input signal, a histidine kinase will drive dephosphorylation of its cognate response regulator, thereby suppressing an unwanted output.

of bacterial genetics and ^{32}P labeled assays are routinely complemented by GFP fusions and bioinformatics analyses in systematic studies of two-component pathways.

Although the canonical architecture of a TCS pathway involves one kinase and one regulator, in some pathways these components are arranged into more complex architectures. In bacterial chemotaxis the histidine kinase CheA phosphorylates two separate response regulators, CheB and CheY (Stock et al., 2000). Phosphorylated CheY regulates the flagellar motor complex to enact movement, whereas CheB is a methyltransferase that regulates adaptation to a persistent chemical signal (Porter et al., 2011). In other pathways, a phosphorelay is used to pass the phosphate through two histidine and two aspartate residues in the sequence His-Asp-His-Asp. In some systems the histidine and asparate residues may be found in a single protein, resulting in a hybrid kinase. In other systems the intermediate components of the phosphorelay are separate proteins, either a response regulator domain or an Hpt that can transfer phosphoryl groups but cannot autophosphorylate. Phosphorelays are less prevalent than canonical TCS pathway and are only found in 20% of characterized bacterial genomes (Schaller et al., 2011). They are typically involved in the regulation of complex developmental processes, including cell cycle progression in *Caulobacter crescentus* (Biondi et al., 2006), sporulation in *Bacillus subtilis* (Hoch, 1993), and fruiting body formation in *Myxococcus xanthus* myxococcus (Lee et al., 2010).

Two-component pathways are believed to have originated in bacteria and spread to archaea and eukarya by horizontal gene transfer (Koretke et al., 2000). Although their absence from metazoan species makes them an attractive target for antibiotics, no successful inhibitors have been developed to date (Worthington et al., 2013). Two-component pathways are present in fungi, slime molds, and plants, and several have been well characterized in model organisms. These pathways use His-Asp phosphotransfer, but the proteins are often arranged in more complicated three- and four- component architectures. For example, in *Saccharomyces cerevisiae* the multi-step phosphorelay Sln1/Ypd1/Ssk1 is used for osmosensing and the regulation of a MAP kinase cascade (Inouye and Dutta, 2003). The plant *Arabidopsis thaliana* has 8 histidine kinases and 23 response regulators used for sensing growth signals, including cytokinins and ethylene, as well as for regulating circadian rhythms (Schaller et al., 2011). Clearly the two-component signaling pathways in this organism are not arranged in the canonical 1:1 organization, but rather have one-to-many connectivity.

Although they both utilize phosphotransfer for signaling, histidine kinases and the eukaryotic serine, threonine, and tyrosine kinases are not evolutionarily related. The phosphoramidate bond used by two-component pathways is unstable relative to the phosphoesters used by Ser/Thr/Tyr protein kinases, explaining why it is used as a phosphotransfer intermediate rather than a site for protein recognition (Gao and Stock, 2009). Further in contrast to the eukaryotic kinases, histidine kinases are constitutive homodimers activated by external signal rather than by phosphorylation (Alberts, 2008).

Signal perception by histidine kinases

The canonical histidine kinase is an integral membrane protein with a variable number of signaling domains. For example, 25 out of 27 *E. coli* histidine kinases have a transmembrane portion that serves as an anchor (Inouye and Dutta, 2003). Genomic sequencing has revealed great variety in the way that histidine kinases sense their inputs and these ways can be subdivided into three general classes (Krell et al., 2010).

The largest of these classes involves using sensory domains inside or outside the cell. Most cytoplasmic sensory domains adopt a Per-ARNT-Sim (PAS) fold comprised of a central antiparallel five-stranded β sheet core surrounded by several α helices (Möglich et al., 2009b). PAS domains are highly diverged with pairwise sequence identity on average less than 20% and the input signal is unknown in most cases. Some previously identified examples include FixL (oxygen), NtrB (nitrogen), and LovK (blue light) (Krell et al., 2010). A second class of histidine kinases can sidestep a sensory domain by sensing signals directly via the transmembrane region. This mechanism is exemplified by the two-component system DesK/DesR, which senses temperature and membrane fluidity using its transmembrane helices (Cybulski et al., 2004). A final class of signal input is comprised of histidine kinases that use auxiliary proteins to sense the environment. Methyl-accepting chemotaxis proteins (MCPs) sense a variety of extracellular chemical signals and transduce them to the chemotaxis phosphorelay system, which in turn coordinates the movement of bacterial flagella (Krell et al., 2010).

The various sensory domains bind their ligands with affinities that correspond to physiological levels. For example, the histidine kinase NarX has a tight K_d of 35 μM for nitrate and a much higher K_d of 3,500 μM for nitrite (Krell et al., 2010), allowing the cell to fine-tune its nitrogen metabolism. Although the diverse sensory architectures used by histidine kinases have been well catalogued, many of the stimuli for these systems remain unknown. One way to study a system without knowing its physiological signal is using a chimeric protein. In the 1980s the sensory domain of EnvZ was swapped for the extracellular domain from Tar, a membrane bound chemoreceptor for asparate, resulting in a functional signal-transduction pathway (Utsumi et al., 1989). More recently, Möglich et al. replaced the heme-binding PAS sensor domain of FixL with the LOV blue light sensor from YtvA to create a light-sensing histidine kinase (Möglich et al., 2009a), and the synthetic biology community has readily adopted this tool.

Histidine kinase architecture and function

The cytosolic core of a histidine kinase consists of a Dimerization and Histidine phosphotransfer (DHp) domain and a C-terminal Catalytic and ATP binding (CA) domain (Fig. 1.4) (Gao and Stock, 2009). The DHp domain is comprised of two long α helices, which are used for kinase dimerization and for interaction with the partner response regulator. Whereas the CA domain exists as a monomer in solution (Gao and Stock, 2009) and the self-association between sensor domains is weak (Cheung and Hendrickson, 2010), the truncated DHp domain is enough for dimerization. Biochemical experiments from the Inouye lab demonstrated that the DHp domain is necessary and sufficient for the dimerization of EnvZ (Park et al., 1998). Bacterial histidine kinases typically form homodimers, although an instance of heterodimerization has been suggested in the pathogen *Pseudomonas aeuroginosa* (Goodman et al., 2009). Recent work has demonstrated that even the evolutionarily related kinases EnvZ and RstB are insulated to avoid

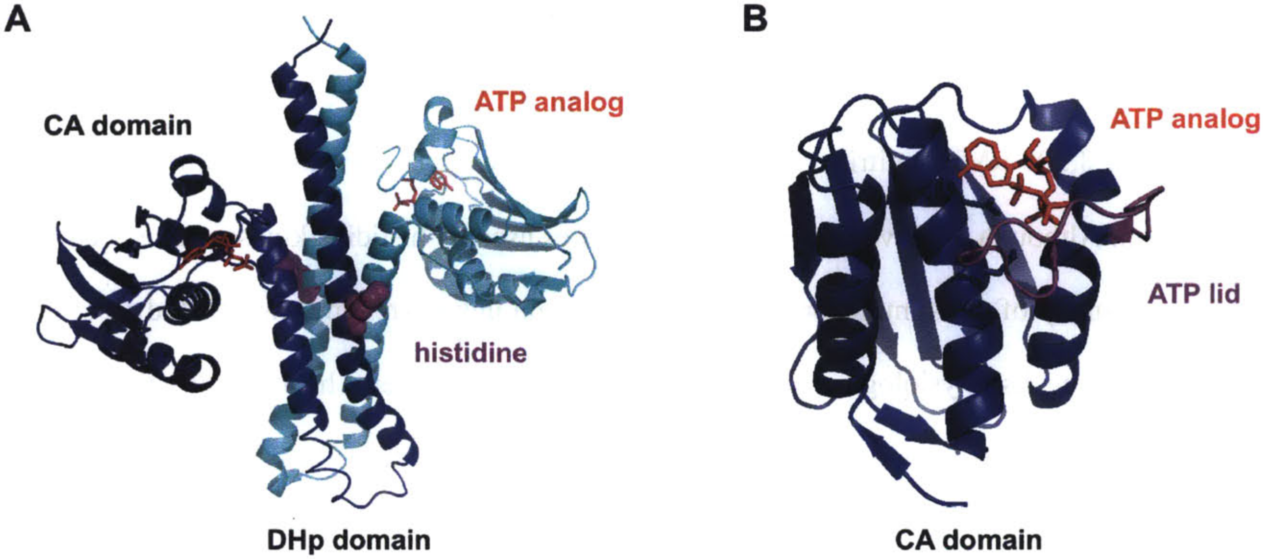


Figure 1.4 Crystal structures of cytoplasmic histidine kinase domains.

(A) Full-length cytoplasmic structure of a histidine kinase dimer (*Thermotoga maritima* HK853; PDB: 2C2A). The two monomers are colored in blue and cyan; the catalytic histidine is shown as purple spheres on either subunit; and the ATP analog AMPPNP captured in the crystal is shown as red sticks. (B) Structure of a single CA domain (*Escherichia coli* PhoQ; PDB: 1ID0). The ATP lid is colored in purple and the ATP analog is shown as red sticks. Adapted from [47].

heterodimerization and showed that the residues responsible for dimerization specificity are localized at the base of the DHp domain hairpin (Ashenberg et al., 2011).

The DHp domain harbors a conserved H-box region, which contains the histidine residue involved in phosphotransfer (Fig. 1.4A). The sequence motif of the H-box is used to classify the protein into one of five families, with most bacterial histidine kinases belonging to the Type I family and harboring the motif `HEhR-P` (H is the invariable histidine residue; 'h' designates a conserved hydrophobic residue) (Kim and Forst, 2001). The DHp domain in the kinase and the first of five alpha helices in the regulator make up the interaction interface in these proteins. Binding experiments have shown that the truncated DHp domain of the kinase PhoQ interacts with the partner PhoP with a K_d that is only slightly higher than the K_d for the entire cytosolic portion of the kinase (Castelli et al., 2003). This suggests that while the other domains may stabilize the interaction, the bulk of interaction specificity between the kinase and regulator resides alongside the catalytic histidine in the center of the DHp domain.

The CA domain, also called the kinase domain due to its roles in binding ATP and autophosphorylation, is typically 350 amino acids in length and less variable in sequence than the DHp domain (Fig. 1.4B). It is part of a superfamily that includes topoisomerases, DNA repair proteins, and molecular chaperones (Dutta and Inouye, 2000). It binds ATP with a conserved K_d of 100-300 μM (Krell et al., 2010) and the ATP-binding region is surrounded by conserved sequence motifs called the N, G1 and G2 boxes, as well as a more variable F box (Kim and Forst, 2001). The ATP lid is located between the F and G2 boxes, and adopts a different conformation in crystal structures depending on whether nucleotide is bound to the CA domain (Gao and Stock, 2009). Mutations to these conserved elements have predictably deleterious effects on nucleotide binding and kinase activity (Zhu and Inouye, 2002). Surprisingly, these mutations also affect phosphatase activity, suggesting that the bound nucleotide must somehow participate in the dephosphorylation reaction.

Signal transduction by canonical histidine kinases

Crystallographic studies have revealed that conformational changes triggered by stimulus perception are relayed to the rest of the protein by piston-like or rotation movements, or a combination of both (Casino et al., 2010). One recent study dissected the mechanism used by the histidine kinase LuxQ, which is involved in quorum sensing. At low cell densities the transmembrane portions of LuxQ form a symmetrical four-helix bundle that is poised for autophosphorylation *in trans* (Neiditch et al., 2006). At higher cellular densities binding of the extracellular molecule AI-2 to the sensory domain breaks this symmetry and prevents autophosphorylation. Many transmembrane histidine kinases include a cytosolic coiled-coil HAMP domain, which aids in propagating structural changes to the site of autophosphorylation (Ferris et al., 2012). The structural changes relayed from the sensory domain result in changes to the interface between the CA and DHp domains. The rigid body movement of the CA domain closer to the DHp domain results in the proper positioning of the ATP molecule for the autophosphorylation and phosphotransfer reactions (Casino et al., 2010).

In principle, the CA domain of one subunit in a histidine kinase homodimer could phosphorylate the catalytic histidine residue in its own protein, or in the partner protein. Early work with CheA, EnvZ, and NRII showed that these well-studied histidine kinases autophosphorylate *in trans* and this became the prevailing view in the field (Krell et al., 2010). Recent crystallographic data appeared inconsistent with this model and indeed the kinases HK853 and PhoR were shown to autophosphorylate *in cis* (Casino et al., 2009). Factors that determine whether the reaction will proceed *in trans* or *in cis* include the handedness and the length of the loop at the base of the DHp domain helices, as well as the length of the loop connecting the DHp and CA domains (Ashenberg et al., 2013; Casino et al., 2010). A recent study by the Marina group showed that changing the loop between helices α1 and α2 in EnvZ to match the loop found in HK853 is sufficient to switch the autophosphorylation reaction from *trans* to *cis* (Casino et al., 2014). This preference appears to be functionally important, because it is conserved among orthologs (Ashenberg et al., 2013).

Most histidine kinases are bi-functional as they can drive phosphorylation of their cognate response regulators and act as phosphatases that stimulate dephosphorylation of the cognate partner (Huynh et al., 2010; Igo et al., 1989; Willett and Kirby, 2012). The isolated DHp domain of EnvZ can completely dephosphorylate OmpR~P in 20 minutes, whereas the complete cytosolic protein can dephosphorylate it in 10 minutes (Zhu et al., 2000). The DHp domain is responsible for binding the cognate regulator in both phosphotransfer and phosphatase reactions, with the CA domain making some additional contacts with the regulator. Interactions between the two cytosolic domains determine whether the protein autophosphorylates, or engages in phosphotransfer and phosphatase reactions. Indeed, a deletion that partially overlaps the G2 box in the histidine kinase VanS abolished phosphatase activity without affecting autophosphorylation (Depardieu et al., 2003). Although it was initially believed that the phosphatase reaction is simply the reverse of phosphotransfer, this may not be the case. All mutations to the catalytic histidine in EnvZ abolish kinase activity, but a subset of these keep phosphatase activity perfectly intact (Hsing and Silhavy, 1997), indicating that the local chemical environment rather than the specific residue is important.

Response regulator architecture and function

Response regulators are typically comprised of a receiver (REC) domain and an effector domain that regulates a cellular process in response to phosphorylation (Fig. 1.5). Given the large number of extant two-component systems, there are inevitably exceptions to this canonical architecture. Approximately 17% of bacterial and 50% of archaeal response regulators have a single receiver domain, which modulates the activity of other two-component signaling pathways via protein-protein interactions (Galperin, 2010). Typical receiver domains have a $(\beta\alpha)_5$ topology and an active site with five highly conserved residues, which are involved in phosphotransfer and the coordination of a magnesium ion necessary for the reaction (Bourret, 2010). Differences in active site residues determine the half-life of phosphorylated receiver domains. These range from 23 seconds for the chemotaxis regulator CheY to 180 minutes for Spo0F, which regulates sporulation (Depardieu et al., 2003).

Response regulators can be phosphorylated by histidine kinases or by low-weight promiscuous phosphodonors such as phosphoramidate and acetyl phosphate (Bourret, 2010). The latter is a high-energy intermediate produced by the AckA-Pta metabolic pathway and present in cells at sufficient levels to phosphorylate response regulators *in vivo* (Klein et al., 2007). Although acetyl phosphate has been proposed to be a global signaling molecule in *Escherichia coli* (McCleary et al., 1993), there remains controversy regarding its physiological role in modulating two-component signaling. Beryllofluoride is another small molecule that binds to response regulators and mimics phosphorylation *in vitro*. Its activity is not relevant *in vivo*, but it has greatly facilitated crystallographic studies of response regulators in the activated state (Yan et al., 1999).

Response regulators act as a "phosphorylation-activated switch", because they can adopt one of two states. In response to phosphorylation response regulators commonly form dimers or higher-order oligomeric species (Gao and Stock, 2009). In members of the OmpR/PhoB family phosphorylation triggers triggers an allosteric change to the α4-β5-α5 face at the C-terminal end of the receiver domain and subsequent homodimerization (Fig. 1.5A). Recent work from the Stock lab has shown that response regulators exhibit specificity in dimerization and heterodimers are avoided (Gao et al., 2008). This specificity is achieved via conserved residues involved in hydrophobic interactions and salt bridges across the regulator-regulator interface.

Response regulators have evolved a variety of effector domains for responding to the various signals sensed by histidine kinases. In a subset of response regulators the effector domain is used for enzymatic reactions, such as the GGDEF diguanylate cyclase domain that is involved in c-di-GMP synthesis (Römling et al., 2005). The majority of effector domains are DNA-binding proteins used to regulate gene expression and can be further classified according to their secondary structural elements (Fig. 1.5B). The largest of these subfamilies is represented by the

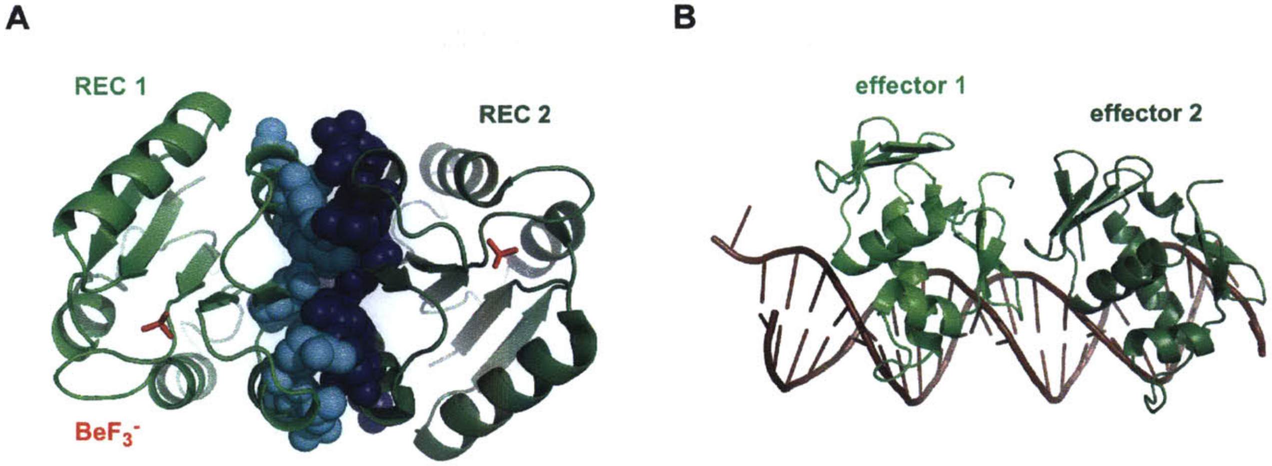


Figure 1.5 Crystal structures of response regulator domains.

(A) Structure of the dimer formed by two receiver (REC) domains (*Escherichia coli* PhoB; PDB: 1ZES). The two domains are colored with different shades of green and the phosphoryl analogue beryllofluoride is shown as red sticks. The highly conserved residues involved in hydrophobic contacts or salt bridges across the dimer interface are shown as spheres. (B) Structure of effector domains bound to DNA (*Escherichia coli* PhoB; PDB: 1GXP). Panel (A) adapted from (Gao et al., 2008).

OmpR/PhoB winged-helix domain, which is found in some of the best-characterized two-component signaling pathways. These proteins contain a DNA recognition helix that binds specifically to the major groove and two loops, or "wings", that bind to adjacent regions in the minor groove (Martínez-Hackert and Stock, 1997).

Although all members of the OmpR/PhoB response regulator family bind DNA in response to phosphorylation, they use different mechanisms to regulate gene expression. When OmpR binds to DNA it enacts transcription by interacting with the α subunit of RNA polymerase, whereas PhoB interacts with σ^{70} to regulate transcription (Stock et al., 2000). These mechanistic differences may be an evolutionary adaptation to prevent response regulators from inappropriately regulating promoters with similar binding sites. OmpR can bind non-specifically to KdpE binding sites, but it is not positioned correctly for contacting polymerase and cannot activate transcription (Ohashi et al., 2005). Interestingly, just a few mutations in OmpR can rewire it to activate the KdpE regulon and further systematic studies are needed to understand regulator-DNA specificity.

Specificity in two-component signal transduction

Bacterial genomes have an average of 52 two-component systems, ranging from just a few pathways in obligate intracellular parasites and endosymbionts to over a hundred pathways in species found in rapidly changing environments (Capra and Laub, 2012). The presence of so many structurally similar proteins inside a crowded bacterial cell could result in either rampant cross talk between parallel signaling pathways, or imply the existence of mechanisms for cross talk prevention. This ability to avoid deleterious cross talk is critical to the faithful transmission of signals inside bacterial cells. There are three key mechanisms for ensuring the specificity of two-component pathways at the level of phosphotransfer: molecular recognition, phosphatase activity, and substrate competition.

The predominant mechanism for enforcing specificity is molecular recognition, the intrinsic ability of an autophosphorylated histidine kinase to recognize its cognate partner to the exclusion of all possible non-cognate partners. Early kinetic studies with the Enterococcus kinase VanS demonstrated that it preferentially phosphorylates its cognate regulator VanR relative to the E. coli regulator PhoB. The k_{cat}/K_M ratio, or specificity constant, for transfer to VanR is 10^4-fold higher than to PhoB (Fisher et al., 1996). More recently, systematic analyses of phosphotransfer from a given kinase to all possible regulators encoded in a genome have demonstrated that histidine kinases typically harbor a global and strong kinetic preference for their cognate response regulator in vitro (Skerker et al., 2005). This ability to discriminate cognate from non-cognate partners in the absence of other cellular components, such as scaffolds, indicates that specificity is encoded primarily at the molecular level. The recognition of the cognate partner is driven by a small set of residues located primarily in one alpha helix of each molecule (Casino et al., 2009; Skerker et al., 2008) and is discussed in the next section.

The specificity of two-component pathways is further reinforced *in vivo* through the phosphatase reaction, which is used to modulate the level of pathway output and to inhibit the pathway after an activating signal has subsided (Huynh and Stewart, 2011). Importantly, the phosphatase activity of a histidine kinase also serves to minimize unwanted cross talk by dephosphorylating the cognate response regulator when it is inappropriately phosphorylated by another kinase or a small molecule phosphodonor (Fig. 1.6A). Consequently, mutations that eliminate the phosphatase activity of a histidine kinase, including deletion of the histidine kinase gene, can lead to the inappropriate activation of the kinase's cognate response regulator under non-inducing conditions (Fig. 1.6A) (Siryaporn and Goulian, 2008).

Specificity is further enhanced by the relative cellular concentrations of histidine kinases and their cognate response regulators, and by competition between regulators for phosphorylated kinases (Fig. 1.6B). For most two-component pathways, abundance of the response regulator likely exceeds that of the cognate kinase. The well-characterized *E. coli* kinase EnvZ and its partner OmpR are found at a ratio of about ~1:35, and other pathways are reported to have similar ratios (Cai and Inouye, 2002; Miyashiro and Goulian, 2008). The higher abundance of the response regulators creates a scenario in which a given regulator effectively outcompetes non-cognate regulators for binding to a cognate kinase, further preventing unwanted phosphotransfer events. Consequently, deleting a given response regulator can lead to inappropriate cross talk from its cognate kinase to other response regulators (Fig. 1.6B) (Groban et al., 2009; Siryaporn and Goulian, 2008).

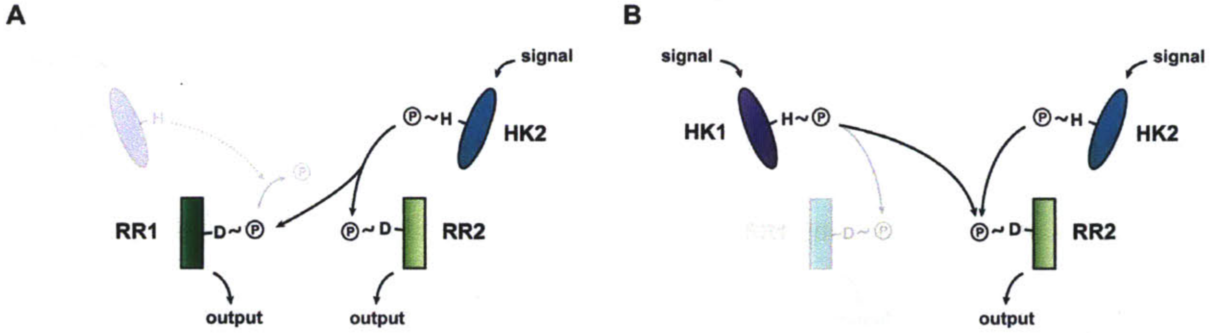


Figure 1.6 Multiple mechanisms ensure the specificity of two-component signaling.

(A) In addition to molecular recognition, phosphotransfer specificity is enforced by the phosphatase activity of histidine kinases. Unwanted cross talk from a non-cognate kinase (HK2) to a response regulator (RR1) is normally eliminated by the phosphatase activity of the cognate kinase (HK1). Deleting a kinase (depicted in faded color) can, consequently, lead to spurious activation of a pathway. (B) Competition between response regulators can further enhance the specificity of phosphotransfer. When a kinase (HK1) is autophoshorylated, its cognate response regulator (RR1) will better recognize, and hence outcompete, other response regulators for phosphotransfer. Deleting a regulator (RR1 depicted in faded color) can therefore allow its cognate kinase to phosphorylate a non-cognate substrate (RR2).

In addition to these three mechanisms, specificity could also arise through temporal or spatial restriction of pathways. For example, in *Rhodobacter capsulatus* the subcellular localization of chemotaxis proteins to either polar or mid-cell clusters helps prevent cross talk (Scott et al., 2010). Although the expression of different pathways at different times could help to prevent unwanted cross talk, there are no clear examples of this mechanism. Collectively, three primary mechanisms – molecular recognition, phosphatase activity, and substrate competition – ensure that two-component signaling pathways are insulated from one another at the level of phosphotransfer. Branched pathways with physiologically relevant one-to-many or many-to-one connectivity have been documented, but in most cases the mechanisms outlined above enforce specific, one-to-one relationships between kinases and their cognate regulators (Laub and Goulian, 2007).

Identification and characterization of specificity residues

The ability of histidine kinases and response regulators to preferentially recognize their cognate partners relies on a limited set of amino acids in each protein. These specificity-determining residues were identified initially through computational analyses of amino acid covariation in large sets of cognate, co-operonic two-component proteins (Fig. 1.7A) (Capra et al., 2010; Skerker et al., 2008; Weigt et al., 2009). This statistical approach identifies pairs of amino acids that covary, or change in a concerted manner over the course of evolution, to maintain the interaction between the partner proteins (Ashenberg and Laub, 2013). In some cases, these pairs are located in the same protein, where they make intramolecular contacts necessary for structural integrity or for promoting certain protein conformations. In other cases, the amino acids are located in opposite proteins, and likely have coevolved to preserve the interaction of a cognate kinase and regulator pair (Fig. 1.7A).

These intermolecular, coevolving residues were subsequently demonstrated to be critical specificity determinants. Mutating these residues in a model histidine kinase, *E. coli* EnvZ, to match those found in other *E. coli* kinases was sufficient to endow EnvZ with the ability to specifically phosphorylate other *E. coli* response regulators rather than its usual cognate partner, OmpR (Skerker et al., 2008). Similarly, response regulators have been rationally rewired to receive phosphoryl groups from non-cognate kinases (Bell et al., 2010; Capra et al., 2010), solidifying the notion that these coevolving amino acids are indeed specificity-determining residues. The phosphotransfer specificity of EnvZ can be rewired to match the specificity of the *E. coli* kinase RstB through just three substitutions. Subsequent analysis of the three single and three double mutant intermediates separating EnvZ and RstB indicated that different intermediates harbor substantially different specificities. Of the three double mutants, one does

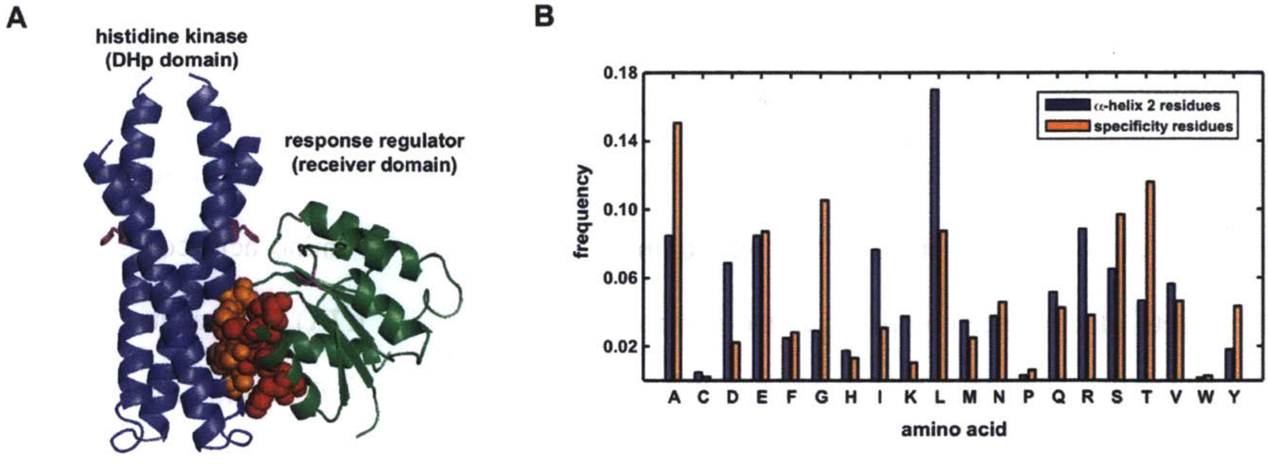


Figure 1.7 Amino acid residues important for phosphotransfer specificity.

(A) Residues that strongly coevolve in cognate pairs of histidine kinases and response regulators are shown on a crystal structure of the *T. maritima* HK853 in complex with RR468 (PDB: 3DGE). Only the DHp domain of HK853 is shown. Specificity residues on the kinase and regulator are shown with space-filling spheres in orange and red, respectively. The conserved histidine and aspartate that participate in phoshotransfer are shown as sticks. (B) Histogram showing amino acid frequencies for the six key specificity residues from a-helix 1 of the kinase and for all residues from a-helix 2, which does not play a prominent role in specificity. Frequencies were computed using a sequence alignment of > 6500 histidine kinases.

not phosphorylate either OmpR or RstA (the cognate partner of RstB), one still phosphorylates only OmpR, albeit weakly, and the third robustly phosphorylates both regulators (Capra et al., 2010).

The first solved crystal structure of a histidine kinase in complex with its cognate regulator, the *Thermatoga maritima* pair HK853-RR468, demonstrated that the phosphotransfer specificity residues lie mainly at the interface formed by these proteins and reside on the surface of an alpha helix in each protein (Casino et al., 2009). The docking of these helical surfaces and the inter-digitation of specificity residues promotes an orientation of the kinase and regulator in which the conserved histidine and aspartate side-chains are ideally positioned for phosphotransfer or dephosphorylation (Fig. 1.7A). Two other structures of kinases and response regulators in complex confirm the central position of the specificity residues at the interaction interface (Yamada et al., 2009; Zapf et al., 2000). Globally, the distribution of amino acid frequencies for the specificity residues in >6500 histidine kinases indicates a preponderance of small, hydrophobic and polar residues and a corresponding paucity of bulky and charged residues (Fig. 1.7B). For instance, alanine, glycine, serine, and threonine are each overrepresented in the set of specificity residues relative to the distribution of amino acids found in α-helix 2 of the DHp domain, which does not significantly impact partner specificity. Conversely, the charged residues aspartate, lysine, and arginine are each underrepresented in the set of kinase specificity residues. These patterns support the notion that kinase-regulator interfaces are likely mediated primarily by hydrophobic and van der Waals interactions that promote steric, rather than charge, complementarity.

Evolution of two-component signaling specificity

Why do specificity residues in two-component signaling proteins covary in the first place? The answer appears, in many cases, to be gene duplication events and the birth of new pathways (Capra and Laub, 2012; Capra et al., 2012). Phylogenetic analyses indicate that, for most species, the majority of new two-component pathways emerge through gene duplication (Alm et al., 2006; Whitworth and Cock, 2009). Immediately after duplication of a kinase-regulator pair, the two signaling pathways are identical, such that each kinase can interact with each regulator. After the pathways diverge with respect to signal inputs and downstream outputs, there is a need to avoid cross talk via changes in the specificity residues of one or both of the recently duplicated kinases (Fig. 1.8A). Such mutations must then be compensated through mutations in the cognate response regulators. This intermolecular coevolution enables the insulation of the two new pathways while maintaining phosphotransfer within each system. For instance, while

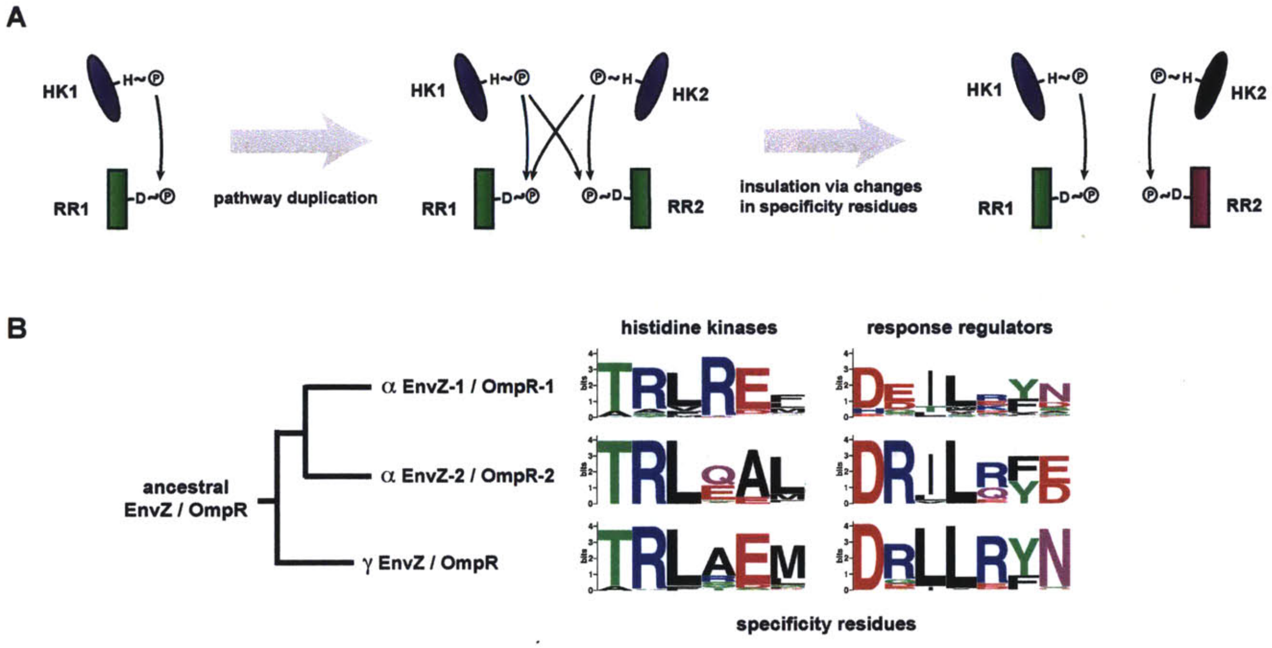


Figure 1.8 The process of pathway insulation following gene duplication has resulted in two-component pathways without significant cross talk.

(A) Duplication of a two-component pathway initially produces two identical pathways that engage in cross talk. To insulate the new pathways from one another, the specificity residues in one or both histidine kinases must change, along with compensatory changes in their cognate response regulators, or vice versa. (B) The EnvZ-OmpR system, present in single copy in γ-proteobacteria, was duplicated in an ancestor of the α-proteobacteria. The duplicates subsequently became insulated at the level of phosphotransfer specificity. Sequence logos of the specificity residues for each group of α-EnvZ and α-OmpR orthologs indicate the changes that likely led to insulation; logos for γ-EnvZ and γ-OmpR orthologs are included for comparison.

there is a single copy of EnvZ-OmpR in γ-proteobacteria, there are two copies in most α-proteobacteria (Fig. 1.8B). These two systems are insulated from one another at the level of phosphotransfer and have different specificity residues. Importantly, the specificity residues of each system are well conserved, indicating that once insulated following duplication, there is likely strong purifying selective pressure on these residues.

The insulation of recently duplicated pathways may also require changes in other existing two-component pathways (Capra et al., 2012). For instance, in α-proteobacteria a duplication of the

NtrB-NtrC system produced the NtrY-NtrX system, and the specificity residues of NtrY-NtrX subsequently diverged from those of NtrB-NtrC to yield two insulated pathways. However, the accumulated changes in NtrY-NtrX likely led to cross talk with the PhoR-PhoB system in α-proteobacteria, driving adaptive substitutions in the specificity residues of that system to insulate it from NtrY-NtrX. Reverting these putative adaptive substitutions in the PhoR of an extant α-proteobacterium, *Caulobacter crescentus*, leads to cross talk with NtrX and a significant fitness disadvantage relative to the wild-type strain. Thus, the avoidance of cross talk between pathways appears to be a major selective pressure driving the diversification of specificity residues following gene duplication events. This process of pathway insulation following duplication has resulted in extant organisms harboring large sets of two-component pathways that can transduce signals without significant cross talk (Laub and Goulian, 2007; Skerker et al., 2005; Yamamoto et al., 2005).

The sequence space of two-component signaling pathways

The key properties of two-component signaling specificity, namely their use of molecular recognition via a small subset of residues, makes them a great system for tackling questions regarding protein sequence space. From previous work done in this area, it is clear that histidine kinases and response regulators can accommodate diverse sets of residues at their interface and the evolution of these proteins is constrained in sequence space. These findings have prompted several interesting research questions: What is the mutational tolerance of the specificity positions and how does this compare with single point mutations of these residues? How are neutral substitutions accommodated at the interface? What is the complete sequence space underlying a productive HK-RR interaction?

During the course of my PhD research I addressed these questions using two suitable experimental systems. In Chapter 2 I describe the results from a successful international collaboration that elucidated how new residues are physically accommodated at the kinase-regulator interaction. This work was done with the *Thermotoga maritima* two-component signaling pathway HK853-RR468 that has previously proved amenable to crystallization. Next I developed and implemented high-throughput assays to systematically map the sequence space underlying the PhoQ-PhoP pathway from *Escherichia coli*. Unlike many two-component systems this pair responds to a know set of input signals and regulates a well-defined set of genes. In Chapter 3 I describe how I identified functional PhoQ variants from a pool of sequences completely randomized at four interface positions and characterized the extent of degeneracy and epistasis for this protein-protein interface. The sequence space of this protein suggested a model for why nature has sampled only a limited number of functional sequences. Finally, in Chapter 4 I propose a list of research directions for future studies.

References

Alberts, B. (2008). Molecular biology of the cell. (New York, Garland Science,).

Allen, B.D., Nisthal, A., and Mayo, S.L. (2010). Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc Natl Acad Sci U S A *107*, 19838-19843.

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.

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., and Laub, M.T. (2013). Using analyses of amino Acid coevolution to understand protein structure and function. Methods Enzymol *523*, 191-212.

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.

Baker, M. (2011). Microarrays, megasynthesis. Nat Methods *8*, 457-460.

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. (2006). Regulation of the bacterial cell cycle by an integrated genetic circuit. Nature *444*, 899-904.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc Natl Acad Sci U S A *103*, 5869-5874.

Boder, E.T., and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. Nat Biotechnol *15*, 553-557.

Bornberg-Bauer, E., and Chan, H.S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci U S A *96*, 10689-10694.

Bourret, R.B. (2010). Receiver domain structure and function in response regulator proteins. Curr Opin Microbiol *13*, 142-149.

Cai, S.J., and Inouye, M. (2002). EnvZ-OmpR interaction and osmoregulation in Escherichia coli. J Biol Chem *277*, 24155-24161.

Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

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., Miguel-Romero, L., and Marina, A. (2014). Visualizing autophosphorylation in histidine kinases. Nat Commun *5*, 3258.

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.

Castelli, M.E., Cauerhff, A., Amongero, M., Soncini, F.C., and Vescovi, E.G. (2003). The H box-harboring domain is key to the function of the Salmonella enterica PhoQ Mg2+-sensor in the recognition of its partner PhoP. J Biol Chem *278*, 23579-23585.

Cheung, J., and Hendrickson, W.A. (2010). Sensor domains of two-component regulatory systems. Curr Opin Microbiol *13*, 116-123.

Church, G.M., Gao, Y., and Kosuri, S. (2012). Next-generation digital information storage in DNA. Science *337*, 1628.

Cunningham, B.C., and Wells, J.A. (1989). High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. Science *244*, 1081-1085.

Cybulski, L.E., del Solar, G., Craig, P.O., Espinosa, M., and de Mendoza, D. (2004). Bacillus subtilis DesR functions as a phosphorylation-activated switch to control membrane lipid fluidity. J Biol Chem *279*, 39340-39347.

Depardieu, F., Courvalin, P., and Msadek, T. (2003). A six amino acid deletion, partially overlapping the VanSB G2 ATP-binding motif, leads to constitutive glycopeptide resistance in VanB-type Enterococcus faecium. Mol Microbiol *50*, 1069-1083.

Dryden, D.T., Thomson, A.R., and White, J.H. (2008). How much of protein sequence space has been explored by life on Earth? J R Soc Interface *5*, 953-956.

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

Ferris, H.U., Dunin-Horkawicz, S., Hornig, N., Hulko, M., Martin, J., Schultz, J.E., Zeth, K., Lupas, A.N., and Coles, M. (2012). Mechanism of regulation of receptor histidine kinases. Structure *20*, 56-66.

Firnberg, E., and Ostermeier, M. (2013). The genetic code constrains yet facilitates Darwinian evolution. Nucleic Acids Res *41*, 7420-7428.

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

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat Methods *7*, 741-746.

Galperin, M.Y. (2010). Diversity of structure and function of response regulator output domains. Curr Opin Microbiol *13*, 150-159.

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

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

Gong, L.I., Suchard, M.A., and Bloom, J.D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. Elife *2*, e00631.

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.

Gregoret, L.M., and Sauer, R.T. (1993). Additivity of mutant effects assessed by binomial mutagenesis. Proc Natl Acad Sci U S A *90*, 4246-4250.

Grigoryan, G., Reinke, A.W., and Keating, A.E. (2009). Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature *458*, 859-864.

Groban, E.S., Clarke, E.J., Salis, H.M., Miller, S.M., and Voigt, C.A. (2009). Kinetic buffering of cross talk between bacterial two-component sensors. J Mol Biol *390*, 380-393.

Guntas, G., Purbeck, C., and Kuhlman, B. (2010). Engineering a protein-protein interface using a computationally designed library. Proc Natl Acad Sci U S A *107*, 19296-19301.

Harding, S.E., and Rowe, A.J. (2010). Insight into protein-protein interactions from analytical ultracentrifugation. Biochem Soc Trans *38*, 901-907.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N. (2011). Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A *108*, 7896-7901.

Hoch, J.A. (1993). Regulation of the phosphorelay and the initiation of sporulation in Bacillus subtilis. Annu Rev Microbiol *47*, 441-465.

Hodges, R.S., and Merrifield, R.B. (1974). Synthetic study of the effect of tyrosine at position 120 of ribonuclease. Int J Pept Protein Res *6*, 397-405.

Hsing, W., and Silhavy, T.J. (1997). Function of conserved histidine-243 in phosphatase activity of EnvZ, the sensor for porin osmoregulation in Escherichia coli. J Bacteriol *179*, 3729-3735.

Hu, C.D., Chinenov, Y., and Kerppola, T.K. (2002). Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. Mol Cell *9*, 789-798.

Hughes, M.D., Nagel, D.A., Santos, A.F., Sutherland, A.J., and Hine, A.V. (2003). Removing the redundancy from randomised gene libraries. J Mol Biol *331*, 973-979.

Huynh, T.N., Noriega, C.E., and Stewart, V. (2010). Conserved mechanism for sensor phosphatase control of two-component signaling revealed in the nitrate sensor NarX. Proc Natl Acad Sci U S A *107*, 21140-21145.

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

Igo, M.M., Ninfa, A.J., Stock, J.B., and Silhavy, T.J. (1989). Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. Genes Dev *3*, 1725-1734.

Inouye, M., and Dutta, R. (2003). Histidine kinases in signal transduction (San Diego, Calif. ; London: Academic Press).

Jäckel, C., Kast, P., and Hilvert, D. (2008). Protein design by directed evolution. Annu Rev Biophys *37*, 153-173.

Kastritis, P.L., and Bonvin, A.M. (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface *10*, 20120835.

Keefe, A.D., and Szostak, J.W. (2001). Functional proteins from a random-sequence library. Nature *410*, 715-718.

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

Klein, A.H., Shulla, A., Reimann, S.A., Keating, D.H., and Wolfe, A.J. (2007). The intracellular concentration of acetyl phosphate in Escherichia coli is sufficient for direct phosphorylation of two-component response regulators. J Bacteriol *189*, 5574-5581.

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.

Kosuri, S., Eroshenko, N., Leproust, E.M., Super, M., Way, J., Li, J.B., and Church, G.M. (2010). Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. Nat Biotechnol *28*, 1295-1299.

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. Annu Rev Microbiol *64*, 539-559.

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

Lee, B., Schramm, A., Jagadeesan, S., and Higgs, P.I. (2010). Two-component systems and regulation of developmental progression in Myxococcus xanthus. Methods Enzymol *471*, 253-278.

Livnah, O., Bayer, E.A., Wilchek, M., and Sussman, J.L. (1993). Three-dimensional structures of avidin and the avidin-biotin complex. Proc Natl Acad Sci U S A *90*, 5076-5080.

Mardis, E.R. (2013). Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto Calif) *6*, 287-303.

Martínez-Hackert, E., and Stock, A.M. (1997). Structural relationships in the OmpR family of winged-helix transcription factors. J Mol Biol *269*, 301-312.

McCleary, W.R., Stock, J.B., and Ninfa, A.J. (1993). Is acetyl phosphate a global signal in Escherichia coli? J Bacteriol *175*, 2793-2798.

Miyashiro, T., and Goulian, M. (2008). High stimulus unmasks positive feedback in an autoregulated bacterial signaling circuit. Proc Natl Acad Sci U S A *105*, 17457-17462.

Morrison, K.L., and Weiss, G.A. (2001). Combinatorial alanine-scanning. Curr Opin Chem Biol *5*, 302-307.

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

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

Neiditch, M.B., Federle, M.J., Pompeani, A.J., Kelly, R.C., Swem, D.L., Jeffrey, P.D., Bassler, B.L., and Hughson, F.M. (2006). Ligand-induced asymmetry in histidine sensor kinase complex regulates quorum sensing. Cell *126*, 1095-1108.

Neylon, C. (2004). Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. Nucleic Acids Res *32*, 1448-1459.

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.

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.

Park, H., Saha, S.K., and Inouye, M. (1998). Two-domain reconstitution of a functional protein histidine kinase. Proc Natl Acad Sci U S A *95*, 6728-6732.

Phizicky, E.M., and Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. Microbiol Rev *59*, 94-123.

Porter, S.L., Wadhams, G.H., and Armitage, J.P. (2011). Signal processing in complex chemotaxis pathways. Nat Rev Microbiol *9*, 153-165.

Povolotskaya, I.S., and Kondrashov, F.A. (2010). Sequence space and the ongoing expansion of the protein universe. Nature *465*, 922-926.

Reidhaar-Olson, J.F., and Sauer, R.T. (1990). Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. Proteins *7*, 306-316.

Roberts, R.W., and Szostak, J.W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci U S A *94*, 12297-12302.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol *10*, 866-876.

Römling, U., Gomelsky, M., and Galperin, M.Y. (2005). C-di-GMP: the dawning of a novel bacterial signalling system. Mol Microbiol *57*, 629-639.

Salisbury, F.B. (1969). Natural selection and the complexity of the gene. Nature *224*, 342-343.

Sarma, V., and Reeves, P. (1977). Genetic locus (ompB) affecting a major outer-membrane protein in Escherichia coli K-12. J Bacteriol *132*, 23-27.

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.

Schreiber, G., and Fersht, A.R. (1995). Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. J Mol Biol *248*, 478-486.

Scott, K.A., Porter, S.L., Bagg, E.A., Hamer, R., Hill, J.L., Wilkinson, D.A., and Armitage, J.P. (2010). Specificity of localization and phosphotransfer in the CheA proteins of Rhodobacter sphaeroides. Mol Microbiol *76*, 318-330.

Sidhu, S.S. (2000). Phage display in pharmaceutical biotechnology. Curr Opin Biotechnol *11*, 610-616.

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.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315-1317.

Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563-564.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature *321*, 674-679.

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

Trudeau, D.L., Smith, M.A., and Arnold, F.H. (2013). Innovation by homologous recombination. Curr Opin Chem Biol *17*, 902-909.

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.

Virnekäs, B., Ge, L., Plückthun, A., Schneider, K.C., Wellnhofer, G., and Moroney, S.E. (1994). Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res *22*, 5600-5607.

Wang, X., Minasov, G., and Shoichet, B.K. (2002). Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. J Mol Biol *320*, 85-95.

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.

Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., *et al.* (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol *30*, 543-548.

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

Willett, J.W., and Kirby, J.R. (2012). Genetic and biochemical dissection of a HisKA domain identifies residues required exclusively for kinase and phosphatase activities. PLoS Genet *8*, e1003084.

Worthington, R.J., Blackledge, M.S., and Melander, C. (2013). Small-molecule inhibition of bacterial two-component systems to combat antibiotic resistance and virulence. Future Med Chem *5*, 1265-1284.

Yamada, S., Sugimoto, H., Kobayashi, M., Ohno, A., Nakamura, H., and Shiro, Y. (2009). Structure of PAS-linked histidine kinase and the response regulator complex. Structure *17*, 1333-1344.

Yan, D., Cho, H.S., Hastings, C.A., Igo, M.M., Lee, S.Y., Pelton, J.G., Stewart, V., Wemmer, D.E., and Kustu, S. (1999). Beryllofluoride mimics phosphorylation of NtrC and other bacterial response regulators. Proc Natl Acad Sci U S A *96*, 14789-14794.

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.

Zhu, Y., and Inouye, M. (2002). The role of the G2 box, a conserved motif in the histidine kinase superfamily, in modulating the function of EnvZ. Mol Microbiol *45*, 653-663.

Zhu, Y., Qin, L., Yoshida, T., and Inouye, M. (2000). Phosphatase activity of histidine kinase EnvZ without kinase catalytic domain. Proceedings of the National Academy of Sciences *97*, 7808.

Chapter 2

Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling

This work was published as Podgornaia, A.I.*, Casino, P.*, Marina, A., Laub, M.T. *Structure 21*, 1636-1647, September 3, 2013.

AIP, PC, AM, and MTL conceived the experiments and wrote the manuscript. AIP performed mutagenesis and phosphotransfer experiments; PC performed crystallography. AIP and PC contributed equally to the work.

Summary

Two-component signal transduction systems typically involve a sensor histidine kinase that specifically phosphorylates a single, cognate response regulator. This protein-protein interaction relies on molecular recognition via a small set of residues in each protein. To better understand how these residues determine the specificity of kinase-substrate interactions, we rationally rewired the interaction interface of a *Thermotoga maritima* two-component system, HK853-RR468, to match that found in a different two-component system, *E. coli* PhoR-PhoB. The rewired proteins interacted robustly with each other, but no longer interacted with the parent proteins. Analysis of the crystal structures of the wild-type and mutant protein complexes, along with a systematic mutagenesis study, reveals how individual mutations contribute to the rewiring of interaction specificity. Our approach and conclusions have implications for studies of other protein-protein interactions, protein evolution, and the design of novel protein interfaces.

Highlights

- *T. maritima* HK853-RR468 was rewired to harbor alternative specificity residues
- the rewired proteins dock similarly to the parent proteins, but with a slight rotation
- the rewired interface is repacked to promote binding and catalytic functions
- systematic analyses show limited mutational trajectories for converting specificity

Introduction

Interacting protein partners must recognize each other while avoiding unproductive interactions within the crowded milieu of the cell. The residues important for a given protein-protein interface must therefore both promote interaction between cognate proteins and prevent, or at least minimize, all possible non-cognate pairings. The challenge of maintaining specificity is particularly acute for proteins that belong to large paralogous protein families, which often share significant similarity to one another at the sequence and structural levels (Gao and Stock, 2009; Keskin et al., 2008).

In bacteria, two-component signal transduction proteins are a prevalent mechanism for sensing and responding to the environment. These signaling pathways rely on a sensor histidine kinase that can autophosphorylate and transfer its phosphoryl group to a cognate response regulator (Stock et al., 2000). Many histidine kinases are bi-functional and can directly dephosphorylate their cognate response regulators (Huynh and Stewart, 2011; Igo et al., 1989). Histidine kinases and response regulators are two of the largest protein families in bacteria, with most organisms encoding tens to hundreds of each type of protein (Alm et al., 2006; Galperin, 2005). However, most histidine kinases phosphorylate only a single cognate response regulator and there is very little cross-talk observed between non-cognate partners (Capra et al., 2012; Laub and Goulian, 2007). Systematic studies of phosphotransfer have demonstrated that histidine kinases typically exhibit a strong kinetic preference for their cognate response regulators *in vitro*, suggesting that the interaction specificity of these signaling pathways is driven largely by molecular recognition rather than the cellular context (Fisher et al., 1996; Skerker et al., 2005).

Previous studies have demonstrated that interaction specificity is dictated by a small subset of residues on each protein (Capra et al., 2010; Skerker et al., 2008). These studies relied on the identification of coevolving amino acids in large multiple sequence alignments of cognate kinase-regulator pairs from a diverse range of bacterial species (Codoner and Fares, 2008). The importance of these residues was validated through the rational rewiring of phosphotransfer specificity. Substituting the specificity residues in the *E. coli* histidine kinase EnvZ with those found in other kinases was sufficient to drive phosphotransfer toward previously non-cognate response regulators. A similar rewiring of the response regulator OmpR allowed it to receive phosphoryl groups from other histidine kinases. These coevolving specificity residues were confirmed as critical to molecular recognition when the first structure of a histidine kinase in complex with its cognate response regulator was solved (Casino et al., 2009; Casino et al., 2010). The complex of *Thermotoga maritima* kinase HK853 bound to a phosphorylated form of RR468 demonstrated that the primary basis of interaction involves the docking of helix $\alpha1$ ($\alpha1$) in the response regulator with both helices of the DHp (Dimerization and Histidine phosphotransfer) domain in the kinase. Nearly all of the specificity residues identified via coevolution studies are found within these helices (Figure 2.1A).

Although two-component proteins have been successfully rewired, it remains unclear how a newly introduced set of specificity residues is accommodated at the molecular interface formed by a histidine kinase and a response regulator. How do individual residues contribute to the rewired specificity of a complex? How do the new residues pack together? Do changes at the interface affect other, distal regions of the proteins? To tackle these questions, we rationally rewired the interaction interface of *Thermotoga maritima* proteins HK853 and RR468 to harbor the specificity-determining residues of an unrelated two-component pathway, *E. coli* PhoR and PhoB. We solved crystal structures of complexes formed by the rewired proteins, as well as the structures of the rewired HK853 and RR468 alone. Comparison of these structures with the native HK853-RR468 complex, along with a systematic mutational analysis of the interface, helps reveal the structural basis of specificity in two-component signaling proteins. More generally, they provide insight into the rules of molecular recognition and coevolution in protein-protein interfaces.

Results

HK853-RR468 and PhoR-PhoB have different phosphotransfer specificities

To investigate the structural consequences of rewiring a kinase-substrate interface, we rationally mutated the specificity residues of the *T. maritima* two-component pathway HK853-RR468 to match those of another two-component system. Previous work has shown that unlike the model kinase EnvZ, the HK853 homodimer autophosphorylates *in cis*, such that the histidine on a given chain is autophosphorylated by the ATP-binding domain of the same chain (Casino et al., 2009). We therefore aimed to reprogram the specificity of the HK853-RR468 system to match that of another system in which the kinase autophosphorylates *in cis*, the *Escherichia coli* system PhoR-PhoB (Ashenberg et al., 2013; Casino et al., 2009). HK853 and *E. coli* PhoR are ~32% identical at the amino acid level across their DHp and CA (Catalytic and ATP binding) domains and share four identities at the nine specificity positions. RR468 and *E. coli* PhoB are ~38% identical across their receiver domains and share two identities at the seven specificity positions (Figure 2.1A).

To confirm that the different specificity residues in HK853-RR468 and PhoR-PhoB yield different phosphotransfer specificities, we purified His_6-tagged versions of each protein. For HK853 and PhoR, we truncated the transmembrane domains, purifying only the soluble, cytoplasmic portions of each kinase (see Methods). RR468 has only a receiver domain, whereas PhoB has a receiver domain and a DNA-binding domain; we purified the receiver domain portion of each regulator. We first autophosphorylated each kinase in the presence of $[\gamma\text{-}^{32}P]$ATP and then added cognate substrates to examine phosphotransfer. At room temperature, HK853 rapidly phosphorylated RR468 (Figure 2.1B). Because HK853 is bi-functional with strong phosphatase activity for RR468~P (Casino et al., 2009), the combination of phosphotransfer and

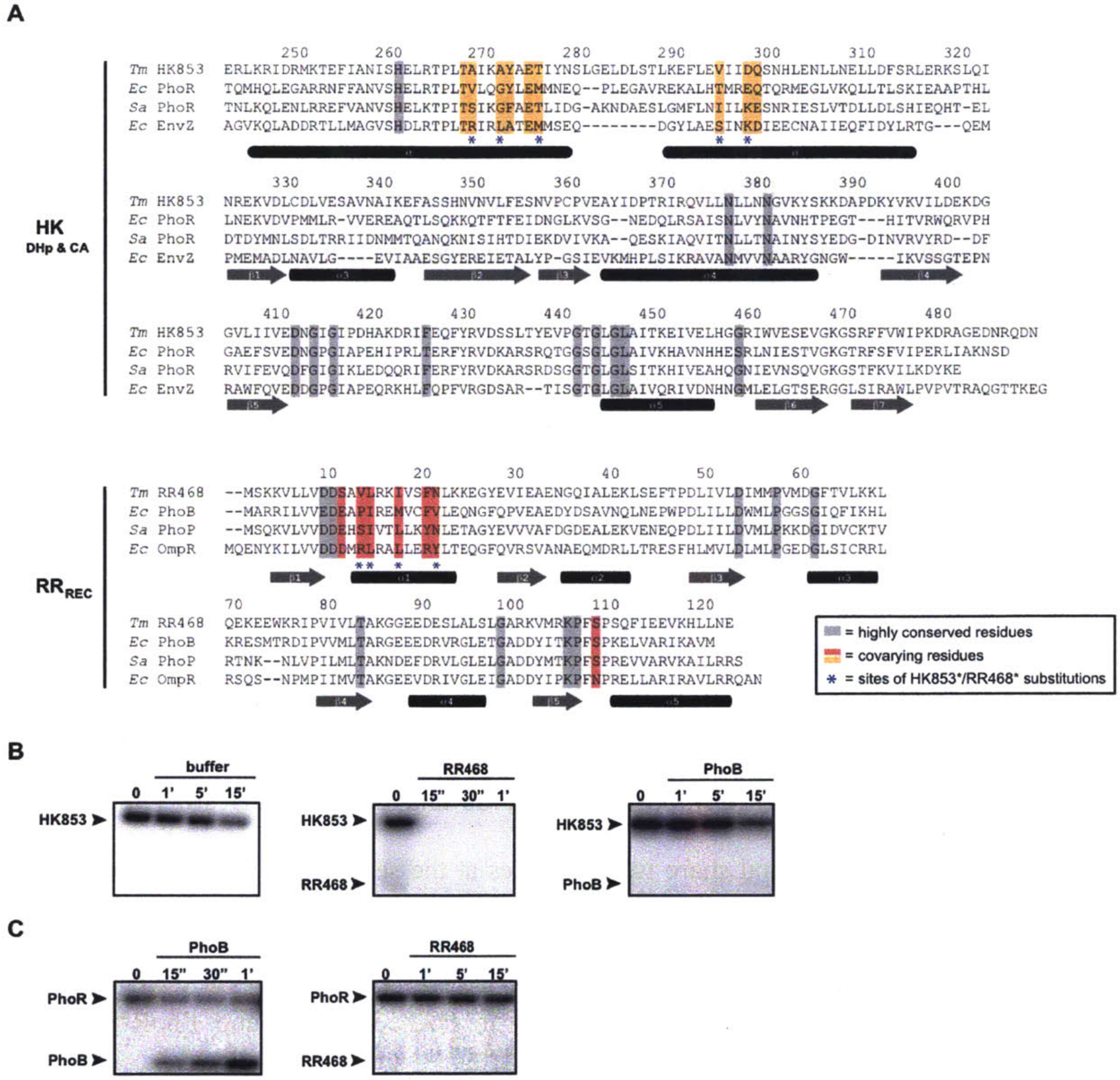

A
250 260 270 280 290 300 310 320
Tm HK853 ERLKRIDRMKTEFIANISHELRTPLTAIKAYAETIYNSLGELDLSTLKEFLEVIIDQSNHLENLLNELLDFSRLERKSLQI
Ec PhoR TQMHQLEGARRNFFANVSHELRTPLTVLQGYLEMMNEQ--PLEGAVREKALHTMREQTQRMEGLVKQLLTLSKIEAAPTHL
Sa PhoR TNLKQLENLRREFVANVSHELKTPITSIKGFAETLIDG-AKNDAESLGMFLNIILKESNRIESLVTDLLDLSHIEQHT-EL
Ec EnvZ AGVKQLADDRTLLMAGVSHDLRTPLTRIRLATEMMSEQ--------DGYLAESINKDIEECNAIIEQFIDYLRTG---QEM
330 340 350 360 370 380 390 400
Tm HK853 NREKVDLCDLVESAVNAIKEFASSHNVNVLFESNVPCPVEAYIDPTRIRQVLLNLLNNGVKYSKKDAPDKYVKVILDEKDG
Ec PhoR LNEKVDVPMMLR-VVEREAQTLSQKKQTFTFEIDNGLKVSG--NEDQLRSAISNLVYNAVNHTPEGT---HITVRWQRVPH
Sa PhoR DTDYMNLSDLTRRIIDNMMTQANQKNISIHTDIEKDVIVKA--QESKIAQVITNLLTNAINYSYEDG-DINVRVYRD--DF
Ec EnvZ PMEMADLNAVLG----EVIAAESGYEREIETALYP-GSIEVKMHPLSIKRAVANMVVNAARYGNGW-----IKVSSGTEPN
410 420 430 440 450 460 470 480
Tm HK853 GVLIIVEDNGIGIPDHAKDRIFEQFYRVDSSLTYEVPGTGLGLAITKEIVELHGGRIWVESEVGKGSRFFVWIPKDRAGEDNRQDN
Ec PhoR GAEFSVEDNGPGIAPEHIPRLTERFYRVDKARSRQTGGSGLGLAIVKHAVNHHESRLNIESTVGKGTRFSFVIPERLIAKNSD
Sa PhoR RVIFEVQDFGIGIKLEDQQRIFERFYRVDKARSRDSGGTGLGLSITKHIVEAHQGNIEVNSQVGKGSTFKVILKDYKE
Ec EnvZ RAWFQVEDDGPGIAPEQRKHLFQPFVRGDSAR--TISGTGLGLAIVQRIVDNHNGMLELGTSERGGLSIRAWLPVPVTRAQGTTKEG
HK
DHp & CA
10 20 30 40 50 60
Tm RR468 --MSKKVLLVDDSAVLRKIVSFNLKKEGYEVIEAENGQIALEKLSEFTPDLIVLDIMMPVMDGFTVLKKL
Ec PhoB --MARRILVVEDEAPIREMVCFVLEQNGFQPVEAEDYDSAVNQLNEPWPDLILLDWMLPGGSGIQFIKHL
Sa PhoP --MSQKVLVVDDEHSIVTLLKYNLETAGYEVVVAFDGDEALEKVENEQPDLIILDVMLPKKDGIDVCKTV
Ec OmpR MQENYKILVVDDDMRLRALLERYLTEQGFQVRSVANAEQMDRLLTRESFHLMVLDLMLPGEDGLSICRRL
70 80 90 100 110 120
Tm RR468 QEKEEWKRIPVIVLTAKGGEEDESLALSLGARKVMRKPFSPSQFIEEVKHLLNE
Ec PhoB KRESMTRDIPVVMLTARGEEEDRVRGLETGADDYITKPFSPKELVARIKAVM
Sa PhoP RTNK--NLVPILMLTAKNDEFDRVLGLELGADDYMTKPFSPREVVARVKAILRRS
Ec OmpR RSQS--NPMPIIMVTAKGEEVDRIVGLEIGADDYIPKPFNPRELLARIRAVLRRQAN
RR_{REC}
▨ = highly conserved residues
🟥 = covarying residues
* = sites of HK853*/RR468* substitutions
B
buffer
0 1' 5' 15'
HK853
RR468
0 15" 30" 1'
HK853
RR468
PhoB
0 1' 5' 15'
HK853
PhoB
C
PhoB
0 15" 30" 1'
PhoR
PhoB
RR468
0 1' 5' 15'
PhoR
RR468

Figure 2.1 Specificity residues in two-component signaling proteins.

(A) Multiple sequence alignment of histidine kinases (DHp and CA domains) and cognate response regulators (receiver domain only), with specificity residues and highly conserved residues highlighted. Species abbreviations: *(Ec) Escherichia coli*; *(Sa) Staphylococcus aureus*; *(Tm) Thermotoga maritima.* Sequences are numbered according to the *Tm* proteins, with the last digit of each number positioned above the relevant amino acid residue. (B) HK853 and (C) PhoR phosphotransfer specificity. Each histidine kinase construct was autophosphorylated with [^{32}P-γ]-ATP and then incubated with a response regulator or buffer at room temperature. Samples were taken at the time points indicated and phosphotransfer assessed by SDS-PAGE and phosphorimaging. Arrowheads indicate the position of autophosphorylated kinase or phosphorylated response regulator.

subsequent dephosphorylation of RR468 led to a loss of radiolabeled HK853 and RR468 within 15 seconds (Figure 2.1B). The rapid disappearance of phosphorylated HK853 resulted from phosphotransfer to the regulator, not simply dephosphorylation of HK853 through hydrolysis. When HK853 was incubated with RR468(D53A), which cannot be phosphorylated, the kinase remained phosphorylated for extended periods of time (Casino et al., 2009; Figure 2.3A). Like HK853, the histidine kinase PhoR rapidly phosphorylated its cognate partner, PhoB (Figure 2.1C). As the PhoR construct harbors only modest phosphatase activity for PhoB~P, continuous phosphotransfer resulted in the accumulation of phosphorylated PhoB over a 1-minute time course. Finally, we examined phosphotransfer from HK853 and PhoR to the non-cognate regulators PhoB and RR468, respectively. Neither HK853 nor PhoR phosphorylated the non-cognate substrate (Figure 2.1B-C). The modest decrease in intensity of the HK853 band likely results from dephosphorylation, not phosphotransfer (Fig. 2.3B). These experiments demonstrate that HK853-RR468 and PhoR-PhoB have different phosphotransfer specificities, consistent with their different specificity residues.

Rewiring the specificity of HK853-RR468 to match that of PhoR-PhoB

To rewire HK853-RR468, we substituted the specificity residues of HK853 and RR468 with those found in PhoR and PhoB, respectively, producing HK853* (A268V, A271G, T275M, V294T, D297E) and RR468* (V13P, L14I, I17M, N21V) (Figure 2.1A). We also substituted the specificity residues of PhoB with those found in RR468, producing PhoB* (P13V, I14L, M17I, V21N). The substitutions introduced into HK853 did not significantly affect kinase autophosphorylation (Figure 2.3B). We then examined phosphotransfer from HK853 and HK853* to the response regulators RR468, RR468*, PhoB, and PhoB* (Figure 2.2).

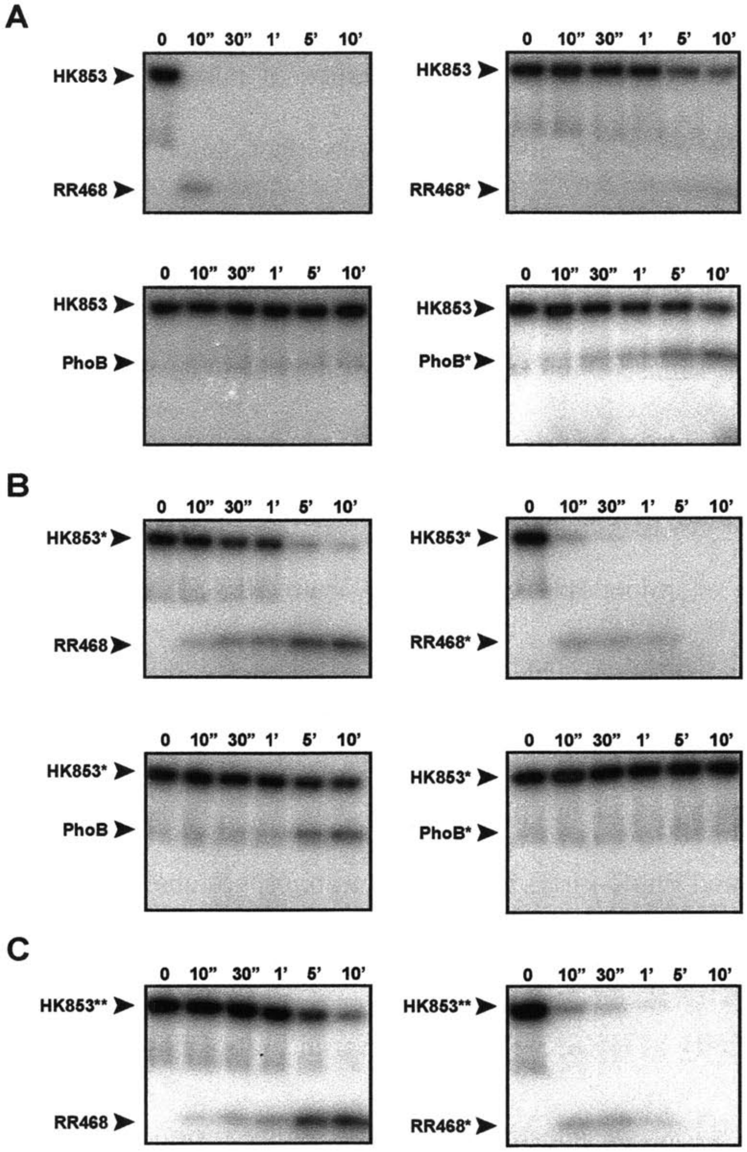


Figure 2.2 Rational rewiring of phosphotransfer specificity.

Phosphotransfer assays for wild-type and mutant two-component proteins. In each panel, the histidine kinase indicated was autophosphorylated with [^{32}P-γ]-ATP and then incubated with the response regulator indicated at 4 °C. Samples were taken at the time points indicated and phosphotransfer assessed by SDS-PAGE and phosphorimaging. (A) Wild-type HK853 and (B) HK853* (which harbors the substitutions A268V, A271G, T275M, V294T, D297E) were tested for phosphotransfer to RR468, RR468*, PhoB, and PhoB*. RR468* contains the substitutions V13P, L14I, I17M, and N21V. PhoB* contains the substitutions P13V, I14L, M17I, and V21N. (C) HK853**, which harbors the substitutions A268V, A271G, and T275M was tested for phosphotransfer to RR468 and RR468*.

Because phosphotransfer from HK853 to RR468 is so rapid at room temperature (Figure 2.1B), we performed these assays at 4°C to facilitate the comparison of relative phosphotransfer rates.

As before, HK853 rapidly phosphorylated and dephosphorylated RR468, with a complete loss of radiolabel within 30 seconds (Figure 2.2A). HK853 phosphorylated RR468* at an extremely slow rate, indicating that the substitutions introduced into RR468* disrupted the cognate interaction. As expected, wild-type HK853 did not transfer to the non-cognate regulator PhoB under these reaction conditions, even after 10 minutes. However, HK853 was capable of phosphorylating PhoB*, indicating that the introduction of RR468-like specificity residues into PhoB was sufficient to promote phosphorylation by HK853 (Figure 2.2A).

For HK853*, we observed significantly reduced rates of phosphotransfer to RR468, indicating that changes to the specificity residues of HK853 had diminished the cognate pairing of HK853 and RR468 (Figure 2.2B). Strikingly, however, HK853* rapidly phosphorylated and dephosphorylated RR468*, indicating that the substitutions in RR468* restored a robust interaction with HK853*. Consistent with its new specificity residues, we found that HK853* could phosphorylate PhoB, but not PhoB* (Figure 2.2B). We confirmed that HK853* has phosphatase activity toward RR468*, but not RR468, by using [^{32}P] acetyl-phosphate to radiolabel RR468*; subsequent addition of HK853* led to a rapid loss in radiolabel compared to buffer alone (Figure 2.3C). This result indicates that HK853* stimulates the dephosphorylation of RR468*.

A mutant of HK853, designated HK853**, harboring only three of the five specificity mutations (A268V, A271G, T275M) behaved similar to HK853*, suggesting that these three residues in the middle of $\alpha1$ were sufficient to rewire interaction specificity (Figure 2.2C). Collectively, these findings demonstrate that introduction of PhoR-like and PhoB-like substitutions into HK853 and

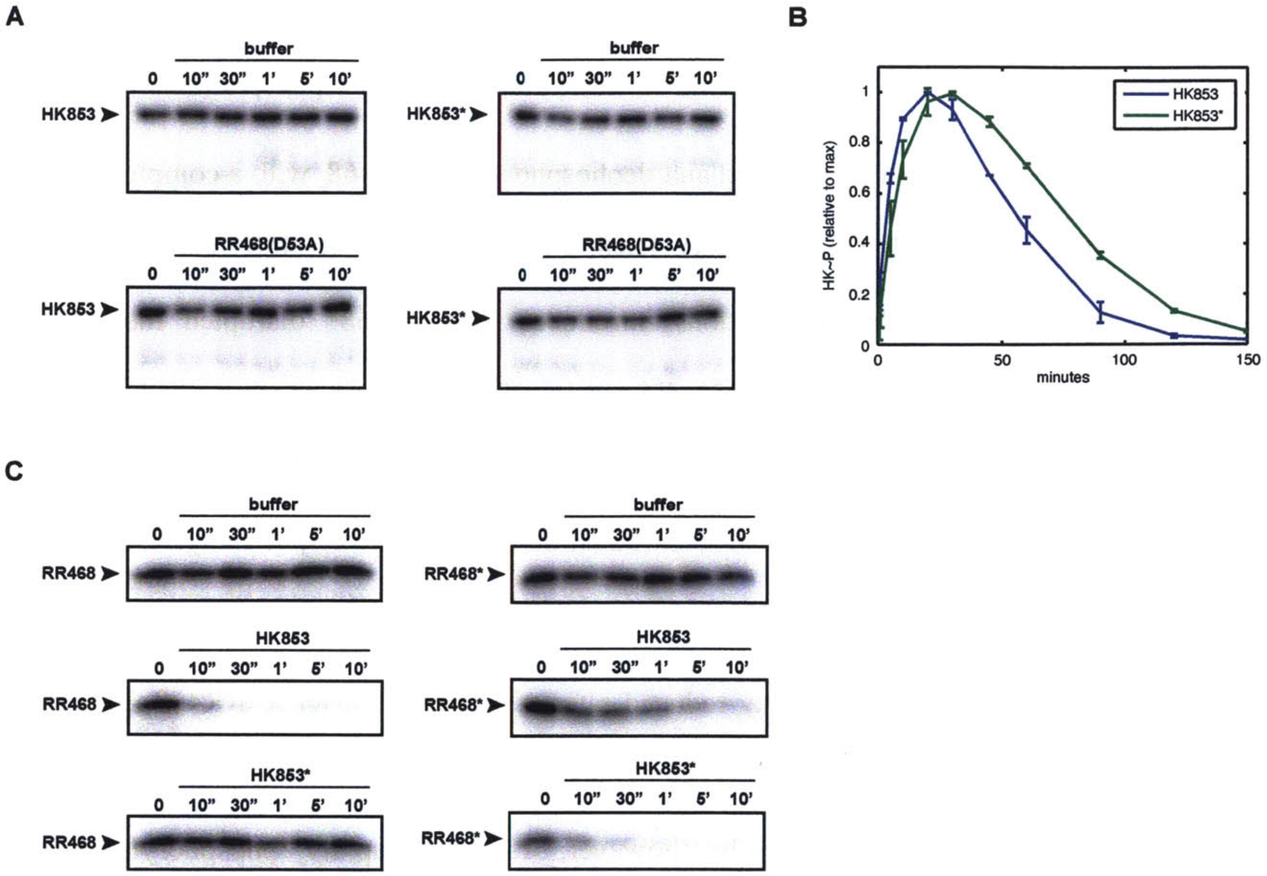


Figure 2.3 Kinase and regulator dephosphorylation.

(A) Time courses illustrating the stability of autophosphorylated HK853 and HK853* at 4 °C. Experiments were performed as in Figure 2.2. (B) Time course of HK853 and HK853* autophosphorylation at room temperature, starting immediately after the addition of [^{32}P-γ]-ATP to 5 μM kinase. Quantified autophosphorylation intensity is shown as percentage of max for each kinase. Error bars show ± SD of two experiments. (C) Phosphatase experiments with RR468 and RR468*. The regulators were phosphorylated using [^{32}P] acetyl-phosphate and incubated with buffer, HK853, or HK853* at 4 °C. We used 5 μM of each protein (regulator and kinase) in the RR468 experiments and 10 μM of each protein in the RR468* experiments.

RR468 was sufficient to reprogram their phosphotransfer specificity while maintaining phosphotransfer and dephosphorylation rates comparable to those seen in the wild-type proteins.

Structural characterization of the rewired complex reveals changes in relative orientation between interacting proteins

To investigate the effects of the new specificity residues on partner recognition, we solved an X-ray crystal structure of the rewired complex formed by HK853* and RR468* (HK*-RR*) (Table 2.1), which can engage in both phosphotransfer and phosphatase reactions (Figure 2.2B, Figure 2.3C). The rewired complex preserves the stoichiometry of the original complex (HK-RR; PDB:3DGE) (Casino et al., 2009): HK853* forms a homodimer that interacts with two molecules of RR468* (HK_2-2RR) (Figure 2.4). In the rewired complex, the HK_2-2RR is generated by crystallographic two-fold symmetry while in the HK-RR complex the asymmetric unit contained HK_2-2RR (Table 2.1). Further, the structures of the individual components are highly similar. The kinases superimpose with an rmsd value of 1.82 Å (Table 2.2, Figure 2.5A). The superposition of individual domains shows even more similarity, with rmsd values of 1.0 Å and 0.87 Å for the DHp (residues 245-317) and CA (residues 320-480) domains, respectively. The slightly greater rmsd value for the kinases relative to individual domains results from a 19.7° rigid body rotation of the HK853* CA domain toward $\alpha1$ of the DHp domain, which agrees with the structural plasticity and dynamic nature reported for the CA domain (Albanesi et al., 2009). RR468* has an almost identical structure to RR468 in the two functional complexes, HK-RR and HK*-RR*, with an rmsd of 0.6 Å (Table 2.3).

Despite high similarity at the domain level, the arrangement and relative orientation of the kinase with respect to the regulator differs in the rewired complex. In particular, RR468* is rotated 17.5° and translated 0.5 Å along the DHp domain, such that the angle between $\alpha1$ of HK853* (residues 246-279) and $\alpha1$ of RR468* (residues 12-26), the two key structural elements for

Table 2.1 Crystallographic data and refinement statistics

Processed data	**HKf***	**RRf***	**HK*-RR***	**HK-RR***
Wavelength (Å)	0.92	0.87	0.98	0.87
Resolution (Å)	72.98-2.70 (2.85-2.70)	35.74-1.80 (1.88-1.79)	48.34-3.00 (3.16-3.00)	46.32-3.10 (3.27-3.10)
R_{merge} (%)	0.057 (0.337)	0.078 (0.279)	0.060 (0.389)	0.071 (0.401)
R_{pim} (%)	0.023 (0.13)	0.030 (0.107)	0.036 (0.226)	0.035 (0.198)
Mean I/δ(I)	21.0 (5.6)	19.7 (7.6)	15.0 (3.7)	17.9 (4.1)
N° reflections (observed/unique)	60367/8277 (8550/1162)	84433/10950 (11971/1566)	43627/12140 (6406/1750)	108863/ 22038 (15794/3160)
Completeness (%)	99.5 (99.2)	100.0 (100.0)	98.3 (98.9)	99.9 (100.0)
Redundancy	7.3 (7.4)	7.7 (7.6)	3.6 (3.7)	4.9 (5.0)
Space group	$C222_1$	I222	I222	$C222_1$

Cell dimensions (Å)	a=81.96
	b=160.38
	c=43.89
		a=53.98
		b=58.06
		c=71.47
			a=75.71
			b=85.31
			c=185.59
				a=119.32
				b=143.93
				c=138.97

Refined data				
R_{factor} (%)	0.236	0.186	0.209	0.202
R_{free} (%)	0.279	0.223	0.252	0.253
Asymmetric unit composition	1 HK	1RR	1HK:1RR	2HK:2RR
N° protein atoms	1890	976	2856	5842
N° water molecules	44	91	13	13
N° ligand/ion	1	5	4	14
RMSD				
Bond deviation (Å)	0.009	0.008	0.005	0.004
Angle deviation (°)	1.3	1.3	1.0	0.9
Media B-factor (Å^2)				
Main chain	70.3	11.4	72.7	71.3
Side chain	71.2	13.5	75.2	73.2
All atoms	70.7	12.4	73.9	72.2
Ramachandran Map (%)				
Favoured	96.88	99.15	94.89	96.81
Allowed	3.12	0.85	4.83	3.06
Disallowed region	0	0	0.28	0.14
PDB accession code	4JAU	4JA2	4JAS	4JAV

Values in parentheses correspond to data for the highest resolution shell

Rmerge = ΣhklΣi | I(hkl)i - <I(hkl)> | / ΣhklΣi<I(hkl)i>

Rpim = Σhkl√(1/(n-1)) Σi | I(hkl)i - <I(hkl)> | /ΣhklΣi I(hkl)i

Rfactor= Σ||Fo|−|Fc||/Σ|Fo|

Rfree is the Rfactor calculated with 5 to 7% of the total unique reflections chosen randomly and omitted from refinement.

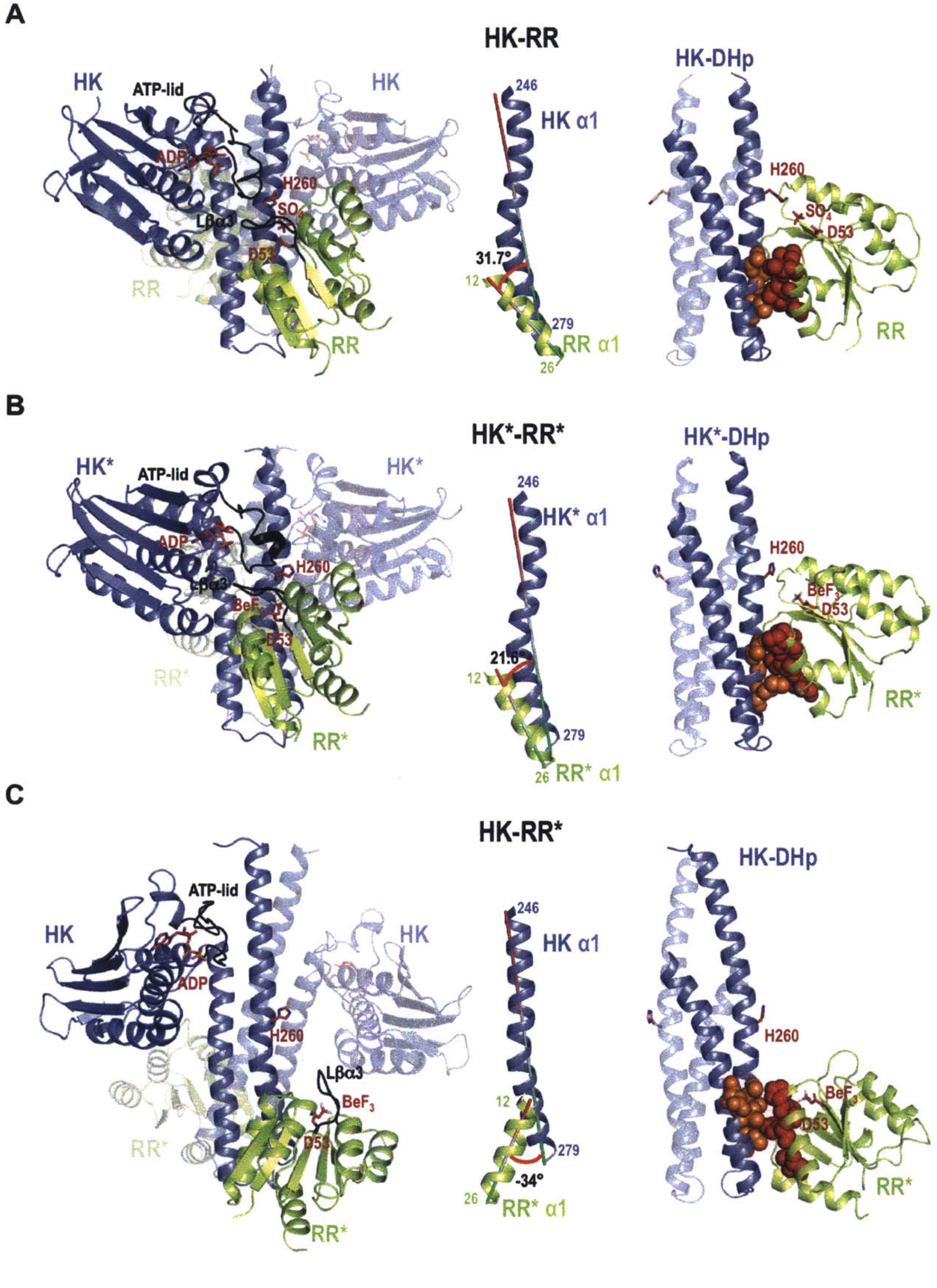


Figure legend on next page.

Figure 2.4 Crystal structures of the wild-type complex HK-RR, the rewired and functional HK*-RR* complex, and the impaired HK-RR* complex.

Cartoon representation of (A) HK-RR complex with HK853 bound to ADP and RR468 D53 bound to SO_4, (B) HK*-RR* complex with HK853* bound to ADP and RR468* D53 bound to BeF_3^- and (C) HK-RR* complex with HK853 bound to ADP and RR468* D53 bound to BeF_3^-. *Left*; cartoon representations of the overall structure of the three complexes formed by a homodimeric HK (blue colored with one subunit transparent) bound to two molecules of RR (yellow-green colored with one molecule transparent). In each complex, the ATP-lid in the HK and the β3-α3 linker in the RR are colored in black; the phosphorylatable residues H260 and D53 as well as bound ligands ADP, sulfate (SO_4) and beryllium trifluoride (BeF_3) are shown as sticks. *Middle;* the angle formed between the interacting helices HK α1 (246-279) and RR α1 (12-26) for each complex is shown. *Right*; HK-RR interface shown by the DHp domain (with one subunit transparent) bound to one RR with the critical specificity residues (13, 14, 17 and 21 in red for RR468 and 268, 271, 275, 294 and 297 in orange for HK853) highlighted in space-filling spheres.

Table 2.2 RMSD differences in the HK component (residues 245-480)

Rmsd (Å)	HK-RR	HK*-RR*	HK-RR*	HKf	HKf*
	(3DGE)[1]			(2C2A)[1]	
HK-RR	-	1.82	5.6	5.2	4.95
HK*-RR*	-	-	4.8	4.35	4.1
HK-RR*	-	-	-	2.16	1.92
HKf	-	-	-	-	0.89

[1] PDB codes for the previously published structures.

Table 2.3 RMSD differences in the RR component (residues 2-121)

Rmsd (Å)	HK-RR	HK*-RR*	HK-RR*	RRf*
HK-RR	-	0.61	0.69	0.67
HK*-RR*	-	-	0.6	0.54
HK-RR*	-	-	-	0.47

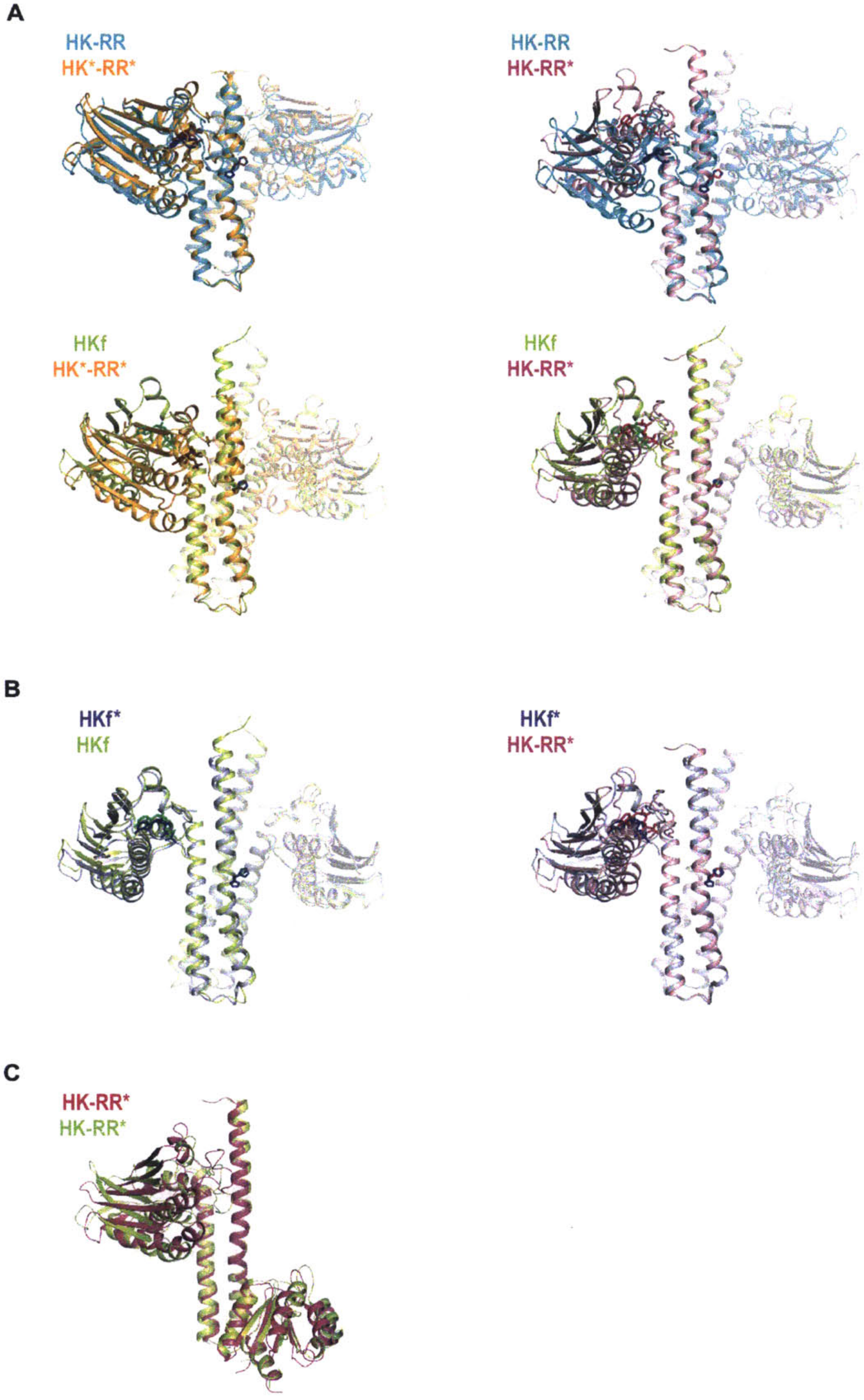


Figure legend on next page.

Figure 2.5 Superposition of free and complex structures.

(A) Top row: superposition of the HK853 component of HK-RR (cyan; PDB:3DGE) with the corresponding component of HK*-RR* (orange) and HK-RR* (magenta). Bottom row: free HK853, HKf, (yellow-green; PDB:2C2A) superposed with the corresponding component of HK*-RR* (orange) and HK-RR* (magenta). (B) Superposition of HKf (yellow-green) with HKf* (blue) and with HK-RR* (magenta). The phosphorylatable His260, as well as the ADP ligand, are shown as sticks in similar color to the cartoon representation of the structure from which they are extracted. (C) Superimposition of the individual HK-RR complexes (green and magenta) present in the asymmetric unit for the HK-RR* structure.

complex formation, is 21.6° (Figure 2.4B). In the HK-RR complex this same angle is 31.7°. Because of this rotation, the HK853*-RR468* complex has a more parallel orientation between α1 in the regulator and the kinase helical bundle. In this orientation RR468* loses some interactions with the DHp α-helix 2 and increases interactions with the CA domain. In particular, the β2-α2 and β3-α3 loops of RR468* now interact directly with the α3-β1 loop and the ATP-lid of the HK853* CA domain. These new contacts lead to a larger interaction surface area in the HK*-RR* complex relative to HK-RR (2690 $Å^2$ *vs* 1866 $Å^2$) (Figure 2.6 and Table 2.4). However, the new contacts specific to the HK*-RR* complex appear largely dispensable for phosphotransfer and dephosphorylation. The differences in orientation are likely due to the mutations introduced and not to differences in crystal packing since RR468* in the HK*-RR* complex could acquire the disposition of RR468 in HK-RR without steric crystallographic clashes.

Although HK853 phosphorylates and dephosphorylates the mutant RR468* very slowly (Figure 2.2A, Figure 2.3C), we were able to solve a structure of these proteins in complex (HK-RR*). The asymmetric unit showed HK_2-2RR stoichiometry with almost two-fold symmetry (Table 2.1), broken due to a slightly different relative disposition of the CA domain with respect to the

Table 2.4 Intermolecular contacts in the three different complexes.

HK-RR	HK*-RR*	HK-RR*
H260-M55 (H)	H260-K85 (H)	
	H260-A84 (H)	
R263-K105 (P)	R263-K105 (P)	
R263-A84 (H)	R263-A84 (H)	
	T264-S11 (H)	
L266-P106 (H)		
T267-P106 (P)	T267-P106 (H)	
T267-L14 (H)	T267-I14 (H)	
T267-F107 (H)	T267-F107 (H)	
	T267-K105 (H)	
A268-V13 (H)	V268-P13 (H)	A268-P13 (H)
	V268-I14 (H)	
K270-P106 (H)		
K270-F107 (H)		
A271-I17 (H)		
A271-P109 (H)	G271-P109 (H)	
Y272-V13 (P)		
Y272-I17 (H)	Y272-M17 (P)	Y272-M17 (P)
E274-S108 (P)	E274-S108 (P)	
E274-P109 (H)	E274-P109 (H)	
T275-N21 (P)		
T275-I17 (H)	M275-M17 (H)	
T275-F20 (H)	M275-F20 (H)	
N278-K24 (P)		
S279-F20 (H)		
S279-K24 (P)	S279-K24 (P)	
E282-K24 (H)		E282-Q111 (P)
		E282-S110 (P)
L283-F20 (H)		
		T287-F20 (H)
E290-K16 (SB)		E290-F20 (H)
F291-K16 (H)		
F291-I17 (H)		F291-M17 (H)
F291-F20 (H)	F291-F20 (H)	
V294-V13 (H)		
E303-P106 (H)	E303-P106 (H)	
R314-E89 (P)		
R314-G87 (H)		
	R317-E89 (P)	
Q321-E88 (P)		
	S346-E33 (H)	
	H347-P57 (H)	
K387-P57 (H)	K387-M59 (H)	
	K387-Q36 (P)	
	K387-V58 (H)	
	D388-Q36 (P)	
E438-P57 (H)	E438-M56 (P)	
E438-M55 (H)	E438-M55 (H)	
	E438-K85 (SB)	
	P440-M56 (H)	

Type of interactions: (H) Hydrophobic, (P) Polar and (SB) Salt bridge.
Areas of interaction: DHp in magenta, CA in green and with the other subunit in blue.

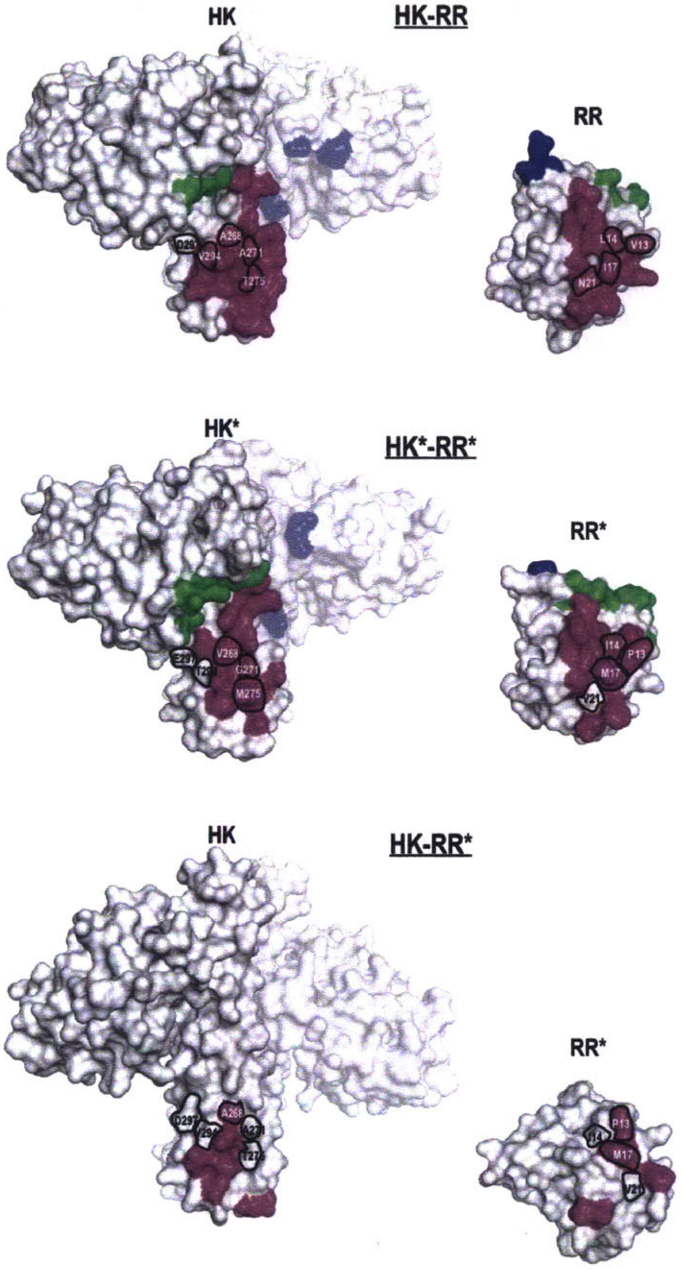


Figure 2.6 Surface of interaction in the HK-RR, HK*-RR* and HK- RR* complexes.

Surface representation of the three complexes in white, showing a homodimer HK, with one subunit transparent, and one RR molecule separated from the complex and rotated 180° with respect to the HK for visualization purposes. Interactions between the RR and the DHp domain are colored magenta; interactions between the RR and the CA domain are colored green; interactions between the RR and the other subunit of the HK are colored blue. Residues subjected to mutagenesis in the HK are indicated by thick black lines, labeled in white if they interact in the complex and in black if they do not.

DHp domain (~18.8°), confirming the previously mentioned plasticity of the CA domain (Figure 2.5C). Both HK subunits in the HK-RR* complex adopt a conformation that is more similar to the free form of HK853 (HKf; PDB:2C2A; rmsd = 2.2 Å) than to the kinase in the HK-RR complex (rmsd=5.6 Å) (Table 2.2, Figure 2.5). The structure of RR468* in the HK-RR* complex is similar to RR468* in the HK*-RR* complex (rmsd=0.6 Å, Table 2.3), but the RR468* molecule in HK-RR* adopts a totally different position relative to the kinase. In the HK-RR* complex, RR468* α1 is rotated ~55° and slightly displaced (~1.0 Å) relative to RR468* and RR468 in the rewired HK*-RR* and native HK-RR complexes, respectively (Figure 2.4C). Many of the intermolecular contacts seen in the productive HK-RR and HK*-RR* complexes are lost in the HK-RR* complex, consistent with its greatly diminished phosphotransfer rate (Table 2.4). Finally, the phosphorylatable residues in RR468* (Asp53) and HK853 (His260) are extremely far apart (19.0 Å) and improperly oriented for any catalytic reaction (Figure 2.4C).

Given the structural data for the functional complexes, HK-RR and HK*-RR*, we conclude that there is a permissible range of rotational motion of the kinase relative to the regulator that allows proper positioning of the active site while accommodating new interfacial residues.

Differential contacts in the rewired functional complex are largely dispensable

To test whether new interactions between the CA domain and RR* in the HK*-RR* complex affect phosphotrasfer, we made mutations at several positions in HK and HK*, introducing either smaller or bulkier residues to try and disrupt the newly observed contacts. For the two mutations in the kinase, S346Y and H347G, there were no significant effects on phosphotransfer or dephosphorylation kinetics (Fig. 2.7A-B). Two mutants in RR468, D60G and D60E, significantly slowed the rate of dephosphorylation but affected both wild type and mutant HK853 similarly (Fig. 2.7). As none of the mutations made affected only the HK*-RR* complex,

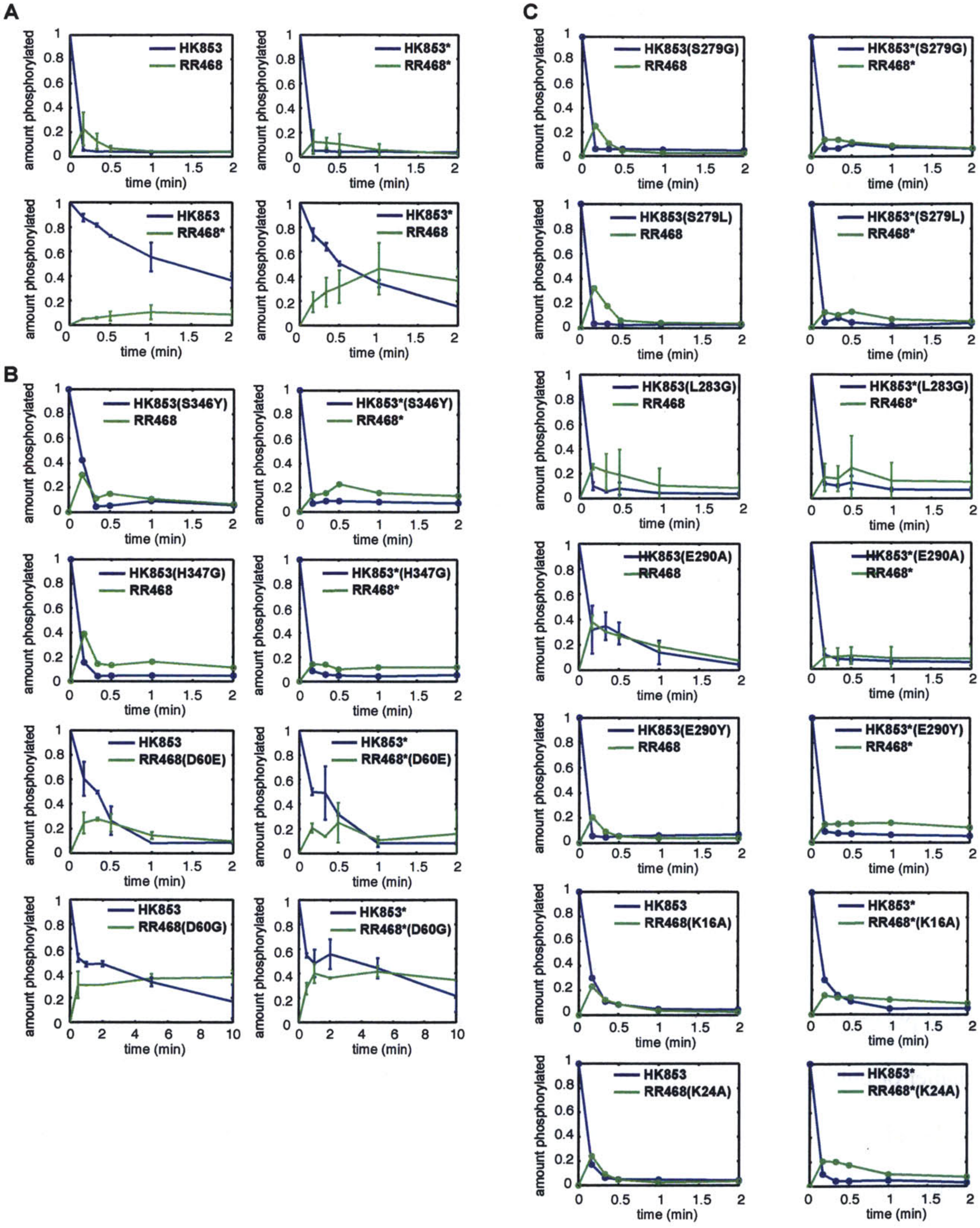


Figure legend on next page.

Figure 2.7 Phosphotransfer time-courses for selected mutations.

Each histidine kinase construct was autophosphorylated with [32P-γ]-ATP and then incubated with the response regulator indicated at 4 °C. Samples were taken at the time points indicated and phosphotransfer assessed by SDS-PAGE and phosphorimaging. Bands corresponding to phosphorylated kinase (blue points) and regulator (green points) were quantified and normalized by the initial amount of autophosphorylated kinase. (A) Wild-type and rewired HK853 and RR468 (also see Fig. 2.2). (B-C) HK853 and HK853* harboring mutations at (B) the interface between the CA domain and the response regulator or (C) the interface between the DHp stem and the response regulator. Where shown, error bars represent ± SD of two experiments.

we conclude that these new contacts observed in the HK*-RR* structure are likely not critical to the activity of HK853*.

In contrast, the HK-RR* complex has lost all contacts with the CA domain but maintains three DHp-RR α1 interactions (A268-P13, Y272-M17 and F291-M17) conserved in the three complexes. The HK-RR* complex also shows new interactions between HK α1 with RR α1 (T287-F20 and E290-F20) and with β5 strand (E282-S110 and E282-Q111). As a result, RR468* has recognized a similar region to anchor to the kinase but the lack of complementarity with the kinase has forced the RR to slide around the DHp domain.

Although the key specificity residues pack against each other in both HK-RR and HK*-RR* complexes, certain intermolecular interactions involving the base of the DHp domain and receiver domain are lost in HK*-RR* due to the relative rotation of the RR*. We made mutations at sites outside the rewired specificity residues to test the importance of these lost interactions (Fig. 2.1). The mutations K16A and K24A in the response regulator and S279G, S279L, and E290Y in the kinase had no effect on phosphotransfer kinetics of either the HK-RR or the HK*-RR* complexes (Fig. 2.7C). Mutation L283G, which is located in the α1-α2 linker of the histidine kinase, led to a minor decrease in phosphatase activity for both HK-RR and HK*-RR* complexes. Thus, the residues at the bottom of the DHp domain are in contact in the HK-RR complex, but are not critical for the interaction of TM853 and RR468. In sum, the structural differences between the HK-RR and HK*-RR* structures do not result in significant differences in phosphotransfer or dephosphorylation by the two protein complexes. Despite having a new set of specificity residues, HK*-RR* is functionally and structurally very similar to HK-RR.

The active center of the rewired complex is a snapshot of a new intermediate state in phosphotransfer

A closer view of the active sites in the HK*-RR* and HK-RR* complexes shows the phosphomimetic beryllium trifluoride (BeF_3^-) bound to the catalytic Asp53 of RR* in both cases. In HK*-RR* the BeF_3^- is placed similarly to the sulfate found in the active site of the HK-RR complex, but the Be atom is slightly closer to the phosphorylatable His260 (Be-His260 Cα=7.85 Å) than the sulphur atom of the sulfate in the HK-RR complex (S-His260 Cα=8.30 Å) (Figure 2.8). However, Met55 of RR468* is interposed between the His and the BeF_3^- in the HK*-RR* structure, forcing an alternative rotamer for His260 that points away from the active site. Thus, the structure of the HK*-RR* mutant complex may represent an earlier phase of the phosphatase reaction, when the phosphoryl group is still bound to the response regulator. Alternatively, if His260 acquired the rotamer conformation observed in the HK-RR complex, the distance between the phosphoacceptor nitrogen of this residue and the Be atom in the HK*-RR* complex would be 3.6 Å, a distance compatible with phosphotransfer (Figure 2.8). This observation suggests that the conformation captured in the crystal could, instead, correspond to the end of the phosphotransfer reaction just prior to complex dissociation. Because the phosphatase reaction is not the reverse of a phosphotransfer reaction (Hsing and Silhavy, 1997), the active center observed in the crystal could correspond to either of these reactions.

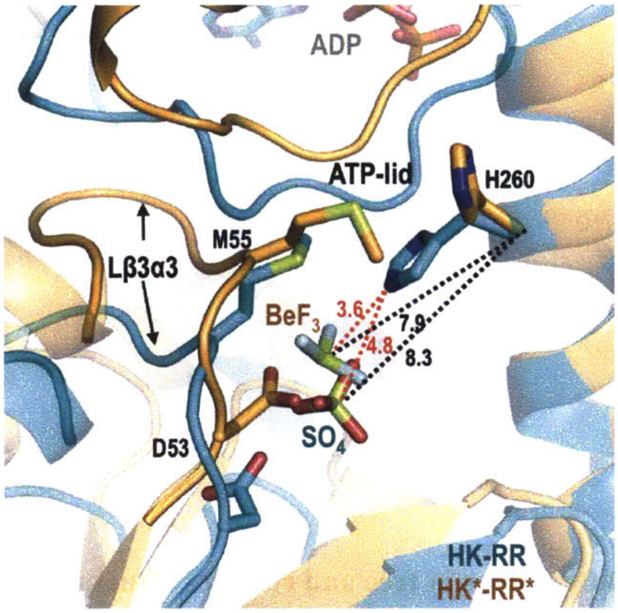


Figure 2.8 Comparison of the active center in the HK-RR and HK*-RR* complexes.

Close-up view of the active center with a superposition of HK-RR (in cyan) and HK*-RR* (in orange) in cartoon representation. The phosphorylatable residues H260 and D53, residue M55, and the bound ligands (sulfate (SO_4) in the HK-RR complex and beryllium trifluoride (BeF_3) in the HK*-RR* complex) are shown as sticks. Distances are shown by dashed lines; in black color for Cα of H260 in HK-RR with the sulfur atom of SO_4 (8.3 Å) and for Cα of H260 in HK*-RR* with the Be atom of BeF_3 (7.9 Å); in red color for εN of H260 in HK-RR with the sulfur atom of SO_4 (4.8 Å) and with the Be atom of BeF_3^- (3.6 Å).

Introduction of new specificity residues does not affect global structural integrity

Structural differences between the different HK-RR complexes could be imposed by specific requirements for partner recognition, or could reflect intrinsic changes in the individual proteins resulting from the point mutations. To address this issue, we solved the structures of the free forms of HK853* (HKf*) bound to ADP and RR468* (RRf*) bound to BeF_3^- (Table 2.1). The conformations of HKf* and HKf are almost identical (rmsd = 0.9 Å for the superposition of the structures; Table 2.2, Figure S2.5B). However, there are local changes around the mutated residues that can be attributed to interactions between DHp α-helices 1 and 2 mediated by the

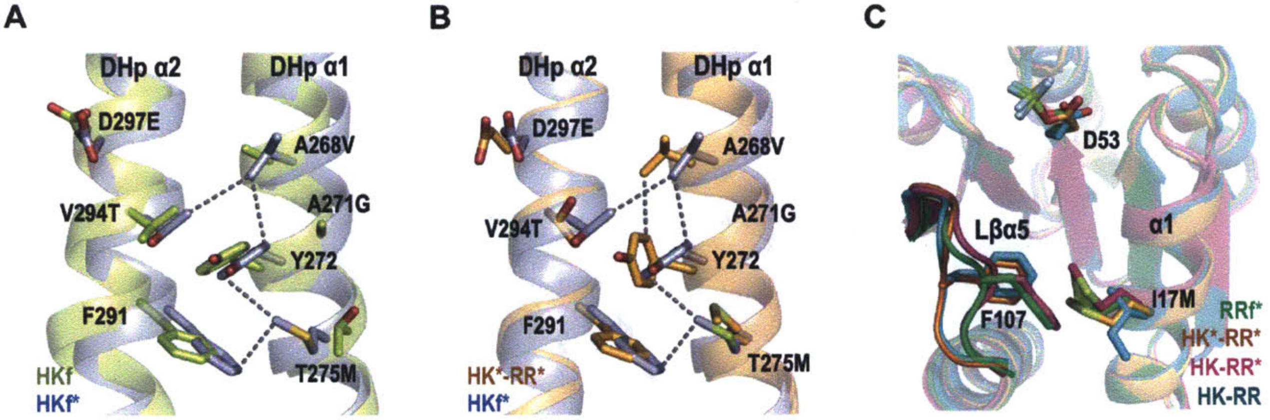


Figure 2.9 Differential interactions in free and complex structures.

(A) DHp domains with a superposition of HKf (in green) and HKf* (in blue) to show new interactions resulting from the mutations introduced into HKf*: A268V, A271G, T275M, V294T and D297E. (B) DHp domains with a superposition of HKf* and HK853* from the HK*-RR* complex (in orange) to show changes in the interactions for M275. (C) Superposition of RR468 in RRf* (green), HK-RR (cyan), HK*-RR* (orange) and HK-RR* (magenta) structures shows the interaction M17-F107 and the different conformations of F107 and the β5-α5 linker in the RR alone or in complex. All the structures are shown in cartoon representation with the selected residues labeled in black and drawn as sticks.

new side chains. In HKf* V268 and M275, together with T294 and F291 from α2 and Y272 from α1, generate a hydrophobic network (Figure 2.9A). These interactions induce changes in the exposed recognition surface and could impair interaction with the regulator. The flexibility of the new Met side chain seems to play a key role in this network, since it is sandwiched between the Y272 and F291 aromatic rings (Figure 2.9A). The interaction of V268 and M275 with the complementary surface provided by the mutated residues of RR468* promotes the disappearance of this hydrophobic network in the productive complex (Figure 2.9B).

As with HK853*, comparison of RR468* in isolation and in the HK*-RR* complex shows minimal structural differences (Table 2.3) that are localized around the mutated residues. The I17M mutation seems to have the biggest structural impact on RR468* as the new Met side chain now mediates a hydrophobic interaction with F107 that was not present in RR468 (Figure 2.9C). This interaction involves a 2.4 Å displacement of F107 toward M17, a movement that brings the β5-α5 linker (Lβα5) closer to α1. In the HK-RR complex, these two structural elements of RR468 clamp α1 of the kinase's DHp domain. Although the M17-F107 interaction occurs in both RRf* and HK-RR*, F107 is positioned perpendicular to the M17 side chain in the HK*-RR* complex, suggesting that mutations in HK* may compensate for the presence of M17 (Figure 2.9C). The positioning of F107 in HK*-RR* is similar to that seen in HK-RR, suggesting this conformation may be important for the formation of a productive complex.

Systematic mutational characterization of the rewired interface

To further analyze the effects of individual substitutions at the HK853-RR468 interface (Figure 2.10A), we constructed all possible mutational intermediates separating the HK-RR and HK*-RR* pairings. Consistent with our previous finding that only three of the five mutations (A268V, A271G, T275M) in HK853 are necessary to rewire the interface, residues V294 and D297 do not make inter-protein contacts in the HK*-RR* structure (Table 2.2). We therefore constructed three single and three double mutants in HK853; along with the wild type and triple mutant, there were eight different residue combinations for the kinase. For RR468 we made four single, six double, and four triple mutants; along with the wild type and quadruple mutant (V13P, L14I, I17M, N21V), there was a total of 16 specificity residue combinations for the regulator. For simplicity, we refer to each of the mutants according to the identities of their specificity residues.

We systematically tested all 128 pairwise combinations of the 8 kinases and 16 regulators for activity by incubating each autophosphorylated kinase with each regulator for 15 seconds (Figure 2.10B). Residue combinations that supported both phosphotransfer and dephosphorylation, such as the wild-type residues, led to a depletion of radiolabeled kinase with minimal accumulation of

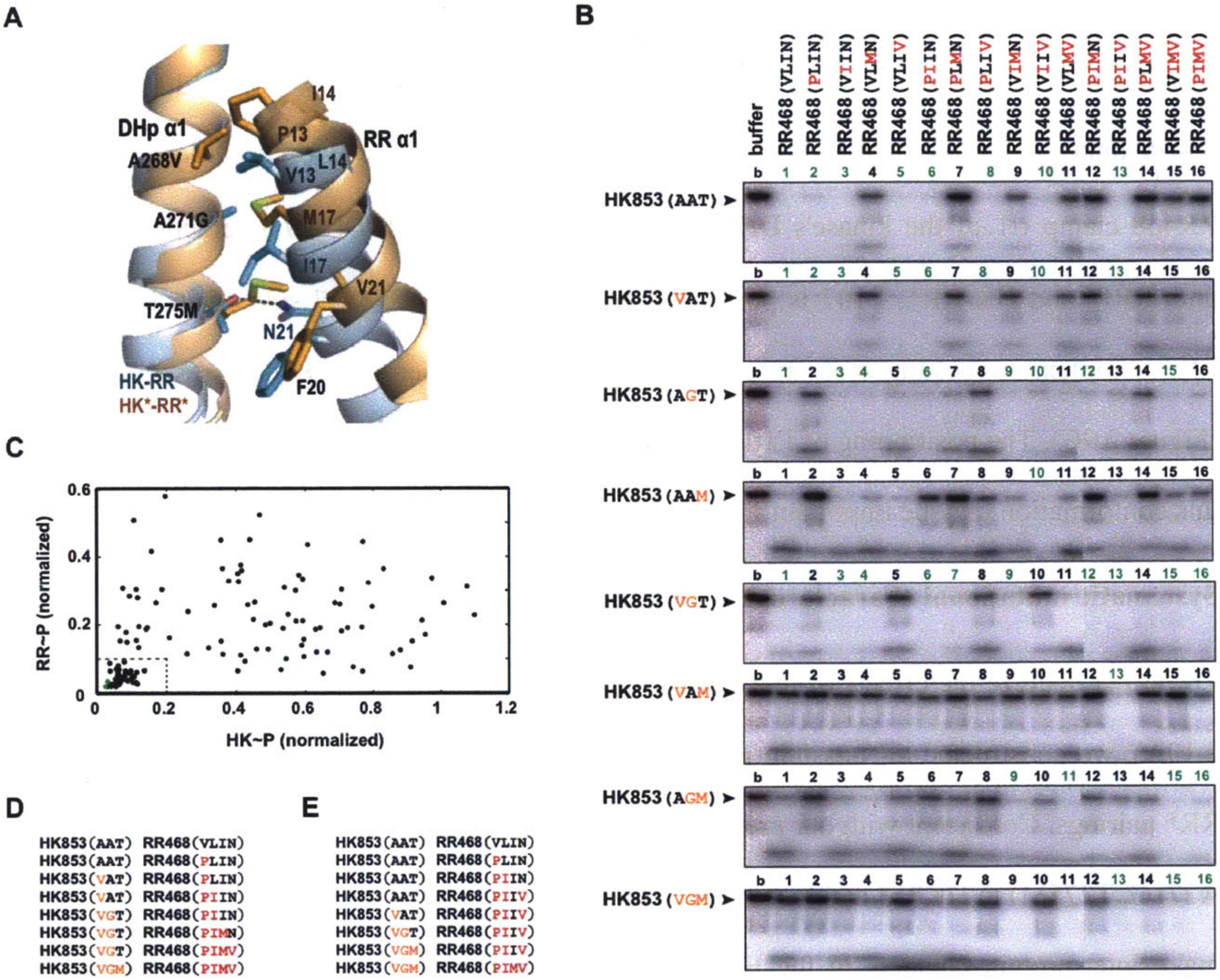


Figure 2.10 Phosphotransfer between all possible mutational intermediates separating the complexes HK-RR and HK*-RR*.

(A) Superposition of HK-RR (in cyan) and HK*-RR* (in orange) complexes highlighting how mutations in DHp α1 (A268V, A271G and T275M) affect interactions with mutations in RR α1 (V13P, L14I, I17M and N21V) and F20. A dashed line represents a polar interaction between N21 and T275 in the HK-RR complex. (B) Phosphotransfer assays for wild-type HK853 and HK853 harboring all possible combinations of one, two, or three PhoR-like specificity substitutions present in HK853** (A268V, A271G and T275M). Each lane represents the incubation of the indicated autophosphorylated kinase with the indicated response regulator for 15 seconds at room temperature. Reactions 1-11 and 12-16 were run on separate SDS-PAGE gels; the resulting phosphorimages were contrasted identically and stitched together. (C) The histidine kinase (HK) and response regulator (RR) bands from the phosphotransfer experiments in panel B were quantified and plotted. For each mutational pairing, the x-axis value indicates the intensity of the autophosphorylated HK band (HK~P) normalized to the intensity of the autophosphorylated kinase band and the y-axis value indicates the intensity of the phosphorylated response regulator band (RR~P). In each case, band intensities were normalized to the intensity of the autophosphorylated kinase incubated without response regulator (lane 1 of each gel in panel B). Green points indicate the pairs HK853-RR468 and HK853*-RR468*. The box in the lower left indicates pairings deemed functional; a low level of both the kinase and regulator bands reflects efficient phosphotransfer and dephosphorylation. The 43 functional pairings are underlined in panel B. (D) One example of a mutational path from the wild type to the rewired complex in which each intermediate state is functional. (E) An example of a mutational path in which all mutations to the kinase occur in three successive steps.

radiolabeled regulator. Combinations that supported only phosphotransfer led to an accumulation of phosphorylated regulator, and unproductive combinations retained phosphorylated kinase. Although the *in vivo* role of HK853/RR468 is unknown, studies with other two-component systems have demonstrated that both phosphotransfer and phosphatase activity are important for proper signal transduction (Huynh and Stewart, 2011). Therefore we identified functional kinase-regulator pairs for which (i) the intensity of the phosphorylated kinase band in the presence of regulator was less than 20% the intensity of the kinase band in the presence of buffer alone and (ii) the intensity of the phosphorylated regulator band was less than 10% the intensity of the autophosphorylated kinase band in the presence of buffer. Of the 128 combinations, 43 pairs satisfied these criteria (Figure 2.10B-C).

Our systematic mutagenesis shed light on the amino acid combinations permissible at this protein interface. For example, the profile of HK853 against the 16 regulator mutants demonstrates that robust phosphotransfer and dephosphorylation are retained for all regulators except those harboring the I17M mutation (Figure 2.10B). For most of these pairings, the defect resulting from an I17M substitution can be rescued by the substitution A271G in the kinase. Given the similarity in size and nature between isoleucine and methionine, it is difficult to rationalize the strong impact of this mutation. However, our structural data indicated that the interaction of M17 with F107 could compromise proper packing of RR468* with HK853 DHp α1, thereby precluding a productive interaction (Figure 2.10C). Similarly, substituting a methionine residue into the kinase disrupted phosphatase activity, as evidenced by the inability of HK853(AAM), HK853(VAM), and HK853(AGM) to dephosphorylate the majority of regulator mutants (Figure 2.10B). In the HK*-RR* structure, the side chain of M275 lies within the hydrophobic pocket generated by RR468* residues M17, F20 and V21 (Figure 2.10A). A comparison with HK-RR shows that the residues M275 and M17 would have clashed if the displacement of RR468* α1 had not generated the hydrophobic pocket in HK*-RR*. In the new interface, the polar bond between T275 and N21 is replaced by a hydrophobic interaction between M275 and M17. This arrangement is further facilitated by the A271G mutation in HK*, as the original alanine side chain would have clashed with M17 in RR*. Together, these findings illuminate the conformational restrictions imposed by individual residues that underlie amino acid coevolution, which in turn enabled the identification of specificity residues by statistical covariation analyses (Skerker et al., 2008).

Some combinations of residues yielded more promiscuous proteins. For instance, HK853(VGT) phosphorylated and dephosphorylated 10 of the 16 regulators, including both the wild-type regulator and the quadruple mutant (Figure 2.10B). By comparison, wild-type HK853 phosphorylates and dephosphorylates 8 of the regulators, including its partner RR468 but not the quadruple mutant. The flexibility introduced by glycine (A271G) may lead to productive interactions with many regulators. Conversely, some residue combinations interact with a very limited number of regulators. For instance, RR468(PLMV) interacted poorly with almost all kinase partners, and HK853(VAM) could phosphorylate and dephosphorylate only a single regulator.

Analysis of these mutational intermediates also demonstrated the interdependence of interface residues. The double mutant HK853(VAM) can phosphorylate all of the regulators to some extent, but its phosphatase activity is limited to RR468(PIIV) (Figure 2.10B). However, either the A268V or the T275M mutation in HK853 by itself allows the kinase to both phosphorylate and dephosphorylate eight or six of the regulators, respectively. As another example, the mutation A268V has little effect on the wild-type kinase and its phosphotransfer profile looks essentially identical to the starting protein (Figure 2.10B). However, when introduced in the context of HK853(AGT), this mutation affects interactions with several of the mutant regulators. Thus, the effect of individual substitutions on specificity is highly context-dependent and difficult to predict from the behavior of the individual mutations.

Mutational trajectories between the wild-type and mutant interfaces

Having a complete characterized set of mutational intermediates also shed light on the evolution of two-component signaling protein specificity following duplication and divergence. There are 5,040 possible mutational paths between two sequences that are 7 letters in length, assuming that only one amino acid is mutated at each step. Given our criteria for productive phosphotransfer and dephosphorylation (Figure 2.10C), we calculated that only 200 of these 5,040 possible paths, or 4%, retain functional kinase-regulator pairs along the entire path. Interestingly, two of the 43 kinase-substrate combinations that we deemed functional do not appear in any of the 200 mutational paths. For example, the pair HK853(AAM) and RR468(VILV) can be reached from a wild-type starting point, but all possible subsequent mutations produce a non-functional pair.

As noted above, the mutations T275M in the kinase and I17M in the regulator are often deleterious (Figure 2.10B). Consequently, these substitutions are often found along dead-end mutational paths. Of the 200 paths that maintain a productive interaction, 97 introduce T275M into the kinase as the final mutation, 64 paths introduce I17M into the regulator as the last mutation, and 29 paths end with the regulator mutation V13P. These patterns suggest that other mutations must be introduced first to effectively prime the interface for introduction of residues that restrict conformational freedom, such as methionine or proline. More generally, we infer that these residues contribute to specificity by eliminating interactions with non-cognate regulators, rather than promoting interaction with the cognate regulator. For instance, HK853(VGT) phosphorylates and dephosphorylates 10 of the 16 regulators; subsequent introduction of the T275M mutation eliminates interaction with 7 of these 10 regulators, but does not substantially improve interaction with the fully rewired, quadruple RR468 mutant (Figure 2.10B).

Although the majority of paths connecting the HK-RR and HK*-RR* pairs that maintain a functional interaction involve alternating mutations in the two proteins (Figure 2.10D), there are paths in which either the kinase or the regulator is completely changed through successive mutations while the other protein remains fixed. For instance, if the first three mutations produce RR468(PIIV) in any order, HK853 can then tolerate three successive mutations A268V followed by A271G/T275M in either order (Figure 2.10E). There are a total of 12 mutational paths in which the kinase accumulates all of its mutations in succession. Similarly, if the first two mutations in a path yield the double mutant HK853(VGT), then the regulator can accumulate four successive mutations, in seven different orders; there are 14 such paths, in which the regulator accumulates all mutations in succession. In both of these examples, the protein that remains fixed while the other protein accumulates substitutions is highly tolerant of mutations in its partner: RR468(PIIV) fully interacts with 5 out of 8 mutant kinases and HK853(VGT) fully interacts with 10 out of the 16 mutant regulators (Figure 2.10B). Such tolerant intermediates may play important roles in the rewiring of two-component signaling interfaces that occurs during evolutionary processes such as duplication and divergence.

Discussion

Despite the importance of protein-protein interaction specificity to the operation of cells, it remains relatively unclear how proteins use a finite set of amino acids to specifically recognize cognate partners. We addressed this question using two-component signaling proteins, which utilize a limited and known set of amino acids for partner recognition. The identification of these residues has guided the rational rewiring of two-component signaling pathways (Bell et al., 2010; Capra et al., 2010; Skerker et al., 2008), but a structural understanding of how rewiring is achieved was lacking. Here, we reprogrammed the structurally characterized complex *T. maritima* HK853-RR468 by introducing nine specificity residues from *E. coli* PhoR-PhoB. Although highly specific, phosphotransfer and dephoshorylation rates of PhoB-PhoR are slower than those of HK853-RR468. Strikingly, the introduction of PhoR-PhoB specificity residues did not impair the rapid reaction rates of HK853-RR468, despite the change in interaction specificity. This finding supports the notion that the specificity residues are required for recognition and proper positioning of the two partners, but other residues set the rates of the reactions (Pazy et al., 2009; Zapf et al., 1998).

The structure of the rewired complex HK*-RR* demonstrates that these functionally rewired two-component proteins preserve the overall structure of the wild-type complex along with catalytic activity. The kinase domains can reposition themselves slightly relative to each other and relative to the regulator to accommodate the foreign interfacial residues, but the overall complex retains wild-type character. Thus, rewiring leads primarily to local spatially restricted changes in the regions of each protein directly engaged in molecular docking.

Although the HK-RR and HK*-RR* complexes are similar, the HK-RR* complex harbors a completely different intermolecular orientation such that the phosphorylatable residues are no longer in close proximity. We do not yet have a structure of the HK*-RR complex, which can participate in slow phosphotransfer but not dephosphorylation. Because HK853* is competent as a phosphatase for other partners, the lack of activity with respect to RR is probably due to a mismatch at the interaction interface, which could force the regulator to dock in the wrong orientation. Alternatively, the regulator may bind at the correct position but too weakly to support rapid phosphotransfer and dephosphorylation.

How do individual residues contribute to specificity? In some cases, disruptive mutations that change the size or nature of a residue can be restored by balancing mutations at neighboring inter- or intra-molecular positions. This sort of intermolecular compensation is consistent with the extensive amino acid coevolution previously documented for two-component signaling proteins (Skerker et al., 2008). However, it is difficult to ascribe a role to individual specificity residues, as the effect of a given substitution can be highly context-dependent. In general, our systematic mutagenesis study indicated that individual residues do not typically contribute equally or additively to specificity. This interdependence of specificity residues resonates with other recent studies suggesting that amino acid epistasis in proteins is extensive and common (Breen et al., 2012; Levin et al., 2009; Ortlund et al., 2007).

Our work also has implications for understanding the evolutionary processes of duplication and divergence, which underlie the massive expansion of two-component signaling protein families in bacteria (Capra and Laub, 2012; Capra et al., 2012). Post duplication, paralogous signaling proteins must become insulated with respect to phosphotransfer while retaining an interaction with their cognate partners. Examples of the mutational trajectories that proteins follow, and any constraints they face, are largely unknown. Although *T. maritima* HK853-RR468 and *E. coli* PhoR-PhoB are not closely related, our systematic analysis of mutational trajectories between them provides insights into how specificity evolves. Our finding that only 4% of the theoretically possible evolutionary paths retain a functional interaction suggests that trajectories through sequence space may be severely constrained. Previous work on mutational trajectories that convert β-lactamase from a drug-sensitive to drug-resistant state also found strong constraints with only 8 of 120, or 6%, of paths permissible (Weinreich et al., 2006).

None of the mutational combinations we tested were fully insulated from both the starting HK-RR and the final HK*-RR* complex, *i.e.* none forms an orthogonal interacting pair. This finding highlights the potential evolutionary importance of promiscuous states of the binding interface (Aharoni et al., 2005; Matsumura and Ellington, 2001). We speculate that ancestral two-component proteins may have harbored such promiscuity; following duplication and divergence the paralogs could have gained specificity simply through the accumulation of mutations that disrupt a subset of interactions seen in the ancestral state. This process of sub-functionalization represents a rapid route to specificity and could help explain the apparent ease with which paralogous protein families have expanded.

Our results also have implications for protein design efforts. As noted, there are dependencies between neighboring residues on one molecule, and between residues on two different protein partners. Thus, the HK853-RR468 interface can tolerate certain substitutions only in combination – an effect not easily predicted based on a consideration of how the individual substitutions behave. Consequently, efforts to design or engineer novel protein-protein interfaces will have to tackle this combinatorial problem. Our findings also underline the importance of subtle backbone flexibility in protein design (Humphris and Kortemme, 2008; Smith and Kortemme, 2008), as a static complex would not properly accommodate the introduction of a new set of specificity residues. In sum, our studies have provided important new insights into the molecular and structural basis of two-component signaling specificity, but also highlight the significant challenges that remain in computationally predicting (Chen and Keating, 2012) the effects of mutations and in designing interfaces *de novo*.

Materials and Methods

Cloning, mutagenesis, and protein purification

All site-directed mutagenesis (see Table 2.5 for primers) was done with Gateway (Invitrogen) cloning vectors as described previously (Skerker et al., 2008). Mutagenized and sequence-verified protein sequences were moved from pENTR vectors into pDEST vectors using the Gateway LR reaction (pDEST-His$_6$-MBP for HK853 and PhoR; pDEST-His$_6$-TRX for PhoB; pDEST-His$_6$ for RR468). For crystallization assays, the complete cytoplasmic portion of HK853* (232-489) was recloned from pDEST-His$_6$-MBP into pET24b and full-length RR468* (1-122) was recloned from pDEST-His$_6$ into pET22b, using in both cases, the In-Fusion HD cloning technology (Clontech) (Table 2.6). Expression and purification was carried out as described previously (Casino et al., 2009; Skerker et al., 2005).

Phosphotransfer assays

Autophosphorylation and phosphotransfer assays were performed as described previously (Capra et al., 2010). Histidine kinases diluted to 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 $MgCl_2$ were autophosphorylated with 500 μM ATP and 0.5 μCi [$\gamma^{32}P$]-ATP (from a stock at ~6000 Ci/mmol, Perkin Elmer). PhoR was autophosphorylated for 1 hour at 30°C; HK853 and all HK853 mutants were autophosphorylated for 20 minutes at room temperature (see Figure 2.3B for autophosphorylation time course). The autophosphorylated kinase mixture was added directly to response regulator (at 5 μM in HKEDG buffer supplemented with 5 mM $MgCl_2$). Reactions were quenched with 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 reactions carried out at

4°C, the autophosphorylated histidine kinase was incubated at 4°C for 5 minutes prior to addition of chilled response regulator. Radiolabeled bands were quantified with ImageJ software.

Table 2.5 Mutagenesis primers

Primer	Sequence (5' → 3')
HK853(A268V, A271G, T275M)	ACGCCTTTAACGGTCATAAAAGGTTATGCGGAAATGATTTACAACAGT
HK853(V294T, D297E)	GGAGTTCCTCGAGACGATAATAGAGCAAAGCAACCACCTCG
HK853(A268V)	GCTCAGAACGCCTTTAACGGTCATAAAAGCTTATGC
HK853(A271G)	CGCCTTTAACGGCCATAAAAGGTTATGCGGAAAC
HK853(T275M)	GCCATAAAAGCTTATGCGGAAATGATTTACAACAGTCTGGGAG
HK853(M275T) on (A268V, A271G, T275M)	GGTTATGCGGAAACAATTTACAACAGTCTGGGAGAACTGG
HK853(V268A) on (A268V, A271G, T275M)	CAGAACGCCTTTAACGGCCATAAAAGGTTATGCG
HK853(G271A) on (A268V, A271G, T275M)	CCTTTAACGGTCATAAAAGCTTATGCGGAAATGATTTACAAC
HK853(L283G)	CAGTCTGGGAGAAGGGGATCTCAGCACCCTC
HK853(E290A)	GCACCCTCAAGGCGTTCCTCGAGGTG
HK853(E290A) on (V294T)	GCACCCTCAAGGCGTTCCTCGAGACG
HK853(S346Y)	GATCAAAGAATTTGCTTCATATCACAACGTGAATGTTCTCTTTG
HK853(H347G)	GATCAAAGAATTTGCTTCATCTGGCAACGTGAATGTTCTCTTTG
HK853(S279G)	GGAAACAATTTACAACGGTCTGGGAGAACTGGATC
HK853(S279L)	GCGGAAACAATTTACAACCTTCTGGGAGAACTGGATCTC
HK853(E290Y)	GGATCTCAGCACCCTCAAGTACTTCCTCGAGGTG
HK853(V294L)	CAAGGAGTTCCTCGAGCTGATAATAGATCAAAGCAAC
HK853(S279G) on (A268V, A271G, T275M, V294T, D297E)	GGAAATGATTTACAACGGTCTGGGAGAACTGGATC
HK853(S279L) on (A268V, A271G, T275M, V294T, D297E)	GCGGAAATGATTTACAACCTTCTGGGAGAACTGGATCTC
HK853(E290Y) on (A268V, A271G, T275M, V294T, D297E)	GGATCTCAGCACCCTCAAGTACTTCCTCGAGACG
HK853(V294L) on (A268V, A271G, T275M, V294T, D297E)	CTCAAGGAGTTCCTCGAGCTGATAATAGAGCAAAGCAAC
RR468(V13P, L14I)	GTTGATGACTCGGCACCTATCAGAAAAATCGTTTC
RR468(I17M, N21V) on (V13P, L14I)	GGCACCTATCAGAAAAATGGTTTCTTTCGTTCTGAAAAAAGAAGG
RR468(D53A)	CTGATAGTTCTCGCCATAATGATGCCCGTG
RR468(V13P)	CTCGTTGATGACTCGGCGCCTCTGAGAAAAATCGTTTC
RR468(L14I)	CTCGTTGATGACTCGGCGGTTATCAGAAAAATCGTTTCTTTC
RR468(I17M)	CTCGGCGGTTCTGAGAAAAATGGTTTCTTTCAATCTG
RR468(N21V)	CGGTTCTGAGAAAAATCGTTTCTTTCGTTCTGAAAAAAGAAGGTTACG
RR468(I17M) on RR468(N21V)	CTCGGCGGTTCTGAGAAAAATGGTTTCTTTCGTTCTG
RR468(V13P, I17M)	GTTGATGACTCGGCGCCTCTGAGAAAAATGGTTTCTTTCAATCTG
RR468(I17M) on (L14I)	GACTCGGCGGTTATCAGAAAAATGGTTTCTTTCAATCTG
RR468(V21N) on (V13P, L14I, I17M, N21V)	GCACCTATCAGAAAAATGGTTTCTTTCAATCTGAAAAAAGAAGGTTACG
RR468(I14L) on (V13P, L14I, I17M, N21V)	GTTGATGACTCGGCACCTCTGAGAAAAATGGTTTCTTTCG
RR468(P13V) on (V13P, L14I, I17M, N21V)	CTCGTTGATGACTCGGCAGTTATCAGAAAAATGGTTTCTTTCG
RR468(D60E)	GATGCCCGTGATGGAAGGATTCACCGTG
RR468(D60G)	GATGCCCGTGATGGGTGGATTCACCGTG
RR468(K16A)	GACTCGGCGGTTCTGAGAGCAATCGTTTCTTTCAATCTG
RR468(K24A)	CGTTTCTTTCAATCTGAAAGCAGAAGGTTACGAAGTGATAGAAGC
RR468(K16A) on (V13P, L14I, I17M, N21V)	GACTCGGCACCTATCAGAGCAATGGTTTCTTTCGTTCTG
RR468(K24A) on (V13P, L14I, I17M, N21V)	GGTTTCTTTCGTTCTGAAAGCAGAAGGTTACGAAGTGATAGAAGC
PhoB(M17I, V21N)	GCTCCAATTCGCGAAATCGTCTGCTTCAACCTCGAACAAAATGGCTTTC
PhoB(P13V, I14L) on (M17I, V21N)	GGTCGTAGAAGATGAAGCTGTACTTCGCGAAATCGTCTGCTTC

Table 2.6 Cloning primers

Primer	Sequence (5' → 3')
HK853-pET24b (insert)-FW -RV	AGGAGATATACCATGGAAAATGTGACAGAATCAAAAGA CTTTTGGGATCCACACAAAGA
HK853-pET24b (vector)-FW -RV	TGTGGATCCCAAAAGACCGT CATGGTATATCTCCTTCTTAAAG
RR468-pET22b (insert)-FW -RV	GAAGGAGATATACATATGTCTAAAAAAGTTCTTCTCGTT GTGGTGGTGCTCGAGTCATTCATTTAATAGATGCTTCAC
RR468-pET22b (vector)-FW -RV	CTCGAGCACCACCACCAC ATGTATATCTCCTTCTTAAAGTTA

Phosphatase assays

To perform the phosphatase assays in Figure 2.3C, [^{32}P] acetyl-phosphate was freshly synthesized as described previously (Jagadeesan et al., 2009). RR468 (at 10 μM in HKEDG buffer supplemented with 5 mM $MgCl_2$) was mixed 1:1 with [^{32}P] acetyl-phosphate and incubated for 1 hour at room temperature. The mixture was washed three times with cold HKEDG buffer and the concentration of $MgCl_2$ subsequently adjusted to 5 mM. The phosphorylated regulator was chilled at 4°C for 5 minutes and then incubated with buffer or with an equimolar amount of kinase (both pre-chilled) for the times indicated. Because RR468* autophosphorylates more poorly than RR468 using this method, both regulator and kinase concentrations were doubled to 20 μM starting concentration to measure RR468*~P dephosphorylation.

Crystallization, data collection, and model building

Crystallization of HKf* and RRf* proteins, HK*-RR* and HK-RR* complexes was achieved by the vapor diffusion method, using the sitting drop technique, mixing 0.6 μL of protein and 0.6 μL of reservoir solution. Crystals of HKf* were obtained in 8% PEG4000, 0.8 M LiCl and Tris pH 8.5 by mixing 10 mg/mL of protein, 4 mM ADP and 4 mM $MgCl_2$. Crystals of RRf* in complex with BeF_3^- were obtained in 50% PEG400, NaAc pH 4.6 and 0.2 M $Li_2(SO_4)$ mixing 15 mg/mL of protein, 30 mM NaF, 5 mM $BeSO_4$ and 7 mM $MgCl_2$. Crystals of the complexes HK*-RR* and HK-RR* were obtained in 2.2 M $(NH_4)_2SO_4$ and Bis-Tris pH 5.5 by cocrystallization mixing 10 mg/mL of HK853* or HK853, 7.5 mg/mL of RR468*, 4 mM ADP, 30 mM NaF, 5 mM $BeSO_4$ and 7 mM $MgCl_2$. Crystals of HKf* were cryoprotected by increasing PEG4000 to 16% and by addition of 20% sucrose while crystals of the complexes were cryoprotected by addition of 35% sucrose. X-ray diffraction data was collected at Diamond Light Source I04-1 (Oxfordshire, UK) for HKf*, at European Synchrotron Facility ID23-2 (ESRF, Grenoble, France) for RRf*, and at ESRF ID23-1 and ID23-2 for the complexes HK*-RR* and HK-RR*, respectively. Data reduction was performed using XDS, Pointless and Scala to a Bragg space of 2.7 Å for HKf*, 1.8 Å for RR*, 3.0 Å for HK*-RR* complex and 3.1 Å for HK-RR* complex. Phases were obtained by molecular replacement using Phaser and the final models were obtained by subsequent cycles of refinement with Refmac5 and model building with the program Coot (Emsley and Cowtan, 2004). Despite the limited resolution for the complexes, the quality of the maps (Figure 2.11) allowed model building and unambiguous side chain assignments except for the first and last residues in HK* (232 to 243; 481 to 490) and HK (chain A 232 to 233; 480 to 490 and chain B 232 to 235; 480 to 490) where electronic density was absent, which reflects the elevated flexibility of these regions. Crystallographic data and refinement statistics are in Table 2.1. The programs Pointless, Scala, Phaser and Refmac are contained in CCP4 Suite. Figures 2.4-2.6, 2.8, 2.9 and 2.11 were produced using PyMOL (http://www.pymol.org). Superimpositions were carried out with Superpose from CCP4 Suite. Movement analysis was performed using the program Dyndom (Lee et al., 2003). In the HK-RR and HK-RR* complexes, where the asymmetric unit is formed by HK_2-2RR, HK chain A and RR chain C were chosen for the structural comparisons. The 3D structures are deposited with the following PDB accession codes: 4JAU for HKf*, 4JA2 for RRf*, 4JAS for HK*-RR* and 4JAV for HK-RR*. PDB accession codes for additional structures analyzed in the manuscript were 2C2A for HKf and 3DGE for HK-RR.

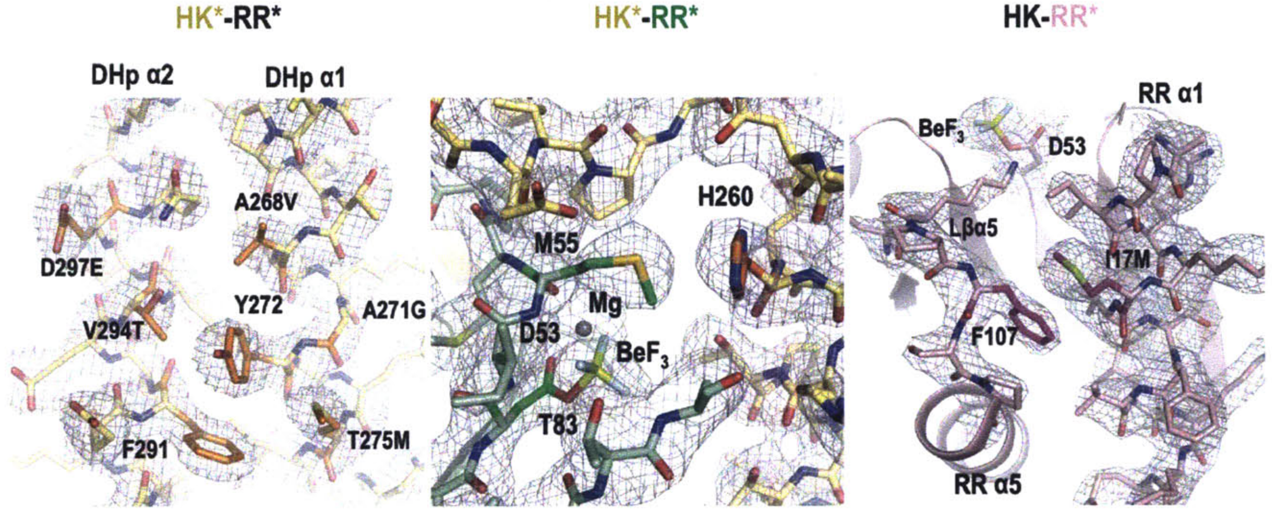


Figure 2.11 Quality of the electron density maps.

The electron density from $2F_o$-F_c maps contoured at 1.0 σ are shown around the complex interface (left), the active center of the HK*-RR* complex (center) and the RR468* contact interface of the HK-RR* complex (right). Notice that the interface views (left and right) are shown in the same orientation as in Figure 2.9. More relevant side chains and structural elements are labeled and colored in a darker hue.

Acknowledgements

We thank A. Keating and members of the Laub laboratory for helpful comments on the manuscript. We acknowledge the European Synchrotron Radiation Facility (ESRF) and Diamond Light Source (DLS) for provision of synchrotron radiation facilities and we would like to thank ID23-1, ID23-2 (ESRF) and I04-1 (DLS) beamlines staff for assistance during data collection. M.T.L is an Early Career Scientist at the Howard Hughes Medical Institute. This work was supported by an NSF CAREER award (MCB-0844442) to M.T.L.; grants BIO2010-15424 from the Ministerio de Ciencia e Innovación and ACOMP2012/044 from Generalitat Valenciana to A.M.; and NSF Graduate Research Fellowship to A.P. P.C. is recipient of a postdoctoral junior grant “Juan de la Cierva” from the Ministry of Economy and Competitiveness.

References

Aharoni, A., Gaidukov, L., Khersonsky, O., McQ Gould, S., Roodveldt, C., and Tawfik, D.S. (2005). The 'evolvability' of promiscuous protein functions. Nat Genet *37*, 73-76.

Albanesi, D., Martín, M., Trajtenberg, F., Mansilla, M.C., Haouz, A., Alzari, P.M., de Mendoza, D., and Buschiazzo, A. (2009). Structural plasticity and catalysis regulation of a thermosensor histidine kinase. Proc Natl Acad Sci U S A *106*, 16185-16190.

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.

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.

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.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535-538.

Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems. Annu Rev Microbiol *66*, 325-347.

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.

Chen, T.S., and Keating, A.E. (2012). Designing specific protein-protein interactions using computation, experimental library screening, or integrated methods. Protein Sci *21*, 949-963.

Codoner, F., and Fares, M. (2008). Why should we care about molecular coevolution? Evol Bioinform Online *4*, 29-38.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126-2132.

Fisher, S.L., Kim, S.K., Wanner, B.L., and Walsh, C.T. (1996). Kinetic comparison of the specificity of the vancomycin resistance kinase VanS for 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.

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

Hsing, W., and Silhavy, T.J. (1997). Function of conserved histidine-243 in phosphatase activity of EnvZ, the sensor for porin osmoregulation in Escherichia coli. J Bacteriol *179*, 3729-3735.

Humphris, E.L., and Kortemme, T. (2008). Prediction of protein-protein interface sequence diversity using flexible backbone computational protein design. Structure *16*, 1777-1788.

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

Igo, M.M., Ninfa, A.J., Stock, J.B., and Silhavy, T.J. (1989). Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. Genes Dev *3*, 1725-1734.

Jagadeesan, S., Mann, P., Schink, C.W., and Higgs, P.I. (2009). A novel "four-component" two-component signal transduction mechanism regulates developmental progression in Myxococcus xanthus. J Biol Chem *284*, 21435-21445.

Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008). Principles of protein-protein interactions: what are the preferred ways for proteins to interact? Chem Rev *108*, 1225-1244.

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

Lee, R.A., Razaz, M., and Hayward, S. (2003). The DynDom database of protein domain motions. Bioinformatics *19*, 1290-1291.

Levin, K.B., Dym, O., Albeck, S., Magdassi, S., Keeble, A.H., Kleanthous, C., and Tawfik, D.S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. Nature Struct & Mol Biol *16*, 1049-1055.

Matsumura, I., and Ellington, A.D. (2001). In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. J Mol Biol *305*, 331-339.

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.

Pazy, Y., Wollish, A.C., Thomas, S.A., Miller, P.J., Collins, E.J., Bourret, R.B., and Silversmith, R.E. (2009). Matching biochemical reaction kinetics to the timescales of life: structural determinants that influence the autodephosphorylation rate of response regulator proteins. J Mol Biol *392*, 1205-1220.

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.

Smith, C.A., and Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol *380*, 742-756.

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

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.

Zapf, J., Madhusudan, M., Grimshaw, C.E., Hoch, J.A., Varughese, K.I., and Whiteley, J.M. (1998). A source of response regulator autophosphatase activity: the critical role of a residue adjacent to the Spo0F autophosphorylation active site. Biochemistry *37*, 7725-7732.

Chapter 3

Pervasive degeneracy and epistasis in a protein-protein interface

This work has been submitted for publication.

AIP and MTL conceived the approach, analyzed the data, and wrote the paper. AIP performed all experiments and computational analyses.

Abstract

Comprehensive mapping of protein sequence space is a long-standing problem. Identification of all sequences of length N that support a given biochemical function necessitates the analysis of 20^N combinations, often a technical challenge for $N > 2$. Here, we leverage next-generation sequencing to systematically map the sequence space critical to the protein-protein interface formed by the *Escherichia coli* protein kinase PhoQ and its cognate substrate PhoP. As this interaction relies primarily on just four residues in each protein, we generated a library containing all 160,000 (20^4) variants of PhoQ at these positions and used a two-step selection and deep-sequencing process to identify 1,659 fully functional variants. Our results reveal a striking degree of degeneracy in the PhoQ-PhoP interface. Additionally, we find that individual positions in PhoQ frequently exhibit strong interdependencies such that the functionality of most mutants cannot be easily predicted from the behavior of single mutants. This epistasis, along with a highly non-uniform density of functional variants across sequence space, appears to strongly constrain PhoQ evolution. Consequently, the natural diversity of PhoQ orthologs pales in comparison to the full diversity of functional PhoQ variants. Our findings have wide-ranging implications in understanding protein function, evolution, and design.

Protein-protein interactions drive the operation and functions of cells and organisms. These interactions involve the formation of a molecular interface comprising a subset of amino acids from each protein. For a given protein with N interfacial residues, there are 20^N possible variants. In addition to the wild-type residues, how many combinations of N residues support a functional interaction, and how are these combinations distributed and connected in sequence space (Smith, 1970)? Interfacial residues often vary between orthologs indicating some ability to accommodate change (DePristo et al., 2005; Harms and Thornton, 2013), but this natural variability may not capture the full plasticity of a given interface. Mutagenesis studies can help to assess the tolerance of an interface to substitutions, but the total number of mutants that can be tested typically limits such efforts. Saturation mutagenesis of single sites has become routine, but does not test combinations of mutations at different sites (Fowler et al., 2010; Hietpas et al., 2011; McLaughlin et al., 2012; Whitehead et al., 2012). Even studies examining the mutational intermediates separating orthologs, or separating ancestral and evolved proteins, typically exclude residues not present in either protein (Weinreich et al., 2006). Here, we take advantage of fully randomized gene synthesis, an *in vivo* screen, and deep sequencing to comprehensively map the sequence space underlying a protein-protein interaction operating in its native context.

As a model system, we examined the interface formed by two-component signal transduction proteins, the predominant form of signaling in bacteria (Casino et al., 2009; Stock et al., 2000). In particular, we focused on *E. coli* PhoQ, a sensor histidine kinase that autophosphorylates upon extracellular magnesium limitation and subsequently transfer its phosphoryl group to the response regulator PhoP, which activates gene expression (Groisman, 2001). When not stimulated to autophosphorylate, PhoQ binds to and stimulates the dephosphorylation of PhoP.

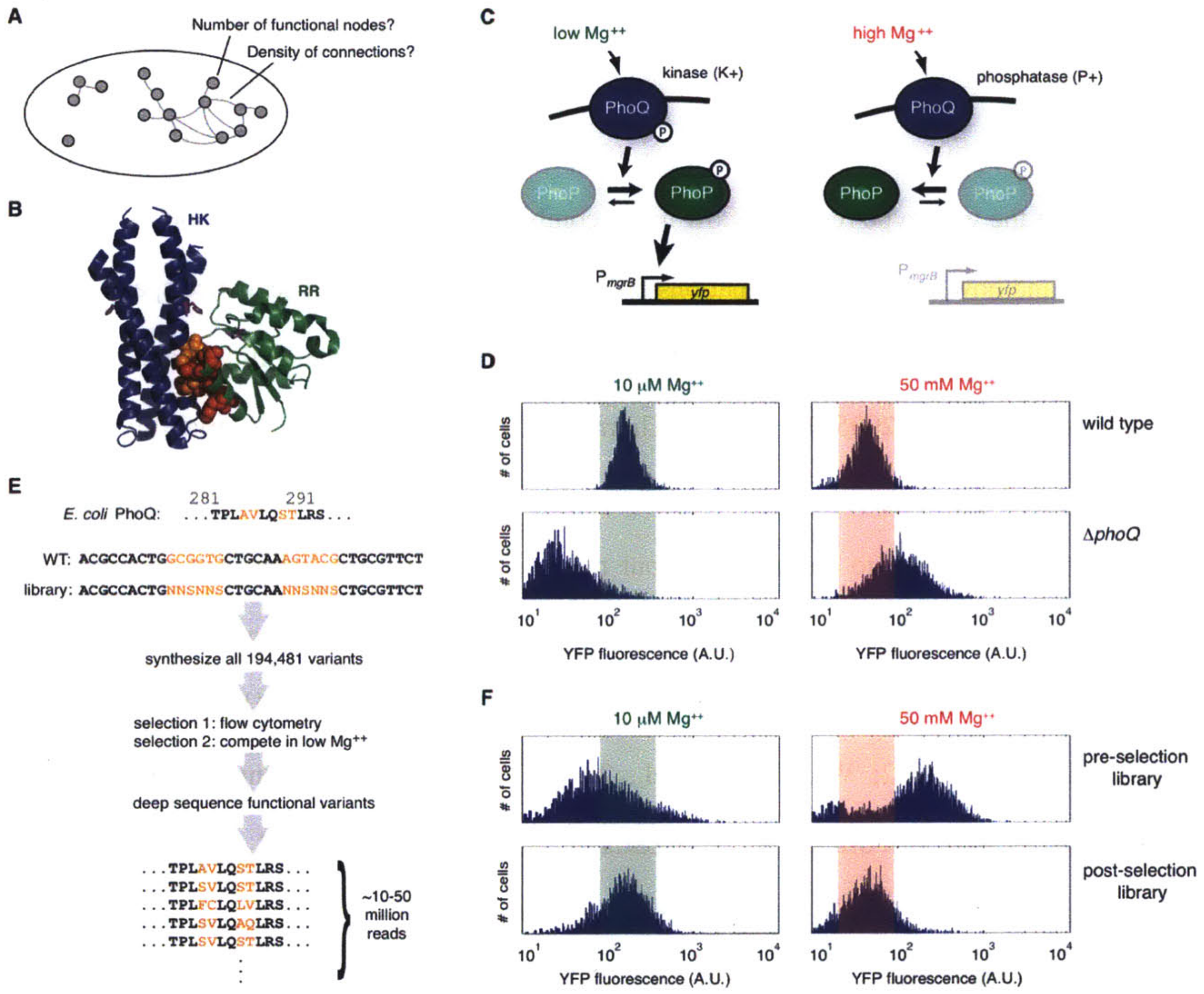


Figure 3.1 Mapping a protein sequence space.

(A) A hypothetical model of protein sequence space with nodes representing functional variants and lines connecting variants differing by one residue. (B) The interaction interface between a *Thermotoga maritima* histidine kinase (HK853) and its cognate response regulator (RR468) from PDB:3DGE. Specificity-determining, interfacial residues are space-filled in orange (kinase) or red (regulator). (C) PhoQ is a bifunctional histidine kinase that phosphorylates or dephosphorylates PhoP depending on extracellular magnesium concentration. (D) YFP levels measured by flow cytometry for ~5,000 cells from strains expressing wild-type *phoQ* from a low-copy plasmid or an isogenic strain lacking *phoQ*. Shaded regions indicate wild-type YFP levels. (E) Experimental pipeline for mapping PhoQ sequence space. (F) YFP levels in cells harboring the *phoQ* library on a low-copy plasmid, pre- and post-selection.

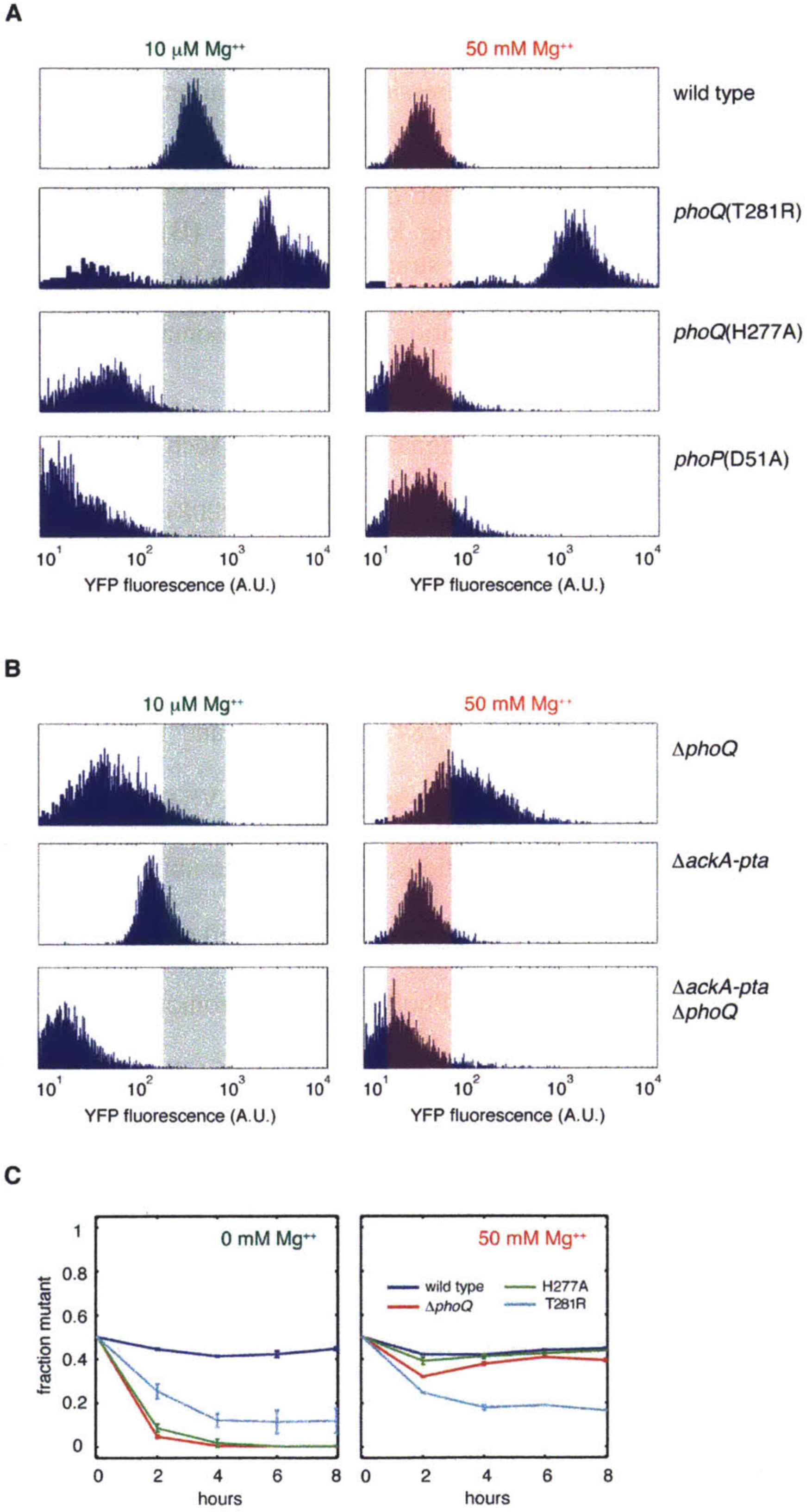


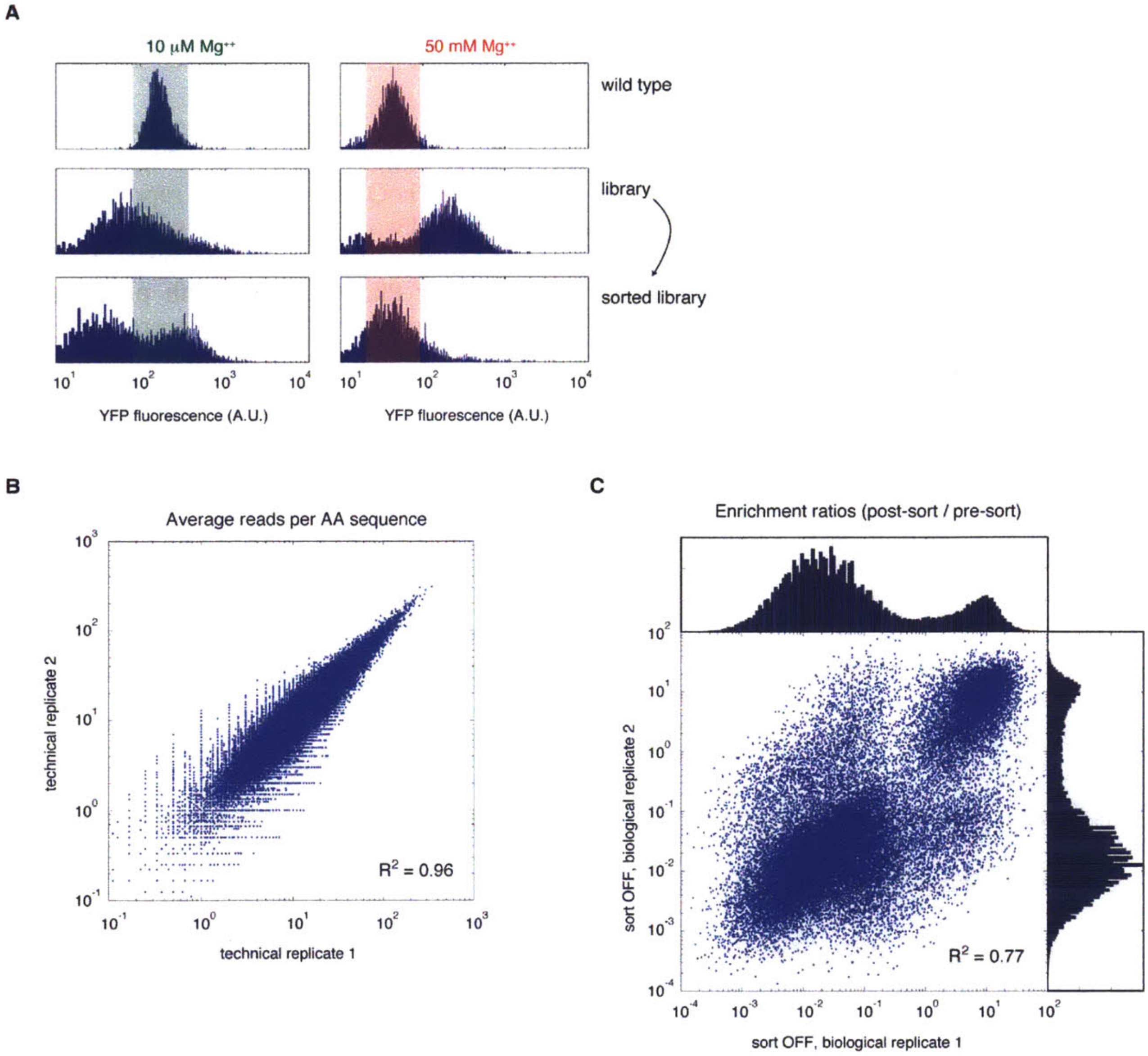
Figure legend on next page.

Figure 3.2 The behavior of kinase- and phosphatase- mutants of PhoQ.

(A-B) Flow cytometry profiles of strains expressing the *phoQ* variant indicated following growth in 10 μM or 50 mM Mg^{++}. Shaded regions indicate YFP levels for a strain producing wild-type PhoQ. The mutations H277A and D51A target the catalytic histidine and aspartate in PhoQ and PhoP, respectively, and prevent YFP induction in 10 mM Mg^{++}. The mutation T281R in PhoQ inhibits phosphatase activity while retaining kinase activity. (B) The Δ*phoQ* strain shows elevated levels of YFP relative to the H277A strain because acetyl phosphate (produced by *ackA-pta* gene products) can drive PhoP phosphorylation. (C) Head-to-head competitions of wild-type *E. coli* against strains with the indicated mutation in the chromosomally-encoded PhoQ.

The interface formed by two-component signaling proteins, such as PhoQ-PhoP, involves a limited number of residues from each protein (Skerker et al., 2008) (Figure 3.1B). For histidine kinases, mutating just three or four interfacial residues to match those found in another kinase is often sufficient to reprogram partner specificity (Capra et al., 2010; Podgornaia et al., 2013).

For PhoQ these key interfacial residues are A284, V285, S288, and T289, referred to hereafter as AVST for simplicity (Figure 3.1E). Given 20 amino acids, AVST is one of 160,000 possible mutational combinations at these four positions. To assess the ability of each combination to promote a functional PhoQ-PhoP interface *in vivo*, we developed a high-throughput screen using a strain in which *yfp* is expressed from a PhoP-dependent promoter, P_{mgrB} (Figures 3.1C-D). Given low extracellular Mg^{++}, PhoQ is predominantly a kinase, driving PhoP phosphorylation and YFP production; in high extracellular Mg^{++}, PhoQ is mainly a phosphatase, stimulating PhoP dephosphorylation and preventing YFP production (Groisman, 2001; Miyashiro and Goulian, 2007). Point mutants of PhoQ that selectively ablate kinase or phosphatase activity yielded the expected decrease or increase, respectively, in YFP levels (Figure 3.2A), and cells lacking *phoQ* fail to properly activate or repress YFP production in low (10 μM) or high (50 mM) Mg^{++} conditions, respectively. To systematically probe the PhoQ interface, we constructed a library in which the four key interfacial residues were fully randomized and transformed this library into a



D

	Number of quality-filtered reads	Number of DNA sequences with > 5 quality reads	Number of AA sequences with > 5 quality reads*
Initial library (tech replicate 1)	14,533,541	740,493	140,517
Initial library (tech replicate 2)	14,062,861	728,248	139,513
Sorted off at 50 mM (bio replicate 1)	38,225,614	85,315	20,519
Sorted off at 50 mM (bio replicate 2)	46,400,504	91,155	21,955

* at least one underlying DNA sequence has > 5 reads. Does not include stop codon-containing sequences.

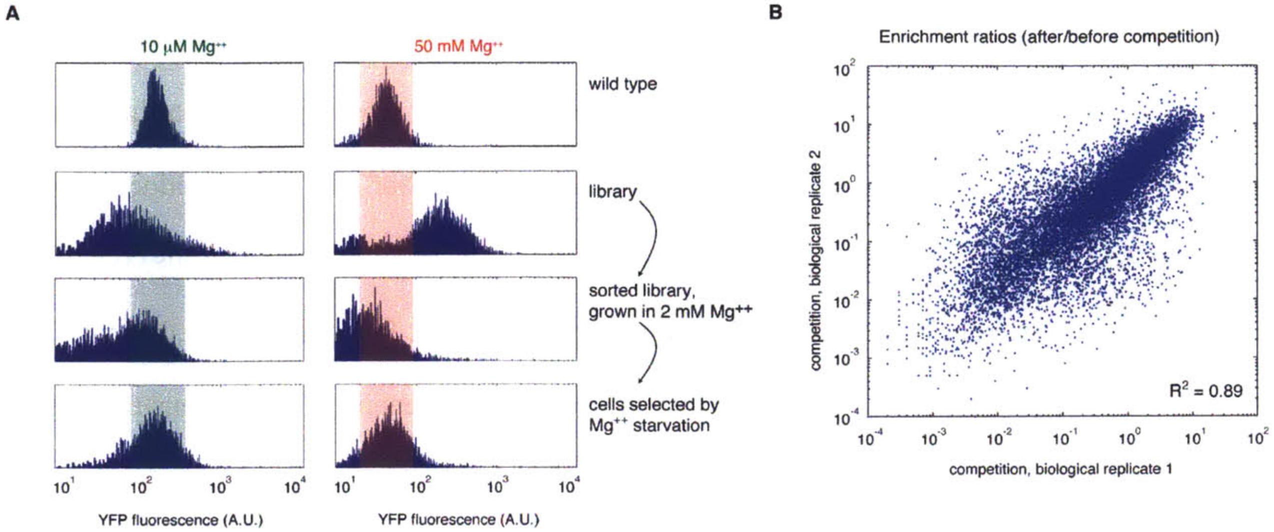
Figure legend on next page.

Figure 3.3 FACS-based screen for PhoQ variants with phosphatase activity comparable to wild-type PhoQ.

(A) Flow cytometry profiles for the wild-type control, initial library, and after FACS-based selection for cells producing YFP similar to the wild-type control in 50 mM Mg^{++}. Shaded regions indicate YFP levels for a strain producing wild-type PhoQ. (B) Technical reproducibility of sequencing-based assessment of PhoQ variant abundance. Two aliquots of cells harboring the PhoQ library were processed separately for sequencing. R^2 is the Pearson correlation coefficient. (C) Biological reproducibility of FACS-based selection for PhoQ variants with phosphatase activity comparable to wild-type PhoQ. Two independent aliquots of the library were subjected to FACS and sequencing. R^2 is Spearman correlation. (D) Numbers of quality-filtered deep sequencing reads for each sample.

ΔphoQ strain harboring the P_{mgrB}-*yfp* reporter. The library was grown for 6 hours in medium with low or high extracellular Mg^{++} to stimulate PhoQ kinase or phosphatase activity, respectively, and then subjected to fluorescence-activated cell sorting (FACS) to isolate those mutants that behaved similarly to a strain with the wild-type PhoQ.

This screen proved more stringent in selecting mutants with wild-type phosphatase activity because cells deficient in PhoQ kinase activity still accumulate some phosphorylated PhoP, using acetyl-phosphate as a phosphodonor (Figure 3.2B). We therefore performed a second screen that required cells to exhibit robust kinase activity: cells selected for phosphatase activity (Figure 3.3) were starved of extracellular Mg^{++} for 18 hours and then recovered in Mg^{++} replete medium for 6 hours. This competitive growth scheme selects for PhoQ mutants with wild-type kinase activity because induction of the PhoP regulon is necessary for cells to survive without Mg^{++} (Figure 3.2C). Flow-cytometry analysis demonstrated that our selection process strongly enriched for mutants able to phosphorylate PhoP in low Mg^{++} conditions, and two biological replicates exhibited high reproducibility (Figures 3.1F and 3.4).



C

	Number of quality-filtered reads	Number of DNA sequences with > 5 quality reads	Number of AA sequences with > 5 quality reads*
Pre-competition (bio replicate 1)	23,288,830	70,236	18,588
Pre-competition (bio replicate 2)	18,985,124	67,698	17,743
Post-competition (bio replicate 1)	20,486,707	44,660	12,974
Post-competition (bio replicate 2)	22,010,298	50,920	14,565

* at least one underlying DNA sequence has > 5 reads. Does not include STOP codon-containing sequences.

D

	Total number of cells	
	Biological replicate 1	Biological replicate 2
1. Growth from frozen	793,000,000	696,000,000
2. Overnight starvation	215,000,000	229,000,000
3. Recovery to log phase	8,400,000	24,400,000

* Cells count using CFU platings at start of indicated time period.

Figure 3.4 Competitive growth-based screen for PhoQ variants with kinase activity comparable to wild-type PhoQ.

(A) Flow cytometry profiles for the wild-type control, initial library, library after sorting in 50 mM Mg^{++} (see Fig. S2), and after overnight Mg^{++} starvation. Shaded regions corresponds to the position of the wild type YFP levels. (B) Biological reproducibility of sorting approach. Two independent aliquots of the sorted library were subjected to Mg^{++} starvation and sequenced. R^2 is Spearman correlation. (C) Numbers of quality-filtered deep sequencing reads for each sample. (D) Total number of cells present at each step of the experiment, tabulated to ensure oversampling of the ~1x10^6 possible library variants.

To identify the interfacial residues that promote a PhoQ-PhoP interaction, we deep-sequenced the relevant region of *phoQ* from cells that passed our two-step selection. The starting library used NNS codons (N=A, C, G, or T; S=C or T) to randomize the four interfacial residues. Hence, the theoretical diversity of the library is 194,481 variants, with 160,000 combinations containing no stop codons. Sequencing of the starting library indicated that 93% of the 160,000 possible protein variants were present in >3 reads, and independent sequencing runs demonstrated high technical reproducibility (R^2 = 0.96) (Figure 3.3). For the post-selection library, we sought sequences enriched relative to the starting library, using an unbiased binary classifier to optimize inclusion of true positives while minimizing the number of false negatives (Figures 3.3 and 3.4). The training set for the classifier consisted of library data for (i) the 34,481 PhoQ variants harboring one or more stop codons, each of which would produce a truncated, non-functional PhoQ, and (ii) data collected individually for each possible single mutant of PhoQ at the first three positions randomized in our library. Of these 57 single point mutants, 13 exhibited nearly wild-type flow cytometry profiles and could successfully compete with the wild type under conditions of Mg^{++} starvation (Figures 3.5 and 3.6A). The binary classification tree identified 1,659 unique, functional PhoQ variants in our library with an estimated false positive rate <1%.

To validate the functionality of the novel PhoQ variants identified, we isolated and tested 20 individual mutants. Flow-cytometry analysis demonstrated that each mutant enabled PhoP-dependent gene expression to approximately wild-type levels in the presence of low Mg^{++} and each supported the suppression of PhoP activity in high Mg^{++}, indicating that these PhoQ variants harbored both kinase and phosphatase activity (Figure 3.6B). We purified four of these variants, harboring specificity residues ALAV, SLSS, SVAQ, and SVSY, and confirmed that

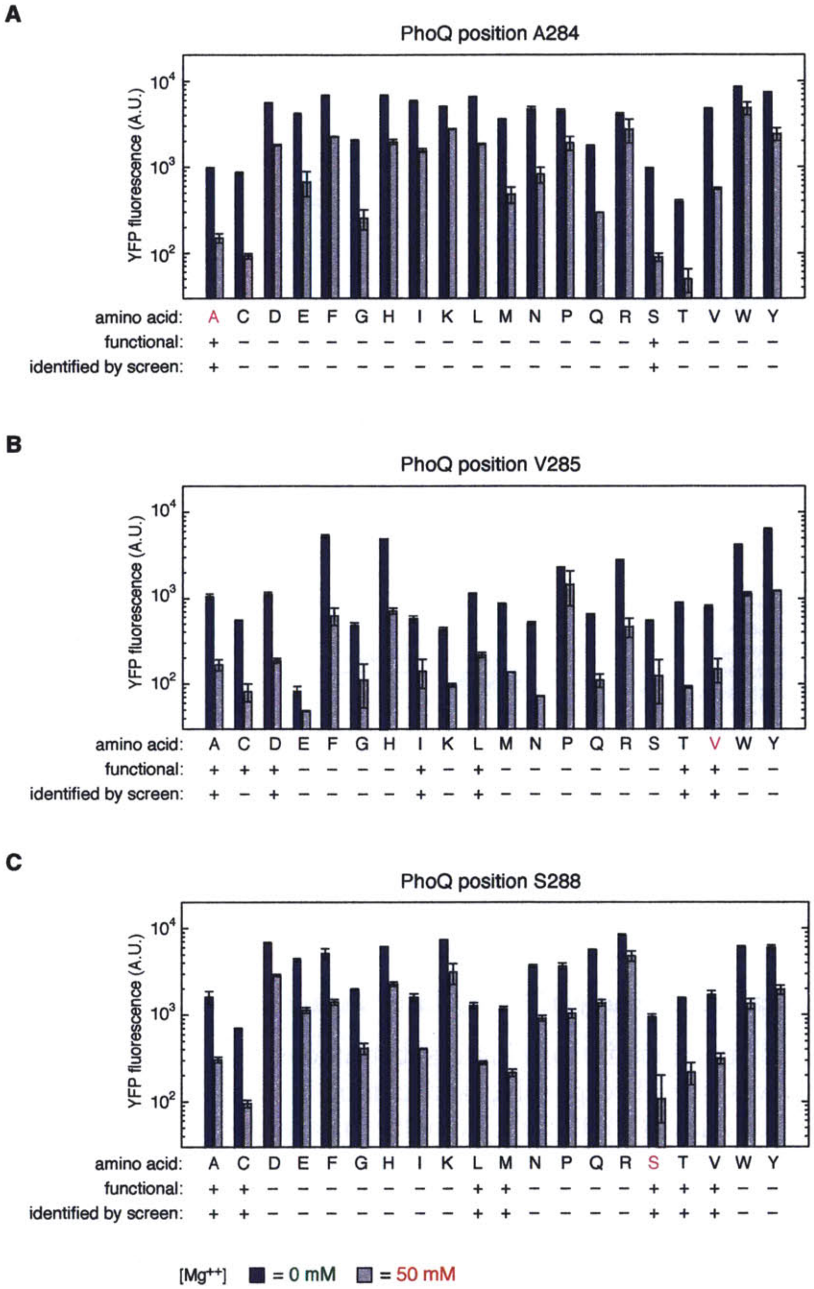
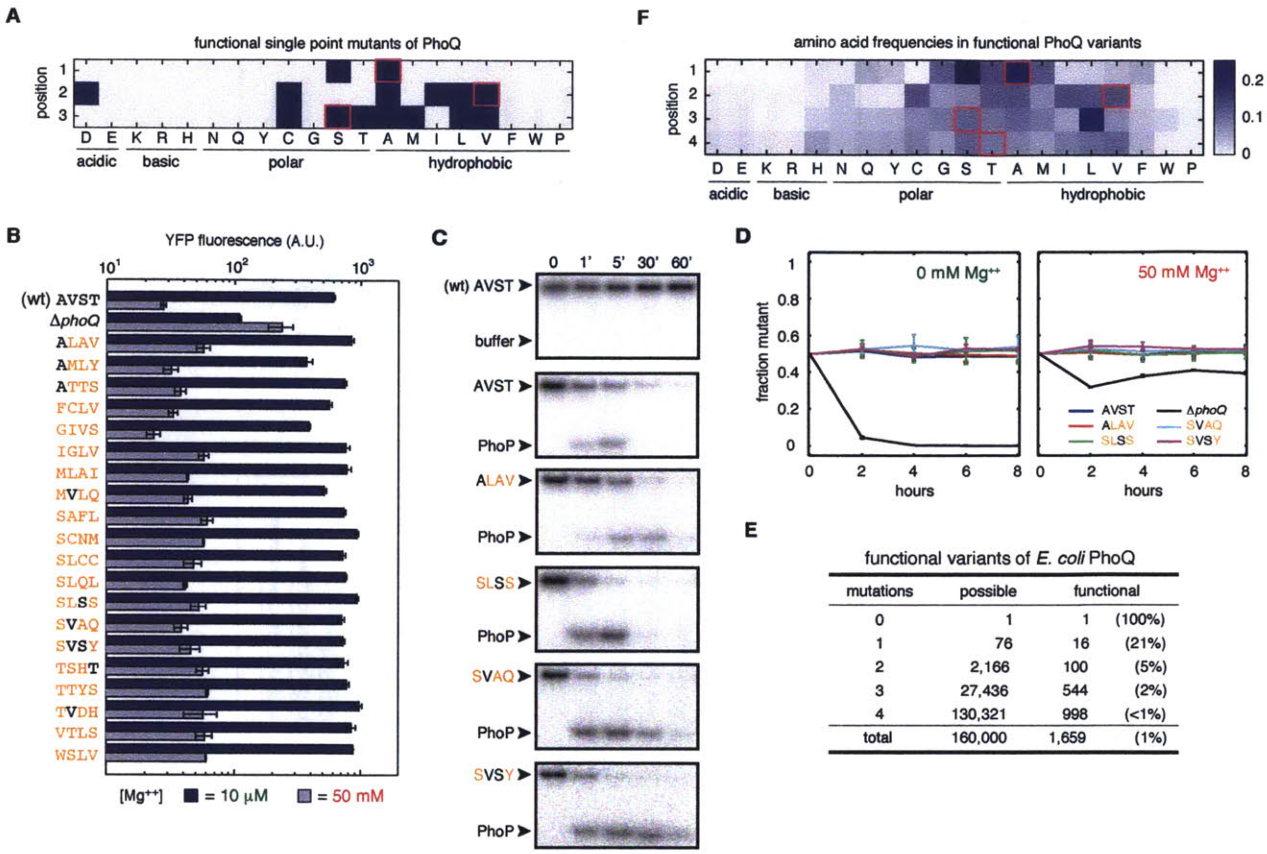


Figure 3.5 Behavior of individual PhoQ point mutations.

Mean YFP value calculated from flow cytometry measurements of strains harboring wild-type PhoQ or each of 19 possible amino acid substitutions at positions 284 (A), 285 (B), and 288 (C) in PhoQ. Wild-type residues are A284, V285, and S288 (shown in magenta). Error bars show ± SD for two biological replicates. Pluses and minuses indicate (middle row below each graph) whether each variant was signal responsive, as judged by flow cytometry analysis, and fit, as judged through competitions with a wild-type control strain, and (bottom row below each graph) whether each variant was identified in the two-step selection process performed on the entire library of PhoQ variants.



functional variants of *E. coli* PhoQ

mutations	possible	functional	
0	1	1	(100%)
1	76	16	(21%)
2	2,166	100	(5%)
3	27,436	544	(2%)
4	130,321	998	(<1%)
total	160,000	1,659	(1%)

Figure 3.6 Functional degeneracy of PhoQ interfacial residues.

(A) Summary of the functionality of single substitutions in PhoQ assessed individually. Blue indicates a variant was signal-responsive and competitive in Mg^{++} starvation; magenta boxes indicate wild-type residues. (B) Flow cytometry measurements of YFP levels for 20 PhoQ variants in 10 μM and 50 mM Mg^{++}. Error bars indicate SD for two replicates. (C) The PhoQ variants indicated were autophoshorylated *in vitro* and tested for phosphotransfer to and dephosphorylation of PhoP. (D) Head-to-head competitions of wild-type *E. coli* against strains expressing the indicated PhoQ variants from the chromosome. Δ*phoQ* is shown for comparison (also see Figure 3.2). (E) Summary of functional PhoQ variants identified. (F) Heatmap indicating the frequency of each amino acid at each position in the 1,659 functional PhoQ variants; magenta boxes indicate wild-type residues.

each exhibited kinase and phosphatase activity *in vitro*, as manifest by the robust transfer of phosphate from autophosphorylated PhoQ to PhoP, and then subsequent PhoP dephosphorylation (Figure 3.6C). We also introduced these sets of interfacial residues onto the chromosome of a wild-type strain and confirmed that these mutants can compete with wild-type when co-cultured in low or high magnesium conditions (Figure 3.6D).

The identification of 1,659 PhoQ variants that are signal-responsive and competitively fit in our two-step selection, hereafter referred to as functional variants, indicates extensive degeneracy or plasticity of the PhoQ-PhoP interfacial residues. In addition to 16 single point mutants, there were 100 double, 544 triple, and 998 quadruple mutants, indicating that many combinations of residues, some of which differ significantly from wild type PhoQ, can support a functional interaction with PhoP (Figure 3.6E).

Tabulating amino acid frequencies in the set of 1,659 functional variants showed enrichment of hydrophobic and small polar amino acids at each position (Figure 3.6F). Not surprisingly, proline was the least common residue at each position, as it would break the alpha-helix critical to PhoQ-PhoP interactions. Most other bulky and charged residues also appeared at low frequencies, indicating they can be tolerated, but only in certain contexts. For example, the mutation A284H (yielding variant HVST) abolished the phosphatase activity of PhoQ both *in vivo* and *in vitro*, indicating a significant disruption of the PhoQ-PhoP interaction (Figures 3.7A and B). However, the quadruple mutant HSLV was fully functional, indicating that the A284H substitution can be tolerated in this particular context. Similarly, we found that the substitution S288Q alone (variant AVQT) was non-functional, but permissible in variants also harboring the A284S substitution, such as the quadruple mutant SIQC (Figures 3.7A-C).

Conversely, there are substitutions that were individually permissible, but not functional in combination. For instance, the single substitutions A284S, V285T, S288L, and T289V each support a functional PhoQ-PhoP interface, but are non-functional when combined (STLV)

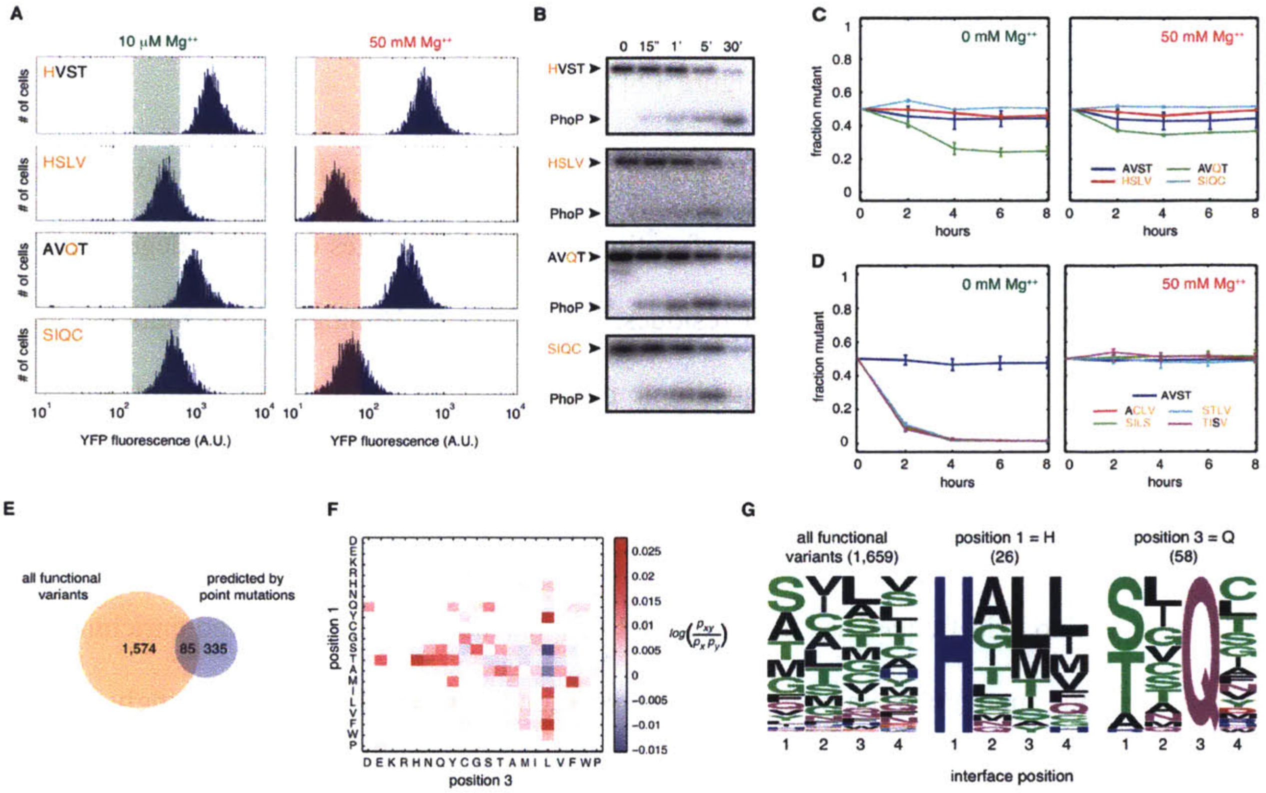


Figure 3.7 Epistasis and interdependencies between PhoQ interfacial residues.

(A) Flow cytometry measurements of YFP for the PhoQ variants indicated. Shaded regions indicate wild-type YFP levels. (B) *In vitro* analysis of PhoP phosphorylation and dephosphorylation for variants from (A). (C-D) Head-to-head competitions of wild-type *E. coli* against strains producing the indicated PhoQ variants from the chromosome. (E) Venn diagram comparing the number of functional PhoQ variants identified with the number predicted from single mutant analysis, assuming position independence. (F) Heatmap showing frequency of all residue pairs at positions 1 and 3 relative to frequency expected if residues occurred independently (for other positions, see Extended Data Fig. 5). (G) Sequence logos indicating frequencies of amino acids at each position in the sets of PhoQ indicated. Residues are colored according to chemical properties.

(Figure 3.7D). Similarly, we found that the combinations ACLV, TISV, and SILS, each involving residues found at high frequency in the set of 1,659 functional variants, were severely impaired in competition against wild-type PhoQ (Figures 3.7D and 3.8B). Collectively, our

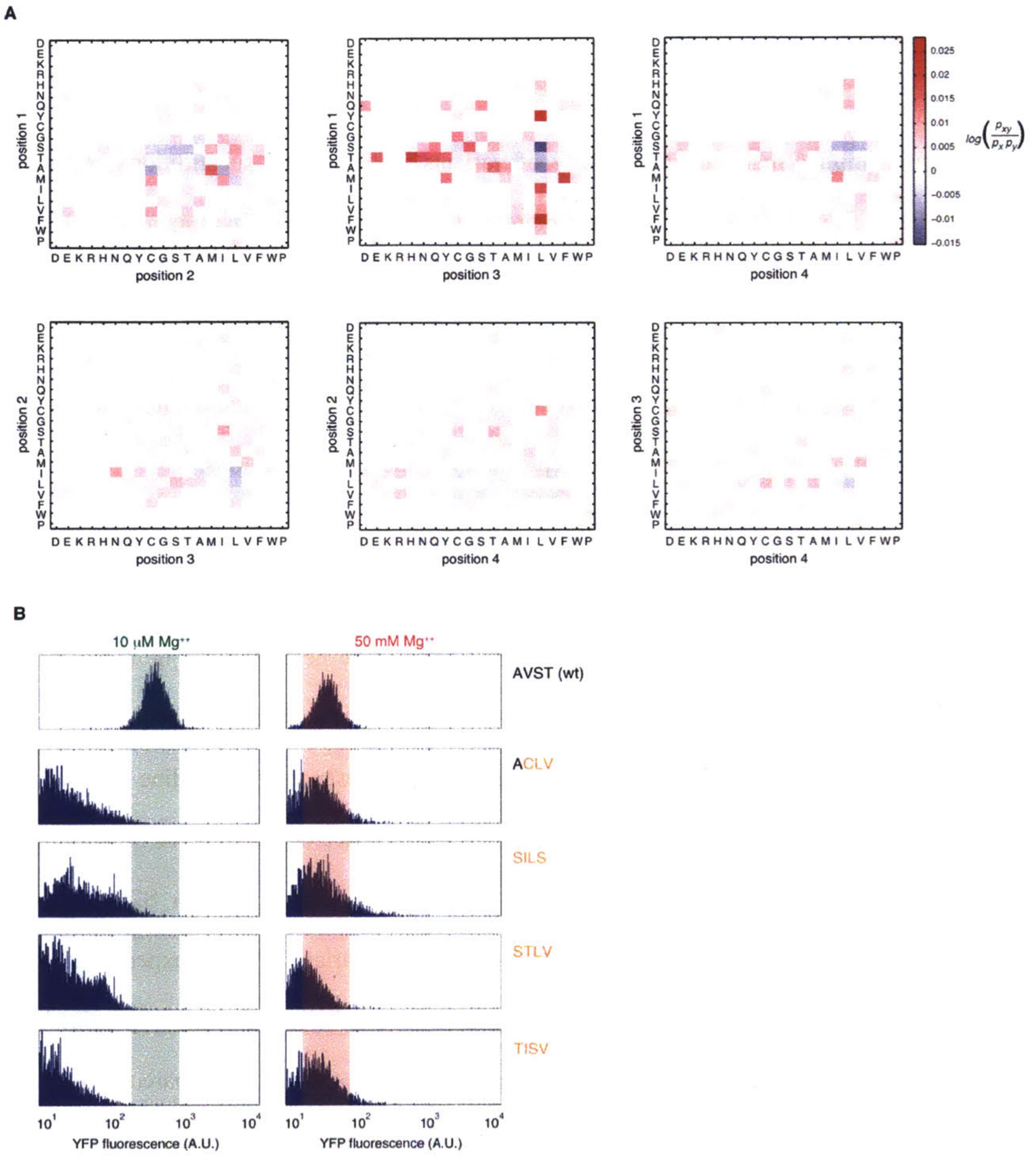


Figure 3.8 Epistasis between PhoQ interface residues.

(A) Heatmaps showing, for each pair of positions in PhoQ that were randomized in the library, the frequency of all residue pairs relative to frequency expected if residues occurred independently. (B) Flow cytometry profiles for PhoQ variants that contain residues found at high overall frequency in the post-selection library but that are not functional in combination (also see Fig. 3f).

results demonstrate that the effects of individual substitutions are often highly context-dependent, or epistatic, in promoting a PhoQ-PhoP interaction.

This epistasis implies that the functionality of variants with multiple substitutions cannot be easily predicted from the behavior of single point mutants or site-saturation mutagenesis. If each position contributed independently, our single mutant data (Figure 3.6A) would predict 420 functional combinations: [AS][ACDILV][ACLMSTV][RTVWY]. However, our two-step screen recovered only 85 of these 420, and revealed an additional 1,574 functional combinations (Figure 3.7E), emphasizing the strong interdependency of individual positions.

To further assess interdependencies in PhoQ, we measured mutual information between each pair of positions in the set of 1,659 functional variants. The strongest coupling was observed between positions 1 and 2 and between positions 1 and 3. The heatmaps in Figures 3.7F and 3.8A indicate how much the frequency of each possible pair of amino acids differs from that expected if the positions were independent. For instance, the presence of a histidine at position 1 in functional PhoQ variants is highly correlated with the presence of a leucine or methionine at position 3 (Figure 3.7F). The combination of A284H together with S288L or S288M occurs about three times more often in the set of 1,659 functional variants than is expected if these substitutions occurred independently. This functional interdependency is further emphasized in a frequency logo built using only sequences with a histidine at position 1 compared to the logo built using all 1,659 sequences (Figure 3.7G). Similarly, a glutamine at position 3 is strongly correlated with either a serine or threonine at position 1 (Figures 3.7F and G). These dependencies likely arise from constraints on the packing of adjacent residues in the PhoQ-PhoP interface, as also suggested by structural studies of a rewired two-component signaling system (Podgornaia et al., 2013).

The epistasis observed between interfacial residues significantly constrains the mutational trajectories that PhoQ can follow through sequence space, assuming PhoQ must retain the ability to interact productively with PhoP. For instance, of the 100 functional double mutants of PhoQ, only 23 represent cases where both single mutants are functional (Figure 3.9A). In 46 cases, only one of the single mutants is functional meaning that the mutational paths to these double mutants are constrained, requiring a certain order of substitutions. In the remaining 31 cases, the double mutant is functional even though neither single mutant is functional. Thus, trajectories connecting the wild-type combination AVST to these double mutants, if they exist, must be indirect, requiring the acquisition of a transient additional mutation. For example, the double mutant TVNT supports a PhoQ-PhoP interaction, but neither TVST nor AVNT is functional. However, the wild-type AVST could transition first to AVTT; this S288T mutation permits the subsequent substitution A284T to create the functional intermediate TVTT. The substitution S288N can then simultaneously undo the transient mutation S288T and yield the functional double mutant TVNT (Figure 3.9A). False negatives in our data could mean that some double mutants without a direct path from AVST are, in fact, connected by a functional intermediate, but the rate is low enough (Figure 3.5) that only a few cases are likely to be impacted.

To systematically explore the impact of epistasis on mutational trajectories, we quantified the shortest path connecting the wild-type combination AVST with each of the 1,658 functional variants (Figure 3.9B). For 428 variants the Hamming distance (in amino acids) from AVST equals the number of mutational steps necessary to reach it. For example, AVST can be changed to the quadruple mutant SLCC using four consecutive substitutions. However, due to epistasis, these substitutions cannot be introduced in any order; only 3 of the 24 possible direct paths are permissible (Figure 3.9C). Strikingly, nearly 70% (1,151) of the functional combinations require

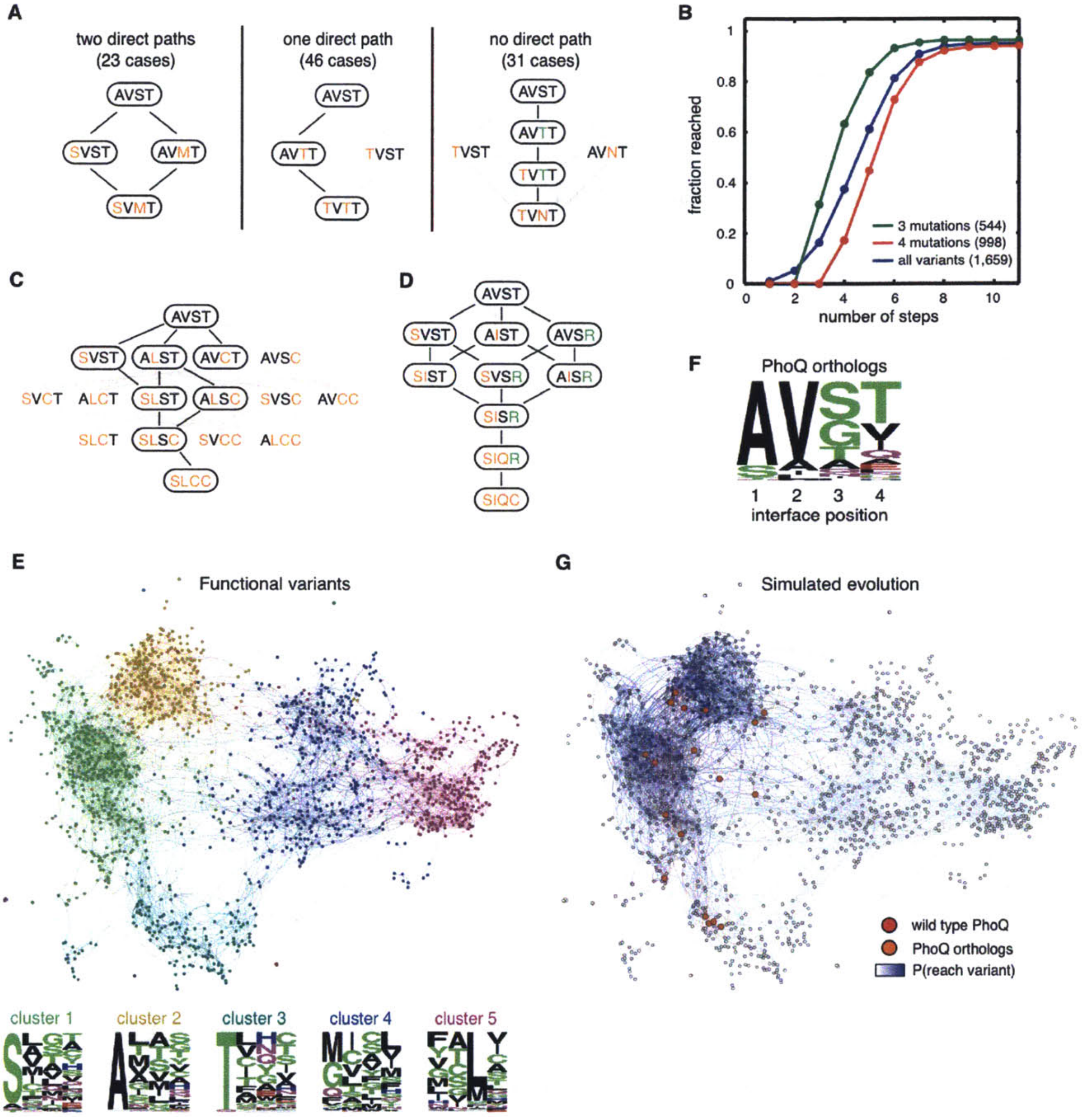


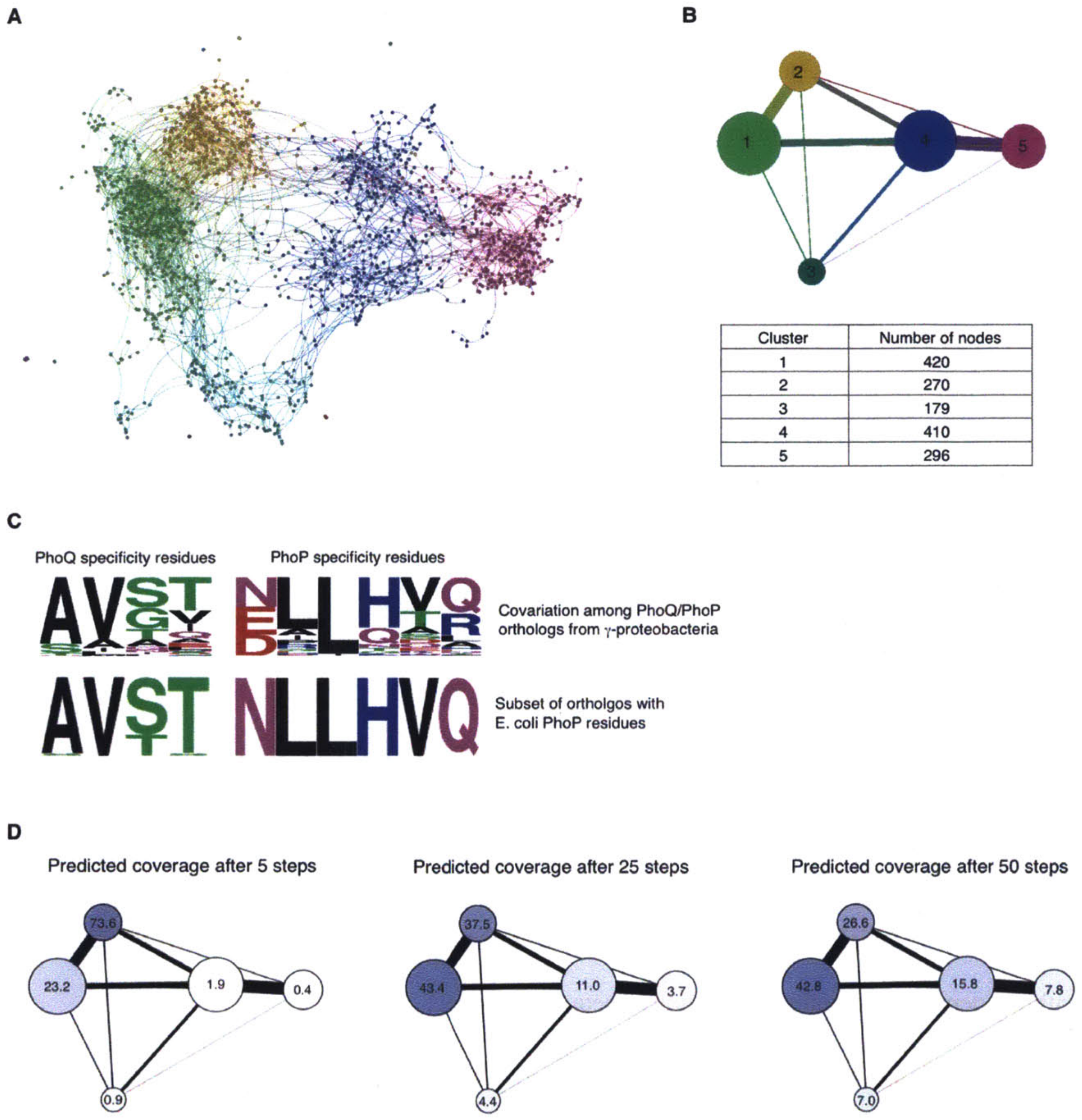
Figure 3.9 PhoQ sequence space.

(A) Tabulation and examples of double mutants that can be reached by 2, 1, or 0 direct paths. Functional PhoQ variants are circled. Variants differing by one residue are connected by lines (black for accessible paths; gray for inaccessible paths). Orange text indicates a residue present in the double mutant; green indicates a mutation not present in either terminal node (also, see Extended Data Fig. 6a). (B) Cumulative number of variants reached in a given number of mutational steps from wild-type PhoQ. Curves are shown for all functional variants and for subsets harboring 3 or 4 mutations. (C-D) Trajectories between AVST (wild-type) and SLCC (C) or SIQC (D) plotted as in (A). **e**, Force-directed graph of PhoQ sequence space: each node is a functional variant and edges connect variants differing by one residue. The graph was clustered for maximal modularity, and each cluster colored. (F) Frequency logo for interfacial residues of PhoQ orthologs from g-proteobacteria. (G) The graph from (E) but with the probability of a given node being reached after 25 mutational steps indicated by the color scale shown. Orange nodes indicate variants found in PhoQ orthologs.

more mutational steps than their Hamming distance from wild type, indicating that many mutational paths in sequence space are indirect, requiring the transient introduction of mutations that make other mutations permissible. For instance, with the quadruple mutant SIQC noted above (Figure 3.7), there are six possible paths connecting it to the wild-type AVST (Figure 3.9D). Each path requires five mutational steps and involves the transient introduction of arginine at the fourth position. These analyses underscore the context-dependency of individual substitutions in PhoQ interfacial residues and indicate that epistasis may fundamentally constrain PhoQ evolution.

There are 79 PhoQ variants that cannot be reached from the wild-type sequence without passing through a non-functional intermediate. Most (57) of these have no functional neighbors with Hamming distance 1 suggesting that they are mutationally inflexible. By contrast, the wild-type sequence has 16 neighbors differing by only one amino acid, indicating that *E. coli* PhoQ resides within a densely interconnected region of sequence space.

To better visualize the connectivity of sequence space, we generated force-directed graphs in which nodes represent functional PhoQ variants and edges connect nodes whose sequences differ by one amino acid (Figures 3.9E and 3.10A). These graphs reveal five densely interconnected clusters within sequence space (Figure 3.10B). The cluster containing the wild-type combination, AVST, involves sequences typically harboring alanine at the first position. The graph contains 13 residue combinations found in PhoQ orthologs, 12 of which reside within just two clusters.



Cluster	Number of nodes
1	420
2	270
3	179
4	410
5	296

Figure 3.10 Analysis of PhoQ library and ortholog diversity.

(A) Graph from Fig. 4e with nucleotide changes rather than amino acid changes linking the functional nodes. (B) The number of nodes in each cluster shown in Fig. 4e. (C) Frequency logos for specificity-determining interfacial residues in PhoQ and PhoP orthologs from g-proteobacteria. Logos were built using 90 orthologous, co-operonic PhoQ-PhoP pairs *(top)* or the subset of 34 orthologous pairs that have PhoP residues matching those found in *E. coli* (*bottom*). (D) Diagrams showing probability (as percentage) that each cluster will be reached after a given number of mutational steps.

Some residue combinations found in PhoQ orthologs are not present in the graph. However, these orthologs often have PhoP partners quite divergent from *E. coli* PhoP (Figure 3.10C); consequently, these PhoQ orthologs are constrained differently than *E. coli* PhoQ and therefore likely have a different set of functionally degenerate variants in sequence space.

In general, the natural diversity of interfacial residues in PhoQ orthologs (Figure 3.9F), even those with divergent PhoP partners, is much more limited than the diversity in our selected, functional variants. This difference may indicate that some PhoQ variants, and even entire clusters, identified here as functional have subtle defects that confer a disadvantage in the wild. Alternatively, mutational trajectories may be fundamentally constrained both by epistasis (Breen et al., 2012; DePristo et al., 2005; Lehner, 2011; Ortlund et al., 2007; Weinreich et al., 2005) and by the dense interconnectivity of clusters within sequence space such that nature has not sampled certain regions of sequence space. To test this idea we simulated PhoQ mutational trajectories starting from the *E. coli* combination AVST, with each step involving a single mutation to a functional PhoQ variant at Hamming distance 1. Even after 25 mutational steps, a relatively limited region of sequence space was explored (Figures 3.9G and 3.10D). The region most densely sampled in these simulations included the PhoQ ortholog residue combinations. These results support the notion that the local density and interconnectivity of functional variants has limited nature's exploration of sequence space, reflected in the limited diversity of PhoQ orthologs (Breen et al., 2012; Povolotskaya and Kondrashov, 2010).

Collectively, our comprehensive map of PhoQ sequence space reveals greater mutational flexibility and degeneracy than would be expected either by site-saturation mutagenesis, which assumes position independence, or by studies of natural diversity. Further, our work demonstrates the pervasive impact of epistasis on protein function. In sum, the approach described here for systematically mapping sequence space provides new insights into the fundamental nature of protein function and evolution, and will significantly impact efforts in protein design and engineering.

Materials and Methods

Growth media and strains

Cultures were grown at 37°C in M9 minimal medium consisting of 1x M9 salts, 100 μM $CaCl_2$, 0.2% glucose, and 0.1% casamino acids, with $MgSO_4$ added at the concentrations indicated. Antibiotics were added, when appropriate at the following concentrations: carbenicillin, 50 μg/mL; kanamycin, 30 μg/mL; spectinomycin, 50 μg/mL.

The base strain for all flow cytometry experiments was ML2195, which was constructed by transforming *E. coli* strain TIM175 (MG1655 Δ*phoPQ* Δ*lacZYA* *attλ*::[P_{mgrB}-*yfp*+] *attHK*::[P_{tetA}-*cfp*+]) with a pBR322 plasmid (amp^R) containing P_{mgrB}-*yfp*+ and P_{tetA}-*cfp*+. This strain was subsequently transformed with a low-copy plasmid pLPQ2 (pSC101 origin; spec^R) expressing *phoPQ* from the *lacUV5* promoter (Lippa and Goulian, 2009; Miyashiro and Goulian, 2007). Point mutations in *phoQ* were introduced into plasmid pLPQ2 using PCR-based site-directed mutagenesis as previously described (Ashenberg et al., 2011).

Flow cytometry measurements

Overnight cultures were washed with PBS, diluted 1:100 into M9 with 10 μM Mg^{++} or 1:250 into M9 with 50 mM Mg^{++}, and grown for 6 hours. Freeze-thawed library samples were compared to freeze-thawed aliquots of isogenic strains harboring wild-type *phoQ* on plasmid pLPQ2. Cultures were diluted to OD_{600} of 0.001 in PBS and flow cytometry measurements made for ~10^4 cells per strain. The following settings were used on the LSRII-HTS cytometer: FITC voltage 540; AmCyan voltage 510; SSC 220 (threshold at 400); FSC 530.

Plasmid library construction

The *phoQ* library was generated following previously published methods(Kinney et al., 2010). Briefly, a DNA cassette containing *ccdB-camR* flanked by BsmBI restriction sites was cloned between nucleotides 678 and 964, corresponding to residues 226 and 322 of the *phoQ* gene on plasmid pLPQ2, creating plasmid pLPQ2-*ccdB4*. A pool of double-stranded DNA (from DNA2.0; Menlo Park, CA) corresponding to residues 227-321 of PhoQ with randomized NNS codons for residues 284, 285, 288, and 289, was synthesized with flanking BsaI restriction sites. The library and pLPQ2-*ccdB4* were each digested for 1 hour with type II nucleases BsaI and BsmBI (NEB), respectively, to create overhangs 'AAAA' and 'CTGC', purified (Qiagen PCR-purification kit), and then ligated together at equimolar (175 fmol) concentration for 16 hours at 16°C. The ligation reaction was then dialyzed for 90 minutes using 0.025μm filters (Millipore) and electroporated into DH10B *E. coli* (Invitrogen). The electroporation, yielding ~10^8 unique transformants (~10^5 transformants in no insert control), was recovered overnight in LB and plasmids then midi-prepped (Qiagen). The library was then transformed into strain ML2195, yielding ~$5x10^7$ unique transformants, with cells recovered in LB with selection for the library and reporter plasmids for six hours (final OD_{600} ~ 0.5) (Santoro and Schultz, 2002), and aliquots flash-frozen in 10% DMSO.

Library sorting and competition

Cultures were grown as described above for flow cytometry analysis and diluted to OD_{600} 0.1 in PBS. Cells with YFP levels comparable to those seen in cells harboring wild-type *phoQ* measured on the same day were selected using a FACS Aria (50 mM Mg^{++}). Approximately $2x10^7$ cells, representing 20x coverage (Denault and Pelletier, 2007) of the library, were sorted and those behaving like the wild-type control were plated on LB with plasmid selection, and allowed to recover overnight. The resulting colonies were pooled, mixed to uniformity in LB, and aliquots frozen in 10% DMSO. Two biological replicates were done.

A freeze-thawed aliquot of sorted cells was inoculated into M9 media containing 2 mM Mg^{++} and grown for ~5 hours to OD_{600} ~ 0.5. Cells were washed with PBS and diluted into M9 media with 0 mM Mg^{++} overnight, representing ~18 hours of Mg^{++} starvation. Cells were then diluted 1:10 in fresh M9 media with 2 mM Mg^{2+} and allowed to grow to OD_{600} ~ 0.5 (10 hours), at which point aliquots were flash-frozen for future analysis. One of the biological replicates from the FACS-based screen above was subjected to two independent competition replicates. A sufficient number of cells were sampled at every step to avoid bottlenecks. Cell counts were tallied using an average of CFU counts plated at two different 10-fold dilutions on LB with carbenicillin and spectinomycin (Figure 3.4D).

Illumina sample preparation and sequencing

Plasmids were mini-prepped (Qiagen) from frozen aliquots of cells and used as template for PCR reactions (20 cycles) with custom barcoded primers containing Illumina flowcell adaptor sequences (Table 3.1). The samples were multiplexed (4–6 samples per run) and sequenced on an Illumina HiSeq instrument.

Analysis of Illumina data for sorted library

Illumina reads from a single lane were grouped using exact match to one of several four-letter barcodes. The reads were quality-filtered for exact match to the sequence NNSNNSCTGCAANNSNNSCTGCGTTCTCTGCGTAGT, where N is any nucleotide and S is G or C. Read counts were normalized such that each sequenced pool had the same total. DNA sequences with < 5 quality-filtered reads in the initial library were omitted from further analysis.

For each DNA sequence we calculated the enrichment ratios of post- to pre- sorted library samples, and the enrichment ratios of post- to pre- competition samples. These were then converted to enrichment ratios for each amino acid sequence by averaging over the ratios for underlying DNA sequences. This ensured that amino acid sequences with a large number of reads (e.g. 'SSSS', 81 possible unique DNA sequences) were not preferentially selected over sequences represented with a few DNA sequences.

We tabulated the enrichment ratios for two biological replicates of the library sorting and two biological replicates of the library competition for each amino acid sequence. We used a binary classification tree, implemented in MATLAB, to classify the library sequences. The training set was derived from our data for 57 individually tested point mutants of PhoQ at positions 284, 285, and 288, and the behavior of the 34,481 PhoQ variants harboring one or more stop codons in our two-step selection. Of the 57 point mutants, 13 were deemed functional, *i.e.* signal-responsive and competitive against the wild-type, and hence true positives, with the 44 remaining point mutants deemed non-functional, and hence true negatives. The 34,4841 PhoQ variants containing a stop codon, each of which produces a truncated and therefore non-functional protein, were also considered true negatives. The binary classifier seeks to chose thresholds for enrichment ratios of individual sequences in the post-selection library that maximize inclusion of true positives while minimizing inclusion of true negatives.

Competition assays

Previously described strains MC4100-CFP and MC4100-YFP were used for competition experiments (Hegreness et al., 2006). Chromosomal mutations in *phoQ* were constructed in MG1655 using λ Red-mediated recombination, transduced into MC4100-YFP using P1 phage, and the FRT-flanked kanamycin-resistance cassette subsequently removed by expressing

FLP(Datsenko and Wanner, 2000). Overnight bacterial cultures of MC4100-YFP harboring a given *phoQ* mutant and MC4100-CFP harboring wild-type *phoQ* were each diluted into fresh M9 medium with 2 mM Mg^{++} and grown for ~2 hrs to exponential phase. The cells were then washed with PBS, diluted to OD_{600} ~ 0.002 and mixed 1:1 with a competitor strain in 20 mL M9 containing 0 or 50 mM Mg^{++}. Samples were collected every 2 hours, fixed using paraformaldehyde, and the proportions of YFP- and CFP-tagged cells tallied using flow cytometry. Each competitor was present at 35-65% of the population at the start of each competition. Data was normalized to a 50% initial ratio for visualization.

To generate fitness data for the 57 single point mutants of PhoQ at positions 284, 285, and 288, we conducted head-to-head competition assays using strains ML2231 (MC4100-CFP Δ*phoPQ*) and ML2232 (MC4100-YFP Δ*phoPQ*) harboring pLPQ2 or its variants. The competition experiments were set up as described above, with the exception that the strains were competed in 0 mM Mg^{++} overnight, recovered for 6 hours in 2 mM Mg^{++}, and then tallied by flow cytometry. Fitness was calculated as the ratio of the percentage mutant strain before and after the competition; strains that remained at > 85% of their starting relative number were called as fit (Figure 3.5).

Protein purification and phosphotransfer assays

Expression, protein purification, and phosphotransfer experiments were carried out as previously described (Skerker et al., 2005). Only the cytoplasmic portion (residues 238-486) of PhoQ and its variants was used. The kinase was autophosphorylated for 1 hour at 30°C and incubated at a 1:8 ratio (Sanowar and Le Moual, 2005) with full-length PhoP (1 μM PhoQ and 8 μM PhoP in 10 μL reaction).

Identification of orthologs and force-directed graphing

A reciprocal best BLAST hit search was implemented using Python and queried against fully sequenced bacterial genomes available in GenBank as of August 2013. The list of hits was curated to contain only co-operonic PhoQ-PhoP pairs (as indicated by the gene identifier numbers) with less than 95% overall similarity to the *E.coli* proteins. This resulted in 89 orthologs from γ-proteobacteria, of which 35 are unique sequences.

Force-directed graphs were generated using Gephi network visualization software (Bastian et al., 2009). The layout was generated by running the Force Atlas algorithm to completion; the clusters were generated based on modularity with the Resolution parameter set to 2. For the graph in Figure 3.9E the edges were generated for every two sequences Hamming 1 distance apart; for the graph in Figure 3.10A edges connect sequences that can be reached by nucleotide substitutions.

Table 3.1 Primers used for sequencing on Illumina HiSeq.

Primer Name	**Sequence**
Illumina_FW	AATGATACGGCGACCACCGAGATCTGGGATGTGCTGCATAGTCTGAAAACGCCACTG
Illumina_RV1	CAAGCAGAAGACGGCATACGAGATACTACAGACTACGCAGAGAACGCAG
Illumina_RV2	CAAGCAGAAGACGGCATACGAGATACTCAGTACTACGCAGAGAACGCAG
Illumina_RV3	CAAGCAGAAGACGGCATACGAGATACTGATGACTACGCAGAGAACGCAG
Illumina_RV4	CAAGCAGAAGACGGCATACGAGATACTTGTCACTACGCAGAGAACGCAG
Illumina_RV5	CAAGCAGAAGACGGCATACGAGATACTAACCACTACGCAGAGAACGCAG
Illumina_RV6	CAAGCAGAAGACGGCATACGAGATACTGGATACTACGCAGAGAACGCAG
Illumina_seq	GGGATGTGCTGCATAGTCTGAAAACGCCACTG

* underlined positions correspond to barcodes for multiplexing.

References

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.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature *490*, 535-538.

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.

Datsenko, K.A., and Wanner, B.L. (2000). One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A *97*, 6640-6645.

Denault, M., and Pelletier, J.N. (2007). Protein library design and screening: working out the probabilities. Methods Mol Biol *352*, 127-154.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nature reviews. Genetics *6*, 678-687.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat Methods *7*, 741-746.

Groisman, E.A. (2001). The pleiotropic two-component regulatory system PhoP-PhoQ. J Bacteriol *183*, 1835-1842.

Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nature reviews. Genetics *14*, 559-571.

Hegreness, M., Shoresh, N., Hartl, D., and Kishony, R. (2006). An equivalence principle for the incorporation of favorable mutations in asexual populations. Science *311*, 1615-1617.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N. (2011). Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A *108*, 7896-7901.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc Natl Acad Sci U S A *107*, 9158-9163.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends in genetics : TIG *27*, 323-331.

Lippa, A.M., and Goulian, M. (2009). Feedback inhibition in the PhoQ/PhoP signaling system by a membrane peptide. PLoS Genet *5*, e1000788.

McLaughlin, R.N., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature *491*, 138-142.

Miyashiro, T., and Goulian, M. (2007). Stimulus-dependent differential regulation in the Escherichia coli PhoQ PhoP system. Proc Natl Acad Sci U S A *104*, 16305-16310.

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.

Podgornaia, A.I., Casino, P., Marina, A., and Laub, M.T. (2013). Structural Basis of a Rationally Rewired Protein-Protein Interface Critical to Bacterial Signaling. Structure.

Povolotskaya, I.S., and Kondrashov, F.A. (2010). Sequence space and the ongoing expansion of the protein universe. Nature *465*, 922-926.

Sanowar, S., and Le Moual, H. (2005). Functional reconstitution of the Salmonella typhimurium PhoQ histidine kinase sensor in proteoliposomes. Biochem J *390*, 769-776.

Santoro, S.W., and Schultz, P.G. (2002). Directed evolution of the site specificity of Cre recombinase. Proc Natl Acad Sci U S A *99*, 4185-4190.

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.

Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature *225*, 563-564.

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

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.

Weinreich, D.M., Watson, R.A., and Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution; international journal of organic evolution *59*, 1165-1174.

Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., *et al.* (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol *30*, 543-548.

Chapter 4

Conclusions and Future Directions

The data presented here are unpublished.

Conclusions

In the course of my graduate studies I have characterized the interaction specificity of two-component signaling pathways at the molecular level. Previous work from the Laub lab had identified the residues that maintain interaction specificity in these signaling pathways (Skerker et al., 2008). Additionally, it had been shown that changing as few as three of these residues is enough to "rewire" the interaction preference towards a non-cognate partner (Capra et al., 2010). In my PhD work I built upon this knowledge by answering the following two questions: How are new residues physically accommodated at the protein-protein interface? How many residue combinations can be used at a particular interface without altering the interaction specificity?

To address the first question I initiated a research collaboration with the Marina group, which solved the first co-crystal structure of a kinase bound to a regulator (Casino et al., 2009). Our collaboration combined the two labs' expertise in protein rewiring and X-ray crystallography to arrive at a structural description of a rewired two-component signaling interface. As shown in Chapter 2, I found that the non-native residues introduced during rewiring are accommodated through a repacking of the interfacial residues and by a modest rigid body rotation of the kinase and regulator molecules relative to each other. Furthermore, I generated a rich mutagenesis data set and used it to understand the extent of epistasis at the interface and its effect on mutational trajectories.

The finding that two-component proteins exhibit epistasis at their interface motivated me to undertake an ambitious project to systematically map the sequence space of interfacial residues. In Chapter 3 I described the experimental pipeline necessary to achieve this goal, including the creation of a large and diverse protein library, high-throughput screening for both binding and *in vivo* fitness, and deep sequencing to identify enriched variants. I performed extensive experimental validation of selected functional variants using transcriptional reporters, *in vitro* biochemical experiments, and head-to-head competitions against the wild-type variant. Although it was previously known that two-component systems could be rationally rewired, this project showed that these proteins' interfaces are highly degenerate and can utilize a large number of different interfacial residues. Furthermore, I confirmed that the interface is characterized by extensive epistasis between individual residues, highlighting the non-additive contributions made by each interface position to partner binding and signal transmission. Using the data generated from this project I was able to construct a network map of a histidine kinase sequence space and generate a model explaining why nature has sampled only a subset of the available functional variants.

The results presented here have solidified our understanding of two-component signaling specificity and also provided generalizable rules for protein design and evolution. The data sets generated in the course of my thesis work can be used in the future for computational modeling of the protein-protein interface. I have also established a framework for studying protein sequence space that can be extended to other studies, as described below.

Future Directions

Systematic mapping of response regulator sequence space

The experimental and bioinformatics approaches for studying the interfacial residues of the histidine kinase PhoQ (Chapter 3) can be readily applied to an analysis of the interfacial residues on the cognate response regulator PhoP. Previous work had identified six specificity residues in the first α-helix of the response regulator receiver domain (Capra et al., 2010; Skerker et al., 2008). Preliminary flow cytometry and *in vitro* phosphotransfer data for alanine point mutations

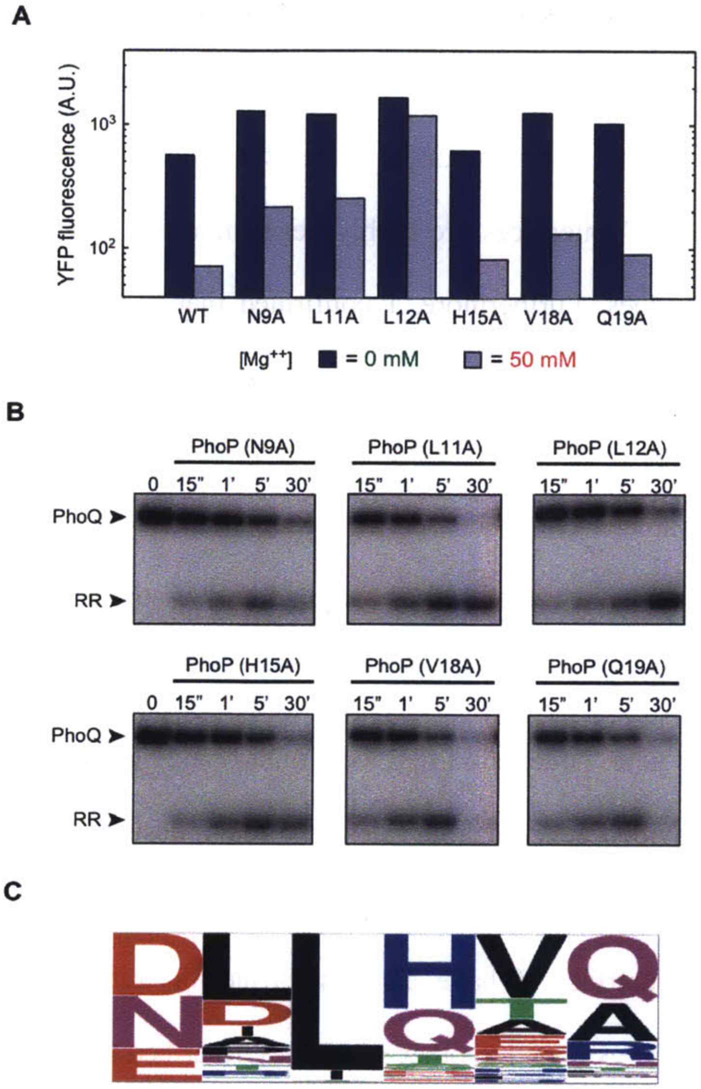


Figure 4.1 Alanine scan of *E. coli* PhoP specificity residues.

(A) Flow cytometry measurements of YFP levels for wild-type and indicated alanine mutants of PhoP in 0 mM and 50 mM Mg^{++}. (B) PhoQ was autophosphoryalted *in vitro* and tested for phosphotransfer and dephosphorylation of indicated PhoP mutants. (C) Sequence logo of PhoP orthologs. The positions correspond to residues 9, 11, 12, 15, 18, and 19 in *E. coli* PhoP.

in PhoP show varying effects on kinase and phosphatase activity (Figures 4.1A and B). In order to have a manageable library size, the PhoP library could be limited to just 3-4 positions as was done for PhoQ. I would choose positions N9, L11, L12, and H15, where the alanine substitutions have the greatest effect. Additionally, these positions show varying levels of conservation among

PhoP orthologs (Figure 4.1C) and it would be interesting to see if these patterns hold up in the entire sequence space.

Although some aspects of this follow-up project are straightforward (*i.e.* we can expect the PhoP side of the interface to be as degenerate and epistatic as the PhoQ side), there is a lot to learn about the biology of response regulators. How do mutations at the interface affect the stability of phosphorylated PhoP? Do these mutations affect dimerization and the subsequent activation of DNA transcription? If we perform this screen in a Δ*phoQ* background, we can learn if PhoP is activated by cross talk from other histidine kinases. Although we routinely test whether a histidine kinase can phosphorylate non-cognate response regulators using *in vitro* phosphotransfer, it is difficult to test for cross talk directed at a particular response regulator due to the need to purify and individually autophosphorylate dozens of histidine kinases. Thus, using the established *in vivo* screening tools for the PhoQ-PhoP pathway can lead to novel insights into two-component signaling.

Screen for novel two-component signaling interfaces

A more ambitious application of the approaches developed in Chapter 3 involves randomizing positions in both histidine kinase and response regulator to identify novel two-component signaling interfaces. There are a number of technical challenges involved in generating and screening the necessary starting library, as randomizing just three positions on each side of the interface results in 20^6, or ~$6x10^7$ different protein variants. This number is at the limit of current DNA synthesis and screening capabilities, especially when considered in light of the need to oversample at each step of the high-throughput screen. The positions of the mutant residues in the two proteins presents another challenge, as they are spaced over 1 kbp apart and cannot be captured by a single short Illumina read. This spatial separation will require a custom cloning approach to adapt it to current deep sequencing protocols.

One solution that avoids the combinatorial challenge pointed out above is to screen only a subset of all possible combinations and identify a small number of novel interfaces. Even a single new interface would serve as proof of concept for the method and could be extensively interrogated by low-throughput approaches. For example, does the novel pair cross talk with other signaling pathways in *E. coli*? Has this particular set of residues been observed in any other two-component pair from sequenced bacterial genomes? This project stands to teach us a lot about protein engineering as well as the "systems biology" of bacterial signaling pathways.

A computational model of the PhoQ-PhoP interaction interface

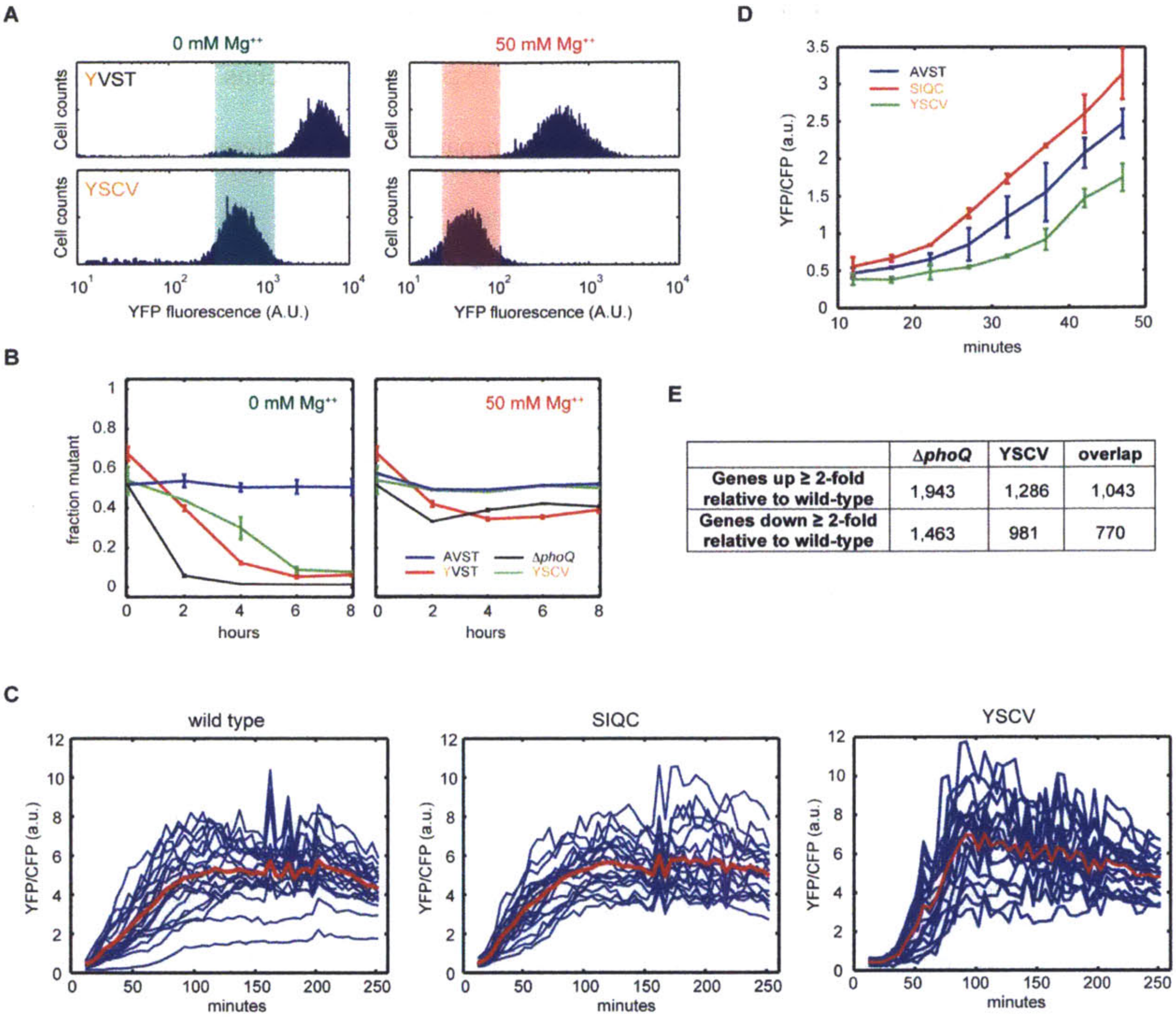
In Chapters 2 and 3 I have presented experimental mutagenesis data for two different two-component signaling pathways. One way to extend my findings to other signaling pathways is to build a predictive computational model of residues that are permitted at the kinase-regulator interface. We could use the experimental data for 128 interface combinations (Figure 2.10) in *Thermotoga maritima* proteins HK853 and RR468 to train modeling software such as Rosetta to recognize functional combinations. Given a good enough model, we can conduct a high-throughput mutagenesis study of the interface *in silico*. There is currently no PhoQ-PhoP complex crystal structure, but when one becomes available it will be possible to use the negative and positive data for 160,000 different variants (Chapter 3) to better model the interface.

One challenge to computational modeling is that we do not know exactly how the mutations affect the interface. We saw in Chapter 2 that accommodation of the residues can involve rigid-body rotations of the kinase and regulator relative to each other. A suitable model must sample backbone movements, although the full extent of allowed rotations is not understood. A further challenge is the need to understand the structural differences between kinase and phosphatase states of a histidine kinase. Ultimately, high-throughput experimental screens and computational models of two-component interfaces should advance in parallel to completely characterize the biochemical properties of these pathways.

Dynamics of two-component signaling activation

In addition to various high-throughput screens, the data and tools presented in Chapter 3 suggest a complementary, low-throughput approach for studying two-component signaling. From the behavior of different PhoQ mutants I learned that it is not only the steady state levels of phosphorylated PhoP that matter during magnesium starvation, but also the temporal dynamics of activation. One variant that I identified in an early version of the library screen has the residues "YSCV" at the interface and exhibits signal-responsive behavior despite the bulky, individually deleterious A284Y substitution (Figure 4.2A). Surprisingly, this variant fared very poorly in head-to-head competitions against chromosomally encoded PhoQ (Figure 4.2B). I monitored the dynamics of this variant alongside wild type and the signal-responsive variant "SIQC" (Chapter 3) using time-lapse fluorescence microscopy. Interestingly, I saw that the YSCV variant is initially slower at activating the PhoP regulon, but eventually reaches wild-type steady-state YFP levels (Figure 4.2C-E). In sum, these results suggest that activation dynamics are very important for this two-component system.

Although in Chapter 3 I focused primarily on the functional variants of PhoQ, we can use non-functional variants like YSCV to learn more about the overall properties of the signaling pathway. Specifically, which aspect of histidine kinase activity do deleterious mutations affect and how does this activity contribute to the overall level of transcriptional output? The simplest explanation is that incompatible interface mutations impair binding to and subsequent



	ΔphoQ	YSCV	overlap
Genes up ≥ 2-fold relative to wild-type	1,943	1,286	1,043
Genes down ≥ 2-fold relative to wild-type	1,463	981	770

Figure 4.2 Defects in PhoQ activation dynamics can affect organismal fitness.

(A) Flow cytometry measurements of YFP for the PhoQ variants indicated. Shaded regions indicate wild-type YFP levels. (B) Head-to-head competitions of wild-type *E. coli* against strains producing the indicated PhoQ variants from the chromosome. (C) Time-lapse microscopy for PhoQ variants at 0 mM Mg^{++}. Each blue trace represents a cell; the red trace represents an average over n = 20 cells. (D) Time-lapse microscopy comparing activation at 0 mM Mg^{++}. Each trace is an average of two biological replicates, each in turn averaged over n = 20 cells. Error bars show ± SD. (E) Microarray analysis: number of genes up- and down- regulated in Δ*phoQ* and YSCV (relative to wild-type) after 45 minutes in 0 mM Mg^{++}.

phosphorylation of PhoP. Alternatively, a non-functional variant could bind PhoP too tightly and sequester the phosphorylated regulator from interacting with DNA and activating gene expression. These variants could also be structurally destabilized or have impaired autophosphophorylation kinetics.

To understand the relationship between histidine kinase activity and transcriptional output we can measure the system using fluorescence time-lapse microscopy under different stimulatory conditions. Coupled with a microfluidic device, microscopy can also reveal differences in the response to constant vs. pulsed or gradually increasing signal. Using tools originally created in the Goulian lab (Miyashiro and Goulian, 2007), we can compare the temporal changes in a PhoP-regulated promoter fused to YFP to the control *tetA* promoter driving CFP. By analyzing the microscopy data, we can dissect the influence of each regulatory component in the PhoQ-PhoP signaling pathway, including positive auto-regulation, negative feedback from the small protein MgrB (Lippa and Goulian, 2009), and other cellular components such as acetyl-phosphate and non-cognate histidine kinases (Figure 4.3). With these data we can build a mathematical model and learn generalizable principles that govern two-component signaling.

Concluding Remarks

Bacterial model systems have long offered tractable experimental set-ups for addressing fundamental questions in biology. I have demonstrated how applying cutting edge tools to a model organism can yield exciting insights into protein function As a result of my work we have a network map of the PhoQ sequence space whose properties (*e.g.* density of nodes near the *E. coli* wild-type sequence) explain why nature has not sampled all of the functional PhoQ variants. One caveat is that my definition of "functional" is limited to those variants that survive

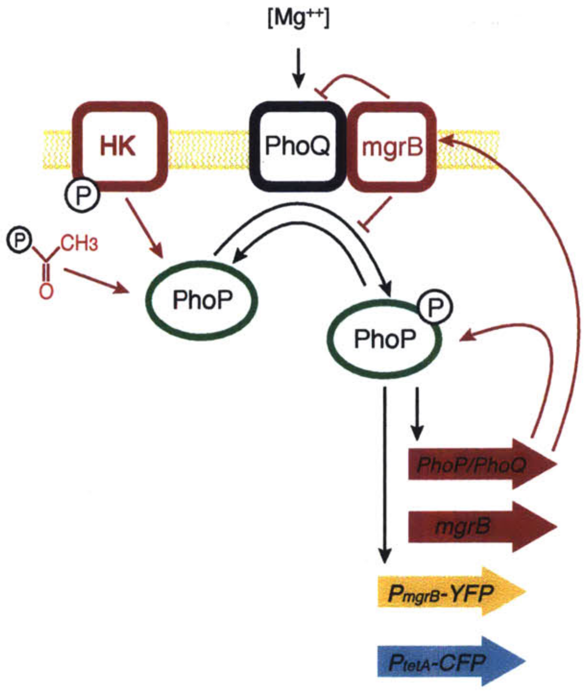


Figure 4.3 Components of the *E. coli* PhoQ-PhoP signaling pathway.

The PhoQ-PhoP signaling pathway is subject to positive and negative feedback, as well as potential phosphorylation by acetyl-phosphate and other histidine kinases.

laboratory conditions; they are not subjected to the barrage of environmental stresses witnessed by bacteria in their natural environment. To understand whether the properties of the network graph for *E. coli* PhoQ generalize to other signaling pathways, we could extend the high-throughput screen to an orthologous pair, such as PhoQ-PhoP from another bacterial species. Furthermore, we could set up the same screen for a completely different two-component system and see whether the extent of explored sequence space matches what I have found for PhoQ.

I anticipate that such studies of protein sequence space will become more commonplace as DNA synthesis and sequencing technologies continue to advance. It would be very interesting to learn the extent to which other types of protein-protein interfaces exhibit degeneracy and epistasis.

One could imagine that degeneracy would be inversely correlated with binding strength and that epistasis would vary for residues found in different types of protein secondary structure elements. Although my time in the Laub lab is ending, current members of the lab are actively pursuing all of the future directions outlined above in order to further our understanding of protein interactions at the atomic, molecular, and organismal levels.

References

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.

Lippa, A.M., and Goulian, M. (2009). Feedback inhibition in the PhoQ/PhoP signaling system by a membrane peptide. PLoS Genet *5*, e1000788.

Miyashiro, T., and Goulian, M. (2007). Stimulus-dependent differential regulation in the Escherichia coli PhoQ PhoP system. Proc Natl Acad Sci U S A *104*, 16305-16310.

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.