Title

The Study and Engineering of Cellular Signaling Pathways

Author

Groban, Eli S.

Publication Date

2008

The study and engineering of cellular signaling pathways.

by

Eli S. Groban

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Graduate Group in Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

[redacted]

Acknowledgements

This work would not have been possible without the help and support of various members of my family, my friends, my advisors, and my colleagues. First, I would like to acknowledge Christopher Voigt and Matthew Jacobson who provided thoughtful ideas for various projects throughout my tenure at UCSF. Chris Voigt created a laboratory that produces world-class synthetic biology and he taught me how to produce a clean story from my many experiments. Matt Jacobson provided an amazing role model for how to run a research group, how to be a productive mentor, and how to get almost anyone excited about working with him. Moreover, Professor Charles Lovett from Williams College really got me excited about science and research and without him I would not have made it to this point. Next, I would like to thank all of my professional contacts over the years who contributed to this work and to my success. Most notable is Christopher Anderson, who spent tireless hours teaching me how to efficiently navigate the laboratory space and apply less conventional, but more advanced, methods in order to speed up scientific progress. Sergio Wong, Petri Fast, and Katarzyna Bernacki provided excellent guidance and scientific discussion in the laboratory, on the ice, and while drinking a glass of really good, or really bad wine. Unfortunately, I do not have enough space to list all of my collaborators, but I wish to thank each of them and their role for helping me to achieve this goal. Most importantly, I would like to thank my wife Kim, my brother Matt, and my parents for their unending support throughout this process.

Chapter 1 is a paper published in the April, 2006 issue of *PLoS Computational Biology*. Arjun Narayanan and I split the research required for this paper. Arjun Narayanan, Matthew Jacobson, and I wrote and edited the final manuscript.

Chapter 2 is a paper published in the January, 2007 issue of the *Journal of the American Chemical Society*. I performed all of the implicit solvent calculations for this paper but had a limited role in the molecular dynamics (Daniel Mandell and Ilya Chorny) and Quantam Mechanics (Sergio Wong) portions. Dan Mandell wrote a majority of the paper and I helped to edit it along with Matthew Jacobson.

Chapter 3 is a paper published in the April, 2008 issue of *Chemistry and Biology*. I contributed a homology model for this work. Edwin Tan performed all of the experiments in this work and wrote a majority of the paper. Matthew Jacobson and I provided some suggestions during the editing process.

Chapter 4 is a finished manuscript that was recently submitted to *ACS Chemical Biology*. I contributed a few homology models for this work. Edwin Tan performed most of the experiments and wrote a majority of the paper. Again, Matthew Jacobson and I provided some suggestions during the editing process.

Chapter 5 is a book chapter that will be published in *Systems Biology and Biotechnology of E. coli*. Jeffrey Tabor and I wrote the chapter. Christopher Voigt, Jeffrey Tabor, and I edited the chapter.

Chapter 6 is a manuscript in preparation. I conducted all of the experiments and wrote the paper. Howard Salis contributed a computational model. Elizabeth Clarke contributed reporter plasmids from her work that made my work possible. Susan Miller helped to design the kinetic experiments and provided valuable feedback on the manuscript. Howard Salis, Elizabeth Clarke, Susan Miller, Christopher Voigt and I edited the manuscript.

Chapter 7 is a manuscript in preparation. Ryan Clark, Elizabeth Clarke, and I performed the experiments. Elizabeth Clarke and I wrote the paper. This research and this paper are not yet complete. Elizabeth Clarke plans to complete the research and submit the final manuscript.

This dissertation is dedicated to my wife Kim, my brother Matt,

and my parents, Nora and Bob.

Abstract

Cells use a variety of different mechanisms to sense and respond to the constantly changing environment. Single celled organisms use signaling pathways to find food, escape toxins, and protect themselves in dangerous or overcrowded environments. The failure of cell signaling in these organisms usually leads to the death of the cell. In a multi-celled organism, individual cells rely on signaling pathways to cooperate with the rest of the organism. A failure in cell to cell signaling in higher organisms may lead to the brief survival of the cell but the eventual failure and death of the organism as a whole. We chose to apply various methods to study cell signaling and communication in both prokaryotic and eukaryotic organisms. The first two approaches we took to this problem were entirely computational. We developed parameters for phosphorylated amino acids and used these parameters to predict structural changes as a function of phosphorylation. In addition, we showed that both phosphate charge and the geometry by which this phosphate interacts with other residues determine the energy gained or lost as a result of this interaction. Next, we joined computation and experiment to successfully predict agonists and antagonists for a G-protein coupled receptor. In a follow up study, we used the information gained on the G-protein coupled receptor to investigate selectivity among a set of similar receptors from different mammals. In the final section of this work, we use a mixture of computation and experiment to show that two component signaling pathways do not interfere with one another *in viv*o. We then apply this knowledge to deconstruct the bacterial chemotaxis pathway into two separate,

orthogonal signaling systems.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Background and Overview of Signaling Systems

Eukaryotic and prokaryotic cells respond to changes in both the intracellular and extracellular environment through a multitude of protein regulatory networks that function as complex circuits [1]. Sensor proteins in the cell membrane are linked to gene expression through a protein relay, which integrates multiple incoming signals with metabolic and cell state information [2-4]. The transfer of information amongst the relay proteins and to and from transcription factors mediates their activities through structural, dynamic, and oligomeric changes [5-8]. These changes act to perform computational operations on the input signals thereby allowing the cell to choose an appropriate response. Cellular signaling pathways must filter noise while responding quickly when the correct level of signal is present. Examples of signaling pathways range from the two component systems, used in single celled organisms to control chemotaxis and virulence, to the complex signaling cascade which controls the mammalian cell entry into mitosis [9-12].

Cells sense and respond to their external environment using a series of protein relay systems to transmit information from the cellular surface to the transcriptional machinery. Bacterial, yeast, and mammalian cells have a variety of sensors on the cell surface, analogous to the eyes and nose of higher organisms. Bacterial sensors activate in the presence of an amazing number of environmental and metabolic signals in order to survive environmental stress. They also have the ability to sense cell density in their immediate vicinity in order to produce more complicated synchronous behavior analogous to that of multi-cellular organisms. Mammalian sensors tend to respond to signals sent from other cells, either small molecules or hormones circulating throughout the body, in order to remain in sync with the entire organism. Within each cell, different genes encode different sensors. Therefore, cells with different genomic composition have the ability to sense different inputs.

All of the pathways described in the first two paragraphs allow the cell to adapt and survive various types of stress. While these pathways may seem vast and complicated, there is one thread tying most of them together. A simple molecule composed of one phosphorous atom and four oxygen atoms, phosphate. The phosphate group carries a negative charge at physiological pH, allowing it to impart a local or global change in the structure of regulatory proteins. Therefore, the currency for 70% of vital processes in the cell is phosphate. This work has seven chapters, all seemingly discussing topics that are in no way related to each other. However, all of the systems studied in this collection of work on signaling pathways use phosphate as a signaling molecule in some part of the process. In this work I discuss computational structural biology, G-protein coupled receptors, and synthetic biology, but all of the signaling systems evolved to use a common signal carrier, a phosphate group.

Prediction of Structural Changes as a Result of Phosphorylation

In Chapter 1, in collaboration with Arjun Narayan, we used a computational algorithm to determine the effect of placing a phosphate group on the end of an amino acid side chain in a subset of proteins. The algorithm used in this work was an extension of a loop prediction algorithm written by Matthew Jacobson. Instead of using the program to predict the structure of a protein, we used it to predict the local perturbations in structure upon phosphorylation. This is a direct extention of homology modeling, a technique that predicts the three dimensional structure of a protein based on the structure of a close relative. In homology modeling, most amino acids are the same and it is only small differences in the local environment that leads to differences in protein shape. In this work, we realize that the phosphorylated version of a protein can be treated as a homolog of the unphosphorylated version, having an almost identical amino acid composition except for the addition of a charged group that perturbs the energy landscape and, therefore, the three dimensional shape of the protein. In Chapter 1, we correctly predict protein conformational changes upon phosphorylation.

Investigating the properties of phosphorylated amino acid side chains

In Chapter 2, in collaboration with Dan Mandel, Ilya Chorny, and Sergio Wong, we use computation to investigate the properties of phosphate and its propensity to interact with different amino acids. Most importantly, we investigated the differences between arginine and lysine as hydrogen bond donors. Not only do these two amino acids have a different structure, which leads to a different charge density, but arginine has the ability to participate in a bidentate hydrogen bonding configuration with the phosphate while lysine does not. Moreover, we believe that phosphate can assume either a -2 or -1 charge state since the pKa of one of its protons is ~6.5, very close to physiological pH. This variety of charge states and variety of interaction geometries creates a versatile signaling molecule. Unlike in an experimental lab, we can use computation to fix geometries and determine interaction energies of the phosphate with different side chains, all while having different charges. We did this using a variety of tools including molecular dynamics based potentials of mean force, implicit solvent calculations, and quantam mechanics.

Designing agonists and antagonists of a G-protein coupled receptor

In Chapter 3, in collaboration with Edwin Tan and Thomas Scanlan, we designed small molecules that served as agonists and antagonists for an orphan G-protein coupled receptor (GPCR). GPCRs are Eukaryotic sensors that respond to small molecules and initiate a cellular response. In the case of the B2 aminergic GPCR that we investigated in this study, the native ligand is unknown. Moreover, at the time this work was completed, there was no known structure for a B2 aminergic receptor, making rational drug design a very hard problem. In a conversation between Eli and Edwin, they decided that using computation to develop a homology model for the B2 receptor could help to develop potent drugs to both activate and inhibit the action of this protein. We constructed the homology model of the B2 aminergic GPCR using the structure of bovine rhodopsin, another GPCR, as a model template. Edwin used this model as a guide to develop high affinity agonists and antagonists towards the B2 aminergic GPCR.

Molecular discrimination between the mouse and rat TAAR1

In Chapter 4, in collaboration with Edwin Tan, John Naylor, James Bunzow, David Grandy, and Thomas Scanlan, we investigated the molecular discrimination inherent in two GPCR orthologs in mouse and rat. Mice, rats, humans, and other mammals contain the trace amine associated receptor 1 (TAAR1). Although these proteins are related from one species to another, they have subtle differences in the binding site which cause differences in the recognition of one ligand over another. Once again, it was quite hard to rationally design molecules that could act on one form of the protein without an atomic level view of the binding site. Therefore, we constructed a homology model of the rat and mouse TAAR1 based on the recently solved structure of the B2 aminergenic receptor. Using these models, Edwin synthesized molecules that acted on one form of the protein and not on the other. Moreover, we were also able to rationalize the selectivity based on the homology models of the binding site.

The use of non-native parts and circuits to program *Escherichia coli*

In Chapter 5, in collaboration with Jeffrey Tabor and Christopher Voigt, we review the current state of the new field of synthetic biology, more specifically, the use of synthetic DNA to reprogram the E. coli bacterium. Synthetic biology is a new field conceived in the prospect of engineering new functions into a given cell by transferring genetic pieces for sensors and effectors (output generators) from various organisms into the cell. Instead of synthesizing new genes, which is beyond the scope of our current knowledge, combinations of DNA parts optimized by evolution are expanded, allowing them to perform many other tasks. Many different groups across the globe have developed interesting new sensors, circuits, and effectors to engineer function into the E. coli bacterium. In this review, we provide a healthy background for the reader and then lead him through interesting efforts in bacterial redesign.

Kinetic buffering between two-component systems prevents cross talk

In Chapter 6, in collaboration with Howard Salis, Elizabeth Clarke, and Christopher Voigt, we determine that two, two-component signaling pathways in the osmotic shock response system in *E. coli* are remarkable insulated. Two-component signaling systems are a ubiquitous method by which bacteria sense and responds to their environment. In this canonical system, a sensor histidine kinase responds to signal by autophosphorylating and then transferring this phosphate to a cytoplasmic response regulator protein. Upon phosphorylation, the response regulator binds to a specific DNA operator sequence and either activates or represses transcription of a single gene, or a set of genes. E. coli alone has at least 32 of these systems and a major question in the field is whether they interact and, if they do not, what is the mechanism that prevents unwanted phosphorylation, or cross talk. We determined that a combination of kinetic preference for a cognate histidine kinase, combined with the phosphatase activity inherent in the histidine kinase, prevents unwanted phosphorylation of many different response regulator proteins by a single histidine kinase and, conversely, the phosphorylation of a single response regulator by multiple histidine kinases.

Remote control of bacterial chemotaxis

In Chapter 7, in collaboration with Elizabeth Clarke, Ryan Clark, Matthew Eames, Tanja Kortemme, and Christopher Voigt, we redesigned the native *E. coli* chemotaxis machinery to respond to a signal of our choosing. Bacterial chemotaxis is a robust signaling system that guides the cell towards food and away from toxins. While this system is amazingly robust, it is unable to distinguish between different nutrients and toxins. This makes sense, as there was never any evolutionary pressure to migrate towards one nutrient over another. We created two orthogonal signaling pathways in the bacterium, one that responds to aspartate and another that responds to serine. Using a DNA encoded switch with a memory function, we are able to remotely control whether the bacterium swims towards one nutrient over another. In order to complete this work, we needed to validate a switch, validate its function in vivo, develop new signaling pathways, and then develop a bacterial strain compatible with the new switching and signaling machinery. Overall, this project was a great example of the power of synthetic biology.

Future Directions

Although all of the work described here spans different fields and different aspects of computational and experimental biology, the common thread is the study of signaling pathways involving phosphorylation. As one can see, there are many different approaches to study these pathways and all provide a different set of information. Computational studies can answer questions on a detailed atomic level, but have a harder time addressing global questions of protein interaction and specificity within larger signaling systems. Experiments answer these larger questions but often are lacking in extreme atomic description of the systems involved. Moving forward, a mixture of experiments and computation will provide powerful tools to answer important questions in biology, medicine, and genetic engineering. Unfortunately, very few researchers rely on both computational and experimental methods to evaluate biological systems. With hope, experimental and computational collaboration will increase to provide a better picture of the biological world.

Reference

1. Tyson, J.J., K.C. Chen, and B. Novak, *Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell.* Curr Opin Cell Biol, 2003. **15**(2): p. 221-31.
2. Fabret, C., V.A. Feher, and J.A. Hoch, *Two-component signal transduction in Bacillus subtilis: how one organism sees its world.* J Bacteriol, 1999. **181**(7): p. 1975-83.
3. Mizuno, T., *Compilation of all genes encoding two-component phosphotransfer signal transducers in the genome of Escherichia coli.* DNA Res, 1997. **4**(2): p. 161-8.
4. Mizuno, T., T. Kaneko, and S. Tabata, *Compilation of all genes encoding bacterial two-component signal transducers in the genome of the cyanobacterium, Synechocystis sp. strain PCC 6803*. DNA Res, 1996. **3**(6): p. 407-14.
5. Wang, L., et al., *Dissection of the functional and structural domains of phosphorelay histidine kinase A of Bacillus subtilis.* J Bacteriol, 2001. **183**(9): p. 2795-802.
6. Stephenson, S.J. and M. Perego, *Interaction surface of the Spo0A response regulator with the Spo0E phosphatase.* Mol Microbiol, 2002. **44**(6): p. 1455-67.
7. Varughese, K.I., *Molecular recognition of bacterial phosphorelay proteins.* Curr Opin Microbiol, 2002. **5**(2): p. 142-8.
8. Zapf, J., et al., *A transient interaction between two phosphorelay proteins trapped in a crystal lattice reveals the mechanism of molecular recognition and phosphotransfer in signal transduction.* Structure Fold Des, 2000. **8**(8): p. 851-62.
9. Msadek, T., *When the going gets tough: survival strategies and environmental signaling networks in Bacillus subtilis.* Trends Microbiol, 1999. **7**(5): p. 201-7.
10. Novak, R., et al., *Emergence of vancomycin tolerance in Streptococcus pneumoniae.* Nature, 1999. **399**(6736): p. 590-3.
11. Liu, X., et al., *The MAP kinase pathway is required for entry into mitosis and cell survival.* Oncogene, 2004. **23**(3): p. 763-76.
12. Hoffmaster, A.R. and T.M. Koehler, *Control of virulence gene expression in Bacillus anthracis.* J Appl Microbiol, 1999. **87**(2): p. 279-81.

Chapter 1: Conformational Changes in Protein Loops and Helices Induced by Post-Translational Phosphorylation

Eli S. Groban[1,2,‡], Arjun Narayanan[1,2,‡], and Matthew P. Jacobson[1,2]

[1] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America,

[2] Graduate Group in Biophysics, University of California San Francisco, San Francisco, California, United States of America

‡ These authors contributed equally to this work

Abstract

Post-translational phosphorylation is a ubiquitous mechanism for modulating protein activity and protein-protein interactions. In this work, we examine how phosphorylation can modulate the conformation of a protein by changing the energy landscape. We present a molecular mechanics method in which we phosphorylate proteins in silico and then predict how the conformation of the protein will change in response to phosphorylation. We apply this method to a test set comprised of proteins with both phosphorylated and non-phosphorylated crystal structures, and demonstrate that it is possible to predict localized phosphorylation-induced conformational changes, or the absence of conformational changes, with near-atomic accuracy in most cases. Examples of proteins used for testing our methods include kinases and prokaryotic response regulators. Through a detailed case study of cyclin-dependent kinase 2, we also illustrate how the computational methods can be used to provide new understanding of how phosphorylation drives conformational change, why substituting Glu or Asp for a phosphorylated amino acid does not always mimic the effects of phosphorylation, and how a phosphatase can "capture" a phosphorylated amino acid. This work illustrates how computational methods can be used to elucidate principles and mechanisms of post-translational phosphorylation, which can ultimately help to bridge the gap between the number of known sites of phosphorylation and the number of structures of phosphorylated proteins.

Synopsis

Many proteins are chemically modified after they are synthesized in the cell. These post-translational modifications can modulate the ability of a protein to perform chemical reactions and to interact with other proteins. At the cellular level, for example, these chemical modifications are critical for allowing the cell to respond to its environment and control its division. One of the most common mechanisms by which proteins can be modified is by phosphorylation—the addition of a phosphate group to an amino acid side chain of the protein. Thousands of proteins are known to be modified by phosphorylation, but only for a small minority of these do we have any detailed understanding of how the chemical modification regulates the function of the protein. The authors describe a computational method that can make testable predictions about the structural changes that occur in a protein induced by post-translational phosphorylation. Their results show that the method can produce structural models of the phosphorylated proteins with near-atomic accuracy, and provide insight into the energetics of conformational switches driven by phosphorylation. As such, the computational method complements experiments aimed at understanding the mechanisms of protein regulation by phosphorylation.

Introduction

Post-translational phosphorylation is a ubiquitous mechanism for cellular regulation, playing a role in such diverse processes as signal transduction, transport, cytoskeletal regulation, and metabolism. A variety of amino acids can be phosphorylated, but serine, threonine, and tyrosine are the most important sites of phosphorylation in eukaryotes, whereas histidine and aspartate play the central role in the “two-component” signaling pathways of prokaryotes. Several thousand sites of post-translational phosphorylation are now known [1], and this number will continue to grow quickly. Estimates of the fraction of proteins that are phosphorylated in vivo range as high as 30% [1]; higher values are associated with particular stages of the cell cycle or responses to external stimuli. Protein kinases catalyze post-translational phosphorylation, and many kinases are themselves regulated by phosphorylation, leading to complex signaling and regulatory networks. Kinases are targets of aggressive drug development efforts [2] aimed at treating cancer and other diseases such as diabetes.

Despite the huge amount of research related to post-translational phosphorylation, the detailed role that specific sites of post-translational phosphorylation play in the function of individual proteins remains poorly understood in most cases. Structural information is particularly limited, due in part to the difficulty of obtaining sufficient purified protein in a specific modification state. X-ray crystallography has determined atomic-resolution structures for a few tens of phosphorylated proteins [3], whereas nuclear magnetic resonance (NMR) experiments have elucidated structure and dynamics of a similar number of phosphorylated peptides and small proteins in solution [4–10]. Electron paramagnetic resonance (EPR) and circular dichroism experiments can also provide information on conformational change due to phosphorylation [11], but do not provide atomic detail. The number of known phosphorylation sites will almost certainly continue to grow much more quickly than the number of structurally well-characterized phosphoproteins.

We believe that computational studies can play a central role in elucidating principles and mechanisms of post-translational modification, and ultimately help to bridge the gap between the number of known sites of phosphorylation and the number of structures of phosphorylated proteins. Only a small number of computational modeling studies have been published on protein phosphorylation, particularly in relation to the thousands of molecular/cellular biology studies published every year on the topic. A few studies have focused on model systems, including studies from the McCammon group that examined the role of phosphorylation in stabilizing the N-termini of helices [12] and conformational/dynamical changes due to phosphorylation of a Ser residue in a tetrapeptide [13,14]. Luo et al. investigated the strengths of solvent-exposed salt bridges, including model systems for phosphate interactions with lysine and arginine, using a Generalized Born solvent model [15]. A few modeling studies of specific phosphorylated proteins have also been reported. Maurer et al. investigated the influence of phosphorylation on the docking of a peptide to bovine thrombin [16].

Zhou and Abagyan calculated the binding energy of phosphotyrosine-containing peptides to SH2 and PTB domains [17]. Several molecular dynamics [18–30] and homology modeling [27,31,32] studies have been reported involving phosphorylated proteins [18–20,24,25,31–33] or peptides [21–23,26,27,34]. A Car-Parinello study has been reported [35], as well as a creative application of docking algorithms to investigate conformational changes upon phosphorylation of the N-terminal tail in phenylalanine hydroxylase [36].

In this work, we examine how phosphorylation can modulate the conformation of a protein by changing the energy landscape. We limit our attention in this work to relatively small conformational changes involving a phosphorylated loop and its surroundings, and in one case, conformational changes in a helix and its surroundings. At physiological pH, the phosphate group predominantly carries a −2 charge, whereas the site of phosphorylation is neutral before modification (in the case of Ser, Thr, and Tyr). The conceptual foundation of our work is shown in Figure 1. We view the phosphate as a perturbation to the energy landscape of the protein. Our goal in this work is to predict the conformational changes caused by this perturbation. That is, given the structure (or an accurate model) of the unphosphorylated protein, can we predict the structure of the phosphorylated protein? The critical tools we need to understand and predict the structural effects of post-translational phosphorylation are as follows:

One tool that is needed is an energy function that captures the essential physics, especially for the modified residues. We use a molecular mechanics energy function consisting of the OPLS-AA force field and a Generalized Born implicit solvent model. The use of a physics-based energy function is important for two reasons. First, there are too few structures of phosphorylated proteins to make knowledge-based energy functions feasible. Second, conformational changes induced by phosphorylation are largely driven by the electrostatic perturbation induced by the phosphate group. We use the energy function not only to predict conformational changes induced phosphorylation, but also to better understand the key physical effects underlying these conformational changes, especially in our case study of cyclin dependent kinase 2 (CDK2).

Another necessary tool is sampling algorithms capable of exploring critical degrees of freedom. After phosphorylating a protein in silico, we need to explore the new energy landscape and identify the new global energy minimum, if in fact it is significantly different than that of the unphosphorylated protein. In principle, molecular dynamics could be used for this purpose, and in some cases, this type of strategy has been successful, in studies by ourselves and others [18–30]. However, the timescales for converting from the non-phospho to the phosphorylated form are unknown, and in fact are not physically meaningful because the in silico phosphorylation is alchemical, i.e., we do not attempt to mimic a kinase actually performing the phosphorylation. For relatively large conformational changes, or those involving high-energy barriers, e.g., Pro peptide bond isomerization, the relevant timescale could be quite long—microseconds or more. Instead, we use a strategy in which we have adapted algorithms that we previously developed for homology modeling, i.e., sampling methods for side chains [37,38], loops [39–41], and helices [42]. The essence of the approach is to combine dihedral angle sampling methods, which enable large energy barriers to be surmounted, with direct minimization to enumerate many local minima. We discuss our strategy in detail in Materials and Methods.

The key conclusion of this study is that our molecular mechanics–based methods appear to be capable of reproducing conformational changes induced by post-translational phosphorylation, with near-atomic resolution in most cases considered here, which are limited to relatively modest conformational changes and not, e.g., more drastic order–disorder transitions. This work thus represents a significant step toward a broadly applicable method for predicting structural effects of phosphorylation. Through a detailed case study of CDK2, we also illustrate how the computational methods can be used to provide new understanding of how phosphorylation drives conformational change, why substituting Glu or Asp for a phosphorylated amino acid does not always mimic the effects of phosphorylation, and how a phosphatase can "capture" a phosphorylated amino acid.

Results and Discussion

We first present an in-depth study of activation loop phosphorylation in CDK2, both to introduce the computational methods and to highlight the ability of these methods to gain new insights into how phosphorylation drives conformational changes. We then present a broader survey of our ability to predict phosphorylation-driven conformational changes.

Case Study: CDK2

CDK2 is a member of the cyclin-dependent kinase protein family [43], whose members play a central role in cell cycle regulation. CDK2 is activated through binding of cyclin A and post-translational phosphorylation of a threonine residue on the activation loop, which lies close to the catalytic site. Compared to other phosphorylated proteins, there is a wealth of structural information for CDK2 in both its phosphorylated and unphosphorylated forms [3]. In this case study, we examine the ability of our method to: (1) reconstruct and predict the active conformation of the activation loop in CDK2 bound to cyclin A [44]; (2) reconstruct the cyclin A dependence of CDK2 activation [44,45]; (3) discriminate between two potential phosphate localization sites in the CDK2/kinase-associated phosphatase (KAP) complex [46]; and (4) examine the effects of substituting Thr160 with a Glu, which results in only partial activation relative to phosphorylation of Thr160 [47].

A structural alignment and superposition of the active and inactive [48] forms of CDK2 bound to cyclin A reveals that the global Cα root mean square deviation

(RMSD) between the two structures is only 0.5 Å if residues 152–163 are removed from the calculation. The phosphorylatable threonine residue itself moves approximately 10 Å upon phosphorylation, as measured at the γ oxygen. In the unphosphorylated structure, Thr160 is well solvated and somewhat disordered (B-factors > 50). In the phosphorylated structure, pThr160 localizes to interact with a cluster of Arg residues (50, 126, and 150). We will refer to this and similar conformations as "active" conformations of the loop.

As described in Materials and Methods, we use a hierarchical loop prediction algorithm [40] based upon dihedral angle backbone sampling, rotamer-based side chain optimization, an all-atom force field and a Generalized Born solvation model [49,50] to predict the structural consequences of phosphorylation on loops and their surroundings. The hierarchical prediction algorithm allows us to quickly prune the conformational space and focus sampling on energetically favorable regions of conformation space (Figure 2). Unlike Monte Carlo and molecular dynamics sampling schemes, the algorithm itself has no knowledge of the starting conformation of the loop. The entire conformational space of the loop is sampled with varying sampling resolution in a hierarchical manner. This method has proven successful in recreating crystallographic conformations of unphosphorylated loops to loop lengths of 12 residues. In this paper, we expand on this loop prediction method to predict the structures of phosphorylated loops.

Before attempting to predict the phosphorylated structure from the unphosphorylated structure, we first ensure that we can predict the conformation of the phosphorylated loop with high accuracy when all other portions of the phosphorylated structure are taken from the crystal structure. We refer to this test as "loop reconstruction." This tests the suitability of the energy function for identifying the correct conformation of the phosphorylated loop and the suitability of the sampling function for generating native-like structures. Much of our analysis will focus on the accuracy with which the phosphate group is positioned, which is measured by the distance of the phosphorus atom to its crystallographic position. We also provide other measures of accuracy for the overall loop, including a backbone RMSD measure (calculated based on the N, Cα, and C atoms), and the RMSD calculated over all heavy atoms, including side chains. For all of these measures of accuracy, the predicted structure is first aligned to the reference crystal structure by a least-squares superposition of all Cα atoms, excluding the loop being simulated. Thus, structural differences outside of the loop region being predicted can affect the reported measures of accuracy.

For the complex of phosphorylated CDK2 with cyclin A, we are able to reproduce the conformation of the activation loop with less than 2 Å RMSD overall. The phosphate group is placed with particularly high accuracy, less than 1 Å error as measured by the position of the phosphorous atom (Figures 2 and 3, and Table 2). We also wish to highlight the differences between Figure 2B and 2C, which plot the energy of different local minima identified during the four-stage prediction algorithm versus the backbone RMSD (Figure 2B) and the error in the phosphate position (Figure 2C). In the energy versus backbone RMSD plot, an energy "funnel" is clearly evident. That is, the lowest energy identified decreases as the loop, as a whole, approaches native-like conformations. On the other hand, the phosphate group itself appears to respond to a narrow, deep energy well. That is, the phosphate locates the Arg cluster very early in the simulation (during the first of the four stages of the loop prediction algorithm), and remains there while the rest of the loop adjusts to the new position of the phosphate group.

Next, we examine our ability to predict the phosphorylated loop conformation starting from the unphosphorylated structure. Naively attempting to predict the conformation of the phosphorylated activation loop from the unphosphorylated structure, holding the nearby side chains fixed, results in a poor prediction, with an error in the position of the phosphorus atom of approximately 9 Å and a backbone RMSD of approximately 7 Å (Table 2). The reason for this failure is clear. A comparison of the phosphorylated and unphosphorylated structures reveals that the backbone of the phosphorylated conformation of the activation loop passes directly through the side chain of Tyr179 in the unphosphorylated form (Figure 4). This single misplaced side chain can prevent the loop from assuming the correct activated conformation. Thus, in this case, introduction of the phosphate group not only changes the activation loop conformation, but also rearranges the side chain hydrogen bonding network well beyond the loop itself.

Thus, sampling of loop conformations must be coupled with, at least, the sampling of side chain conformations in the vicinity of the loop to permit accurate prediction. As described in Materials and Methods, we have adopted a strategy in which side chains near the loop are removed during the loop build-up, and side chains both on the loop and its vicinity are optimized and energy minimized simultaneously for the representative loop from each cluster. Applying this method to the loop reconstruction test increases the error by a small amount. Overall, the error in the phosphorus position increased by about 1 Å, and the backbone RMSD for the activation loop increased an insignificant amount (Figure 5 and Table 2). We note that Figure 5A clearly shows two major "basins" of attraction for the phosphorylated amino acid. The one with the lower energy places the phosphate in contact with the Arg cluster; the higher energy state places the phosphate in solution. These two conformations, which we will call "active" and "inactive," respectively, will be examined further below.

When predicting the structural change upon phosphorylation of CDK2, this strategy permits Tyr179, as well as the three arginine residues that form salt bridges with pThr160, to re-optimize around the loop. The prediction with the simultaneous optimization of the loop and surrounding side chains results in a phosphorous atom error of 0.8 Å and a backbone RMSD of 2.9 Å (Figures 3 and 5B, Table 2). The phosphorus atom position is quite accurate, whereas the remainder of the loop may require the optimization of other structural elements to assume the active conformation. The prediction clearly captures the qualitative change in conformation upon phosphorylation, however. Additionally, we performed the above calculations again, except this time substituting the three Arg residues of the arginine cluster with Lys residues (unpublished results). In this case, the phosphate does not localize to the new “Lys cluster,” and we predict that the kinase would not be active. The preference of phosphorylated side chains for interacting with Arg, relative to Lys, has been noted previously, and we will explore this issue further in a separate publication.

CDK2 Phosphorylation in the Absence of Cyclin Binding

Cyclin A binding is required for the full activation of CDK2 because phosphorylated CDK2 without cyclin A exhibits only 0.3% of the activity of the fully active, phosphorylated, cyclin A–bound structure [3]. The crystal structure of phosphorylated CDK2 in the absence of cyclin A shows that the activation loop is disordered [45]. The loop thus likely adopts several dynamically interconverting conformations with the pThr fully solvated, instead of bound by the arginine cluster. The activation loop in unphosphorylated CDK2 in the absence of cyclin A is ordered, albeit with somewhat elevated B-factors (roughly 60).

The results of predicting the loop in the phosphorylated, cyclin A–unbound structure [45], showed that the phosphate did not localize to the Arg cluster, which is qualitatively consistent with phosphorylation not contributing to activation (results not shown). However, the low-energy predicted loop conformations showed the phosphate localized at a different position, and the results did not provide any evidence that the loop should in fact be disordered. This is due to a deliberate bias in the loop prediction algorithm, which was originally designed to predict well-structured loops, not ensembles of structures characterizing a disordered loop. A full solution will require an algorithm that obeys detailed balance and thus correctly treats the loop conformational entropy; we have recently developed a Monte Carlo version of the loop sampling algorithm that accomplishes this goal, and we will report results in due course. As a preliminary step, we have performed the calculations with one key modification to the existing loop prediction algorithm. Specifically, the algorithm constrains the loop conformations to be relatively close to the body of the protein using a simple distance cutoff; more than 99% of loops in the Protein Data Bank (PDB) satisfy this criterion [40]. However, this restriction prevents disordered loops from exploring conformations in which the loop is well solvated and makes few contacts to the remainder of the protein. By simply turning off this screening, the loop prediction for phosphorylated CDK2 in the absence of cyclin A does show evidence of diverse low-energy conformations (Figure 6B), in the sense that the 20 lowest-energy conformations are highly diverse; the phosphorylated Thr160 is well solvated in most of these. Turning off the screening does not significantly affect the results for the in silico phosphorylation of CDK2 with cyclin A bound: pThr160 still localizes strongly to the Arg cluster (Figure 6A) in all 20 of the lowest-energy conformations. Thus, in this case, the computer simulations can provide at least a qualitative prediction that phosphorylation of CDK2 in the absence of cyclin binding leads to disorder. More work is clearly required to test this capability in other cases of phosphorylation-induced disorder; the remainder of our results focuses exclusively on cases in which the phosphorylated loops are well ordered.

Substitution of Glu for the Phosphorylated Thr in the Activation Loop of CDK2

Among proteins that require phosphorylation for activation, there are many examples of a constitutively active mutant in which the phosphorylatable residue is substituted with either aspartate or glutamate [51–55]. Connell-Crowley et al., however, show that in the CDK2/cyclin A complex, a T160E mutation confers no additional activity to the complex [47]. To examine why, we use the active form of CDK2/cyclin A to create a CDK2/cyclin A T160E mutant in silico, and then use our prediction protocol to predict the conformation of the activation loop, residues 152–163 in the mutant protein. Using the protocol in which side chains surrounding the loop are optimized concurrently with the loop itself, the lowest-energy structure places the Cβ atom of Glu160 9 Å away from its position in the wild-type active loop (Figure 7). The side chain of Glu160 is fully solvated, and the overall loop conformation is reminiscent of the "inactive" conformations identified in the loop prediction results with pThr160. On the other hand, the second lowest-energy conformation places the side chain of Glu160 in a position that is highly analogous to the position of pThr160 in the CDK2/cyclin A complex (i.e., contacting the Arg cluster; see Figure 7). We refer to this as the "active" conformation.

As shown in Table 3, the energy difference between these two states is small for the T160E mutant, with the inactive conformation slightly more stable than the active conformation. For the pThr160 structure, as discussed above, the active structure is correctly predicted to be lower in energy, and the energy gap is somewhat larger (18 kcal/mol). These relatively small differences in overall energies of the active and inactive conformations mask very large differences in the individual energy components. Specifically, the active conformation is strongly favored by the Coulomb electrostatics term in the molecular mechanics energy (172 kcal/mol for the Glu160 case and 298 kcal/mol for the pThr160 case), primarily due to the interaction of the negatively charged side chain of Glu160 or pThr160 with the Arg cluster. On the other hand, the solvation free energy computed from the Generalized Born solvent model strongly favors the inactive conformations, primarily because both Glu160/pThr160 and the Arg cluster are much more solvent exposed in this conformation. The covalent term favors the inactive form; in other words, the active form involves significant internal strain which is relieved in the more open inactive state. In the Glu160 case, the energy differences in the Coulomb and solvation terms largely cancel, and the inactive state winds up slightly lower in energy. In the pThr160 case, the difference in the Coulomb term is significantly larger than the solvation term, causing the active state to be favored by a significant margin.

The differences between Glu160 and pThr160 of course derive primarily from the difference in the total charge on the side chains. The −2 charge on pThr nearly doubles the favorable Coulombic electrostatic attraction to the Arg cluster relative to the singly charged carboxylate group on Glu. The more highly charged pThr side chain also has a more favorable solvation free energy in the inactive form, but the overall difference in solvation free energy does not change by as great a factor, in part because greater solvation of the Arg cluster in the inactive state represents a significant contribution to the difference in solvation free energy, and this component is essentially independent of whether Glu or pThr is present at residue 160.

Phosphorylated CDK2 in Complex with Its Phosphatase

Finally, we examine conformational changes in the activation loop necessary for dephosphorylation to occur. KAP is responsible for the dephosphorylation, and subsequent inactivation, of CDK2 [3,46]. A crystal structure of the CDK2/KAP complex shows that pThr160 localizes to the N-terminus of a helix in the KAP protein, and interacts in that position with a single arginine side chain. The phosphate group must be stable enough in this position to allow the phosphatase to "pull" pThr160 from the Arg cluster on CDK2. Given the highly favorable electrostatic interactions between pThr160 and the three Arg side chains in the cluster, this did not seem intuitively obvious.

Nonetheless, the loop predictions in the CDK2/KAP complex did in fact place the phosphorous atom within 0.5 Å of its position in the crystal structure (Table 2); the lowest energy structure with pThr160 placed in contact with the Arg cluster is approximately 25 kcal/mol higher in energy. Table 3 helps to explain why. The phosphate placed close to the Arg cluster does in fact have much more favorable electrostatic interactions, as measured by the Coulombic term in the molecular mechanics energy function (111 kcal/mol difference). However, the solvation term strongly favors the pThr160 in its correct position in contact with KAP, due largely to the improved solvent accessibility of the Arg cluster after the pThr160 is removed from contact. In addition, the pThr160 is not as buried in the KAP complex. Finally, the covalent terms again indicate significantly higher internal strain in the activation loop when the pThr160 is in contact with the Arg cluster, and much of this strain is relieved when the loop adopts the phosphatase-bound conformation (this might be referred to as a spring-loaded mechanism).

Predicting Conformational Changes of Loops in Response to Phosphorylation

Encouraged by our results on CDK2, we have performed similar calculations on a larger, more diverse test set (Table 1). Our results show that in all cases of loop reconstruction, we reproduce the structure of the phosphorylated loop with high accuracy (Table 4). We reconstruct seven of the loops to within 1 Å backbone RMSD with respect to the original crystal structure and predict the phosphorus atom to within 0.5 Å of its crystallographic position. Although our predictions are not error free, the relatively small magnitude of the error gives us some confidence that we can predict the correct conformation of a phosphorylated loop using our hierarchical sampling methodology, all-atom force field, and Generalized Born solvation model.

In cases in which suitable unphosphorylated and phosphorylated structures of the same protein exist, we sought to extend this methodology to the more realistic situation in which we wish to predict the phosphorylated structure from the unphosphorylated structure. In these cases, we start from the crystal structure of the unphosphorylated protein, phosphorylate the residue of interest in silico, and predict the structure of the phosphorylated loop and its surroundings. We refer to this test as "loop prediction." In this test, sampling only the conformation of the loop itself (Table 5, "loop only") is successful in some cases, such as histidine-containing phosphocarrier protein (HPr), in which the conformational changes upon phosphorylation are small. In other cases, this strategy performs poorly because the phosphorylation induces substantial conformational changes in the surroundings of the loop, as discussed for CDK2 above. In order to capture these essential structural rearrangements in the region surrounding the loop, we also optimize the conformations of all side chains that have at least one atom within 4.5 Å of the loop, as described in Materials and Methods. This method permits all cases, excepting extracellular signal-regulated kinase 2 (ERK2), to be predicted with near-atomic accuracy, especially the phosphate group itself. Encouragingly, our molecular mechanics methods appear to be capable of not only predicting when significant conformational changes occur, as in CDK2, but also when phosphorylation induces little conformational change, as in HPr.

ERK2 is unsuccessful because our existing sampling methods do not permit us to sample global structural changes like the changes in domain orientation that occur upon dual phosphorylation of ERK2 [3]. Given that we can accurately reconstruct the active conformation of the ERK2 activation loop, this inability to account for the domain reorientation is likely the cause of the poor prediction of the activation loop structure. This test case highlights a fundamental limitation of our current prediction strategy; the only solution is to expand the sampling to explicitly sample domain orientations, either in a manner analogous to our helix prediction methods or by using normal modes or other methods capable of sampling low-frequency, global motions of the protein.

The SpoIIAA case also requires explanation. The initial loop prediction and reconstruction tests on this case, using an unprotonated (−2) phosphate group as in the other cases, gave very poor results, listed in parentheses in Tables 4 and 5. Examining the structure of the phosphorylated protein suggested a possible explanation. The pSer58 side chain lies within approximately 5 Å of the side chain of Asp56. The pKa of the phosphorylated amino acids is approximately 6, which implies that the predominant charge state is generally −2 under physiological conditions, and the good results we obtain for the other test cases using a −2 phosphate seem to confirm this assumption. However, it is well known that the pKa's of titratable groups can be significantly shifted in macromolecules due to desolvation and/or the electrostatic field from the rest of the protein. In this case, the close proximity of Asp56 is likely to shift the pKa lower and favor the −1, protonated state of the phosphate group on pSer58 (the pH used in the crystallization conditions is 6.5). In fact, performing the loop predictions with a protonated phosphate group leads to excellent results. In current work, we are implementing a new version of our molecular mechanics algorithms that allows automatic identification of the optimal protonation state for phosphorylated amino acids and other titratable groups (especially His), using a thermodynamic cycle analogous to that employed by the numerous Poisson-Boltzmann based pKa prediction algorithms [56].

Predicting Conformational Changes of Helices in Response to Phosphorylation

To test our ability to model conformational responses that extend beyond a single loop and its surroundings, we consider the repacking of helices that is known to occur upon phosphorylation of response regulator proteins of prokaryotes [3]. The response regulator proteins are the second part of the two-component signaling system that bacteria use to sense and respond to their environment. In most two-component systems, a sensor histidine kinase autophosphorylates a His upon sensing signal and transfers this phosphate to an Asp on a response regulator. Upon phosphorylation, certain structural changes occur, most notably a helix shift and loop rearrangement, that ultimately allow the response regulator to activate transcription of a set of genes.

We used our existing helix prediction algorithm to model the structural changes of the response regulator protein FixJ (from *Sinorhizobium meliloti* [57]), specifically, the loop-helix-loop region that undergoes a helix shift and loop rearrangements. Another response regulator protein, Spo0A, has crystal structures in both the phosphorylated and unphosphorylated forms, but the crystal structure for the unphosphorylated form of Spo0A is a domain-swapped dimer, solved at a pH of 4.5. As discussed in Materials and Methods, the helix prediction algorithm employs rigid body sampling for the helix (residues 87–94) combined with the hierarchical loop sampling algorithm for predicting the connecting loops [42] (residues 79–86 and 95–99). The overall backbone RMSD for reconstructing this region in the phosphorylated crystal structure is 0.5 Å, with most of the error arising from the first flanking loop, which is longer, closest to the phosphate, and more exposed to solvent

(Table 4 and Figure 8). The results from in silico phosphorylation of the unphosphorylated structure are also quite good, with an overall RMSD of 1.6 Å.

Conclusion

We have described our initial efforts to predict and understand how the structure of a protein is modulated by post-translational phosphorylation. We believe that this work has practical significance in that we demonstrate that it is possible to make testable predictions concerning the structure of phosphorylated proteins, given the structure of the unphosphorylated protein and a known site of phosphorylation. In this work, we have restricted our efforts to predicting relatively modest, localized conformational changes, and we have assumed knowledge of which portions of the protein undergo significant conformational change (those portions closest to the phosphorylated amino acid). Despite these limitations, this modeling technology can be used to create hypotheses about mechanisms of regulation by phosphorylation that can then be tested experimentally, e.g., by site-directed mutagenesis. Applications of this type are underway. In addition, our in-depth case study of CDK2 illustrates how the computational methods can be used to obtain new insights into the energetic underpinnings of phosphorylation-induced conformational change even in a well-studied system.

Materials and Methods

Dataset selection.

We searched the PDB [58] for phosphorylated protein structures determined by X-ray crystallography (with better than 2.5 Å resolution) that are phosphorylated on well-ordered loop structures that were less than 15 residues in length. We exclude structures in which phosphorylation causes a large, global rearrangement of the protein structure, such as a hinge-bending movement or domain rearrangement, as is the case with glycogen phosphorylase [59] and insulin receptor tyrosine kinase [60,61]. In order to test the limitations of our method, we include one test case, ERK2, in which phosphorylation causes a domain rearrangement [3]. We also include a prokaryotic response regulator, FixJ, in which phosphorylation of an Asp induces a significant conformational change in the orientation of a helix [57,62]; this case is successfully treated with an extension of our methods described below. The test set is listed in Table 1.

We determined the loop length that we would predict for each phosphorylated structure using visual inspection. In the cases in which crystallographic structures are available for both the unphosphorylated and phosphorylated protein, these structures were superimposed and the residues to be predicted were defined as the portion of the loop that deviated in the superposition. For the reconstruction of phosphorylated structures without knowledge of the unphosphorylated form, the loop residues to be optimized were determined using a combination of visual inspection of secondary structure and crystallographic B-factors.

Molecular mechanics energy function.

All energy calculations use the OPLS-AA force field [37,63,64] and the Surface Generalized Born (SGB) model of solvation [49,50]. The molecular mechanics energy function represents electrostatics by a relatively simple model of fixed atomic partial charges interacting through the Coulomb approximation. The solvent model captures key effects of desolvation with relatively modest computational expense. Despite the simplicity of the energy function (i.e., it neglects polarizability contributions to electrostatics, and implicit solvent models have well-known limitations), it performs well in predicting conformations of phosphorylated loops (see Results).

The force field parameters for the phosphorylated amino acids were generated by an automated atom-typing algorithm provided in the Impact software package. The atomic partial charges for the phosphorylated amino acid side chains were adjusted slightly from the default values by performing quantum chemistry calculations. The partial charges for phosphoserine (pSer) and phosphothreonine (pThr) were taken from previous work by Wong et al. [65], whereas charges for phosphotyrosine (pTyr) and phosphoaspartate (pAsp) were determined in their −2 and −1 charge states by performing quantum mechanical calculations with the software program Jaguar [66]. Methyl-benzyl-phosphate was used to represent the pTyr side chain, and acetyl phosphate was used for pAsp. Geometry optimization of the phosphate ion was carried out at the HF/6-31G** level, incorporating a condensed-phase environment via a self-consistent reaction field (SCRF) algorithm

[67,68]. Single point calculations were performed at the LMP2/cc-pvtz(-f) level, also with SCRF treatment of solvation. Electrostatic potential fitting was used to determine the partial charges. The atomic partial charges for all four phosphorylated amino acids are provided in Tables S1–S4.

Loop prediction methodology.

This study uses the method of Jacobson et al. [40] for predicting loop conformations. In brief, the loop prediction methodology uses an ab initio dihedral sampling scheme to enumerate conformations of the loop backbone that are free from steric clashes. Other methods have employed similar dihedral angle sampling schemes, including ICM [69], CONGEN [70], and the work of DePristo, et al [71]. Unlike Monte Carlo and Molecular Dynamics sampling schemes, the algorithm itself has no knowledge of the starting conformation of the loop, and therefore, does not start predictions based on a starting structure of the loop. These closed backbone conformations are clustered, and a single member of each cluster is then selected for side chain addition and optimization, followed by complete energy minimization. The lowest energy structure is selected as the output of the loop prediction algorithm. The algorithm also permits explicit treatment of crystal packing. We have used this capability in this work (in the "loop reconstruction" cases, as described below), but did not identify any clear crystal-packing artifacts relevant to the conformations of the phosphorylated loops.

The Jacobson et al. paper [40] describes a hierarchical refinement procedure in which multiple iterations of the loop prediction algorithm are used to reduce errors caused by insufficient sampling. The parameters in this scheme have been slightly modified in this study due to the inclusion of some rather long loops (up to 15 residues) in our test set. The first stage allows for unrestrained sampling. After this stage is complete, the top ten lowest energy structures are passed to a first refinement stage in which more extensive sampling is performed around these low-energy basins. Specifically, loop conformations in this stage are only retained if all of the Cα atoms in the loop are within 10 Å of the starting loop structure (which is one of the ten lowest-energy structures from the initial stage). The five resulting lowest-energy structures from this stage are subjected to a second round of refinement, in which the maximum Cα deviation is restricted to 5 Å. Finally, the five lowest-energy structures from this stage are subjected to a third and final round of refinement, in which the maximum Cα deviation is restricted to 2.5 Å. In all, this procedure provides a rank-ordered list consisting of about 250 loops and their associated energies in the context of the full protein. The final prediction is the loop conformation with the lowest energy.

Most published tests of loop prediction methods, including the Jacobson et al. paper [40], evaluate their success by the ability to reconstruct loops in a native protein structure. In these tests, all portions of the protein other than the loop in question remain in the native conformation during the simulation. Success of a prediction methodology in such a test is an important prerequisite for more realistic applications. We perform loop reconstruction tests in this work to assess the ability of the molecular mechanics energy function to identify correct conformations of phosphorylated loops, and to ensure that the sampling methods are sufficient to generate near-native conformations. However, predicting conformational changes induced by phosphorylation, i.e., by phosphorylating a protein in silico, is qualitatively more challenging. In the cases we consider, the sites of phosphorylation are located on loops, and most of the conformational change is localized to that loop. However, there is always some degree of conformational rearrangement in the vicinity of the loop, especially in the conformations of side chains contacting it. Similarly, predicting the conformation of a loop in a homology model is more challenging than reconstructing a loop in a native protein structure because the environment surrounding any given loop in a homology model contains errors that can affect the loop prediction accuracy. We address this issue by performing rotamer optimization and minimization of side chains in the immediate vicinity of the loop concurrently with the optimization of the side chains on the loop itself. In the test set presented here, this strategy performs well in predicting local structural changes, despite the fact that there are also some changes in backbone conformation in the surroundings. We speculate that small changes in the conformations of the surrounding side chains can compensate for not explicitly allowing backbone relaxation. We also use this strategy in the control studies of reconstructing phosphorylated loops in which, interestingly, it sometimes improves the accuracy.

Helix prediction methodology.

The helix prediction algorithm used in this paper is based on the work of Li et al [42]. Briefly, the helix backbone is treated as a rigid body and sampled in six degrees of freedom (three translations and three rotations), and the two flanking loops are sampled using the loop prediction algorithm described above. Again, the method broadly samples the possible configurations of the helix and surrounding loops, independent of any starting configuration, and then hierarchically samples more finely around low-energy basins. As with the loop prediction algorithm, side chains on the loop-helix-loop region, and the surroundings if desired, are sampled using a rotamer-based optimization algorithm.

Development of "rotamer" libraries for phosphorylated residues.

The loop and helix prediction algorithms require the use of a rotamer library for sampling side chain conformations. However, the number of phosphorylated residues in the PDB is insufficient to construct these libraries by statistical analysis of observed side chain conformations. Instead, we obtained a rotamer library by exhaustively exploring the energy landscape of the side chains of phosphorylated amino acid dipeptides, and retaining all conformations below an energy threshold that ensures inclusion of all experimentally observed rotamers. We use the same energy function as in the rest of the work, i.e., OPLS-AA/SGB. The pSer and pThr side chains are sampled at 10° resolution, whereas pTyr and pAsp are sampled at 30° resolution, due to the larger number of rotatable bonds. The total number of rotamers are 802 for pSer, 501 for pThr, 858 for pTyr, and 1,008 for pAsp. These values do not include the rotation of the phosphate group about the phosphoester bond, which is sampled uniformly at the same resolution as the other bonds. The rotamer libraries are provided in Tables S5–S8.

Accession Numbers

The Protein Data Bank (http://www.rcsb.org/pdb) accession numbers for the proteins discussed in this paper are as follows: β PGM, (PDB ID: 1LVH); CDK2, phosphorylated (PDB ID: 1JST); CDK2, unphosphorylated (PDB ID: 1FIN); CDK2/KAP complex (PDB ID: 1FQ1); cyclin A (PDB ID: 1B39); ERK2, phosphorylated (PDB ID: 2ERK); ERK2, unphosphorylated (PDB ID: 1ERK); FixJ, phosphorylated (PDB ID: 1D5W); FixJ, unphosphorylated (PDB ID: 1DCK); GLM, (PDB ID: 1MKI); Hpr, phosphorylated (PDB ID: 1FU0); Hpr, unphosphorylated (PDB ID: 1PTF); LCK, (PDB ID: 3LCK); Pim1, (PDB ID: 2BIK); PMM, (PDB ID: 1K35); Psp, phosphorylated (PDB ID: 1J97); Psp, unphosphorylated (PDB ID: 1L7O); P38 γ, (PDB ID: 1CM8); SpoIIAA, phosphorylated (PDB ID: 1H4X); and SpoIIAA, unphosphorylated (PDB ID: 1H4Z); Spo0A, (PDB ID: 1QMP).

References

1. Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res 27: 237–239.
2. Bridges AJ (2001) Chemical inhibitors of protein kinases. Chem Rev 101: 2541–2571.
3. Johnson LN, Lewis RJ (2001) Structural basis for control by phosphorylation. Chem Rev 101: 2209–2242.
4. Keane NE, Chavanieu A, Quirk PG, Evans JS, Levine BA, et al. (1994) Structural determinants of substrate selection by the human insulin-receptor protein-yyrosine kinase. Eur J Biochem 226: 525–536.
5. Quirk PG, Patchell VB, Colyer J, Drago GA, Gao Y (1996) Conformational effects of serine phosphorylation in phospholamban peptides. Eur J Biochem 236: 85–91.
6. Quirk PG, Patchell VB, Gao Y, Levine BA, Perry SV (1995) Sequential phosphorylation of adjacent serine residues on the N-terminal region of cardiac troponin-I: Structure-activity implications of ordered phosphorylation. FEBS Lett 370: 175–178.
7. Chavanieu A, Keane NE, Quirk PG, Levine BA, Calas B, et al. (1994) Phosphorylation effects on flanking charged residues: Structural implications for signal-transduction in protein-kinases. Eur J Biochem 224: 115–123.
8. Tholey A, Lindemann A, Kinzel V, Reed J (1999) Direct effects of phosphorylation on the preferred backbone conformation of peptides: A nuclear magnetic resonance study. Biophys J 76: 76–87.
9. Tholey A, Pipkorn R, Bossemeyer D, Kinzel V, Reed J (2001) Influence of myristoylation, phosphorylation, and deamidation on the structural behavior of the N-terminus of the catalytic subunit of CAMP-dependent protein kinase. Biochemistry 40: 225–231.
10. Coadou G, Evrard-Todeschi N, Gharbi-Benarous J, Benarous R, Girault JP (2001) Conformational analysis by NMR and molecular modelling of the 41-62 hydrophilic region of HIV-1 encoded virus protein U (Vpu). Effect of the phosphorylation on sites 52 and 56. Comptes Rendus De L Academie Des Sciences Serie Ii Fascicule C-Chimie 4: 751–758.
11. Hwang I, Thorgeirsson T, Lee J, Kustu S, Shin YK (1999) Physical evidence for a phosphorylation-dependent conformational change in the enhancer-binding protein NtrC. Proc Natl Acad U S A 96: 4880–4885.
12. Smart JL, McCammon JA (1999) Phosphorylation stabilizes the N-termini of alpha-helices. Biopolymers 49: 225–233.
13. Shen TY, Wong CF, McCammon JA (2001) Atomistic Brownian dynamics simulation of peptide phosphorylation. J Am Chem Soc 123: 9107–9111.
14. Hamelberg D, Shen T, McCammon JA (2005) Phosphorylation effects on cis/trans isomerization and the backbone conformation of serine-proline motifs: Accelerated molecular dynamics analysis. J Am Chem Soc 127: 1969–1974.

15. Luo R, David L, Hung H, Devaney J, Gilson MK (1999) Strength of solvent-exposed salt-bridges. J Phys Chem B 103: 727–736.
16. Maurer MC, Peng JL, An SS, Trosset JY, Henschen-Edman A, et al. (1998) Structural examination of the influence of phosphorylation on the binding of fibrinopeptide A to bovine thrombin. Biochemistry 37: 5888–5902.
17. Zhou Y, Abagyan R (1998) How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains. Fold Des 3: 513–522.
18. Feng MH, Philippopoulos M, MacKerell AD, Lim C (1996) Structural characterization of the phosphotyrosine binding region of a high-affinity SH2 domain-phosphopeptide complex by molecular dynamics simulation and chemical shift calculations. J Am Chem Soc 118: 11265–11277.
19. Stultz CM, Levin AD, Edelman ER (2002) Phosphorylation-induced conformational changes in a mitogen-activated protein kinase substrate. Implications for tyrosine hydroxylase activation. J Biol Chem 277: 47653–47661.
20. Roche P, Mouawad L, Perahia D, Samama JP, Kahn D (2002) Molecular dynamics of the FixJ receiver domain: movement of the beta 4-alpha 4 loop correlates with the in and out flip of Phe101. Protein Sci 11: 2622–2630.
21. Peters GH, Frimurer TM, Andersen JN, Olsen OH (2000) Molecular dynamics simulations of protein-tyrosine phosphatase 1B. II. Substrate-enzyme interactions and dynamics. Biophys J 78: 2191–2200.
22. Schneider ML, Post CB (1995) Solution structure of a band 3 peptide inhibitor bound to aldolase: A proposed mechanism for regulating binding by tyrosine phosphorylation. Biochemistry 34: 16574–16584.
23. Wozniak-Celmer E, Oldziej S, Ciarkowski J (2001) Theoretical models of catalytic domains of protein phosphatases 1 and 2A with Zn2+ and Mn2+ metal dications and putative bioligands in their catalytic centers. Acta Biochim Pol 48: 35–52.
24. Tomoo K, Shen X, Okabe K, Nozoe Y, Fukuhara S, et al. (2003) Structural features of human initiation factor 4E, studied by X-ray crystal analyses and molecular dynamics simulations. J Mol Biol 328: 365–383.
25. Young MA, Gonfloni S, Superti-Furga G, Roux B, Kuriyan J (2001) Dynamic coupling between the SH2 and SH3 domains of c-Src and hck underlies their inactivation by C-terminal tyrosine phosphorylation. Cell 105: 115–126.
26. Phan-Chan-Du A, Hemmerlin C, Krikorian D, Sakarellos-Daitsiotis M, Tsikaris V, et al. (2003) Solution conformation of the antibody-bound tyrosine phosphorylation site of the nicotinic acetylcholine receptor beta-subunit in its phosphorylated and nonphosphorylated states. Biochemistry 42: 7371–7380.
27. Fu Z, Aronoff-Spencer E, Backer JM, Gerfen GJ (2003) The structure of the inter-SH2 domain of class IA phosphoinositide 3-kinase determined by site-directed spin labeling EPR and homology modeling. Proc Natl Acad U S A 100: 3275–3280.
28. Powell DW, Rane MJ, Joughin BA, Kalmukova R, Hong JH, et al. (2003) Proteomic identification of 14-3-3zeta as a mitogen-activated protein kinase-activated protein kinase 2 substrate: Role in dimer formation and ligand binding. Mol Cell Biol 23: 5376–5387.

29. Mendieta J, Gago F (2004) In silico activation of Src tyrosine kinase reveals the molecular basis for intramolecular autophosphorylation. J Mol Graph Model 23: 189–198.
30. Bartova I, Otyepka M, Kriz Z, Koca J (2004) Activation and inhibition of cyclin-dependent kinase-2 by phosphorylation: A molecular dynamics study reveals the functional importance of the glycine-rich loop. Protein Sci 13: 1449–1457.
31. Peng B, Morrice NA, Groenen LC, Wettenhall REH (1996) Phosphorylation events associated with different states of activation of a hepatic cardiolipin protease-activated protein kinase: Structural identity to the protein kinase N-type protein kinases. J Biol Chem 271: 32233–32240.
32. Miller M, Ginalski K, Lesyng B, Nakaigawa N, Schmidt L, et al. (2001) Structural basis of oncogenic activation caused by point mutations in the kinase domain of the MET proto-oncogene: Modeling studies. Proteins 44: 32–43.
33. Kubala M, Obsil T, Obsilova V, Lansky Z, Amler E (2004) Protein modeling combined with spectroscopic techniques: An attractive quick alternative to obtain structural information. Physiol Res 53 (Suppl 1): S187–197.
34. Hansson T, Nordlund P, Aqvist J (1997) Energetics of nucleophile activation in a protein tyrosine phosphatase. J Mol Biol 265: 118–127.
35. Dal Peraro M, Alber F, Carloni P (2001) Ser133 phosphate-KIX interactions in the CREB-CBP complex: An ab initio molecular dynamics study. Eur Biophys J 30: 75–81.
36. Miranda FF, Teigen K, Thorolfsson M, Svebak RM, Knappskog PM, et al. (2002) Phosphorylation and mutations of Ser(16) in human phenylalanine hydroxylase - Kinetic and structural effects. J Biol Chem 277: 40937–40943.
37. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS (2002) Force field validation using protein side chain prediction. J Phys Chem B 106: 11673–11680.
38. Jacobson MP, Friesner RA, Xiang ZX, Honig B (2002) On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 320: 597–608.
39. Coutsias EA, Seok CL, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. J Comput Chem 25: 510–528.
40. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, et al. (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55: 351–367.
41. Guallar V, Jacobson MP, McDermott A, Friesner RA (2004) Computational modeling of the catalytic reaction in triose phosphate isomerase. J Mol Biol 337: 227–239.
42. Li X, Jacobson MP, Friesner RA (2004) High-resolution prediction of protein helix positions and orientations. Proteins 55: 368–382.
43. Harper JW, Adams PD (2001) Cyclin-dependent kinases. Chem Rev 101: 2511–2526.
44. Russo AA, Jeffrey PD, Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. Nat Struct Biol 3: 696–700.

45. Brown NR, Noble MEM, Lawrie AM, Morris MC, Tunnah P, et al. (1999) Effects of phosphorylation of threonine 160 on cyclin-dependent kinase 2 structure and activity. J Biol Chem 274: 8746–8756.
46. Song HW, Hanlon N, Brown NR, Noble MEM, Johnson LN, et al. (2001) Phosphoprotein-protein interactions revealed by the crystal structure of kinase-associated phosphatase in complex with phosphoCDK2. Mol Cell 7: 615–626.
47. Connell-Crowley L, Solomon MJ, Wei N, Harper JW (1993) Phosphorylation independent activation of human cyclin-dependent kinase 2 by cyclin A in vitro. Mol Biol Cell 4: 79–92.
48. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, et al. (1995) Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 376: 313–320.
49. Gallicchio E, Zhang LY, Levy RM (2002) The SGB/NP hydration free energy model based on the Surface Generalized Born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem 23: 517–529.
50. Ghosh A, Rapp CS, Friesner RA (1998) Generalized Born model based on a surface integral formulation. J Phys Chem B 102: 10983–10990.
51. Charbon G, Breunig KD, Wattiez R, Vandenhaute J, Noel-Georis I (2004) Key role of Ser562/661 in Snf1-dependent regulation of Cat8p in *Saccharomyces cerevisiae* and *Kluyveromyces lactis*. Mol Cell Biol 24: 4083–4091.
52. Kassenbrock CK, Anderson SM (2004) Regulation of ubiquitin protein ligase activity in c-Cbl by phosphorylation-induced conformational change and constitutive activation by tyrosine to glutamate point mutations. J Biol Chem 279: 28017–28027.
53. Huang W, Erikson RL (1994) Constitutive activation of Mek1 by mutation of serine phosphorylation sites. Proc Natl Acad Sci U S A 91: 8960–8963.
54. Klose KE, Weiss DS, Kustu S (1993) Glutamate at the site of phosphorylation of nitrogen-regulatory protein NTRC mimics aspartyl-phosphate and activates the protein. J Mol Biol 232: 67–78.
55. McCabe TJ, Fulton D, Roman LJ, Sessa WC (2000) Enhanced electron flux and reduced calmodulin dissociation may explain “calcium-independent” eNOS activation by phosphorylation. J Biol Chem 275: 6123–6128.
56. Georgescu RE, Alexov EG, Gunner MR (2002) Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. Biophys J 83: 1731–1748.
57. Birck C, Mourey L, Gouet P, Fabry B, Schumacher J, et al. (1999) Conformational changes induced by phosphorylation of the FixJ receiver domain. Structure Fold Des 7: 1505–1515.
58. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
59. Sprang SR, Acharya KR, Goldsmith EJ, Stuart DI, Varvill K, et al. (1988) Structural changes in glycogen phosphorylase induced by phosphorylation. Nature 336: 215–221.

60. Hubbard SR (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. EMBO J 16: 5572–5581.
61. Hubbard SR, Wei L, Ellis L, Hendrickson WA (1994) Crystal structure of the tyrosine kinase domain of the human insulin receptor. Nature 372: 746–754.
62. Gouet P, Fabry B, Guillet V, Birck C, Mourey L, et al. (1999) Structural transitions in the FixJ receiver domain. Structure Fold Des 7: 1517–1526.
63. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105: 6474–6487.
64. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118: 11225–11236.
65. Wong SE, Bernacki K, Jacobson M (2005) Competition between intramolecular hydrogen bonds and solvation in phosphorylated peptides: Simulations with explicit and implicit solvent. J Phys Chem B 109: 5249–5258.
66. Schrodinger LLC (2002) Jaguar 5.0 [computer program]. Portland (Oregon): Schrodinger, L.L.C.
67. Marten B, Kim K, Cortis C, Friesner RA, Murphy RB, et al. (1996) New model for calculation of solvation free energies: Correction of self-consistent reaction field continuum dielectric theory for short-range hydrogen-bonding effects. J Phys Chem 100: 11775–11788.
68. Tannor DJ, Marten B, Murphy R, Friesner RA, Sitkoff D, et al. (1994) Accurate first principles calculation of molecular charge distributions and solvation energies from ab initio quantum mechanics and continuum dielectric theory. J Am Chem Soc 116: 11875–11882.
69. Abagyan R, Totrov M (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 235: 983–1002.
70. Bruccoleri RE, Karplus M (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26: 137–168.
71. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: Efficient generation of accurate, representative ensembles. Proteins 51: 41–55.
72. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.
73. Audette GF, Engelmann R, Hengstenberg W, Deutscher J, Hayakawa K, et al. (2000) The 1.9 A resolution structure of phospho-serine 46 HPr from *Enterococcus faecalis*. J Mol Biol 303: 545–553.
74. Jia Z, Vandonselaar M, Hengstenberg W, Quail JW, Delbaere LT (1994) The 1.6 A structure of histidine-containing phosphotransfer protein HPr from *Streptococcus faecalis*. J Mol Biol 236: 1341–1355.

75. Seavers PR, Lewis RJ, Brannigan JA, Verschueren KH, Murshudov GN, et al. (2001) Structure of the *Bacillus* cell fate determinant SpoIIAA in phosphorylated and unphosphorylated forms. Structure 9: 605–614.
76. Cho H, Wang W, Kim R, Yokota H, Damo S, et al. (2001) BeF(3)(-) acts as a phosphate analog in proteins phosphorylated on aspartate: Structure of a BeF(3)(-) complex with phosphoserine phosphatase. Proc Natl Acad Sci U S A 98: 8525–8530.
77. Wang W, Cho HS, Kim R, Jancarik J, Yokota H, et al. (2002) Structural characterization of the reaction pathway in phosphoserine phosphatase: Crystallographic "snapshots" of intermediate states. J Mol Biol 319: 421–431.
78. Canagarajah BJ, Khokhlatchev A, Cobb MH, Goldsmith EJ (1997) Activation mechanism of the MAP kinase ERK2 by dual phosphorylation. Cell 90: 859–869.
79. Zhang F, Strand A, Robbins D, Cobb MH, Goldsmith EJ (1994) Atomic structure of the MAP kinase ERK2 at 2.3 A resolution. Nature 367: 704–711.
80. Lewis RJ, Brannigan JA, Muchova K, Barak I, Wilkinson AJ (1999) Phosphorylated aspartate in the structure of a response regulator protein. J Mol Biol 294: 9–15.
81. Regni C, Tipton PA, Beamer LJ (2002) Crystal structure of PMM/PGM: An enzyme in the biosynthetic pathway of *P. aeruginosa* virulence factors. Structure 10: 269–279.
82. Bellon S, Fitzgibbon MJ, Fox T, Hsiao HM, Wilson KP (1999) The structure of phosphorylated p38gamma is monomeric and reveals a conserved activation-loop conformation. Structure Fold Des 7: 1057–1065.
83. Lahiri SD, Zhang G, Dunaway-Mariano D, Allen KN (2002) Caught in the act: The structure of phosphorylated beta-phosphoglucomutase from *Lactococcus lactis*. Biochemistry 41: 8351–8359.
84. Yamaguchi H, Hendrickson WA (1996) Structural basis for activation of human lymphocyte kinase Lck upon tyrosine phosphorylation. Nature 384: 484–489.

Chapter 1: Table 1. Proteins Used in Our Test Set

Protein	Phosphorylated Structures (PDB ID)	Phosphorylated Residue(s)	Residues Predicted	Unphosphorylated Structures (PDB ID)
Hpr	1FU0 [73]	Ser46	42–48	1PTF [74]
SpoIIAA	1H4X [75]	Ser57	49–58	1H4Z [75]
Psp	1J97 [76]	Asp11	10–18	1L7O [77]
CDK2	1JST [44]	Thr160	152–163	1FIN [48]
ERK2	2ERK [78]	Thr183, Tyr185	172–186	1ERK [79]
FixJ	1D5W [57]	Asp54	79–99	1DCK [62]
Spo0A	1QMP [80]	Asp55	54–63	
GLM	1MKI	Ser74	64–75	
PMM	1K35 [81]	Ser108	105–114	
Pim1	2BIK	Ser261	251–262	
P38 γ	1CM8 [82]	Thr183,	177–189	
β PGM	1LVH [83]	Asp8	7–16	
LCK	3LCK [84]	Tyr394	389–402	

Caption for Table 1.

Unphosphorylated structures are listed only for the cases in which we attempt to predict the phosphorylated structure starting from the unphosphorylated structure. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the phosphorylated amino acid is not located within this region. We include this case to highlight the ability to extend our methods to regions larger than loops and their immediate surroundings.

β PGM, β-phosphoglucomutase; GLM, probable glutaminase from *Bacillus subtilis;* LCK, human lymphocyte kinase; Pim1, proto-oncogene serine/threonine protein kinase pim1; PMM, phosphomannomutase; Psp, phosphoserine phosphatase; Spo0A, stage 0 sporulation factor A.

Chapter 1: Table 2. Results from the CDK2 Case Study

Case	Deviations from Phosphorylated Crystal Structure (Å)			
	Loop Only		Loop + Surroundings	
	P	BB	P	BB
Reconstruction	0.6	1.7	1.7	1.8
Prediction	8.8	6.7	0.8	2.9
CDK2/KAP	1.1	2.3	0.5	1.3

Caption for Table 2.

The case study is summarized using two measures: the deviation of the phosphate group from its position in the crystal structure of phosphorylated CDK2 (“P”), and the overall backbone RMSD of the predicted loop compared to the same structure (“BB”). For the CDK2/KAP case, the reference structure is that of phosphorylated CDK2 in complex with KAP (1fq1), whereas the other cases are compared to the crystal structure of phosphorylated CDK2 bound to cyclin A (1JST). “Loop Only” refers to predicting only the loop in question; these results are provided only for comparison. “Loop + Surroundings” refers to the protocol in which we predict the loop residues as well as side chains of residues within 4.5 Å of any atom in the loop; these are the results that should be used to evaluate the success of the method. The different prediction cases are as follows. “Reconstruction” refers to predicting residues 152–163 in the crystal structure of phosphorylated CDK2 in complex with cyclin A. “Prediction” refers to the prediction of residues 152–163 after in silico phosphorylation of unphosphorylated CDK2/cyclin A (1fin). Finally, “CKD2/KAP” refers to predicting residues 155–165 in the structure of the phosphorylated activation loop of CDK2 when in complex with its phosphatase, KAP. It is important to note that the prediction methodology being used has no knowledge of the starting structure of the loop.

Chapter 1: Table 3. Energy Differences (kcal/mol) between “Active” and “Inactive” Conformations of the Activation Loop

	$\Delta E = E_{active} - E_{inactive}$		
Energy Component	pThr160	T160E	CDK2/KAP
Covalent	15.3	16.4	10.8
Coulomb	−298.2	−172.3	−110.9
Lennard-Jones	23.8	−2.4	30.1
Solvation	240.6	161.2	94.6
Total	−18.5	2.9	24.6

Caption for Table 3

The energy differences are broken down by different components of the molecular mechanics energy: covalent (bonds, bond-angles, and torsions), the Coulombic electrostatic term, Lennard-Jones, and the solvation free energy. The active form is defined by pThr160 being localized near the Arg cluster. In the pThr160 case, this is the lowest-energy–predicted structure, i.e., as represented in Table 2. In the Thr160Glu (T160E) substitution, the lowest-energy structure has the Glu side chain pointed out into solution; we refer to this as the “inactive” conformation. The second highest-energy structure has the Glu side chain in a position analogous to the pThr side chain, and we refer to this as the “active” structure for T160E. For CDK2/KAP, active/inactive take on different meanings. The active conformation is analogous to the other active structures, i.e., the phosphate group is localized at the Arg cluster.

The inactive conformation corresponds to lowest-energy conformation, in which the phosphate is correctly localized near the N-terminus of a helix in KAP.

Chapter 1: Table 4. Reconstruction of Phosphorylated Loop Conformations

		Loop Only				Loop + Surroundings			
Protein	Loop Length	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range
HPr	7	2.9	0.4	1.7	0.2–3.3	2.8	0.3	0.8	0.2–3.5
SpoIIAA	10	0.7 (2.1)	0.3 (0.4)	1.4	0.2–3.8	0.6 (2.2)	0.3 (0.8)	1.2	0.2–6.6
Psp	9	0.2	0.2	0.5	0.2–3.2	0.3	0.2	1.0	0.2–4.0
CDK2	12	0.6	1.7	3.5	1.2–8.4	1.7	1.8	1.7	1.0–12.5
ERK2	15	0.3/0.3	3.8	5.2	1.8–10.8	2.3/1.0	4.7	3.2	2.6–14.2
FixJ	8/10/5	0.1	0.5	1.2	0.4–4.7	0.3	0.8	1.0	0.4–4.7
Spo0A	10	0.3	0.2	0.8	0.2–3.1	0.8	0.4	0.9	0.2–4.1
GLM	12	0.1	0.2	1.0	0.2–6.2	0.2	2.1	1.6	0.4–5.8
PMM	10	0.3	0.4	1.0	0.3–5.4	0.3	1.0	1.1	0.4–5.9
Pim1	12	1.2	1.2	2.9	1.2–14.9	1.3	1.4	2.2	0.4–14.8
P38γ	12	0.4/0.5	2.3	2.8	1.1–5.7	0.7/5.6	1.4	1.6	1.2–8.2
β-PGM	10	0.9	0.6	1.3	0.2–7.1	0.6	0.2	0.6	0.2–8.8
LCK	14	1.2	3.0	3.3	1.6–10.1	2.5	2.7	2.4	1.8–11.9

Caption for Table 4

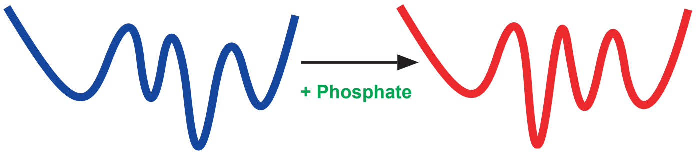
"Reconstruction" refers to the ab initio prediction of loop residues with the rest of the protein in the phosphorylated conformation. "Loop Only" refers to the prediction of only the loop in question, whereas "Loop + Surroundings" refers to the prediction of the loop in question in addition to surrounding side chains with at least one atom within 4.5 Å of any atom in the loop. The accuracy of the prediction is assessed by three measures: the error in the predicted phosphorus atom position ("P Error"), a backbone RMSD measure ("BB RMSD"), and the RMSD for all heavy atoms included in the prediction (“Heavy RMSD”). See the text for details. The “Sampled BB RMSD Range” column indicates the range of backbone RMSDs of sampled loops, emphasizing that our ab initio build-up procedure samples many different conformations of the loop to be predicted, and the final prediction is the lowest-energy conformation sampled. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the backbone and heavy atom RMSDs are calculated over the entire 23-residue region. For SpoIIAA, the predictions using a phosphate group with a −2 charge (as in the other test cases) gave poor results (in parentheses), whereas using a protonated phosphate group gave much better results; see text for details. Abbreviations used for the protein names are defined in Table 1.

Chapter 1: Table 5. Prediction of Phosphorylated Loop Conformations

		Loop Only				Loop + Surroundings			
Protein	Loop Length	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range	P Error	BB RMSD	Heavy RMSD	Sampled BB RMSD Range
HPr	7	1.0	0.4	1.3	0.4–6.1	1.0	0.5	1.1	0.4–7.0
SpoIIAA	10	0.3 (6.1)	0.7 (5.7)	1.7	0.6–7.1	0.9 (6.5)	0.8 (3.9)	1.5	0.5–6.7
Psp	9	2.1	0.5	0.8	0.4–5.9	0.4	0.4	1.0	0.4–5.9
CDK2	12	8.8	6.7	8.1	3.2–9.9	0.8	2.9	3.6	2.0–11.0
ERK2	15	15.4/14.8	8.8	10.7	3.6–10.9	15.4 / 3.0	6.7	5.0	3.6–13.1
FixJ	8/10/5	1.3	1.1	1.5	0.4–4.8	1.0	1.6	1.7	0.4–4.8

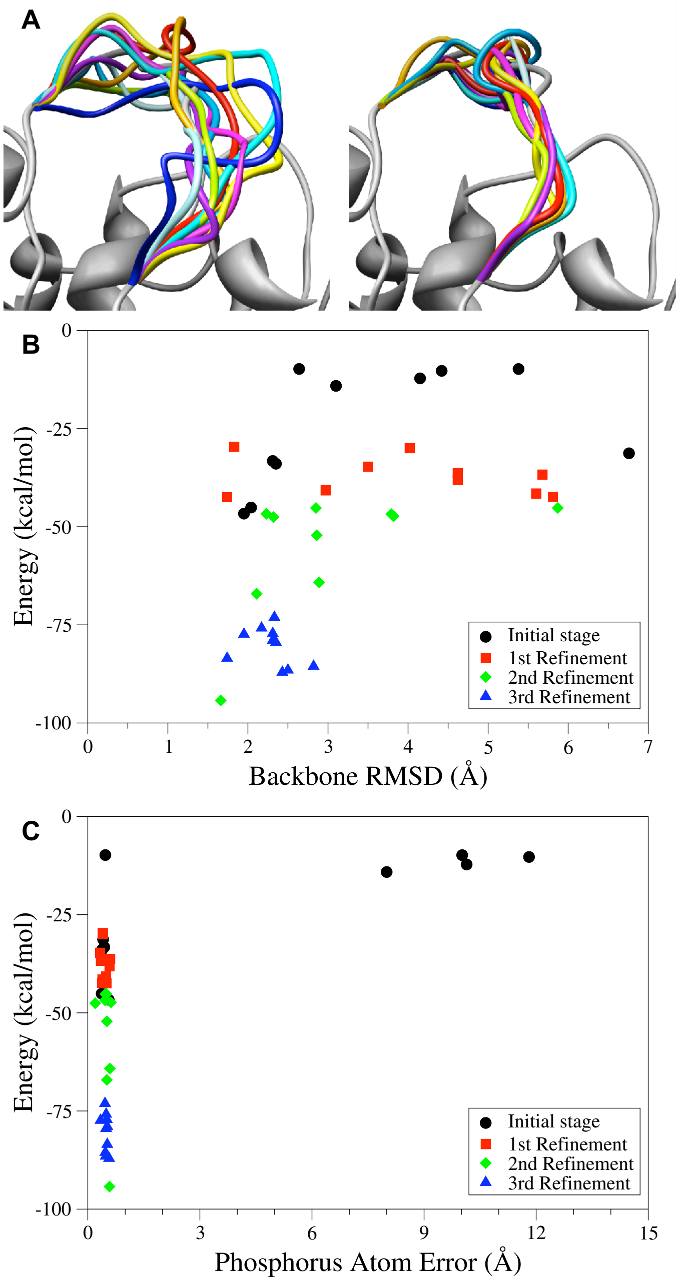
Caption for Table 5

"Prediction" refers to the ab initio prediction of loop residues with the rest of the protein held in the unphosphorylated conformation. "Loop Only" refers to the prediction of only the loop in question; these results are provided only for comparison. "Loop + Surroundings" refers to the prediction of the loop in question in addition to surrounding side chains with at least one atom within 4.5 Å of any atom in the loop; these are the results that should be used in evaluating the success of our method. Definitions of the columns are provided in Table 4, and abbreviations used for the protein names are defined in Table 1. In the case of FixJ, the predicted region contains a helix (ten residues) and two surrounding loops (eight and five residues), and the backbone and heavy atom RMSDs are calculated over the entire 23-residue region. For SpoIIAA, the predictions using a phosphate group with a −2 charge (as in the other test cases) gave poor results (in parentheses), whereas using a protonated phosphate group gave much better results; see text for details. ERK2, as described in the text, undergoes a domain reorientation as well as a loop movement, which is likely the cause of the poor predictions. The improvement of the CDK2 prediction when optimization of surrounding side chains is performed in concert with the loop prediction highlights the need to incorporate side chain flexibility in the calculation to successfully predict the phosphorylated loop conformation from the unphosphorylated structure.



Chapter 1: Figure 1. Phosphorylation Can Perturb the Energy Landscape of a Protein to Cause Changes in Conformation and Dynamics

This figure is a visual representation of such changes. In this work, we predict the structural change of proteins due to phosphorylation by locating the new global energy minimum of the energy surface.



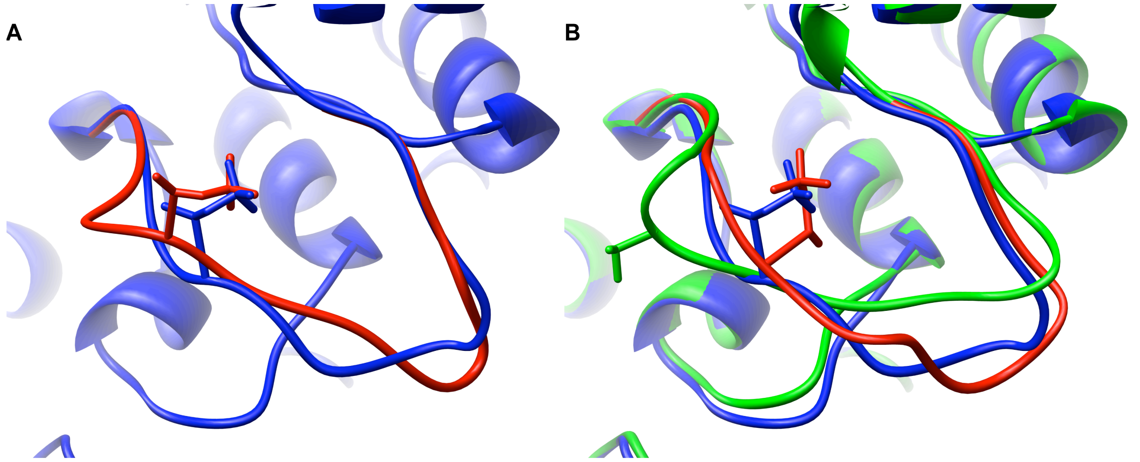
Chapter 1: Figure 2. Example of the Hierarchical Loop Prediction, Applied to Reconstructing the Phosphorylated Activation Loop of CDK2/Cyclin A

In this example, only the loop is predicted, and the remainder of the protein is held rigid in its crystallographic conformation.

(A) Backbone traces for the ten lowest-energy loops sampled in the initial build-up stage (left) and the third and final refinement stage (right) of the hierarchical loop prediction protocol (figures prepared with Chimera [72]).

(B) Energies as a function of backbone RMSD values of the 20 lowest-energy loops sampled in the four stages of the loop prediction protocol.

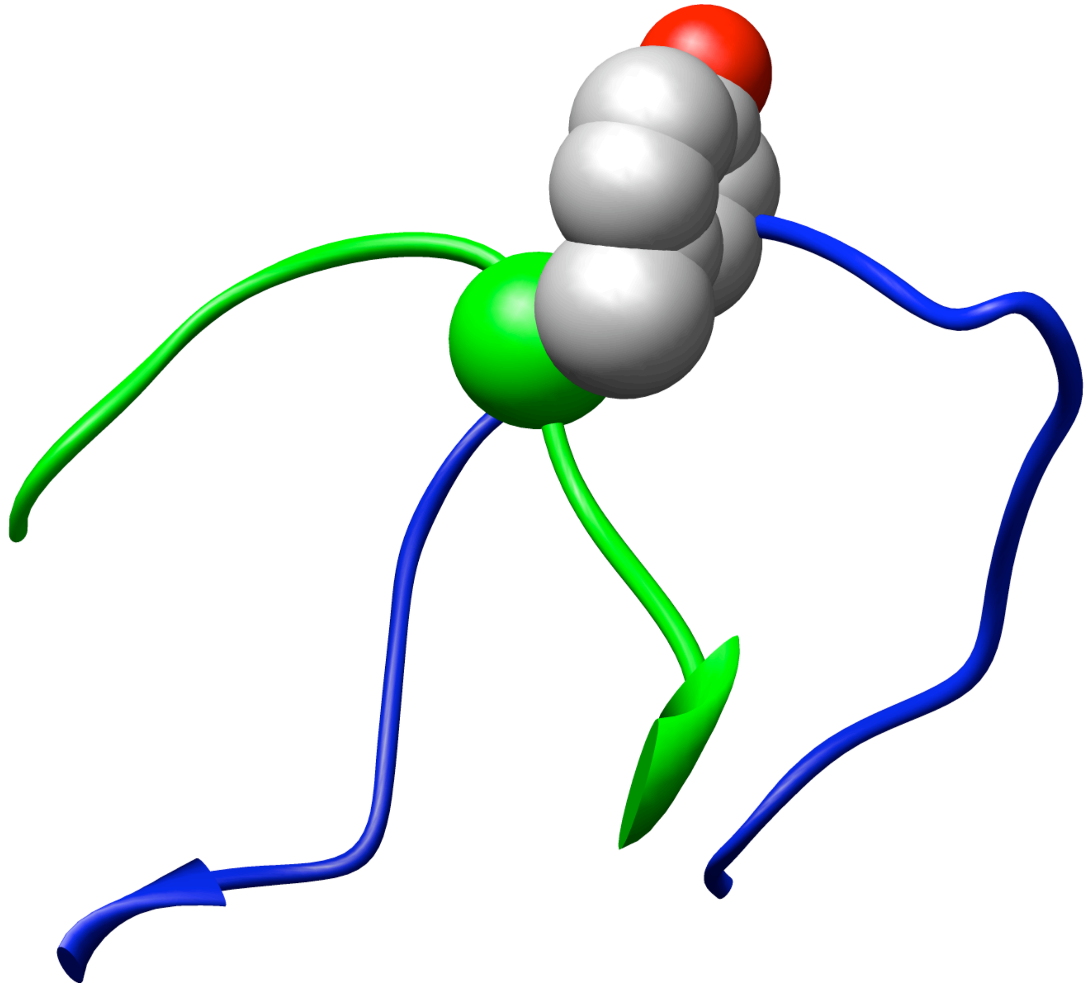
(C) Energies as a function of the error in the phosphorus atom position, relative to the crystal structure, for the 20 lowest-energy loops sampled in the four stages of the loop prediction protocol.



Chapter 1: Figure 3. Loop Reconstruction and Prediction in CDK2/Cyclin A

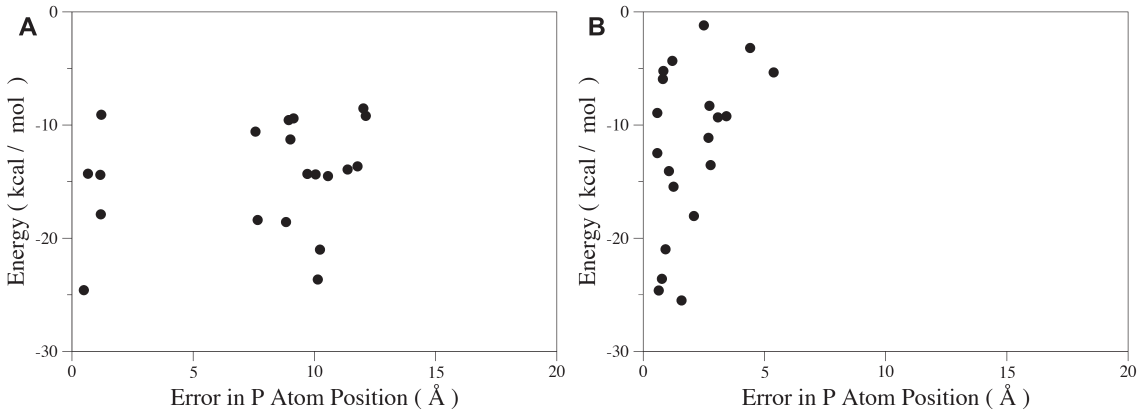
(A) Reconstruction: Crystal structure of phosphorylated CDK2/cyclin A (blue) and the predicted loop structure (red). The starting structure for the prediction was the phosphorylated structure, with the only difference being the loop region.

(B) Prediction: Crystal structure of phosphorylated CDK2/cyclin A (blue), the crystal structure of the unphosphorylated CDK2/cyclin A (green), and the predicted loop structure upon in silico phosphorylation of the unphosphorylated CDK2/cyclin A structure (red). Figures prepared with Chimera [72].



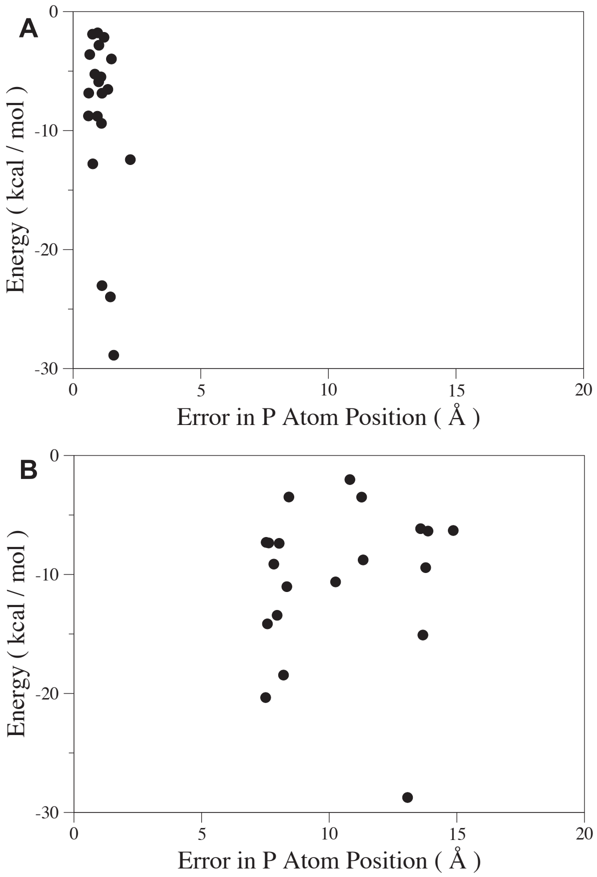
Chapter 1: Figure 4. Detailed View of One Portion of the Structural Superposition between the Phosphorylatedand Unphosphorylated Crystal Structures of CDK2/Cyclin A

The phosphorylated activation loop (blue) passes through the middle of Tyr179 (CPK) when inserted into the non-phosphorylated structure (green) of CDK2. Figure prepared with Chimera [72].



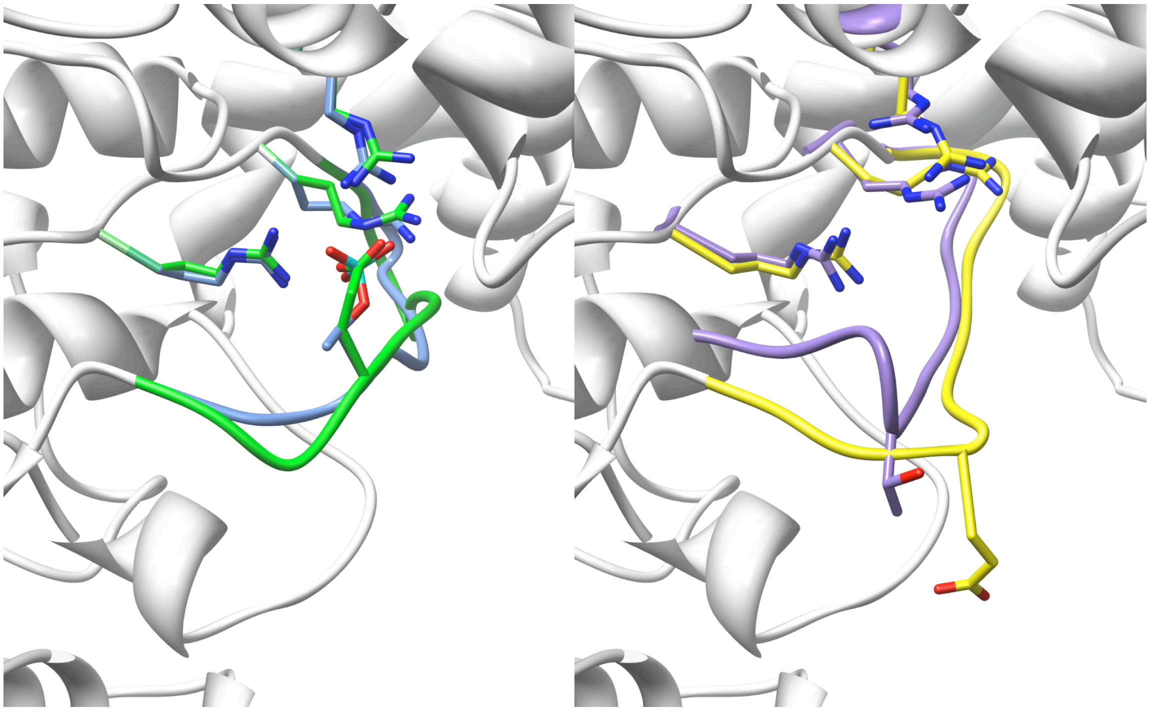
Chapter 1: Figure 5. CDK2 Case Study

The 20 lowest-energy loops predicted for (A) reconstruction of residues 152–163 in phosphorylated, cyclin-bound CDK2; and (B) prediction of the structure of residues 152–163 upon in silico phosphorylation of Thr160 in the unphosphorylated, cyclin-bound CDK2. In each case, the loop and surrounding side chains are optimized simultaneously as described in Materials and Methods.



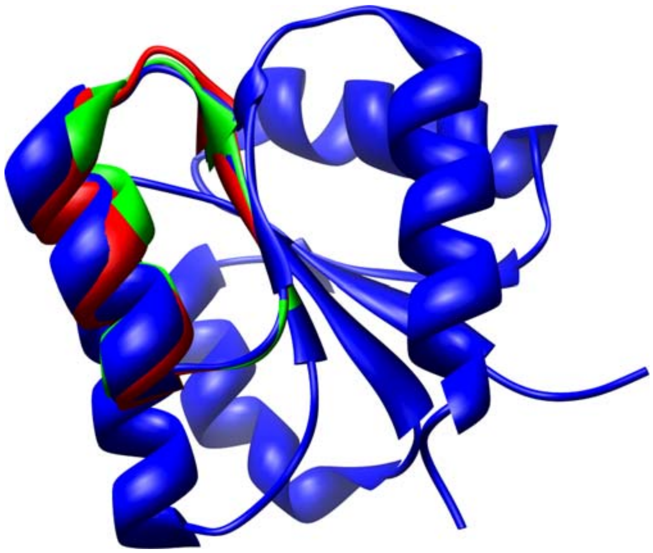
Chapter 1: Figure 6. Differences in Conformational Predictions for Phosphorylated CDK2 with and without Cyclin A Bound

The 20 lowest-energy loops are considered for predicting the conformation of residues 152–163 upon in silico phosphorylation of Thr160 (A) in cyclin-bound CDK2 and (B) CDK2 in the absence of cyclin A. In (B), the *x*-axis represents the deviation of the phosphate from its position in the fully activated CDK2/cyclin A complex; the activation loop and pThr160 are disordered in the crystal structure of phosphorylated CDK2 in the absence of cyclin A. These calculations were performed with a consistent set of parameters that do not bias the results toward well-ordered loop structures. In the absence of cyclin A, the phosphate does not localize to the Arg cluster as in the cyclin bound case, and the 20 lowest-energy structures show considerable diversity in conformation.



Chapter 1: Figure 7. Active and Inactive Conformations of the CDK2 Activation Loop

Left: Blue represents the crystal structures of the phospho-CDK2/cyclin A complex, and green represents the T160E-predicted active-like conformation. The Arg cluster is shown in stick representation. The carboxylate group of Glu160 and the phosphate group of pThr160 are almost exactly superimposed. Right: Purple represents the crystal structure of unphosphorylated CDK2/cyclin A, and yellow represents the predicted inactive conformation of T160E. These two structures are qualitatively similar in that Thr160 and Glu160 both point out into solvent and the Arg cluster is better solvated. Figures prepared with Chimera [72]



Chapter 1: Figure 8. Helix Reconstruction and Prediction of FixJ

The loop-helix-loop region of FixJ was predicted starting from either the phosphorylated structure (blue) or the unphosphorylated structure (unpublished results). The reconstruction (starting from the phosphorylated structure) is in red, and the prediction (starting from the unphosphorylated structure) is in green. Figure prepared with Chimera [72].

Chapter 1: Supplementary Table 1: Partial atomic charges employed for pThr.

Atom ID	Phosphorylated Threonine Charge in -2 state	Charge in -1 state
CB	0.280	0.280
HB	-0.030	-0.030
CG2	-0.210	-0.210
1HG2	0.060	0.060
2HG2	0.060	0.060
3HG2	0.060	0.060
OG1	-0.700	-0.500
P	1.570	1.530
O1P	-1.030	-1.000
O2P	-1.030	-1.000
O3P	-1.030	-0.750
HO	NA	0.500

Chapter 1: Supplementary Table 2: Partial atomic charges employed for pSer.

Atom ID	Phosphorylated Serine Charge in -2 state	Charge in -1 state
CB	0.280	0.280
1HB	-0.030	0.035
2HB	-0.030	0.035
OG	-0.700	-0.600
P	1.570	1.500
O1P	-1.030	-1.000
O2P	-1.030	-1.000
O3P	-1.030	-0.750
HO	NA	0.500

Chapter 1: Supplementary Table 3: Partial atomic charges employed for pTyr.

	Phosphorylated Tyrosine	
Atom ID	Charge in -2 state	Charge in -1 state
CB	-0.005	-0.005
1HB	0.060	0.060
2HB	0.060	0.060
CG	-0.115	-0.115
CD1	-0.115	-0.115
HD1	0.115	0.115
CE1	-0.115	-0.115
HE1	0.115	0.115
CZ	0.220	0.240
CE2	-0.115	-0.115
HE2	0.115	0.115
CD2	-0.115	-0.115
HD2	0.115	0.115
OH	-0.700	-0.600
P	1.570	1.550
O1P	-1.030	-0.970
O2P	-1.030	-0.970
O3P	-1.030	-0.750
HO	NA	0.500

Chapter 1: Supplementary Table 4: Partial atomic charges employed for pAsp.

	Phosphorylated Aspartate	
Atom ID	Charge in -2 state	Charge in -1 state
CB	-0.470	-0.440
1HB	0.220	0.230
2HB	0.220	0.230
CG	0.940	0.870
OD1	-0.590	-0.540
OD2	-0.730	-0.670
P	1.470	1.500
O1P	-1.020	-0.950
O2P	-1.020	-0.950
O3P	-1.020	-0.790
HO	NA	0.510

Chapter 1: Supplementary Table 5: Rotamer library used for sampling pThr. Numbers should be multiplied by 10 to obtain the torsion angle in degrees. The columns represent the sampled heavy-atom torsion angles, other than the rotation about the O-P bond in the phosphate group, which is sampled uniformly. c_1 represents rotation about the C_a-C_b bond, and c_2 is the rotation about the C_b-O_g bond.

-18	-18
-18	6
-18	7
-18	8
-18	9
-18	10
-18	11
-18	12
-18	13
-18	14
-18	15
-18	16
-18	17
-18	18
-17	-18
-17	6
-17	7
-17	8
-17	9
-17	10
-17	11
-17	12
-17	13
-17	14
-17	15
-17	16
-17	17
-17	18
-16	-18
-16	6
-16	7
-16	8
-16	9
-16	10
-16	11
-16	12
-16	13
-16	14
-16	15
-16	16
-16	17
-16	18
-15	-18
-15	6
-15	7
-15	8
-15	9
-15	10
-15	11
-15	12
-15	13
-15	14
-15	15
-15	16
-15	17
-15	18
-14	-18
-14	5
-14	6
-14	7
-14	8
-14	9
-14	10
-14	11
-14	12
-14	13
-14	14
-14	15
-14	16
-14	17
-14	18
-13	-18
-13	5
-13	6
-13	7
-13	8
-13	9
-13	10
-13	11
-13	12
-13	13
-13	14
-13	15
-13	16
-13	17
-13	18
-12	-18
-12	5
-12	6
-12	7
-12	8
-12	9
-12	10
-12	11
-12	12
-12	13
-12	14
-12	15
-12	16
-12	17
-12	18
-11	-18
-11	-6
-11	-5
-11	6
-11	7
-11	8
-11	9
-11	10
-11	11
-11	12
-11	13
-11	14
-11	15
-11	16
-11	17
-11	18
-10	-18
-10	-6
-10	7
-10	8
-10	9
-10	10
-10	11
-10	12
-10	13
-10	14
-10	15
-10	16
-10	17
-10	18
-9	-18
-9	-6
-9	7
-9	8
-9	9
-9	10
-9	11
-9	12
-9	13
-9	14
-9	15
-9	16
-9	17
-9	18
-8	-18
-8	-6
-8	8
-8	9
-8	10
-8	11
-8	12
-8	13
-8	14
-8	15
-8	16
-8	17
-8	18
-7	-18
-7	-6
-7	8
-7	9
-7	10
-7	11
-7	12
-7	13
-7	14
-7	15
-7	16
-7	17
-7	18
-6	-18
-6	8
-6	9
-6	10
-6	11
-6	12
-6	13
-6	14
-6	15
-6	16
-6	17
-6	18
-5	-18
-5	8
-5	9
-5	10
-5	11
-5	12
-5	13
-5	14
-5	15
-5	16
-5	17
-5	18
-4	-18
-4	-6
-4	7
-4	8
-4	9
-4	10
-4	11
-4	12
-4	13
-4	14
-4	15
-4	16
-4	17
-4	18
-3	-18
-3	-6
-3	-5
-3	7
-3	8
-3	9
-3	10
-3	11
-3	12
-3	13
-3	14
-3	15
-3	16
-3	17
-3	18
-2	-18
-2	-6
-2	-5
-2	7
-2	8
-2	9
-2	10
-2	11
-2	12
-2	13
-2	14
-2	15
-2	16
-2	17
-2	18
-1	-18
-1	-6
-1	6
-1	7
-1	8
-1	9
-1	10
-1	11
-1	12
-1	13
-1	14
-1	15
-1	16
-1	17
-1	18
0	-18
0	-6
0	6
0	7
0	8
0	9
0	10
0	11
0	12
0	13
0	14
0	15
0	16
0	17
0	18
1	-18
1	7
1	8
1	9
1	10
1	11
1	12
1	13
1	14
1	15
1	16
1	17
1	18
2	-18
2	7
2	8
2	9
2	10
2	11

2	12
2	13
2	14
2	15
2	16
2	17
2	18
3	-18
3	8
3	9
3	10
3	11
3	12
3	13
3	14
3	15
3	16
3	17
3	18
4	-18
4	8
4	9
4	10
4	11
4	12
4	13
4	14
4	15
4	16
4	17
4	18
5	-18
5	8
5	9
5	10
5	11
5	12
5	13
5	14
5	15
5	16
5	17
5	18
6	-18
6	8
6	9
6	10
6	11
6	12
6	13
6	14
6	15
6	16
6	17
6	18
7	-18
7	8
7	9
7	10
7	11
7	12
7	13
7	14
7	15
7	16
7	17
7	18
8	-18
8	8
8	9
8	10
8	11
8	12
8	13
8	14
8	15
8	16
8	17
8	18
9	-18
9	8
9	9
9	10
9	11
9	12
9	13
9	14
9	15
9	16
9	17
9	18
10	-18
10	8
10	9
10	10
10	11
10	12
10	13
10	14
10	15
10	16
10	17
10	18
11	-18
11	7
11	8
11	9
11	10
11	11
11	12
11	13
11	14
11	15
11	16
11	17
11	18
12	-18
12	7
12	8
12	9
12	10
12	11
12	12
12	13
12	14
12	15
12	16
12	17
12	18
13	-18
13	6
13	7
13	8
13	9
13	10
13	11
13	12
13	13
13	14
13	15
13	16
13	17
13	18
14	-18
14	6
14	7
14	8
14	9
14	10
14	11
14	12
14	13
14	14
14	15
14	16
14	17
14	18
15	-18
15	6
15	7
15	8
15	9
15	10
15	11
15	12
15	13
15	14
15	15
15	16
15	17
15	18
16	-18
16	6
16	7
16	8
16	9
16	10
16	11
16	12
16	13
16	14
16	15
16	16
16	17
16	18
17	-18
17	6
17	7
17	8
17	9
17	10
17	11
17	12
17	13
17	14
17	15
17	16
17	17
17	18
18	-18
18	6
18	7
18	8
18	9
18	10
18	11
18	12
18	13
18	14
18	15
18	16
18	17
18	18

Chapter 1: Supplementary Table 6: Rotamer library used for sampling pSer. Numbers should be multiplied by 10 to obtain the torsion angle in degrees. The columns represent the sampled heavy-atom torsion angles, other than the rotation about the O-P bond in the phosphate group, which is sampled uniformly. c_1 represents rotation about the C_a-C_b bond, and c_2 is the rotation about the C_b-O_g bond.

-18	-18
-18	-17
-18	-16
-18	-15
-18	-14
-18	-13
-18	-12
-18	-11
-18	-10
-18	7
-18	8
-18	9
-18	10
-18	11
-18	12
-18	13
-18	14
-18	15
-18	16
-18	17
-18	18
-17	-18
-17	-17
-17	-16
-17	-15
-17	-14
-17	-13
-17	-12
-17	-11
-17	-10
-17	-9
-17	7
-17	8
-17	9
-17	10
-17	11
-17	12
-17	13
-17	14
-17	15
-17	16
-17	17
-17	18
-16	-18
-16	-17
-16	-16
-16	-15
-16	-14
-16	-13
-16	-12
-16	-11
-16	-10
-16	-9
-16	7
-16	8
-16	9
-16	10
-16	11
-16	12
-16	13
-16	14
-16	15
-16	16
-16	17
-16	18
-15	-18
-15	-17
-15	-16
-15	-15
-15	-14
-15	-13
-15	-12
-15	-11
-15	-10
-15	-9
-15	6
-15	7
-15	8
-15	9
-15	10
-15	11
-15	12
-15	13
-15	14
-15	15
-15	16
-15	17
-15	18
-14	-18
-14	-17
-14	-16
-14	-15
-14	-14
-14	-13
-14	-12
-14	-11
-14	-10
-14	-9
-14	6
-14	7
-14	8
-14	9
-14	10
-14	11
-14	12
-14	13
-14	14
-14	15
-14	16
-14	17
-14	18
-13	-18
-13	-17
-13	-16
-13	-15
-13	-14
-13	-13
-13	-12
-13	-11
-13	-10
-13	-9
-13	-8
-13	6
-13	7
-13	8
-13	9
-13	10
-13	11
-13	12
-13	13
-13	14
-13	15
-13	16
-13	18
-12	-18
-12	-17
-12	-16
-12	-15
-12	-14
-12	-13
-12	-12
-12	-11
-12	-10
-12	-9
-12	-8
-12	-7
-12	7
-12	8
-12	9
-12	10
-12	11
-12	12
-12	13
-12	14
-12	15
-12	16
-12	17
-12	18
-11	-18
-11	-17
-11	-16
-11	-14
-11	-13
-11	-12
-11	-11
-11	-10
-11	-9
-11	-8
-11	-7
-11	-6
-11	8
-11	9
-11	10
-11	11
-11	12
-11	13
-11	14
-11	15
-11	16
-11	17
-11	18
-10	-18
-10	-17
-10	-16
-10	-15
-10	-14
-10	-13
-10	-12
-10	-11
-10	-10
-10	-9
-10	-8
-10	-7
-10	-6
-10	8
-10	9
-10	10
-10	11
-10	12
-10	13
-10	15
-10	16
-10	17
-10	18
-9	-18
-9	-17
-9	-16
-9	-15
-9	-14
-9	-13
-9	-12
-9	-11
-9	-10
-9	-9
-9	-8
-9	-7
-9	-6
-9	8
-9	9
-9	10
-9	11
-9	12
-9	13
-9	14
-9	15
-9	16
-9	17
-9	18
-8	-18
-8	-17
-8	-16
-8	-15
-8	-14
-8	-13
-8	-12
-8	-11
-8	-10
-8	-9
-8	-8
-8	-7
-8	9
-8	10
-8	11
-8	12
-8	13
-8	14
-8	15
-8	17
-8	18
-7	-18
-7	-16
-7	-15
-7	-14
-7	-12
-7	-11
-7	-10
-7	-9
-7	-8
-7	-7
-7	9
-7	10
-7	11
-7	12
-7	13
-7	14
-7	16
-7	17
-7	18
-6	-18
-6	-17
-6	-16
-6	-15
-6	-14
-6	-13
-6	-12
-6	-11
-6	-10
-6	-9
-6	-8
-6	-7
-6	9
-6	10

-6 11
-6 12
-6 13
-6 14
-6 15
-6 16
-6 17
-6 18
-5 -18
-5 -17
-5 -16
-5 -15
-5 -14
-5 -13
-5 -12
-5 -11
-5 -10
-5 -9
-5 -8
-5 -7
-5 9
-5 10
-5 11
-5 12
-5 13
-5 14
-5 15
-5 16
-5 17
-5 18
-4 -18
-4 -17
-4 -16
-4 -15
-4 -13
-4 -12
-4 -11
-4 -10
-4 -9
-4 -8
-4 -7
-4 8
-4 9
-4 10
-4 11
-4 12
-4 13
-4 14
-4 15
-4 16
-4 17
-4 18
-3 -18
-3 -17
-3 -16
-3 -15
-3 -14

-3 -13
-3 -12
-3 -11
-3 -10
-3 -9
-3 -8
-3 -7
-3 8
-3 9
-3 10
-3 11
-3 12
-3 13
-3 14
-3 15
-3 16
-3 17
-3 18
-2 -18
-2 -17
-2 -16
-2 -15
-2 -14
-2 -13
-2 -12
-2 -10
-2 -9
-2 -8
-2 -7
-2 -6
-2 8
-2 9
-2 10
-2 11
-2 12
-2 13
-2 14
-2 15
-2 16
-2 17
-2 18
-1 -18
-1 -17
-1 -16
-1 -15
-1 -14
-1 -13
-1 -12
-1 -11
-1 -10
-1 -9
-1 -8
-1 -7
-1 8
-1 9
-1 10
-1 11

-1 12
-1 13
-1 14
-1 15
-1 16
-1 17
-1 18
0 -18
0 -17
0 -16
0 -15
0 -14
0 -13
0 -12
0 -11
0 -10
0 -9
0 -8
0 -7
0 7
0 8
0 9
0 10
0 11
0 12
0 13
0 14
0 15
0 16
0 17
0 18
1 -18
1 -17
1 -16
1 -15
1 -14
1 -13
1 -12
1 -11
1 -10
1 -9
1 -8
1 8
1 9
1 10
1 11
1 12
1 13
1 14
1 15
1 16
1 17
1 18
2 -18
2 -17
2 -16
2 -15

2 -14
2 -13
2 -12
2 -11
2 -10
2 -9
2 -8
2 8
2 9
2 10
2 11
2 12
2 13
2 14
2 15
2 16
2 17
2 18
3 -18
3 -17
3 -16
3 -15
3 -14
3 -13
3 -12
3 -11
3 -10
3 -9
3 -8
3 9
3 10
3 11
3 12
3 13
3 14
3 15
3 16
3 17
3 18
4 -18
4 -17
4 -16
4 -15
4 -14
4 -13
4 -12
4 -11
4 -10
4 -9
4 -8
4 9
4 10
4 11
4 12
4 13
4 14
4 15

4 16
4 17
4 18
5 -18
5 -17
5 -16
5 -15
5 -14
5 -13
5 -12
5 -10
5 -9
5 -8
5 9
5 10
5 11
5 12
5 13
5 14
5 15
5 16
5 17
5 18
6 -18
6 -17
6 -16
6 -15
6 -14
6 -13
6 -12
6 -11
6 -10
6 -9
6 9
6 10
6 11
6 12
6 13
6 14
6 15
6 16
6 17
6 18
7 -18
7 -17
7 -16
7 -15
7 -14
7 -13
7 -12
7 -11
7 -10
7 -9
7 9
7 10
7 11
7 12

7 13
7 14
7 15
7 16
7 17
7 18
8 -18
8 -17
8 -16
8 -15
8 -14
8 -13
8 -12
8 -11
8 -10
8 -9
8 9
8 10
8 11
8 12
8 14
8 15
8 16
8 17
8 18
9 -18
9 -17
9 -16
9 -15
9 -14
9 -13
9 -12
9 -11
9 -10
9 -9
9 -8
9 9
9 10
9 11
9 12
9 13
9 14
9 15
9 16
9 17
9 18
10 -18
10 -17
10 -16
10 -15
10 -14
10 -13
10 -12
10 -11
10 -10
10 -9
10 -8

10	9
10	10
10	11
10	12
10	13
10	14
10	15
10	16
10	18
11	-18
11	-17
11	-16
11	-15
11	-14
11	-13
11	-12
11	-11
11	-10
11	-9
11	-8
11	9
11	10
11	11
11	12
11	13
11	14
11	15
11	16
11	17
11	18
12	-18
12	-17
12	-16
12	-15
12	-14
12	-13
12	-12
12	-11
12	-10
12	-9
12	-8
12	8
12	9
12	10
12	11
12	12
12	13
12	14
12	15
12	16
12	17
12	18
13	-18
13	-17
13	-16
13	-15
13	-14
13	-13
13	-12
13	-11
13	-10
13	-9
13	8
13	9
13	10
13	11
13	12
13	13
13	14
13	15
13	16
13	17
13	18
14	-18
14	-17
14	-16
14	-15
14	-14
14	-13
14	-12
14	-11
14	-10
14	-9
14	7
14	8
14	9
14	10
14	12
14	13
14	14
14	15
14	16
14	17
14	18
15	-18
15	-17
15	-16
15	-15
15	-14
15	-13
15	-12
15	-11
15	-10
15	-9
15	8
15	9
15	10
15	11
15	12
15	13
15	14
15	15
15	16
15	17
15	18
16	-18
16	-17
16	-16
16	-15
16	-14
16	-13
16	-12
16	-11
16	-10
16	-9
16	7
16	8
16	9
16	10
16	12
16	13
16	14
16	15
16	16
16	17
16	18
17	-18
17	-17
17	-16
17	-15
17	-14
17	-13
17	-12
17	-11
17	-10
17	-9
17	7
17	8
17	9
17	10
17	11
17	12
17	14
17	15
17	16
17	17
17	18
18	-18
18	-17
18	-16
18	-15
18	-14
18	-13
18	-12
18	-11
18	-10
18	7
18	8
18	9
18	10
18	11
18	12
18	13
18	14
18	15
18	16
18	17
18	18

Chapter 1: Supplementary Table 7: Rotamer library used for sampling pTyr. Numbers should be multiplied by 30 to obtain the torsion angle in degrees. The columns represent the sampled heavy-atom torsion angles, other than the rotation about the O-P bond in the phosphate group, which is sampled uniformly. c_1 represents rotation about the C_a-C_b bond, c_2 is the rotation about C_b-C_g, and c_3 is rotation about the C_z-O bond.

-6	-6	-4
-6	-6	-3
-6	-6	-2
-6	-6	2
-6	-6	3
-6	-6	4
-6	-5	-4
-6	-5	-3
-6	-5	-2
-6	-5	2
-6	-5	3
-6	-5	4
-6	-4	-4
-6	-4	-3
-6	-4	-2
-6	-4	2
-6	-4	3
-6	-4	4
-6	-3	-4
-6	-3	-3
-6	-3	-2
-6	-3	2
-6	-3	3
-6	-3	4
-6	-2	-4
-6	-2	-3
-6	-2	-2
-6	-2	2
-6	-2	3
-6	-2	4
-6	0	-4
-6	0	-3
-6	0	-2
-6	0	2
-6	0	3
-6	0	4
-6	1	-4
-6	1	-3
-6	1	-2
-6	1	2
-6	1	3
-6	1	4
-6	2	-4
-6	2	-3
-6	2	-2
-6	2	2
-6	2	3
-6	2	4
-6	3	-4
-6	3	-3
-6	3	-2
-6	3	2
-6	3	3
-6	3	4
-6	4	-4
-6	4	-3
-6	4	-2
-6	4	2
-6	4	3
-6	4	4
-6	6	-4
-6	6	-3
-6	6	-2
-6	6	2
-6	6	3
-6	6	4
-5	-6	-4
-5	-6	-3
-5	-6	-2
-5	-6	2
-5	-6	3
-5	-6	4
-5	-5	-4
-5	-5	-3
-5	-5	-2
-5	-5	2
-5	-5	3
-5	-5	4
-5	-4	-4
-5	-4	-3
-5	-4	-2
-5	-4	2
-5	-4	3
-5	-4	4
-5	-3	-4
-5	-3	-3
-5	-3	-2
-5	-3	2
-5	-3	3
-5	-3	4
-5	-2	-4
-5	-2	-3
-5	-2	-2
-5	-2	2
-5	-2	3
-5	-2	4
-5	0	-4
-5	0	-3
-5	0	-2
-5	0	2
-5	0	3
-5	0	4
-5	1	-4
-5	1	-3
-5	1	-2
-5	1	2
-5	1	3
-5	1	4
-5	2	-4
-5	2	-3
-5	2	-2
-5	2	2
-5	2	3
-5	2	4
-5	3	-4
-5	3	-3
-5	3	-2
-5	3	2
-5	3	3
-5	3	4
-5	4	-4
-5	4	-3
-5	4	-2
-5	4	2
-5	4	3
-5	4	4
-5	6	-4
-5	6	-3
-5	6	-2
-5	6	2
-5	6	3
-5	6	4
-4	-6	-4
-4	-6	-3
-4	-6	-2
-4	-6	2
-4	-6	3
-4	-6	4
-4	-5	-4
-4	-5	-3
-4	-5	-2
-4	-5	2
-4	-5	3
-4	-5	4
-4	-4	-4
-4	-4	-3
-4	-4	-2
-4	-4	2
-4	-4	3
-4	-4	4
-4	-3	-4
-4	-3	-3
-4	-3	-2
-4	-3	2
-4	-3	3
-4	-3	4
-4	-2	-4
-4	-2	-3
-4	-2	-2
-4	-2	2
-4	-2	3
-4	-2	4
-4	0	-4
-4	0	-3
-4	0	-2
-4	0	2
-4	0	3
-4	0	4
-4	1	-4
-4	1	-3
-4	1	-2
-4	1	2
-4	1	3
-4	1	4
-4	2	-4
-4	2	-3
-4	2	-2
-4	2	2
-4	2	3
-4	2	4
-4	3	-4
-4	3	-3
-4	3	-2
-4	3	2
-4	3	3
-4	3	4
-4	4	-4
-4	4	-3
-4	4	-2
-4	4	2
-4	4	3
-4	4	4
-4	6	-4
-4	6	-3
-4	6	-2
-4	6	2
-4	6	3
-4	6	4
-3	-6	-4
-3	-6	-3
-3	-6	-2
-3	-6	2
-3	-6	3
-3	-6	4
-3	-5	-4
-3	-5	-3
-3	-5	-2
-3	-5	2
-3	-5	3
-3	-5	4
-3	-4	-4
-3	-4	-3
-3	-4	-2
-3	-4	2
-3	-4	3
-3	-4	4
-3	-3	-4
-3	-3	-3
-3	-3	-2
-3	-3	2
-3	-3	3
-3	-3	4
-3	-2	-4
-3	-2	-3
-3	-2	-2
-3	-2	2
-3	-2	3
-3	-2	4
-3	0	-4
-3	0	-3
-3	0	-2
-3	0	2
-3	0	3
-3	0	4
-3	1	-4
-3	1	-3
-3	1	-2
-3	1	2
-3	1	3
-3	1	4
-3	2	-4
-3	2	-3
-3	2	-2
-3	2	2
-3	2	3
-3	2	4
-3	3	-4
-3	3	-3
-3	3	-2
-3	3	2
-3	3	3
-3	3	4
-3	4	-4
-3	4	-3
-3	4	-2
-3	4	2
-3	4	3
-3	4	4
-3	6	-4
-3	6	-3
-3	6	-2
-3	6	2
-3	6	3
-3	6	4
-2	-6	-4
-2	-6	-3
-2	-6	-2
-2	-6	2
-2	-6	3
-2	-6	4
-2	-5	-4
-2	-5	-3
-2	-5	-2
-2	-5	2
-2	-5	3
-2	-5	4

```
-2 -4 -4
-2 -4 -3
-2 -4 -2
-2 -4 2
-2 -4 3
-2 -4 4
-2 -3 -4
-2 -3 -3
-2 -3 -2
-2 -3 2
-2 -3 3
-2 -3 4
-2 -2 -4
-2 -2 -3
-2 -2 -2
-2 -2 2
-2 -2 3
-2 -2 4
-2 0 -4
-2 0 -3
-2 0 -2
-2 0 2
-2 0 3
-2 0 4
-2 1 -4
-2 1 -3
-2 1 -2
-2 1 2
-2 1 3
-2 1 4
-2 2 -4
-2 2 -3
-2 2 -2
-2 2 2
-2 2 3
-2 2 4
-2 3 -4
-2 3 -3
-2 3 -2
-2 3 2
-2 3 3
-2 3 4
-2 4 -4
-2 4 -3
-2 4 -2
-2 4 2
-2 4 3
-2 4 4
-2 6 -4
-2 6 -3
-2 6 -2
-2 6 2
-2 6 3
-2 6 4
-1 -6 -4
-1 -6 -3
-1 -6 -2
```

```
-1 -6 2
-1 -6 3
-1 -6 4
-1 -5 -4
-1 -5 -3
-1 -5 -2
-1 -5 2
-1 -5 3
-1 -5 4
-1 -4 -4
-1 -4 -3
-1 -4 -2
-1 -4 2
-1 -4 3
-1 -4 4
-1 -3 -4
-1 -3 -3
-1 -3 -2
-1 -3 2
-1 -3 3
-1 -3 4
-1 -2 -4
-1 -2 -3
-1 -2 -2
-1 -2 2
-1 -2 3
-1 -2 4
-1 0 -4
-1 0 -3
-1 0 -2
-1 0 2
-1 0 3
-1 0 4
-1 1 -4
-1 1 -3
-1 1 -2
-1 1 2
-1 1 3
-1 1 4
-1 2 -4
-1 2 -3
-1 2 -2
-1 2 2
-1 2 3
-1 2 4
-1 3 -4
-1 3 -3
-1 3 -2
-1 3 2
-1 3 3
-1 3 4
-1 4 -4
-1 4 -3
-1 4 -2
-1 4 2
-1 4 3
-1 4 4
```

```
-1 6 -4
-1 6 -3
-1 6 -2
-1 6 2
-1 6 3
-1 6 4
0 -6 -4
0 -6 -3
0 -6 -2
0 -6 2
0 -6 3
0 -6 4
0 -5 -4
0 -5 -3
0 -5 -2
0 -5 2
0 -5 3
0 -5 4
0 -4 -4
0 -4 -3
0 -4 -2
0 -4 2
0 -4 3
0 -4 4
0 -3 -4
0 -3 -3
0 -3 -2
0 -3 2
0 -3 3
0 -3 4
0 -2 -4
0 -2 -3
0 -2 -2
0 -2 2
0 -2 3
0 -2 4
0 0 -4
0 0 -3
0 0 -2
0 0 2
0 0 3
0 0 4
0 1 -4
0 1 -3
0 1 -2
0 1 2
0 1 3
0 1 4
0 2 -4
0 2 -3
0 2 -2
0 2 2
0 2 3
0 2 4
0 3 -4
0 3 -3
0 3 -2
```

```
0 3 2
0 3 3
0 3 4
0 4 -4
0 4 -3
0 4 -2
0 4 2
0 4 3
0 4 4
0 6 -4
0 6 -3
0 6 -2
0 6 2
0 6 3
0 6 4
1 -6 -4
1 -6 -3
1 -6 -2
1 -6 2
1 -6 3
1 -6 4
1 -5 -4
1 -5 -3
1 -5 -2
1 -5 2
1 -5 3
1 -5 4
1 -4 -4
1 -4 -3
1 -4 -2
1 -4 2
1 -4 3
1 -4 4
1 -3 -4
1 -3 -3
1 -3 -2
1 -3 2
1 -3 3
1 -3 4
1 -2 -4
1 -2 -3
1 -2 -2
1 -2 2
1 -2 3
1 -2 4
1 0 -4
1 0 -3
1 0 -2
1 0 2
1 0 3
1 0 4
1 1 -4
1 1 -3
1 1 -2
1 1 2
1 1 3
1 1 4
```

```
1 2 -4
1 2 -3
1 2 -2
1 2 2
1 2 3
1 2 4
1 3 -4
1 3 -3
1 3 -2
1 3 2
1 3 3
1 3 4
1 4 -4
1 4 -3
1 4 -2
1 4 2
1 4 3
1 4 4
1 6 -4
1 6 -3
1 6 -2
1 6 2
1 6 3
1 6 4
2 -6 -4
2 -6 -3
2 -6 -2
2 -6 2
2 -6 3
2 -6 4
2 -5 -4
2 -5 -3
2 -5 -2
2 -5 2
2 -5 3
2 -5 4
2 -4 -4
2 -4 -3
2 -4 -2
2 -4 2
2 -4 3
2 -4 4
2 -3 -4
2 -3 -3
2 -3 -2
2 -3 2
2 -3 3
2 -3 4
2 -2 -4
2 -2 -3
2 -2 -2
2 -2 2
2 -2 3
2 -2 4
2 0 -4
2 0 -3
2 0 -2
```

```
2 0 2
2 0 3
2 0 4
2 1 -4
2 1 -3
2 1 -2
2 1 2
2 1 3
2 1 4
2 2 -4
2 2 -3
2 2 -2
2 2 2
2 2 3
2 2 4
2 3 -4
2 3 -3
2 3 -2
2 3 2
2 3 3
2 3 4
2 4 -4
2 4 -3
2 4 -2
2 4 2
2 4 3
2 4 4
2 6 -4
2 6 -3
2 6 -2
2 6 2
2 6 3
2 6 4
3 -6 -4
3 -6 -3
3 -6 -2
3 -6 2
3 -6 3
3 -6 4
3 -5 -4
3 -5 -3
3 -5 -2
3 -5 2
3 -5 3
3 -5 4
3 -4 -4
3 -4 -3
3 -4 -2
3 -4 2
3 -4 3
3 -4 4
3 -3 -4
3 -3 -3
3 -3 -2
3 -3 2
3 -3 3
3 -3 4
```

```
3 -2 -4
3 -2 -3
3 -2 -2
3 -2 2
3 -2 3
3 -2 4
3 0 -4
3 0 -3
3 0 -2
3 0 2
3 0 3
3 0 4
3 1 -4
3 1 -3
3 1 -2
3 1 2
3 1 3
3 1 4
3 2 -4
3 2 -3
3 2 -2
3 2 2
3 2 3
3 2 4
3 3 -4
3 3 -3
3 3 -2
3 3 2
3 3 3
3 3 4
3 4 -4
3 4 -3
3 4 -2
3 4 2
3 4 3
3 4 4
3 6 -4
3 6 -3
3 6 -2
3 6 2
3 6 3
3 6 4
4 -6 -4
4 -6 -3
4 -6 -2
4 -6 2
4 -6 3
4 -6 4
4 -5 -4
4 -5 -3
4 -5 -2
4 -5 2
4 -5 3
4 -5 4
4 -4 -4
4 -4 -3
4 -4 -2
4 -4 2
4 -4 3
4 -4 4
4 -3 -4
4 -3 -3
4 -3 -2
4 -3 2
4 -3 3
4 -3 4
4 -2 -4
4 -2 -3
4 -2 -2
4 -2 2
4 -2 3
4 -2 4
4 0 -4
4 0 -3
4 0 -2
4 0 2
4 0 3
4 0 4
4 1 -4
4 1 -3
4 1 -2
4 1 2
4 1 3
4 1 4
4 2 -4
4 2 -3
4 2 -2
4 2 2
4 2 3
4 2 4
4 3 -4
4 3 -3
4 3 -2
4 3 2
4 3 3
4 3 4
4 4 -4
4 4 -3
4 4 -2
4 4 2
4 4 3
4 4 4
4 6 -4
4 6 -3
4 6 -2
4 6 2
4 6 3
4 6 4
5 -6 -4
5 -6 -3
5 -6 -2
5 -6 2
5 -6 3
5 -6 4
5 -5 -4
5 -5 -3
5 -5 -2
5 -5 2
5 -5 3
5 -5 4
5 -4 -4
5 -4 -3
5 -4 -2
5 -4 2
5 -4 3
5 -4 4
5 -3 -4
5 -3 -3
5 -3 -2
5 -3 2
5 -3 3
5 -3 4
5 -2 -4
5 -2 -3
5 -2 -2
5 -2 2
5 -2 3
5 -2 4
5 0 -4
5 0 -3
5 0 -2
5 0 2
5 0 3
5 0 4
5 1 -4
5 1 -3
5 1 -2
5 1 2
5 1 3
5 1 4
5 2 -4
5 2 -3
5 2 -2
5 2 2
5 2 3
5 2 4
5 3 -4
5 3 -3
5 3 -2
5 3 2
5 3 3
5 3 4
5 4 -4
5 4 -3
5 4 -2
5 4 2
5 4 3
5 4 4
5 6 -4
5 6 -3
5 6 -2
5 6 2
5 6 3
5 6 4
6 -6 -4
6 -6 -3
6 -6 -2
6 -6 2
6 -6 3
6 -6 4
6 -5 -4
6 -5 -3
6 -5 -2
6 -5 2
6 -5 3
6 -5 4
6 -4 -4
6 -4 -3
6 -4 -2
6 -4 2
6 -4 3
6 -4 4
6 -3 -4
6 -3 -3
6 -3 -2
6 -3 2
6 -3 3
6 -3 4
6 -2 -4
6 -2 -3
6 -2 -2
6 -2 2
6 -2 3
6 -2 4
6 0 -4
6 0 -3
6 0 -2
6 0 2
6 0 3
6 0 4
6 1 -4
6 1 -3
6 1 -2
6 1 2
6 1 3
6 1 4
6 2 -4
6 2 -3
6 2 -2
6 2 2
6 2 3
6 2 4
6 3 -4
6 3 -3
6 3 -2
6 3 2
6 3 3
6 3 4
6 4 -4
6 4 -3
6 4 -2
6 4 2
6 4 3
6 4 4
6 6 -4
6 6 -3
6 6 -2
6 6 2
6 6 3
6 6 4
```

Chapter 1: Supplementary Table 8: Rotamer library used for sampling pAsp. Numbers should be multiplied by 30 to obtain the torsion angle in degrees. The columns represent the sampled heavy-atom torsion angles, other than the rotation about the O-P bond in the phosphate group, which is sampled uniformly. c_1 represents rotation about the C_a-C_b bond, c_2 is the rotation about C_b-C_g, and c_3 is rotation about the C_g-O bond.

-5	-5	-5
-5	-5	-1
-5	-5	0
-5	-5	1
-5	-5	2
-5	-5	5
-5	-5	6
-5	-4	-5
-5	-4	-1
-5	-4	0
-5	-4	1
-5	-4	2
-5	-4	5
-5	-4	6
-5	-3	-5
-5	-3	-1
-5	-3	0
-5	-3	1
-5	-3	2
-5	-3	5
-5	-3	6
-5	-2	-5
-5	-2	-1
-5	-2	0
-5	-2	1
-5	-2	2
-5	-2	5
-5	-2	6
-5	-1	-5
-5	-1	-1
-5	-1	0
-5	-1	1
-5	-1	2
-5	-1	5
-5	-1	6
-5	0	-5
-5	0	-1
-5	0	0
-5	0	1
-5	0	2
-5	0	5
-5	0	6
-5	1	-5
-5	1	-1
-5	1	0
-5	1	1
-5	1	2
-5	1	5
-5	1	6
-5	2	-5
-5	2	-1
-5	2	0
-5	2	1
-5	2	2
-5	2	5
-5	2	6
-5	3	-5
-5	3	-1
-5	3	0
-5	3	1
-5	3	2
-5	3	5
-5	3	6
-5	4	-5
-5	4	-1
-5	4	0
-5	4	1
-5	4	2
-5	4	5
-5	4	6
-5	5	-5
-5	5	-1
-5	5	0
-5	5	1
-5	5	2
-5	5	5
-5	5	6
-5	6	-5
-5	6	-1
-5	6	0
-5	6	1
-5	6	2
-5	6	5
-5	6	6
-4	-5	-5
-4	-5	-1
-4	-5	0
-4	-5	1
-4	-5	2
-4	-5	5
-4	-5	6
-4	-4	-5
-4	-4	-1
-4	-4	0
-4	-4	1
-4	-4	2
-4	-4	5
-4	-4	6
-4	-3	-5
-4	-3	-1
-4	-3	0
-4	-3	1
-4	-3	2
-4	-3	5
-4	-3	6
-4	-2	-5
-4	-2	-1
-4	-2	0
-4	-2	1
-4	-2	2
-4	-2	5
-4	-2	6
-4	-1	-5
-4	-1	-1
-4	-1	0
-4	-1	1
-4	-1	2
-4	-1	5
-4	-1	6
-4	0	-5
-4	0	-1
-4	0	0
-4	0	1
-4	0	2
-4	0	5
-4	0	6
-4	1	-5
-4	1	-1
-4	1	0
-4	1	1
-4	1	2
-4	1	5
-4	1	6
-4	2	-5
-4	2	-1
-4	2	0
-4	2	1
-4	2	2
-4	2	5
-4	2	6
-4	3	-5
-4	3	-1
-4	3	0
-4	3	1
-4	3	2
-4	3	5
-4	3	6
-4	4	-5
-4	4	-1
-4	4	0
-4	4	1
-4	4	2
-4	4	5
-4	4	6
-4	5	-5
-4	5	-1
-4	5	0
-4	5	1
-4	5	2
-4	5	5
-4	5	6
-4	6	-5
-4	6	-1
-4	6	0
-4	6	1
-4	6	2
-4	6	5
-4	6	6
-3	-5	-5
-3	-5	-1
-3	-5	0
-3	-5	1
-3	-5	2
-3	-5	5
-3	-5	6
-3	-4	-5
-3	-4	-1
-3	-4	0
-3	-4	1
-3	-4	2
-3	-4	5
-3	-4	6
-3	-3	-5
-3	-3	-1
-3	-3	0
-3	-3	1
-3	-3	2
-3	-3	5
-3	-3	6
-3	-2	-5
-3	-2	-1
-3	-2	0
-3	-2	1
-3	-2	2
-3	-2	5
-3	-2	6
-3	-1	-5
-3	-1	-1
-3	-1	0
-3	-1	1
-3	-1	2
-3	-1	5
-3	-1	6
-3	0	-5
-3	0	-1
-3	0	0
-3	0	1
-3	0	2
-3	0	5
-3	0	6
-3	1	-5
-3	1	-1
-3	1	0
-3	1	1
-3	1	2
-3	1	5
-3	1	6
-3	2	-5
-3	2	-1
-3	2	0
-3	2	1
-3	2	2
-3	2	5
-3	2	6
-3	3	-5
-3	3	-1
-3	3	0
-3	3	1
-3	3	2
-3	3	5
-3	3	6
-3	4	-5
-3	4	-1
-3	4	0
-3	4	1
-3	4	2
-3	4	5
-3	4	6
-3	5	-5
-3	5	-1
-3	5	0
-3	5	1
-3	5	2
-3	5	5
-3	5	6
-3	6	-5
-3	6	-1
-3	6	0
-3	6	1
-3	6	2
-3	6	5
-3	6	6
-2	-5	-5
-2	-5	-1
-2	-5	0
-2	-5	1
-2	-5	2
-2	-5	5
-2	-5	6
-2	-4	-5
-2	-4	-1
-2	-4	0
-2	-4	1
-2	-4	2
-2	-4	5
-2	-4	6
-2	-3	-5
-2	-3	-1
-2	-3	0
-2	-3	1
-2	-3	2
-2	-3	5
-2	-3	6
-2	-2	-5
-2	-2	-1
-2	-2	0
-2	-2	1
-2	-2	2
-2	-2	5
-2	-2	6
-2	-1	-5
-2	-1	-1

```
-2 -1 0
-2 -1 1
-2 -1 2
-2 -1 5
-2 -1 6
-2 0 -5
-2 0 -1
-2 0 0
-2 0 1
-2 0 2
-2 0 5
-2 0 6
-2 1 -5
-2 1 -1
-2 1 0
-2 1 1
-2 1 2
-2 1 5
-2 1 6
-2 2 -5
-2 2 -1
-2 2 0
-2 2 1
-2 2 2
-2 2 5
-2 2 6
-2 3 -5
-2 3 -1
-2 3 0
-2 3 1
-2 3 2
-2 3 5
-2 3 6
-2 4 -5
-2 4 -1
-2 4 0
-2 4 1
-2 4 2
-2 4 5
-2 4 6
-2 5 -5
-2 5 -1
-2 5 0
-2 5 1
-2 5 2
-2 5 5
-2 5 6
-2 6 -5
-2 6 -1
-2 6 0
-2 6 1
-2 6 2
-2 6 5
-2 6 6
-1 -5 -5
-1 -5 -1
-1 -5 0
-1 -5 1
-1 -5 2
-1 -5 5
-1 -5 6
-1 -4 -5
-1 -4 -1
-1 -4 0
-1 -4 1
-1 -4 2
-1 -4 5
-1 -4 6
-1 -3 -5
-1 -3 -1
-1 -3 0
-1 -3 1
-1 -3 2
-1 -3 5
-1 -3 6
-1 -2 -5
-1 -2 -1
-1 -2 0
-1 -2 1
-1 -2 2
-1 -2 5
-1 -2 6
-1 -1 -5
-1 -1 -1
-1 -1 0
-1 -1 1
-1 -1 2
-1 -1 5
-1 -1 6
-1 0 -5
-1 0 -1
-1 0 0
-1 0 1
-1 0 2
-1 0 5
-1 0 6
-1 1 -5
-1 1 -1
-1 1 0
-1 1 1
-1 1 2
-1 1 5
-1 1 6
-1 2 -5
-1 2 -1
-1 2 0
-1 2 1
-1 2 2
-1 2 5
-1 2 6
-1 3 -5
-1 3 -1
-1 3 0
-1 3 1
-1 3 2
-1 3 5
-1 3 6
-1 4 -5
-1 4 -1
-1 4 0
-1 4 1
-1 4 2
-1 4 5
-1 4 6
-1 5 -5
-1 5 -1
-1 5 0
-1 5 1
-1 5 2
-1 5 5
-1 5 6
-1 6 -5
-1 6 -1
-1 6 0
-1 6 1
-1 6 2
-1 6 5
-1 6 6
0 -5 -5
0 -5 -1
0 -5 0
0 -5 1
0 -5 2
0 -5 5
0 -5 6
0 -4 -5
0 -4 -1
0 -4 0
0 -4 1
0 -4 2
0 -4 5
0 -4 6
0 -3 -5
0 -3 -1
0 -3 0
0 -3 1
0 -3 2
0 -3 5
0 -3 6
0 -2 -5
0 -2 -1
0 -2 0
0 -2 1
0 -2 2
0 -2 5
0 -2 6
0 -1 -5
0 -1 -1
0 -1 0
0 -1 1
0 -1 2
0 -1 5
0 -1 6
0 0 -5
0 0 -1
0 0 0
0 0 1
0 0 2
0 0 5
0 0 6
0 1 -5
0 1 -1
0 1 0
0 1 1
0 1 2
0 1 5
0 1 6
0 2 -5
0 2 -1
0 2 0
0 2 1
0 2 2
0 2 5
0 2 6
0 3 -5
0 3 -1
0 3 0
0 3 1
0 3 2
0 3 5
0 3 6
0 4 -5
0 4 -1
0 4 0
0 4 1
0 4 2
0 4 5
0 4 6
0 5 -5
0 5 -1
0 5 0
0 5 1
0 5 2
0 5 5
0 5 6
0 6 -5
0 6 -1
0 6 0
0 6 1
0 6 2
0 6 5
0 6 6
1 -5 -5
1 -5 -1
1 -5 0
1 -5 1
1 -5 2
1 -5 5
1 -5 6
1 -4 -5
1 -4 -1
1 -4 0
1 -4 1
1 -4 2
1 -4 5
1 -4 6
1 -3 -5
1 -3 -1
1 -3 0
1 -3 1
1 -3 2
1 -3 5
1 -3 6
1 -2 -5
1 -2 -1
1 -2 0
1 -2 1
1 -2 2
1 -2 5
1 -2 6
1 -1 -5
1 -1 -1
1 -1 0
1 -1 1
1 -1 2
1 -1 5
1 -1 6
1 0 -5
1 0 -1
1 0 0
1 0 1
1 0 2
1 0 5
1 0 6
1 1 -5
1 1 -1
1 1 0
1 1 1
1 1 2
1 1 5
1 1 6
1 2 -5
1 2 -1
1 2 0
1 2 1
1 2 2
1 2 5
1 2 6
1 3 -5
1 3 -1
1 3 0
1 3 1
1 3 2
1 3 5
1 3 6
1 4 -5
1 4 -1
1 4 0
1 4 1
1 4 2
1 4 5
1 4 6
1 5 -5
1 5 -1
1 5 0
1 5 1
1 5 2
1 5 5
1 5 6
1 6 -5
1 6 -1
1 6 0
1 6 1
1 6 2
1 6 5
1 6 6
2 -5 -5
2 -5 -1
2 -5 0
2 -5 1
2 -5 2
2 -5 5
2 -5 6
2 -4 -5
2 -4 -1
2 -4 0
2 -4 1
2 -4 2
2 -4 5
2 -4 6
2 -3 -5
2 -3 -1
2 -3 0
2 -3 1
2 -3 2
2 -3 5
2 -3 6
2 -2 -5
2 -2 -1
2 -2 0
2 -2 1
2 -2 2
2 -2 5
2 -2 6
2 -1 -5
2 -1 -1
2 -1 0
2 -1 1
2 -1 2
2 -1 5
2 -1 6
2 0 -5
```

```
2 0 -1
2 0 0
2 0 1
2 0 2
2 0 5
2 0 6
2 1 -5
2 1 -1
2 1 0
2 1 1
2 1 2
2 1 5
2 1 6
2 2 -5
2 2 -1
2 2 0
2 2 1
2 2 2
2 2 5
2 2 6
2 3 -5
2 3 -1
2 3 0
2 3 1
2 3 2
2 3 5
2 3 6
2 4 -5
2 4 -1
2 4 0
2 4 1
2 4 2
2 4 5
2 4 6
2 5 -5
2 5 -1
2 5 0
2 5 1
2 5 2
2 5 5
2 5 6
2 6 -5
2 6 -1
2 6 0
2 6 1
2 6 2
2 6 5
2 6 6
3 -5 -5
3 -5 -1
3 -5 0
3 -5 1
3 -5 2
3 -5 5
3 -5 6
3 -4 -5
3 -4 -1
3 -4 0
3 -4 1
3 -4 2
3 -4 5
3 -4 6
3 -3 -5
3 -3 -1
3 -3 0
3 -3 1
3 -3 2
3 -3 5
3 -3 6
3 -2 -5
3 -2 -1
3 -2 0
3 -2 1
3 -2 2
3 -2 5
3 -2 6
3 -1 -5
3 -1 -1
3 -1 0
3 -1 1
3 -1 2
3 -1 5
3 -1 6
3 0 -5
3 0 -1
3 0 0
3 0 1
3 0 2
3 0 5
3 0 6
3 1 -5
3 1 -1
3 1 0
3 1 1
3 1 2
3 1 5
3 1 6
3 2 -5
3 2 -1
3 2 0
3 2 1
3 2 2
3 2 5
3 2 6
3 3 -5
3 3 -1
3 3 0
3 3 1
3 3 2
3 3 5
3 3 6
3 4 -5
3 4 -1
3 4 0
3 4 1
3 4 2
3 4 5
3 4 6
3 5 -5
3 5 -1
3 5 0
3 5 1
3 5 2
3 5 5
3 5 6
3 6 -5
3 6 -1
3 6 0
3 6 1
3 6 2
3 6 5
3 6 6
4 -5 -5
4 -5 -1
4 -5 0
4 -5 1
4 -5 2
4 -5 5
4 -5 6
4 -4 -5
4 -4 -1
4 -4 0
4 -4 1
4 -4 2
4 -4 5
4 -4 6
4 -3 -5
4 -3 -1
4 -3 0
4 -3 1
4 -3 2
4 -3 5
4 -3 6
4 -2 -5
4 -2 -1
4 -2 0
4 -2 1
4 -2 2
4 -2 5
4 -2 6
4 -1 -5
4 -1 -1
4 -1 0
4 -1 1
4 -1 2
4 -1 5
4 -1 6
4 0 -5
4 0 -1
4 0 0
4 0 1
4 0 2
4 0 5
4 0 6
4 1 -5
4 1 -1
4 1 0
4 1 1
4 1 2
4 1 5
4 1 6
4 2 -5
4 2 -1
4 2 0
4 2 1
4 2 2
4 2 5
4 2 6
4 3 -5
4 3 -1
4 3 0
4 3 1
4 3 2
4 3 5
4 3 6
4 4 -5
4 4 -1
4 4 0
4 4 1
4 4 2
4 4 5
4 4 6
4 5 -5
4 5 -1
4 5 0
4 5 1
4 5 2
4 5 5
4 5 6
4 6 -5
4 6 -1
4 6 0
4 6 1
4 6 2
4 6 5
4 6 6
5 -5 -5
5 -5 -1
5 -5 0
5 -5 1
5 -5 2
5 -5 5
5 -5 6
5 -4 -5
5 -4 -1
5 -4 0
5 -4 1
5 -4 2
5 -4 5
5 -4 6
5 -3 -5
5 -3 -1
5 -3 0
5 -3 1
5 -3 2
5 -3 5
5 -3 6
5 -2 -5
5 -2 -1
5 -2 0
5 -2 1
5 -2 2
5 -2 5
5 -2 6
5 -1 -5
5 -1 -1
5 -1 0
5 -1 1
5 -1 2
5 -1 5
5 -1 6
5 0 -5
5 0 -1
5 0 0
5 0 1
5 0 2
5 0 5
5 0 6
5 1 -5
5 1 -1
5 1 0
5 1 1
5 1 2
5 1 5
5 1 6
5 2 -5
5 2 -1
5 2 0
5 2 1
5 2 2
5 2 5
5 2 6
5 3 -5
5 3 -1
5 3 0
5 3 1
5 3 2
5 3 5
5 3 6
5 4 -5
5 4 -1
5 4 0
5 4 1
5 4 2
5 4 5
5 4 6
5 5 -5
5 5 -1
5 5 0
5 5 1
5 5 2
5 5 5
5 5 6
5 6 -5
5 6 -1
5 6 0
5 6 1
5 6 2
5 6 5
5 6 6
6 -5 -5
6 -5 -1
6 -5 0
6 -5 1
6 -5 2
6 -5 5
6 -5 6
6 -4 -5
6 -4 -1
6 -4 0
6 -4 1
6 -4 2
6 -4 5
6 -4 6
6 -3 -5
6 -3 -1
6 -3 0
6 -3 1
6 -3 2
6 -3 5
6 -3 6
6 -2 -5
6 -2 -1
6 -2 0
6 -2 1
6 -2 2
6 -2 5
6 -2 6
6 -1 -5
6 -1 -1
6 -1 0
6 -1 1
6 -1 2
6 -1 5
6 -1 6
6 0 -5
6 0 -1
6 0 0
6 0 1
6 0 2
6 0 5
6 0 6
```

```
6  1  -5
6  1  -1
6  1  0
6  1  1
6  1  2
6  1  5
6  1  6
6  2  -5
6  2  -1
6  2  0
6  2  1
6  2  2
6  2  5
6  2  6
6  3  -5
6  3  -1
6  3  0
6  3  1
6  3  2
6  3  5
6  3  6
6  4  -5
6  4  -1
6  4  0
6  4  1
6  4  2
6  4  5
6  4  6
6  5  -5
6  5  -1
6  5  0
6  5  1
6  5  2
6  5  5
6  5  6
6  6  -5
6  6  -1
6  6  0
6  6  1
6  6  2
6  6  5
6  6
```

Chapter 2: Strengths of Hydrogen Bonds Involving Phosphorylated Amino Acid Side Chains

Daniel J. Mandell[1], Ilya Chorny[2], Eli S. Groban[3], Sergio E. Wong[3], Elisheva Levine[4], Chaya S. Rapp[4], and Matthew P. Jacobson[2]

[1] Graduate Program in Biological and Medical Informatics, University of California, San Francisco

[2] Department of Pharmaceutical Chemistry, Box 2240, University of California, San Francisco, CA 94143-2240

[3] Graduate Group in Biophysics, University of California, San Francisco

[4] Department of Chemistry, Stern College for Women, Yeshiva University, New York, New York 10016

Abstract

Post-translational phosphorylation plays a key role in regulating protein function. Here we provide a quantitative assessment of the relative strengths of hydrogen bonds involving phosphorylated amino acid side chains (pSer, pAsp) with several common donors (Arg, Lys, and backbone amide groups). We utilize multiple levels of theory, consisting of explicit solvent molecular dynamics, implicit solvent molecular mechanics, and quantum mechanics with a self-consistent reaction field treatment of solvent. Because the ~6 pKa of phosphate suggests that -1 and -2 charged species may coexist at physiological pH, hydrogen bonds involving both protonated and deprotonated phosphates for all donor–acceptor pairs are considered. Multiple bonding geometries for the charged–charged interactions are also considered. Arg is shown to be capable of substantially stronger salt bridges with phosphorylated side chains than Lys. A pSer hydrogen bond acceptor tends to form more stable interactions than a pAsp acceptor. The effect of phosphate protonation state on the strengths of the hydrogen bonds is remarkably subtle, with a more pronounced effect on pAsp than pSer.

Introduction

Protein phosphorylation is a key signaling mechanism in diverse cellular processes including metabolism[1], ion channel regulation[2, 3], and cell cycle progression[4-6]. In eukaryotes, the main sites of phosphorylation are tyrosine, serine and threonine side chains, while aspartate and histidine side chains are phosphorylated in the "two-component" signaling pathways of prokaryotes[6, 7]. Addition of the phosphate group, which typically carries a -2 charge at physiological pH, perturbs the local electrostatic potential in the protein and often induces conformational changes that influence function[6] or modulate protein-protein interactions[1].

A critical property of phosphorylated residues is their propensity to accept hydrogen bonds through their phosphate oxygens, frequently with positively charged side chains to form "salt bridges". Salt bridge energetics depend sensitively on the identity, proximity and orientation of the participating side chains and their surrounding environment. Quantitative measurements of salt bridge contributions to protein stability have provided different results depending on the experimental system and experimental design, with estimates ranging from -5.0 to -0.5 kcal/mol of stabilization in T4 lysozyme[8, 9], to 2.0 to 4.0 kcal/mol of destabilization in coiled coils and Arc repressor[10, 11]. There have also been a number of previous computational studies aimed at quantifying the strength of interaction of small ions[12-15]. In particular, Masunov and Lazaridis[15] used molecular dynamics methods to estimate the free energies of salt bridges between likely orientations of all charged naturally occurring amino acid side chains.

This work investigates the strengths of hydrogen bonds and salt bridges involving phosphorylated amino acid side chains using small molecule analogs for common acceptors (methyl phosphate for pSer and pThr, acetyl phosphate for pAsp) and donors (butyl ammonium for a Lys side chain, propyl guaninidinium for Arg, and N-methylacetamide for backbone amide NH groups). Interactions of all donors with propionic acid (Glu analog) are also considered for comparison to a carboxylate receptor with -1 charge.

We utilize multiple levels of theory, including explicit solvent molecular dynamics (MD), implicit solvent molecular mechanics (Poisson-Boltzmann), and quantum mechanics with a self-consistent reaction field treatment of solvent. This approach allows us to identify trends that are consistent across the methods, as well as uncover the sensitivity of each method to different forces governing hydrogen bond strengths. Continuum solvent methods, primarily those based on the Poisson-Boltzmann equation[16] or more heuristic methods such as Generalized Born[17], offer substantial speed advantages relative to explicit solvent models in applications such as molecular dynamics. However, treating the solvent as a continuum dielectric is an approximation and neglects important first-shell solvation effects related to the finite size and asymmetry of a water molecule[18-20]. The results of the explicit solvent calculations cannot be considered free of error either. Notably, molecular mechanics methods using fixed-charged force fields, as employed in the present molecular dynamics calculations, ignore the effect of electronic polarizability on hydrogen bond strengths. Even in high dielectric solvent, the strong electric field exerted by a -2 phosphate group can be expected to lead to significant polarization of the electrons on nearby molecules. To assess the potential impact of electronic polarizability on the strengths of the hydrogen bonds considered here, we have used quantum mechanics with a large basis set, electron correlation treated at the "local" Moller-Plesset second order perturbation theory[21] (LMP2) level, and solvent treated using a self-consistent reaction field (SCRF) method.

The central results of this work consist of potentials of mean force (PMFs) from the explicit solvent molecular dynamics calculations, which are one-dimensional free energy landscapes for a pair of interacting groups as a function of distance between the phosphate and hydrogen bond donors. The PMFs cannot be used trivially to predict the *absolute* free energies of association of the small molecules in solution (which requires extensive averaging over translational and rotational degrees of freedom) or the absolute strengths of hydrogen bonds in a protein environment (which depend on the local environment, such as solvent accessibility). However, the PMFs do provide insight into the *relative* intrinsic strengths of the various types of hydrogen bonds considered, and help to address the following issues:

(1) *The conditions under which conditions Arg or Lys make stronger hydrogen bonds with a phosphorylated side chain.* There is ample albeit indirect evidence that the ability of guanidinium ions to form bidentate hydrogen bonds with carboxylate or phosphate ions leads to particularly strong interactions[22-24]. This property of

guanidinium ions has been extensively employed in the design of synthetic receptors for phosphate-containing ligands[25]. Bidentate hydrogen bonds between Arg and pSer/pThr are also commonly observed in the relatively small number of crystal structures of phosphorylated proteins[6]. However, previous computational studies have suggested that interactions of phosphorylated groups with Lys may be intrinsically stronger[26, 27]. We revisit this issue here using multiple levels of theory.

(2) *The effect of phosphate protonation state on hydrogen bond strength.* The ~6 pKa of phosphate suggests that both -1 and -2 charged species may coexist at physiological pH. We investigate the effect of phosphate protonation state on all hydrogen bonding interactions considered.

(3) *The energetic consequences of substituting a carboxylate be for a phosphate.* In cases where phosphorylation of a protein leads to its activation, it is frequently useful to engineer a constitutively active mutant, e.g., for use in *in vitro* studies. Simply substituting a Glu or Asp for the phosphorylated residue(s) is sufficient to achieve constitutive activation in many cases[28-32], but in other cases this simple strategy results in only partial activation or none at all[33, 34]. The relative strengths of hydrogen bonds involving carboxylates and phosphates is also relevant to the design of inhibitors of SH2 domains[35, 36], which bind phosphorylated peptides. Here we examine in some detail the differences in the intrinsic hydrogen bond strengths of carboxylates *vs.* phosphates with common hydrogen bond donors. These results provide a foundation for understanding why Asp/Glu can

sometimes substitute for phosphorylated amino acids, although other physicochemical differences will undoubtedly also play a role.

We do not directly address the strengths of hydrogen bonds to the phosphate backbone of RNA[37] and DNA[38], although the results we present may have some relevance to this issue.

Materials and Methods

MD Simulation Materials and Parameters

MD simulations were performed using GROMACS 3.2.1[39]. The 2001 OPLS all atom force field[40] was used for stretch, bend, torsional and Lennard-Jones parameters, as well as partial charges for the glutamate, acetate, arginine, lysine, and N-methylacetamide backbone analogs. Partial charges for pSer (-1, -2) and pAsp (-1, -2) analogs were obtained from quantum mechanical calculations[41, 42] and these are presented together with the phosphate OPLS atom types in Supporting Information. Molecules were solvated with TIP3P[43] water molecules in a 40 Å cubic box under periodic boundary conditions. The cut-off distance for the short-range neighbor list was set to 10 Å. Long-range electrostatics were calculated with particle mesh Ewald[44] with a real-space cutoff of 9 Å for nonbonded interactions. Na^+ counterions were fixed at the corners of the solvent box as necessary to obtain electroneutrality. Prior to running molecular dynamics, the potential energy of each configuration was relaxed by steepest descent minimization, followed by 100 ps of molecular dynamics equilibration. Molecular dynamics was then run for 2.1 ns with a time step of 2 fs. Interaction energies and atomic coordinates were recorded every 500 fs. The system was propagated in time with a velocity version of the Verlet algorithm[45, 46]. During and subsequent to equilibration, Nose-Hoover temperature coupling[47] and Berendsen pressure coupling[48] were used to maintain system temperature and pressure, with a reference temperature of 298 K, a reference pressure of 1.0 atm, a time constant of 1 ps, and an isothermal compressibility of 1.1 x 10^{-6} (kcal mol^{-1} $Å^{-3}$). The various acceptors, donors, charge states, and geometries comprised a total of 25 hydrogen bonding configurations representing more than 1 ms of molecular dynamics simulation time.

MD Simulation Constraints

Umbrella sampling[49] using distance and position restraints was employed to calculate a one-dimensional PMF for one of two common interaction geometries, either a coplanar approach or a collinear approach (Figure 1). These orientations consistently arose in simulations of the side chain analogs without position restraints (data not shown). For collinear geometries, the donated hydrogen and its covalently bonded atom, and the accepting oxygen and its covalently bonded atom, were constrained to move on a line. For coplanar geometries, donated hydrogens and accepting oxygens were constrained to a plane, and two additional heavy atoms from each molecule were constrained to move along a line. These constraints kept the interacting moieties facing each other, and in the case of lysine allowed the donated hydrogens to rotate slightly through the plane of interaction to sample optimal hydrogen bonding orientations. Hydrogen atom–heavy atom covalent bond lengths were constrained only in the backbone analog using the LINCS algorithm[50] to stabilize the N–H bond.

The distance between the molecules was constrained using a biasing potential at 0.5 Å intervals. A nearby heavy atom from the donor molecule was constrained to that of the acceptor using a quadratic biasing potential,

$$V(r) = k(r - r_i) \qquad \text{Eq 1.}$$

where k = 143.5 (kcal/Å^2) is the biasing force constant and r_i is the point about which the molecules are constrained. In cases with guanidinium in a coplanar orientation, additional simulation with a higher force constant of 263.0, 430.4, or 860.8 (kcal/Å^2) was necessary to sample adequately around the solvation barrier.

MD Analysis

The potential of mean force was calculated using probability distributions of the constrained distance, r, obtained from the umbrella sampling, as described by Souaille and Roux[51]. Trajectories from each window i were converted to biased population distributions $P_i(r)$ with a bin width of 0.1 Å. The weighted histogram analysis method (WHAM)[52] was used to merge histograms $P_i(r)$ into a single unbiased curve $P(r)$. The algorithm was considered to converge after the free energy constants for all the windows changed by less with 0.01 kcal/mol. The PMF, D$G(r)$, was then calculated using the standard relationship

$$\mathrm{D}G(r) = -k_b T \ln[P(r)] \qquad \text{Eq 2.}$$

where k_b is Boltzmann's constant. Each PMF was shifted vertically so that the average potential between 10 Å and 11 Å was 0 kcal/mol. Block averaging[53] was used to calculate errors. The trajectories in each of the constrained windows were divided into N shorter trajectories. The distribution $P(r)$ for each of the N trajectories was then calculated using the methods described above. The resulting N $P(r)$ were averaged and used to calculate the standard deviation, which we report as the error. N=20 was chosen to minimize the correlation between neighboring blocks.

Implicit Solvent Continuum Electrostatics

Implicit solvent calculations were performed using the DelPhi program[54] to solve the linearized Poisson-Boltzman equation. In order to compare with the explicit solvent PMFs, DelPhi calculations were performed on each configuration used in the MD simulations except that the molecules were fixed at a defined distance from 2.5 Å to 11.0 Å at 0.25 Å intervals. The Coulombic and solvation (reaction field) components of the free energy from DelPhi were added to Lennard-Jones energies calculated separately to obtain the implicit solvent PMF. The nonpolar component of the solvation free energy was computed with a solvent-accessible surface area model. The partial charges and atomic radii in the explicit and implicit solvent simulations were the same; that is, we did not use the default charges and radii from Delphi, in order to compare more directly with the molecular dynamics results. The Delphi calculations used 4 grid points per Å, an internal dielectric of 1, an external dielectric of 80, and an ionic strength of zero.

Quantum Mechanics

Quantum mechanics calculations were performed using the Jaguar software package[55]. *Ab initio* single-point energy calculations were performed on the same coordinates employed for the implicit solvent molecular mechanics results. A self-consistent reaction field (SCRF) method[56] was used to mimic the condensed phase environment. The procedure started by calculating atomic charges for the molecule in a vacuum using electrostatic fitting[57, 58]. This step entailed a Hartree-Fock calculation and subsequent electron correlation correction to evaluate the electrostatic potential. The response from the surrounding dielectric and corresponding surface charges was calculated. Atomic charges were then re-calculated taking into account the dielectric response. The solvation energy was calculated at each iteration until it converged. The basis set for the Hartree-Fock and electron correlation calculations was cc-pVTZ(-f). Electron correlation was treated at the level of local Moller-Plesset second order perturbation theory (LMP2)[21].

Results and Discussion

Each level of theory employed in this study captures different aspects of hydrogen bonding with varying computational expense. Although not free from error, the explicit solvent molecular dynamics results include substantial averaging over conformational and rotational degrees of freedom for both solute and solvent, and comprise the core of our results. Our explicit solvent PMFs take a form typical to those of oppositely charged ions (e.g., Figure 2). The lowest free energy is generally seen when the ions are directly in contact (separation distance roughly equal to the sum of the van der Waals radii); we refer to this minimum as the “contact minimum”. As the separation between the ions increases, the free energy rises sharply to a solvation barrier. In many cases, as the separation between the molecules increases further the free energy reaches a second minimum, in which the solute ions are separated by approximately one water molecule, which we refer to as the “solvent-separated minimum”. In cases with particularly well-ordered waters, a second barrier and second solvent-separated minimum may exist with further separation. At distances beyond these features the potential energy approaches zero. The free energy of the hydrogen bond or salt bridge is calculated as the difference in energy between the largest separation sampled (11 Å) and the contact minimum. The free energies of all orientations, charge states, and acceptor–donor pairs as calculated by MD in explicit solvent are summarized in Table 1.

Control Study of NaCl

As a control, we computed a one-dimensional PMF between Na^+ and Cl^- ions (Supporting Materials), a very well studied system. This experiment allowed for comparisons to previous studies without the rotational degrees of freedom and multiple partial charges inherent to polyatomic systems. Our calculations yielded a free energy at the contact minimum of -1.7 kcal/mol and a free energy at the top of the solvation barrier of 1.7 kcal/mol. Masunov and Lazaridis[15] found a contact minimum free energy of about -1.3 kcal/mol and a solvation barrier free energy of about 2.0 kcal/mol when using a Spherical Solvent Boundary Potential[59] for long-range electrostatics, and -2.0 kcal/mol and 1.5 kcal/mol respectively for the free energy when using Ewald summation. Other comparable studies from Smith and Deng[60], Lyubartsev and Laaksonen[12] (at 0.5 M concentration), and Martorana *et al.*[13] found comparable results, with contact minima of about -1 to -2 kcal/mol.

Effect of Interaction Geometry on Energy Landscapes

Examples of the collinear and coplanar geometries used for the PMFs are depicted in Figure 1. In cases where the phosphate is protonated, the proton is placed on an oxygen not involved in the hydrogen bonding; it serves only to change the overall charge on the phosphate. In the coplanar orientation the molecules are constrained to move along the axis of the distance constraint, and the planarity constraint prevents rotation about this axis. In contrast, the collinear orientation allows for rotation about the collinear axis. The height of the solvation barriers and the stability of the solvent-separated minima depend on the relative rotation of the molecules about the collinear axis, so by integrating out this degree of freedom we produce a PMF with less pronounced maxima, and frequently no secondary minima, in comparison to the coplanar PMFs (all PMFs are available in Supporting Information). Note that the free energy computed for the hydrogen bond in this type of PMF explicitly includes entropic contributions from rotations about the collinear axis.

Salt Bridge Interactions of Glu/Asp

The strengths of salt bridges between positively charged Lys or Arg with negatively charged carboxylate groups on Glu and Asp have been considered in several previous studies[15, 61]. We include our own data here only to obtain internally consistent comparison with the results involving phosphate groups. Taking a collinear approach, propyl guanidinium (Arg) and butyl ammonium (Lys) yield similar free energies as hydrogen bond donors to propionic acid (Glu), at -3.4 kcal/mol and -3.1 kcal/mol respectively. In a planar approach, Arg is much more stable, at -8.5 kcal/mol. The enhanced stability of Arg is largely due to the ability of its guanidinium moiety to form nearly idealized bidentate hydrogen bonds with the carboxylate moiety of Glu. Moreover, the free energy of the bidentate bonds is slightly greater than double that of the single collinear bond, suggesting some cooperativity as there is less of an entropic cost to pay in forming the second hydrogen bond. Entropic cooperativity in bidentate bonding interactions, sometimes referred to as the chelate effect, has been observed elsewhere experimentally[62].

Masunov and Lazaridis computed one-dimensional PMFs between ionizable side chain analogs similar to the configurations we have used[15]. In particular, our Glu–Arg coplanar approach (Supporting Materials) used similar ionic structures and positional restraints as employed by Masunov and Lazaridis in a corresponding calculation. Their analysis yielded a contact minimum free energy and solvation barrier free energy of about -4.3 kcal/mol and 3.0 kcal/mol respectively, roughly half the magnitude of our results. This discrepancy is likely attributable to differences in force fields and simulation protocol. Importantly, their simulations were run with the CHARMM 19 force field[63] for side chain analogs, while ours used the OPLS-AA 2001 force field (see Methods). Notably, the charges for the donated Arg hydrogens are 0.35*e* in CHARMM as opposed to 0.46*e* in OPLS, and the accepting Glu oxygens take a charge of -0.60*e* in CHARMM in contrast to -0.80*e* in OPLS. Further, Masunov and Lazaridis handled long-range electrostatics with a Spherical Solvent Boundary Potential[59] (SSBP) over a spherical cluster of 200 waters with an 11 Å radius, while we employed particle mesh Ewald over a 40 Å cubic box of roughly 2150 water molecules. Figure 2 of the Masunov paper shows that SSBP produces a contact minimum about 0.7 times the depth as Ewald summation when calculating a one-dimensional PMF for Na^+ and Cl^- ions. Rodinger *et al.*[64] found that the interface between an explicit water droplet and a continuum solvent field appearing in models like SSBP polarizes the explicit waters up to 10 Å away from the solvent-vacuum boundary. Lattice summation methods like particle mesh Ewald can also introduce small artifacts related to periodicity-induced perturbations in Coulombic and solvation energies, although given the quantity and permittivity of the solvent in the present study these perturbations should nearly cancel each other[65]. Additionally, the earlier paper reports the typical use of 7 umbrella sampling windows constructed at 1 Å intervals from 3 Å to 9 Å and simulated for 200 ps. In the present study, 18 windows were used, ranging from 2.5 Å to 11 Å at 0.5 Å intervals, and each window was equilibrated for 200 ps followed by 2.1 ns of simulation.

Rozanska and Chipot[61] calculated a PMF for guanidinium and acetate using Ewald lattice summation for long-range electrostatics, which corresponds geometrically to our Glu–Arg coplanar orientation. Their simulations produced a contact minimum free energy of -2.7 kcal/mol and a solvation barrier free energy of 3.4 kcal/mol. The atomic partial charges were computed from quantum mechanics and more closely resembled those of OPLS than CHARMM, with the donated hydrogens at 0.49*e* and the accepting oxygens at -0.87*e*. Importantly, their guanidinium and acetate moieties were biased to face one another through torsional restraints. This would allow some degree of rotation through the plane of interaction, in contrast to the C-C–C-N linearity constraint imposed on the corresponding ions of the present study, which enforces nearly constant bidentate hydrogen bonding. We have found that removing this linearity constraint reduces the stability of the salt bridge by at least 1.5 kcal/mol. Other notable differences include Rozanska and Chipot's use of 4 umbrella sampling windows rather than 18, the AMBER force field[66] for the potential energy function rather than OPLS-AA, and the TIP4P water model rather than TIP3P.

Salt Bridge Interactions of pSer

We now turn our attention to hydrogen bond interactions involving phosphorylated amino acid acceptors, the central results of this work. As is the case with carboxylate serving as the acceptor ion, butyl ammonium (Lys) and propyl guanidinium (Arg) donors yield similar contact minima free energies in a collinear orientation with protonated methyl phosphate ($pSer^{-1}$), at -3.7 kcal/mol. When pSer is deprotonated, the energies of the collinear orientation become more distinguishable, with $pSer^{-2}$–Arg forming a salt bridge worth -4.7 kcal/mol, compared to -4.2 kcal/mol for $pSer^{-2}$–Lys. In the planar approach, Lys and Arg produce substantially different free energy profiles. When the accepting pSer is protonated, the contact minimum reaches -9.3 kcal/mol with Arg. If pSer takes a -2 charge, the free energy for this geometry is -10.6 kcal/mol.

The depth of the pSer–Arg contact minima further demonstrate the significance of bidentate hydrogen bonding in a coplanar approach. Similar to the effects observed with Glu–Arg, a coplanar pSer–Arg salt bridge provides more than twice the stability of one that is collinear. Coplanar pSer–Lys salt bridges yield similar energies to their collinear counterparts, which was also observed with Glu–Lys. PMFs with all permutations of pSer charge states with different donors also suggest that the effects of pSer protonation are small but significant. In a collinear approach, deprotonation of pSer stabilizes a pSer–Lys salt bridge by -0.5 kcal/mol and pSer–Arg by -1.0 kcal/mol. The effect appears stronger in a coplanar approach, with deprotonation stabilizing pSer–Lys by -1.0 kcal/mol and pSer–Arg by -1.3 kcal/mol.

Mavri and Vogel[26] used PM3 semi-empirical molecular orbital calculations with the SM3 reaction field treatment of solvent[67] to investigate the strengths of interactions for methylammonium and methylguanadinium with mono- and divalent methylphosphate in several orientations. Their coplanar calculations correspond geometrically with our coplanar pSer–Lys and pSer–Arg orientations. Although the authors conclude that phosphate interactions with Lys are generally stronger than with Arg, their PM3-SM3 calculations show interaction free energies of +6.1 kcal/mol with coplanar pSer^{-1}-Lys and +3.6 kcal/mol for coplanar pSer^{-1}–Arg. The authors also found that phenyphosphate^{-2} in complex with Lys or Arg produces an interaction free energy stronger than -28.0 kcal/mol in both cases. The findings clearly contradict our results, both those generated using molecular mechanics and using quantum mechanical methods, which are largely consistent with each other. It should be noted that the authors of this study speculated that the semi-empirical quantum mechanical model may not have been well parameterized for phosphate groups.

Luo et al.[14] computed strengths of salt bridge interactions for Arg and Lys with phosphate using the CHARMM 22.0 empirical force field and a generalized Born implicit solvent model. In particular, they computed a PMF between monovalent phosphate and guanidinium that corresponds geometrically to our coplanar pSer^{-1}–Arg orientation. The authors found a contact minimum of about -3.8 kcal/mol, less than half the depth of our contact minimum of -9.3 kcal/mol for pSer^{-1}–Arg. Several differences in simulation protocol may contribute to this discrepancy. The CHARMM force field places a weaker charge on the unprotonated phosphate oxygens (-0.82*e*) than the charge we obtained from electrostatic potential fitting (-1.032*e*). Further, the authors used the protonated phosphate oxygen to accept one of the planar hydrogen bonds, while we used deprotonated oxygens to accept both hydrogen bonds. Additionally, because the authors were trying to mimic experiments involving ion pairs in high ionic strength (1 mol/L) aqueous solution, they applied an ionic shielding correction to their ion pair calculations that weakened their interactions on the order of 1 kcal/mol. Finally, the authors used a generalized Born implicit solvent model instead of Poisson-Boltzmann and explicit solvent used here.

Some efforts have also aimed to quantify the strengths of phosphate interactions with charged side chain analogs experimentally. Springs and Haake[68] extracted free energies of association for guanidinium–phosphate and butylamine–phosphate from pK$_a$ shifts. The absolute free energies cannot be directly compared to our work, because the experimental free energies are for free ions, whereas the computational PMFs represent constrained geometries, which is more appropriate to understanding hydrogen bonding in a macromolecule. In addition, the experiments were carried out in solution with 1 mol/L ionic strength, creating significant ionic shielding. However, the relative free energies can be profitably compared.

Specifically, the experimentally determined free energies show a stronger interaction for guanidinium–phosphate (-0.6 kcal/mol) than butylamine–phosphate (-0.4 kcal/mol), in agreement with our results.

Salt Bridge Interactions of pAsp

Response regulators in bacterial "two component" signaling systems use Asp side chains to accept a phosphate group from a sensor histidine kinase[69]. Resonance structures involving the pi orbitals of the covalently linked carboxylate and phosphate groups suggest that the electron density on the phosphate group on pAsp may be significantly different from that of pSer/pThr[1]. This conjecture is confirmed by the partial charges we obtained by electrostatic potential fitting, as discussed in Methods. We examined the effect of this difference on hydrogen bond strengths with the panel of hydrogen bond donors.

As with propionic acid (Glu) and methyl phosphate (pSer) acceptors, the energies for acetyl phosphate (pAsp) with propyl guanidinium (Arg) and butyl ammonium (Lys) in a collinear approach are very similar (within 1 kcal/mol). The more striking comparison arises from different charge states of pAsp. Deprotonating pAsp when accepting from collinear Lys stabilizes the interaction by -1.4 kcal/mol, and by -1.5 kcal/mol when the donor is Arg. In contrast, deprotonating pSer in a collinear salt bridge stabilizes the interaction by only -0.5 kcal/mol with Lys and -1.0

[1] A similar argument could be made regarding pTyr, i.e., that there could be some conjugation between the pi electronic systems in the benzene ring and on the phosphate. However, quantum calculations on benzyl phosphate followed by electrostatic potential fitting (see Methods) suggested that the electron density on the phosphate group in pTyr is minimally different than in methyl phosphate (pSer), and we did not pursue this issue further.

kcal/mol with Arg. The deprotonation effect increases in the coplanar approach only for lysine. Coplanar pAsp^{-2}–Lys shows a 2.5 kcal/mol stabilization over pAsp^{-1}–Lys, while coplanar pAsp^{-2}–Arg yields only a 1.0 kcal/mol stabilization over pAsp^{-1}–Arg. Bidentate hydrogen bonding continues to produce strong effects, with coplanar Arg showing a 3.3 kcal/mol stronger salt bridge than a collinear approach with pAsp^{-1} and a 2.8 kcal/mol stronger interaction than collinear with pAsp^{-2}. Coplanar Arg also bonds stronger than coplanar Lys to pAsp in the -1 and -2 charge states by 3.9 kcal/mol and 2.4 kcal/mol respectively, as Lys cannot form bidentate bonds due to the same geometric constraints inhibiting them with Glu and pSer acceptors.

The greater sensitivity to charge state of pAsp over pSer might be attributed to differences in the charge distribution for the -1 and -2 ions. The quantum mechanically calculated partial charges for the acceptor oxygens on pSer remain at -1.032*e* regardless of protonation of the remaining phosphate oxygen. In contrast, partial charges of the corresponding oxygens of pAsp decrease to -1.016*e* from -0.949*e* upon phosphate deprotonation. We attribute these differences in partial charges to protonation effects on electron density due to resonance between the carboxylate and phosphate pi-electron systems; this effect of course does not occur in methyl phosphate (pSer).

Hydrogen Bond Interactions with Amide NH Groups

An earlier survey (data not shown) identified backbone amides as the second most common hydrogen bond partner after Arg with phosphates in phosphorylated proteins in the Protein Data Bank[70]. We used N-methylacetamide (CH_3–NH–CO–CH_3) as an analog of the protein backbone to investigate the free energy of backbone hydrogen bonds to the carboxylate and phosphate hydrogen bond acceptors. In all simulations the amide hydrogen was placed in a collinear geometry with its acceptor. The glutamate analog was truncated to acetate because hydrophobic interactions were observed between the aliphatic tail of Glu and the methyl caps of N-methylacetamide. In contrast to the behavior observed with Arg or Lys donors, the interaction weakens slightly when the acceptor is deprotonated. A weakened hydrogen bond arising from a stronger P–O dipole may appear counterintuitive. However, hydrogen bond formation depends on a delicate balance between the free energy gain of bonded pairs and the loss of hydrogen bonds to surrounding waters, and the desolvation penalty is significantly lower for the protonated phosphate group. This effect was also observed by Wong *et al.*[41] in a study involving phosphate–amide interactions with phosphate acceptors possessing both -1 and -2 charges.

Implicit Solvent Poisson-Boltzmann Calculations

While continuum solvent models can provide substantial speed advantages relative to explicit solvent in performing free energy calculations, the merits and shortcomings of implicit solvent models remain a subject of interest and some contention. Treating the solvent as a dielectric continuum neglects important first-shell solvation effects arising from the finite size and asymmetry of a water molecule[18-20]. In particular, the very strong ionic interactions between a -2 charged phosphate with positively charged ions presents a challenging test of implicit solvent models.

We performed Poisson-Boltzmann (PB) implicit solvent calculations on all configurations (see Methods) to compare with the explicit solvent MD simulations. We retained the same geometries, atomic radii, Lennard-Jones parameters, and partial charges in these calculations. Figure 2 shows one example of a comparison between the explicit solvent PMF and the implicit solvent results; the others are available in Supporting Information. As has been seen in other work[15, 18], the implicit solvent potentials generally contain less structure than the explicit solvent PMFs, i.e., no secondary minima, consistent with the fact that the implicit solvent model treats water as a continuum. In addition, we observe that the implicit solvent results tend to exaggerate the energy barrier required for separating the ions from contact to infinite separation.

Our primary concern here, however, is the depth of the contact minima (Table 2), and in this respect the implicit solvent results generally recapitulate most of the key trends observed in the explicit solvent PMFs. In particular, the implicit solvent results agree that protonating the phosphate group weakens hydrogen bonds with charged donors but strengthens interactions with the amide NH group, and that the strongest hydrogen bond of the phosphate group is the bidentate interaction with guanidinium. Overall, the implicit solvent calculations predict stronger hydrogen bonding interactions of the phosphate group than explicit solvent MD, with the largest discrepancies observed for Arg forming bidentate interactions with unprotonated phosphate. It may be possible to reduce the general over-prediction of the hydrogen bond strengths by empirically adjusting the radii used to define the dielectric surface in the implicit solvent calculation, but we have not performed such an optimization at this time. It should also be reiterated that the explicit solvent PMFs cannot be considered to be free of error either and will depend on the choice of explicit solvent model and other simulation parameters.

Self-Consistent Reaction Field Quantum Mechanics Calculations

To investigate the possible effects of electronic polarization on the energetics of hydrogen bonding we performed quantum mechanics (QM) calculations using a self-consistent reaction field to mimic the condensed phase (see Methods). One limitation of this method with respect to the MD calculations is the use of implicit solvent. An advantage, however, is that atomic partial charges are recomputed at each distance to account for electronic polarizability. For instance, the donated Arg hydrogens in the pSer^{-2}–Arg coplanar configuration increase in charge from 0.54*e* and 0.51*e* at 11.0 Å separation to 0.69*e* and 0.62*e* at the contact minimum distance of 4.25 Å.

As with the PB analysis, the QM calculations were carried out on fixed orientations at 0.25 Å intervals. The total interaction energies, computed from the difference between potentials at 11.0 Å separation and the contact minima, are presented in Table 3. A superposition of MD, PB, and QM energy landscapes is shown in Figure S3. As observed in the explicit solvent MD results, bidentate hydrogen bonds with coplanar arginine tend to produce salt bridges about twice as strong as the monodentate collinear approach. On average, QM found 1.5 kcal/mol stronger interactions than MD with a fixed charge force field and explicit solvent. The largest differences occur for configurations with a -2 charged receptor. The mean difference in free energy between QM and MD for -2 charged receptors is -3.1 kcal/mol, while the same figure for -1 charged receptors is -0.9 kcal/mol. It is of course reasonable that the larger charge on the -2 anions would induce larger polarization effects.

Since the quantum mechanical and PB calculations both employ an implicit solvent model, it is informative to compare trends between these methods as well. The average difference between QM and PB is -1.4 kcal/mol, similar to the difference observed when comparing to MD. As expected, the QM calculations are substantially more sensitive to polarization effects than PB. The mean free energy difference between QM and PB for -2 charged receptors is -3.0 kcal/mol, while for -1 charged receptors it is 1.0 kcal/mol.

Overall, the quantum mechanical calculations show a significant role for polarization in hydrogen bond stability, and suggest that the explicit solvent molecular dynamics simulations, using a fixed charge force field, might systematically underestimate the strengths of hydrogen bonds involving the phosphate group with a -2 charge, relative to -1 phosphate or carboxylate groups. Generally, the quantum mechanical calculations support the conclusions from the explicit solvent molecular dynamics. One key limitation of the quantum calculations, however, is that solvent is treated as a dielectric continuum, as in the implicit solvent results. Quantum mechanical simulations are possible with explicit solvent, but extensive sampling of the water is required to obtain reasonable free energies of solvation, making this approach computationally extremely intensive. A more tractable way to assess electronic polarizability effects in explicit solvent may be to perform molecular dynamics using the new generation of polarizable force fields[71] with polarizable explicit water[72].

Conclusions

Calculating hydrogen bond free energies using multiple levels of theory has provided an internally consistent survey of hydrogen bond strengths for common hydrogen bonding partners, charge states, and geometries involving phosphorylated amino acid side chains. Additionally, the results suggest relative merits and shortcomings of each level of theory for this application. We conclude now by returning to the issues raised in the introduction:

(1) *The conditions under which Arg or Lys make stronger hydrogen bonds with a phosphorylated side chain.* Lys forms as strong or slightly stronger hydrogen bonds than Arg with most of the acceptors studied in the collinear approach, probably because the e-amino group of lysine has a denser positive charge field than the arginine guanidinium moiety. However, the results are unambiguous that the bidentate interactions available to guanidinium (Arg) with phosphate provide much stronger interactions than can be formed between ammonium ions (Lys) and phosphate in either a monodentate or bidentate geometry.

(2) *The effect of phosphate protonation state on hydrogen bond strength.* Phosphate protonation (i.e., to the -1 charge state) produces a small but significant destabilizing effect with Arg and Lys donors (~1 kcal/mol for pSer), particularly when the acceptor is pAsp (up to 2.5 kcal/mol). In contrast, the interactions with the amide NH group were mildly stabilized by acceptor protonation (~0.6 kcal/mol); this is consistent with previous work of Wong *et al.*[41]. Altogether, however, the hydrogen bond strengths of the phosphate groups in the -2 and -1 charge states are strikingly

similar. This is remarkable because the hydrogen bond strength results largely from near-cancellation of two very large quantities: the strong Coulombic attraction between the ions, and the dielectric screening (and first-shell solvation effects) exerted by the water. Changing the charge state of the phosphate ion perturbs both of these quantities significantly, but apparently the changes are such that the overall strengths of the hydrogen bonds are not greatly affected.

(3) *The energetic consequences of substituting a carboxylate for a phosphate.* All of the pSer^{-2} orientations with charged residue donors form stronger salt bridges than these charged residue donors do with glutamate suggesting that Asp/Glu substitution might not mimic phosphorylation. In contrast, the strengths of the hydrogen bonds of the phosphate groups in the -1 charge state are generally closer to the corresponding hydrogen bonds of the carboxylate group, especially for pAsp^{-1}. Table 4 lists hydrogens bonds with phosphate acceptors that are at least 0.5 kcal/mol as strong when the acceptor is a carboxylate. Of course, in the protein microenvironment the local electrostatic field, steric restrictions, exposure to various solvent and ion concentrations, departure from ideal orientations, and other factors will significantly impact the hydrogen bond strengths computed here. Nevertheless, these calculations provide a quantitative framework for beginning to assess when substitution with Glu or Asp might mimic phosphorylation, at least when a protein structure is available, and suggest that the protonation state of the phosphate may be a critical parameter.

References

1. Audette, G. F.; Engelmann, R.; Hengstenberg, W.; Deutscher, J.; Hayakawa, K.; Quail, J. W.; Delbaere, L. T. *J Mol Biol* **2000,** 303, (4), 545-53.
2. Patel, A. J.; Honore, E. *Trends Neurosci* **2001,** 24, (6), 339-46.
3. Martens, J. R.; Kwak, Y. G.; Tamkun, M. M. *Trends Cardiovasc Med* **1999,** 9, (8), 253-8.
4. Vermeulen, K.; Van Bockstaele, D. R.; Berneman, Z. N. *Cell Prolif* **2003,** 36, (3), 131-49.
5. Feng, M. H.; Philippopoulos, M.; MacKerell, A. D.; Lim, C. *Journal of the American Chemical Society* **1996,** 118, (45), 11265-11277.
6. Johnson, L. N.; Lewis, R. J. *Chemical Reviews* **2001,** 101, (8), 2209-2242.
7. Johnson, L. N.; Oreilly, M. *Current Opinion in Structural Biology* **1996,** 6, (6), 762-769.
8. Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. *Biochemistry* **1990,** 29, (9), 2403-2408.
9. Sun, D. P.; Sauer, U.; Nicholson, H.; Matthews, B. W. *Biochemistry* **1991,** 30, (29), 7142-53.
10. Schneider, J. P.; Lear, J. D.; DeGrado, W. F. *Journal of the American Chemical Society* **1997,** 119, (24), 5742-5743.
11. Waldburger, C. D.; Schildbach, J. F.; Sauer, R. T. *Nature Structural Biology* **1995,** 2, (2), 122-128.
12. Lyubartsev, A. P.; Laaksonen, A. *Physical Review E* **1997,** 55, (5), 5689-5696.
13. Martorana, V.; La Fata, L.; Bulone, D.; San Biagio, P. L. *Chemical Physics Letters* **2000,** 329, (3-4), 221-227.
14. Luo, R.; David, L.; Hung, H.; Devaney, J.; Gilson, M. K. *Journal of Physical Chemistry B* **1999,** 103, 727-736.
15. Masunov, A.; Lazaridis, T. *Journal of the American Chemical Society* **2003,** 125, 1722-1730.
16. Gilson, M. K. *Current Opinion in Structural Biology* **1995,** 5, (2), 216-223.
17. Ghosh, A.; Rapp, C. S.; Friesner, R. A. *Journal of Physical Chemistry B* **1998,** 102, (52), 10983-10990.
18. Chorny, I.; Dill, K. A.; Jacobson, M. P. *Journal of Physical Chemistry B* **2005,** 2005, 24056-24060.
19. Asthagiri, D.; Schure, M. R.; Lenhoff, A. M. *Journal of Physical Chemistry B* **2000,** 104, 8753-8761.
20. Yu, Z.; Jacobson, M. P.; Rapp, C. S.; Friesner, R. A. *Journal of Chemical Physics* **2004,** 108, 6643-6654.
21. Saebo, S.; Pulay, P. *Annual Review of Physical Chemistry* **1993,** 44, 213-236.
22. Singh, J.; Thornton, J. M.; Snarey, M.; Campbell, S. F. *FEBS Letters* **1987,** 224, (1), 161-171.
23. Saenger, W.; Wagner, K. G. *Acta Crystallographica* **1972,** B28, 2237-2244.
24. Lewin, S. *Journal of Theoretical Biology* **1969,** 23, (2), 279-284.
25. Schug, K. A.; Lindner, W. *Chemical Reviews* **2005,** 105, (1), 67-113.

26. Mavri, J.; Vogel, H. J. *Proteins-Structure Function and Genetics* **1996,** 24, (4), 495-501.
27. Deerfield, D. W.; Nicholas, H. B.; Hiskey, R. G.; Pedersen, L. G. *Proteins-Structure Function and Genetics* **1989,** 6, (2), 168-192.
28. Charbon, G.; Breunig, K. D.; Wattiez, R.; Vandenhaute, J.; Noel-Georis, I. *Mol Cell Biol* **2004,** 24, (10), 4083-91.
29. Kassenbrock, C. K.; Anderson, S. M. *J Biol Chem* **2004,** 279, (27), 28017-27.
30. Huang, W.; Erikson, R. L. *Proc Natl Acad Sci U S A* **1994,** 91, (19), 8960-3.
31. Klose, K. E.; Weiss, D. S.; Kustu, S. *J Mol Biol* **1993,** 232, (1), 67-78.
32. McCabe, T. J.; Fulton, D.; Roman, L. J.; Sessa, W. C. *J Biol Chem* **2000,** 275, (9), 6123-8.
33. Zhang, J.; Zhang, F.; Ebert, D.; Cobb, M. H.; Goldsmith, E. J. *Structure* **1995,** 3, (3), 299-307.
34. Mansour, S. J.; Candia, J. M.; Matsuura, J. E.; Manning, M. C.; Ahn, N. G. *Biochemistry* **1996,** 35, (48), 15529-36.
35. Garcia-Echeverria, C. *Current Medicinal Chemistry* **2001,** 8, (13), 1589-1604.
36. Cody, W. L.; Lin, Z. W.; Panek, R. L.; Rose, D. W.; Rubin, J. R. *Current Pharmaceutical Design* **2000,** 6, (1), 59-98.
37. Calnan, B. J.; Tidor, B.; Biancalana, S.; Hudson, D.; Frankel, A. D. *Science* **1991,** 252, (5010), 1167-71.
38. Frigyes, D.; Alber, F.; Pongor, S.; Carloni, P. *Journal of Molecular Structure (Theochem)* **2001,** 574, 39-45.
39. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J Comput Chem* **2005,** 26, (16), 1701-18.
40. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *Journal of Physical Chemistry B* **2001,** 105, 6474-6487.
41. Wong, S. E.; Bernacki, K.; Jacobson, M. P. *Journal of Physical Chemistry B* **2005,** 109, (11), 5249-5258.
42. Groban, E. S.; Narayanan, A.; Jacobson, M. P. *PLoS Computational Biology* **2006**, 2, 534-553.
43. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *Journal of Chemical Physics* **1983,** 79, (2), 926-935.
44. Darden, T.; York, D.; Pedersen, L. *Journal of Chemical Physics* **1993,** 98, (12), 10089-10092.
45. Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *Journal of Chemical Physics* **1982,** 76, (1), 637-649.
46. Verlet, L. *Physical Review* **1967,** 159, (1), 98.
47. Hoover, W. G. *Physical Review. A* **1985,** 31, (3), 1695-1697.
48. Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, a.; Haak, J. R. *Journal of Chemical Physics* **1984,** 81, (8), 3684-3690.
49. Torrie, G. M.; Valleau, J. P. *Journal of Computational Physics* **1977,** 23, 187-199.
50. Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *Journal of Computational Chemistry* **1997,** 18, (12), 1463-1472.
51. Souaille, M.; Roux, B. *Comput. Phys. Commun* **2001,** 135, 40-57.

52. Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P.; Rosenberg, J. M. *Journal of Computational Chemistry* **1992,** 13, (8), 1011-1021.
53. Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids.* Oxford University Press: 1989.
54. Nicholls, a.; Honig, B. *Journal of Computational Chemistry* **1991,** 12, (4), 435-445.
55. *Jaguar*, 5.0; Schrodinger, L.L.C.: Portland, OR, 1991-2003.
56. Tannor, D. J.; Marten, B.; Murphy, R.; Friesner, R. A.; Sitkoff, D.; Nicholls, A.; Ringnalda, M.; Goddard, W. A.; Honig, B. *Journal of the American Chemical Society* **1994,** 116, (26), 11875-11882.
57. Chirlian, L. E.; Francl, M. M. *Journal of Computational Chemistry* **1987,** 8, (6), 894-905.
58. Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H. *Journal of Computational Chemistry* **1990,** 11, (3), 297-310.
59. Beglov, D.; Roux, B. *Journal of Chemical Physics* **1994,** 100, (12), 9050-9063.
60. Smith, D. E.; Dang, L. X. *Journal of Chemical Physics* **1994,** 100, (5), 3757-3766.
61. Rozanska, X.; Chipot, C. *Journal of Chemical Physics* **2000,** 112, (22), 9691-9694.
62. Breslow, R.; Belvedere, S.; Gershell, L.; Leung, D. *Pure Appl. Chem.* **2000,** 72, (3), 333-342.
63. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983,** 4, (2), 187-217.
64. Rodinger, T.; Howell, P. L.; Pomes, R. *Journal of Chemical Physics* **2005,** 123, (3).
65. Hunenberger, P. H.; McCammon, J. A. *Biophysical Chemistry* **1999,** 78, (1-2), 69-88.
66. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *Journal of the American Chemical Society* **1995,** 117, (19), 5179-5197.
67. Cramer, C. J.; Truhlar, D. G. *Journal of Computer-Aided Molecular Design* **1992,** 6, (6), 629-666.
68. Springs, B.; Haake, P. *Bioorganic Chemistry* **1977,** 6, (2), 181-190.
69. Hoch, J. A.; Silhavy, T. J., *Two-Component Signal Transduction.* ASM Press: Washington, DC, 1995.
70. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res* **2000,** 28, (1), 235-42.
71. Halgren, T. A.; Damm, W. *Current Opinion in Structural Biology* **2001,** 11, (2), 236-242.
72. Yu, H.; van Gunsteren, W. F. *Journal of Chemical Physics* **2004,** 121, (19), 9549-9564.

Chapter 2: Table 1. Hydrogen bonding free energies with standard errors in kcal/mol computed from explicit solvent MD and WHAM.

	Lys Collinear	**Arg Collinear**	**Lys Coplanar**	**Arg Coplanar**	**Amide NH Collinear**
Glu	-3.1±0.3	-3.4±0.2	-2.6±0.4	-8.5±0.1	-1.8±0.6
pSer(-1)	-3.7±0.2	-3.7±0.3	-3.5±0.1	-9.3±0.4	-1.6±0.7
pSer(-2)	-4.2±0.3	-4.7±0.3	-4.5±0.3	-10.6±0.6	-1.0±0.6
pAsp(-1)	-3.2±0.3	-3.0±0.4	-2.4±0.4	-6.3±0.4	-1.8±0.7
pAsp(-2)	-4.6±0.2	-4.5±0.3	-4.9±0.4	-7.3±0.2	-1.1±0.6

Chapter 2: Table 2. Hydrogen bonding free energies in kcal/mol computed from continuum electrostatics.

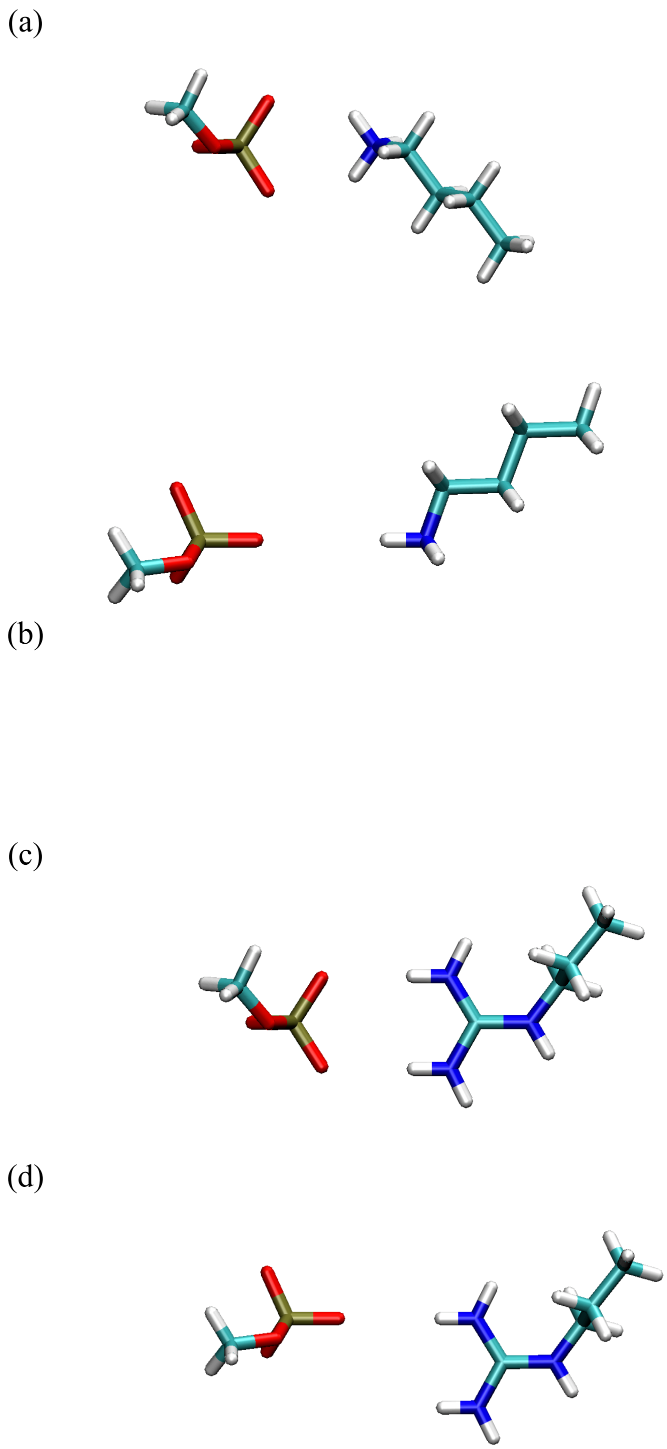
	Lys Collinear	**Arg Collinear**	**Lys Coplanar**	**Arg Coplanar**	**Amide NH Collinear**
Glu	-5.1	-3.9	-2.7	-10.6	-2.3
pSer(-1)	-6.7	-6.1	-3.6	-13.0	-2.1
pSer(-2)	-7.7	-6.7	-5.6	-15.4	-1.2
pAsp(-1)	-5.5	-5.1	-2.4	-10.8	-2.6
pAsp(-2)	-7.2	-6.5	-5.8	-15.5	-1.3

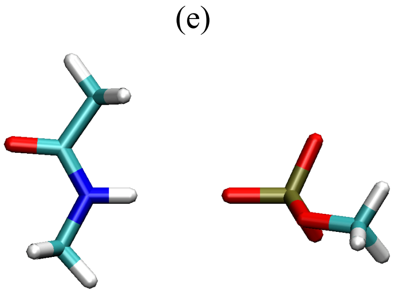
Chapter 2: Table 3. Hydrogen bonding free energies in kcal/mol computed from SCRF quantum mechanics.

	Lys Collinear	Arg Collinear	Lys Coplanar	Arg Coplanar	Amide NH Collinear
Glu	-4.5	-3.7	-4.3	-10.2	-2.1
pSer(-1)	-4.4	-4.1	-2.5	-8.1	-0.8
pSer(-2)	-8.6	-6.8	-8.8	-13.6	-2.6
pAsp(-1)	-3.8	-3.5	-2.4	-8.2	-0.8
pAsp(-2)	-7.8	-6.4	-8.5	-12.6	-2.7

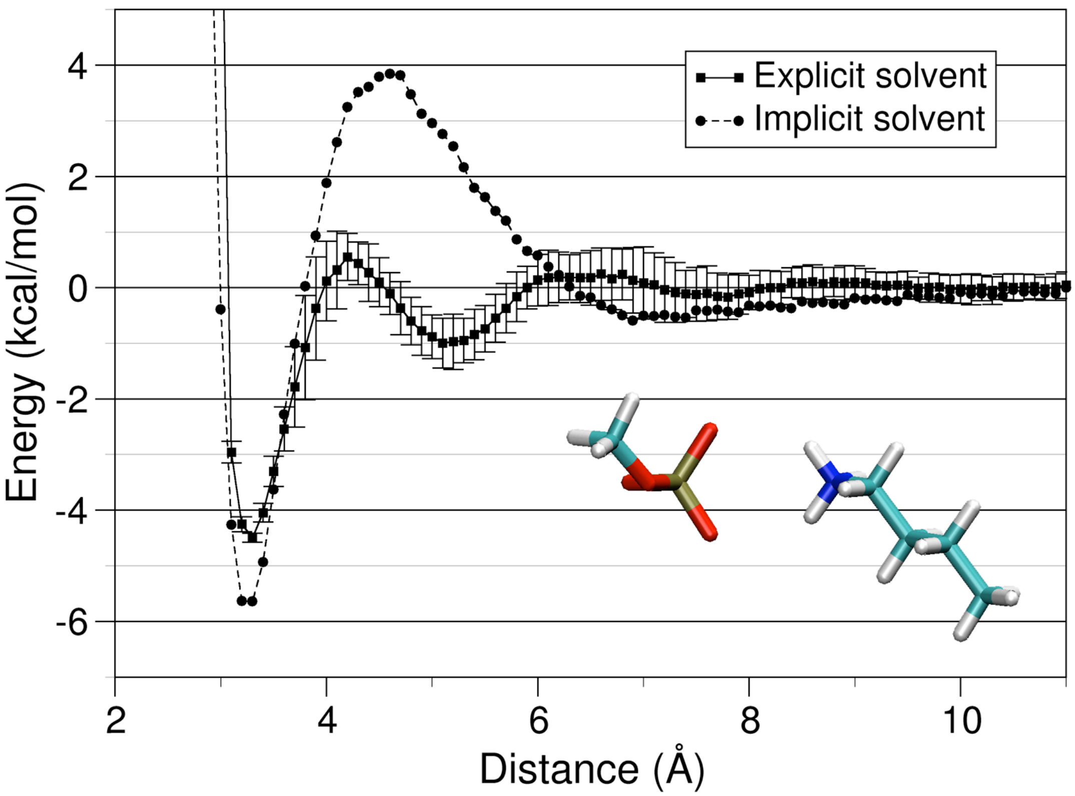
Chapter 2: Table 4. Change in salt bridge free energy (kcal/mol) when a carboxylate is substituted for a phosphate. Charged–charged ion pairs with carboxylate acceptor substitutions worth at least 0.5 kcal/mol as with a phosphate acceptor are shown.

Ion Pair	Orientation	ΔΔG with Glu acceptor
pAsp(-1)–Lys	Collinear	0.1
pAsp(-1)–Lys	Coplanar	-0.2
pAsp(-1)–Arg	Collinear	-0.5
pAsp(-1)–Arg	Coplanar	-2.2
pAsp(-2)–Arg	Coplanar	-1.2
pSer(-1)–Arg	Collinear	0.3


(a)
(b)
(c)
(d)



Chapter 2: Figure 1. Hydrogen bonding geometries considered in this work. Only the unprotonated methyl phosphate acceptor (representing the pSer side chain with a -2 charge) is shown. In the case of protonated phosphate groups, the hydrogen is placed on one of the oxygens not directly involved in hydrogen bonding. We refer to these geometries as (a) “Lys coplanar”, (b) “Lys collinear”, (c) “Arg coplanar”, (d) “Arg collinear”, and (e) “amide NH collinear”.



Chapter 2: Figure 2. PMFs for a coplanar interaction between methyl phosphate (representing pSer^{-2}) and butyl ammonium (Lys) computed using MD in explicit solvent (squares with error bars, computed as described in Methods), and using molecular mechanics with implicit solvent (dotted line and circles). The distance along the x-axis is measured between the phosphorus atom and the ammonium nitrogen. The PMFs for the other 24 hydrogen bonding interactions are available in Supporting Materials.

Supporting Information Available: Partial atomic charges and OPLS atom types used in the calculations; explicit solvent PMF for the NaCl control calculation; sample PMF using the quantum SCRF method, compared with implicit and explicit solvent molecular mechanics results; explicit and implicit solvent PMFs for all 25 hydrogen bonding interactions, with insets depicting the geometries.

Phosphate partial charges and OPLS atom types

CH3-PO4H(-1) (protonated pSer)

atom name	OPLS atom type	quantum mechanically calculated charge
P	opls_445	1.500e
OD1	opls_447	-0.505e
OP1	opls_446	-1.032e
OP2	opls_446	-1.032e
OP3	opls_434	-0.750e
HO1	opls_155	0.500e
CD	opls_157	0.139e
HD1	opls_140	0.060e
HD2	opls_140	0.060e
HD3	opls_140	0.060e

CH3-PO4(-2) (unprotonated pSer)

atom name	OPLS atom type	quantum mechanically calculated charge
P	opls_445	1.500e
OD1	opls_447	-0.723e
OP1	opls_446	-1.032e
OP2	opls_446	-1.032e
OP3	opls_446	-1.032e
CD	opls_157	0.139e
HD1	opls_140	0.060e
HD2	opls_140	0.060e
HD3	opls_140	0.060e

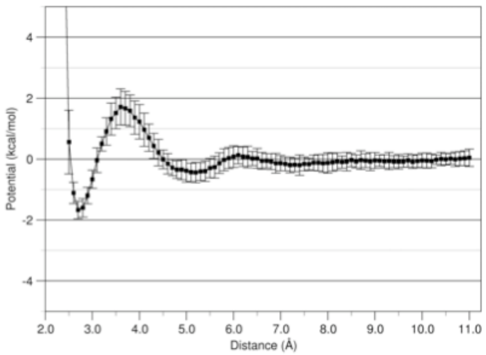
CH3-COO-PO3H(-1) (protonated pAsp)

atom name	OPLS atom type	quantum mechanically calculated charge
CB	opls_274	-0.436e
HB1	opls_140	0.155e
HB2	opls_140	0.155e
HB3	opls_140	0.155e
CG	opls_271	0.870e
OD1	opls_272	-0.540e
OD2	opls_272	-0.671e
P	opls_445	1.487e
O1P	opls_446	-0.949e
O2P	opls_446	-0.949e
O3P	opls_434	-0.787e
HO1	opls_155	0.510e

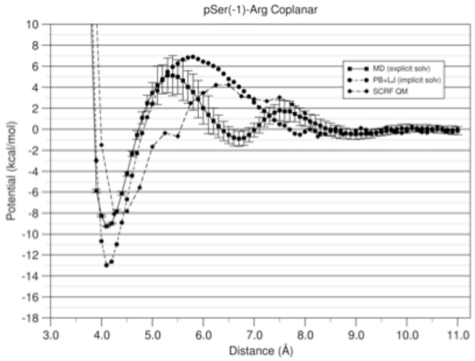
CH3-COO-PO3(-2) (unprotonated pAsp)

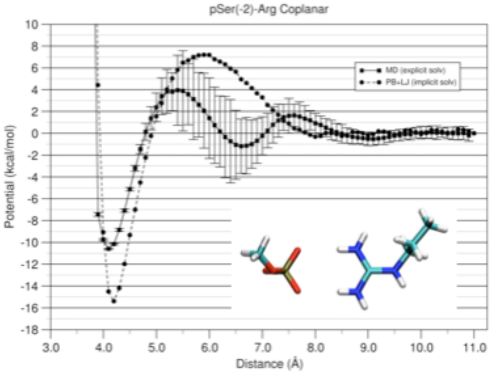
atom name	OPLS atom type	quantum mechanically calculated charge
CB	opls_274	-0.465e
HB1	opls_140	0.146e
HB2	opls_140	0.146e
HB3	opls_140	0.146e
CG	opls_271	0.936e
OD1	opls_272	-0.588e
OD2	opls_272	-0.732e
P	opls_445	1.459e
O1P	opls_446	-1.016e
O2P	opls_446	-1.016e
O3P	opls_446	-1.016e

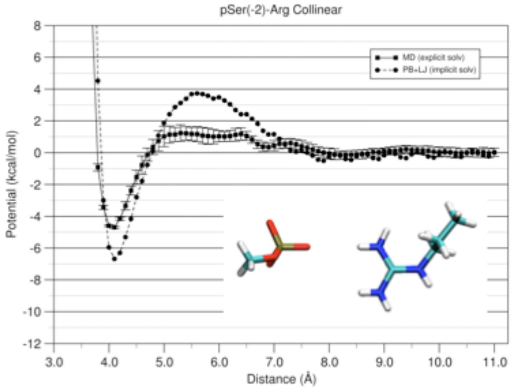
Control simulation of NaCl PMF

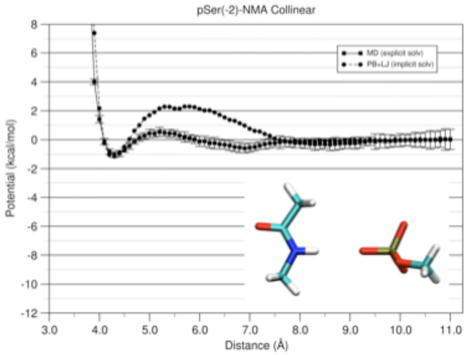


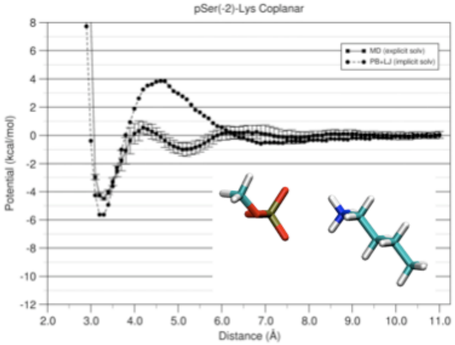
Comparison of explicit and implicit solvent molecular mechanics PMFs with quantum mechanical results with SCRF solvent

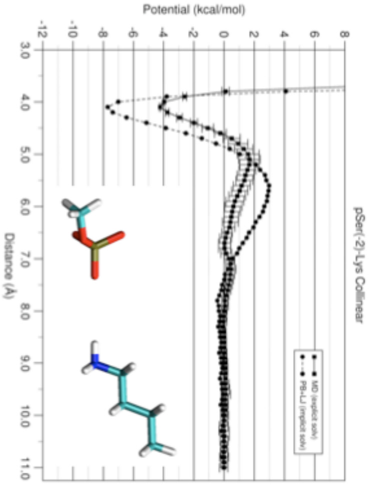

pSer(-1)-Arg Coplanar

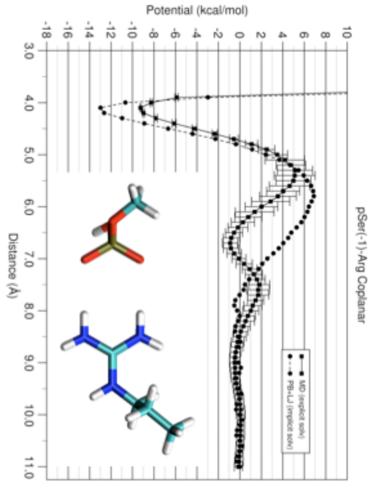

pSer(-2)-Arg Coplanar

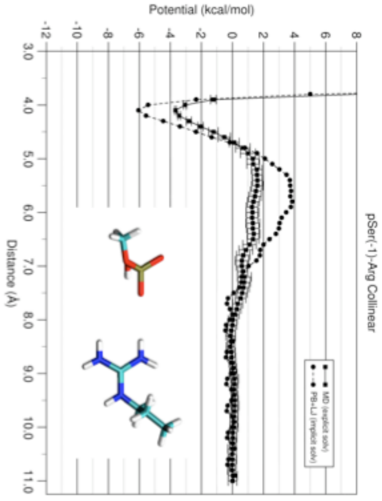

pSer(-2)-Arg Collinear

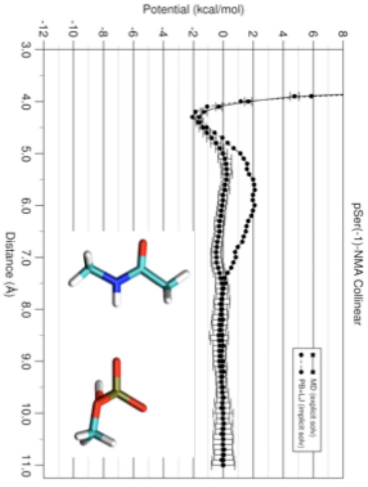

pSer(-2)-NMA Collinear

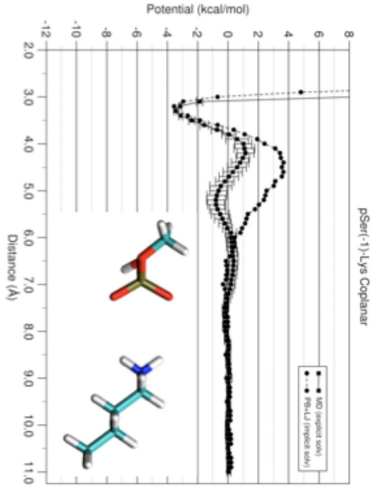

pSer(-2)-Lys Coplanar

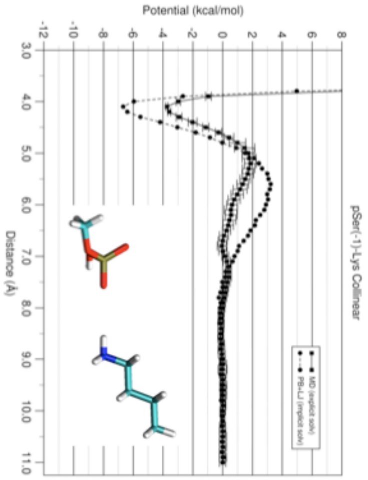

pSer(-2)-Lys Collinear

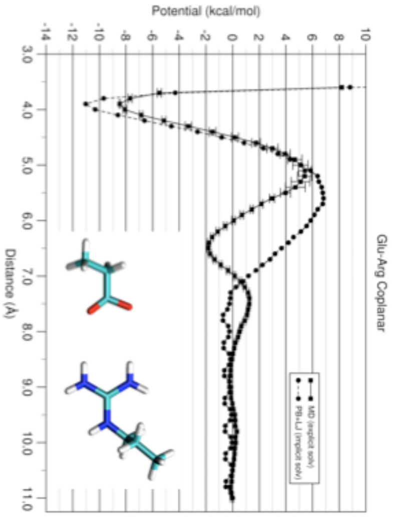

pSer(-1)-Arg Coplanar

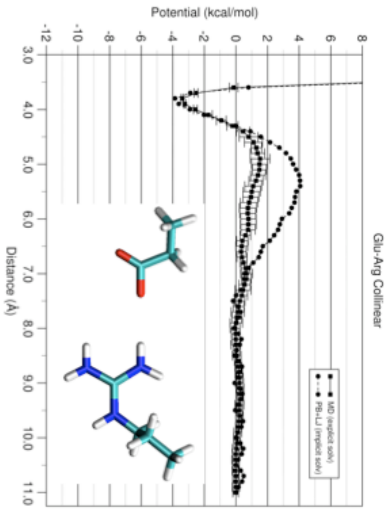

pSer(-1)-Arg Collinear

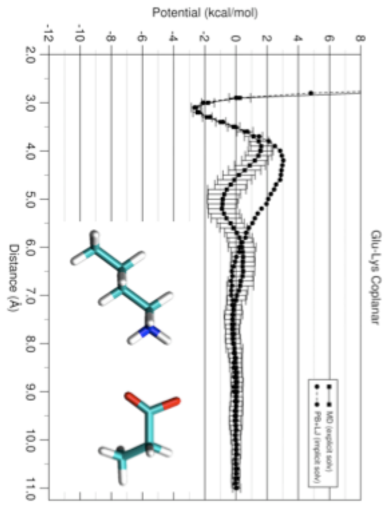

pSer(-1)-NMA Collinear

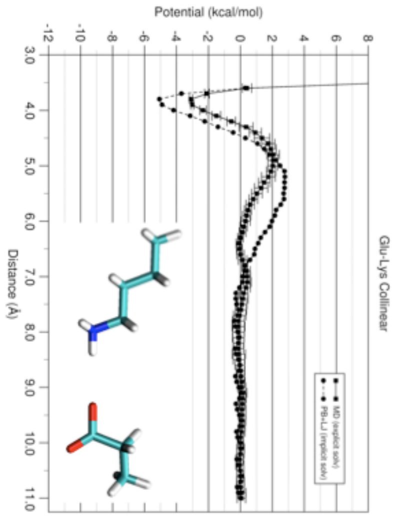

pSer(-1)-Lys Coplanar

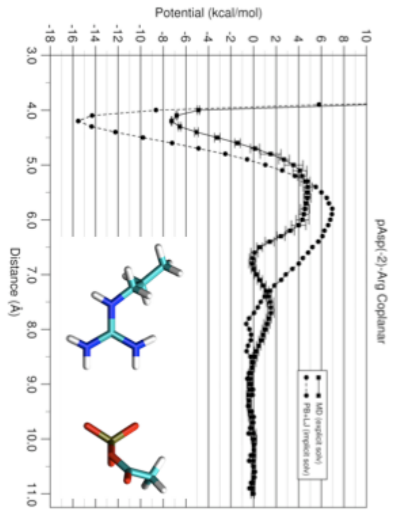

pSer(-1)-Lys Collinear

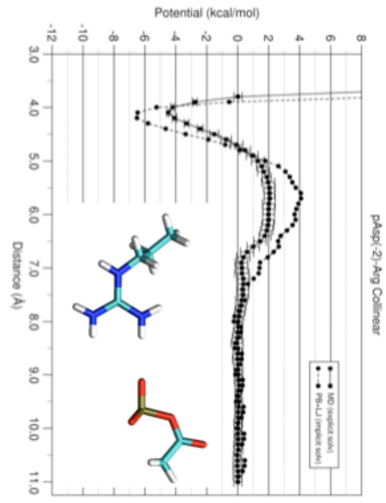

Glu-Arg Coplanar
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

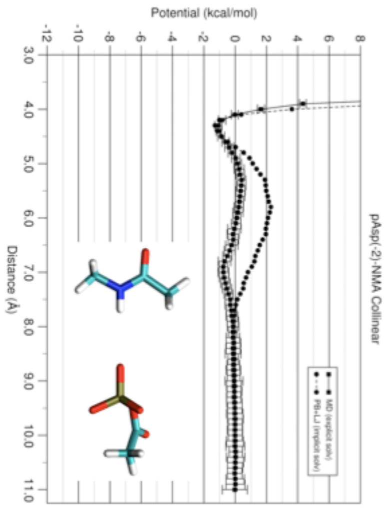

Glu-Arg Collinear
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

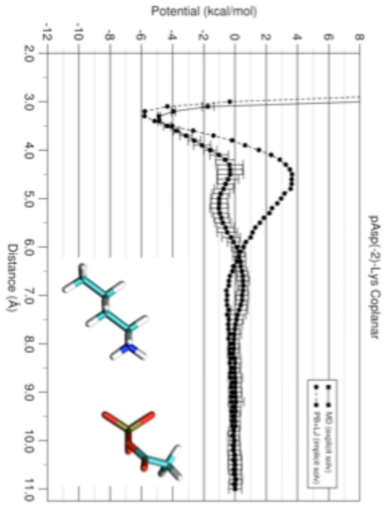

Glu-Lys Coplanar
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

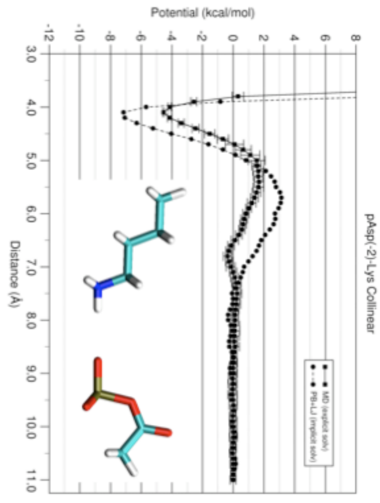

Glu-Lys Collinear
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

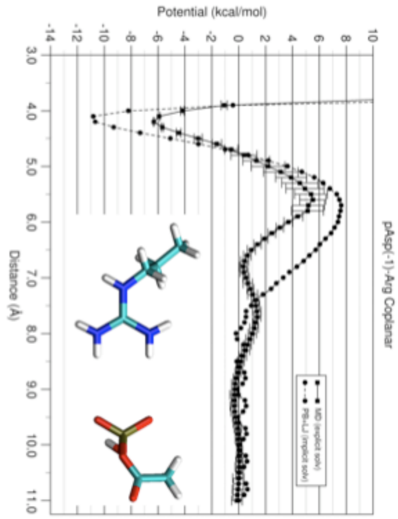

pAsp(-2)-Arg Coplanar
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

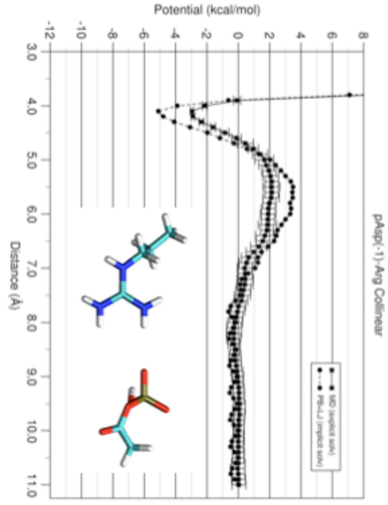

pAsp(-2)-Arg Collinear
Potential (kcal/mol)
Distance (Å)
MD (explicit solv)
PB-LJ (implicit solv)

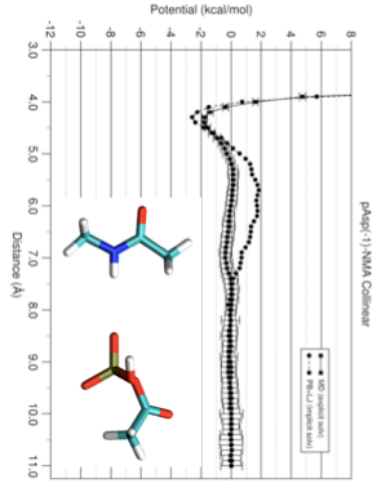

pAsp(-2)-NMA Collinear
Potential (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)

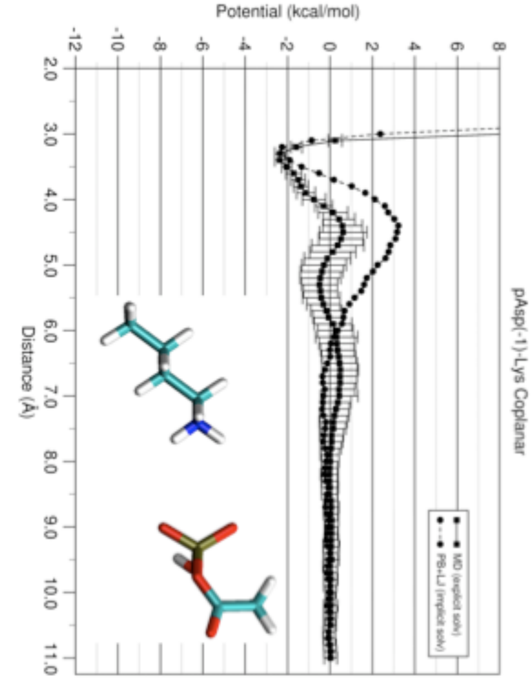

pAsp(-2)-Lys Coplanar
Potential (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)

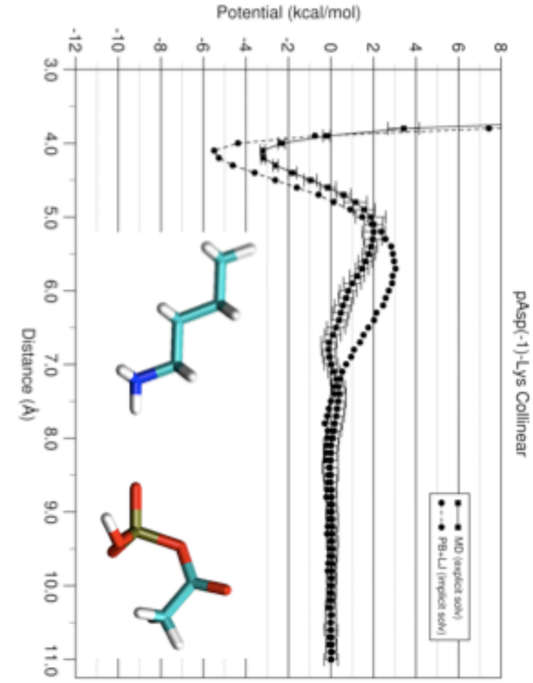

pAsp(-2)-Lys Collinear
Potential (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)

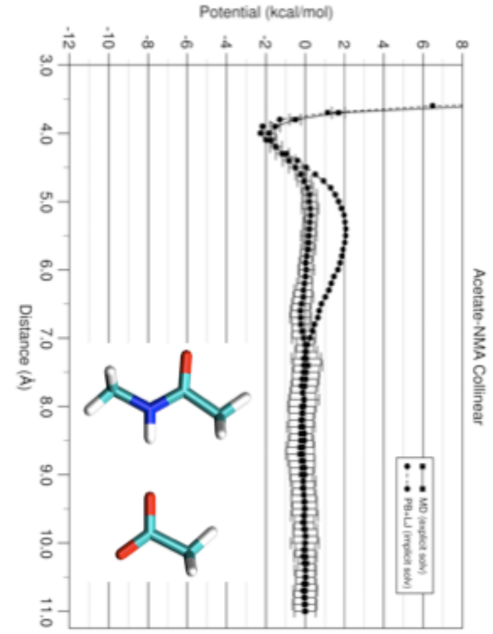

pAsp(-1)-Arg Coplanar
Potential (kcal/mol)
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
Distance (Å)
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)


pAsp(-1)-Arg Collinear
Potential (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)


pAsp(-1)-NMA Collinear
Potential (kcal/mol)
-12
-10
-8
-6
-4
-2
0
2
4
6
8
Distance (Å)
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
MD (explicit solv)
PB+LJ (implicit solv)


pAsp(-1)-Lys Coplanar
Potential (kcal/mol)
8
6
4
2
0
-2
-4
-6
-8
-10
-12
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
Distance (Å)
MD (explicit solv)
PB+LJ (implicit solv)


pAsp(-1)-Lys Collinear
Potential (kcal/mol)
8
6
4
2
0
-2
-4
-6
-8
-10
-12
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
Distance (Å)
MD (explicit solv)
PB+LJ (implicit solv)


Acetate-NMA Collinear
Potential (kcal/mol)
8
6
4
2
0
-2
-4
-6
-8
-10
-12
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
Distance (Å)
MD (explicit solv)
PB+LJ (implicit solv)

Chapter 3: Towards Deciphering the Code to Aminergic G-Protein Coupled Receptor Drug Design.

Edwin S. Tan[1]*, Eli S. Groban*[2]*, Matthew P. Jacobson*[3]*, Thomas S. Scanlan*[4]

[1]Chemistry and Chemical Biology Graduate Program, University of California, San Francisco, San Francisco, CA 94158-2280.

[2]Graduate Group in Biophysics, University of California, San Francisco, San Francisco CA 94158-2240.

[3]Department of Pharmaceutical Chemistry, University of California at San Francisco, 600 16th Street, San Francisco, CA 94158-2240

[4]Department of Physiology and Pharmacology, Oregon Health and Science University, 3181 Southwest Sam Jackson Park Road, Mail Code, L334, Portland, Oregon 97239

Summary

The trace amine-associated receptor 1 (TAAR$_1$) is a biogenic amine G-protein coupled receptor (GPCR) that is potently activated by 3-iodothyronamine (1, T$_1$AM) *in vitro*. Compound 1 is an endogenous derivative of the thyroid hormone thyroxine that rapidly induces hypothermia, anergia, and bradycardia when administered to mice. To explore the role of TAAR$_1$ in mediating the effects of 1, we rationally designed and synthesized rat TAAR$_1$ superagonists and lead antagonists using the rotamer toggle switch model of aminergic GPCR activation. The functional activity of a ligand is proposed to be correlated to its probable interactions with the rotamer switch residues; agonists allow the rotamer switch residues to toggle to their active conformation while antagonists interfere with this conformational transition. These agonist and antagonist design principles provide a conceptual model for understanding the relationship between the molecular structure of a drug and its pharmacological properties.

Introduction

3-Iodothyronamine (1, T_1AM; Fig. 1a) is an endogenous, decarboxylated, and deiodinated metabolite of the thyroid hormone thyroxine (T_4; Fig. 1a) that is found in the brain, heart, liver and blood (Scanlan et al., 2004). When administered to mice intraperitoneally, 1 rapidly induces hypothermia, anergia, and bradycardia; effects of which are opposite those observed with hyperthyroidism. *In vitro*, 1 induces the production of cAMP (adenosine 3′,5′-cyclic monophosphate) in HEK293 (human embryonic kidney 293) cells stably transfected with the GPCR known as $TAAR_1$ (Hart et al., 2006; Scanlan et al., 2004; Wainscott et al., 2007; Zucchi et al., 2006). Additionally, 1 has been found to inhibit neurotransmitter reuptake by the dopamine (DAT) and norepinephrine transporter (NET), and inhibits vesicular packaging by the vesicular monoamine transporter 2 (VMAT2) (Snead et al., 2007). To understand the role of $TAAR_1$ in mediating the effects of 1, we sought to develop small molecules that regulate the activity of $TAAR_1$.

Rat $TAAR_1$ ($rTAAR_1$) is homologous to the β_2 adrenergic (β_2AR), dopamine, and serotonin receptors and belongs to the biogenic amine subfamily of class A rhodopsin like GPCRs (Borowsky et al., 2001; Bunzow et al., 2001; Lindemann et al., 2005). GPCRs are seven transmembrane (TM) proteins with an extracellular amino terminus and an intracellular carboxy terminus (Fig. 1b-c) (Gether, 2000; Wess, 1998). The binding site of aminergic GPCRs is located within the TM region and is primarily composed of the extracellular half of transmembranes 3, 5, 6, and 7 (Cherezov et al., 2007; Rasmussen et al., 2007; Rosenbaum et al., 2007; Tota et al., 1991). Elegant pharmacological and mutagenesis studies on β$_2$AR suggest that epinephrine binds to β$_2$AR with aspartic acid 3.32 (D3.32) acting as the counterion for the charged amine, serine residues 5.42, 5.43, and 5.46 (S5.42, S5.43, and S5.46, respectively) interacting with the catechol hydroxyls, phenylalanines 6.51 and 6.52 (F6.51 and F6.52) interacting with the catechol ring, and asparagine 6.55 (N6.55) as the partner for the β-hydroxy group (Fig. 1d) (see the Experimental Procedures section for a description of the residue indexing system) (Liapakis et al., 2000; Shi and Javitch, 2002; Strader et al., 1989; Strader et al., 1994; Strader et al., 1988; Strader et al., 1989; Wieland et al., 1996; Zuurmond et al., 1999).

Previous work with the β$_2$AR suggests that agonist binding toggles a rotamer switch to its active configuration and induces a conformational change in TM6 (Fig. 2) (Shi et al., 2002). The movement of the cytoplasmic end of TM6 away from TM3 is thought to break an ionic lock interaction that is present in the inactive state of the receptor (Fig. 2a). This exposes G-protein recognition sites in the intracellular surface of the receptor that activate G-proteins and initiate the signaling cascade (Ballesteros et al., 2001; Yao et al., 2006). The rotamer switch is partly composed of a tryptophan (W6.48) and phenylalanine (F6.52) residues in TM6 that toggle concertedly between their inactive (Fig. 2a) and active (Fig. 2b) rotamer configurations to modulate the bend angle of the kink in TM6 formed by proline 6.50 (P6.50). The ionic lock involves highly conserved aspartic acid (D3.49) and arginine (R3.50) residues in TM3 and a glutamic acid (E6.30) residue in TM6. The absolute conservation of the rotamer switch and ionic lock residues in $rTAAR_1$ suggests a mechanism of activation for $rTAAR_1$ similar to β_2AR.

Studies probing the mechanism of agonist induced conformational changes in the β_2AR have found that agonist binding occurs in a sequential process involving a series of conformational intermediates that have increasing numbers of interactions with the agonist as the receptor moves toward the fully active state (Kobilka et al., 2007). The binding site of β_2AR is not prearranged to simultaneously interact with all of the functional groups of a given agonist like epinephrine (Fig. 1d). Upon binding, only a few structural elements of epinephrine (i.e. the amine and catechol moiety) are proposed to be engaged with the β_2AR. These initial interactions induce a conformation transition to an intermediate that reveals additional contact points that interact with the β-hydroxyl and/or N-methyl groups. The functional groups of epinephrine have a synergistic effect on binding affinity and receptor activation and collectively influence the overall conformation of the active receptor (Liapakis et al., 2004). The ensemble of active receptor states induced by different agonists may have disparate functional properties and have different capacity to activate downstream effector molecules such as G_s protein, GPCR receptor kinase, and/or arrestin (Swaminath et al., 2004).

Despite being a major drug target and having insights into the molecular mechanism of GPCR activation and agonist induced conformational changes, the nature of the ligand-receptor interaction is not fully understood. Although there have been many successful campaigns into GPCR drug design, it is surprising to find that there are no general postulates that can serve as guiding principles in the process of agonist and/or antagonist development without requiring extensive structure-activity relationship (SAR) data to develop a pharmacophore for the receptor of interest. Even with pharmacophore models in hand, the code to aminergic GPCR drug design is still unknown. Presently, it is unclear what inherent structural features of a ligand are responsible for endowing agonistic or antagonistic properties or how and why those structural elements lead to receptor activation or inhibition.

Based on the rotamer toggle switch model, we hypothesized that the functional properties of a compound are determined by the nature of its interaction with the rotamer switch residues. If a compound allows the rotamer switch to toggle and/or has more favorable interactions with the active state of the receptor, it will act as an agonist (Fig. 2b). In contrast, a compound will behave as an antagonist if it can sterically occlude the rotamer switch and/or has more favorable interactions with the inactive state of the receptor (Fig. 2a). Herein we describe the rational design and synthesis of $rTAAR_1$ superagonists (agonists that are more potent and/or more efficacious than 1) and lead antagonists guided by the rotamer toggle switch model of aminergic GPCR activation.

Results

Development of $rTAAR_1$ superagonists

The ligand binding site of $rTAAR_1$ differs from that of the β_2AR in that two hydrophobic residues, alanine (A5.42) and phenylalanine (F5.43) (Fig. 3b), replace the serine residues S5.42 and S5.43 in TM5 (Fig. 1d). By analogy to the catecholamines (epinephrine, norepinephrine, and dopamine), we speculate that 2 (Fig. 3a) (Tan et al., 2007), a potent $rTAAR_1$ agonist, is anchored into the binding site by the salt bridge interaction between the charged amine and D3.32, and the hydrogen bond interaction between the biaryl ether oxygen and S5.46 (Fig. 3b). To experimentally test this hypothesis, a series of derivatives of 2 containing functional groups at the β-phenyl ring (ring C in Fig. 3a) were synthesized. We specifically incorporated polar functional groups (3-7) (Table 1) capable of forming hydrogen bond interactions because our homology model of $rTAAR_1$ (see Experimental Procedures section for a description of how the model was generated), which was based on the crystal structure of bovine rhodopsin, showed that the surrounding residues around the β-phenyl ring would be asparagines (N7.35 & N7.39), a methionine (M6.55) and a cysteine (C6.54) (Fig. 3b). Therefore, if 2 binds in this orientation, having functional groups that can interact with these residues should theoretically enhance binding affinity and thus increase potency. Additionally, fluorine substituted analogs of 2 (8, 9) (Table 1) were also synthesized to determine the effects of decreasing the electron density of the β-phenyl ring on $rTAAR_1$ activation. Compounds 3-9 were synthesized from 4-bromodiphenylether and a mono-substituted benzaldehyde in 4 to 7 steps (Supplementary Schemes 1 and 2). Detailed synthetic procedures for compounds 2-56 are described in the Supplementary Data.

In HEK293 stable cell lines, 2 (EC_{50} = 28 ± 2 nM, E_{max} = 103 ± 4 %) activates the stimulatory G protein coupled $rTAAR_1$ at the same level as 1 (EC_{50} = 33 ± 3 nM, E_{max} = 100 ± 0 %) (Table 1) (Tan et al., 2007). Representative dose-response curves of agonists for $rTAAR_1$ are shown in Supplementary Figure 1a. Appending a methoxy group at the *para* (3) or *ortho* (4) position of the β-phenyl ring in 2 was detrimental, decreasing the potency ~5-6-fold and the efficacy 19-35 % (3, EC_{50} = 142 ± 40 nM, E_{max} = 68 ± 8 %, and 4, EC_{50} = 163 ± 18 nM, E_{max} = 84 ± 2 %) (Table 1). A hydroxyl group at the β-phenyl ring was well tolerated by $rTAAR_1$ but only at the *para* position. The potency of the *para* hydroxy derivative 5 increased ~4.5-fold (EC_{50} = 6 ± 1 nM) and its efficacy was slightly enhanced (E_{max} = 114 ± 9 %). When the hydroxyl substituent was located at the *ortho* (6) or *meta* (7) position, the potency decreased ~7.5-16.5-fold (EC_{50} = 467 ± 107 nM and 212 ± 39 nM, respectively) while the efficacy either decreased or was unaffected (E_{max} = 70 ± 6 % and 106 ± 7 %, respectively). Similarly, $rTAAR_1$ somewhat prefers a fluorine group at the *para* over the *meta* position as the potency was the same for 8 (EC_{50} = 28 ± 6 nM) but decreased 2-fold for 9 (EC_{50} = 57 ± 6 nM). The efficacy of 8 and 9 (E_{max} = 99 ± 9 % and 110 ± 2 %, respectively) were unaffected by fluorination and were similar to that of 2. All compounds with stereogenic centers (2-50 and 52-56) were evaluated as racemic mixtures. The observed activities of all compounds tested (1-56) were found to be rTAAR$_1$-dependent, as all compounds showed no cAMP accumulation when exposed to an empty vector control cell line (data not shown).

In an effort to improve the potency of 5, we explored its tolerance for methylation at the amine, iodination of the inner ring, and hydroxylation of the outer ring. These modifications, individually or in combination, have previously been found to be beneficial for rTAAR$_1$ activation (Hart et al., 2006). Mono-methylation of the amine in 5 provided 10 while mono-iodination of the inner ring yielded 11 (Supplementary Schemes 3 and 4). Adding a hydroxyl group to the *para* or *meta* position of the outer ring in 11 gave 12 and 13, respectively (Supplementary Scheme 4).

When screened for agonist activity, some of the 5 derivatives were more efficacious but none were more potent. N-Methylation of 5 (10) was beneficial increasing the efficacy 13 % (E_{max} = 127 ± 2 %) but it did not improve potency (EC_{50} = 5 ± 1 nM) (Table 1). Mono-iodination of the inner ring (11) was unfavorable decreasing potency ~3-fold (EC_{50} = 17 ± 2 nM) without significantly affecting efficacy (E_{max} = 107 ± 8 %). In the presence of an outer ring *para* hydroxyl group (12), the rTAAR$_1$ activity improved back to the level of 5 (EC_{50} = 4 ± 1 nM, E_{max} = 115 ± 2 %). In contrast, a *meta* hydroxyl group on the outer ring of 11 (13) had no effect on potency and efficacy (EC_{50} = 22 ± 2 nM and E_{max} = 111 ± 9 %).

Development of $rTAAR_1$ lead antagonist

According to our proposed binding orientation of 2 in $rTAAR_1$ (Fig. 3b), the rotamer switch residues are located in the vicinity of position 2 of the inner ring (ring B in Fig. 3a). Using the toggle switch model of aminergic GPCR activation as a guideline (Fig. 2), we attempted to convert 2 into an antagonist by appending functional groups at the 2 position to theoretically interfere with the rotamer switch residues. An alcohol group was installed into the 2 position (R_5, Table 2) of 2 (14) to serve as a handle for synthesizing a panel of ethers (15-24) and esters (25 and 26) varying in steric bulk, rigidity, topology and polarity (Table 2, Supplementary Schemes 5 and 6).

The effects of the ether and ester substituents on receptor agonist activity were variable. The core scaffold 14 and ethyl ether 16 were decent agonists activating to the same efficacy level as 2 (E_{max} = 108 ± 1 % and 95 ± 5 %, respectively) but at ~3-5-fold lower potency (EC_{50} = 96 ± 10 nM and 144 ± 31 nM, respectively) (Table 2). On the other hand, the methyl ether 15 showed the opposite trend being equipotent to 2 (EC_{50} = 35 ± 4 nM) but less efficacious (E_{max} = 82 ± 8 %). The unsaturated alkene and alkyne counterparts of the propyl ether 17 appear to be well tolerated by $rTAAR_1$ as 22 (EC_{50} 169 ± 6 nM) and 23 (EC_{50} = 138 ± 37 nM) were at least 6-fold more potent than 17 (EC_{50} >1 μM). The efficacies of 17, 22, and 23 were comparable to each other (E_{max} = 69 ± 5 %, 71 ± 4 %, and 78 ± 1 %, respectively). Further increasing the size of the ether substituents (18-21 and 24) desirably decreased potency (EC_{50} > 1 μM) but it did not completely abolish the agonist activity (E_{max} ≤ 10 %) of the compounds. These compounds activated $rTAAR_1$ between 15 % and 62 % efficacy. Similarly the ester substituents (25 and 26) decreased the potency of 2 (EC_{50} = 143 ± 4 nM and 234 ± 43 nM, respectively) but did not reduce its efficacy below 10 % (E_{max} = 57 ± 5 % and 74 ± 3 %, respectively) (Table 2).

The observed agonist activities of 14-26 were consistent with the idea that the inner ring functional groups of these compounds were not properly interfering with the rotamer switch residues. In compound 14, rotation of the inner ring about the β carbon and the biaryl ether oxygen axis renders position 2 and 6 indistinguishable (Table 2). Within the binding site, it's possible that the inner rings of 15-26 have rotated 180^{o} and are actually orienting the position 2 functional group towards the extracellular surface of $rTAAR_1$ around methionine 6.55 (M6.55) instead of the intracellular region near the rotamer switch residues. In this alternate binding orientation, these compounds would be predicted to have some agonist activity as the ether or ester appendage would not be able to interfere with the rotamer switch residues.

To test this hypothesis, the core scaffold of 14 was modified to have the phenoxy group moved one carbon over to the *meta* position with respect to the ethylamine chain (28) (Table 3, Supplementary Scheme 7). In this orientation, the 2 and 6 positions of the inner ring are now structurally distinct. Having a *meta* phenoxy group should not be detrimental to binding affinity because the isomer of 2 with the phenoxy group at the *meta* position (27) was found to be a slightly better agonist than

2 for $rTAAR_1$ (EC_{50} = 19 ± 2 nM, E_{max} = 131 ± 7 %) (Fig. 3c) (Tan et al., 2007). With this modification we synthesized 21 compounds (29-49) with an ether or ester appendage at the 2 position that again varied in steric bulk, rigidity, topology, and polarity (Table 3, Supplementary Schemes 7-9).

For the ether series (29-46), an interesting correlation was observed between the size of the position 2 substituent (R_6, Table 3) and the agonist activity of the compound. The core scaffold 28 was ~12-fold less potent (EC_{50} = 232 ± 8 nM) and 43 % less efficacious (E_{max} 88 ± 9 %) compared to 27 (Table 3). Methylating the phenol of 28 (29) increased the potency ~2-fold (EC_{50} = 102 ± 26 nM) but had no effect on efficacy (E_{max} = 88 ± 1 %). When the ether group was an ethyl ether or larger (30-37), the potency of the compound was poor (>1 μM). The efficacy showed a different profile. When the ether group was less than 5 atom units long (30, 31, 36, and 40), the compound still had some degree of agonist activity (E_{max} = 26 % to 66 %). As the ether group increased in size equal to or greater than 5 atom units long (32-35 and 41-46), the compounds became non-agonists activating $rTAAR_1$ at less than 10 % efficacy. An exception to this trend was 37. Although its isobutoxy group is only four atom units long, 37 activated below 10 % efficacy (E_{max} = 6 ± 1 %). Compared to 31 (EC_{50} = >1 μM, E_{max} = 41 ± 0 %), introducing an unsaturated alkene (38) or alkyne (39) into the position 2 group increased both potency (EC_{50} = 602 ± 10 nM and 182 ± 46 nM, respectively) and efficacy (E_{max} = 79 ± 5 % and 103 ± 0 %, respectively).

In the ester series (47-49), the potency of the compounds was greater than 1 μM when the position 2 functional group was 5 atom units long (47 and 49) but less than 1 μM when 4 atom units long (48, EC_{50} = 599 ± 165 nM) (Table 3). The efficacy of the 47-49 were between 33-53 %.

Since there are currently no binding assays available for $rTAAR_1$, the antagonist activity of the 11 non-agonists (32-35 and 41-46) was determined by testing for the inhibition of cAMP production of $rTAAR_1$ in stably transfected HEK293 cells treated with EC_{50} concentration (33 nM) of 1. Representative dose response curves of antagonists against $rTAAR_1$ are shown in Supplementary Figure 1b. This competition assay was validated in the β_2AR where the antagonist propranolol was able to inhibit the cAMP production induced by the agonist isoproterenol (data not shown).

The 11 non-agonists antagonized 1 induced $rTAAR_1$ activation to varying degrees. The butyl ether 32 showed ca. 75 % antagonism with a half maximal inhibitory concentration (IC_{50}) of 8 ± 2 μM (Table 3). Isobutyl ether 37 was also a weak antagonist showing 50 % inhibition and a potency of >10 μM. The longer pentyl and hexyl ethers (33 and 34, respectively) were better antagonists reducing the 1 signal to 3-6 % at a potency of ~4-5 μM. The cyclohexylmethyl ether 41 was equally potent (IC_{50} = 5 ± 1 μM) but somewhat less inhibitory (I_{max} = 9 ± 1 %). Compared to the benzyl ether 35 (IC_{50} = 5 ± 1 μM, I_{max} = 6 ± 3 %), the heterocyclic pyridine methyl ethers (44-46) were less potent (IC_{50} ≥ 7 μM) and inhibitory (I_{max} ≥ 15 %). The cyanoalkyl ethers 42 and 43 were poor antagonists inhibiting the 1 signal no lower than 23 % with an IC_{50} value >10 μM. The inhibitory effects of these compounds were neither due to inhibition of adenyl cylcase nor cytotoxicty (data not shown); suggesting that these compounds are *bona fide* $rTAAR_1$ antagonists.

Structure-activity relationship of $rTAAR_1$ lead antagonist

The agonist and antagonist properties of 27 and 34, respectively, suggested that the hexyloxy group is essential for antagonism. To determine if the outer ring (ring A in Fig. 3a) and β-phenyl ring are also necessary for antagonism, we synthesized analogs of 34 lacking the outer ring (50) or the β-phenyl ring (51) (Table 4, Supplementary Schemes 10 and 11). In an attempt to improve the potency of 34, we also explored the effects of N-methylation (52) and functionalization of the outer ring (53-56) (Table 4, Supplementary Schemes 8 and 10).

Removing the outer ring or β-phenyl ring of 34 was detrimental to $rTAAR_1$ antagonism. In the absence of the outer ring (50), 34 was converted into a weak agonist (Table 4, Fig 3c). Similarly, 34 became an agonist without the β-phenyl ring (51, EC_{50} = 201 ± 23 nM, EC_{50} = 59 ± 6 %) (Fig. 3c).

Mono-methylating the amine (52) or inserting electron withdrawing groups on the outer ring (54-56) preserved the antagonist activity of 34. When screened for agonist activity, these compounds did not activate $rTAAR_1$ (Table 4). In the antagonist assay, the potency of 52 was unaffected (IC_{50} = 5 ± 1 μM) but the antagonist activity slightly decreased (IC_{50} = 10 ± 4 %). The potency and inhibitory capacity of 34 was also not significantly affected by introducing a *para*-fluoro, *meta*-

fluoro, or *meta*-cyano group into the outer ring (54, 55, and 56 respectively). The IC_{50} and I_{max} values of these compounds were ~3 μM and ≤ 2 %, respectively. Interestingly, inserting a *para*-hydroxy group into the outer ring (53) endowed some agonist activity to 34 activating $rTAAR_1$ at >1 μM potency and 16 ± 3 % efficacy.

Discussion

The rotamer toggle switch model of aminergic GPCR activation (Fig. 2) has proven to be a useful guideline in the design and synthesis of $rTAAR_1$ agonists and antagonists. Previous SAR studies on the ethylamine portion of 1 for $rTAAR_1$ provided 2 as a promising scaffold for developing $rTAAR1_1$ superagonists, which we define as compounds that are more potent and/or efficacious than 1 (Tan et al., 2007). In addition to being as potent and efficacious as 1, 2 provides the added benefit of having many potential sites for derivitization. By analogy to the assumed binding mode of epinephrine to β_2AR (Fig. 1d), we deduced 2 to bind to $rTAAR_1$ in a similar fashion with the charged amine forming a salt bridge interaction with D3.32 and the biaryl ether oxygen hydrogen bonding to S5.46 (Fig. 3b). The β-phenyl ring is proposed to occupy a pocket near the interface of TM6 and TM7.

In the context of the rotamer toggle switch model, our analysis of the ligand-receptor interaction of β_2AR agonists showed that agonists generally lack functional groups in the region of the molecule that is predicted to be located in the vicinity of the rotamer switch residues. Structurally, most of these agonists appear to have functional groups that complement the physicochemical properties of the residues within the binding site. Following this lead, we attempted to improve the agonist properties of 2 by incorporating functional groups in the regions of the molecule (e.g. β-phenyl ring, charged amine, outer ring, and position 5 of the inner ring (Fig. 3a,b)) away from rotamer switch residues. In the β-phenyl ring, SAR studies presented here showed a clear preference for a hydroxyl group at the *para* position. The *para* hydroxyl analog (5) was 24-78-fold and 8-46 % more potent and efficacious, respectively, compared to the *ortho* or *meta* hydroxyl analogs (6 and 7) and *ortho* or *para* methoxy (3 and 4) analogs. Additionally, the *para* hydroxyl improved the potency and efficacy of 2 ~4.5-fold and 11 %, respectively. We believe that this enhancement in agonist activity is a reflection of an increase in the binding affinity of 2 for $rTAAR_1$ due to hydrogen bond interactions of the *para* hydroxyl with N7.39 and/or N7.35 (Fig. 3b). Mutating residue 7.39 in the β_2AR has previously been found to perturb the binding affinity of agonists and antagonists (Suryanarayana et al., 1991). In the recently determined crystal structure of the β_2AR, N7.39 of β_2AR was involved in hydrogen bond interactions with the β-carbon hydroxyl group of the partial inverse agonist carazolol (Cherezov et al., 2007; Rasmussen et al., 2007; Rosenbaum et al., 2007).

In the presence of the *para* hydroxyl, mono-methylating the charged amine (10) or incorporating a 1 moiety into the molecule (12) was tolerated but it had modest effects on agonist activity, if any at all. N-methyl 10 was equipotent to 5 but 13 % more efficacious. On the other hand, 12 essentially has the same potency and efficacy as 5. The comparable levels of agonist activity of 5, 10, and 12, suggest that these compounds have similar interactions with $rTAAR_1$ and possibly elicit the same final active conformation of the receptor.

In contrast to the β_2AR agonists, our analysis of the SAR and potential binding modes of antagonists for the dopamine 1-like and 2-like receptors revealed the presence of structural moieties within these compounds that could conceivably sterically occlude the rotamer toggle switch residues from assuming their active conformation. Applying this hypothesis to $rTAAR_1$ we attempted to convert 2 into an antagonist by installing ethers and esters at the 2 position of the inner ring that varied in steric bulk, rigidity, topology, and polarity (15-26). Based on our proposed binding orientation of 2 (Fig. 3b), this position was identified to be the prime location for presenting groups that could interfere with the rotamer switch residues in $rTAAR_1$. Unfortunately, none of these compounds turned out to be antagonists. Presumably 15-26 were still able to activate $rTAAR_1$ between 15 % and 95 % efficacy because the variable position 2 groups (R_5, Table 2) are positioned away from the rotamer switch residues within the binding site due to rotation of the inner ring about the β-carbon and biaryl ether oxygen axis.

To circumvent this problem, we modified the core scaffold by moving the phenoxy group from the *para* (14) to the *meta* (28) position (Tables 2 and 3). With this modification, the agonist activity of the compound decreased as the size of the ether substituent increased (Fig. 3c). When the ether group was ≥ 5 atom units long (32-35, and 41-46), the agonist activity of the compound was completely abolished (≤ 10 % efficacy). Compounds with substituents less than 5 atom units long (29-31) were weak agonists activating $rTAAR_1$ between 41 to 88 % efficacy. The composition of the substituent appears to be important as an ester group that is 5 atom units long (47 and 49) was still an agonist (EC_{50} = 33 to 53 %). When the non-agonists (32-35 and 41-46) were screened for antagonist activity in a competition assay with 1 at its EC_{50} concentration (33 nM), all compounds were found to inhibit 1 induced cAMP production to varying degrees at 10 μM. Compound 34 was the best antagonist showing >90 % inhibition of $rTAAR_1$ activation with an IC_{50} value of 4 μM. The antagonist activity of 32-35 and 41-46 are thought to arise from the ether substituents sterically occluding F6.52 and/or W6.48 of the rotamer switch residues from assuming their active conformation.

The hexyloxy group, outer ring, and β-phenyl ring of 34 are all necessary for antagonism. In the absence of any one of these groups, the resulting compounds lose their antagonist activity and become agonists. Since the transformation of 34 to 27 yielded the greatest increase in agonist potency and efficacy, the hexyloxy group is the most important of the three structural elements in terms of decreasing agonist activity and conferring antagonist properties to 34 (Fig. 3c). This is consistent with the notion that the outer ring and β-phenyl ring are essential scaffolding elements that assures 34 docks into the $rTAAR_1$ binding site in the proper orientation to position the hexyloxy group, the molecular basis of antagonism, to interfere with the rotamer switch residues (Fig. 3d).

Significance

The rotamer toggle switch model of aminergic GPCR activation is a useful model for understanding the molecular basis of $rTAAR_1$ activation by 1 and related analogs. It has proven helpful in the development of $rTAAR_1$ agonists and antagonists providing superagonists 5, 10, and 12 and lead antagonists 34, 54 and 55. This structure-activity relationship study suggests that agonist or antagonist properties of aminergic GPCR drugs could arise from probable drug interactions with the rotamer switch residues. Agonists allow the rotamer switch to toggle and/or have more favorable interactions with the active state of the receptor while antagonists sterically occlude the rotamer switch and/or have more favorable interactions with the inactive state of the receptor.

These agonist and antagonist design principles have the potential to accelerate and increase the efficiency of the drug discovery and development process for GPCRs. Having insights into the critical ligand-receptor interactions important for receptor activation or inhibition facilitates the interpretation of SAR data and correlation of pharmacophore models with the molecular properties of the receptor binding site. This information then provides a map of the binding site landscape and presents a drug design blueprint for identifying promising scaffolds, recognizing compatible functional groups to incorporate, and evaluating the contribution of individual structural elements of a given compound towards its binding affinity, selectivity, and functional properties. We envision these principles to supplement all current GPCR drug design strategies (e.g. ligand-based drug design, focused library screening, virtual screening, structure based drug design, etc.) (Evers et al., 2005; Evers et al., 2005; Klabunde et al., 2005; Klabunde et al., 2002) and help generate predictive rules and guidelines that would prove to be a useful and general method for designing activators or inhibitors for biogenic amine GPCRs and possibly other rhodopsin like GPCRs.

Experimental Procedures

Residue indexing system. Residues are labeled relative to the most conserved amino acid in the transmembrane segment in which it is located (Ballesteros et al., 1995). Tryptophan 6.48, for example, is located in transmembrane 6 and precedes the most conserved residue by 2 positions. Arginine 3.50 is the most conserve residue in TM3. This system simplifies the identification of corresponding residues in different GPCRs.

Homology model of $rTAAR_1$. The sequence of $rTAAR_1$ was aligned to 26 human biogenic amine GPCRs (i.e. dopamine, α-adrenergic, β-adrenergic, and serotonin receptors) and the sequence for bovine rhodopsin (Protein Data Bank accession code 1F88) (Palczewski et al., 2000) using the program MUSCLE (Edgar, 2004). We constructed our homology model of $rTAAR_1$ based on the crystal structure of the inactive state bovine rhodopsin as a template and used our in house software PLOP (commercially available as Prime from Schrödinger Inc). The modeling program did not modify conserved residues, leaving each atom in these residues at their original PDB coordinates. Non-conserved side chains were built onto the structure using the backbone coordinates for bovine rhodopsin as a reference point. All chain breaks or gaps were closed using a previously published loop building and optimization algorithm (Jacobson et al., 2004). After building the complete model, side chain optimization, followed by backbone and side chain energy minimization, was performed on all non-conserved residues. The homology modeling program relies on the OPLS all atom force field (Jacobson et al., 2002; Jorgensen et al., 1996;

Kaminski et al., 2001) and a Generalized Born solvent model (Gallicchio et al., 2002; Ghosh et al., 1998) to evaluate the energy of different conformations and select the lowest energy structure as the final model.

Synthesis. Detailed synthetic procedures and chemical compound information are described in the supplementary information.

In Vitro cAMP Assays-Agonist activity assay. After incubating in fresh medium for at least 2 h, HEK293 cells stably transfected with $rTAAR_1$ were harvested in Krebs-Ringer-HEPES buffer (KRH) and preincubated with 200 μM 3-isobutyl-1-methylxanthine (IBMX) for 20-30 minutes. Cells were incubated in KRH with 133 μM IBMX and 3μL of the test compound, forskolin (10 μM), or vehicle (dimethyl sulfoxide, DMSO) for 1 h at 37 °C (300 μL total volume). The cells were boiled for 20 min after addition of 100 μL 0.5 mM sodium acetate buffer. The cell lysate was centrifuged to remove cellular debris, and an aliquot (30 μL) was transferred to an opaque, flat bottom 96-well plate (Corning #3917). The cAMP content of the aliquot was measured by use of the Hithunter^{TM} cAMP XS kit (DiscoveRX, Fremont, CA). The plate was shaken on a titer plate shaker for 2 min after addition of 20 μL of cAMP XS antibody/lysis mix. After incubation in the dark for 1h, 20 μL of cAMP XS ED reagent was added and the plate was shaken for 2 min. After another hour of incubation in the dark, 40 μL of cAMP XS EA/CL substrate mix was added and the plate was shaken for 2 min. The plate was sealed with an acetate plate sealer (Thermo Scientific #3501) and allowed to incubate in the dark for 15-18 h before luminescence was measured (3 readings/well at 0.33

s/reading) on an Analyst™ AD Assay Detection System (LJL Biosystems) or a Packard Fusion Microplate Reader. Data were reported relative to **1** and expressed as %T_1AM. The activity of **1** at 10μM was set as 100 %T_1AM. Concentration-response curves were plotted and EC_{50} values were calculated with Prism software (GraphPad, San Diego, CA). Standard error of the mean was calculated from the EC_{50} and E_{Max} values of each independent triplicate experiment by use of Prism Software.

In Vitro cAMP Assays-Antagonist activity assay. Same as the agonist activity assay procedure described above with the following changes: cells that were harvested in KRH buffer and preincubated with IBMX for 20-30 minutes were incubated in KRH with 133 μM IBMX and 3μL of the putative antagonist, or vehicle (DMSO) for 30 min at 37 °C (300 μL total volume). 3μL of the competing agonist (T_1AM, EC_{50} concentration (33 nM) as the final concentration), T_1AM (10 μM), forskolin (10 μM) or vehicle (DMSO) was then added to the reactions before incubating for 1h at 37°C. The cells were then processed as described in the agonist activity assay. Concentration-response curves were plotted and IC_{50} values were calculated with Prism software (GraphPad, San Diego, CA).

References

Ballesteros, J., Jensen, A., Liapakis, G., Rasmussen, S., Shi, L., Gether, U., and Javitch, J. (2001). Activation of the beta(2)-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. J. Biol. Chem. 276, 29171-29177.

Ballesteros, J., and Weinstein, H. (1995). Integrated methods for the construction of three-dimensional models of structure-function relations in G protein-coupled receptors. Methods Neuroscience 25, 366-428.

Borowsky, B., Adham, N., Jones, K., Raddatz, R., Artymyshyn, R., Ogozalek, K., Durkin, M., Lakhlani, P., Bonini, J., Pathirana, S., et al. (2001). Trace amines: Identification of a family of mammalian G protein-coupled receptors. Proc. Natl. Acad. Sci. U. S. A. 98, 8966-8971.

Bunzow, J., Sonders, M., Arttamangkul, S., Harrison, L., Zhang, G., Quigley, D., Darland, T., Suchland, K., Pasumamula, S., Kennedy, J., et al. (2001). Amphetamine, 3,4-methylenedioxymethamphetamine, lysergic acid diethylamide, and metabolites of the catecholamine neurotransmitters are agonists of a rat trace amine receptor. Mol. Pharmacol. 60, 1181-1188.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., et al. (2007). High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 318, 1258-1265.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797.

Evers, A., Hessler, G., Matter, H., and Klabunde, T. (2005). Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtual screening protocols. J. Med. Chem. 48, 5448-5465.

Evers, A., and Klabunde, T. (2005). Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the Alpha1A adrenergic receptor. J. Med. Chem. 48, 1088-1097.

Gallicchio, E., Zhang, L. Y., and Levy, R. M. (2002). The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. J. Comput. Chem. 23, 517-529.

Gether, U. (2000). Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. Endocr. Rev. 21, 90-113.

Ghosh, A., Rapp, C., and Friesner, R. (1998). Generalized born model based on a surface integral formulation. J. Phys. Chem. B 112, 10983-10990.

Hart, M., Suchland, K., Miyakawa, M., Bunzow, J., Grandy, D., and Scanlan, T. (2006). Trace amine-associated receptor agonists: Synthesis and evaluation of thyronamines and related analogues. J. Med. Chem. 49, 1101-1112.

Jacobson, M., Kaminski, G., Friesner, R., and Rapp, C. (2002). Force field validation using protein side chain prediction. J. Phys. Chem. B 106, 11673-11680.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., and Friesner, R. A. (2004). A hierarchical approach to all-atom protein loop prediction. Proteins: Struct., Funct., Bioinf. 55, 351-367.

Jorgensen, W., Maxwell, D., and Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118, 11225-11236.

Kaminski, G., Friesner, R., Tirado-Rives, J., and Jorgensen, W. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J. Phys. Chem. B 105, 6474-6487.

Klabunde, T., and Evers, A. (2005). GPCR antitarget modeling: Pharmacophore models for biogenic amine binding GPCRs to avoid GPCR-mediated side effects. ChemBioChem 6, 876-889.

Klabunde, T., and Hessler, G. (2002). Drug design strategies for targeting G-protein-coupled receptors. ChemBioChem 3, 929-944.

Kobilka, B., and Deupi, X. (2007). Conformational complexity of G-protein-coupled receptors. Trends Pharmacol. Sci. 28, 397-406.

Liapakis, G., Ballesteros, J., Papachristou, S., Chan, W., Chen, X., and Javitch, J. (2000). The forgotten serine - A critical role for Ser-203(5.42) in ligand binding to and activation of the beta(2)-adrenergic receptor. J. Biol. Chem. 275, 37779-37788.

Liapakis, G., Chan, W., Papadokostaki, M., and Javitch, J. (2004). Synergistic contributions of the functional groups of epinephrine to its affinity and efficacy at the beta(2) adrenergic receptor. Mol. Pharmacol. 65, 1181-1190.

Lindemann, L., Ebeling, M., Kratochwil, N., Bunzow, J., Grandy, D., and Hoener, M. (2005). Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. Genomics 85, 372-385.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., et al. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. Science 289, 739-745.

Rasmussen, S. G., Choi, H. J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R., Sanishvili, R., Fischetti, R. F., et al. (2007). Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature 450, 383-387.

Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Yao, X. J., Weis, W. I., Stevens, R. C., et al. (2007). GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. Science 318, 1266-1273.

Scanlan, T., Suchland, K., Hart, M., Chiellini, G., Huang, Y., Kruzich, P., Frascarelli, S., Crossley, D., Bunzow, J., Ronca-Testoni, S., et al. (2004). 3-Iodothyronamine is an endogenous and rapid-acting derivative of thyroid hormone. Nat. Med. 10, 638-642.

Shi, L., and Javitch, J. (2002). The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. Annu. Rev. Pharmacol. Toxicol. 42, 437-467.

Shi, L., Liapakis, G., Xu, R., Guarnieri, F., Ballesteros, J., and Javitch, J. (2002). Beta(2) adrenergic receptor activation - Modulation of the proline kink in transmembrane 6 by a rotamer toggle switch. J. Biol. Chem. 277, 40989-40996.

Snead, A. N., Santos, M. S., Seal, R. P., Miyakawa, M., Edwards, R. H., and Scanlan, T. S. (2007). Thyronamines inhibit plasma membrane and vesicular monoamine transport. ACS Chem. Biol. 2, 390-398.

Strader, C., Candelore, M., Hill, W., Sigal, I., and Dixon, R. (1989). Identification of 2 serine residues involved in agonist activation of the beta-adrenergic-receptor. J. Biol. Chem. 264, 13572-13578.

Strader, C., Fong, T., Tota, M., Underwood, D., and Dixon, R. (1994). Structure and function of G-protein coupled receptors. Annu. Rev. Biochem. 63, 101-132.

Strader, C., Sigal, I., Candelore, M., Rands, E., Hill, W., and Dixon, R. (1988). Conserved aspartic-acid residue-79 and residue-113 of the beta-adrenergic-receptor have different roles in receptor function. J. Biol. Chem. 263, 10267-10271.

Strader, C., Sigal, I., and Dixon, R. (1989). Structural basis of beta-adrenergic-receptor function. FASEB J. 3, 1825-1832.

Suryanarayana, S., Daunt, D., von Zastrow, M., and Kobilka, B. (1991). A point mutation in the 7th hydrophobic domain of the alpha-2-adrenergic-receptor increases its affinity for a family of beta-receptor-antagonists. J. Biol. Chem. 266, 15488-15492.

Swaminath, G., Xiang, Y., Lee, T., Steenhuis, J., Parnot, C., and Kobilka, B. (2004). Sequential binding of agonists to the beta(2) adrenoceptor - Kinetic evidence for intermediate conformational states. J. Biol. Chem. 279, 686-691.

Tan, E., Miyakawa, M., Bunzow, J., Grandy, D., and Scanlan, T. (2007). Exploring the structure-activity relationship of the ethylamine portion of 3-iodothyronamine for rat and mouse trace amine-associated receptor 1. J. Med. Chem. 50, 2787-2798.

Tota, M. R., Candelore, M. R., Dixon, R. A., and Strader, C. D. (1991). Biophysical and genetic analysis of the ligand-binding site of the beta-adrenoceptor. Trends Pharmacol. Sci. 12, 4-6.

Wainscott, D., Little, S., Yin, T., Tu, Y., Rocco, V., He, J., and Nelson, D. (2007). Pharmacologic characterization of the cloned human trace amine-associated receptor1 (TAAR1) and evidence for species differences with the rat TAAR1. J. Pharmacol. Exp. Ther. 320, 475-485.

Wess, J. (1998). Molecular basis of receptor/G-protein-coupling selectivity. Pharmacol. Ther. 80, 231-264.

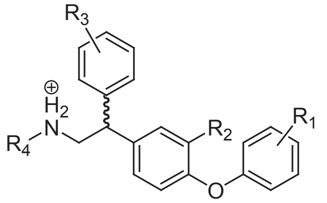
Wieland, K., Zuurmond, H., Krasel, C., Ijzerman, A., and Lohse, M. (1996). Involvement of Asn-293 in stereospecific agonist recognition and in activation of the beta(2)-adrenergic receptor. Proc. Natl. Acad. Sci. U. S. A. 93, 9276-9281.

Yao, X., Parnot, C., Deupi, X., Ratnala, V., Swaminath, G., Farrens, D., and Kobilka, B. (2006). Coupling ligand structure to specific conformational switches in the beta(2)-adrenoceptor. Nat. Chem. Bio. 2, 417-422.

Zucchi, R., Chiellini, G., Scanlan, T., and Grandy, D. (2006). Trace amine-associated receptors and their ligands. Br. J. Pharmacol. 149, 967-978.

Zuurmond, H., Hessling, J., Bluml, K., Lohse, M., and Ijzerman, A. (1999). Study of interaction between agonists and Asn293 in helix VI of human beta(2)-adrenergic receptor. Mol. Pharmacol. 56, 909-916.

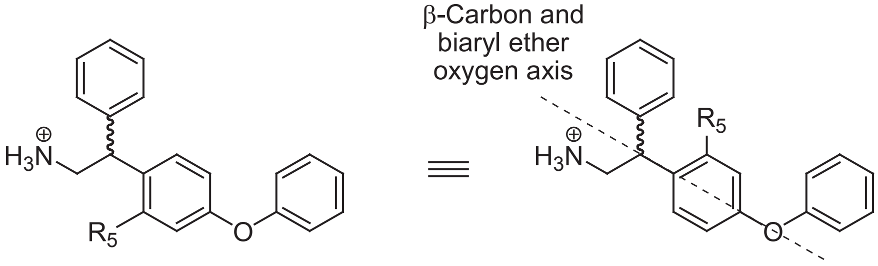
Table 1. Agonist activity of compounds **1-13** on rTAAR$_1$.



Compd	R$_1$	R$_2$	R$_3$	R$_4$	EC$_{50}$^{a} ± SEM (nM)	E_{max}^{b} ± SEM (%)	N^{c}
1	See **Fig. 1**				33 ± 3	100 ± 0	5
2	H	H	H	H	28 ± 2	103 ± 4	3
3	H	H	*p*-OMe	H	142 ± 40	68 ± 8	3
4	H	H	*o*-OMe	H	163 ± 18	84 ± 2	3
5	H	H	*p*-OH	H	6 ± 1	114 ± 9	4
6	H	H	*o*-OH	H	467 ± 107	70 ± 6	3
7	H	H	*m*-OH	H	212 ± 39	106 ± 7	3
8	H	H	*p*-F	H	28 ± 6	99 ± 9	3
9	H	H	*m*-F	H	57 ± 6	110 ± 7	3
10	H	H	*p*-OH	Me	5 ± 1	127 ± 2	4
11	H	I	*p*-OH	H	17 ± 2	107 ± 8	4
12	*p*-OH	I	*p*-OH	H	4 ± 1	115 ± 2	6
13	*m*-OH	I	*p*-OH	H	22 ± 2	111 ± 9	4

[a] EC$_{50}$ is the half-maximal effective concentration of a compound. [b] E$_{\text{max}}$ is the maximum stimulation achieved at a concentration of 10 μM and was calculated by use of Prism software. EC$_{50}$ and E$_{\text{max}}$ values represent the average of N independent experiments in triplicate and were calculated by use of Prism software as described in the Experimental Procedures section. The standard error of the mean (SEM) were calculated from the EC$_{50}$ and E$_{\text{Max}}$ values of each independent triplicate experiment by use of Prism software. E$_{\text{max}}$ = 100 % is defined as the activity of **1** at 10 μM. [c] N is the number of independent experiments in triplicate that were performed and used to calculate the EC$_{50}$ and E$_{\text{max}}$ values.

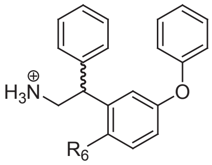
Table 2. Agonist activity of compounds **14-26** on $rTAAR_1$.



Compd	R_5	EC_{50}[a] ± SEM (nM)	E_{max}[b] ± SEM (%)	*N*[c]
14	OH	96 ± 10	108 ± 1	3
15	OMe	35 ± 4	82 ± 8	3
16	OEt	144 ± 31	95 ± 5	3
17	OPr	>1000	69 ± 5	2
18	OBu	>1000	31 ± 1	2
19	OBn	>1000	58 ± 2	2
20	O-*i*Pr	>1000	62 ± 2	2
21	O-*i*Bu	>1000	15 ± 4	2
22	OCH_2CHCH_2	169 ± 6	71 ± 4	2
23	OCH_2CCH	138 ± 37	78 ± 1	2
24	$OCH_2CO_2CH_3$	>1000	56 ± 0	2
25	$O_2CCH_2CH_2Cl$	143 ± 4	57 ± 5	2
26	O_2CCH_3	234 ± 43	74 ± 3	2

[a-c] See footnotes for **Table 1**.

Table 3. Agonist and antagonist activity of compounds **27**-**49** on rTAAR$_1$.



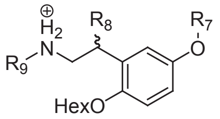
		Agonist Activity			Antagonist Activity		
Compd	R_6	$EC_{50}{}^{a}$ ± SEM (nM)	$E_{max}{}^{b}$ ± SEM (%)	N^{c}	$IC_{50}{}^{d}$ ± SEM (μM)	$I_{max}{}^{e}$ ± SEM (%)	N^{c}
27	H	19 ± 2	131 ± 7	3	-	-	-
28	OH	232 ± 8	88 ± 9	2	-	-	-
29	OMe	102 ± 26	88 ± 1	3	-	-	-
30	OEt	>1000	66 ± 3	2	-	-	-
31	OPr	>1000	41 ± 0	2	-	-	-
32	OBu	>1000	3 ± 0	2	8 ± 2	12 ± 3	2
33	OPent	>1000	0 ± 3	2	5 ± 0	6 ± 1	2
34	OHex	>1000	0 ± 1	2	4 ± 0	3 ± 1	4
35	OBn	>1000	9 ± 1	2	5 ± 1	6 ± 3	2
36	O-*i*Pr	>1000	40 ± 1	2	-	-	-
37	O-*i*Bu	>1000	6 ± 1	2	>10	23 ± 6	2
38	OCH_2CHCH_2	602 ± 10	79 ± 5	2	-	-	-
39	OCH_2CCH	182 ± 46	103 ± 0	2	-	-	-
40	OCH_2-Cyclopropyl	>1000	26 ± 7	3	-	-	-
41	OCH_2-Cyclohexyl	>1000	0 ± 3	2	5 ± 1	9 ± 1	-
42	$OCH_2CH_2CH_2CN$	>1000	3 ± 4	2	>10	30 ± 2	3
43	$OCH_2CH_2CH_2CH_2CN$	>1000	2 ± 2	2	>10	23 ± 4	3
44	OCH_2-(4-Pyridinyl)	>1000	4 ± 2	2	7 ± 1	15 ± 2	3
45	OCH_2-(3-Pyridinyl)	>1000	9 ± 0	2	>10	33 ± 1	3
46	OCH_2-(2-Pyridinyl)	>1000	6 ± 1	2	7 ± 0	16 ± 2	3
47	$O_2CCH_2CH_2Cl$	>1000	50 ± 8	2	-	-	-
48	$O_2CCH(CH_3)_2$	599 ± 165	53 ± 6	2	-	-	-
49	$O_2CCH_2CH(CH_3)_2$	>1000	33 ± 7	2	-	-	-

[a-c] See footnotes for Table 1.

[d] IC_{50} is the half-maximal inhibitory concentration of a compound at inhibiting the signal of fixed concentration of 1 (33 nM) in a competition assay. [e] I_{max} is the maximum stimulation achieved by a fixed concentration of 1 (33 nM) when competed with a 10 μM dose of a compound. IC_{50} and I_{max} values represent the average of N independent experiments in triplicate and were calculated by use of Prism software as described in the Experimental Procedures section. The standard

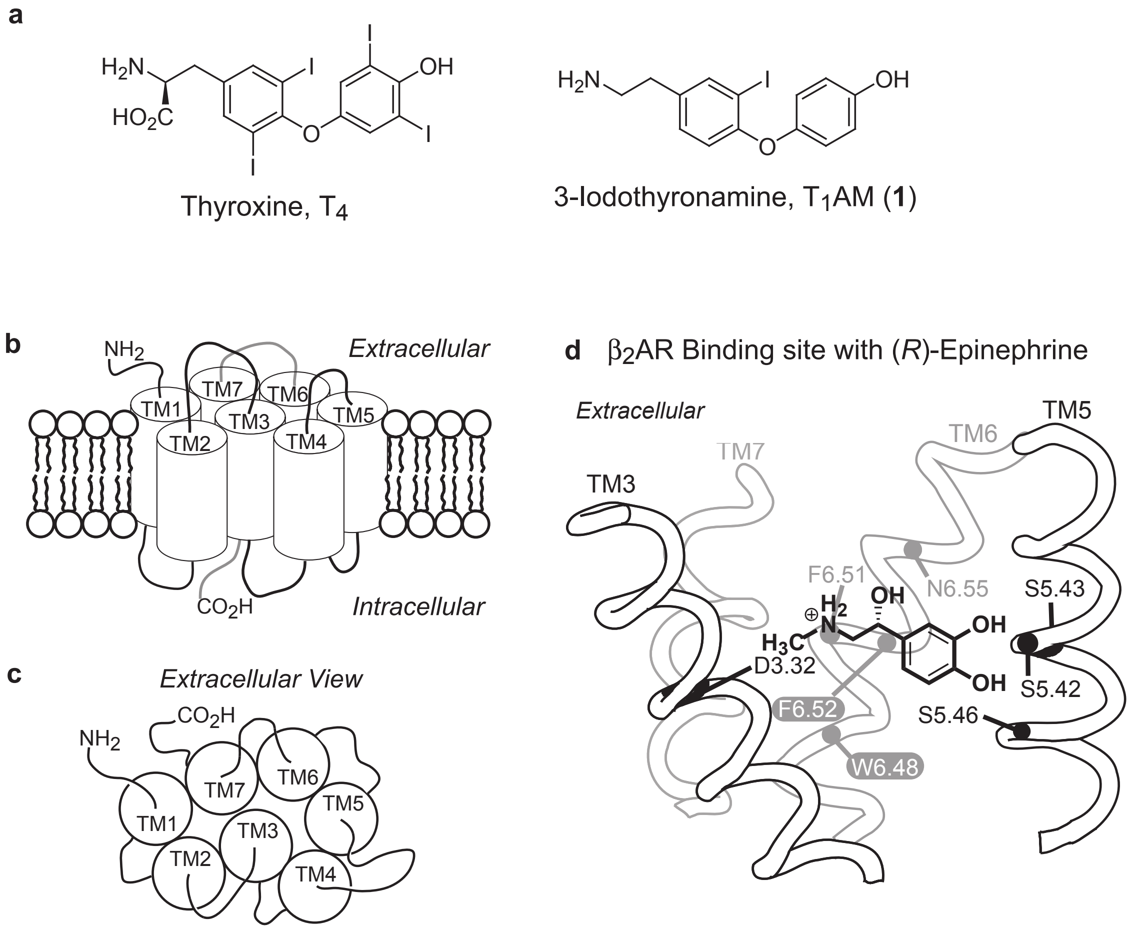
error of the mean (SEM) were calculated from the IC_{50} and I_{Max} values of each independent triplicate experiment by use of Prism software. I_{max} = 100 % is defined as the activity of 1 at 10 μM. I_{max} of T_1AM at 33 nM was 45 ± 5 %. [f]N is the number of independent experiments in triplicate that were performed and used to calculate the IC_{50} and I_{max} values.

Table 4. Agonist and antagonist activity of compounds **50**-**56** on $rTAAR_1$.

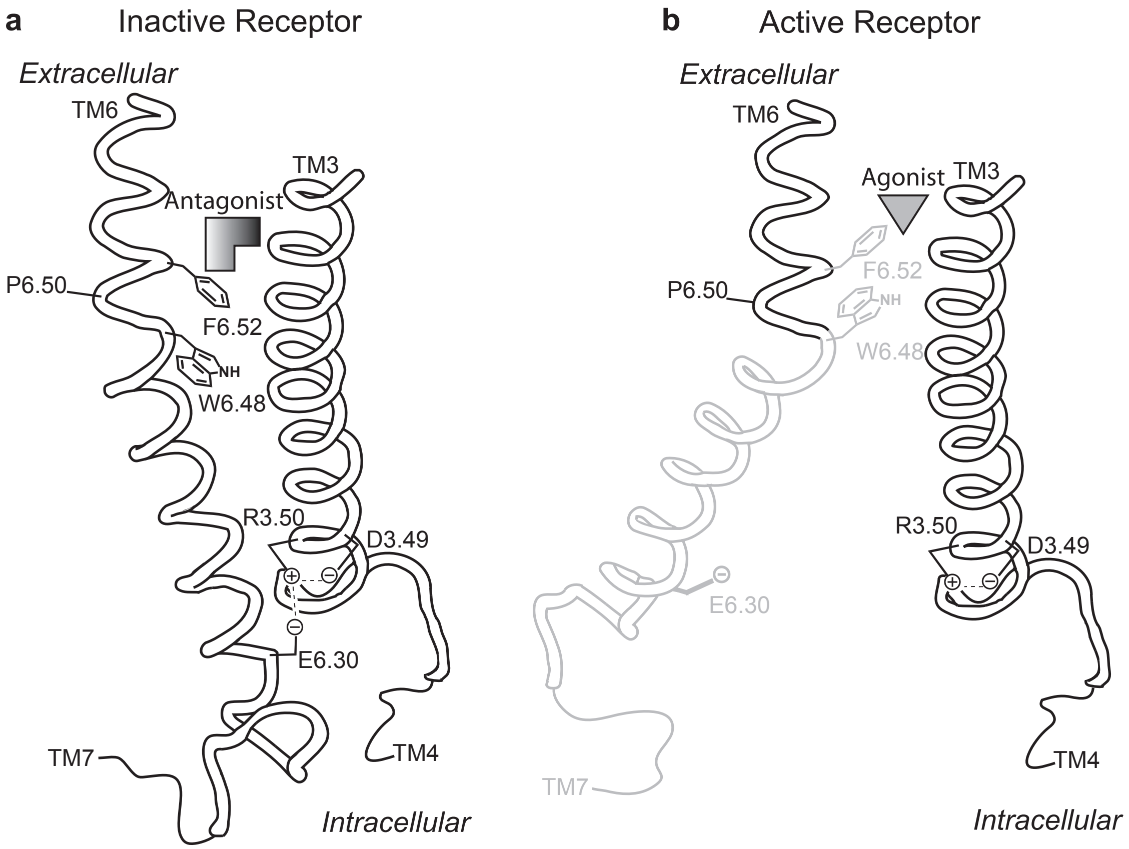


				Agonist Activity			Antagonist Activity		
Compd	R_7	R_8	R_9	EC_{50}[a] ± SEM (nM)	E_{max}[b] ± SEM (%)	*N*[c]	IC_{50}[d] ± SEM (μM)	I_{max}[e] ± SEM (%)	*N*[c]
50	H	Ph	H	>1000	37 ± 9	2	-	-	-
51	Ph	H	H	201 ± 23	59 ± 6	2	-	-	-
52	Ph	Ph	CH_3	>1000	0 ± 3	2	5 ± 1	10 ± 4	3
53	*p*-OH-Ph	Ph	H	>1000	16 ± 3	3	-	-	-
54	*p*-F-Ph	Ph	H	>1000	0 ± 3	2	3 ± 0	0 ± 3	3
55	*m*-F-Ph	Ph	H	>1000	0 ± 4	2	3 ± 1	0 ± 5	3
56	*m*-CN-Ph	Ph	H	>1000	0 ± 4	2	3 ± 1	2 ± 4	3

[a-f] See footnotes for Table 3.

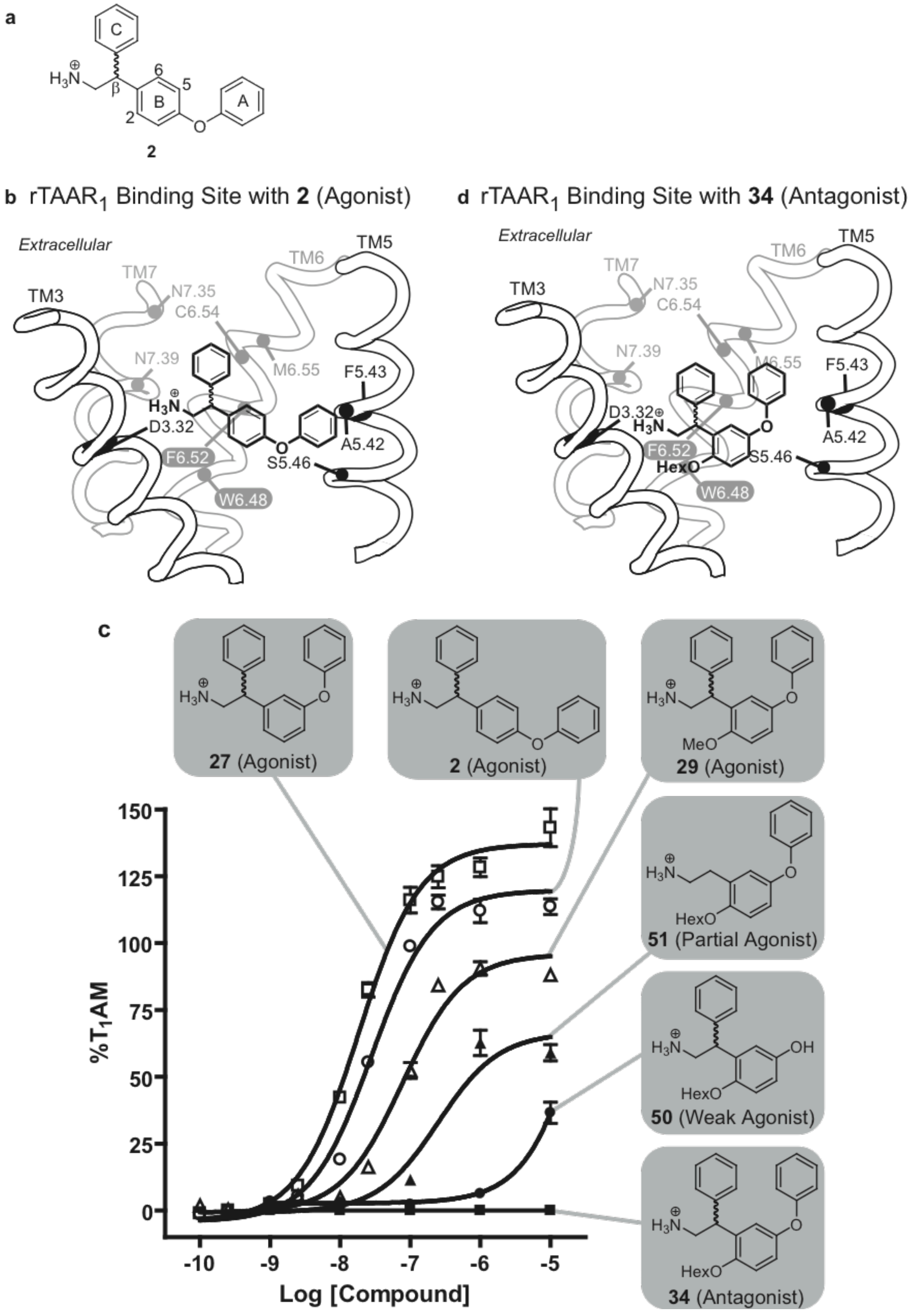


Chapter 3: Figure 1. Hormones, metabolites, and biogenic amine GPCR. (a) Structures of thyroxine (T_4) and 3-iodothyronamine (1, T_1AM). (b) Schematic representations of the helical arrangement of GPCRs viewed from the cell membrane and (c) extracellular surface. (d) Binding orientation of (*R*)-epinephrine in the binding site of the β_2AR viewed from the perspective of TM4. The location of the rotamer switch residues (white letters) (see Fig. 2) and residues known to interact with (*R*)-epinephrine are labeled. The residue indexing system is described in the Experimental Procedures section.



Chapter 3: Figure 2. Rotamer toggle switch model of aminergic GPCR activation.

(a) Inactive state of the receptor with an antagonist sterically occluding the rotamer switch residues (W6.48 and F6.52) from assuming their active conformation. (b) Agonist binding toggles the rotamer switch to its active conformation and induces a conformational change in TM6 that breaks the ionic lock interaction (D3.49, R3.50, & E6.30) present in the inactive state of the receptor. (a) and (b) are viewed from the perspective of TM7, see Fig. 1b-c.



b rTAAR$_1$ Binding Site with **2** (Agonist)

d rTAAR$_1$ Binding Site with **34** (Antagonist)

Chapter 3: Figure 3. SAR of rTAAR$_1$ ligands and their proposed binding mode in rTAAR$_1$.

(**a**) Structure of **2**. The A, B, and C rings correspond to its outer, inner, and β-phenyl rings, respectively. (**b**) Proposed binding orientation of **2** in the binding site of rTAAR$_1$, viewed from the perspective of TM4. The rotamer switch residues (white letters), proposed binding and specificity determinant residues are labeled. (**c**) Agonist dose response curves of **2** (○), **27** (□), **29** (Δ), **34** (■), **50** (●), and **51** (▲). (**d**) Proposed binding orientation of **34** in the binding site of rTAAR$_1$, viewed from the perspective of TM4.

Chapter 4: Molecular Recognition of 3-Iodothyronamine and Related Analogs by the Rat and Mouse Trace Amine Associated Receptor 1

Edwin S. Tan,[a] John C. Naylor,[d] Eli S. Groban,[b,c] James R. Bunzow,[d] Matthew P. Jacobson,[c] David K. Grandy,[d] and Thomas S. Scanlan[d]

[a]Chemistry and Chemical Biology Graduate Program, University of California, San Francisco, San Francisco, CA 94158-2280.

[b]Graduate Group in Biophysics, University of California, San Francisco, San Francisco CA 94158-2240.

[c]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA

[d]Department of Physiology and Pharmacology, Oregon Health and Science University, Portland, OR 97239, USA

Abstract

The trace amine associated receptor 1 ($TAAR_1$) is an aminergic G protein-coupled receptor (GPCR) potently activated by 3-iodothyronamine (1), an endogenous metabolite of thyroid hormone. Structure activity relationship studies on 1 and related agonists showed the rat and mouse species of $TAAR_1$ to be able to accommodate structural modification and functional groups on the ethylamine portion and the biaryl ether moiety of the molecule. However, each receptor clearly exhibited distinct ligand preferences. In this study, we generated single and double mutants of rat and mouse $TAAR_1$ to understand the molecular recognition of agonists and the underlying basis for the ligand selectivity of rat and mouse $TAAR_1$. Key, non-conserved specificity determinant residues in transmembranes 4 and 7 within the binding appear to be the primary source for some of the observed ligand preferences. Residue 7.39 in transmembrane 7 dictated the preference for a β-phenyl ring while residue 4.56 in transmembrane 4 was partially responsible for the lower potency of 1 and tyramine for the mouse receptor. Addtionally, 1 and tyramine were found to have the same binding mode in rat $TAAR_1$ despite structure activity relationship data suggesting the possibility of each molecule having different docking orientations in the binding site. These findings expand the body of knowledge regarding the critical binding site residues involved in the ligand-receptor interaction that can influence compound selectivity and functional activity of aminergic GPCRs.

Introduction

3-Iodothyronamine (1, T_1AM; Fig. 1) is an endogenous metabolite of the thyroid hormone thyroxine (T_4; Fig. 1) detected in the brain, heart, liver, blood (*1, 2*). It has profound physiological effects *in vivo* that are opposite to thyroid hormone. Compound 1 induces anergia, hypothermia, bradycardia, hyperglycemia, and hyperinsulinemia when administered to mice and rapidly reduces cardiac output in an *ex vivo* working rat heart preparation (*2-6*). Additionally, 1 can increase food intake and influence energy metabolism by favoring lipid utilization over carbohydrate consumption (*7, 8*).

These physiological responses to 1 may be mediated by more than one molecular target. *In vitro*, 1 can activate aminergic G protein-protein coupled receptors (GPCRs) in the biogenic amine subfamily; stimulating the production of cAMP via the trace amine-associated receptor 1 ($TAAR_1$) and the degradation of cAMP via the α_{2A} adrenergic receptor (*2, 6, 9, 10*). Additionally, 1 has a neuromodulatory role inhibiting neurotransmitter reuptake by the dopamine (DAT) and norepinephrine transporter (NET), and inhibiting vesicular packaging by the vesicular monoamine transporter 2 (VMAT2) (*11*).

To date 1 is the most potent endogenous molecule that can activate rodent $TAAR_1$. $TAAR_1$ is a member of the trace amine associated receptor family of orphan GPCRs, which consists of 19 rat, 16 mouse, and 9 human subtypes (*12-14*). It's homologous to the dopamine, adrenergic, and serotonin receptors and expressed in multiple tissues including the heart, kidney, liver, spleen, and pancreas. Although the biological function of many TAAR subtypes have yet to be defined, some mouse TAARs have recently been implicated as a secondary class of chemosensory receptors expressed in the olfactory epithelium (*15*).

In an effort to determine the role of $TAAR_1$ in mediating the effects of 1, we have previously explored the structure activity relationship (SAR) of 1 and developed a number of analogs with improved agonist activity (*9, 16*). These SAR studies showed rat and mouse $TAAR_1$ ($rTAAR_1$ and $mTAAR_1$, respectively) could tolerate structural modifications and substitutents on the ethylamine portion and biaryl ether moiety of 1. However the two receptors clearly have distinct structural preferences. In the ethylamine chain, for example, $rTAAR_1$ favored unsaturated hydrocarbon substitutents while $mTAAR_1$ preferred polar and hydrogen bond acceptors (*16*). Compound 1 and tyramine (Table 1), an endogenous phenethylamine metabolite, also exhibited some degree of species variability being ~7 to 10-fold less potent for $mTAAR_1$ compared to $rTAAR_1$ (*9*). These distinct ligand selectivity of rat and mouse $TAAR_1$ were remarkable given that the two receptors are 93% similar. Since the rodent receptors are only 83%-85% similar to human $TAAR_1$, understanding the molecular basis of species variability will undoubtedly have important implications in the development of activators and inhibitors for human $TAAR_1$. In this study, we explored the molecular recognition of 1 and related agonists by $TAAR_1$ and identified specificity determinant residues that give rise to the disparate ligand preferences between rat and mouse $TAAR_1$.

Results and Discussion

Our initial structure activity relationship study identified the β-phenylphenoxyphenethylamine (2 & 3) and phenoxynaphethylamine (4) as promising scaffolds for the development of small molecule regulators of $TAAR_1$ (Fig. 1) (*16*). Compounds 2 and 3 were both $rTAAR_1$ selective being 357- to 526-fold more potent for $rTAAR_1$ compared to $mTAAR_1$ (Table 1). Compound 4, on the other hand, was more or less equally beneficial for both rat and mouse $TAAR_1$ with a potency disparity between the two receptors within 5-fold. Using the toggle switch model of aminergic GPCR activation as a guideline, 2 was further developed to give superagonists (agonists that are more potent and/or efficacious than 1) 5, 6, and 7 (Table 1) (*17*). These superagonists remained $rTAAR_1$ selective exhibiting potencies that are $\geq$ 1600- to 2500-fold better in $rTAAR_1$ versus $mTAAR_1$.

The poor agonist activity of the β-phenylphenoxyphenethylamines for $mTAAR_1$ can be attributed to the β-phenyl group of the molecule and not its outer ring moiety (Fig. 1). Removing the outer rings of 2 and 5 (8 and 9, respectively) did not increase the potency for $mTAAR_1$ (Table 1). Both 8 and 9 still activated $mTAAR_1$ poorly with a potency $\geq$10μM. In contrast, eliminating the β-phenyl ring of 2, 5, and 7 to give 1 or 10 improved the potency for $mTAAR_1$ ~24- to 32-fold and decreased the $rTAAR_1$ selectivity from $\geq$ 350- to 2500-fold down to ~7-10-fold. Detailed synthetic procedures for 8 and 9 are described in the supplementary information.

Aminergic GPCRs are heptahelical transmembrane proteins with an extracellular amino terminus and an intracellular carboxy terminus (Fig. 2a). The binding site of aminergic GPCRs are located within the transmembrane (TM) region of the receptor and is predominantly composed of residues from TM 3, 5, 6 and 7 (*18-21*). Based on pharmacological and mutagenesis studies, epinephrine is proposed to bind to the β_2-adrenergic receptor (β_2AR) with aspartic acid 3.32 (D3.32) acting as the counterion for the charged amine, serine residues 5.42, 5.43, and 5.46 (S5.42, S5.43, and S5.46, respectively) interacting with the catechol hydrodroxyls, phenylalanines 6.51 and 6.52 (F6.51 and F6.52) interacting with the catechol ring, and asparagines 6.55 (N6.55) as the partner for the β-hydroxy group. (Fig. 2b) (See experimental procedures for a description of the residue indexing system) (*22-29*). By analogy to the catecholamine receptors (dopamine, epinephrine, and norepinephrine) we have previously deduced 2 to bind to rTAAR$_1$ with the charged amine forming a salt bridge interaction with D3.32 and the biaryl ether oxygen hydrogen bonding to S5.46 (Fig. 2c) (*17*). In this binding orientation, our homology model of rTAAR$_1$ showed the β-phenyl ring to be positioned near the interface between TM 6 and 7 and surrounded by cysteine 6.54, methionine 6.55, and asparagines 7.35 and 7.39 (C6.54, M6.55, N7.35, and N7.39, respectively).

In mTAAR$_1$, C6.54 and N7.35 are conserved but not M6.55 and N7.39. Instead mTAAR$_1$ has a threonine and tyrosine at 6.55 and 7.39, respectively (T6.55 and Y7.39) (Fig. 2d). In the β_2AR, the residues at 6.55 and 7.39 have been previously shown to interact with the ligand and alter the receptor's ligand specificity (*28, 30*).

Since the β-phenyl ring is proposed to be in the vicinity of 6.55 and 7.39, and both residues are not conserved between rat and mouse $TAAR_1$, we hypothesized that one or both of these residues are specificity elements that influence compatibility with β-phenyl ring of the β-phenylphenoxyphenethylamine scaffold. We tested this hypothesis by measuring the activity of 3 and 5, representative β-phenylphenoxyphenethylamines, against rat and mouse $TAAR_1$ HEK293 cells stably expressing single or double swap mutants at 6.55 and/or 7.39.

Residue 7.39 controls specificity for the β-phenyl ring of $TAAR_1$ ligands

Swapping residue 6.55 of rat and mouse $TAAR_1$ had minor effects on the activity and selectivity of 3 and 5. In the $rTAAR_1$ 6.55 single mutant [$rTAAR_1$(M6.55T)] the potency and efficacy of 3 (EC_{50} = 55 ± 20 nM, E_{max} = 119 ± 6%) and 5 (EC_{50} = 11 ± 1 nM, E_{max} = 117 ± 20%) decreased 2- to 3-fold and ≤ 12%, respectively (Table 2). When T6.55 was mutated to a methionine in $mTAAR_1$ [$mTAAR_1$(T6.55M)], both compounds were still poor agonists activating the receptor with potencies >10μM and efficacies ranging from 20% to 45%. Despite the mutation at residue 6.55, both 3 and 5 remained considerably $rTAAR_1$ selective (182- to 909-fold) (Fig. 3).

The single swap mutants at residue 7.39 had opposing effects on the activity of 3 and 5. The potency of both compounds decreased 167- to 526-fold in the $rTAAR_1$ 7.39 mutant [$rTAAR_1$(N7.39Y)] (EC_{50} = 10 μM and 1 μM for 3 and 5, respectively) but increased 60- to 100-fold in its $mTAAR_1$ 7.39 mutant counterpart [mTAAR$_1$(Y7.39N)] (EC$_{50}$ = 102 ± 21 nM and 176 ± 32 nM for 3 and 5, respectively) (Table 2). The efficacy of 3 (E$_{max}$ = 95 ± 6%) and 5 (E$_{max}$ = 91 ± 11%) both increased 29% to 89% in mTAAR$_1$(Y7.39N). For rTAAR$_1$(N7.39Y) the efficacy of 5 (E$_{max}$ = 126%) increased 12% while that of 3 (E$_{max}$ = 85%) decreased 46% compared to wild type. Interestingly, swapping the residues at 7.39 converted both compounds to good mTAAR$_1$ agonists that were now 6- to 98-fold selective for mTAAR$_1$ over rTAAR$_1$ (Fig. 3).

The activity profile of 3 and 5 for the double swap mutants was similar to that of the 7.39 single mutants but showed some enhancements in both potency and efficacy. The potency of 3 and 5 decreased 333- to 526-fold for the rTAAR$_1$ 6.55 and 7.39 double mutant [rTAAR$_1$(M6.55T/N7.39Y)] but increased 70- to 320-fold for the mTAAR$_1$ 6.55 and 7.39 double mutant equivalent [mTAAR$_1$(T6.55M/Y7.39N)] (Table 2). The same trend was also observed with regards to efficacy; decreasing 44% to 85% for rTAAR$_1$(M6.55T/N7.39Y) and increasing 39% to 95% for mTAAR$_1$(T6.55M/Y7.39N). Compared to the 7.39 single mutant, the mTAAR$_1$ selectivity of 3 and 5 in the double mutant was more pronounced. Compounds 3 and 5 were now twice as potent (14- to 233-fold vs 6- to 98-fold) and more efficacious (10% to 64% versus 31%) for mTAAR$_1$ than rTAAR$_1$ (Fig. 3).

The decrease in activity of 3 and 5 for the 6.55 and/or 7.39 single and double mutants in rTAAR$_1$ cannot be attributed to the introduced mutations compromising the functional competency of the receptors because the activity of the positive controls (1 and 4) for the same mutants only changed ≤ 4-fold and ≤ 16% in terms of potency and efficacy, respectively (Table 2). Likewise, the enhanced activity of 3 and 5 for the mTAAR$_1$ single and double swap mutants is not a consequence of the mutations rendering the receptors constitutively active and more responsive to agonists because the potency of the positive controls (1, 4, and/or 11) only changed ≤ 2-fold and the efficacy were essentially identical compared to the wild type receptor (Table 1 and 2). Compound 11 is a novel agonist for rat and mouse TAAR$_1$ that can be considered a halogen free analog of 1. Detailed synthetic procedures for 11 are described in the supplementary information.

Swapping residues at 7.39 was sufficient to convert 3 and 5 from a rTAAR$_1$ into a mTAAR$_1$ selective agonist. The TAAR$_1$ binding site appears to be able to accommodate a phenyl ring in the interface between TM 6 and 7 within the binding site near residue 7.39. Compounds 3 and 5 are poor agonists for wild type mTAAR$_1$ because this phenyl pocket near 7.39 is occupied by the phenol group of Y7.39 from the receptor and unavailable to the β-phenyl rings of 3 and 5. In contrast, 3 and 5 are excellent agonists for wild type rTAAR$_1$ because the smaller aspargine residue at 7.39 is less sterically encumbering and does not compete with the β-phenyl rings of 3 and 5 for the phenyl pocket near 7.39 of the receptor.

It is should be noted that the configuration of the outer ring does not significantly affect the specificity of TAAR$_1$ for β-phenyl ring bearing compounds. Regardless of whether the outer ring is at the *meta-* (3) or *para-* (5) position relative to the ethylamine chain, both 3 and 5 were affected similarly by mutations at 6.55 and/or 7.39 in rat and mouse TAAR$_1$; indicating that the β-phenyl ring of both molecules occupy the same binding pocket within the binding site. This result supports the assumptions we had proposed regarding the location of the antagonistic groups of the lead $rTAAR_1$ antagonists previously developed with 3 as the core scaffold (*17*).

Compound 1 and Tyramine have similar binding modes

In addition to being a superagonist for $rTAAR_1$, 7 (EC_{50} = 4 ± 1 nM, E_{max} = 115 ± 2%) is also an interesting molecule because it embodies both tyramine (EC_{50} = 65 ± 1 nM, E_{max} = 119 ± 7%) and 1 (EC_{50} = 33 ± 3 nM, E_{max} = 100 ± 0%) (Fig. 1 and 4). This hybrid compound potentially explains how two molecules with very different molecular volumes can elicit similar responses. If the β-phenyl group of 7 represents the aromatic ring of tyramine, then tyramine would occupy the phenyl pocket near 7.39 and thus have a binding mode different from 1 (Fig. 4). On the other hand, if the inner ring of 7 represents the tyramine aromatic ring then tyramine and 1 would have similar binding modes. The $rTAAR_1$ agonist activity of 9 (EC_{50} = 115 ± 12 nM, E_{max} = 105 ± 5%) supports the feasibility of tyramine potentially having two alternate binding modes in the rat receptor (Table 1). Since the tyrosine residue at 7.39 in $mTAAR_1$ abolished the phenyl pocket near 7.39, tyramine must share the same binding mode as 1 in the mouse species of $TAAR_1$.

To determine if tyramine and 1 have similar or distinct binding modes, we mutated alanine 5.42 in $rTAAR_1$ to a threonine, leucine, or isoleucine [$rTAAR_1$(A5.42T), $rTAAR_1$(A5.42L), and $rTAAR_1$(A5.42I), respectively] and examined the effects of the mutations on the potency of tyramine and 1. The idea behind these 5.42 single mutants is to perturb the pocket occupied by the outer ring of 1, 7 and other biaryl ether containing molecules by introducing more sterically demanding residues than alanine. Threonine was chosen because SAR studies on human $TAAR_1$, which has a threonine at residue 5.42, showed 1 to be a poor agonist.(*10*).

In the $rTAAR_1$(A5.42L) mutant, the potency of 1 (EC_{50} = 58 ± 16 nM) and tyramine (EC_{50} = ~2 μM) decreased 2- and 31- fold, respectively (Table 3a). When A5.42 was mutated to an isoleucine, there was a 3- and 154-fold decrease in the potency of 1 (EC_{50} = 108 ± 14 nM) and tyramine (EC_{50} = >10 μM), respectively. The efficacy of tyramine in the leucine mutant decreased 53% to 70% compared to wild type $rTAAR_1$. A similar trend was observed for the $rTAAR_1$(A5.42T) mutant where 1 (EC_{50} = 88 ± 13 nM) and tyramine (EC_{50} = >10 μM) were 3- and 154-fold less potent. Additionally, tyramine was 50% less efficacious for this receptor than wild type. Since the activity of the positive control (11) for all 5.42 mutants were essentially unaffected, the reduced activities of 1 and tyramine for these mutants cannot be due to a compromised activation capacity of these receptors.

The observed effects on the activity of 1 and tyramine in the $rTAAR_1$ 5.42 mutants are consistent with both compounds having similar binding modes. If tyramine preferentially occupied the phenyl pocket near residue 7.39, its potency would not have been affected by changes to residue 5.42. Since its activity decreased 31- to 154-fold when residue 5.42 was mutated, tyramine is probably binding to

$rTAAR_1$ with its hydroxyl group engaged in hydrogen bond interactions with S5.46, the residue one turn below 5.42 in TM5(Fig. 2d). This binding mode corresponds to the same binding orientation of 1 for rat and mouse $TAAR_1$.

Although the leucine and isoluecine mutations increased the steric bulk around residue 5.42, it's interesting that the agonist activity of the smaller tyramine was more severely affected than the larger 1 or 11; especially when these mutations were intended to block the binding pocket for the outer ring. These results suggest that despite being bigger than alanine, leucine and isoleucine are not large enough to completely abolish the outer ring binding pocket. The small effects of the mutations on the agonist activity of 1 and 11 can potentially be attributed to their larger number of interactions with the receptor compared to tyramine. Like tyramine, 1 and 11 should both be anchored in the binding site of $rTAAR_1$ by a salt bridge interaction between the charged amine and D3.32, and a hydrogen bond interaction with S5.46. However, the extra functional groups present in 1 and 11 (i.e. outer ring, iodine, and naphthyl ring) (Table 1) can make additional interactions with TM 5 and 6 that are not available to tyramine. With more contacts to the receptor, 1 and 11 would be less sensitive to structural changes at residue 5.42 than tyramine.

Residue 4.56 is partially responsible for the lower potency of 1 for $mTAAR_1$

Within the thyronamine series, rat and mouse $TAAR_1$ had the same rank order potency but the potency values of individual compounds for the two receptors were not identical. In general, thyronamines are ~10-fold less potent for $mTAAR_1$

compared to $rTAAR_1$ (*2, 9*). A possible explanation for this potency disparity can be attributed to a lower G-protein coupling efficiency for $mTAAR_1$ versus $rTAAR_1$. If this were the case, then it would be impossible to have an equipotent agonist for both receptors because $mTAAR_1$ would be inherently less active than $rTAAR_1$. Since 12 was found to be an equipotent agonist for $rTAAR_1$ (EC_{50} = 65 ± 6 nM, E_{max} = 115 ± 2%) and $mTAAR_1$ (EC_{50} = 82 ± 17 nM, E_{max} = 112 ± 3%), the G-protein coupling efficiency of $mTAAR_1$ is comparable to that of $rTAAR_1$ (Table 1). Detailed synthetic procedures for 12 are described in the supplementary information.

We hypothesized the potency disparity of thyronamines to be brought about by non-conserved amino acid(s) at key specificity determinant residues within the binding site. In particular, we speculated that tyrosine 4.56 (Y4.56) was primarily responsible for the ~10-fold lower potency of 1 for $mTAAR_1$. This residue was deductively identified through a process of elimination using the following logic: (1) since the binding sites of GPCRs are located within the transmembrane regions of the receptor, all intracellular and extracellular loops as well as the amino- and carboxy-terminus were eliminated (*21*); (2) amino acid differences in TM 1 and 2 were eliminated because the binding site is primarily composed to TM 3, 4, 5, 6, and 7 (*23*); (3) since the ethylamine chain of 1 and 12 are exactly the same, non-conserved residues in TM 3, 6 and 7 cannot be responsible; (4) TM5 was eliminated because it is absolutely conserved between the two species; (5) Y4.56 was the only non-conserved residue remaining when the intracellular half of TM 4 was eliminated because the binding site of GPCRs is located in the extracellular half of the transmembrane region (Fig. 5). In our homology model of rat and mouse $TAAR_1$, residue 4.56 was found to be in the vicinity of the purported outer ring binding pocket and could conceivably make contacts with the bound ligand. To test the importance of residue 4.56, we generated rat and mouse $TAAR_1$ single swap mutants at this location.

When Y4.56 of $mTAAR_1$ was converted to phenylalanine [$mTAAR_1$(Y4.56F)], the potency of 1 increased 5-fold to from 314 ± 43 nM to 67 ± 17 nM (Table 1 and 3b). Interestingly, the potency of 1 (EC_{50} = 38 ± 11 nM) for the tyrosine mutant of $rTAAR_1$ [$rTAAR_1$(F4.56Y)] was comparable to wild type $rTAAR_1$ (33 ± 3 nM). The same trend was also observed for tyramine; where its potency increased 7-fold in $mTAAR_1$(Y4.56F) (EC_{50} = 40 ± 13 nM) but unaffected in $rTAAR_1$(F4.56Y) (EC_{50} = 54 ± 11 nM). Since the agonist activity of the positive controls (4 and 11) for the mutants was comparable to that of the wild type receptors for both species, the mutations did not compromise the functional capacity of either receptors.

Residue 4.56 appears to play an important role for the lower potency of 1 for $mTAAR_1$. Swapping this residue with that found in $rTAAR_1$ increased the potency of 1 in $mTAAR_1$ to be within two fold of the potency value for wild type $rTAAR_1$. Interestingly, the potency of tyramine for $mTAAR_1$(Y4.56F) was equivalent to that of wild type $rTAAR_1$. Although mutating residue 4.56 improved the potency in the $mTAAR_1$ mutants, the reciprocal effect of 1 and tyramine becoming less potent for $rTAAR_1$(F4.56Y) was not observed. This indicates that 4.56 is only partially responsible and not the sole basis for the observed potency disparity of 1 and tyramine between the rat and mouse $TAAR_1$. Overall, these results implicate residues in TM4 to possibly make contacts with the ligand and affect receptor activation.

Conclusion and Significance

The disparate ligand structural preferences exhibited by rat and mouse TAAR_1 can be attributed to key, non-conserved specificity determinant residues within the binding site. Residue 7.39 appears to dictate the specificity for a β-phenyl ring; the bulky tyrosine residue at 7.39 in mTAAR_1 sterically clashed with the β-phenyl ring whereas the smaller asparagine at the same location in rTAAR_1 was more compatible and able to accomodate a β-phenyl moiety. The lower potency of 1 in mTAAR_1 was partly caused by the presence of a tyrosine at residue 4.56 rather than a phenylalanine. Although compound 7 implied the possibility of 1 and tyramine having different binding modes in the binding site of rTAAR_1, 1 and tyramine appear to have the same docking orientation.

With the recent developments and accomplishments in structure determination of the β_2 adrenergic receptor, the practicality of a structure-based drug design approach towards developing activators and inhibitors for aminergic GPCRs has never been so promising. A critical aspect to the success of this strategy will depend on having insights into the molecular basis of ligand recognition, the mechanism of GPCR activation, and the relationship of how these ligand-receptor interactions are relayed and translated into receptor activation or inhibition. The information presented herein should prove beneficial towards this cause as it provides valuable information regarding the binding site residues involved in ligand-receptor interactions that can influence compound specificity and functional activity of an aminergic GPCR.

Methods

Residue Indexing Scheme. Residues are labeled relative to the most conserved amino acid in the transmembrane segment in which it is located.(*31*) Asparagine 7.39, for example, is located in transmembrane 7 and precedes the most conserved residue by 11 positions. Proline 6.50 is the most conserved residue in TM6. This system simplifies the identification of corresponding residues in different GPCRs.

$TAAR_1$ Site-Directed Mutagenesis. $TAAR_1$ mutants were generated by using the QuikChange Site-Directed Mutagenesis Kit (Stratagene, La Jolla , CA). Primers were designed that coded for the desired mutation flanked with 10-15 base pairs of sequence. Complementary oligonucleotides were then used in PCR using an expression plasmid containing the desired receptor as template. The PCR product was then digested with Dpn I and transformed into XL1 Blue competent cells. Colonies were picked and the DNA isolated was sequenced to confirm the mutation.

The DNA for the mutants was then used to transfect HEK293 (human embryonic kidney) cells using Fugene (Roche, Indianapolis, IN) and stable cell lines were made for further assays under G418 selection.

Homology Model of rat and mouse $TAAR_1$. The sequence of rTAAR1 was imported into the Prime Software package (commercially available from Schrödinger Inc.) and aligned with the sequence of the human B2-adrenergic G protein coupled receptor, for which a structure was recently solved (Protein Data Bank Accesion code 2RH1) (cite the Cherezov, Rosenbaum, 2007 paper "High resolution crystal structure of human..). Instead of using a multiple alignment to create a sequence profile, the align GPCR program in Prime was used to create an optimal sequence alignment. We constructed our homology model of $rTAAR_1$ based on the crystal structure of human B2-adrenergic G protein coupled receptor as a template and used the homology modeling suite in Prime. The modeling program did not modify conserved residues, leaving each atom in these residues at their original PDB coordinates. Non-conserved side chains were built onto the structure using the backbone coordinates for bovine rhodopsin as a reference point. All chain breaks or gaps were closed using a previously published loop building and optimization algorithm (*34*). After building the complete model, side chain optimization, followed by backbone and side chain energy minimization, was performed on all non-conserved residues. The homology modeling program relies on the OPLS all atom force field (*35-37*) and a Generalized Born solvent model (*38, 39*) to evaluate the energy of different conformations and select the lowest energy structure as the final model. This procedure was repeated with the sequence of mTAAR1 to generate a model for the mouse ortholog of this receptor.

Synthesis. Detailed synthetic procedures and chemical compound information of novel molecules are described in the supplemental information.

In Vitro cAMP Agonist Activity Assay. Compounds were tested as described previously(*17*) Data were reported relative to **1** and expressed as $\%T_1AM$. The activity of **1** at 10μM was set as 100 $\%T_1AM$. Concentration-response curves were plotted and EC_{50} values were calculated with Prism software (GraphPad, San Diego, CA).

Standard error of the mean was calculated from the EC_{50} and E_{Max} values of each independent triplicate experiment by use of Prism Software.

References

1 Piehl, S., Heberer, T., Balizs, G., Scanlan, T. S., Smits, R., Koksch, B., and Kohrle, J. (2008) Thyronamines are isozyme-specific substrates of deiodinases. *Endocrinology 149*, 3037-45.

2 Scanlan, T., Suchland, K., Hart, M., Chiellini, G., Huang, Y., Kruzich, P., Frascarelli, S., Crossley, D., Bunzow, J., Ronca-Testoni, S., Lin, E., Hatton, D., Zucchi, R., and Grandy, D. (2004) 3-Iodothyronamine is an endogenous and rapid-acting derivative of thyroid hormone. *Nat. Med. 10*, 638-642.

3 Chiellini, G., Frascarelli, S., Ghelardoni, S., Carnicelli, V., Tobias, S., DeBarber, A., Brogioni, S., Ronca-Testoni, S., Cerbai, E., Grandy, D., Scanlan, T., and Zucchi, R. (2007) Cardiac effects of 3-iodothyronamine: a new aminergic system modulating cardiac function. *FASEB J. 21*, 1597-1608.

4 Doyle, K., Suchland, K., Ciesielski, T., Lessov, N., Grandy, D., Scanlan, T., and Stenzel-Poore, M. (2007) Novel thyroxine derivatives, thyronamine and 3-iodothyronamine, induce transient hypothermia and marked neuroprotection against stroke injury. *Stroke 38*, 2569-2576.

5 Frascarelli, S., Ghelardoni, S., Chiellini, G., Vargiu, R., Ronca-Testoni, S., Scanlan, T. S., Grandy, D. K., and Zucchi, R. (2008) Cardiac effects of trace amines: pharmacological characterization of trace amine-associated receptors. *Eur. J. Pharmacol. 587*, 231-6.

6 Regard, J. B., Kataoka, H., Cano, D. A., Camerer, E., Yin, L., Zheng, Y. W., Scanlan, T. S., Hebrok, M., and Coughlin, S. R. (2007) Probing cell type-specific functions of Gi in vivo identifies GPCR regulators of insulin secretion. *J. Clin. Invest. 117*, 4034-43.

7 Braulke, L. J., Klingenspor, M., DeBarber, A., Tobias, S. C., Grandy, D. K., Scanlan, T. S., and Heldmaier, G. (2008) 3-Iodothyronamine: a novel hormone controlling the balance between glucose and lipid utilisation. *Journal of Comparative Physiology B 178*, 167-77.

8 Dhillo, W. S., Bewick, G. A., White, N. E., Gardiner, J. V., Thompson, E. L., Bataveljic, A., Murphy, K. G., Roy, D., Patel, N. A., Scutt, J. N., Armstrong, A., Ghatei, M. A., and Bloom, S. R. (2008) The thyroid hormone derivative 3-iodothyronamine increases food intake in rodents. *Diabetes, Obesity and Metabolism*.

9 Hart, M., Suchland, K., Miyakawa, M., Bunzow, J., Grandy, D., and Scanlan, T. (2006) Trace amine-associated receptor agonists: Synthesis and evaluation of thyronamines and related analogues. *J. Med. Chem. 49*, 1101-1112.

10 Wainscott, D., Little, S., Yin, T., Tu, Y., Rocco, V., He, J., and Nelson, D. (2007) Pharmacologic characterization of the cloned human trace amine-associated receptor1 (TAAR1) and evidence for species differences with the rat TAAR1. *J. Pharmacol. Exp. Ther. 320*, 475-485.

11 Snead, A., Santos, M., Seal, R., Miyakawa, M., Edwards, R., and Scanlan, T. (2007) Thyronamines inhibit plasma membrane and vesicular monoamine transport. *ACS Chem. Biol. 2*, 390-398.

12 Borowsky, B., Adham, N., Jones, K., Raddatz, R., Artymyshyn, R., Ogozalek, K., Durkin, M., Lakhlani, P., Bonini, J., Pathirana, S., Boyle, N., Pu, X., Kouranova, E., Lichtblau, H., Ochoa, F., Branchek, T., and Gerald, C. (2001) Trace amines: Identification of a family of mammalian G protein-coupled receptors. *Proc. Natl. Acad. Sci. U. S. A. 98*, 8966-8971.

13 Bunzow, J., Sonders, M., Arttamangkul, S., Harrison, L., Zhang, G., Quigley, D., Darland, T., Suchland, K., Pasumamula, S., Kennedy, J., Olson, S., Magenis, R., Amara, S., and Grandy, D. (2001) Amphetamine, 3,4-methylenedioxymethamphetamine, lysergic acid diethylamide, and metabolites of the catecholamine neurotransmitters are agonists of a rat trace amine receptor. *Mol. Pharmacol. 60*, 1181-1188.

14 Lindemann, L., Ebeling, M., Kratochwil, N., Bunzow, J., Grandy, D., and Hoener, M. (2005) Trace amine-associated receptors form structurally and functionally distinct subfamilies of novel G protein-coupled receptors. *Genomics 85*, 372-385.

15 Liberles, S., and Buck, L. (2006) A second class of chemosensory receptors in the olfactory epithelium. *Nature 442*, 645-650.

16 Tan, E., Miyakawa, M., Bunzow, J., Grandy, D., and Scanlan, T. (2007) Exploring the structure-activity relationship of the ethylamine portion of 3-iodothyronamine for rat and mouse trace amine-associated receptor 1. *J. Med. Chem. 50*, 2787-2798.

17 Tan, E. S., Groban, E. S., Jacobson, M. P., and Scanlan, T. S. (2008) Toward deciphering the code to aminergic G protein-coupled receptor drug design. *Chem. Biol. 15*, 343-53.

18 Bridges, T. M., and Lindsley, C. W. (2008) G-protein-coupled receptors: from classical modes of modulation to allosteric mechanisms. *ACS Chem. Biol. 3*, 530-41.

19 Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., and Stevens, R. C. (2007) High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science 318*, 1258-65.

20 Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Yao, X. J., Weis, W. I., Stevens, R. C., and Kobilka, B. K. (2007) GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. *Science 318*, 1266-73.

21 Tota, M. R., Candelore, M. R., Dixon, R. A., and Strader, C. D. (1991) Biophysical and genetic analysis of the ligand-binding site of the beta-adrenoceptor. *Trends Pharmacol. Sci. 12*, 4-6.

22 Liapakis, G., Ballesteros, J., Papachristou, S., Chan, W., Chen, X., and Javitch, J. (2000) The forgotten serine - A critical role for Ser-203(5.42) in ligand binding to and activation of the beta(2)-adrenergic receptor. *J. Biol. Chem. 275*, 37779-37788.

23 Shi, L., and Javitch, J. (2002) The binding site of aminergic G protein-coupled receptors: The transmembrane segments and second extracellular loop. *Annu. Rev. Pharmacol. Toxicol. 42*, 437-467.

24 Strader, C., Candelore, M., Hill, W., Dixon, R., and Sigal, I. (1989) A single amino-acid substitution in the beta-adrenergic-receptor promotes partial agonist activity from antagonists. *J. Biol. Chem. 264*, 16470-16477.

25 Strader, C., Candelore, M., Hill, W., Sigal, I., and Dixon, R. (1989) Identification of 2 serine residues involved in agonist activation of the beta-adrenergic-receptor. *J. Biol. Chem. 264*, 13572-13578.

26 Strader, C., Fong, T., Tota, M., Underwood, D., and Dixon, R. (1994) Structure and function of G-protein-couple receptors. *Annu. Rev. Biochem. 63*, 101-132.

27 Strader, C., Sigal, I., Candelore, M., Rands, E., Hill, W., and Dixon, R. (1988) Conserved aspartic-acid residue-79 and residue-113 of the beta-adrenergic-receptor have different roles in receptor function. *J. Biol. Chem. 263*, 10267-10271.

28 Wieland, K., Zuurmond, H., Krasel, C., Ijzerman, A., and Lohse, M. (1996) Involvement of Asn-293 in stereospecific agonist recognition and in activation of the beta(2)-adrenergic receptor. *Proc. Natl. Acad. Sci. U. S. A. 93*, 9276-9281.

29 Zuurmond, H., Hessling, J., Bluml, K., Lohse, M., and Ijzerman, A. (1999) Study of interaction between agonists and Asn293 in helix VI of human beta(2)-adrenergic receptor. *Mol. Pharmacol. 56*, 909-916.

30 Suryanarayana, S., Daunt, D., Von Zastrow, M., and Kobilka, B. (1991) A point mutation in the 7th hydrophobic domain of the alpha-2 adrenergic-receptor increases its affinity for a family of beta-receptor-antagonists. *J. Biol. Chem. 266*, 15488-15492.

31 Ballesteros, J., and Weinstein, H. (1995) Integrated methods for the construction of three-dimensional models of structure-function relations in G protein-coupled receptors. *Methods Neuroscience 25*, 366-428.

32 Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Le Trong, I., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000) Crystal structure of rhodopsin: A G protein-coupled receptor. *Science 289*, 739-45.

33 Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research 32*, 1792-7.

34 Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., and Friesner, R. A. (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins: Struct., Funct., Bioinf. 55*, 351-67.

35 Jacobson, M., Kaminski, G., Friesner, R., and Rapp, C. (2002) Force field validation using protein side chain prediction. *J. Phys. Chem. B 106*, 11673-11680.

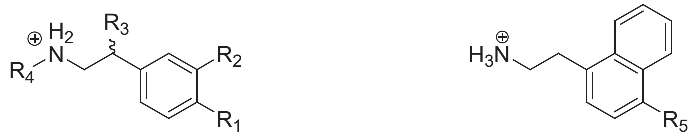
36 Jorgensen, W., Maxwell, D., and Tirado-Rives, J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc. 118*, 11225-11236.

37 Kaminski, G., Friesner, R., Tirado-Rives, J., and Jorgensen, W. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B 105*, 6474-6487.

38 Gallicchio, E., Zhang, L. Y., and Levy, R. M. (2002) The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem. 23*, 517-29.

39 Ghosh, A., Rapp, C., and Friesner, R. (1998) Generalized born model based on a surface integral formulation. *J. Phys. Chem. B 112*, 10983-10990.

Table 1. Agonist activity of **1-12** and **tyramine** on wild type rat and mouse TAAR$_1$



						rTAAR$_1$			mTAAR$_1$		
						EC$_{50}$^{a}	*E*$_{max}$^{b}	*N*^{c}	EC$_{50}$^{a}	*E*$_{max}$^{b}	*N*^{c}
Compd	**R$_1$**	**R$_2$**	**R$_3$**	**R$_4$**	**R$_5$**	**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**	
1 (T$_1$AM)	*p*-OH-Ph	I	H	H	-	33 ± 3	100 ± 0	5	314 ± 43	100 ± 0	5
2 (ET-13)	OPh	H	Ph	H	-	28 ± 2	103 ± 4	3	>10,000	35 ± 8	3
3 (ET-14)	H	OPh	Ph	H	-	19 ± 2	131 ± 7	3	>10,000	15 ± 4	3
5 (ET-36)	OPh	H	*p*-OH-Ph	H	-	6 ± 1	114 ± 9	4	>10,000	62 ± 6	3
6 (ET-64)	OPh	H	*p*-OH-Ph	CH$_3$	-	5 ± 1	127 ± 2	4	>10,000	42 ± 1	2
7 (ET-69)	*p*-OH-Ph	I	*p*-OH-Ph	H	-	4 ± 1	115 ± 2	6	>10,000	34 ± 5	3
8 (ET-71)	OH	H	Ph	H	-	78 ± 9	122 ± 16	3	>10,000	49	1
9 (ET-50)	OH	H	*p*-OH-Ph	H	-	115 ± 12	105 ± 5	3	>10,000	72	1
10 (PTA)	OPh	H	H	H	-	63 ± 7	93 ± 4	3	420 ± 66	85 ± 4	3
Tyramine	OH	H	H	H	-	65 ± 1	119 ± 7	3	271 ± 52	110 ± 2	3
4 (ET-21)	-	-	-	-	Ph	26 ± 1	113 ± 5	3	100 ± 22	104 ± 3	3
11 (ET-102)	-	-	-	-	*p*-OH-Ph	19 ± 3	96 ± 2	3	171 ± 13	98 ± 1	2
12 (1-NEA)	-	-	-	-	H	65 ± 6	115 ± 2	3	82 ± 17	112 ± 3	2

[a]EC$_{50}$ is the half-maximal effective concentration of a compound. [b]*E*max is the maximum stimulation achieved at a concentration of 10 μM and was calculated by use of Prism software. EC$_{50}$ and *E*$_{max}$ values represent the average of N independent experiments in triplicate and were calculated by use of Prism software as described in the Methods section. *E*$_{max}$ = 100 % is defined as the activity of **1** at 10 μM. [c]N is the number of independent experiments in triplicate that were performed and used to calculate the EC$_{50}$ and E$_{max}$ values.

Table 2. Agonist activity of **1**, **3-5**, and **11** on TAAR$_1$ TM 6 and/or 7 mutants

	rTAAR$_1$								
	M6.55T			**N7.39Y**			**M6.55T/N7.39Y**		
	EC$_{50}$^{a}	***E*$_{\text{max}}$^{b}**	***N*^{c}**	**EC$_{50}$^{a}**	***E*$_{\text{max}}$^{b}**	***N*^{c}**	**EC$_{50}$^{a}**	***E*$_{\text{max}}$^{b}**	***N*^{c}**
Compd	**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**	
1	129 ± 43	100 ±0	2	135	100	1	90	100	1
3	55 ± 20	119 ± 6	2	~10,000	85	1	>10,000	46	1
4	59 ± 0	117 ± 0	2	146	129	1	75	111	1
5	11 ± 1	117 ± 20	2	~1,000	126	1	~2,000	70	1

	mTAAR$_1$								
	T6.55M			**Y7.39N**			**T6.55M/Y7.39N**		
	EC$_{50}$^{a}	***E*$_{\text{max}}$^{b}**	***N*^{c}**	**EC$_{50}$^{a}**	***E*$_{\text{max}}$^{b}**	***N*^{c}**	**EC$_{50}$^{a}**	***E*$_{\text{max}}$^{b}**	***N*^{c}**
Compd	**± SEM (nM)**	**± SEM (%)**	***N*^{c}**	**± SEM (nM)**	**± SEM (%)**	***N*^{c}**	**± SEM (nM)**	**± SEM (%)**	***N*^{c}**
1	1418 ± 592	100 ± 0	2	251 ± 43	100 ± 0	5	208 ± 60	100 ± 0	3
3	>10,000	20 ± 4	2	102 ± 21	95 ± 6	5	43 ± 5	110 ± 8	3
4	350 ± 101	96 ± 13	2	173 ± 41	93 ± 15	2	-	-	-
5	>10,000	45 ± 5	2	176 ± 32	91 ± 11	5	138 ± 24	101 ± 9	3
11	-	-	-	179 ± 21	92 ± 6	3	134 ± 43	101 ± 11	3

[a-c] See footnotes for **Table 1**. Compound structures are shown in **Table 1**.

Table 3. Agonist activity of **1**, **4**, **11**, and **tyramine** on TAAR$_1$ TM 4 or 5 mutants

a. TAAR$_1$ TM 5 mutants

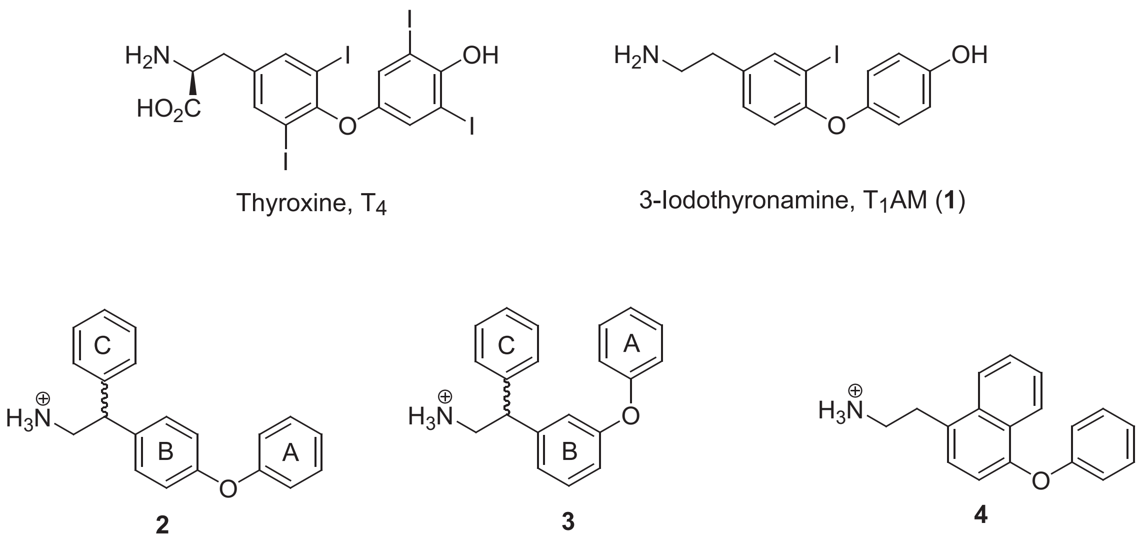
				rTAAR$_1$					
	A5.42L			A5.42I			A5.42T		
	EC$_{50}$^{a}	E_{max}^{b}	N^c	EC$_{50}$^{a}	E_{max}^{b}	N^c	EC$_{50}$^{a}	E_{max}^{b}	N^c
Compd	**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**	
1	58 ± 16	100 ± 0	3	108 ± 14	100 ± 0	3	88 ± 13	100 ± 0	4
Tyramine	~2,000	102 ± 18	3	>10,000	49 ± 7	3	>10,000	69 ± 5	4
11	32 ± 12	101 ± 8	3	29 ± 4	94 ± 2	3	25 ± 3	117 ± 13	4

[a-c] See footnotes for **Table 1**. Compound structures are shown in **Table 1**.

b. TAAR$_1$ TM 4 mutants

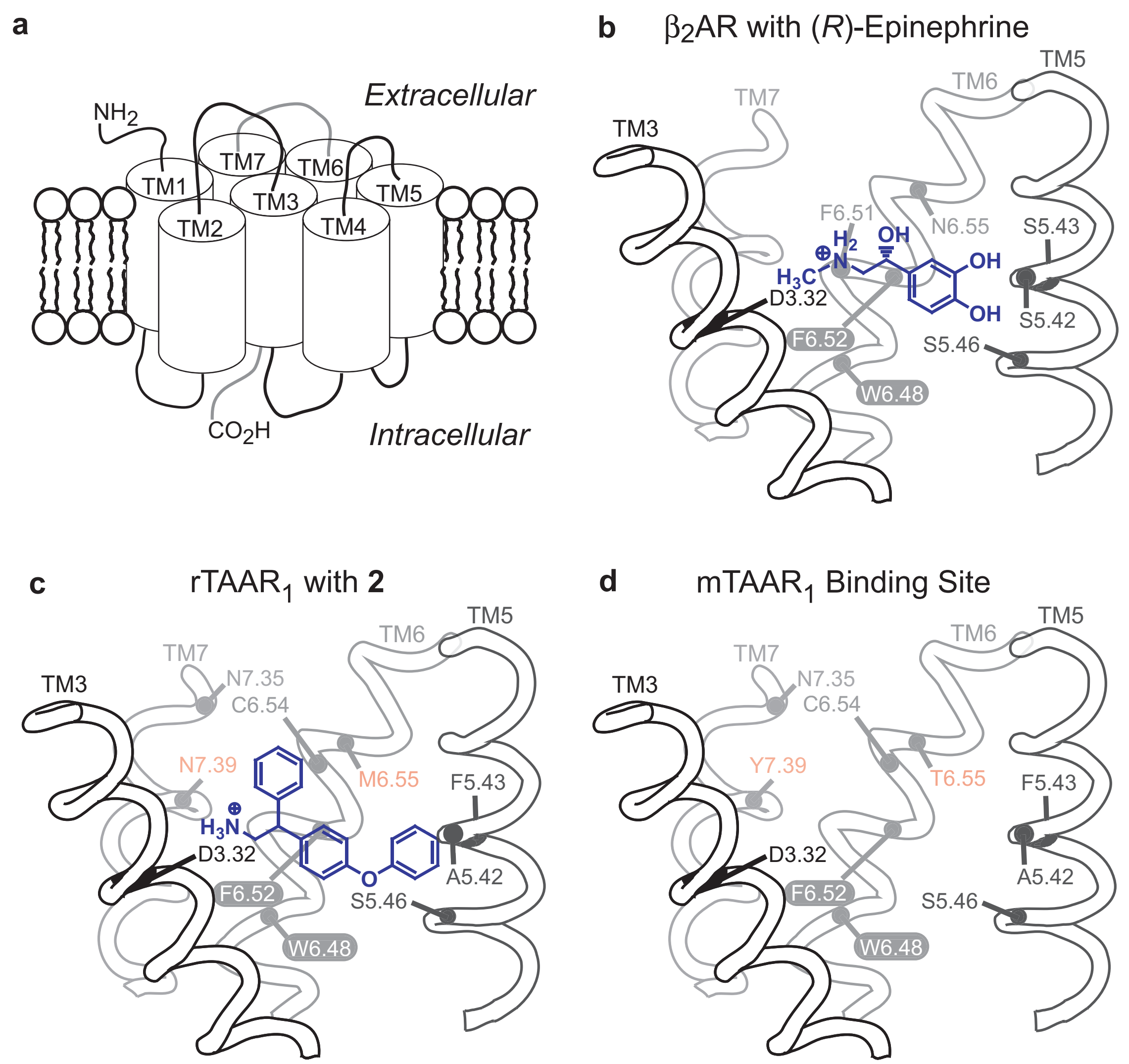
	rTAAR$_1$(F4.56Y)			mTAAR$_1$(Y4.56F)		
	EC$_{50}$^{a}	E_{max}^{b}	N^c	EC$_{50}$^{a}	E_{max}^{b}	N^c
Compd	**± SEM (nM)**	**± SEM (%)**		**± SEM (nM)**	**± SEM (%)**	
1	38 ± 11	100 ± 0	4	67 ± 17	100 ± 0	6
Tyramine	54 ± 11	108 ± 6	4	40 ± 13	118 ± 8	6
4	22 ± 5	101 ± 14	4	100 ± 16	123 ± 7	6
11	27 ± 15	101 ± 6	4	123 ± 41	105 ± 5	6

[a-c] See footnotes for **Table 1**. Compound structures are shown in **Table 1**.



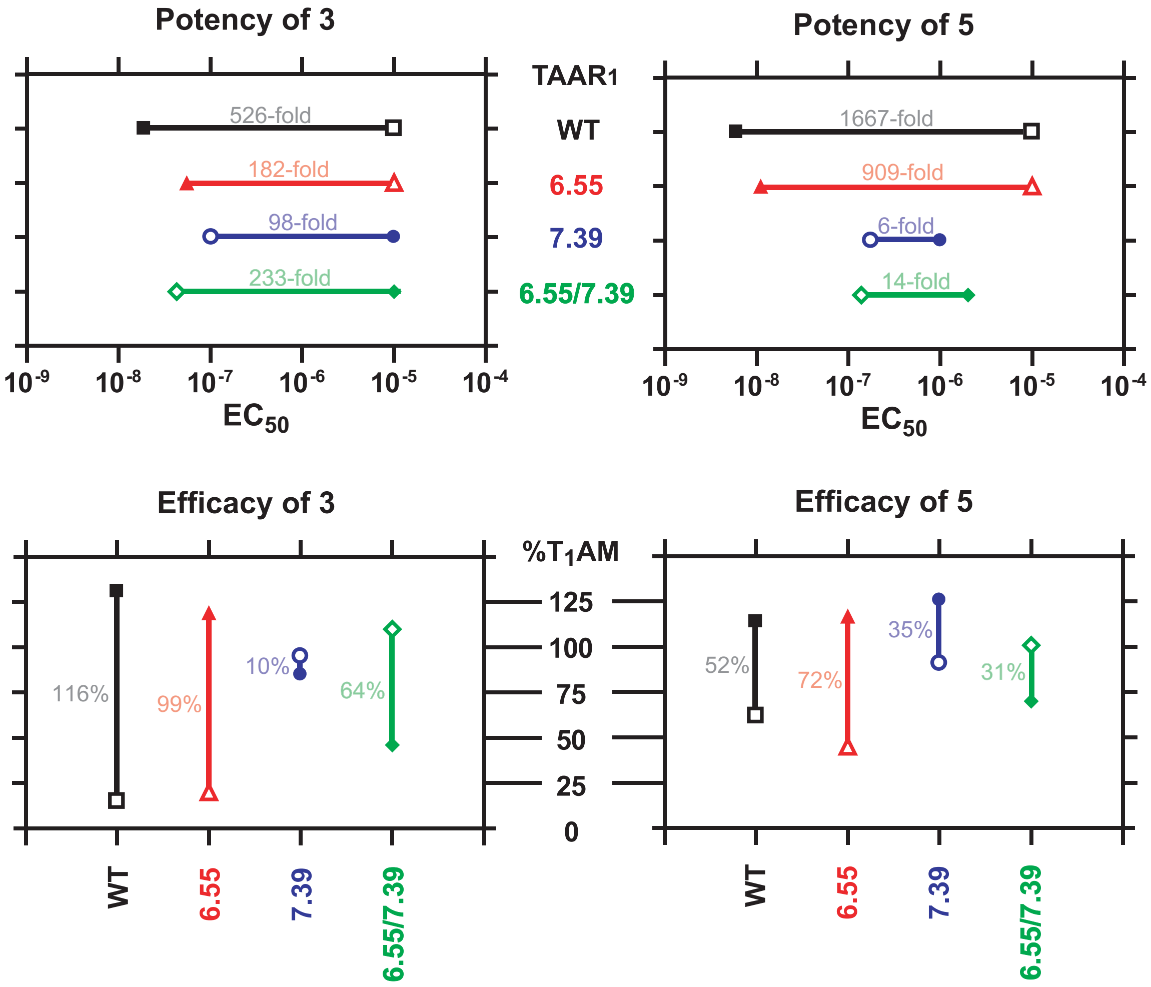
Chapter 4: Figure 1. Thyroid hormone, 3-iodothyronamine, and related analogs.

Structures of thyroxine, 3-iodothyronamine (1), β-phenylphenoxyphenethylamines (2 and 3), and phenoxynaphethylamine (4). The A, B, and C rings of 2 and 3 correspond to the outer, inner, and β-phenyl rings, respectively.



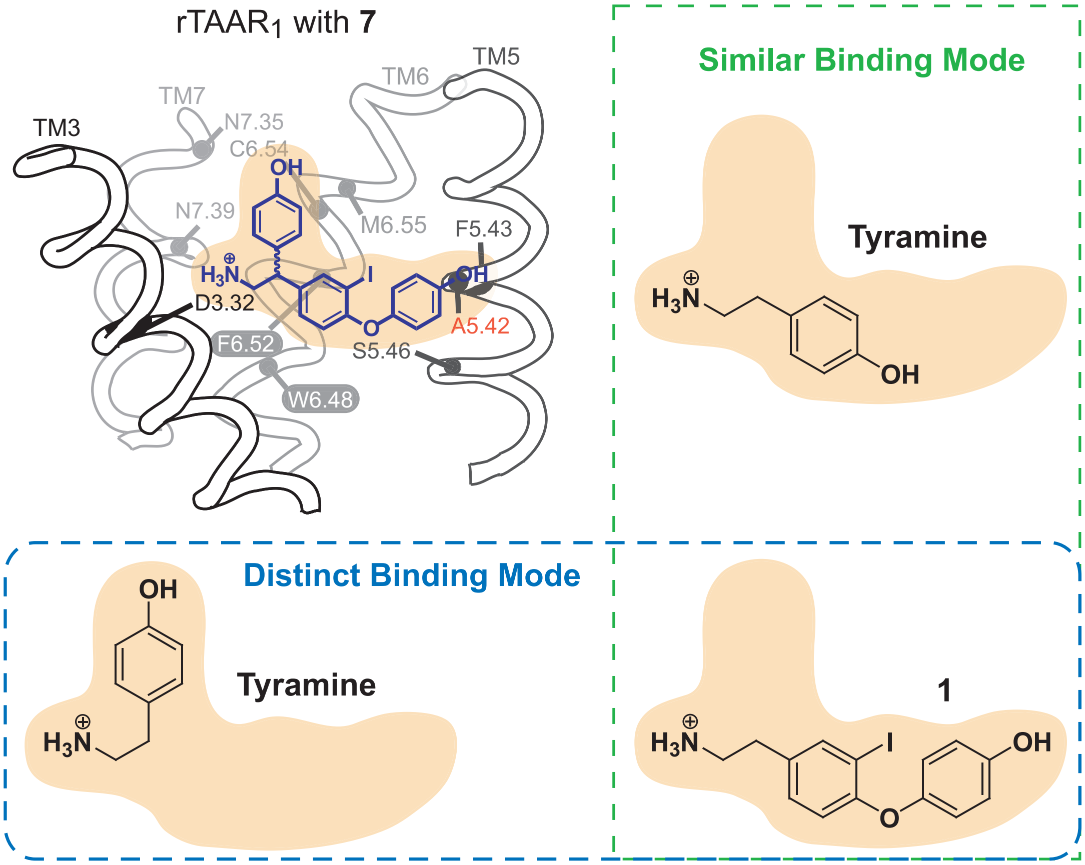
Chapter 4: Figure 2. Biogenic amine GPCR.

(**a**) Schematic representation of the helical arrangement of GPCRs viewed from the cell membrane. (**b**) Binding orientation of (R)-epinephrine in the binding site of the β$_2$AR. (**c**) Proposed binding orientation of **2** in the binding site of rTAAR$_1$. (**d**) Binding site of mTAAR$_1$. The binding sites of β$_2$AR, rTAAR$_1$, and mTAAR$_1$ are viewed from the perspective TM4. The rotamer switch residues (white letters), proposed binding and specificity determinant residues are labeled. Non-conserved residues that were mutated are shown in red.



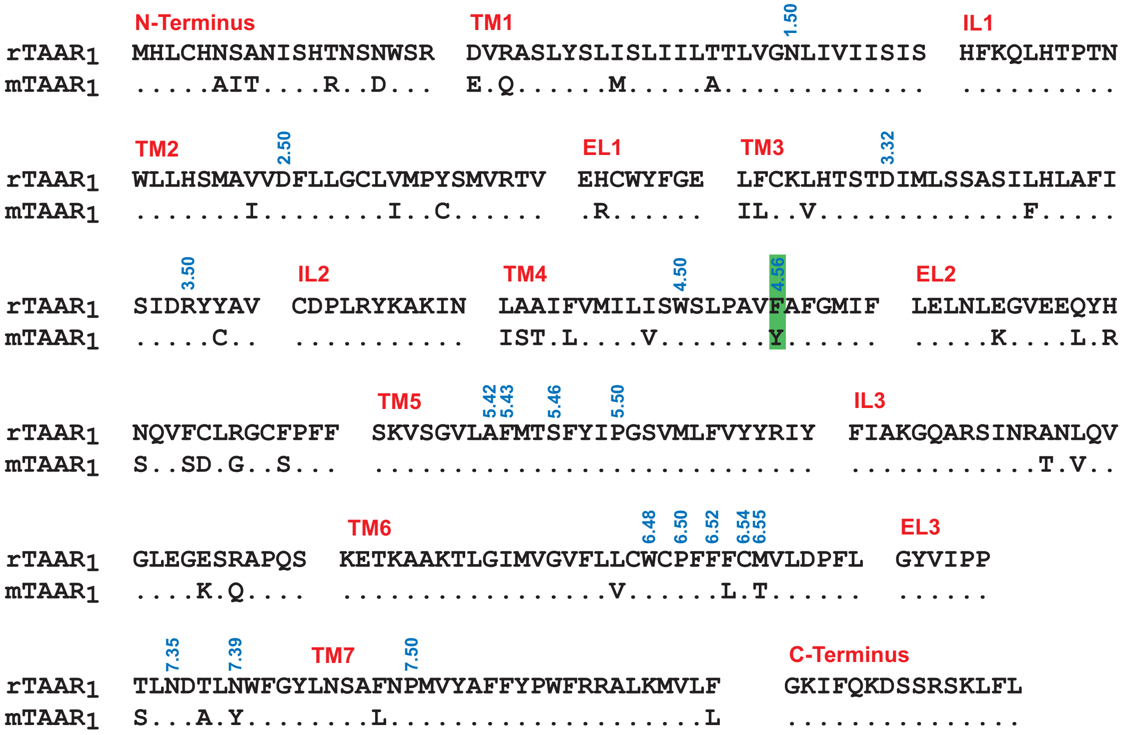
Chapter 4: Figure 3. Activity profiles of **3** and **5** on rat and mouse TAAR$_1$ wild type and mutant receptors.

The top and bottom panels show the potency and efficacy, respectively, of **3** and **5**. The EC$_{50}$ and E$_{max}$ for rTAAR$_1$ and mTAAR$_1$ receptors are depicted in solid and hollow symbols, respectively. WT rTAAR$_1$ [■], rTAAR$_1$(M6.55T) [▲], rTAAR$_1$(N7.39Y) [●], rTAAR$_1$(M6.55T/N7.39Y) [♦], WT mTAAR$_1$ [□], mTAAR$_1$(T6.55M) [Δ], mTAAR$_1$(Y7.39N) [○], and mTAAR$_1$(T6.55M/Y7.39N) [◊]. The lines connecting the solid and hollow symbols represent the difference in potency (top panels) and efficacy (bottom panels) between rat and mouse TAAR1 receptors. The fold difference in potency and percent difference in efficacy are listed above (top panels) and to the left (bottom panels) of the connecting lines, respectively.



Chapter 4: Figure 4. Proposed binding modes of 7, 1, and tyramine in rTAAR$_1$.

The binding site of rTAAR$_1$ is viewed from the perspective of TM4 (see Fig. 2a). The rotamer switch residues (white letters), proposed binding and specificity determinant residues are labeled. Residue A5.42 is shown in red. The similar and distinct binding mode models of 1 and tyramine are outlined in green and blue dashed lines, respectively.



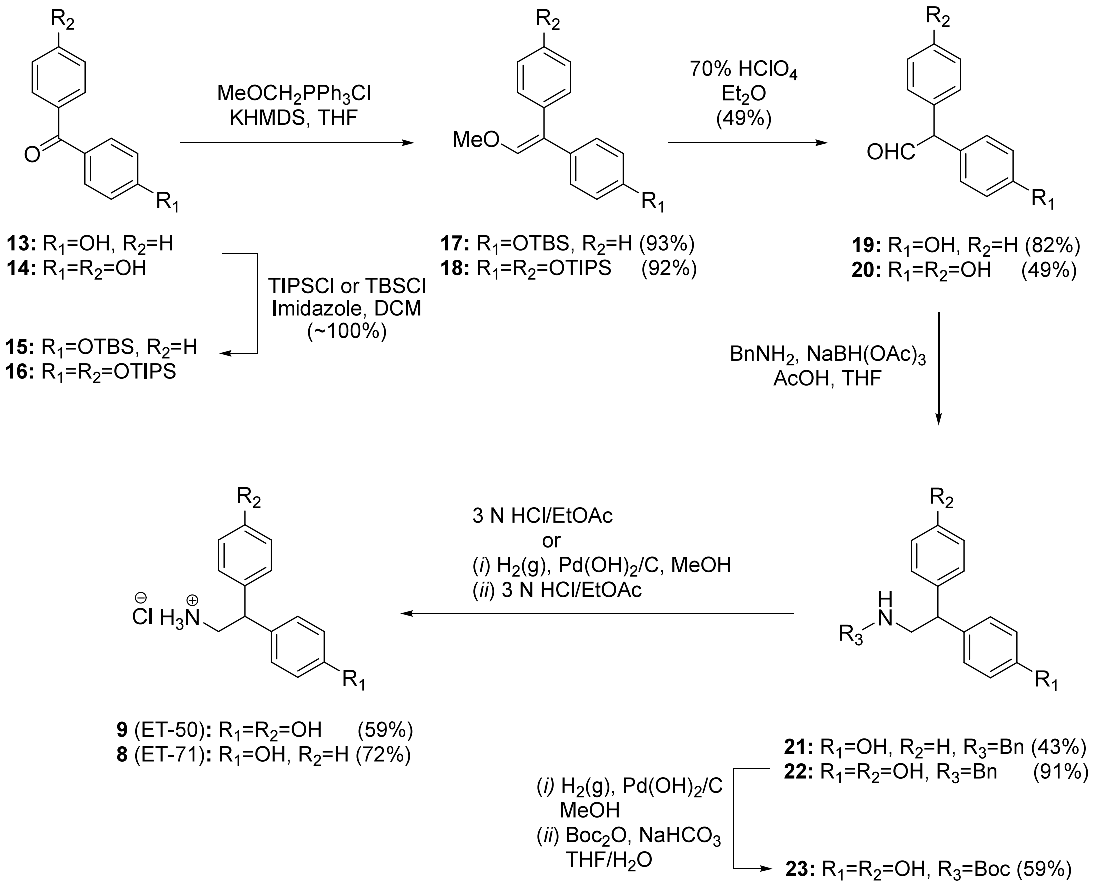
```
         N-Terminus            TM1                    1.50          IL1
rTAAR1   MHLCHNSANISHTNSNWSR  DVRASLYSLISLIILTTLVGNLIVIISIS  HFKQLHTPTN
mTAAR1   .....AIT....R..D...  E.Q......M.....A.............  ..........

         TM2     2.50                    EL1       TM3     3.32
rTAAR1   WLLHSMAVVDFLLGCLVMPYSMVRTV  EHCWYFGE  LFCKLHTSTDIMLSSASILHLAFI
mTAAR1   .......I........I..C......  .R......  IL..V.............F.....

          3.50     IL2          TM4       4.50  4.56      EL2
rTAAR1   SIDRYYAV  CDPLRYKAKIN  LAAIFVMILISWSLPAVFAFGMIF  LELNLEGVEEQYH
mTAAR1   .....C..  ...........  IST.L....V.......Y......  .....K....L.R

                        TM5   5.42 5.43 5.46 5.50           IL3
rTAAR1   NQVFCLRGCFPFF  SKVSGVLAFMTSFYIPGSVMLFVYYRIY  FIAKGQARSINRANLQV
mTAAR1   S..SD.G..S...  ............................  ............T.V..

                    TM6          6.48 6.50 6.52 6.54 6.55     EL3
rTAAR1   GLEGESRAPQS  KETKAAKTLGIMVGVFLLCWCPFFFCMVLDPFL  GYVIPP
mTAAR1   ....K.Q....  .................V......L.T......  ......

          7.35  7.39  TM7  7.50                         C-Terminus
rTAAR1   TLNDTLNWFGYLNSAFNPMVYAFFYPWFRRALKMVLF  GKIFQKDSSRSKLFL
mTAAR1   S...A.Y........L....................L  ...............
```

Chapter 4: Figure 5. Sequence comparison of rat and mouse $TAAR_1$.

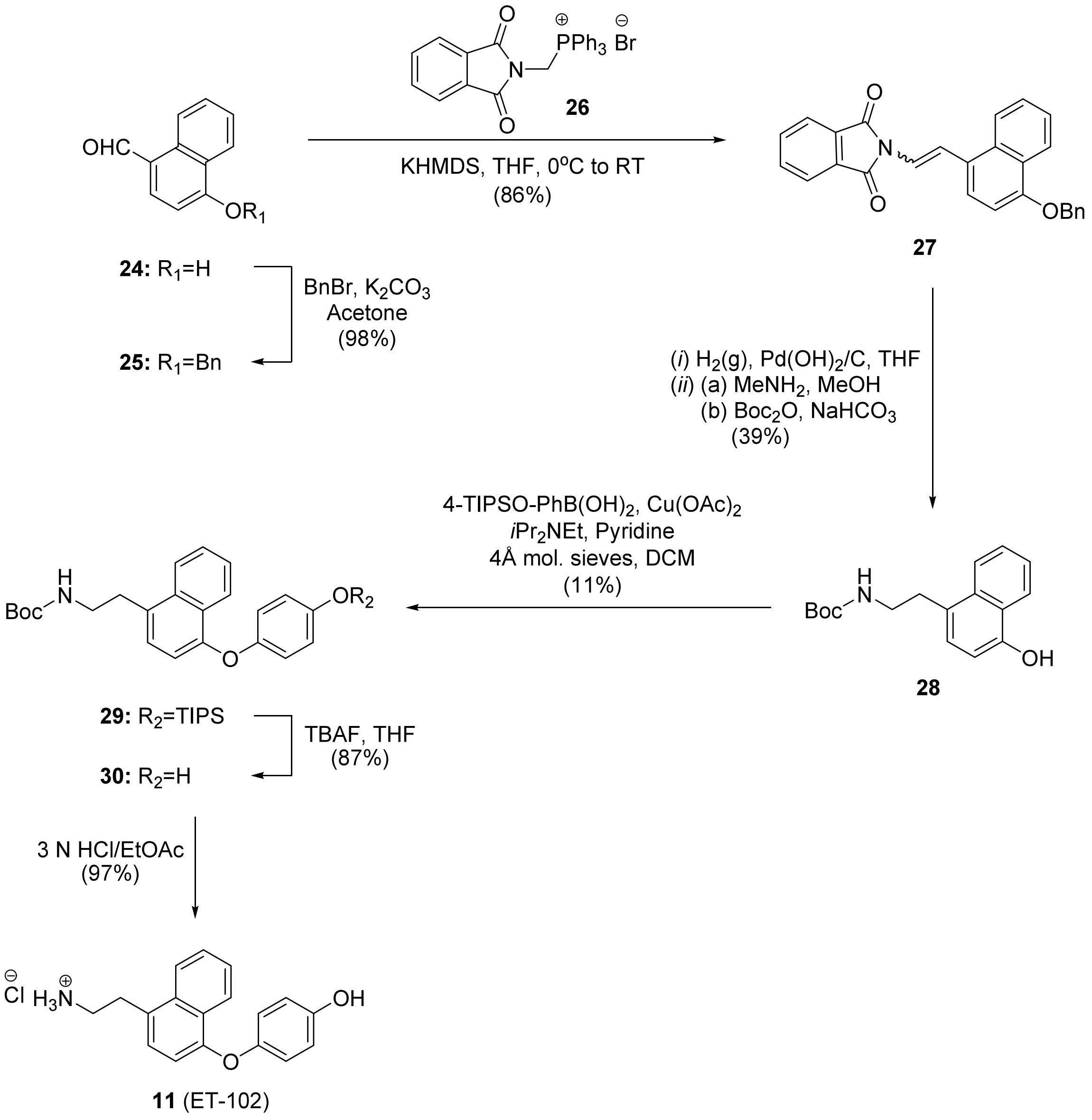
Dots represent conserved residues. Amino and carboxy termini (N-terminus and C-terminus, respectively), intracellular loops (IL), extracellular loops (EL), and transmembrane regions (TM) are labeled. The most conserved residue in each TM region is labeled X.50 and residue 4.56 is highlighted in green.

Supplementary Information

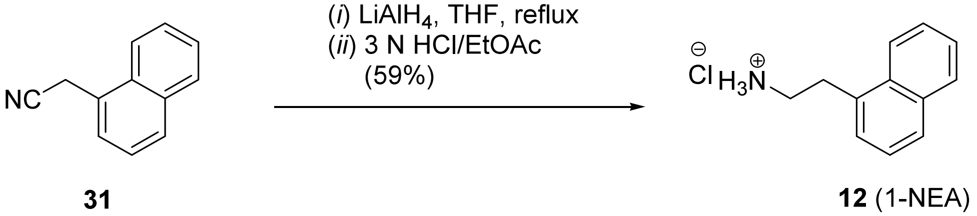
Scheme S1. Synthesis of 8 and 9.



Scheme S2. Synthesis of 11.



Scheme S3. Synthesis of 12.



Supplementary Methods

General. ^{1}H and ^{13}C NMR spectra were taken on a Varian 400 (400 MHz and 100 MHz respectively) or Bruker Avance 400 MHz NMR. Data reported are calibrated to internal TMS (0.0 ppm) for all solvents unless otherwise noted and are reported as follows: chemical shift, multiplicity (app = apparent, br = broad, s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, dd = doublet of doublet, dt = double of triplets, tt = triplet of triplets, dq = doublet of quartets, quin = quintet, sext = sextet, sep = septet), coupling constant, and integration. High resolution mass spectrometry (HRMS) using electrospray ionization was performed by the Mass Spectrometry Laboratory, School of Chemical Sciences, University of Illinois at Urbana-Champaign. Inert atmosphere operations were conducted under argon passed through a drierite drying tube in flame dried or oven dried glassware unless otherwise noted. Anhydrous THF, DCM, diethyl ether, pyridine, and diisopropyl ethyl amine were filtered through two columns of activated basic alumina and transferred under an atmosphere of argon gas in a solvent purification system designed and manufactured by Seca Solvent Systems. Anhydrous DMF was obtained by passing through two columns of activated molecular sieves. All other anhydrous solvents and reagents were purchased from Aldrich, Sigma-Aldrich, Fluka, Alfa Aesar, Acros, Fisher or VWR and were used without any further purification unless otherwise stated. Compounds were purified by flash column chromatography (using Gelduran silica gel: 60Å pore size, 40-64μm particle size, and 7.0±0.5 pH), Biotage SP1™ purification system (FLASH+® silica cartridge: 500m^{2}/g surface area, 60Å pore size, and 40-63μm particle size), or preparatory thin layer chromatography (Analtech prep TLC plates (20x20 cm, 1000μm). During purification, crude products were typically dissolved DCM to improve solubility and facilitate loading onto the silica gel. Final compounds were judged to be >95% pure by ^{1}H NMR analysis.

General Procedure for *t*-Butyldimethylsilyl or Triisopropylsilyl Protection. To a solution of alcohol (4.67 mmol) and imidazole (10.5 mmol) in DCM (50 mL) was added TBSCl or TIPSCl (5.13 mmol). After stirring overnight at room temperature, the reaction was diluted with DCM, washed with water and brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for Formation of a Methyl Enol Ether. To a solution of (methoxymethyl)triphenylphosphonium chloride (27.97 mmol) in THF (80 mL) at 0^{o}C was added KHMDS (27.97 mmol, 0.5M in Toluene) over 15 min. After stirring at 0^{o}C for 30 min, a solution of benzophenone (9.32 mmol) in THF (10 mL) was added over 10 min. The reaction was warmed to room temperature and stirred overnight. The reaction was quenched with water and diluted with ether. The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure. The crude product was triturated with hexanes (or cold Et_2O) and filtered to remove the precipitated triphenylphosphine oxide by products. The crude product was then dissolved in DCM, absorbed to silica, and dry loaded into a SiO_2 column equilibrated with 100% hexanes. The column was then flushed with 100% hexanes (recycling solvents) until all of the triphenylphosphine has eluted.

General Procedure for Methyl Enol Ether Deprotection. To a solution of methyl enol ether (8.58 mmol) in Et_2O (30 mL) was added dropwise 70% aq $HClO_4$ (8.75 mL). After stirring at room temperature for ~2-24 hr, the reaction was quenched with saturated $NaHCO_3$ (91 mL) and diluted with Et_2O. The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for Reductive Amination. To a solution of aldehyde (3.09 mmol) and amine (3.39 mmol) in THF (15 mL) was added $NaBH(OAc)_3$ (4.63 mmol) and acetic acid (3.09 mmol). After stirring at room temperature for 24 h, the reaction was quenched with saturated $NaHCO_3$, made basic with K_2CO_3 (pH ~8-10) and diluted with Et_2O. The organic layer was washed with water and brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for Catalytic Hydrogenation with Pd/C or $Pd(OH)_2/C$. To a solution of benzyl amine, benzyl ether, alkene, or alkyne (1.57 mmol) in MeOH (36 mL) was added Pd/C or $Pd(OH)_2/C$ (0.24 g). After purging all of the air in the reaction by exposure to vacuum, the reaction was stirred overnight at room temperature under $H_2(g)$ atmosphere (1 atm). The Pd/C or $Pd(OH)_2/C$ was filtered through celite and rinsed with MeOH or EtOAc. The filtrate was concentrated under reduced pressure.

General Procedure for *t*-Boc Deprotection. The *t*-Boc protected amine (1.40 mmol) was dissolved in a 3N anhydrous HCl solution in EtOAc (3 mL), and the reaction mixture was stirred at room temperature for 2-16 h. The reaction was exposed to Et_2O and the resulting amine hydrochloride salts were washed with Et_2O. If the amine hydrochloride salts did not form a precipitate, the Et_2O/EtOAc solution was concentrated under reduced pressure and triturated with Et_2O or hexanes to give the hydrochloride salts.

General Procedure for Phthalimide Deprotection and Subsequent *t*-Boc Protection. To a solution of phthalimide (9.08 mmol) in MeOH (30 mL) was added $MeNH_2$ (181.54 mmol, 40% in H_2O). After refluxing for 3 h, the reaction was concentrated under reduced pressure and placed under high vacuum to remove all the $MeNH_2$. To a solution of the crude product (9.03 mmol) in THF (30 mL) was added an aqueous solution of $NaHCO_3$ (9.08 mmol in 15 mL of water) followed by addition of di-*tert*-butyl dicarbonate (9.08 mmol). After stirring at room temperature overnight, the reaction was quenched with water and diluted with Et_2O. The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for Formation of a Biaryl Ether. To a solution of phenol (0.53 g, 1.73 mmol), phenyl boronic acid (0.43 g, 3.47 mmol), copper (II) acetate (0.31 g, 1.73 mmol) and dried 4Å molecular sieves (2.64 g, 5 equiv by starting material weight, flame dried under high vacuum for ~5 min) in DCM (20 mL) was added pyridine (0.70 mL, 8.66 mmol) and Hunig's base (1.50 mL, 8.66 mmol). After stirring under dry air atmosphere at room temperature for 24 h or longer, the reaction was filtered through celite and silica gel, rinsed with EtOAc and washed with 0.5 M HCl, water and brine. The organic layer was dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for *t*-Butyldimethylsilyl or Triisopropylsilyl Deprotection. To a stirred solution of *t*-butyldimethylsilyl protected phenol (155mg, 0.41 mmol) in THF (4 mL) was added dropwise TBAF (0.45 mL, 1M in THF, 0.45 mmol). The reaction mixture was stirred for 15 min at room temperature and diluted with Et_2O. The mixture was washed with water, brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product.

General Procedure for Reduction of a Nitrile to an Amine Hydrochloride. To a suspension of lithium aluminum hydride (26.7 mmol) in THF (56 mL) at 0^oC, was added a solution of nitrile (6.66 mmol) in THF (10 mL). After refluxing under argon for 24 h, the reaction was quenched with water (1.014 mL), 10% aqueous NaOH (2.028 mL) and water (3.043 mL). The reaction was filtered to remove the precipitated aluminum salts. The filtrate was washed with water and brine and extracted with EtOAc. The organic layer was dried over $MgSO_4$, filtered, and concentrated under reduced pressure to give the crude product. The crude mixture was treated with a 3N anhydrous HCl solution in EtOAc (5-10 mL), exposed to Et_2O, and the resulting amine hydrochloride salts were washed with Et_2O. If the amine hydrochloride salts did not form a precipitate, the Et_2O/EtOAc solution was concentrated under reduced pressure and rinsed with Et_2O to give the hydrochloride salts.

Scheme S1. Synthesis of 8 and 9

(4-(*tert*-Butyldimethylsiloxy)phenyl)-(phenyl)methanone (15). Refer to general procedure for *t*-butyldimethylsilyl or triisopropylsilyl protection. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (5%/95%)) to give **15** as a colorless oil (1.59 g, ~100% yield). ^1H NMR (400 MHz, chloroform-*d*) δ 7.77 (d, *J*=8.4 Hz, 2H), 7.76 (dd, *J*=8.4 Hz, 1.3 Hz, 2H), 7.56 (tt, *J*=8.2 Hz, 7.4 Hz, 1.2 Hz, 1H), 7.47 (t, *J*=7.4 Hz, 2H), 6.90 (dt, *J*=8.6 Hz, 2.8 Hz, 1.9 Hz, 2H), 1.00 (s, 9H), 0.25 (s, 6H).

Bis(4-Triisopropylsiloxyphenyl)methanone (16). Refer to general procedure for *t*-butyldimethylsilyl or triisopropylsilyl protection. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (5%/95%)) to give **16** as a colorless oil (4.91 g, ~100% yield). ^1H NMR (400 MHz, chloroform-*d*) δ 7.72 (dt, *J*=9.0 Hz, 2.7 Hz, 2.2 Hz, 4H), 6.93 (dt, *J*=8.8 Hz, 2.7 Hz, 2.2 Hz, 4H), 1.30 (m, 6H), 1.12 (d, *J*=7.3 Hz, 36H).

2-Methoxy-((1-(4-(tert-butyldimethylsiloxy)phenyl))-1-(phenyl))ethene (17). Refer to general procedure for formation of a methyl enol ether. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (0%/100%) to (2%/98%)) to give **17** as a yellow oil (1.61 g, 93% yield). ^1H NMR (400 MHz, chloroform-*d*) δ 7.37 (dd, *J*=8.2 Hz, 1.5 Hz, 1H), 7.20-7.33 (m, 5H), 7.06 (dt, *J*=8.6 Hz, 2.9 Hz, 2.1 Hz, 1H), 6.78 (dt, *J*=8.6 Hz, 2.9 Hz, 2.0 Hz, 1H), 6.75 (dt, *J*=8.4 Hz, 2.9 Hz, 2.0 Hz, 1H), 6.37; 6.39 (s, 1H), 3.76; 3.74 (s, 3H), 0.98 (s, 9H), 0.20 (s, 6H).

2-Methoxy-(1,1-bis(4-Triisopropylsiloxyphenyl))ethene (18). Refer to general procedure for formation of a methyl enol ether. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (0%/100%) to (2%/98%)) to give **18** as a yellow oil (4.76 g, 92% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 7.24 (d, *J*=9.2 Hz, 2H), 7.05 (dt, *J*=8.6 Hz, 2.9 Hz, 2.1 Hz, 2H), 6.79 (dd, *J*=8.8 Hz, 7.0 Hz, 4H), 6.29 (s, 1H), 3.73 (s, 3H), 1.26 (sep, *J*=7.4 Hz, 6H), 1.10 (d, *J*=7.1 Hz, 36H).

2-(4-Hydroxyphenyl)-2-(phenyl)acetaldehyde (19). Refer to general procedure for methyl enol ether deprotection. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (25%/75%)) to give **19** as a yellow oil (0.66 g, 82% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 9.91 (d, *J*=2.6 Hz, 1H), 7.38 (tt, *J*=7.3 Hz, 7.2 Hz, 1.6 Hz, 2H), 7.31 (tt, *J*=8.3 Hz, 7.3 Hz, 1.4 Hz, 1H), 7.21 (dd, *J*=7.1 Hz, 1.7 Hz, 2H), 7.09 (dt, *J*=8.2 Hz, 3.1 Hz, 1.9 Hz, 2H), 6.84 (dt, *J*=8.8 Hz, 2.9 Hz, 2.2 Hz, 2H), 4.91 (s, 1H), 4.83 (t, *J*=2.6 Hz, 1H),

2,2-Bis(4-hydroxy)phenylacetaldehyde (20). Refer to general procedure for methyl enol ether deprotection. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (10%/90%) to (30%/70%)) to give **20** as a reddish yellow oil (0.95 g, 49% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 9.87 (d, *J*=2.4 Hz, 1H), 7.07 (dt, *J*=8.6 Hz, 2.9 Hz, 1.9 Hz, 4H), 6.84 (dt, *J*=8.6 Hz, 2.9 Hz, 2.1 Hz, 4H), 6.72 (s, 1H), 4.76 (d, *J*=2.4 Hz, 1H), 4.73 (s, 1H).

N-Benzyl-2-(4-hydroxyphenyl)-2-phenylethylamine (21). Refer to general procedure for reductive amination. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (30%/70%) to (50%/50%)) to give **21** as a brownish foam (0.40 g, 43% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 7.17-7.31 (m, 10H), 7.03 (d, *J*=8.6 Hz, 2H), 6.68 (d, *J*=8.6 Hz, 2H), 4.16 (t, *J*=7.8 Hz, 1H), 3.81 (s, 2H), 3.19 (d, *J*=7.7 Hz, 2H).

N-Benzyl-2,2-di-(4-hydroxy)phenylethylamine (22). Refer to general procedure for reductive amination. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (30%/70%) to (50%/50%)) to give **22** as a brownish foam (1.21 g, 91% yield). ^{1}H NMR (400 MHz, methanol-*d_4*) δ 7.19-7.32 (m, 5H), 7.00 (dt, *J*=8.4 Hz, 2.9 Hz, 1.8 Hz, 4H), 6.70 (dt, *J*=8.6 Hz, 3.0 Hz, 2.1 Hz, 4H), 4.00 (t, *J*=7.8 Hz, 1H), 3.74 (s, 2H), 3.07 (d, *J*=8.1 Hz, 2H).

N-*t*-Boc-2,2-di-(4-Hydroxy)phenylethylamine (23). Refer to general procedure for catalytic hydrogenation with Pd/C or $Pd(OH)_2$. The crude product was *t*-Boc protected following the general procedure for *t*-Boc protection of free amine. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (30%/70%) to (40%/60%)) to give **4.52** as a clear oil/white foam (0.33 g, 59% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 7.03 (dt, *J*=8.3 Hz, 2.9 Hz, 1.7 Hz, 4H), 6.76 (dt, *J*=8.6 Hz, 2.8 Hz, 2.1 Hz, 4H), 5.27 (br s, 2H), 4.52 (br s, 1H), 4.00 (t, *J*=7.8 Hz, 1H), 3.66 (t, *J*=6.5 Hz, 2H), 1.41 (s, 9H).

2-(4-Hydroxyphenyl)-2-phenylethylamine Hydrochloride (8, ET-71). Refer to general procedure for catalytic hydrogenation with Pd/C or $Pd(OH)_2$. The crude product treated with a 3N anhydrous HCl solution in EtOAc (1 mL), exposed to Et_2O, and the resulting amine hydrochloride salts were washed with Et_2O. Compound **8** was obtained as a brownish powder (0.24 g, 72% yield). ^{1}H NMR (400

MHz, methanol-*d*$_4$) δ 7.27-7.34 (m, 4H), 7.22 (tt, *J*=7.5 Hz, 6.9 Hz, 1.7 Hz, 1H), 7.12 (dt, *J*=8.6 Hz, 3.0 Hz, 2.1 Hz, 2H), 6.75 (dt, *J*=8.6 Hz, 3.0 Hz, 2.0 Hz, 2H), 4.17 (t, *J*=8.2 Hz, 1H), 3.53 (d, *J*=8.2 Hz, 2H). HRMS (EI$^+$) *m/z* for $C_{14}H_{16}NO$ [M + H]$^+$: calcd, 214.1232; found, 214.1227.

2,2-Di-(4-hydroxy)phenylethylamine Hydrochloride (9, ET-50). Refer to general procedure for *t*-Boc deprotection: white solid, 0.15 g, 59% yield. ^{1}H NMR (400 MHz, methanol-*d*$_4$) δ 7.12 (dt, *J*=8.4 Hz, 3.1 Hz, 2.0 Hz, 4H), 6.77 (dt, *J*=8.6 Hz, 2.9 Hz, 2.0 Hz, 4H), 4.08 (t, *J*=8.3 Hz, 1H), 3.49 (d, *J*=8.2 Hz, 2H). HRMS (EI$^+$) *m/z* for $C_{14}H_{16}NO_2$ [M + H]$^+$: calcd, 230.1181; found, 320.1176.

Scheme S2. Synthesis of 11

4-Benzyloxy-naphthaldehyde (25). To a solution of 4-hydroxynaphthaldehyde (3.96 g, 23.02 mmol) in acetone (60 mL) was added K_2CO_3 (3.50 g, 25.32 mmol) and BnBr (5.91 g, 34.53 mmol). After stirring at room temperature for 24 h, the reaction was concentrated under reduced pressure and diluted with water and Et_2O. The organic layer was washed with brine, dried over $MgSO_4$, filtered, and concentrated under reduced pressure. The crude product was purified via Biotage® purification system (EtOAc/Hexanes (0%/100%) to (25%/75%)) to give **25** as a yellow oil (5.90 g, 98% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 10.21 (s, 1H), 9.31 (d, *J*=8.6 Hz, 1H), 8.41 (dd, *J*=8.3 Hz, 0.8 Hz, 1H), 7.91 (d, *J*=8.1 Hz, 1H), 7.71 (dd, *J*=6.8 Hz, 1.5 Hz, 1H), 7.58 (dd, *J*=6.3 Hz, 1.3 Hz, 1H), 7.52 (d, *J*=7.3 Hz, 2H), 7.44 (t, *J*=7.2 Hz, 2H), 7.39 (tt, *J*=8.6 Hz, 7.2 Hz, 1.4 Hz, 1H), 7.00 (d, *J*=8.1 Hz, 1H), 5.35 (s, 2H).

(1,3-dioxo-1,3-dihydro-isoindol-2-ylmethyl)-triphenyl phosphonium bromide (26). To a solution of N-(bromomethyl)phtalimide (24.49 g, 102.02 mmol) in toluene (500 mL) was added PPh_3 (26.76 g, 102.02 mmol). After stirring under reflux for 24 h, the white precipitate that formed were filtered, washed with hexanes, and dried under vacuum.

2-(Isoindolinyl-1,3-dione)-(1-(4-benzyloxy)naphthyl)ethene (27). To a solution of (1,3-dioxo-1,3-dihydro-isoindol-2-ylmethyl)-triphenyl phosphonium bromide (20.79 g, 41.39 mmol) in THF (376 mL) at 0^oC was added KHMDS (82.78 mL, 41.39 mmol, 0.5M in toluene) over 10 min. After stirring at 0^oC for 10 min, a solution of **25** (4.52 g, 17.25 mmol) in THF (10 mL) was added and the reaction was warmed to room temperature. After 24 h, the reaction was quenched with water and diluted with EtOAc. The organic layer was washed with brine and saturated NH_4Cl, dried over $MgSO_4$, filtered, and concentrated under reduced pressure. The crude product was purified via Biotage® purification system (DCM/Hexanes (50%/50%) to (80%/20%)) to give **27** as an orange solid (6.01 g, 86% yield). ^1H NMR (400 MHz, chloroform-*d*) δ 8.40 (dq, *J*=8.3 Hz, 0.7 Hz, 1H), 8.33 (d, *J*=14.9 Hz, 1H), 8.14 (d, *J*=8.6 Hz, 1H), 7.92 (dd, *J*=5.6 Hz, 3.0 Hz, 2H), 7.78 (dd, *J*=5.6 Hz, 3.0 Hz, 2H), 7.50-7.61 (m, 5H), 7.44 (t, *J*=7.3 Hz, 2H), 7.37 (tt, *J*=8.3 Hz, 7.3 Hz, 1.4 Hz, 1H), 7.27 (d, *J*=14.9 Hz, 1H), 6.92 (d, *J*=8.1 Hz, 1H), 5.29 (s, 2H).

N-*t*-Boc-2-(4-hydroxy-naphthyl)ethylamine (28). Refer to general procedure for phthalimide deprotection and subsequent *t*-Boc protection. The crude product was purified via Biotage® purification system (EtOAc/Hexanes (0%/100%) to

(40%/60%)) to give **28** as a brownish white solid (1.65 g, 31% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 8.24 (d, *J*=8.3 Hz, 1H), 8.01 (d, *J*=8.1 Hz, 1H), 7.55 (dd, *J*=6.8 Hz, 1.6 Hz, 1H), 7.50 (dd, *J*=6.8 Hz, 1.4 Hz, 1H), 7.14 (d, *J*=7.3 Hz, 1H), 6.76 (d, *J*=7.6 Hz, 1H), 5.53 (s, 1H), 4.60 (br s, 1H), 3.46 (q, *J*=6.6 Hz, 2H), 3.19 (t, *J*=6.7 Hz, 2H), 1.45 (s, 9H).

N-*t*-Boc-2-(4-(4-triisopropylsiloxy)phenoxynaphthyl)ethylamine (29). Refer to general procedure to formation of a biaryl ether. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (5%/95%) to (10%/90%)) to give **29** as a colorless oil (0.11 g, 11% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 8.35 (d, *J*=8.3 Hz, 1H), 8.06 (d, *J*=8.1 Hz, 1H), 7.58 (dd, *J*=6.7 Hz, 1.3 Hz, 1H), 7.51 (dd, *J*=6.9 Hz, 1.4 Hz, 1H), 7.17 (d, *J*=7.7 Hz, 1H), 6.95 (dd, *J*=6.6 Hz, 2.5 Hz, 2H), 6.88 (dd, *J*=6.5 Hz, 2.4 Hz, 2H), 6.70 (d, *J*=7.9 Hz, 1H), 4.60 (br s, 1H), 3.48 (q, *J*=6.5 Hz, 2H), 3.22 (t, *J*=7.0 Hz, 2H), 1.44 (s, 9H), 1.25 (m, 3H), 1.11 (d, *J*=7.0 Hz, 18H).

N-*t*-Boc-2-(4-(4-hydroxy)phenoxynaphthyl)ethylamine (30). Refer to general procedure for *t*-butyldimethylsilyl or triisopropylsilyl deprotection. The crude product was purified via flash SiO_2 chromatography (EtOAc/Hexanes (20%/80%)) to give **30** as a colorless oil/white foam (66 mg, 87% yield). ^{1}H NMR (400 MHz, chloroform-*d*) δ 8.34 (d, *J*=8.3 Hz, 1H), 8.06 (d, *J*=7.8 Hz, 1H), 7.58 (dd, *J*=7.1 Hz, 1.4 Hz, 1H), 7.52 (t, *J*=7.1 Hz, 1H), 7.17 (d, *J*=7.8 Hz, 1H), 6.97 (d, *J*=8.8 Hz, 2H), 6.84 (d, *J*=8.8 Hz, 2H), 6.72 (d, *J*=7.3 Hz, 1H), 5.07 (br s, 1H), 4.63 (br s, 1H), 3.48 (q, *J*=6.6 Hz, 2H), 3.22 (t, *J*=6.8 Hz, 2H), 1.44 (s, 9H).

2-(4-(4-Hydroxy)phenoxynaphthyl)ethylamine Hydrochloride (11, ET-102). Refer to general procedure for *t*-Boc deprotection: white solid, 66 mg, 97% yield. ^{1}H NMR (400 MHz, methanol-*d$_4$*) δ 8.36 (dd, *J*=8.3 Hz, 0.8 Hz, 1H), 8.08 (d, *J*=8.3 Hz, 1H), 7.65 (dd, *J*=6.8 Hz, 1.4 Hz, 1H), 7.56 (dd, *J*=6.8 Hz, 1.1 Hz, 1H), 7.28 (d, *J*=7.8 Hz, 1H), 6.93 (d, *J*=8.8 Hz, 2H), 6.82 (d, *J*=9.1 Hz, 2H), 6.68 (d, *J*=7.8 Hz, 1H), 3.39 (d, *J*=7.1 Hz, 1H), 3.37 (d, *J*=6.1 Hz, 1H), 3.26 (d, *J*=8.3 Hz, 1H), 3.24 (dd, *J*=10.2 Hz, 1.1 Hz, 1H). HRMS (EI$^+$) *m*/*z* for $C_{18}H_{18}NO_2$ [M + H]$^+$: calcd, 280.1338; found, 280.1332.

Scheme S3. Synthesis of 12

2-(Naphthyl)ethylamine Hydrochloride (12, 1-NEA). Refer to general procedure for reduction of a nitrile to an amine hydrochloride: yellowish brown solid, 0.36 g, 59% yield. ^{1}H NMR (400 MHz, methanol-*d$_4$*) δ 8.09 (d, *J*=8.4 Hz, 1H), 7.90 (dd, *J*=8.2 Hz, 0.6 Hz, 1H), 7.82 (dd, *J*=6.0 Hz, 3.3 Hz, 1H), 7.58 (dd, *J*=6.8 Hz, 1.4 Hz, 1H), 7.51 (dd, *J*=6.8 Hz, 1.1 Hz, 1H), 7.45 (d, *J*=9.5 Hz, 1H), 7.44 (d, *J*=6.2 Hz, 1H), 3.46 (t, *J*=8.0 Hz, 2H), 3.27 (m, 2H). HRMS (EI$^+$) *m*/*z* for $C_{12}H_{14}N$ [M + H]$^+$: calcd, 172.1126; found, 172.1127.

Chapter 5: Performance Characteristics for Sensors and Circuits Used to Program E.coli

Jeffrey J. Tabor[1], Eli Groban[1,2] and Christopher A. Voigt[1,2]

Department of Pharmaceutical Chemistry, University of California

San Francisco.

[1]Department of Pharmaceutical Chemistry, University of California-San Francisco,

600 16th Street, Suite

518, Box 2280, San Francisco, CA 94143-2280, USA

[2]Biophysics Program, San Francisco, CA, USASan Francisco, CA 94158.

Summary

The behavior of *E.coli* can be reprogrammed by the introduction of foreign segments of DNA. Three classes of genetic parts, termed sensors, circuits and actuators comprise the DNA programs. Sensors are gene products which allow the cell to detect physical or chemical information in its environment. Genetic engineers can use sensors directly from nature, modify them in some manner, or design them *de novo* to control cellular processes with extracellular or intracellular signals. Genetic circuits act to process information from sensors in order to dictate the behavior of the cell. They can be designed with combinations of "off the shelf" regulatory parts such as transcription factors and promoters, or in some cases can be used "as is" from nature. Finally, genetic circuits govern the expression of actuators, genes whose products perform some physical function to alter the state or the environment within which the cell exists. Using recent DNA synthesis and assembly technologies, genetic sensors, circuits and actuators can be combined to create programs that command cells to perform a series of tasks. This approach will transform the way that genetic engineers approach problems in biotechnology. This review covers the construction of genetic sensors and circuits for use in *E.coli*, as well as genetic methods to perturb their performance features.

I. Introduction

To program novel behaviors into *E.coli*, handfuls of genetic parts, or segments of DNA with defined functions, are introduced into the cell. In the background, thousands of regulatory and metabolic reactions operate simultaneously and in direct physical contact with the heterologous parts. The engineered components can operate as largely insulated modules or can be functionally integrated with the preexisting networks of the host cell. Despite what would appear to be long odds, surprisingly complex behaviors often with medical, industrial or academic relevance can be achieved.

In this chapter, we will discuss some of the principles which guide the programming of *E.coli*. We define biological programs as strings of genetic parts encoded on segments of DNA which are introduced to the cell on plasmid vectors or integrated into the genome. The designed DNA fragments carry three classes of parts which we will refer to as *sensors*, *circuits* and *actuators* (Voigt 2006). Each of these functions is encoded on a piece of DNA. When combined they create a genetic program that provides a set of instruction that the cell can read and execute. Though the sensor/circuit/actuator construction paradigm can be applied to program any number of genetically tractable organisms (Drubin et al. 2007; Greber and Fussenegger 2007; Sia et al. 2007), this chapter will be limited to a discussion of *E.coli* where much of the foundational work has been accomplished.

Sensors are defined as biological molecules which receive physical or chemical *inputs* from the external or intracellular environment and transfer this information as molecular *output* functions. Sensors can be derived from nature and used “as is”, redesigned for novel input-output functions, or designed *de novo*. This chapter reviews the three most commonly occurring sensors in *E.coli*: cytoplasmic ligand-binding transcription factors, two-component signaling systems and riboregulators.

Sensors transmit information to genetic circuits. *Genetic circuits* are groups of regulatory molecules which control gene expression to program the cellular response to sensory inputs. Genetic circuits are ubiquitous in the genomes of natural organisms and the characterization of their input-output ranges and dynamic and steady-state responses, or *performance features*, can inform the construction of synthetic analogs with defined properties. In some cases the entire DNA segment encoding a natural circuit can be used “out of the box”, or as found in nature, while simply being connected to user defined sensors and actuators. Synthetic genetic circuits are built by designing a piece of DNA which carries a s.eries of regulatory parts which interact in a defined manner.

Genetic circuits drive *actuators* which act to change the state or behavior of the host cell or its environment. Actuators range from simple reporters like Green Fluorescent Protein (GFP) to entire organelles. The programming of reliable and sophisticated behaviors in *E.coli* will require actuator expression and function to be tightly governed by environmental, physiological or metabolic signals which are transmitted through genetic circuits via sensors.

Programs written from sensors circuits and actuators can coordinate sophisticated multistep behaviors with applications in biotechnology (Figure 1). This type of integrated bioprocessing includes, for example, sensing, integrating and responding media conditions or cell growth stages or densities within a fermenter for optimized yields of an industrially relevant natural product.

Historically limited to piecemeal stitching of naturally occurring DNA fragments, modern DNA synthesis and assembly methods allow the arbitrary connection of sensors, circuits and actuators. Very large (genome scale) biological programs can now be written *in silico* and constructed commercially (Endy 2008; Gibson et al. 2008). The reprogramming of genomes will enable streamlining of the cell through the wholesale addition, deletion or modification of regulatory and metabolic pathways. This will in turn increase the stability, efficiency and productivity (Posfai et al. 2006) of engineered cellular processes.

II. Sensors

Genetic sensors typically receive information from the extracellular environment or internal cell state, which then transmitted to gene regulatory networks. Environmental sensing in *E.coli* largely comprises three strategies: classical regulation, two-component sensing and riboregulation. We will discuss some of the best studied and most widely engineered examples of these sensors throughout this section.

Sensors can receive myriad physical and chemical inputs including small or macromolecules, pH, temperature, light and even signals from other cells. This chapter will focus only on small molecule signals which are the most widely used inputs for engineering *E.coli*.

II.A. Classical regulation

Classical regulation is the control of promoter activity by ligand binding proteins (Figure 2a) (Jacob and Monod 1965). The sensor is a cytoplasmic transcription factor which receives an environmental signal by directly binding to a small molecule ligand. Ligand binding triggers a conformational rearrangement which results in increased or decreased affinity of the transcription factor for cognate DNA operator sequences. The sensory output can be transmitted in two ways, by activation or the relief of repression. Activation typically occurs by transcription factor-mediated recruitment of the RNA polymerase complex at the promoter while repression occurs by its occlusion (Wagner, 2000).

Classical transcription factors are the most widely used sensors for programming *E.coli*. This is due to the simplicity of their components, their rapid output (strong transcriptional responses occur on the order of 1 minute (Guzman et al. 1995)), the ease with which their input and output specificities can be re-engineered, and the availability of their inducer compounds (Wagner 2000). Here, common strategies are outlined for re-engineering the specifities and performance features of classically regulated transcription factors. Throughout this section we will focus on a particularly well elaborated example, the tetracycline responsive TetR protein.

Re-engineering Classically Regulated Sensors

The steady-state quantitative relationships between the concentration of input signal and output gene expression, or transfer functions (Weiss et al. 1999; Yokobayashi et al. 2002; Canton et al. 2008), have been characterized for many classically regulated systems. The features of transfer functions arise from the rate of occupation of promoters by transcription factors and RNA polymerases at different input concentrations (Bintu et al. 2005b). The transfer function of a circuit can be measured by linking it to a sensor, varying the amount of input and measuring the output with a reporter gene (Figure 3). Transfer functions are useful in the design of cellular behaviors because they define the minimal and maximal amount of sensory input which generate circuit responses, the magnitude of induction at any given input concentration and the sensitivity of the circuit to input (Bintu et al. 2005a).

The dynamic range of induction, or magnitude of output in the fully activated (ON) state divided by that of the inactive (OFF) state, is a critical feature of any sensor. In many cases, a large dynamic range of induction is desirable because it more clearly differentiates the absence and presence of an environmental input. Increased dynamic ranges can be achieved by increasing the transcription rate of the ON state, decreasing the transcription rate of the OFF state, or both. The ON state can most easily be increased by strengthening the -35 and -10 RNA polymerase recognition sequences while the repressed state can be lowered by changing the configuration of operator sites around the promoter (Lutz and Bujard 1997; Lutz et al. 2001b; Cox et al. 2007). The sensitivity, or rate of increase in transcriptional output as a function of ligand concentration is largely proportional to the cooperativity of binding of the transcription factor at the promoter. We will discuss strategies for programming cooperativity in Section III.

Increasing Dynamic Range

The dynamic range of classical transcription factor systems can be increased by changing the architecture of the output promoter. Traditionally this been accomplished by the addition, deletion or reorganization of the operator sites (de Boer et al. 1983; Brosius et al. 1985; Guzman et al. 1995; Lutz and Bujard 1997). In this section we will discuss efforts specific to the TetR protein.

TetR has been used as the basis for engineering a more tightly repressed and strongly inducible sensing system. To accomplish this two high affinity operator sites were added to an otherwise strong promoter. TetR was then constitutively expressed to repress then promoter in the absence of the input ligand. The system showed virtually no expression in the OFF state, was sensitive to very low levels of input and showed a ~5000-fold dynamic range of induction (Lutz and Bujard 1997). The performance features of the re-engineered system were all marked improvements over the naturally occurring version from which it was derived (Kleckner et al. 1978; de la Torre et al. 1984) and as a result it has become one of the most widely used sensors for programming *E.coli.*

Changing Operator Specificity

Novel transcription factor:promoter pairs can also be derived from natural systems. The introduction of two mutations within the TetR operator sequence can reduce the affinity to levels insufficient for *in vivo* repression. Rational redesign of DNA binding domains or directed evolution can then be used to re-establish the affinity of the transcription factor for the mutant operators. Indeed, such methods have generated novel transcription factor:promoter pairs based on the TetR (Helbl and Hillen 1998; Helbl et al. 1998) and LacR and lambda Cro (Backes et al. 1997) systems as well. Importantly, these novel specificities can be generated with very small numbers of amino acid substitutions in the transcription factors, allowing the rapid generation and screening of many new orthogonal sets in the cellular context. Similar strategies are likely to be amenable to virtually any classically regulated promoter system in *E.coli.*

Changing the input ligand

The input specificities of classically regulated systems can also be reprogrammed. This is typically accomplished by randomly mutating amino acid residues around the ligand binding pocket and screening variants in functional assays *in vivo* (Collins et al. 2005; Collins et al. 2006; Hawkins et al. 2007). We will discuss efforts to reprogram the ligand specificity of the TetR protein in this section.

TetR has been evolved to recognize an alternate ligand with strong preference over the natural ligand (Henssler et al. 2004). Importantly, the novel inducer is not recognized by the wild type TetR protein, a feature which gives rise to two orthogonal input sensors. The combination of novel input and output specificities has the potential to generate completely orthogonal sensing systems which can be used in parallel with one another. Indeed, two TetR variants which sensed different ligands and activated different promoters were recently introduced in the same *E.coli* cell to control the expression of two separate genes (Kamionka et al. 2004). This work demonstrates the value of classically regulated sensing systems as a platform for the construction of genetic control elements with broad applications in biological design.

II.B. Two-component sensing

A common strategy for environmental sensing in bacteria is a process known as two-component sensing. The canonical two-component system consists of a membrane-bound sensor protein that receives an environmental signal at an extracellular sensory domain and passes the information to a cytoplasmic response regulator protein (Figure 2b). This occurs via the transfer of a phosphate moiety from the cytoplasmic kinase domain of the sensor protein to the receiver domain of the response regulator protein, which can then bind to DNA operator sites at a DNA binding domain to activate or repress gene expression (Hoch and Silhavy 1995).

These sensors are slower to respond than their classically regulated counterparts. For example, the well studied EnvZ-OmpR system of *E.coli* reaches half maximal response to the presence of an input signal in about 5 minutes but requires much longer (on the order of 1 hour) to reach steady-state (Batchelor and Goulian 2006). This happens despite the fact that the phosphotransfer event occurs on a seconds time scale at most (Laub et al. 2007).

The re-engineering of two-component systems has been aided by the modularity of the protein structure. Modular systems are those that are composed of multiple interchangeable subcomponents, or modules. In two-component systems, the extracellular sensory domain of the sensor kinase protein can be replaced by the sensor module from a similar protein. Likewise, the kinase domain of a given sensor protein can be swapped with another to change its specificity for a response regulator (Figure 4). Similar to the classically regulated systems, the specificity of the sensor kinase for its input signal can be altered by computational design methods.

Domain swapping

Sensors can easily be rewired to new outputs by domain swapping. This involves fusing non-cognate sensor and kinase domains at a splice site in a linker region. Most two-component engineering efforts to date have been based on domain swapping, a design process by which chimeric proteins are built from the subdomains of two or more pre-existing proteins (Figure 4). This type of engineering allows the sensing pathway to be rewired such that, for example, the output promoter will respond to a completely different input ligand.

The early discovery of a convenient module boundary (Utsumi et al. 1989) made the osmo-responsive EnvZ/OmpR two-component system of *E.coli* a favorite target for many engineering efforts (Baumgartner et al. 1994; Looger et al. 2003; Levskaya et al. 2005). In the natural configuration, the sensor kinase EnvZ phosphorylates the response regulator OmpR in response to changes in osmolarity. Phosphorylated OmpR then binds to operator sites at a promoter, activating or repressing gene expression (Aiba et al., 1989; Aiba and Mizuno 1990; Forst et al., 1989). In the pioneering domain swapping effort, Inoyue and co-workers fused the cytoplasmic domain of EnvZ with the sensory domain of the transmembrane aspartate receptor (TAR), thus rewiring the EnvZ/OmpR pathway to be activated by the amino acid aspartate (Utsumi et al. 1989).

The sensory domain of the chemoreceptor protein Trg has similarly been fused to the cytoplasmic domain of EnvZ (Baumgartner et al. 1994). The Trg sensory domain interacts with periplasmic sugar binding proteins only when they are bound to their ligands to direct *E.coli* chemotaxis. The hybrid Trg-EnvZ protein allowed control of the EnvZ/OmpR pathway with the unnatural ligand ribose via the ribose binding protein (RBP) (Baumgartner et al. 1994).

The sensory domain of other sensor kinases have also been used to control a chemotactic signaling. NarX is a histidine kinase which senses nitrate and nitrite (Williams and Stewart 1997). Replacement of the sensory domain of the TAR protein with the sensory domain of the NarX kinase has programmed E.coli to chemotax away from extracellular nitrate and nitrite (Ward et al. 2002).

In 2005, the osmosensing domain of EnvZ was replaced with a light sensing domain from the *Synechocystis* phytochrome protein Cph1 to program *E.coli* to respond to light (Levskaya et al. 2005). This also required the introduction of a two gene metabolic pathway to produce the chromophore PCB, which binds to the engineered sensor kinase (Gambetta and Lagarias 2001). A confluent lawn of the engineered *E.coli* could then be used as a high resolution film capable of directly converting a two-dimensional light input pattern into a pigment output pattern.

Redesigning ligand specificities

Other efforts have used computational methods to redesign of other periplasmic sugar binding protiens to sense ligands as varied as trinitrotoluene (TNT), L-Lactate, (Looger et al. 2003) and Zn^{2+} (Dwyer et al. 2003) for control of gene expression through the Trg-EnvZ-OmpR pathway. Unlike domain swapping strategies, these studies required detailed knowledge of the three-dimensional structure of the parental proteins. The structural information then guided the authors to consider between 5 and 17 amino acids residues as candidates for mutation, and the computational searches typically yielded small lists of candidate protein sequences which were directly amenable to experimental evaluation.

Designing the Histidine Kinase-Response Regulator Interface

There are at least 32 natural two-component systems in *E.coli* all of which have similar structures at the sensor/response regulator interface (Ulrich et al. 2005). To maintain the fidelity of signal transmission through any one of these pathways the sensors and response regulators have evolved a great deal of pairwise molecular specificity (Skerker et al. 2005). Knowledge of the specificity determinants of the histidine kinase-response regulator interactions could allow rewiring of input-output relationships.

Bioinformatic algorithms have been used to elucidate regions of the histidine kinase proteins responsible for response regulator specificity. This information enabled the rewiring of two-component pathways by mutating sensor:response regulator interaction domains. The substitution of as few as three amino acid resides within a cytoplasmic subdomain of EnvZ reprogrammed its specificity away from OmpR to numerous other response regulators (Skerker et al. 2008). The ability to redesign protein-protein interfaces adds a valuable degree of freedom which will greatly increase the number of possible alternative two-component signaling pathways that can be constructed in *E.coli.*

II.C. Riboregulators

RNA molecules can sense inputs, often through interactions with small or macromolecular ligands, and transmit the information to control gene expression. This typically occurs via the formation of a ligand binding pocket within the regulatory RNA (riboregulator) which triggers an overall change in its secondary structure. These structural rearrangements can hide or liberate regulatory domains which can then modulate gene expression *in cis* or *in trans* (on the same or another gene). To date, 16 *E.coli* genes have been shown to be subject to cis-acting regulation by ligand binding riboregulators termed riboswitches (Barrick and Breaker 2007).

Bacterial riboswitch sensors convert ligand binding into a change in the transcription or translation rate of the mRNA within which they are embedded (Winkler and Breaker 2003). Though not as widely utilized as their protein counterparts, the structural and functional simplicity of RNA makes it a very attractive platform for the engineering of sensing in bacteria (Isaacs et al. 2006). This is because secondary structure, which governs much of the overall shape and function of RNA, can be computationally predicted with good accuracy (Mathews et al. 1999) and can be experimentally verified much more rapidly than can protein structures (Soukup and Breaker 1999c). This enables realistic *in silico* design of riboregulators *de novo*, a monumentally difficult task in the protein world.

Reprogramming

Riboregulation is also compelling because simple base pairing rules and robust directed evolution methods allow the construction of many orthogonal regulators based on a single parent structure (Tang and Breaker 1997; Koizumi et al. 1999; Soukup and Breaker 1999a; Soukup and Breaker 1999b; Jose et al. 2001; Soukup et al. 2001; Isaacs et al. 2004; Bayer and Smolke 2005). The modular structure of riboregulators also allows them to be introduced into many different genes and even ported between vastly different organisms with surprising ease (Yen et al. 2004). Moreover, unlike in two-component engineering the sensory domains of riboregulators need not bear any structural or evolutionary relationship to the regulatory domains to which they are fused (Soukup and Breaker 1999b; Jose et al. 2001; Buskirk et al. 2004; Bayer and Smolke 2005).

As a concise demonstration of the design advantages of riboregulators, a riboswitch was recently designed *de novo* to reprogram *E.coli* chemotaxis (Topp and Gallivan, 2007). In this work an antisense RNA domain was engineered to base pair with and occlude a ribosome binding site (RBS) upstream of the open reading frame of a chemotaxis-dependent gene, inhibiting translation and subsequently chemotaxis (Figure 2c). An ligand binding (aptamer) domain for the small molecule theophylline was included within the riboregulator such that when theophylline was present, a local base pairing rearrangement occurred which liberated the ribosome binding site, allowing translation. In this way, the engineered riboswitch guided *E.coli* to swim up a gradient of a chemical that does not normally function as an attractant. Though domain swapping and directed evolution have enabled the rewiring of chemotaxis at the protein level as well (Ward et al. 2002; Derr et al. 2006), the benefits of riboregulation are manifest in this example as high throughput efforts have allowed rapid increases in the dynamic range of induction of the riboswitch in response to ligand (Lynch et al. 2007; Topp and Gallivan 2008).

II.D. Cell-Cell Communication

Cells also have the ability to sense the presence of other cells in the environment. In bacteria this often occurs through a process known as quorum sensing (Miller and Bassler 2001). In short, cells produce membrane-diffusible signals which diffuse into other cells and function as ligands for classical transcription factors. This type of sensing can drive coordinated decision making in cell communities, which enables more sophisticated behaviors.

Cell-cell communication sensors have been used in *E.coli* to control the density of a bacterial population (You et al. 2004), coordinate the timing and magnitude of gene expression between two different cell types (Brenner et al. 2007), drive multicellular pattern formation (Basu et al. 2004; Basu et al. 2005), coordinate the invasion of a malignant mammalian cell (Anderson et al. 2006) or even create a synthetic ecosystem (Balagadde et al. 2008). Each of these circuits was constructed from the Lux-type quorum sensing circuit of *V. fischeri*. A full review of the engineering applications of this type of cell-cell communication system is reviewed elsewhere (Salis et al. 2009 (in press))

III. Circuits: processing sensory information

Genetic circuits, or networks of interacting regulatory molecules can integrate one or more sensory inputs into logical and dynamic genetic outputs (Hasty et al. 2002; Kaern et al. 2003; Wall et al. 2004). Circuits have previously been constructed in *E.coli* which generate memory (Gardner et al. 2000; Atkinson et al. 2003), oscillations (Elowitz and Leibler 2000; Atkinson et al. 2003) or pulses (Basu et al. 2004) of gene expression. Other circuits have been designed to function as logic gates, capable of integrating information from multiple sensors to produce a single output (Guet et al. 2002; Yokobayashi et al. 2002; Anderson et al. 2007). Genetic circuits can also drive cell-cell communication and community-level decision making (You et al. 2004; Basu et al. 2005; Brenner et al. 2007; Balagadde et al. 2008) This section provides an overview of the performance features and engineering considerations for some of the best characterized and most useful genetic circuit motifs.

III.A. Classical Regulation

The simplest genetic circuits are the classical ligand-inducible transcription systems described in Section II.A. In these simple circuits, the presence of input signal positively influences the transcription of an output gene. The transfer function of classically regulated circuits is important because it describes the level of gene expression out of the circuit in response to a given concentration of input signal. This is important when linking multiple circuits in series, because if the output of one circuit is not quantitatively matched with the input of another, then information transfer through the system breaks down. It is of particular interest to discuss the performance features of classically regulated circuits here as they constitute the foundation of many more complex circuit designs.

III.A.1. Simple promoters

In classically regulated circuits the output abundance typically varies as a positive sigmoidal function of the input concentration (Bintu et al. 2005a) (Figure 5A). This relationship arises because there are two input ranges where the system is non-responsive and one input range under which it is. At low input levels, well below the K_D of the transcription factor for the ligand, there is virtually no change in output. As the input ligand concentration approaches the K_D of the transcription factor, there is a monotonic increase in output protein abundance proportional to input.

The sensitivity (Wall et al. 2004), or slope of the response curve, in this range is largely determined by the cooperativity of binding of the transcription factor at the promoter of interest. Cooperativity refers to an effect where the affinity of a transcription factor for its DNA operator site increases as a consequence of a previous binding event by another transcription factor at a nearby operator (Ptashne and Gann 2002; Bintu et al. 2005a). This is often the result of protein-protein interaction domains which drive multimerization of the transcription factors on the DNA. Finally, as ligand concentrations increase well above the relevant K_D, the pool of transcription factors or relevant DNA operators become saturated and the output does not increase with further increases in input (Figure 5A).

Certain features of the transfer function can be altered by changing the number and type of operator sites near the output promoter in a classically regulated system. For example the sensitivity, or log-log slope of the input-output function in the responsive range, is less than or equal to 1 for promoters with a single operator site. This is true whether the system is regulated by an activator or repressor (Bintu et al. 2005a; Bintu et al. 2005b). The addition of a second operator that enables cooperative binding can significantly increase sensitivity, typically ~2-4 fold (Bintu et al. 2005a; Bintu et al. 2005b). DNA looping can also be used to increase the sensitivity of the response (Vilar and Leibler 2003).

In activator systems, if binding is not cooperative, the sensitivity of the response remains the same with the introduction of a second operator, but the dynamic range of induction increases multiplicatively. In repressor based systems, additional operators which do not elicit cooperative binding can still increase the sensitivity because the presence of a repressor at any the first site can significantly occlude the RNA polymerase, inherently facilitating binding of a repressor at the second site (Bintu et al. 2005a).

Continuous response

Classically regulated transcriptional systems have the property of continuous responsivity. Continuous response means that the abundance of the output gene product in a single cell scales proportionally to the concentration of input signal in the environment. This allows the 'fine-tuning' of output expression levels across an entire population. The fine-tuning of expression also allows the control of protein variance, which has been shown to naturally decrease as protein abundance increase (Bar-Even et al. 2006). As we will discuss in Section III.B.2, many natural genetic circuits lack this property and some have even been intentionally modified to acquire it.

Speed of Response

An important performance feature of any circuit is the rise time, or time required after the addition of an input for the output to reach 50% of its steady-state value. This value, which has been measured for several systems in *E.coli* is approximately 1 cell cycle (45-135 minutes in these studies) (Rosenfeld et al. 2002; Mangan et al. 2006). The time required for an *E.coli* cell to fully respond to an environmental stimulus via the classical mode of regulation is therefore greater than the time required for it to produce a complete copy of itself. The response time of classically regulated circuits can be increased by adding protease tags (Andersen et al. 1998) to speed degradation of regulatory proteins. The slow response times of classically regulated circuits will be compared with those of more complex circuits below.

III.A.2. Complex and Biphasic Promoters

Promoters bearing multiple operator sites, which activate and repress gene expression can result in non-monotonic behavior in response to a monotonic increase in input signal. For example, the P_{RM} promoter of phage λ has three operator sites for the transcription factor CI. CI initially binds at two high affinity sites and has an activating effect on promoter output. When CI reaches higher concentrations, however, it binds to a low affinity site and functions as a repressor. A circuit wherein CI is expressed proportionally to an input can therefore result in an output which is OFF at both low and high input and ON only at intermediate inputs (Michalowski et al. 2004).

The operator for the AraC activator has been added to the *E.coli* lac promoter to generate a two-tiered activation response (Lutz and Bujard 1997). In this design, transcription increases proportional to the concentration of the first input IPTG but saturates at an intermediate level. This response is solely a function of promoter derepression. When provided saturating IPTG, the promoter can then undergo a second tier of activation proportional to the concentration of the activator arabinose. This occurs as a result of AraC mediated recruitment of RNA polymerase at the derepressed promoter. Many mutants of this promoter have also been constructed which offer different performance features as well (Lutz et al. 2001a).

III.A.3. Regulatory Cascades

Multiple classically regulated circuits can be linked in series such that the output of one circuit serves as the input to the next (Figure 5B). Cascades can be used to temporally order the expression of many different output genes in response to a single input stimulus (Kalir et al. 2001), allow cells to respond to increasingly smaller amounts of input (Hooshangi et al. 2005) and filter out transient or noisy input signals.

There are several inherent trade-offs in the use of regulatory cascades. For example, inducer sensitivity and signal amplification can be increased with the number of regulatory steps, but this occurs at a cost to response time. Moreover, lengthening can oftentimes require the redesign of upstream elements to ensure that the output ranges of the existing segment are matched to the input ranges over which the downstream segment can respond (Yokobayashi et al. 2002; Basu et al. 2004; Hooshangi et al. 2005).

Signal amplification and Ultrasensitivity

To directly measure the performance features of genetic cascades, Weiss and coworkers constructed several synthetic genetic circuits that systematically increased the length of a cascade. This included circuits with 1, 2, and 3 repressors connected in series (Figure 5B). As repressors were added to the cascade, the authors observed that the output reached half-maximal response at increasingly lower inducer concentrations, about 40% lower inducer per repressor added. Signal amplification allows cells to respond to inputs which are present in the environment at concentrations below the limit of detection of the natural sensory apparatus.

As with other circuit designs that we have discussed, regulatory cascades can increase sensitivity to the input (Hooshangi et al. 2005; Pedraza and van Oudenaarden 2005) (Figure 5B). In the Weiss example, the range of inducer concentrations required to generate a full response decreased approximately 5-fold upon the addition of the second repressor and 8-fold upon addition of the third. Moreover, a mathematical model indicated that sensitivity would continue to increase as more than three repressors were added to the chain (Hooshangi et al. 2005).

Activation delays

The relaying of an input signal through a multi-step regulatory cascade results in a temporal lag in response (Figure 5B). Whereas a single repressor showed near immediate response and reached a steady-state output at two hours, the two repressor system took greater than six hours to reach steady-state (Hooshangi et al. 2005). The addition of the third repressor delayed signal transmission dramatically. This circuit showed no response to inputs at times less than two hours, and took 10 hours to reach steady-state. Furthermore, a mathematical model showed that with every two additional repressors added the rise time would continue to increase two-fold.

Cascade-Mediated Control of Complex Cellular Processes

The expression of many genes can be temporally ordered if they are regulated by protein regulatory cascades. The *E.coli* flagellum is encoded within 14 operons which contain its structural and regulatory genes. Upon induction, each operon is induced in an order commensurate with the sequence of assembly of the proteins which make up the flagellar apparatus (Kalir et al. 2001). The regulators in this cascade are able to activate each of their target operons in sequence with minutes long lag times in between. This highly regulated sequence of events is probably encoded at the DNA level by variable operator sequences at each promoter for which the regulators have slightly different binding affinities (Kalir et al. 2001). In this scenario, free floating cytoplasmic regulator proteins will occupy stronger operator sites before occupying any given lower affinity operator, allowing rank ordering of gene expression.

Quantitative measurements of gene expression in this system allowed the development of a rigorous computational model which could then be used to make predictive perturbations to circuit behavior (Kalir and Alon 2004). Similarly detailed measurements of the regulatory interactions and their effect on gene expression will be invaluable in the troubleshooting, manipulation and optimization of forward engineered systems as well. Though synthetic biology is far from reliably designing structures as complex as the flagellum, one can envision many smaller scale applications where cascades could be used to time orders of expression in complex processes. For example, timed protein expression could facilitate the stepwise biosynthesis of novel antibiotics (Pfeifer et al. 2001), boost drug production (Keasling 2008) convert agricultural waste into fuel (Service 2007) or even coordinate the expression of complex cellular machines (Temme et al. 2008).

III.B. Feedback and Feed Forward Regulation

Linking the output of a classically regulated circuit back to its input or forward through intermediate regulators can dramatically alter its dynamic and steady-state properties. In this section we review the most common natural and engineered feedback and feed forward circuits, focusing on the impact of overall architecture and key parameters on circuit performance.

III.B.1. Negative feedback

Negative feedback occurs when the output of a given circuit represses its own production (Figure 6). Circuits controlled by negative feedback have unique response characteristics which are critical for certain biological design applications. Though negative feedback can be implemented as an inhibitory step at any point between production and decay of a gene product this section focuses on transcriptional feedback, which has been widely employed in the construction of synthetic circuits.

Response accelerators

The response times of negative feedback circuits are markedly reduced compared to their analogous classically regulated counterparts (Savageau 1974). Using engineered variants of the *tet* system, Alon and coworkers experimentally demonstrated a reduction in rise time from over two hours to 15 minutes upon the introduction of negative feedback (Rosenfeld et al. 2002) (Figure 6). The acceleration of the response is proportional to the strength of repression a parameter which can be engineered by altering the number, strength or location of operator sites (Basu et al. 2004; Cox et al. 2007). Acceleration also increases with the cooperativity of binding of the repressor protein to the promoter (Savageau 1974; Rosenfeld et al. 2002). This term can be also be changed by the addition or removal of operator sites (Bintu et al. 2005a) or by the selection of repressor proteins with different oligomerization properties (Ninfa and Mayo 2004).

Though the negative feedback component reduces response time, it also reduces the absolute steady-state output of a circuit (Rosenfeld et al. 2002; Bashor et al. 2008). The rise time acceleration in negative feedback circuits occurs because shortly after induction the promoter is unrepressed. Only after the accumulation of repressor does the activity of the promoter decrease to steady-state. This is in contrast to the classically regulated promoter which is active at a high level at all times after induction, resulting in a higher steady-state output which takes more time to achieve. The negative feedback circuit architecture is only useful, therefore, if the circuit output is responsive to reduced steady-state expression levels.

Noise buffering

Stochastic fluctuations, or noise, in gene expression is inevitable in genetic circuits and can reduce the fidelity of signal transmission and cellular behavior (McAdams and Arkin 1997). Moreover, as the number of components in an engineered circuit increases, the effects of noise in any one component can be compounded (Hooshangi et al. 2005; Pedraza and van Oudenaarden 2005).

It has been recognized that negative feedback circuit architecture can reduce noise in output gene expression (Savageau 1974; El-Samad and Khammash 2006). To experimentally validate this effect, Becskei and Serrano constructed a synthetic circuit wherein a repressor protein inhibited its own transcription in *E.coli* (Becskei and Serrano 2000). Negative feedback reduced noise, measured as the coefficient of variation in protein expression across a population of cells, up to 70% over a circuit without feedback. Moreover, and as predicted (Savageau 1974) the magnitude of noise buffering was proportional to the strength of feedback.

The reason that negative feedback circuits buffer fluctuations is intuitive. In classically regulated transcriptional systems, fluctuations in any step of protein expression (transcription, mRNA decay, translation, etc.) are amplified by subsequent steps and cause variation in protein abundances between individual cells. In negative feedback circuits, fluctuations that cause increases in the output protein concentration are quickly dampened by increased repression while fluctuations that cause the output levels to decrease reduce repressor abundances and increase transcription rates. The result is that the system returns to steady state more rapidly after random fluctuations.

There is a caveat to the use of negative feedback as a safeguard against noise in engineered circuits. Though noise decreases proportional to feedback strength over a large range of protein abundances (Becskei and Serrano 2000; Thattai and van Oudenaarden 2001), noise can actually increase if the strength of negative feedback becomes too strong (Shahrezaei et al. 2008). This is due to a phenomenon known as the 'small number effect' where the impact of intrinsic fluctuations in chemical reactions increases rapidly as the concentration of reactants becomes very small (Kaern et al. 2005; Bar-Even et al. 2006). That is, at smaller protein concentrations each random protein production or decay event has a larger impact on the mean concentration. This highlights the general biological design principle that increasing the number of proteins in a cell reduces noise in protein abundance (Bar-Even et al. 2006).

III.B.2. Positive Feedback

Positive feedback occurs when the output of a circuit activates its own production (Figure 7A). Circuits with positive feedback can have many features which are valuable in the engineering of more robust, decisive cellular behaviors including ultrasensitivity, bistability, hysteresis and memory. This section describes the performance features of positive feedback loops and how they can be changed by modifying the underlying circuit parameters.

Response Delays

In contrast to negative feedback circuits which accelerate response times, positive feedback circuits have are thought to slow the rise to steady-state. Though it has not been measured in a well-controlled experimental setting, the magnitude of the rise time delay is predicted to be proportional to the strength of the positive feedback step (Savageau 1974). For a transcriptional circuit, feedback strength is governed by the binding affinity of the output transcription factor for its DNA operators, the mode by which the transcription factor interacts with RNA polymerase and its cooperativity (Ninfa and Mayo 2004; Bintu et al. 2005a).

The most direct strategy for manipulating the magnitude of delay in a positive feedback circuit is to vary the DNA operator sites at the promoter to which the activator binds. This can be done by varying the number, spacing and sequence of the operators. Single nucleotide mutations within operator sites can significantly reduce the affinity of a transcription factor for its operator (Takeda et al. 1989; Frank et al. 1997; Falcon and Matthews 2000; Basu et al. 2004). Increasing or decreasing the spacing between multiple operators can affect both binding affinity (Chen and Kadner 2000) and cooperativity of binding (Smith and Sauer 1995).

The introduction of positive feedback increases the steady-state output level of a classical transcriptional circuit. To compensate, one can decrease the production or increase the decay rate of the circuit output. For example, weakening the strength of the self-activating promoter or adding a degradation tag to transcription factor would reduce the steady-state and serve to more closely match the expression levels of a the two circuits.

Ultrasensitivity

It has been demonstrated that regulatory systems with positive feedback are more sensitive to inducer than systems without feedback (Figure 7B) (Savageau 1974; Ferrell and Machleder 1998; Bashor et al. 2008). Positive feedback has since been experimentally verified to impart ultrasensitivity in both natural and engineered circuits. Ultrasensitivity occurs in positive feedback circuits where the strength of the feedback is not so large that the system loses the ability to occupy intermediate output states. The level of ultrasensitivity can be controlled by manipulating the strength of feedback by changing the stability of the activator protein or its cooperativity or binding affinity at the promoter.

Bistability

Positive feedback can create a bistable switch (Ferrell 2002). Bistable circuits can occupy only one of two states, canonically an OFF and an ON state, in response to a continuous range of input concentrations (Figure 7C). This can be very useful in circuit design and will be discussed further in Section III.C. It is challenging to design bistable circuits based on positive feedback (Ajo-Franklin et al. 2007) because if either of the states is quantitatively off balance with the other the system will only be able to occupy one state (Ferrell 2002). For example, leaky transcription of the positive feedback element is often sufficient to trip the switch and keep the circuit in a monostable ON state under all conditions.

A bistable switch based on positive transcriptional feedback has been constructed in *E.coli* (Isaacs et al. 2003). This circuit was composed of a temperature-sensitive transcriptional activator expressed under the control of the promoter which it activates. High kinetic constants of dimerization and transcriptional activation provided the non-linear responsivity which is required for bistability. At permissive temperatures, leaky transcription tripped the feedback switch driving all cells in the population to reach a stable ON state. At destabilizing temperatures, a lack of activator accumulation kept all cells in the OFF state. At intermediate temperatures the population bifurcated such that individual cells occupied either the ON or OFF state. This digital response occurred because intermediate protein expression states in any cell are unstable and small fluctuations are amplified to drive cells to quickly settle in either of the stable states (Ferrell 2002; Isaacs et al. 2003).

Bistable circuits have a unique property in that they can achieve different steady-state output responses under identical input conditions depending on their history (Figure 7D) (Ferrell 2002; Ninfa and Mayo 2004). That is, if the circuit begins in the OFF state it requires a greater input concentration to switch than if it began in the ON state. This characteristic, known as hysteresis is useful in the engineering robust cellular decision making. This is because hysteresis makes circuits with switch-like behaviors less sensitive to fluctuations in input signal near the switch point.

Ninfa and coworkers designed a transcriptional positive feedback circuit with a dominant repressor protein to construct a bistable switch in *E.coli* (Atkinson et al. 2003). In the absence of inducer, the repressor inactivated the feedback loop and the switch was OFF. At activating concentrations of inducer the circuit rapidly switched to the ON state. If the circuit had previously been exposed to high levels of inducer, however, it switched ON at ~70% lower inducer concentrations. Two key circuit parameters drove this system to exhibit hysteresis. First there was very high sensitivity within the switching range making intermediate expression states unstable. Second, the dynamic range of induction was large, on the order of 20-fold. These are the two most critical design requirements in the construction of positive feedback circuits with hysteretic properties (Ferrell 2002; Angeli et al. 2004; Ninfa and Mayo 2004).

Controlling Feedback Saturation

In a bistable switch, the magnitude of output gene expression in the ON state is determined by the protein production and decay parameters of the circuit. The level of gene expression in an activated bistable switch can therefore not be fine tuned. Because the steady-state output level is often an important design consideration in genetic engineering applications, we will discuss several strategies for controlling the magnitude of the ON state, or point of feedback saturation here.

In a simple positive feedback circuit, where an activator protein drives its own promoter, the steady-state output of the fully activated circuit is determined by the maximal rate of production and decay rate of the protein. In the synthetic circuit constructed by Collins and coworkers, the per cell output of the fully activated switch decreased continuously as the activator protein was destabilized (Isaacs et al. 2003). It is likely though that other circuit parameters such as promoter or RBS strength, or mRNA stability could be modified to achieve a similar result.

Eliminating Bistability to Generate a Continuous Response

Sugar inducible systems like *lac* and *ara* are the most widely used elements for engineered genetic control in *E.coli.* They are bistable because sugar-mediated transcriptional activation increases the rate of sugar uptake from the environment, generating a feedback loop. For many engineering applications this bistability is undesirable. Bistability creates discontinuous jumps in output as inducer is added. This hampers the freedom of the genetic engineer to set the circuit at intermediate output phenotypes. Moreover, at intermediate inducer concentrations the population can bifurcate such that some cells occupy the OFF state, some the ON state and none occupy an intermediate state. In many applications in biotechnology, however, it is beneficial for all cells in a population to behave identically. Fortunately, the bistable feedback circuits which nature provides can be modified for continuous input-output control and population homogeneity.

Several groups have shown that by expressing sugar uptake genes constitutively the positive feedback loops can be broken and bistability eliminated, allowing continuous induction over a large range of inducer (Khlebnikov et al. 2000; Khlebnikov et al. 2001; Khlebnikov and Keasling 2002; Morgan-Kiss et al. 2002). The deletion of the sugar catabolic genes from the host also aids in the homogeneity of the response (Morgan-Kiss et al. 2002).

III.B.3. Feed Forward Loops

A common genetic circuit in *E.coli* is the feed forward loop (FFL), where an input is split into two pathways, which then reconverge on an output (Milo et al. 2002; Shen-Orr et al. 2002). In its simplest form, an FFL consists of two regulatory genes (canonically X and Y) and one output gene (Z). Feed forward architecture results when X regulates the production of Y and both in turn regulate the production of Z (Figure 8).

There are two major classes of FFLs. In the first class, termed coherent FFLs, the sign of the regulatory interaction remains the same all the way through the circuit. That is X regulates Y and Z in the same manner that Y regulates Z. Coherent FFLs therefore regulate outputs similarly to single transcription factors, but introduce several quantitative performance differences. In the second class of FFLs, termed incoherent FFLs, the regulatory effect changes after the circuit splits, resulting in opposing regulation at the output. As we will discuss below, this type of circuit can result in interesting dynamic behaviors such as overshoots or pulses of gene expression.

Coherent FFLs: Activation Delays

A FFL is coherent if the regulatory effect of X on Z is the same as the effect of X on Z through Y (Figure 8). Coherent FFLs have been shown to act as sign-sensitive delays in *E.coli* signal processing networks (Mangan and Alon 2003; Mangan et al. 2003; Kalir et al. 2005). 'Sign-sensitive' refers to the fact that the circuits generate a lag in the transcriptional response to either the introduction or removal of an environmental signal, but not both. Activation delays can function as noise filters in that they prevent the circuit from responding to transient pulses of signals. Coherent FFLs are useful tools then for the engineering of sense-response behaviors in which the cell must parse sustained signal from input noise in the environment.

The basis of the delay in this type of FFL is intuitive. Z depends on the presence of X and Y for expression. Though the presence of input signal immediately activates X, Y cannot be expressed until X first accumulates. From that point, Y must then accumulate to a concentration sufficient to activate Z. Indeed, increasing the basal expression level of Y decreases the length of the delay (Mangan et al. 2003). The basal expression rate of an activator protein in a synthetic circuit is simple to tune with promoters or RBSs of different strengths, for example.

The natural arabinose responsive circuit of *E.coli* is a coherent type 1 FFL. This is not true, however, for the minimized pBAD circuit from which one of the natural regulators has been removed (Guzman et al. 1995). The natural arabinose

FFL circuit generates a delay in the activation of transcription after induction (Mangan et al. 2003). In *E.coli*, the absence of glucose increases intracellular cyclic adenosine monophosphate (cAMP) levels which activate the transcription factor CRP (X). CRP activates the expression of the *araC* (Y) gene, the product of which is a transcription factor whose function is dependent upon arabinose. The output araBAD (Z) promoter functions as a logical AND gate, requiring the presence of cAMP:CRP and arabinose:AraC for productive transcription. This FFL results in a ~0.2 cell cycle or 10-20 minute delay in activation of the Z promoter after the onset of inducing conditions. The delay is shown to be sign sensitive as the removal of the stimulus does not result in a delayed inactivation response as compared to a simple AND gate promoter without a feed forward connectivity between the two transcription factors.

Deactivation delays

The sign sensitivity of a FFL mediated delay can be changed by changing the activation logic of the Z promoter from AND to OR (Mangan and Alon 2003). Alon and coworkers proofed this concept by demonstrating that part of the *E.coli* flagellar apparatus is expressed under the control of a coherent FFL in which Z is expressed as a SUM function of X and Y (Kalir et al. 2005). SUM is a modified OR where the influence of X and Y on Z output is additive. Moreover, SUM is a simple operation to engineer in *E.coli*. SUM can be achieved by simply placing two different promoters in series. In this configuration, the first or second promoter can drive expression of the output gene, and if both are active, the rate of production of mRNA is greater.

In the flagellar example, X activates Y and the two transcription factors additively activate the operons that produce the flagellar motor (Kalir et al., 2005). If the input signal is removed and X is transiently inactivated, the circuit prolongs flagellar expression because Y levels linger. The authors show that the delay occurs under a wide range of circuit parameters, and that manipulation of the kinetic parameters of regulation can alter the length of the delay (Kalir and Alon 2004). A similar effect was shown for the Salmonella SPI-1 Type III Secretion System, which contains both a feed forward and split positive feedback loop (Temme et al. 2008).

Incoherent FFLs

An incoherent FFL consists of a circuit where X activates Y and Z but Y represses Z (Figure 8). There are over 100 examples of such a circuit in the E.coli genome (Mangan et al. 2006). This circuit generates several interesting and unique dynamical outputs such as pulses of gene expression and time-derivative sensing (Basu et al. 2004). In this section we will discuss the performance features of incoherent FFLs in *E.coli*, the effect of key molecular parameters on their function, and their application in the construction of some of the most sophisticated synthetic cellular behaviors to date.

ON accelerators

Because X first activates and then indirectly represses the expression of Z, incoherent FFLs result in 'overshoot dynamics' in the expression of Z (Mangan and Alon 2003). This means that Z temporarily reaches abundances greater than the final steady-state. Also, because the strength of a partially repressible promoter driving Z must be stronger than that of a non-repressible promoter capable of generating the same steady-state, the rise time of the output Z is necessarily increased in an I1-FFL (Mangan and Alon 2003). This property is similar to the accelerated response of negative feedback loops as described above. In a natural example, Alon and coworkers have demonstrated that the I1-FFL in the galactose utilization network of *E.coli* results in a 1.75-fold overshoot of the steady-state output and an approximately 3-fold acceleration in rise time (Mangan et al. 2006).

In I1-FFL circuits, important performance features such as the magnitude of response acceleration, the steady-state output and the size of the overshoot are particularly sensitive to the parameters associated with the repressor Y. In general, the higher the expression level of Y and the greater its repressive effects, the greater the acceleration of the circuit (Mangan and Alon 2003).

Pulse generators

A pulse generator is a genetic circuit capable of activating and then completely repressing output gene expression in response to the addition of an input. Incoherent FFLs can generate pulses of gene expression if the repression of Z by Y is very strong. In 2004, Weiss and coworkers constructed a synthetic I1-FFL in *E.coli.* In their design X was the transcription factor LuxR which is activated by the membrane permeable quorum sensing compound AI-1, Y was the strong transcriptional repressor CI and Z was the reporter gene *gfp*.

Because the circuit was constructed *de novo*, the authors could easily investigate the effects of genetic parameters such as the rate of synthesis of Y, and the strength of repression Z by Y. The authors noted that if either of these two parameters was too great, the circuit could never be activated by inducer (Basu et al. 2004). Under a range of permissive kinetic parameters, however, the circuit showed robust pulse generation after addition of inducer. The true pulse of gene expression occurred because the Y protein CI is a very strong repressor of its target promoter, capable of bringing output expression back to zero.

Critical pulse features such as amplitude and duration could be controlled by varying the kinetic parameters of the Y protein or the rate or concentration at which inducer was added. Specifically, the stronger the RBS or the affinity for the output promoter the shorter and smaller the resulting pulse. Furthermore, at fixed Y kinetic parameters, the pulse amplitude varied proportionally to both the absolute concentration and the rate of increase of inducer. This synthetic circuit is an elegant demonstration of the level of behavioral sophistication that can be designed *de novo* and optimized to the specifications of the engineer.

III.B.3. Dynamic Circuits

Several genetic circuits have been engineered which drive dynamic responses. A striking example is the three protein transcriptional ring oscillator known as the "repressilator" (Figure 9) (Elowitz and Leibler 2000). In this circuit, protein A represses protein B, protein B represses protein C and protein C represses protein A. Oscillations occur because the addition of an input signal can cause one of the proteins, say A, to become abundant and repress the next protein in line (B). Because

B is repressed, C begins to rise in abundance and can then in turn repress A. This process continues until A rises again, and in this manner the circuit encodes genetic oscillations. The repressilator was capable of generating three to four oscillations in a given cell, but showed a notable lack of uniformity across the population.

In another example, Ninfa and coworkers constructed a two-component transcriptional oscillator in which a transcription factor first activates itself and then activates its own repressor (Figure 9) (Atkinson et al. 2003). In this circuit an input triggers the activator to initially accumulate. After some time the activator is repressed by the accumulating repressor. As activator levels subsequently fall, so do repressor levels, triggering another round of activator accumulation. This circuit drove dampened oscillations over four periods, which spanned almost 60 hours.

A circuit based on cell-cell communication has been constructed to program population level oscillations in *E.coli* (Balagadde et al. 2005). In this design a gene which triggered cell death was expressed under the control of a quorum sensing circuit. The circuit was OFF at low cell densities but switched ON at high density. Microscopic monitoring demonstrated that *E.coli* expressing this circuit regularly oscillated in density from 1 to 3 cells per picoliter of media with a period of about 20 hours.

As discussed in Section III.B.2, Weiss and coworkers also constructed a dynamic circuit capable of generating a temporal pulse of gene expression in response to a single, step introduction of input signal (Figure 9) (Basu et al. 2004). The amplitude and duration of the pulse could be programmed by changing the strength or production rate of the repressor in the circuit. Moreover, because the circuit input was a membrane diffusible quorum sensing compound, a cell could be triggered to pulse by production of the activator in a nearby cell.

III.C. Switches and Logic

Genetic switches are circuits which rapidly transition between discreet states in response to input. Logical devices are circuits which interpret the states of multiple switches to produce a single, unified output. Switches and logic are useful because they aid the programming of desirable IF/THEN behaviors in *E.coli*. Genetic logic is carried out by circuits which can be rationally designed or combinatorially screened.

Extensibility, or the ease with which a switch or logic device can be connected to a different input or output is a desirable trait in switches and logic devices. Extensibility requires knowledge of the transfer functions of the parts. For example, the output range of a given switch or switches must be matched to the input range of a given logic device in order for signal transmission to proceed properly through the circuit. If expression in the OFF state of a switch is significantly leaky that it is interpreted as ON by the downstream logic gate then the circuit will not properly respond to input signals. If the transfer functions of switches and logic gates are well documented, they can then therefore be used "off the shelf" and connected to other well characterized parts for the rapid design of new systems.

NOT gate

One of the most useful and frequently constructed genetic logic operations is the Boolean NOT gate. Commonly referred to as an inverter, the NOT gate inverts the sign of the regulatory relationship between the input and output of the circuit. In the simplest system, this is accomplished by the introduction of a transcriptional repressor between the input and output (Figure 10A). An input signal which would otherwise activate expression of the output therefore inactivates it via the activation of a repressor. Besides inverting the input/output logic, NOT gates can also known to increase input sensitivity (Hooshangi et al. 2005; Karig and Weiss 2005; Pedraza and van Oudenaarden 2005) and lower sensing thresholds (Karig and Weiss 2005).

Many genetic NOT circuits fail to function properly when constructed. This often occurs because basal expression of the repressor in the absence of input can be sufficient to inhibit the output promoter, constitutively trapping the inverter in the OFF state. The abundance of the repressor protein can then be reduced to match the relevant sensitivity of the output promoter. This can be accomplished by weakening the RBS on the repressor mRNA, weakening the operator sites at the output promoter (Weiss 2001; Yokobayashi et al. 2002; Hooshangi et al. 2005) or randomly mutating the repressor to reduce its strength (Yokobayashi et al. 2002)

Switches and Memory

Memory is required for many sophisticated functions in electronic systems and is also ubiquitous in biology, forming the basis for the burgeoning field of epigenetics. One popular biological design goal which relies on memory is to construct cells that can count how old they are or how many times they and their ancestors have been exposed to some signal over a long period of time. Memory can be implemented as an extreme form of hysteresis in circuits with strong positive feedback. In such systems, previous exposure to high input signal triggers a circuit to remain active even when the signal goes to zero (Ferrell 2002).

In 2000, Collins and coworkers constructed a memory switch constructed in *E.coli* (Gardner et al. 2000). The switch was comprised of two cross-inhibiting transcriptional repressors. If repressor A was expressed, it repressed B and the switch was OFF. If an input was added which inactivated A, B accumulated and in turn, repressed A. This turned the switch ON. This switch generated stable memory over at least 22 hours, allowing a cell many generations away from the ancestor which had received the signal to maintain a stable response. This switch required proper matching of the transfer functions of its two subcomponents. If the expression level of one repressor was too strong in the OFF state the system became monostable. This required the screening of several combinations of promoters and RBSs of different strengths.

Arkin and coworkers have also constructed a memory device based on a permanent genetic rearrangement event. This circuit makes use of the recombinase encoded by the *fimE* gene to flip an improperly oriented promoter into alignment with an output gene (Ham et al. 2006). The DNA reorganization event is permanent, resulting in stable long-term circuit memory. Moreover, because the *fimE* gene can be expressed as the output of any sensor, the *fimE* switch is modular and can potentially generate memory to any input stimulus capable of activating gene expression. An advantage of this circuit is that it produces almost no basal expression when the promoter is in the opposite orientation from the gene it controls.

AND gate

The logical AND operation, where the presence of two inputs (A and B) are required to activate output expression, a useful concept for biological design and can also be applied to the construction of many more sophisticated logical operations. The most parsimonious approach for the construction of a genetic AND gate involves two inter-dependent genetic components which, when expressed simultaneously can initiate a downstream gene expression step. Such a system was recently implemented at the transcriptional level in *E.coli* (Anderson et al. 2007). In this setup, inducible promoter A drives the expression of an mRNA encoding the T7 RNA polymerase (RNAP) gene. The mRNA is non-functional however as two specific stop codons are introduced into the coding sequence. Inducible promoter B drives the expression of a tRNA which encodes an amino acid at those stop codons, rescuing translation of the RNAP. The circuit output is a promoter which is only transcribed by T7 RNAP protein such that it requires the presence of the two inputs A and B (Figure 10B). Importantly, this system was designed to be modular such that any two inducible promoters could be used to drive the AND gate. This modularity allowed the circuit to integrate signals from four different promoters and drive two separate output genes.

In the initial circuit design, the two components of the AND were not properly matched. The basal, or leaky expression rate of the T7 mRNA was significantly high that the circuit produced positive output in the presence of only one input. To reduce leaky expression, the authors randomly mutated the RBS preceding the T7 open reading frame and screened a library for variants dependent upon both inputs for activation. A majority of the variants in the library showed significant dependence on both inputs, indicating that the design was quite robust to variable expression levels. When the promoter driving the T7 mRNA was replaced, however, the new RBS failed to generate enough mRNA to activate the AND gate even when the promoter was fully active. To restore functionality an RBS library was again screened and again produced a viable circuit.

Other Logic

To construct other types of genetic logic, Leibler and coworkers randomly connected five promoters to three classical transcription factors which either activated or repressed them. Two ligands were chosen as inputs and one of the transcription factors was chosen to repress an output reporter gene. Several switch-like logical responses including NAND, NOR and NOT IF arose repeatedly from the circuit library (Guet et al. 2002). Interestingly circuits with similar connectivities, or profiles of regulatory contacts between components, were capable of generating different logical responses while networks with different connectivities were capable of generate the same logic. Many of the constructed circuits also produced intermediate or "fuzzy" logic. This work demonstrates the power of screening random combinations of regulators to achieve a particular desired logic operation.

A large number of intermediate logical operations were also observed in a related study wherein four different transcription factor binding sites were randomly placed in three locations around a single promoter (Cox et al. 2007). This combinatorial approach revealed that activator sites function most effectively when placed directly upstream of the -35 site and function poorly if at all when placed downstream of it. Repressor sites are more tolerant to different locations but are most effective when placed between the -35 and -10 sites.

IV. Actuators: Interfacing Cells with the Environment

A fundamental motivation for the programming of cells is that they have the ability to modify the chemistry and biology of their surrounding environments. Actuators are defined as gene products which elicit any type of cellular process or behavior from the production of a protein, to an enzyme capable of synthesizing drugs or fuels, to the synthesis and control of entire organelles and molecular machines. This section is meant to only briefly outline some of the things that *E.coli* can do.

State reporters

State reporters are molecules whose only function is to be observed or measured. When linked to genetic circuits, reporters can provide a 'print-out' of information coming in from cellular sensors and circuits. In biosensing applications acquiring information about the presence, absence concentration or temporal profile of an input signal in the environment or the cell is itself the goal of the system. Reporters can also provide a physical read out of the solution of computations performed by genetic circuits. The most common reporters are proteins such as β-galactosidase or Green Fluorescent Protein (GFP), the abundances of which can accurately be quantified by standard molecular biological techniques.

Metabolic Engineering

Metabolic engineering involves the expression of enzymes which divert cellular metabolites into alternative pathways to produce desired output products (Lee and Papoutsakis 1999). The enzymes used in metabolic engineering are therefore actuators which can be expressed as the outputs of genetic circuits. A typical metabolic design might employ sensors which detect the presence of upstream metabolites to time the expression of the biosynthetic enzymes which act upon them.

One application of metabolic engineering is the production of liquid fuels (Mielenz 2001; Jarboe et al. 2007; Service 2007; Keasling 2008). To this end, Liao and coworkers recently re-engineered *E.coli* amino acid metabolism for the production of branched chain alcohols, compounds which have desirable fuel properties (Atsumi et al. 2008). This required the expression of one of several two-enzyme clusters which converted intermediate metabolites from amino acid biosynthetic pathways to the various alcohols. Endogenous amino acid metabolic genes could also be overexpressed as complementary actuators to increase flux through the pathways and bump up fuel yields.

Metabolic actuators be used can reprogram *E.coli* for the production of therapeutic compounds as well (Pfeifer et al. 2001; Swartz 2001; Zhang et al. 2006). For example, Keasling and coworkers have introduced a large number of non-native isoprenoid biosynthetic enzymes into *E.coli* to efficiently convert the ubiquitous metabolite acetyl-CoA into artemisinic acid, a direct precursor to the potent and otherwise prohibitively expensive anti-malarial compound artemisinin. Optimization of enzyme expression levels and compensatory engineering to eliminate toxic byproducts has resulted in profound improvements in yield, approximately 1 million fold increase in a 4 years span (Keasling 2008). These efforts are likely to reduce the cost of artemisinin orders of magnitude, to prices compatible with its utilization in many underdeveloped countries with high malarial death rates.

Most metabolic engineering efforts to date have expressed the actuators under classically regulated circuits. These have been chosen for their simplicity and the continuous fine-grained control that they offer over enzyme expression levels. The construction of more sophisticated sensor-circuit-actuator systems should facilitate the design of increasingly ambitious microbial factories and help to optimize yields.

Organelle Transfer

Clusters of genes encoding entire organelles can also be used as actuators. Historically, the ability to manipulate such large fragments of DNA has required the presence of fortuitous restriction sites in the natural organelle sequences or specialized polymerase chain reaction (PCR) based methods. Improved DNA synthesis technologies now allow the *de novo* fabrication of organelle scale fragments.

In the initial demonstration of organelle transfer, 11 genes responsible for the synthesis of cytoplasmic gas vesicles in *B. megaterium* were moved into *E.coli* (Li and Cannon 1998). Expression of this gene cluster from a classically regulated circuit on a standard expression plasmid resulted in the formation of functional vesicles capable of significantly increasing the buoyancy of *E.coli* in aqueous media.

Similar strategies have resulted in the transfer of the fully functional nitrogen fixation (*nif*) (Dixon et al. 1976) and O antigen lipopolysaccharide (Bastin et al. 1991) enzyme clusters from *Klebsiella* and enteropathogenic *E.coli*, as well as the Type III protein secretion organelle from *Salmonella* (Wilson et al. 2007) and the cryptic Type

II organelle from *E.coli* itself (Francetic et al. 2000). These efforts used unmodified, contiguous genomic DNA fragments which were recombined into plasmids and introduced into *E.coli* 'as is'. These strategies therefore relied on expression from the natural promoters and RBSs of the relevant genes, and necessarily introduced the possibility of regulation by undefined control elements. The utility of organelle actuators will undoubtedly benefit from control through sensors and circuits.

Building Genetic Programs

Sensors, circuits and actuators are only truly modular engineering components when they can easily and arbitrarily be linked together. Several methodologies have recently been developed which allow the combination of multiple stretches of DNA without the need for inherent restriction sequences. One example is a universal, iterative cloning method for the assembly of standardized "Biobrick" parts (Shetty et al. 2008). In this method, a DNA part is computationally designed to internally lack several specific restriction sites. These restriction sites are then added to the upstream and downstream regions of the part and used as universal handles for the iterative, arbitrary connection of any two Biobricks. This standardized strategy increases the efficiency and ease with which any two parts can be combined (composability).

A PCR-based strategy termed SLIC has recently been developed for the "one-pot" assembly of up to 10 unrelated stretches of DNA in a specific order (Li and Elledge 2007). This method uses oligonucleotide primers to add specific linker sequences to any piece of DNA which then guide the order of assembly. The benefits to this approach are that no specific restriction sites need be avoided in the internal sequence of any part and that more than two parts can be combined in one step. Other advanced assembly strategies based on large scale oligonucleotide synthesis and polymerase chain reaction (PCR) assembly have allowed the construction of complete viral (Cello et al. 2002; Smith et al. 2003; Tumpey et al. 2005) and even bacterial (Gibson et al. 2008) genomes from computationally designed DNA information.

Standardization and assembly technologies are already helping eliminate barriers between the design and physical construction of DNA (Endy 2005), a process which has been the historical rate limiting step in genetic engineering. A true leap in biological design will occur when these technologies become more widely available and less expensive, allowing true modular assembly of sensors, circuits and actuators. In a watershed example, Collins and coworkers linked a DNA damage sensor to a bistable genetic switch to drive an actuator which triggered biofilm formation in *E.coli*. In this bottom up programming effort, the *E.coli* could stably and strongly switch ON biofilm formation phenotype in the presence of DNA damaging environmental inputs such as UV light or a chemical mutagen (Kobayashi et al. 2004).

Finally, when genetic parts are linked together in a design their quantitative input/output properties must be properly matched (Yokobayashi et al. 2002). As discussed in Section III.C above, if the OFF state of a sensor is sufficiently leaky to activate a downstream genetic circuit, the circuit will not be capable of receiving signaling information from the sensor. There are many strategies for matching the transfer functions of multiple parts, but until universal metrics of genetic activity can be established (Endy 2005) there will always be a significant troubleshooting component in the assembly of functional systems.

V. Conclusions

The vast molecular genetic literature on *E.coli* has made it the subject of choice for many early efforts in synthetic biology. Five decades of work have given genetic engineers a rich repository of parts, often sensors and actuators, which can be taken out of their natural context and used for new, user-defined purposes. More recent efforts have established useful circuit design principles that have further pushed the level of sophistication of behaviors that can be designed.

Complementing the scientific contributions, DNA synthesis and sequencing technologies have become increasingly high throughput and less expensive in the past few years. Further advances will bolster biological design by allowing researchers to bypass the arduous process of physically constructing designed DNA sequences. In the end, *E.coli* synthetic biology serves two major purposes. It enables the goal-oriented engineering of strains which can carry out novel functions of medical, industrial or academic interest and it serves as a bottom-up complement to top-down systems approaches for the elucidation of the molecular principles which govern cellular behavior.

References

Ajo-Franklin CM, Drubin DA, Eskin JA, Gee EP, Landgraf D, Phillips I, Silver PA (2007) Rational design of memory in eukaryotic cells. Genes Dev 21:2271-6

Andersen JB, Sternberg C, Poulsen LK, Bjorn SP, Givskov M, Molin S (1998) New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. Appl Environ Microbiol 64:2240-6

Anderson JC, Clarke EJ, Arkin AP, Voigt CA (2006) Environmentally controlled invasion of cancer cells by engineered bacteria. J Mol Biol 355:619-27

Anderson JC, Voigt CA, Arkin AP (2007) Environmental signal integration by a modular AND gate. Mol Syst Biol 3:133

Angeli D, Ferrell JE, Jr., Sontag ED (2004) Detection of multistability, bifurcations, and hysteresis in a large class of biological positive-feedback systems. Proc Natl Acad Sci U S A 101:1822-7

Atkinson MR, Savageau MA, Myers JT, Ninfa AJ (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in Escherichia coli. Cell 113:597-607

Atsumi S, Hanai T, Liao JC (2008) Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature 451:86-9

Backes H, Berens C, Helbl V, Walter S, Schmid FX, Hillen W (1997) Combinations of the alpha-helix-turn-alpha-helix motif of TetR with respective residues from LacI or 434Cro: DNA recognition, inducer binding, and urea-dependent denaturation. Biochemistry 36:5311-22

Balagadde FK, You L, Hansen CL, Arnold FH, Quake SR (2005) Long-term monitoring of bacteria undergoing programmed population control in a microchemostat. Science 309:137-40

Balagadde FK, Song H, Ozaki J, Collins CH, Barnet M, Arnold FH, Quake SR, You L (2008) A synthetic Escherichia coli predator-prey ecosystem. Mol Syst Biol 4:187

Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N (2006) Noise in protein expression scales with natural protein abundance. Nat Genet 38:636-43

Bashor CJ, Helman NC, Yan S, Lim WA (2008) Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. Science 319:1539-43

Bastin DA, Romana LK, Reeves PR (1991) Molecular cloning and expression in Escherichia coli K-12 of the rfb gene cluster determining the O antigen of an E. coli O111 strain. Mol Microbiol 5:2223-31

Basu S, Mehreja R, Thiberge S, Chen MT, Weiss R (2004) Spatiotemporal control of gene expression with pulse-generating networks. Proc Natl Acad Sci U S A 101:6355-60

Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R (2005) A synthetic multicellular system for programmed pattern formation. Nature 434:1130-4

Batchelor E, Goulian M (2006) Imaging OmpR localization in Escherichia coli. Mol Microbiol 59:1767-78

Baumgartner JW, Kim C, Brissette RE, Inouye M, Park C, Hazelbauer GL (1994) Transmembrane signalling by a hybrid protein: communication from the domain of chemoreceptor Trg that recognizes sugar-binding proteins to the kinase/phosphatase domain of osmosensor EnvZ. J Bacteriol 176:1157-63

Bayer TS, Smolke CD (2005) Programmable ligand-controlled riboregulators of eukaryotic gene expression. Nat Biotechnol 23:337-43

Becskei A, Serrano L (2000) Engineering stability in gene networks by autoregulation. Nature 405:590-3

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R (2005a) Transcriptional regulation by the numbers: applications. Current Opinion in Genetics & Development 15:125-135

Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R (2005b) Transcriptional regulation by the numbers: models. Current Opinion in Genetics & Development 15:116-124

Brenner K, Karig DK, Weiss R, Arnold FH (2007) Engineered bidirectional communication mediates a consensus in a microbial biofilm consortium. Proc Natl Acad Sci U S A 104:17300-4

Brosius J, Erfle M, Storella J (1985) Spacing of the -10 and -35 regions in the tac promoter. Effect on its in vivo activity. J Biol Chem 260:3539-41

Buskirk AR, Landrigan A, Liu DR (2004) Engineering a ligand-dependent RNA transcriptional activator. Chem Biol 11:1157-63

Canton B, Labno A, Endy D (2008) Refinement and standardization of synthetic biological parts and devices. Nat Biotechnol 26:787-93

Cello J, Paul AV, Wimmer E (2002) Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science 297:1016-8

Chen Q, Kadner RJ (2000) Effect of altered spacing between uhpT promoter elements on transcription activation. J Bacteriol 182:4430-6

Collins CH, Arnold FH, Leadbetter JR (2005) Directed evolution of Vibrio fischeri LuxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones. Mol Microbiol 55:712-23

Collins CH, Leadbetter JR, Arnold FH (2006) Dual selection enhances the signaling specificity of a variant of the quorum-sensing transcriptional activator LuxR. Nat Biotechnol 24:708-12

Cox RS, 3rd, Surette MG, Elowitz MB (2007) Programming gene expression with combinatorial promoters. Mol Syst Biol 3:145

de Boer PA, Crossley RE, Rothfield LI (1983) Proc Natl Acad Sci U S A 80:21-25

de la Torre JC, Ortin J, Domingo E, Delamarter J, Allet B, Davies J, Bertrand KP, Wray LV, Jr., Reznikoff WS (1984) Plasmid vectors based on Tn10 DNA: gene expression regulated by tetracycline. Plasmid 12:103-10

Derr P, Boder E, Goulian M (2006) Changing the specificity of a bacterial chemoreceptor. J Mol Biol 355:923-32

Dixon R, Cannon F, Kondorosi A (1976) Construction of a P plasmid carrying nitrogen fixation genes from Klebsiella pneumoniae. Nature 260:268-71

Drubin DA, Way JC, Silver PA (2007) Designing biological systems. Genes Dev 21:242-54

Dwyer MA, Looger LL, Hellinga HW (2003) Computational design of a Zn2+ receptor that controls bacterial gene expression. Proc Natl Acad Sci U S A 100:11255-60

El-Samad H, Khammash M (2006) Regulated degradation is a mechanism for suppressing stochastic fluctuations in gene regulatory networks. Biophysical Journal 90:3749-3761

Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403:335-8

Endy D (2005) Foundations for engineering biology. Nature 438:449-53

Endy D (2008) Genomics. Reconstruction of the genomes. Science 319:1196-7

Falcon CM, Matthews KS (2000) Operator DNA sequence variation enhances high affinity binding by hinge helix mutants of lactose repressor protein. Biochemistry 39:11074-83

Ferrell JE, Jr., Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280:895-8

Ferrell JE, Jr. (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. Curr Opin Cell Biol 14:140-8

Francetic O, Belin D, Badaut C, Pugsley AP (2000) Expression of the endogenous type II secretion pathway in Escherichia coli leads to chitinase secretion. Embo J 19:6697-703

Frank DE, Saecker RM, Bond JP, Capp MW, Tsodikov OV, Melcher SE, Levandoski MM, Record MT, Jr. (1997) Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator: effects of converting a consensus site to a non-specific site. J Mol Biol 267:1186-206

Gambetta GA, Lagarias JC (2001) Genetic engineering of phytochrome biosynthesis in bacteria. Proc Natl Acad Sci U S A 98:10566-71

Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in Escherichia coli. Nature 403:339-42

Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, Merryman C, Young L, Noskov VN, Glass JI, Venter JC, Hutchison CA, 3rd, Smith HO (2008) Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science 319:1215-20

Greber D, Fussenegger M (2007) Mammalian synthetic biology: engineering of sophisticated gene networks. J Biotechnol 130:329-45

Guet CC, Elowitz MB, Hsing W, Leibler S (2002) Combinatorial synthesis of genetic networks. Science 296:1466-70

Guzman LM, Belin D, Carson MJ, Beckwith J (1995) Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J Bacteriol 177:4121-30

Ham TS, Lee SK, Keasling JD, Arkin AP (2006) A tightly regulated inducible expression system utilizing the fim inversion recombination switch. Biotechnol Bioeng 94:1-4

Hasty J, McMillen D, Collins JJ (2002) Engineered gene circuits. Nature 420:224-30

Hawkins AC, Arnold FH, Stuermer R, Hauer B, Leadbetter JR (2007) Directed evolution of Vibrio fischeri LuxR for improved response to butanoyl-homoserine lactone. Appl Environ Microbiol 73:5775-81

Helbl V, Hillen W (1998) Stepwise selection of TetR variants recognizing tet operator 4C with high affinity and specificity. J Mol Biol 276:313-8

Helbl V, Tiebel B, Hillen W (1998) Stepwise selection of TetR variants recognizing tet operator 6C with high affinity and specificity. J Mol Biol 276:319-24

Henssler EM, Scholz O, Lochner S, Gmeiner P, Hillen W (2004) Structure-based design of Tet repressor to optimize a new inducer specificity. Biochemistry 43:9512-8

Hoch J, Silhavy T (1995) Two Component Signal Transduction.

Hooshangi S, Thiberge S, Weiss R (2005) Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc Natl Acad Sci U S A 102:3581-6

Isaacs FJ, Hasty J, Cantor CR, Collins JJ (2003) Prediction and measurement of an autoregulatory genetic module. Proc Natl Acad Sci U S A 100:7714-9

Isaacs FJ, Dwyer DJ, Ding C, Pervouchine DD, Cantor CR, Collins JJ (2004) Engineered riboregulators enable post-transcriptional control of gene expression. Nat Biotechnol 22:841-7

Isaacs FJ, Dwyer DJ, Collins JJ (2006) RNA synthetic biology. Nat Biotechnol 24:545-54

Jarboe LR, Grabar TB, Yomano LP, Shanmugan KT, Ingram LO (2007) Development of ethanologenic bacteria. Adv Biochem Eng Biotechnol 108:237-61

Jose AM, Soukup GA, Breaker RR (2001) Cooperative binding of effectors by an allosteric ribozyme. Nucleic Acids Res 29:1631-7

Kaern M, Blake WJ, Collins JJ (2003) The engineering of gene regulatory networks. Annu Rev Biomed Eng 5:179-206

Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6:451-64

Kalir S, McClure J, Pabbaraju K, Southward C, Ronen M, Leibler S, Surette MG, Alon U (2001) Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science 292:2080-3

Kalir S, Alon U (2004) Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. Cell 117:713-20

Kalir S, Mangan S, Alon U (2005) A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli. Mol Syst Biol 1:2005 0006

Kamionka A, Sehnal M, Scholz O, Hillen W (2004) Independent regulation of two genes in Escherichia coli by tetracyclines and Tet repressor variants. J Bacteriol 186:4399-401

Karig D, Weiss R (2005) Signal-amplifying genetic circuit enables in vivo observation of weak promoter activation in the Rhl quorum sensing system. Biotechnol Bioeng 89:709-18

Keasling JD (2008) Synthetic biology for synthetic chemistry. ACS Chem Biol 3:64-76

Khlebnikov A, Risa O, Skaug T, Carrier TA, Keasling JD (2000) Regulatable arabinose-inducible gene expression system with consistent control in all cells of a culture. J Bacteriol 182:7029-34

Khlebnikov A, Datsenko KA, Skaug T, Wanner BL, Keasling JD (2001) Homogeneous expression of the P(BAD) promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter. Microbiology 147:3241-7

Khlebnikov A, Keasling JD (2002) Effect of lacY expression on homogeneity of induction from the P(tac) and P(trc) promoters by natural and synthetic inducers. Biotechnol Prog 18:672-4

Kleckner N, Barker DF, Ross DG, Botstein D (1978) Properties of the translocatable tetracycline-resistance element Tn10 in Escherichia coli and bacteriophage lambda. Genetics 90:427-61

Kobayashi H, Kaern M, Araki M, Chung K, Gardner TS, Cantor CR, Collins JJ (2004) Programmable cells: interfacing natural and engineered gene networks. Proc Natl Acad Sci U S A 101:8414-9

Koizumi M, Soukup GA, Kerr JN, Breaker RR (1999) Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. Nat Struct Biol 6:1062-71

Laub MT, Biondi EG, Skerker JM (2007) Phosphotransfer profiling: systematic mapping of two-component signal transduction pathways and phosphorelays. Methods Enzymol 423:531-48

Lee SY, Papoutsakis ET (1999) Metabolic Engineering.

Levskaya A, Chevalier AA, Tabor JJ, Simpson ZB, Lavery LA, Levy M, Davidson EA, Scouras A, Ellington AD, Marcotte EM, Voigt CA (2005) Synthetic biology: engineering Escherichia coli to see light. Nature 438:441-2

Li MZ, Elledge SJ (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods 4:251-6

Li N, Cannon MC (1998) Gas vesicle genes identified in Bacillus megaterium and functional expression in Escherichia coli. J Bacteriol 180:2450-8

Looger LL, Dwyer MA, Smith JJ, Hellinga HW (2003) Computational design of receptor and sensor proteins with novel functions. Nature 423:185-90

Lutz R, Bujard H (1997) Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. Nucleic Acids Res 25:1203-10

Lutz R, Lozinski T, Ellinger T, Bujard H (2001a) Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. Nucleic Acids Res 29:3873-81

Lutz R, Lozinski T, Ellinger T, Bujard H (2001b) Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. Nucleic Acids Research 29:3873-3881

Lynch SA, Desai SK, Sajja HK, Gallivan JP (2007) A high-throughput screen for synthetic riboswitches reveals mechanistic insights into their function. Chem Biol 14:173-84

Mangan S, Alon U (2003) Structure and function of the feed-forward loop network motif. Proc Natl Acad Sci U S A 100:11980-5

Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334:197-204

Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of Escherichia coli. J Mol Biol 356:1073-81

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911-40

McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A 94:814-9

Michalowski CB, Short MD, Little JW (2004) Sequence tolerance of the phage lambda PRM promoter: implications for evolution of gene regulatory circuitry. J Bacteriol 186:7988-99

Mielenz JR (2001) Ethanol production from biomass: technology and commercialization status. Curr Opin Microbiol 4:324-9

Miller MB, Bassler BL (2001) Quorum sensing in bacteria. Annu Rev Microbiol 55:165-99

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298:824-7

Morgan-Kiss RM, Wadler C, Cronan JE, Jr. (2002) Long-term and homogeneous regulation of the Escherichia coli araBAD promoter by use of a lactose transporter of relaxed specificity. Proc Natl Acad Sci U S A 99:7373-7

Ninfa AJ, Mayo AE (2004) Hysteresis vs. graded responses: the connections make all the difference. Sci STKE 2004:pe20

Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. Science 307:1965-9

Pfeifer BA, Admiraal SJ, Gramajo H, Cane DE, Khosla C (2001) Biosynthesis of complex polyketides in a metabolically engineered strain of E-coli. Science 291:1790-1792

Posfai G, Plunkett G, 3rd, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, Burland V, Harcum SW, Blattner FR (2006) Emergent properties of reduced-genome Escherichia coli. Science 312:1044-6

Ptashne M, Gann A (2002) Genes & Signals.

Rosenfeld N, Elowitz MB, Alon U (2002) Negative autoregulation speeds the response times of transcription networks. J Mol Biol 323:785-93

Salis H, Tamsir A, Voigt CA (2009 (in press)) Engineering bacterial sensors and signals. Bacterial sensing and signaling

Savageau MA (1974) Comparison of classical and autogenous systems of regulation in inducible operons. Nature 252:546-9

Service RF (2007) Cellulosic ethanol. Biofuel researchers prepare to reap a new harvest. Science 315:1488-91

Shahrezaei V, Ollivier JF, Swain PS (2008) Colored extrinsic fluctuations and stochastic gene expression. Mol Syst Biol 4:196

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 31:64-8

Shetty RP, Endy D, Knight TF, Jr. (2008) Engineering BioBrick vectors from BioBrick parts. J Biol Eng 2:5

Sia SK, Gillette BM, Yang GJ (2007) Synthetic tissue biology: tissue engineering meets synthetic biology. Birth Defects Res C Embryo Today 81:354-61

Skerker JM, Prasol MS, Perchuk BS, Biondi EG, Laub MT (2005) Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. PLoS Biol 3:e334

Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133:1043-54

Smith HO, Hutchison CA, 3rd, Pfannkoch C, Venter JC (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci U S A 100:15440-5

Smith TL, Sauer RT (1995) P22 Arc repressor: role of cooperativity in repression and binding to operators with altered half-site spacing. J Mol Biol 249:729-42

Soukup GA, Breaker RR (1999a) Design of allosteric hammerhead ribozymes activated by ligand-induced structure stabilization. Structure 7:783-91

Soukup GA, Breaker RR (1999b) Engineering precision RNA molecular switches. Proc Natl Acad Sci U S A 96:3584-9

Soukup GA, Breaker RR (1999c) Relationship between internucleotide linkage geometry and the stability of RNA. Rna 5:1308-25

Soukup GA, DeRose EC, Koizumi M, Breaker RR (2001) Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. Rna 7:524-36

Swartz JR (2001) Advances in Escherichia coli production of therapeutic proteins. Curr Opin Biotechnol 12:195-201

Takeda Y, Sarai A, Rivera VM (1989) Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc Natl Acad Sci U S A 86:439-43

Tang J, Breaker RR (1997) Rational design of allosteric ribozymes. Chem Biol 4:453-9

Temme K, Salis H, Tullman-Ercek D, Levskaya A, Hong SH, Voigt CA (2008) Induction and relaxation dynamics of the regulatory network controlling the type III secretion system encoded within Salmonella pathogenicity island 1. J Mol Biol 377:47-61

Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci U S A 98:8614-9

Topp S, Gallivan JP (2008) Random walks to synthetic riboswitches--a high-throughput selection based on cell motility. Chembiochem 9:210-3

Tumpey TM, Basler CF, Aguilar PV, Zeng H, Solorzano A, Swayne DE, Cox NJ, Katz JM, Taubenberger JK, Palese P, Garcia-Sastre A (2005) Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science 310:77-80

Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. Trends Microbiol 13:52-6

Utsumi R, Brissette RE, Rampersaud A, Forst SA, Oosawa K, Inouye M (1989) Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. Science 245:1246-9

Vilar JM, Leibler S (2003) DNA looping and physical constraints on transcription regulation. J Mol Biol 331:981-9

Voigt CA (2006) Genetic parts to program bacteria. Curr Opin Biotechnol 17:548-57

Wagner R (2000) Transcription Regulation in Prokaryotes.

Wall ME, Hlavacek WS, Savageau MA (2004) Design of gene circuits: lessons from bacteria. Nat Rev Genet 5:34-42

Ward SM, Delgado A, Gunsalus RP, Manson MD (2002) A NarX-Tar chimera mediates repellent chemotaxis to nitrate and nitrite. Mol Microbiol 44:709-19

Weiss R, Homsy GE, Knight TF, Jr. (1999) Towards in vivo digital circuits. DIMACS Workshop on Evolution as Computation 1:1-18

Weiss R (2001) Cellular Computation and Communications Using Engineered Genetic Regulatory Networks.

Williams SB, Stewart V (1997) Discrimination between structurally related ligands nitrate and nitrite controls autokinase activity of the NarX transmembrane signal transducer of Escherichia coli K-12. Mol Microbiol 26:911-25

Wilson JW, Coleman C, Nickerson CA (2007) Cloning and transfer of the Salmonella pathogenicity island 2 type III secretion system for studies of a range of gram-negative genera. Appl Environ Microbiol 73:5911-8

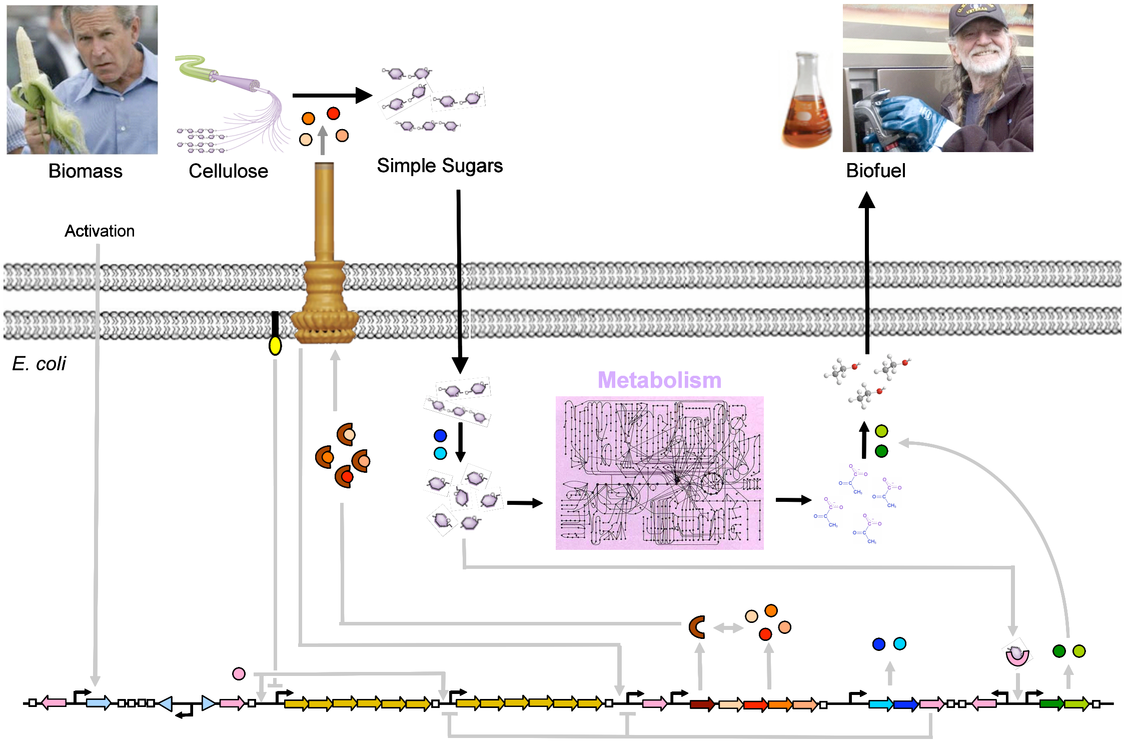
Winkler WC, Breaker RR (2003) Genetic control by metabolite-binding riboswitches. Chembiochem 4:1024-32

Yen L, Svendsen J, Lee JS, Gray JT, Magnier M, Baba T, D'Amato RJ, Mulligan RC (2004) Exogenous control of mammalian gene expression through modulation of RNA self-cleavage. Nature 431:471-6

Yokobayashi Y, Weiss R, Arnold FH (2002) Directed evolution of a genetic circuit. Proc Natl Acad Sci U S A 99:16587-91

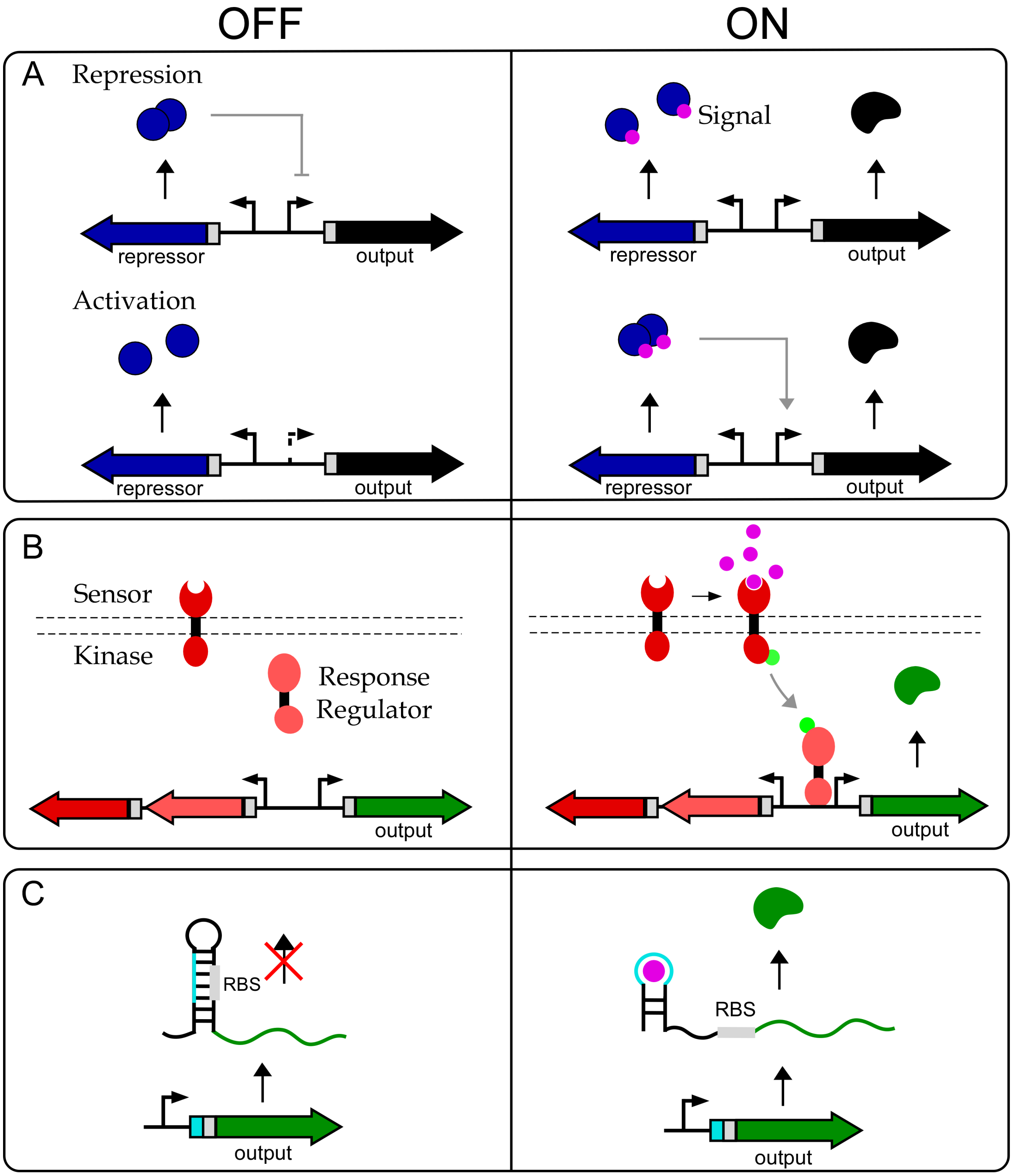
You L, Cox RS, 3rd, Weiss R, Arnold FH (2004) Programmed population control by cell-cell communication and regulated killing. Nature 428:868-71

Zhang W, Ames BD, Tsai SC, Tang Y (2006) Engineered biosynthesis of a novel amidated polyketide, using the malonamyl-specific initiation module from the oxytetracycline polyketide synthase. Appl Environ Microbiol 72:2573-80



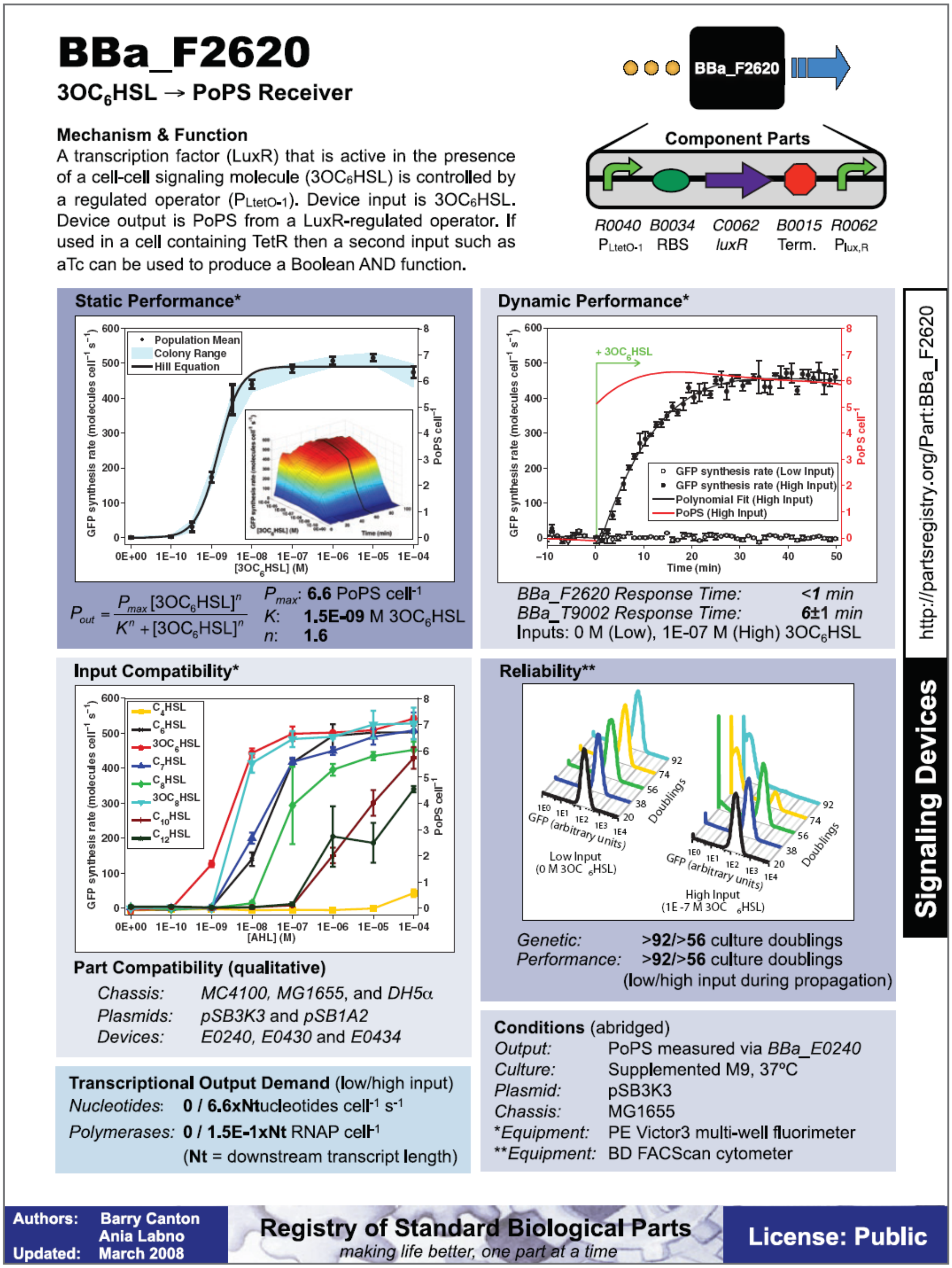
Chapter 5: Figure 1. A hypothetical *E. coli* program to convert biomass to liquid fuel.

Complex plant material requires that multiple enzymes be exported in a timed sequence. The enzymes need to be exported from the cell, in this case using a type III secretion system imported from Salmonella. The build up of simple sugars induces a pathway to break them down into glucose and covert a metabolic product into a fuel. This is an example of integrated bioprocessing, where multiple steps of a manufacturing process are programmed into a single organism. This requires the combination of sensors, circuits, and actuators to control and respond to a sequence of events.



Chapter 5: Figure 2. (A) Classical transcriptional regulation. In classical systems a cytoplasmic transcription factor protein regulates the target genes in response to the presence of input ligand (pink dots). Classical regulation can occur in two forms, repression or activation. With repression, the transcription factor binds to the

promoter in the absence of ligand (left) and undergoes a conformational change upon ligand binding which causes it to dissociate from the DNA, activating transcription (right). With activation, the transcription factor does not associate with the promoter in the absence of ligand (dashed), but does in its presence, increasing the rate of transcription. (B) Two-component sensing. A membrane associated sensor-kinase protein associates with an extracellular ligand at its sensor domain, which drives a structural change in its cytoplasmic kinase domain. This triggers autophosphorylation of the cytoplasmic domain. The phosphate group (green dot) is then transferred to the receiver domain of a cytoplasmic, diffusible response regulator protein (pink). When phosphorylated, the response regulator changes conformation and binds to its cognate operator sites near promoters, activating or repressing gene expression. (C) An engineered riboswitch. A constitutive promoter drives the expression of a gene with an engineered RNA hairpin occluding its ribosome binding site (RBS, grey) and blocking translation. The hairpin also carries an aptamer sequence (turquoise), which can bind to a cognate ligand (purple square), triggering a structural rearrangement which liberates the RBS for productive translation. Adapted from Topp and Gallivan, 2007.



BBa_F2620

$3OC_6HSL$ → PoPS Receiver

BBa_F2620
Component Parts

R0040	B0034	C0062	B0015	R0062
$P_{LtetO-1}$	RBS	*luxR*	Term.	$P_{lux,R}$

Mechanism & Function

A transcription factor (LuxR) that is active in the presence of a cell-cell signaling molecule ($3OC_6HSL$) is controlled by a regulated operator ($P_{LtetO-1}$). Device input is $3OC_6HSL$. Device output is PoPS from a LuxR-regulated operator. If used in a cell containing TetR then a second input such as aTc can be used to produce a Boolean AND function.

Static Performance*

$$P_{out} = \frac{P_{max}[3OC_6HSL]^n}{K^n + [3OC_6HSL]^n}$$

P_{max}: **6.6** PoPS cell^{-1}
K: **1.5E-09** M $3OC_6HSL$
n: **1.6**

Dynamic Performance*

+ 3OC$_6$HSL
GFP synthesis rate (Low Input)
GFP synthesis rate (High Input)
Polynomial Fit (High Input)
PoPS (High Input)
GFP synthesis rate (molecules cell^{-1} s^{-1})
PoPS cell^{-1}
0 100 200 300 400 500 600
0 1 2 3 4 5 6 7 8
−10 0 10 20 30 40 50
Time (min)

BBa_F2620 Response Time:	**<1** *min*
BBa_T9002 Response Time:	**6±1** *min*

Inputs: 0 M (Low), 1E-07 M (High) $3OC_6HSL$

Input Compatibility*

Part Compatibility (qualitative)

Chassis:	*MC4100, MG1655*, and *DH5α*
Plasmids:	*pSB3K3* and *pSB1A2*
Devices:	*E0240, E0430* and *E0434*

Reliability**

Genetic:	**>92/>56** culture doublings
Performance:	**>92/>56** culture doublings

(low/high input during propagation)

Transcriptional Output Demand (low/high input)

Nucleotides: **0 / 6.6xNt**ucleotides cell^{-1} s^{-1}

Polymerases: **0 / 1.5E-1xNt** RNAP cell^{-1}

(**Nt** = downstream transcript length)

Conditions (abridged)

Output:	PoPS measured via *BBa_E0240*
Culture:	Supplemented M9, 37°C
Plasmid:	pSB3K3
Chassis:	MG1655
**Equipment:*	PE Victor3 multi-well fluorimeter
***Equipment:*	BD FACScan cytometer

http://partsregistry.org/Part:BBa_F2620

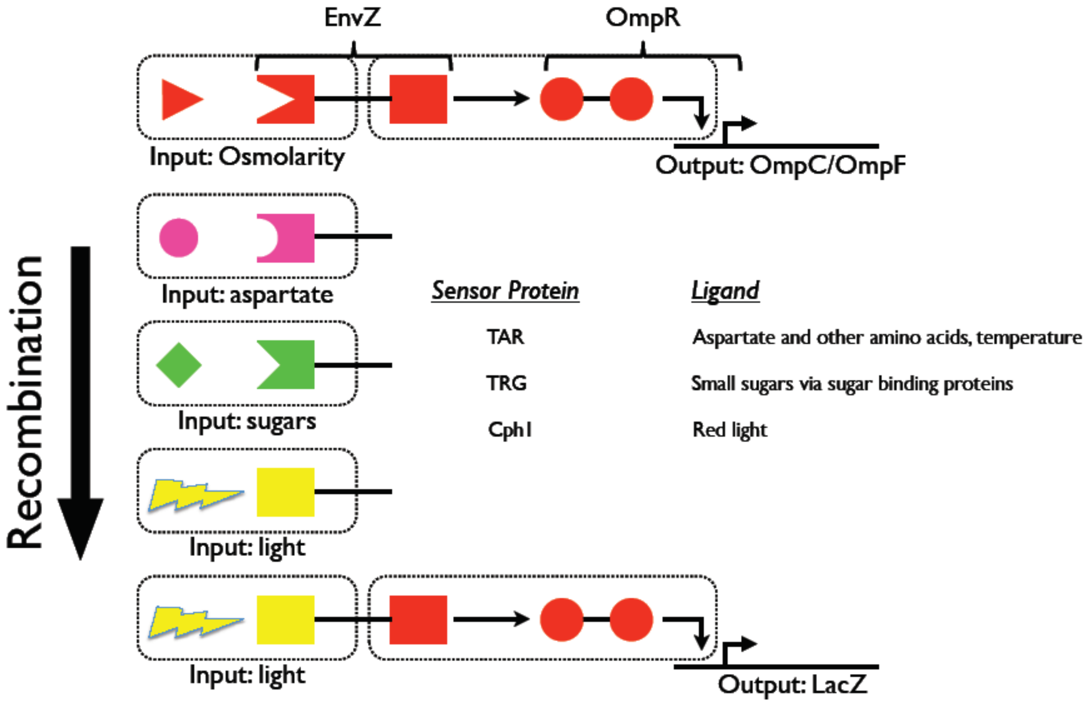
Signaling Devices

Authors: Barry Canton
Ania Labno
Updated: March 2008

Registry of Standard Biological Parts
making life better, one part at a time

License: Public

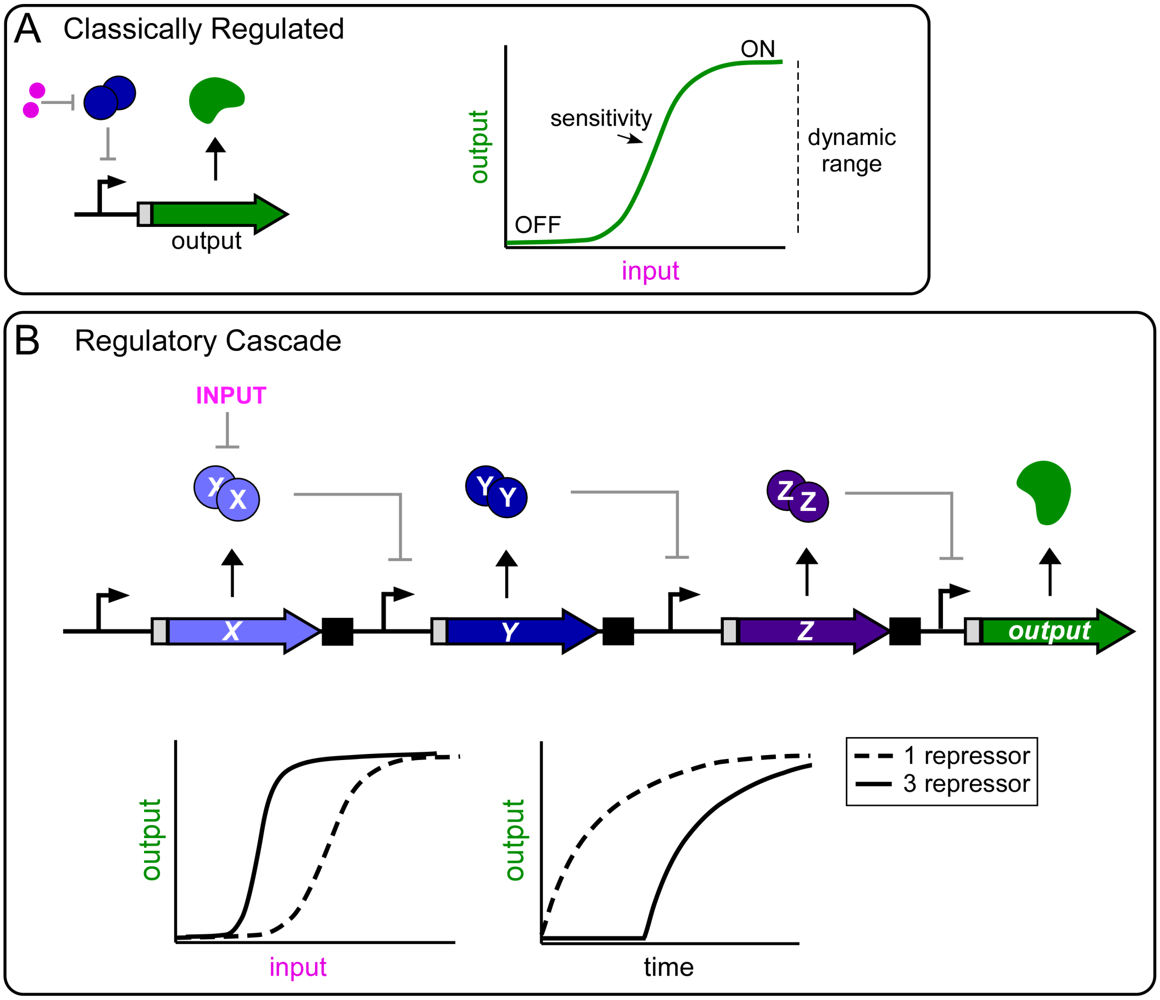
Chapter 5: Figure 3. Performance Feature Specification Sheet (Canton et al. 2008).



Sensor Protein	*Ligand*
TAR	Aspartate and other amino acids, temperature
TRG	Small sugars via sugar binding proteins
Cph1	Red light

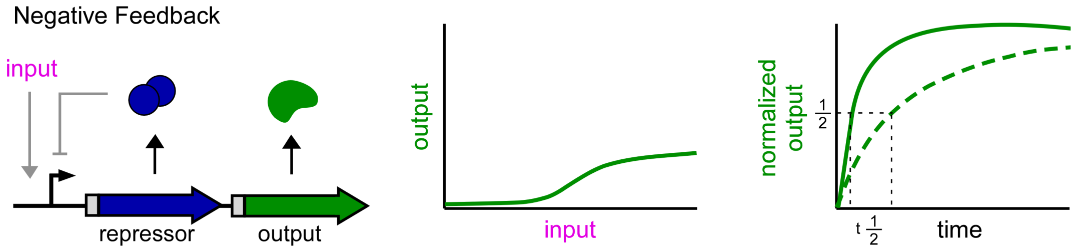
Chapter 5: Figure 4. Domain Swapping.

The sensory domain of EnvZ sense inputs and transfers information to the response regulator OmpR through the kinase domain. Other sensory domains can replace the naturally occurring EnvZ sensory domain to create chimeric sensor proteins.



Chapter 5: Figure 5. (A) Transfer function. Classically regulated promoters typically show sigmoidal response profiles to the concentration of inducer. In region I, well below the K$_D$ of the transcription factor (blue) for the inducer (pink), output changes little changes in input. In region II the concentration of output increases steadily and continuously as a function of input concentration. In region III there is no further increase in output. (B) Regulatory cascade. An input signal inactivates repressor protein X, resulting in the derepression of repressor Y. Upon accumulation of Y, repressor Z is repressed and its levels decline, increasing the concentration of

the output (green).   (Lower left) Cascading results in ultrasensitivity and lower sensing thresholds. A single repressor version of the above circuit (dashed line) shows a standard sigmoidal response. The 3 repressor cascade amplifies signal, reducing the absolute concentration of inducer required to activate the circuit and increasing the sensitivity of the response. (Bottom right) Cascading generates lags in response time. The single repressor circuit (dashed line) responds instantaneously to introduction of inducer while the 3 repressor cascade generates a significant latency in the response.



Chapter 5: Figure 6. Negative Transcriptional Feedback.

A repressor protein (blue) is encoded under the control of the promoter which it regulates. The shape of the input/output curve is the same as above, but the system reaches equal or less output at any given concentration of input. The rise time (vertical dashed lines), or time required for the circuit to reach 50% of its steady-state output (horizontal dashed line) is significantly decreased in negative feedback (solid green line) as compared to classically regulated (dashed green line) systems.

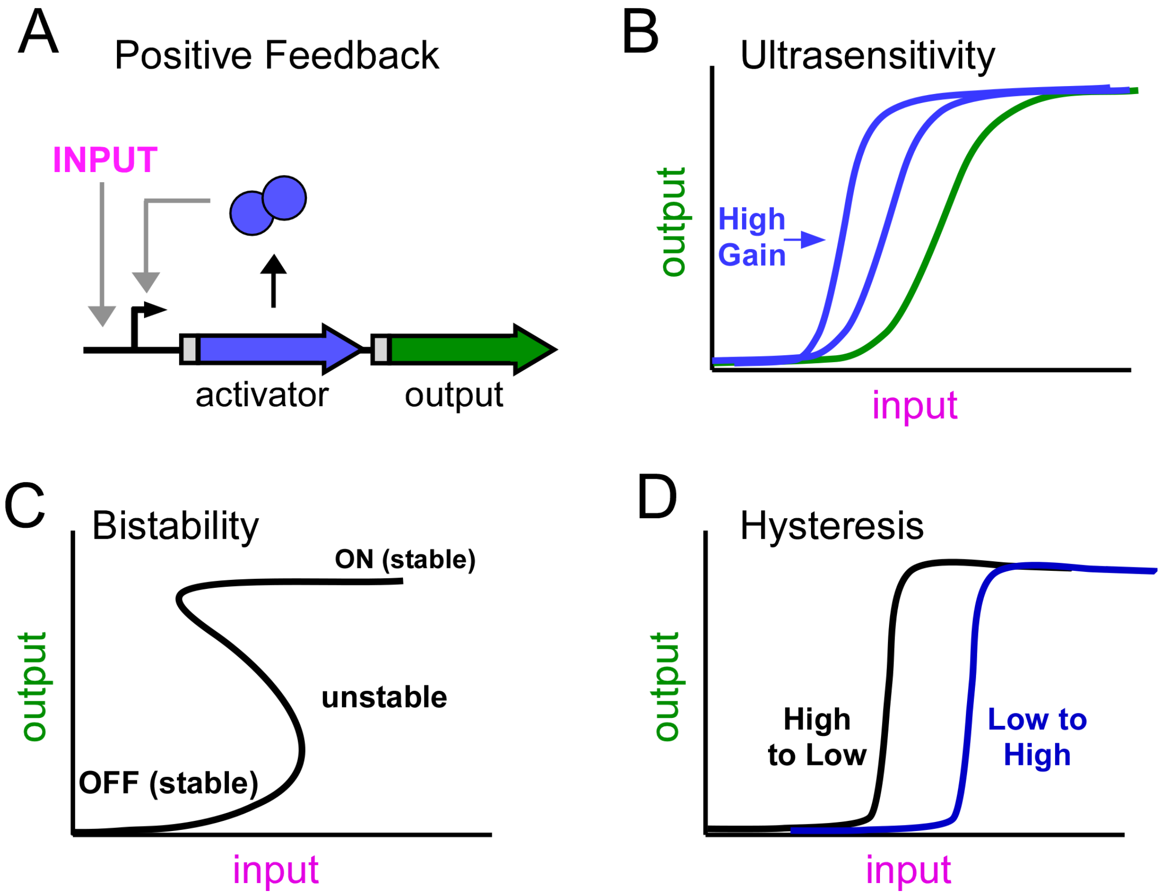
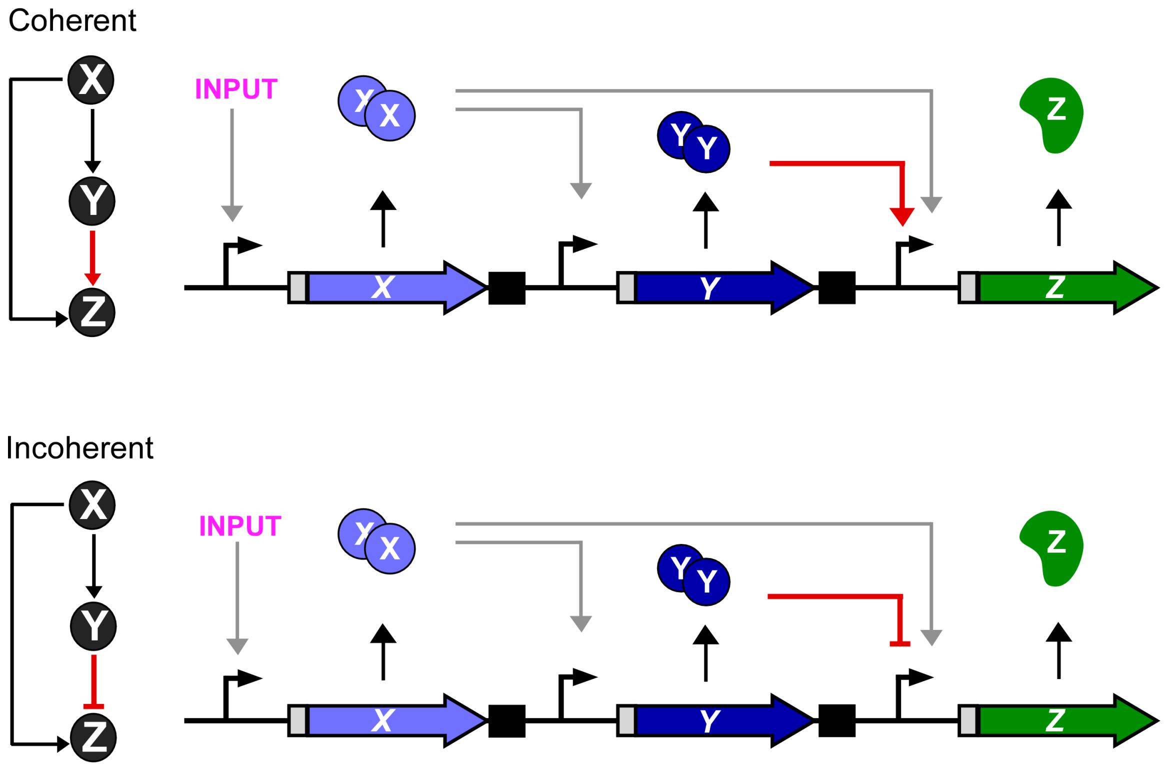


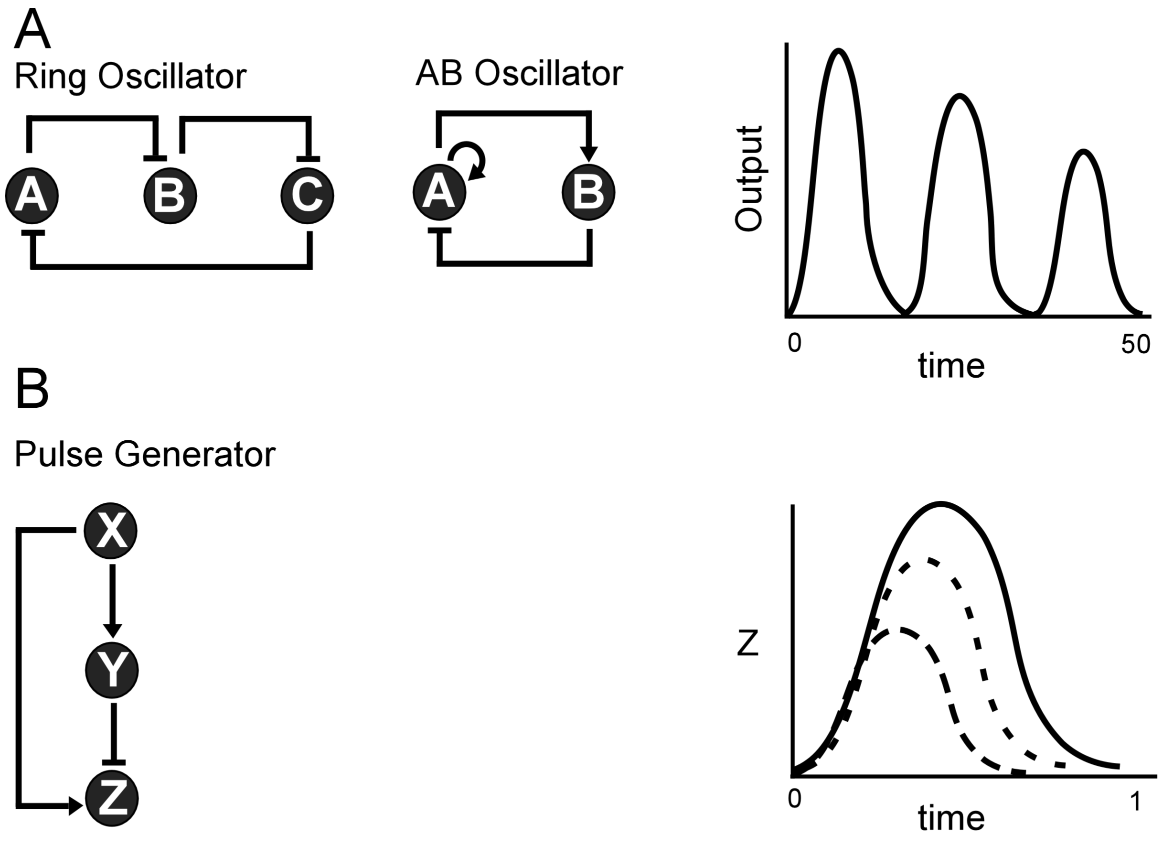
Figure 7. Positive transcriptional feedback.

(A) A self activating protein (blue) is expressed under control of the input. (B) Positive feedback circuits with lower kinetic orders of transcription factor binding and cooperativity result in ultrasensitive responses to inducer. Sensitivity is measured as the slope of the relationship between output and input. This increases from the classically regulated system (green line) to 2 positive feedback systems with increasing kinetic constants of activation (blue lines). (Top) *E.coli* expressing ultrasensitive positive feedback circuits can rest in low, intermediate or high states as a function of input concentration. (C) Positive feedback circuits with very high kinetic orders of activation can achieve bistability. In these circuits, cells rest at low output levels or high output levels but never at intermediate output levels. (D) Hysteresis. When starting at low inducer concentrations and moving higher (blue line) the circuit requires some amount of inducer to switch ON. When starting in the ON state and reducing the concentration of inducer available to the circuit, the switch happens at a significantly lower concentration.



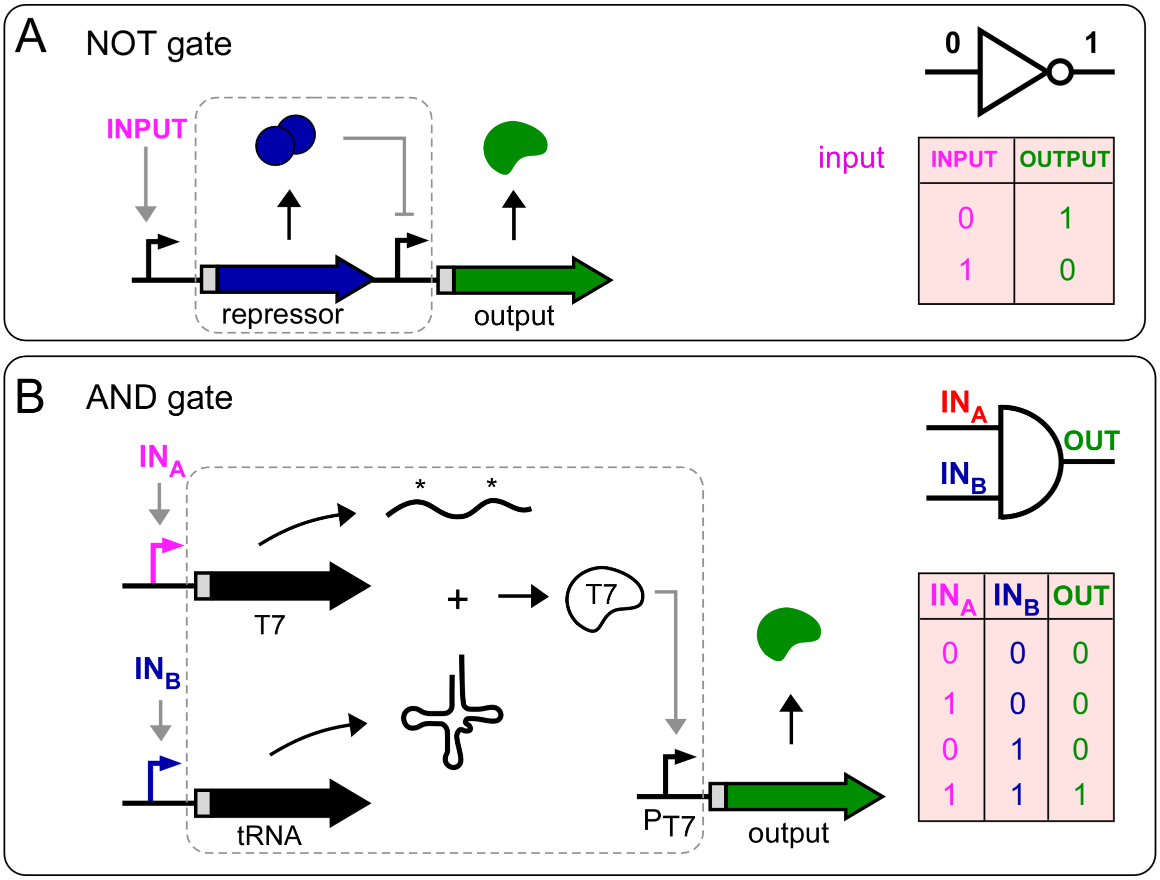
Chapter 5: Figure 8. Feed Forward Loops (FFLs).

FFLs are genetic circuits composed of three proteins, X, Y and Z. X and Y are transcription factors which regulate the expression of Z. The 'feed-forward' connectivity refers to the fact that X also regulates Y. Coherent FFLs result when the regulatory relationship between X and Z is the same as that between X and Y. Incoherent FFLs arise when these two relationships are opposite.



Chapter 5: Figure 9. Dynamic genetic circuits.

(A) Genetic circuits composed of three transcriptional repressors in a closed loop or a self activating protein which also activates its own repressor can cause oscillations in gene expression. (B) Pulse Generator. A type 1 incoherent Feed Forward Loop produces temporal pulses of gene expression. The strength of expression or the kinetic order of repression of the repressor Y can change the duration and amplitude of the pulse (dashed lines).



INPUT	OUTPUT
0	1
1	0

IN_A	IN_B	OUT
0	0	0
1	0	0
0	1	0
1	1	1

Chapter 5: Figure 10. Transcriptional Logic (A) NOT gate. Also known as a genetic inverter, the NOT gate encodes a repressor (blue) under the control of the environmental input (purple dots). The repressor inactivates expression from an otherwise active output promoter. The inverter device (dashed box) comprising the repressor protein and the output promoter is an independent module which can be placed between any input promoter and output gene. The logic of the NOT circuit (upper right) is shown in the truth table (pink). (B) AND gate (dashed box) comprises an untranslatable T7 RNA polymerase mRNA bearing two stop codons (asterisks) in the open reading frame and a suppressor tRNA which incorporates amino acids at the stop codons to allow productive translation. Only when both halves are

transcribed is T7 RNAP produced and does the output promoter become active. Each half of the AND gate can be driven by any inducible promoter, activated by its cognate input signal (blue and red). Adapted from Anderson et al., 2007.

Chapter 6: Kinetic Buffering of Crosstalk between Bacterial Two-Component Sensors

Eli S. Groban[1], Elizabeth J. Clarke[1], Howard Salis[2], Susan M. Miller[1,2], and Christopher A. Voigt[1,2]

[1] Graduate Group in Biophysics

[2] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

Abstract

Two-component systems are a class of sensor that enables bacteria to respond to environmental and cell state signals. The canonical system consists of a membrane-bound sensor histidine kinase that autophosphorylates in response to a signal and transfers the phosphate to an intracellular response regulator. Bacteria typically have dozens of two-component systems. A key question is whether these systems are linear and, if they are, how crosstalk between systems is buffered. Here, we study the EnvZ/OmpR and CpxA/CpxR systems from *Escherichia coli*, which have been shown previously to exhibit slow cross talk *in vitro*. Using *in vitro* radiolabeling and a Rapid Quench Flow System, we experimentally measure ten biochemical parameters capturing the cognate and non-cognate phosphotransfer reactions between the systems. These data are used to parameterize a mathematical model, which is used to predict how crosstalk is affected as different genes are knocked out. It is predicted that significant crosstalk only occurs for the triple mutant: Δ*ompR* Δ*cpxA* Δ*actA-pta*. To test this prediction, all seven combinations of knockouts are made and only this particular strain demonstrates crosstalk, where the *cpxP* promoter is induced 280-fold upon the activation of EnvZ. Further, the behavior of the other knockouts agrees with the model predictions. This model points to a kinetic model of buffering where both the cognate phosphatase activity and competition between regulator proteins for phosphate prevents cross talk *in vivo*.

Introduction

Bacteria use a variety of sensing mechanisms to respond to changes in the environment and the cell state. A ubiquitous motif is the two-component system, where one membrane-bound protein acts as the sensor and another as a signal carrier.[1; 2; 3] Upon receiving a signal, the sensor autophosphorylates at a histidine residue and transfers the phosphate to the signal carrier, a response regulator protein.[1; 2; 4; 5] The response regulator can either affect a larger signaling network by binding to another protein, as in chemotaxis, or it can alter gene expression by binding to DNA.[1; 6; 7; 8; 9] Two-component systems are capable of responding to a wide variety of signals, including temperature, touch, light, metals, dissolved gasses, chemicals, membrane stress, antibiotics, nutrients, pH, and nitrogen sources and are involved in many cellular processes including cell division, motility, pathogenesis, biofilm formation, and cell-cell communication.[9; 10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26]

Two-component systems are remarkably modular – a property that is being increasingly exploited in engineering applications.[27; 28; 29; 30] The N-terminal domain of the sensor kinase, which responds to the extracellular signal, shows the most diversity.[1; 2] However, there is high sequence similarity among response regulator and sensor kinase phosphotransfer domains, in some cases up to 50% identical, and many structural studies show only subtle differences in three-dimensional shape among response regulators in *E. coli*.[31; 32; 33; 34] The conservation of the signaling domains allows domain swapping to be used to rewire pathways to mix-and-match inputs and outputs.[35; 36; 37; 38; 39; 40]

Because of the number of simultaneously expressed systems and the strong conservation in the sequence and structure, it seems plausible that cross reactions would occur frequently between systems.[41] This could either occur as a single kinase phosphorylating multiple response regulators, or conversely, a single regulator could be phosphorylated by multiple kinases. Cross reactions could even be exploited to create a neural network linking inputs and outputs such that a higher level of sensing power is achieved.[42; 43; 44; 45] In contrast, all of the two-component systems could behave linearly with one input being linked to one output. Given the similarity between systems, there would have to be some sort of buffering mechanism for linearity to be preserved.[41; 46; 47]

Previous *in vitro* studies demonstrate some promiscuity among histidine kinases and response regulator proteins. Whiteley *et al* showed that the KinA kinase from *Bacillus subtilis* phosphorylates its target response regulator, Spo0F, with a k_{cat} of 0.083 s^{-1}, and a non-cognate downstream response regulator Spo0A, with a k_{cat} of 1.5 x 10^{-5} s^{-1}.[48] Ishihama and co-workers assayed 25 histidine kinases from *E. coli* against the 34 *E. coli* response regulators.[49] They showed that after a 30 second incubation time, 11 of the 34 response regulators can be phosphorylated by more than one histidine kinase. There are only a few promiscuous kinases, however, so out of 692 possible cross talk pairs, only 3.0% of them showed *in vitro* cross talk. Skerker and Laub assayed the EnvZ, CpxA, and CheA kinases against a panel of all response regulators from *E. coli*.[50] Non-cognate transfer occurs between EnvZ and CxpR, but this is observed at a 60 minute timepoint. This is very slow compared to the cognate transfer, which occurs in 10 seconds. They suggest that the non-cognate transfer is too slow to be relevant *in vivo*.

Several theories have been proposed to explain how two-component systems could buffer *in vivo* the slow crosstalk observed *in vitro*. Savageau proposed that buffering could emerge from the ability for the histidine kinase to both phosphorylate and dephosphorylate the response regulator (a bifunctional interaction).[46] Using a mathematical model, he demonstrated that the phosphatase function could decrease the background phosphorylation of the response regulator, thus reducing cross talk. Laub and co-workers argue that each kinase has a "kinetic preference" for its cognate substrate.[50] They postulate that subtle amino acid differences in the binding interface between the kinase and the response regulator affect the K_m. If the correct response regulator interacts for a longer time with its kinase, this both prevents access to the kinase by the incorrect substrate and drains the kinase of all available phosphate, resulting in phosphorylation of only the correct response regulator. To support this hypothesis, they made small amino acid changes to the kinase at the binding interface, which resulted in a shift in kinetic preference that altered specificity.[42]

In this manuscript, we characterize the interactions between the *E. coli* EnvZ/OmpR and CpxA/CpxR two-component systems.[4; 8; 51] These systems respond to many extracellular signals and are often labeled as osmosensors.[8; 52] They also co-regulate many cellular responses, including flagella assembly, pathogenesis, outer membrane porins, and biofilm formation.[8; 11; 15; 53] OmpR and CpxR often bind to different operators within the same promoter.[54] The phosphotransfer domains share the most similarity amongst *E. coli* two-component systems, sharing 31% amino acid identity for the kinase domain of the sensor histidine kinase and 50% identity between the receiver domain of the response regulators.[55] It has been shown previously that EnvZ can phosphorylate CpxR *in vitro*, albeit at a much slower rate than OmpR.[49; 50]

Both EnvZ and CpxA are bifunctional histidine kinases, containing both kinase and strong phosphatase activities.[56; 57; 58] Interestingly, the OmpR and CpxR response regulator proteins appear to have opposite relationships with their cognate kinases. In the absence of EnvZ, or when EnvZ is not stimulated, OmpR~P levels are very low.[8; 35] EnvZ then acts as a kinase when stimulated. In contrast, CpxR exists in an active phosphorylated state in the absence of CpxA and maintains a basal level of activity in the presence of the CpxA kinase.[59; 60]

We perform a comprehensive kinetic study of the interactions between the EnvZ/OmpR and CpxA/CpxR two-component systems. First, we purified the kinases EnvZ and CpxA and their response regulators OmpR and CpxR. We phosphorylated either the kinase or the response regulator using radiolabeled phosphate and monitored phosphorylation kinetics *in vitro* (Figure 1). Moreover, we also synthesized radiolabeled (^{32}P)-acetylphosphate and used this to phosphorylate the OmpR and CpxR proteins in the absence of kinase to determine autophosphorylation rates. Fully phosphorylated response regulator proteins were then exposed to unlabeled kinase to measure phosphatase activity. After obtaining rate constants for ten reactions, we parameterized a mathematical model and used this model to predict combinations of gene knock outs to make that could induce cross talk *in vivo*. To test the prediction, we constructed seven knock out strains and used promoter-GFP fusions to measure cross talk *in vivo*. Significant cross talk is only observed for a triple knock out predicted by the model.

Results

In vitro phosphotransfer experiments

We measured ten kinetic constants (k_1 – k_{10}, Figure 2) *in vitro*, using purified proteins, radiolabeled phosphate, and a Rapid Quench Flow System for reactions that occur on fast time scales (Materials and Methods). For the kinases, only the cytoplasmic portion was purified as data from previous studies show this portion is active *in vitro*.[8; 49; 50; 61] Cognate (k_1 and k_6) and non-cognate (k_2 and k_7) kinase rates were measured by radiolabeling the histidine kinase and observing phosphotransfer when response regulator is introduced. Purified histidine kinase was labeled by exposing it to ^{32}P-γ-ATP for thirty minutes, allowing it to autophosphorylate. In order to observe single turnover, the phosphorylated kinase was washed to remove all free ATP before adding a 10-fold excess of response regulator protein and allowing the reaction to proceed to completion. A Rapid Quench Flow System was used for reactions that are complete in less than a second. The fraction of phosphorylated response regulator is calculated as a ratio of the amount of phosphorylated protein at a particular time point divided by the highest amount of phosphorylated protein during the course of the reaction.

To measure the cognate (k_3 and k_8) and non-cognate (k_4 and k_9) phosphotase rates, the response regulator was labeled and then mixed with histidine kinase. Previous studies demonstrated that response regulators can be phosphorylated by using acetyl phosphate *in vitro*.[62; 63] As (^{32}P)-acetylphosphate is not commercially available, this small molecule was synthesized from radiolabeled (^{32}P)-phosphate and acetic anhydride.[63] Response regulator proteins were exposed to (^{32}P)-acetylphosphate for one hour and then mixed with histidine kinase. The loss of phosphate from the response regulator is used to calculate the phosphatase rate. These experiments provide data to determine rates of kinase dependent response regulator dephosphorylation. In addition, the ^{32}P actyl phosphate is used to determine the autophosphorylation rates of the response regulators (k_5 and k_{10}). The kinetic constants k_1 to k_{10} were determined by fitting the experimental phosphotransfer curves to the solution of a system of ordinary differential equations (ODEs). For each *in vitro* phosphotransfer assay, a linear ODE is derived that describes the dynamics of the phosphorylation and de-phosphorylation of the response regulator. An analytical algebraic solution to this equation relates the fraction of phosphorylated response regulator to time and the kinetic constants of each reaction in the assay. This procedure is repeated for all ten phosphotransfer assays, resulting in a system of algebraic equations that describe the dynamics of phosphorylated OmpR and CpxR under each reaction condition. Then, using Matlab (Mathworks), we identify the values of the kinetic constants that best reproduce the phosphotransfer curves by minimizing the least-squares residual between the algebraic solution and the experimental data (Materials and Methods). Overall, data fitting required ten different differential equations with ten different parameters, k_1 – k_{10}. However, constants k_3 – k_5 and k_8 – k_{10} appear in multiple equations. The solutions to this fitting procedure are shown as the curves in Figure 2.

EnvZ and CpxA phosphorylate their cognate response regulator proteins very rapidly. The EnvZ/OmpR reaction reaches completion in less than 40 ms, with a forward rate constant of $k_1 = 102\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ (Figure 2A). CpxA phosphorylates CpxR rapidly as well, although with a lower rate constant of $k_6 = 29\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ (Figure 2B). These numbers are faster than estimated in previous studies that could not achieve sub-second resolution.[4; 49; 50]

Non-cognate phosphorylation is much slower. The rate constants for OmpR phosphorylation by CpxA, and CpxR phosphorylation by EnvZ are $k_2 = 0.003\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ and $k_7 = 0.003\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ (Figure 2) which agrees qualitatively with previous studies.[49; 50] This represents a 32,700-fold difference in the rate of cognate versus non-cognate phosphorylation for OmpR and a 9200-fold difference for CpxR.

The cognate and non-cognate phosphatase reactions were also measured. EnvZ dephosphorylates OmpR with a rate constant of $k_3 = 0.003\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ (Figure 2A), which agrees with previous studies.[64] CpxA dephosphorylates CpxR on the same timescale, with a slightly faster rate of $k_8 = 0.003\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ (Figure 2B).[49] However, neither kinase can efficiently dephosphorylate its non-cognate response regulator ($k_4 = 0.0001\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$ and $k_9 = 0.0002\ \mathrm{s}^{-1}\ \mathrm{M}^{-1}$) (Figure 2). It is noteworthy that the cognate dephosphorylation rates (k_3 and k_8) are about the same as the non-cognate phosphorylation rates (k_2 and k_7). This means that the phosphatase activity is just strong enough to remove the phosphate that accumulates from cross reactions, as predicted by Savageau.[46]

The rate of autophosphorylation by acetyl phosphate for OmpR and CpxR is the slowest contribution to response regulator phosphorylation with rate constants of $k_5 = 0.001$ s^{-1} M^{-1} and $k_{10} = 0.\ 001$ s^{-1} M^{-1} respectively (Figure 2). This agrees with published values for OmpR.[62] In the absence of any kinase, both OmpR~P and CpxR~P have an inherent autophosphatase rate that is negligible on the timescales observed. This is slower than previously reported for OmpR~P.[56; 62]

A mathematical model of the EnvZ/OmpR and CpxA/CpxR systems

A mathematical model is developed to determine the global network dynamics that emerge from the set of kinetic parameters obtained *in vitro*. The model focuses on how the steady-state level of CpxR~P is affected by phosphate flux from EnvZ. The model is used to quantify how crosstalk is affected by knocking out different components of the system. The model consists of a set of four differential equations:

$$\boxed{\begin{aligned}
\frac{dC_{\mathrm{OmpR\sim P}}}{dt} &= C_{\mathrm{OmpR}}\left[k_1 C_{\mathrm{EnvZ\sim P}} + k_2 C_{\mathrm{CpxA\sim P}} + k_5\right] - C_{\mathrm{OmpR\sim P}}\left[k_3 C_{\mathrm{EnvZ}} + k_4 C_{\mathrm{CpxA}}\right]\\
\frac{dC_{\mathrm{CpxR\sim P}}}{dt} &= C_{\mathrm{CpxR}}\left[k_6 C_{\mathrm{CpxA\sim P}} + k_7 C_{\mathrm{EnvZ\sim P}} + k_{10}\right] - C_{\mathrm{CpxR\sim P}}\left[k_8 C_{\mathrm{CpxA}} + k_9 C_{\mathrm{EnvZ}}\right]\\
\frac{dC_{\mathrm{EnvZ\sim P}}}{dt} &= \mathrm{I}_{\mathrm{EnvZ}} C_{\mathrm{EnvZ}} - C_{\mathrm{EnvZ\sim P}}\left[k_1 C_{\mathrm{OmpR}} + k_7 C_{\mathrm{CpxR}}\right]\\
\frac{dC_{\mathrm{CpxA\sim P}}}{dt} &= \mathrm{I}_{\mathrm{CpxA}} C_{\mathrm{CpxA}} - C_{\mathrm{CpxA\sim P}}\left[k_6 C_{\mathrm{CpxR}} + k_2 C_{\mathrm{OmpR}}\right]
\end{aligned}}$$

and four mole conservation relations:

$$\boxed{\begin{aligned}
C_{\mathrm{OmpR,TOT}} &= C_{\mathrm{OmpR\sim P}} + C_{\mathrm{OmpR}}\\
C_{\mathrm{CpxR,TOT}} &= C_{\mathrm{CpxR\sim P}} + C_{\mathrm{CpxR}}\\
C_{\mathrm{EnvZ,TOT}} &= C_{\mathrm{EnvZ\sim P}} + C_{\mathrm{EnvZ}}\\
C_{\mathrm{CpxA,TOT}} &= C_{\mathrm{CpxA\sim P}} + C_{\mathrm{CpxA}}
\end{aligned}}$$

where C represents the concentration of each species. $C_{OmpR,TOT}$ and $C_{CpxR,TOT}$ are set to 5 μM and $C_{EnvZ,TOT}$ and $C_{CpxA,TOT}$ are set to 0.5 μM.[65] I_{EnvZ} and I_{CpxA} are parameters that are set to represent the phosphorylation state of the kinase.

The output from our model provides an estimate for the fraction of CpxR~P response regulator protein *in vivo*. In Figure 3, we show the results of eight different simulations based on strains with various components of the two pathways knocked out. To knock out a protein in a given simulation, we set the corresponding kinetic constants to zero. In each simulation we either knock out EnvZ, represented by a gray dashed line, or insert a fully active EnvZ, represented by a black line.

For the wild-type (WT) system, the model reproduces the observation that CpxR~P is insensitive to phosphate flux from EnvZ (Figure 3). When the CpxA protein is removed ($k_2=k_4=k_6=k_8=0$), the system responds in an unexpected way. We were expecting that this would enable crosstalk from EnvZ. However, it produced an interesting inverse phenotype, where phosphate flux from EnvZ decreases the accumulation of CpxR~P (Δ*cpxA*, Figure 3). This occurs because, when present, OmpR acts as the sink for all of the phosphate from EnvZ. This leaves only the very weak phosphatase activity towards CpxR~P, which is very slow *in vitro*. This effect was confirmed *in vivo* (next section).

When OmpR is also removed from the model ($k_1=k_2=k_3=k_4=k_5=0$), this eliminates the inverse effect, and a small amount of crosstalk is observed both in the presence and absence of CpxA (Δ*ompR* Δ*cpxA* and Δ*ompR*, Figure 3). This crosstalk occurs on a much slower timescale than what is observed for cognate transfer between EnvZ and OmpR (Figure 2A and 3A). While we see crosstalk for these perturbations to the model, it is not observed *in vivo* (next section).

Finally, we measured the effect of disrupting the autophosphorylation of OmpR and CpxR. We hypothesized that crosstalk could be masked by the high background of CpxR~P that occurs when CpxA is removed. It has been shown previously that the autophosphylation of response regulators can be disrupted by knocking out acetate kinase A and phosphotransacetylase (Δ*ackA-pta*).[60; 66] This is incorporated into the model by setting $k_5 = k_{10} = 0$.

The disruption of autophosphorylation has no affect on the otherwise intact system (Δ*actA-pta*, Figure 3). It also has little effect on the mutant where OmpR is removed (Δ*ompR* Δ*actA-pta*). In both cases, this is because the CpxR~P level is controlled by the kinase and phosphatase activities of the CpxA kinase. The Δ*actA-pta* Δ*cpxA* simulation shows a slight amount of crosstalk in the presence of active EnvZ yet, in its absence, CpxR is not phosphorylated. This makes sense as all possible sources of phosphate (CpxA, EnvZ, and actyl phosphate) are absent from the system. This small amount of crosstalk is observed *in vivo* (next section).

When OmpR, CpxA, and autophosphorylation activities are all removed, a significant amount of crosstalk is observed (Δ*ackA-pta* Δ*cpxA* Δ*ompR*, Figure 3). By knocking out these pathways, the background concentration of CpxR~P is reduced to zero. With acetyl phosphate synthesis and CpxA knocked out, EnvZ is the only source of phosphate for CpxR. Moreover, both OmpR, which is a competitor for phosphate, and CpxA, which serves as a CpxR~P phosphatase are removed from the system. We sought to confirm this result *in vivo*.

Demonstration of crosstalk *in vivo*

We evaluated different single, double, and triple knock-outs suggested by our model. The model predicts that the system is incredibly robust to cross talk and that only a triple knock out should allow us to observe cross talk *in vivo* (Figure 3B). The triple knock out removes two modes that the cell could use to prevent cross talk, cognate histidine kinase phosphatase activity and competition for phosphorylated kinase between cognate and non-cognate response regulator. Moreover, it also prevents CpxR autoactivation, allowing the observation of cross talk in the absence of the CpxA kinase.

We monitor the *in vivo* phosphorylation state of OmpR and CpxR by following the activity of the *ompC* and *cpxP* promoters, respectively (Figures 4A and 4B). These promoters were chosen because footprint experiments have demonstrated that the response regulator binds to specific sequences upstream of the -35 site.[67; 68] Moreover, these promoters respond to conditions known to activate the systems as shown by both β-galactosidase and GFP reporter systems.[69; 70; 71] We selectively activate the EnvZ/OmpR pathway in order to determine whether *in vivo* cross talk occurs between it and CpxA/CpxR. As there are no known ligands for the EnvZ kinase, we used a second generation version of the hybrid kinase designed by Utsumi and Inouye, that combines the transmembrane aspartate receptor (TAR) with the kinase domain of the EnvZ kinase (TAZ) (Figures 4C and 4D).[57] This hybrid kinase serves to selectively activate the EnvZ kinase domain, and phosphorylate its cognate response regulator OmpR in the presence of the small molecule aspartate. We activate the EnvZ kinase using the TAZ construct and monitor the output levels of both the EnvZ/OmpR and CpxA/CpxR system using the promoters *ompC* and *cpxP* to drive green fluorescent protein (GFP) expression, which is measured at the single-cell level using flow cytometry (Figures 4D and 4G).

TAZ phosphorylates the OmpR response regulator in the presence, but not the absence, of 5 mM aspartate *in vivo* in the wild type system (Figure 4E and Figure 5, first panel). Induction of TAZ via addition of 5 mM aspartate leads to a 34.7 (+/- 1.9) fold increase in induction of fluorescence by the *ompC* promoter showing that the presence of aspartate and TAZ provides an active EnvZ kinase. This activation is TAZ dependent, as a control strain lacking TAZ shows no activity at the *ompC* promoter (Figure 4F). Δ*ompR* does not respond to TAZ showing that the *ompC* promoter is specific to OmpR~P. Our use of GFP as a reporter for system activity, combined with flow cytometry, allowed us to conduct single cell measurements. From this we see that cells containing the *ompC* promoter GFP fusions produce a single distribution in all cases (Figure 4E).

In the wild type system, active EnvZ does not affect the levels of CpxR~P response regulator *in vivo* (Figure 4G-I). By monitoring the abundance of phosphorylated CpxR via the *cpxP* promoter and activating TAZ, we are able to measure the presence or absence of cross talk between EnvZ and CpxR. We are unable to see a significant difference in *cpxP* activity in the presence of active EnvZ.

We then constructed seven knockout mutants containing all possible combinations of Δ*cpxA*, Δ*ompR*, and Δ*ackA-pta* (Methods). In addition, EnvZ is also knocked out and TAZ is introduced on a plasmid. For each of the knockouts, the activity of the promoters is compared for the Δ*envZ* strain (-) and the addition of the TAZ and 5 mM aspartate (+) (Figure 5).

The knockout Δ*cpxA* causes the level of CpxR~P to increase (Figure 5). In the absence of the CpxA kinase, CpxR autophosphorylates using acetyl phosphate as a substrate and reaches a higher steady state. Notably, the experiments recover the decrease in activity from the addition of active EnvZ as observed in the model (Figure 3). The single knockout Δ*ompR* behaves similarly to WT. The double knockout Δ*cpxA* Δ*ompR* strongly activates the *cpxP* promoter and is insensitive to active EnvZ.

The effect of disrupting the acetyl phosphate synthesis pathway was then explored. The Δ*ackA-pta* and Δ*ackA-pta* Δ*ompR* strains display similar behavior to WT. In contrast, the Δ*ackA-pta* Δ*cpxA* strain strongly decreases the amount of CpxR~P and the activity of the *cpxP* reporter. However, there is a small amount of crosstalk observed, where active EnvZ is able to phosphorylate CpxR and induce the *cpxP* promoter 3.6 (+/- 0.2) fold. This is consistent with the prediction of the model (Figure 3).

A Δ*ackA-pta* Δ*cpxA* Δ*ompR* triple knock out shows EnvZ dependent phosphorylation of CpxR *in vivo* (Figure 5, panel 9). Inducing TAZ in this system causes a 370.4 (+/- 79.8) fold increase in *cpxP* activity. This system lacks two crucial mechanisms to prevent cross talk. First, it lacks the phosphatase activity of CpxA. Second, it does not have the OmpR protein, allowing CpxR access to the active EnvZ kinase. These two buffering mechanisms are quite effective. Therefore, cross talk is highly unlikely to be relevant *in vivo*. In the wild type system, CpxA serves as a phosphatase and OmpR quickly removes phosphate from active EnvZ. It is only by removing all possible mechanisms for insulation that we see a potent activation of CpxR by EnvZ *in vivo*.

Discussion

The field of Systems Biology often seeks to model global dynamical behaviors using sets of parameters obtained individually *in vitro*. An open question is whether the conditions *in vitro* sufficiently reflect the cellular environment for the parameters to be relevant. Here, we have completely characterized the phosphotransfer pathway of two interacting two-component systems. A set of kinetic rate constants obtained using purified protein *in vitro* is used to parameterize a mathematical model, which correctly predicts the higher-order dynamical behavior of the system *in vivo*. The effect that the knockouts have on the signaling network could not be predicted from the individual parameters alone.

Here, we started with two systems that share a high degree of sequence similarity and for which crosstalk had been previously demonstrated *in vitro*.[49; 50] Remarkably, it required three knockouts to induce crosstalk, which demonstrates the degree to which two-component systems are buffered from crosstalk. This is consistent with other recent observations, all of which point to phosphotransfer being linear in the native host.[47] There are other mechanisms by which crosstalk can occur, including the inclusion of additional kinase domains in the sensor kinase which interact with other response regulators.[1; 2; 41] There are also examples of cytoplasmic adaptor proteins that integrate multiple sensors and then interact with a downstream response regulator. An example of this is the Spo0F protein in *B. subtilis*, which integrates multiple sensors, and activates the Spo0A response regulator through the histidine phosphotransferase Spo0B.[72; 73] However, unless there are additional domains or proteins involved in phosphotransfer, it would appear that the phosphotransfer that occurs in canonical two-component systems is remarkably linear.

Integrating two-component systems may play a critical role for bacteria to identify an environment. To this end, there are many promoters that contain multiple operators that bind different response regulators.[74; 75; 76] There are also genetic circuits that act as logic to integrate multiple inputs. An emerging conclusion is that most of the signal integration between two-component systems occurs on the transcriptional level.

The remarkable degree of buffering occurs due to a combination of kinetic interactions. We find that the bifunctional phosphatase activities of the sensors are just fast enough to remove the phosphates that get transferred via non-cognate interactions. This supports the theory put forward by Savageau and co-workers.[46] However, when this interaction is removed, crosstalk does not occur. Also, there is evidence that the response regulators act as phosphate sinks by preferentially interacting with their cognate sensor. Previous studies suggest that only 0.02% of EnvZ is phosphorylated *in vitro* and OmpR preferentially interacts with phosphorylated EnvZ, leaving only the unphosphorylated EnvZ to act as a phosphatase to interact with non-cognate response regulators.[49; 64; 77] Our model does not include sequestration and the *in vitro* experiments are with excess substrate, making it impossible to determine a K_m. Still, we see evidence of this effect in the inverse induction when CpxA is knocked out. Also, the removal of OmpR is critical for significant crosstalk to occur.

Programmable sensors are a key component for engineering bacteria.[78] The modularity of two-component systems makes them an intriguing target for engineering, where input signals can be rapidly connected to control different pathways through domain swapping.[35; 36; 37; 38; 39; 40] In addition, two-component systems can generally be moved between species. This enables access to a diversity of sensing functions present in bacteria. However, it has been observed that after transfer, two-component systems can exhibit cross reactions to host systems, which can have deleterious effects.[50] Therefore, it will be critical to understand how natural systems buffer crosstalk such that these interactions can be engineered *de novo*.

Methods

Determination of in vitro kinetic constants: The values of the kinetic constants k_1 to k_{10} were determined by minimizing the difference between the solution of a system of ten linear ordinary differential equations (ODE) and the ten *in vitro* phosphotransfer curves. For each phosphotransfer reaction, we determined the analytical solution for the ODE describing the fraction of phosphorylated response regulator over time, using pseudo-first order kinetics with $C_{\mathrm{EnvZ}_{\mathrm{TOT}}} = C_{\mathrm{CpxA}_{\mathrm{TOT}}} = 0.5\ \mu\mathrm{M}$.

For OmpR phosphotransfer, with $\phi = C_{\mathrm{OmpR\sim P}} / C_{\mathrm{OmpR}_{\mathrm{TOT}}}$, these equations are:

1. cognate kinase: $\phi_1(t) = 1 - \exp\left(-k_1 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}\ t\right)$
2. non-cognate kinase: $\phi_2(t) = 1 - \exp\left(-k_2 C_{\mathrm{CpxA}_{\mathrm{TOT}}}\ t\right)$
3. cognate phosphatase: $\phi_3(t) = \dfrac{k_5 + \exp\left(-t\left(k_5 + k_3 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}\right)\right)k_3 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}}{k_5 + k_3 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}}$
4. non-cognate phosphatase: $\phi_4(t) = \dfrac{k_5 + \exp\left(-t\left(k_5 + k_4 C_{\mathrm{CpxA}_{\mathrm{TOT}}}\right)\right)k_4 C_{\mathrm{CpxA}_{\mathrm{TOT}}}}{k_5 + k_4 C_{\mathrm{CpxA}_{\mathrm{TOT}}}}$
5. acetyl phosphate transfer only: $\phi_5(t) = 1 - \exp\left(-k_5\ t\right)$

For CpxR phosphotransfer, with $\chi = C_{\mathrm{CpxR\sim P}} / C_{\mathrm{CpxR}_{\mathrm{TOT}}}$, these equations are:

6. cognate kinase: $\chi_6(t) = 1 - \exp\left(-k_6 C_{\mathrm{CpxA}_{\mathrm{TOT}}}\ t\right)$
7. non-cognate kinase: $\chi_7(t) = 1 - \exp\left(-k_7 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}\ t\right)$
8. cognate phosphatase: $\chi_8(t) = \dfrac{k_{10} + \exp\left(-t\left(k_{10} + k_8 C_{\mathrm{CpxA}_{\mathrm{TOT}}}\right)\right)k_8 C_{\mathrm{CpxA}_{\mathrm{TOT}}}}{k_{10} + k_8 C_{\mathrm{CpxA}_{\mathrm{TOT}}}}$
9. non-cognate phosphatase: $\chi_9(t) = \dfrac{k_{10} + \exp\left(-t\left(k_{10} + k_9 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}\right)\right)k_9 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}}{k_{10} + k_9 C_{\mathrm{EnvZ}_{\mathrm{TOT}}}}$
10. acetyl phosphate transfer only: $\chi_{10}(t) = 1 - \exp\left(-k_{10}\ t\right)$

A Levenberg-Marquardt algorithm identifies the best fit values of the kinetic constants by minimizing the residual, $R = \sum_{i=1}^{5}\left(\phi_i - \phi_i^{\exp}\right)^2 + \sum_{i=6}^{10}\left(\chi_i - \chi_i^{\exp}\right)^2$, where $\phi^{\exp}$ and $\chi^{\exp}$ are the five experimentally measured phosphotransfer curves for OmpR and CpxR, respectively.

Media and solutions: All cloning was performed in 2YT liquid media (31 g 2YT per 1 liter, Teknova). Selective agar plates were prepared for antibiotic selection using 25 g Luria Bertani Broth, Miller Formulation (Difco) and 18 g Bacto Agar

(Difco) per 1 liter plus appropriate antibiotics (Amp: 100 μg/mL, Kan: 25 μg/mL, Cm: 25 μg/mL). Protein purification buffers: Buffer A (100 mM NaCl, 25 mM Tris-HCl, pH 7.4), Buffer B (Buffer A + 10 mM imidazole), Buffer C (Buffer A + 300 mM imidazole), Storage Buffer (50 mM TRIS-HCL, pH 7.4). Reaction buffer: 50 mM Tris-HCl, 50 mM KCl, 10 mM $MgCl_2$, pH 7.4. Quench buffer: 75 mM Tris-HCl, pH 8.0, 0.6% SDS, 1.2 mM EDTA, 0.025% Bromothymol Blue (BioRad), 15% glycerol. *In vivo* experiments were performed in M9 media (48 mM Na_2HPO_4, 22 mM KH_2PO_4, 16 mM NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 8.6 mM NaCl, pH 7.4) supplemented with 0.4% glucose. 50 mM aspartate (Sigma) was diluted into M9 media with glucose and equilibrated to pH 7.4 before using.

Protein Purification: The cytoplasmic portions of the EnvZ and CpxA kinase and the full OmpR and CpxR proteins were cloned into an in house expression vector in behind a pBAD promoter and prior to a 6X histidine affinity tag. These constructs were transformed into OneShot DH10B cells from Invitrogen and grown for 12 – 16 hours on ampicillin selective LB/Agar. Single colonies were used to inoculate one liter of 2YT media (Teknova) in a 2.8 L baffled flask which was grown at 37^o C until OD_{600} between 0.3 and 0.4 at which point the cultures were moved to 30^o C. At OD_{600} = 0.5, protein expression was induced with 10 mM arabinose and cultures were grown 12 – 16 hours. Cells were pelleted by spinning at 3,738 x g for 20 minutes and resuspended in 25 mL of Buffer A. After addition of lysozyme, cells were frozen at -20 degrees overnight, thawed, sonicated using a Sonic Dismembrator Model 500 (Fisher Scientific) four times for 30 seconds each with 1 second intervals at 25% amplitude. Lysate was cleared by spinning at 19647 x g for 30 minutes after which it was exposed to a 0.8 uM filter (NALGENE) and incubated with 1 mL of TALON resin (Clontech) at 4 degrees with gentle shaking for 1 – 2 hours. Lysate plus resin was poured into a 50 mL column (BioRad) and was allowed to settle. Column was washed with 25 mL of Buffer A, 25 mL of buffer B, and protein was eluted with 10 mL of a mixture of one half Buffer A and one half Buffer C. Elution was dialyzed into Storage buffer and protein was concentrated using an Amicon Ultra-4, Ultracel – 10K centrifugal filter device (Millipore). Final protein concentration was obtained from the absorbance reading at 280 nm and the extinction coefficient calculated using the method of Gill and von Hippel.[79] Glycerol was added to a final concentration of 18.75% and samples were stored at -80 until use.

Kinase phosphorylation: Kinase was diluted to a final concentration of 2 μM in 200 μL of reaction buffer supplemented with 100 μCi of [γ-^{32}P]-ATP (~6000 Ci/mmol, MP Biomedicals) and the reaction was incubated at 37^{o} C for 40 minutes, 10 minutes past the saturation point of the kinase autophosphorylation (data not shown). The protein solution was then diluted to 600 μL using cold reaction buffer and moved to a Microcon Ultracel YM-10 (Millipore). The solution was spun through the filter at 14,000 x g for 20 minutes, after which 400 μL of cold reaction buffer was added. This was repeated 3 times, with the final spin 22 minutes long. Kinase was diluted to 2 μM with cold reaction buffer.

Non-cognate response regulator phosphorylation: Phosphorylated kinase was diluted to 1 μM into a solution containing 10 μM of the non-cognate response regulator protein and 9 μL was removed immediately as the zero point in the reaction. 9 μL aliquots were removed at specified time points, mixed with 9 μL of 2X SDS loading buffer, and kept on ice until all reactions were complete. 15 μL samples were then loaded into 12 well 12% TRIS-Glycine SDS PAGE gels (Lonza) without boiling and run for 1 hour and 10 minutes at 20 milliamps per gel (Bio-Rad). Gels were dried using a Model 583 Gel Dryer (Bio-Rad) and exposed to a Kodak Storage Phosphor Screen (Amersham Biosciences) for 12 – 16 hours before they were imaged using a Typhoon 9400 imaging system (Amersham Biosciences). Densitometry analysis was performed using the free software ImageJ.[80] Further analysis was performed using Microsoft Excel and Matlab.

Cognate response regulator phosphorylation: Phosphorylated kinase was diluted to a final concentration of 2 μM in reaction buffer. Response regulator was diluted to 20 μM in reaction buffer. ~17.5 μL each of kinase and response regulator were mixed and quickly exposed to 82.5 μL of quench buffer after a specified amount of time using a Rapid Quench Flow System (KinTek Corporation). Samples were kept on ice until 15 μL was removed for electrophoreses.

Synthesis of 32-P acetyl phosphate: Synthesis method adapted from the method of McCleary and Stock.[63] 0.240 mL pyridine, 0.063 mL 2M K_2HPO_4, and 0.500 mL of carrier free (^{32}P)-orthophosphate (MP Biomedicals) were combined in a 15 mL FALCON tube and placed on ice for at least 10 minutes. Over the next two minutes

0.0275 mL acetic anhydride was added while mixing on ice. 0.108 mL of 4N LiOH was then added and reaction was mixed for 3 more minutes on ice before addition of 5.75 mL of cold 100% ethanol. Reaction was left to incubate on ice for 1 hour to precipitate product. Product was isolated by spinning at 2,851 x g for 10 minutes and discarding supernatant, followed by two more washes with with 6.0 mL of 100% ethanol. White precipitate was resuspended in 0.75 mL water and 0.35 mL of cold ethanol was added. After a 15 minute incubation on ice mixture was spun at 2,851 x g for 10 minutes and supernatant was collected while precipitate containing impurities was discarded. 5.0 mL of cold ethanol was added to precipitate acetyl phosphate and precipitate was collected by spinning at 2,851 x g for 10 minutes. Purified acetyl phosphate was resuspended in 0.200 mL of 100 mM Tris-HCl, pH 7.0. Acetyl phosphate was stored at -20^o C and used within one week of synthesis. Synthesis method refined and optimized using K_2HPO_4 and comparing to a standard solution of acetyl phosphate (Sigma Aldrich). NMR spectra of synthesized product and purchased mono-acetylated product were identical. NMR spectra were not taken of radioactive product for safety reasons.

Kinase independent response regulator phosphorylation: Response regulator protein was diluted to a concentration of 20 μM in 90 uL of reaction buffer and 10 μL of acetyl phosphate and incubated at 37^o C. 9 μL aliquots were removed at specified timepoints and mixed with 9 μL of SDS loading buffer, 15 μL of which was electrophoresed and imaged in the same fashion as previously described.

Response regulator dephosphorylation: 20 μM response regulator protein was incubated in 90 μL of reaction buffer plus 10 μL of ^{32}P acetyl phosphate for 40 minutes at 37^{o} C. Afterwards, kinase was added to a final concentration of 2 μM with 1 mM ADP and 9 μL aliquots were removed at specified time points, mixed with 9 μL of SDS loading buffer, and placed on ice. 15 μL was removed and electrophoresed in the same method as described previously.

Genomic knock outs: Strain BW28357 was obtained from the Coli Genetic Stock Center at Yale (CGSG) and all genes were knocked out according to the method of Datsenko and Wanner.[81] After confirming each knock out with genomic sequencing, P1 phage was used to transfer each knock out to a recombinase free strain of BW28357. After knocking out the gene of interest we removed the kanamycin antibiotic resistance marker using plasmid pcp20 according to the method of Datsenko and Wanner. All strains were confirmed to be sensitive to kanamycin.

Plasmid construction: The *ompC* promoter region was amplified from *E. coli* genomic DNA using primers gcggccgcATGAAAAGTGTGTAAAGAAGGGT and ggatccGTTATTAACCCTCTGTTATATGCCTTTATTTGC. It was subsequently cut with NotI and BamHI (New England Biolabs) before ligating into plasmid pAC581 immediately prior to GFP Mut3 v.JCA (courtesy of Professor John C. Anderson at the University of California, Berkeley). The *cpxP* promoter was prepared in a similar fashion using primers GCGGCCGCTAATAGGGAAGTCAGCTCTCGGTCATC and GGATCCTTCAGCAGCGTGGCTTAATGAACTGACTGC. Both constructs were shown to be highly selective reporters for the OmpR~P and CpxR~P proteins respectively (data not shown). The TAZ construct was contained on plasmid pTJ003 (courtesy of Masayori Inouye).

Fluorescent Measurements and Cytometry: Cells were transformed with either pAC581-PompC-GFP, pAC581-PcpxP-GFP, or either plasmid and pTJ003 and grown for 12 – 16 hours at 37° C on selective LB/Agar. Control cells containing no plasmid, used to normalize cell fluorescence, were grown for 12 – 16 hours at 37° C on LB/Agar as well. Single colonies were used to inoculate overnight (12 – 16 hours) cultures of 2 mL of M9 media supplemented with appropriate antibiotics. Next, the cultures were diluted 100X into 50 mL of fresh M9 with appropriate antibiotics and grown at 37° C in a shaking water bath to $OD_{600} = 0.2$. They were then split and 18 mL was put into each of two flasks and exposed to 2 mL of either pre-warmed M9 media or pre-warmed M9 media supplemented with 50 mM aspartate, for a final aspartate concentration of 5 mM. Cells were grown at 37° C with shaking for two more hours post induction. At this point a 0.5 mL sample was taken, spun at 3,300 x g for 5 minutes and re-suspended in 1X PBS with kanamycin (2 mg/mL) to stop translation. Cells were diluted into 1X PBS and single cell GFP measurements were made using a BD Biosciences LSRII, courtesy of the Gladstone Research Institute (Laser settings -- FSC: 577, SSC: 335, GFP: 607). Each data set consisted of at least 30,000 bacteria. The FlowJo software package was used to gate the data by FSC-H and SSC-A before calculating a geometric mean GFP fluorescence value (Tree Star Inc.).

Reference

1. Kofoid, E. C. & Parkinson, J. S. (1988). Transmitter and receiver modules in bacterial signaling proteins. *Proc Natl Acad Sci U S A* **85**, 4981-5.
2. Parkinson, J. S. & Kofoid, E. C. (1992). Communication modules in bacterial signaling proteins. *Annu Rev Genet* **26**, 71-112.
3. Silhavy, J. A. H. a. T. J., Ed. (1995). Two-Component Signal Transduction. Herndon, VA: ASM Press.
4. Aiba, H., Mizuno, T. & Mizushima, S. (1989). Transfer of phosphoryl group between two regulatory proteins involved in osmoregulatory expression of the ompF and ompC genes in Escherichia coli. *J Biol Chem* **264**, 8563-7.
5. Ronson, C. W., Nixon, B. T. & Ausubel, F. M. (1987). Conserved domains in bacterial regulatory proteins that respond to environmental stimuli. *Cell* **49**, 579-81.
6. Ninfa, A. J. & Magasanik, B. (1986). Covalent modification of the glnG product, NRI, by the glnL product, NRII, regulates the transcription of the glnALG operon in Escherichia coli. *Proc Natl Acad Sci U S A* **83**, 5909-13.
7. Aiba, H. & Mizuno, T. (1990). Phosphorylation of a bacterial activator protein, OmpR, by a protein kinase, EnvZ, stimulates the transcription of the ompF and ompC genes in Escherichia coli. *FEBS Lett* **261**, 19-22.
8. Forst, S., Delgado, J. & Inouye, M. (1989). Phosphorylation of OmpR by the osmosensor EnvZ modulates expression of the ompF and ompC genes in Escherichia coli. *Proc Natl Acad Sci USA* **86**, 6052-6.
9. Hazelbauer, G. L., Berg, H. C. & Matsumura, P. (1993). Bacterial motility and signal transduction. *Cell* **73**, 15-22.
10. Francez-Charlot, A., Laugel, B., Van Gemert, A., Dubarry, N., Wiorowski, F., Castanie-Cornet, M. P., Gutierrez, C. & Cam, K. (2003). RcsCDB His-Asp phosphorelay system negatively regulates the flhDC operon in Escherichia coli. *Mol Microbiol* **49**, 823-32.
11. Shin, S. & Park, C. (1995). Modulation of flagellar expression in Escherichia coli by acetyl phosphate and the osmoregulator OmpR. *J Bacteriol* **177**, 4696-702.
12. Kleerebezem, M., Quadri, L. E., Kuipers, O. P. & de Vos, W. M. (1997). Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. *Mol Microbiol* **24**, 895-904.
13. Li, Y. H., Lau, P. C., Tang, N., Svensater, G., Ellen, R. P. & Cvitkovitch, D. G. (2002). Novel two-component regulatory system involved in biofilm formation and acid resistance in Streptococcus mutans. *J Bacteriol* **184**, 6333-42.
14. Garmendia, J., Beuzon, C. R., Ruiz-Albert, J. & Holden, D. W. (2003). The roles of SsrA-SsrB and OmpR-EnvZ in the regulation of genes encoding the Salmonella typhimurium SPI-2 type III secretion system. *Microbiology* **149**, 2385-96.
15. Dorel, C., Vidal, O., Prigent-Combaret, C., Vallet, I. & Lejeune, P. (1999). Involvement of the Cpx signal transduction pathway of E. coli in biofilm formation. *FEMS Microbiol Lett* **178**, 169-75.

16. Rabin, R. S. & Stewart, V. (1993). Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in Escherichia coli K-12. *J Bacteriol* **175**, 3259-68.
17. Jung, K., Hamann, K. & Revermann, A. (2001). K+ stimulates specifically the autokinase activity of purified and reconstituted EnvZ of Escherichia coli. *J Biol Chem* **276**, 40896-902.
18. Ferrieres, L. & Clarke, D. J. (2003). The RcsC sensor kinase is required for normal biofilm formation in Escherichia coli K-12 and controls the expression of a regulon in response to growth on a solid surface. *Mol Microbiol* **50**, 1665-82.
19. Ullrich, M., Penaloza-Vazquez, A., Bailey, A. M. & Bender, C. L. (1995). A modified two-component regulatory system is involved in temperature-dependent biosynthesis of the Pseudomonas syringae phytotoxin coronatine. *J Bacteriol* **177**, 6160-9.
20. Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. (1997). A cyanobacterial phytochrome two-component light sensory system. *Science* **277**, 1505-8.
21. Hirakawa, H., Inazumi, Y., Masaki, T., Hirata, T. & Yamaguchi, A. (2005). Indole induces the expression of multidrug exporter genes in Escherichia coli. *Mol Microbiol* **55**, 1113-26.
22. Sato, M., Machida, K., Arikado, E., Saito, H., Kakegawa, T. & Kobayashi, H. (2000). Expression of outer membrane proteins in Escherichia coli growing at acid pH. *Appl Environ Microbiol* **66**, 943-7.
23. Yamamoto, K. & Ishihama, A. (2005). Transcriptional response of Escherichia coli to external copper. *Mol Microbiol* **56**, 215-27.
24. Mascher, T., Zimmer, S. L., Smith, T. A. & Helmann, J. D. (2004). Antibiotic-inducible promoter regulated by the cell envelope stress-sensing two-component system LiaRS of Bacillus subtilis. *Antimicrob Agents Chemother* **48**, 2888-96.
25. David, M., Daveran, M. L., Batut, J., Dedieu, A., Domergue, O., Ghai, J., Hertig, C., Boistard, P. & Kahn, D. (1988). Cascade regulation of nif gene expression in Rhizobium meliloti. *Cell* **54**, 671-83.
26. Carballes, F., Bertrand, C., Bouche, J. P. & Cam, K. (1999). Regulation of Escherichia coli cell division genes ftsA and ftsZ by the two-component system rcsC-rcsB. *Mol Microbiol* **34**, 442-50.
27. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. (2007). Accurate prediction of gene feedback circuit behavior from component properties. *Mol Syst Biol* **3**, 143.
28. Bashor, C. J., Helman, N. C., Yan, S. & Lim, W. A. (2008). Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. *Science* **319**, 1539-43.
29. Dueber, J. E., Mirsky, E. A. & Lim, W. A. (2007). Engineering synthetic signaling proteins with ultrasensitive input/output control. *Nat Biotechnol* **25**, 660-2.
30. Dueber, J. E., Yeh, B. J., Chak, K. & Lim, W. A. (2003). Reprogramming control of an allosteric signaling switch through modular recombination. *Science* **301**, 1904-8.

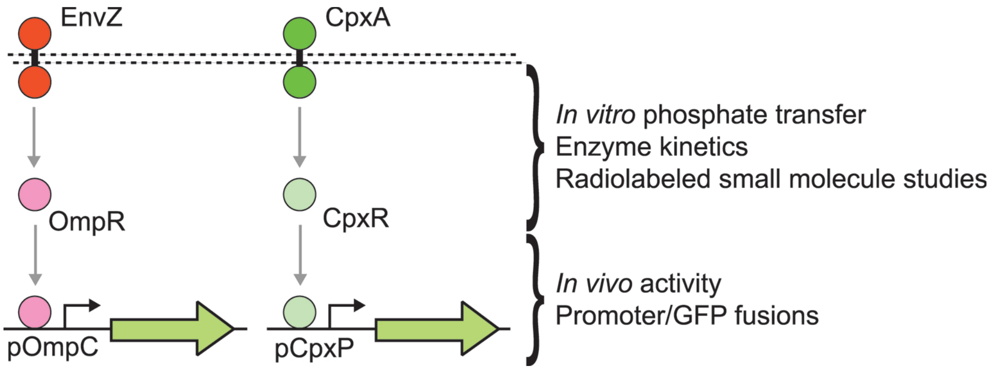
31. Buckler, D. R., Zhou, Y. & Stock, A. M. (2002). Evidence of intradomain and interdomain flexibility in an OmpR/PhoB homolog from Thermotoga maritima. *Structure* **10**, 153-64.
32. Gouet, P., Fabry, B., Guillet, V., Birck, C., Mourey, L., Kahn, D. & Samama, J. P. (1999). Structural transitions in the FixJ receiver domain. *Structure* **7**, 1517-26.
33. Sola, M., Gomis-Ruth, F. X., Serrano, L., Gonzalez, A. & Coll, M. (1999). Three-dimensional crystal structure of the transcription factor PhoB receiver domain. *J Mol Biol* **285**, 675-87.
34. Volz, K. & Matsumura, P. (1991). Crystal structure of Escherichia coli CheY refined at 1.7-A resolution. *J Biol Chem* **266**, 15511-9.
35. Utsumi, R., Brissette, R. E., Rampersaud, A., Forst, S. A., Oosawa, K. & Inouye, M. (1989). Activation of bacterial porin gene expression by a chimeric signal transducer in response to aspartate. *Science* **245**, 1246-9.
36. Baumgartner, J. W., Kim, C., Brissette, R. E., Inouye, M., Park, C. & Hazelbauer, G. L. (1994). Transmembrane signalling by a hybrid protein: communication from the domain of chemoreceptor Trg that recognizes sugar-binding proteins to the kinase/phosphatase domain of osmosensor EnvZ. *J Bacteriol* **176**, 1157-63.
37. Looger, L. L., Dwyer, M. A., Smith, J. J. & Hellinga, H. W. (2003). Computational design of receptor and sensor proteins with novel functions. *Nature* **423**, 185-90.
38. Levskaya, A., Chevalier, A. A., Tabor, J. J., Simpson, Z. B., Lavery, L. A., Levy, M., Davidson, E. A., Scouras, A., Ellington, A. D., Marcotte, E. M. & Voigt, C. A. (2005). Synthetic biology: engineering Escherichia coli to see light. *Nature* **438**, 441-2.
39. Ward, S. M., Delgado, A., Gunsalus, R. P. & Manson, M. D. (2002). A NarX-Tar chimera mediates repellent chemotaxis to nitrate and nitrite. *Mol Microbiol* **44**, 709-19.
40. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2003). Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proc Natl Acad Sci U S A* **100**, 11255-60.
41. Laub, M. T. & Goulian, M. (2007). Specificity in two-component signal transduction pathways. *Annu Rev Genet* **41**, 121-45.
42. Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M. & Laub, M. T. (2008). Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043-54.
43. Stock, J. B., Ninfa, A. J. & Stock, A. M. (1989). Protein phosphorylation and regulation of adaptive responses in bacteria. *Microbiol Rev* **53**, 450-90.
44. Hellingwerf, K. J. (2005). Bacterial observations: a rudimentary form of intelligence? *Trends Microbiol* **13**, 152-8.
45. Hellingwerf, K. J., Postma, P. W., Tommassen, J. & Westerhoff, H. V. (1995). Signal transduction in bacteria: phospho-neural network(s) in Escherichia coli? *FEMS Microbiol Rev* **16**, 309-21.
46. Alves, R. & Savageau, M. A. (2003). Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors:

differences in molecular structure and physiological function. *Mol Microbiol* **48**, 25-51.

47. Siryaporn, A. & Goulian, M. (2008). Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in E. coli. *Mol Microbiol* **70**, 494-506.
48. Grimshaw, C. E., Huang, S., Hanstein, C. G., Strauch, M. A., Burbulys, D., Wang, L., Hoch, J. A. & Whiteley, J. M. (1998). Synergistic kinetic interactions between components of the phosphorelay controlling sporulation in Bacillus subtilis. *Biochemistry* **37**, 1365-75.
49. Yamamoto, K., Hirao, K., Oshima, T., Aiba, H., Utsumi, R. & Ishihama, A. (2005). Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *J Biol Chem* **280**, 1448-56.
50. Skerker, J. M., Prasol, M. S., Perchuk, B. S., Biondi, E. G. & 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.
51. Dong, J., Iuchi, S., Kwan, H. S., Lu, Z. & Lin, E. C. (1993). The deduced amino-acid sequence of the cloned cpxR gene suggests the protein is the cognate regulator for the membrane sensor, CpxA, in a two-component signal transduction system of Escherichia coli. *Gene* **136**, 227-30.
52. De Wulf, P., McGuire, A. M., Liu, X. & Lin, E. C. (2002). Genome-wide profiling of promoter recognition by the two-component response regulator CpxR-P in Escherichia coli. *J Biol Chem* **277**, 26652-61.
53. Raivio, T. L. (2005). Envelope stress responses and Gram-negative bacterial pathogenesis. *Mol Microbiol* **56**, 1119-28.
54. Batchelor, E., Walthers, D., Kenney, L. J. & Goulian, M. (2005). The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. *J Bacteriol* **187**, 5723-31.
55. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403-10.
56. Igo, M. M., Ninfa, A. J., Stock, J. B. & Silhavy, T. J. (1989). Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor. *Genes Dev* **3**, 1725-34.
57. Jin, T. & Inouye, M. (1993). Ligand binding to the receptor domain regulates the ratio of kinase to phosphatase activities of the signaling domain of the hybrid Escherichia coli transmembrane receptor, Taz1. *J Mol Biol* **232**, 484-92.
58. Ninfa, A. J., Ninfa, E. G., Lupas, A. N., Stock, A., Magasanik, B. & Stock, J. (1988). Crosstalk between bacterial chemotaxis signal transduction proteins and regulators of transcription of the Ntr regulon: evidence that nitrogen assimilation and chemotaxis are controlled by a common phosphotransfer mechanism. *Proc Natl Acad Sci U S A* **85**, 5492-6.
59. De Wulf, P. & Lin, E. C. (2000). Cpx two-component signal transduction in Escherichia coli: excessive CpxR-P levels underlie CpxA* phenotypes. *J Bacteriol* **182**, 1423-6.

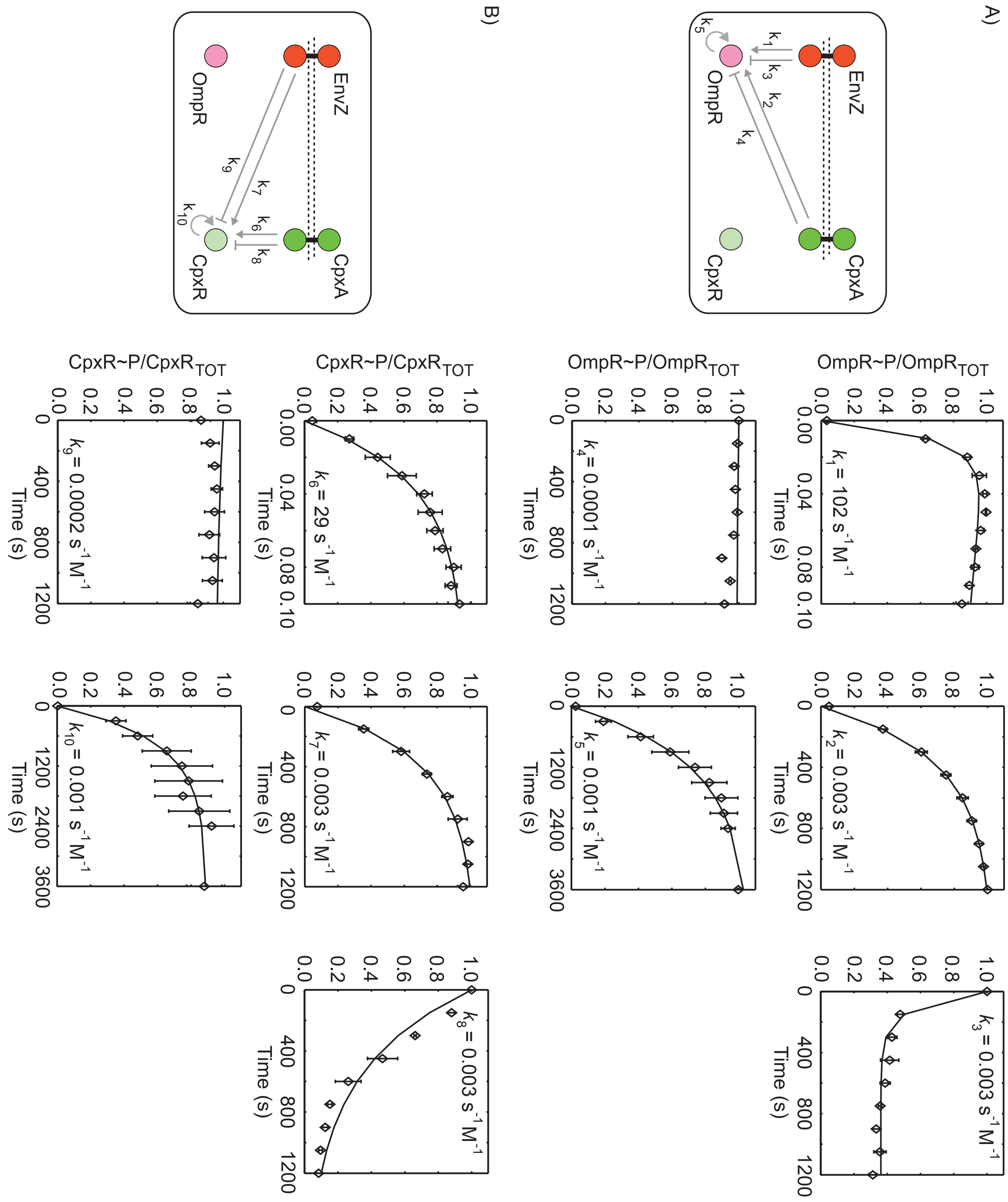
60. Wolfe, A. J., Parikh, N., Lima, B. P. & Zemaitaitis, B. (2008). Signal integration by the two-component signal transduction response regulator CpxR. *J Bacteriol* **190**, 2314-22.
61. Igo, M. M. & Silhavy, T. J. (1988). EnvZ, a transmembrane environmental sensor of Escherichia coli K-12, is phosphorylated in vitro. *J Bacteriol* **170**, 5971-3.
62. Ames, S. K., Frankema, N. & Kenney, L. J. (1999). C-terminal DNA binding stimulates N-terminal phosphorylation of the outer membrane protein regulator OmpR from Escherichia coli. *Proc Natl Acad Sci U S A* **96**, 11792-7.
63. McCleary, W. R. & Stock, J. B. (1994). Acetyl phosphate and the activation of two-component response regulators. *J Biol Chem* **269**, 31567-72.
64. Yoshida, T., Cai, S. & Inouye, M. (2002). Interaction of EnvZ, a sensory histidine kinase, with phosphorylated OmpR, the cognate response regulator. *Mol Microbiol* **46**, 1283-94.
65. Cai, S. J. & Inouye, M. (2002). EnvZ-OmpR interaction and osmoregulation in Escherichia coli. *J Biol Chem* **277**, 24155-61.
66. Wolfe, A. J. (2005). The acetate switch. *Microbiol Mol Biol Rev* **69**, 12-50.
67. Maeda, S. & Mizuno, T. (1990). Evidence for multiple OmpR-binding sites in the upstream activation sequence of the ompC promoter in Escherichia coli: a single OmpR-binding site is capable of activating the promoter. *J Bacteriol* **172**, 501-3.
68. Yamamoto, K. & Ishihama, A. (2006). Characterization of copper-inducible promoters regulated by CpxA/CpxR in Escherichia coli. *Biosci Biotechnol Biochem* **70**, 1688-95.
69. Batchelor, E., Silhavy, T. J. & Goulian, M. (2004). Continuous control in bacterial regulatory circuits. *J Bacteriol* **186**, 7618-25.
70. Otto, K. & Silhavy, T. J. (2002). Surface sensing and adhesion of Escherichia coli controlled by the Cpx-signaling pathway. *Proc Natl Acad Sci U S A* **99**, 2287-92.
71. Batchelor, E. & Goulian, M. (2006). Imaging OmpR localization in Escherichia coli. *Mol Microbiol* **59**, 1767-78.
72. Fabret, C., Feher, V. A. & Hoch, J. A. (1999). Two-component signal transduction in Bacillus subtilis: how one organism sees its world. *J Bacteriol* **181**, 1975-83.
73. Jiang, M., Shao, W., Perego, M. & Hoch, J. A. (2000). Multiple histidine kinases regulate entry into stationary phase and sporulation in Bacillus subtilis. *Mol Microbiol* **38**, 535-42.
74. Zhou, L., Lei, X. H., Bochner, B. R. & Wanner, B. L. (2003). Phenotype microarray analysis of Escherichia coli K-12 mutants with deletions of all two-component systems. *J Bacteriol* **185**, 4956-72.
75. Jubelin, G., Vianney, A., Beloin, C., Ghigo, J. M., Lazzaroni, J. C., Lejeune, P. & Dorel, C. (2005). CpxR/OmpR interplay regulates curli gene expression in response to osmolarity in Escherichia coli. *J Bacteriol* **187**, 2038-49.
76. Oshima, T., Aiba, H., Masuda, Y., Kanaya, S., Sugiura, M., Wanner, B. L., Mori, H. & Mizuno, T. (2002). Transcriptome analysis of all two-component regulatory system mutants of Escherichia coli K-12. *Mol Microbiol* **46**, 281-91.

77. Mattison, K. & Kenney, L. J. (2002). Phosphorylation alters the interaction of the response regulator OmpR with its sensor kinase EnvZ. *J Biol Chem* **277**, 11143-8.
78. Voigt, C. A. (2006). Genetic parts to program bacteria. *Curr Opin Biotechnol* **17**, 548-57.
79. Gill, S. C. & von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**, 319-26.
80. Abramoff, M. D., Magelhaes, P.J., Ram, S.J. (2004). Image Processing with ImageJ. *Biophotonics International* **11**, 36-42.
81. Datsenko, K. A. & 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**, 6640-5.



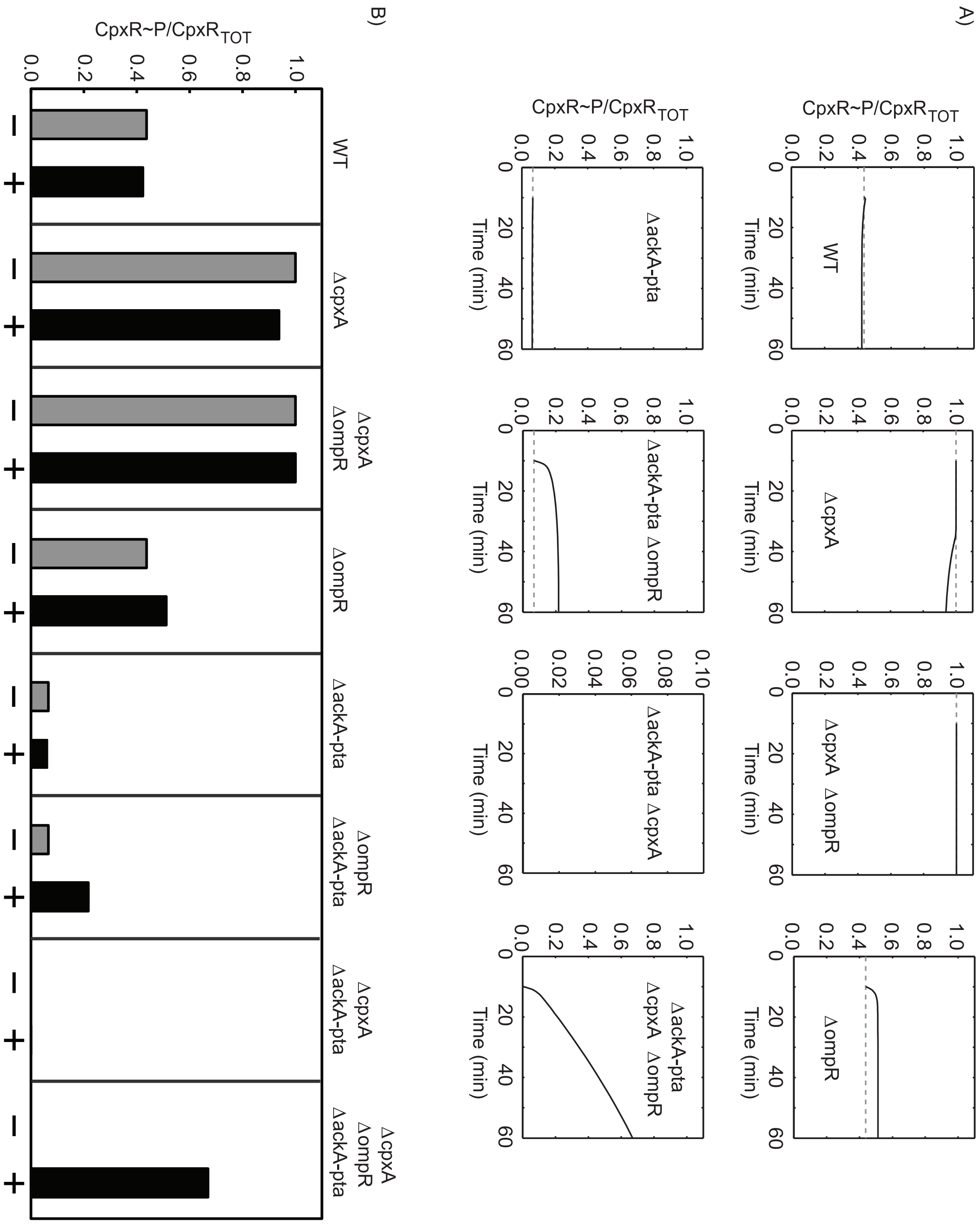
Chapter 6: Figure 1. Overview of the EnvZ/OmpR and CpxA/CpxR two-component systems.

The EnvZ and CpxA histidine kinase (red and green) reside in the inner membrane and respond to a variety of signals. Upon activation, these kinases autophosphorylate and transfer a phosphate to their cognate response regulators, OmpR and CpxR (pink and light green circles). Upon phosphorylation, OmpR~P and CpxR~P activate the *ompC* and *cpxP* promoters respectively. We measure the phosphorelay kinetics *in vitro* using radiolabeled phosphate and purified proteins. The downstream effects are measured *in vivo* using promoter-GFP transcriptional fusions and flow cytometry.


A)
EnvZ
CpxA
OmpR
CpxR
k1
k2
k3
k4
k5
OmpR~P/OmpR$_{TOT}$
Time (s)
k_1 = 102 s^{-1} M^{-1}
k_2 = 0.003 s^{-1} M^{-1}
k_3 = 0.003 s^{-1} M^{-1}
k_4 = 0.0001 s^{-1} M^{-1}
k_5 = 0.001 s^{-1} M^{-1}
B)
EnvZ
CpxA
OmpR
CpxR
k6
k7
k8
k9
k10
CpxR~P/CpxR$_{TOT}$
Time (s)
k_6 = 29 s^{-1} M^{-1}
k_7 = 0.003 s^{-1} M^{-1}
k_8 = 0.003 s^{-1} M^{-1}
k_9 = 0.0002 s^{-1} M^{-1}
k_{10} = 0.001 s^{-1} M^{-1}
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.04
0.08
0.10
0
400
800
1200
2400
3600

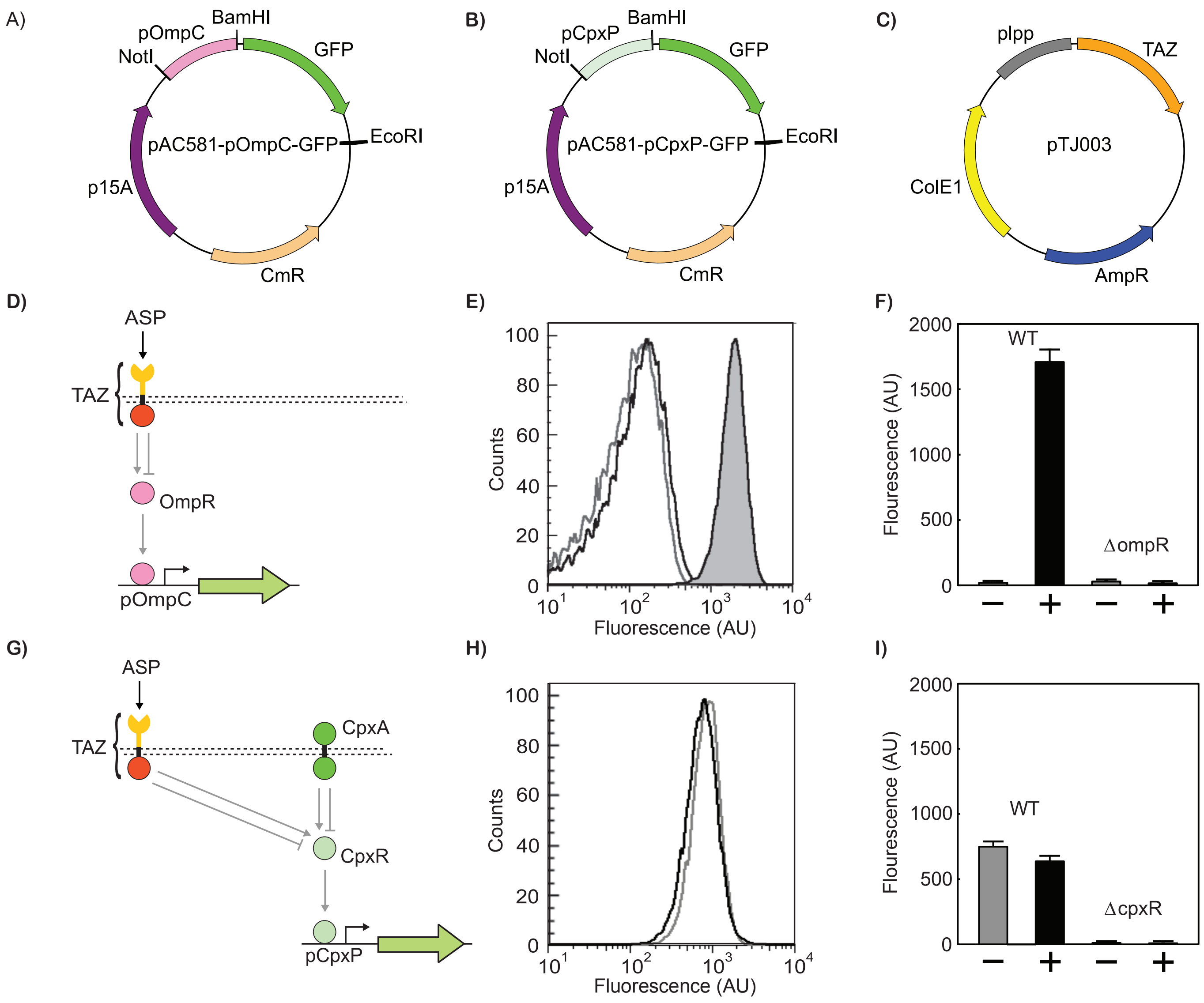
Chapter 6: Figure 2. *In vitro* Kinetic Analysis of the EnvZ/OmpR and CpxA/CpxR systems.

All of the phosphotransfer reactions affecting OmpR (A) and CpxR (B) are shown. The data for each reaction is extracted from gel assays and shown as the fraction of response regulator phosphorylated at each timepoint. Points are the average of four different experiments and error bars are a single standard deviation from the mean. The curves are obtained by fitting for the various kinetic parameters.


A)
CpxR~P/CpxR$_{TOT}$
WT
ΔcpxA
ΔcpxA ΔompR
ΔompR
ΔackA-pta
ΔackA-pta ΔompR
ΔackA-pta ΔcpxA
ΔackA-pta
ΔcpxA ΔompR
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
Time (min)
0
20
40
60
B)
CpxR~P/CpxR$_{TOT}$
0.0
0.2
0.4
0.6
0.8
1.0
WT
ΔcpxA
ΔcpxA ΔompR
ΔompR
ΔackA-pta
ΔompR ΔackA-pta
ΔcpxA ΔackA-pta
ΔcpxA ΔompR ΔackA-pta
-
+

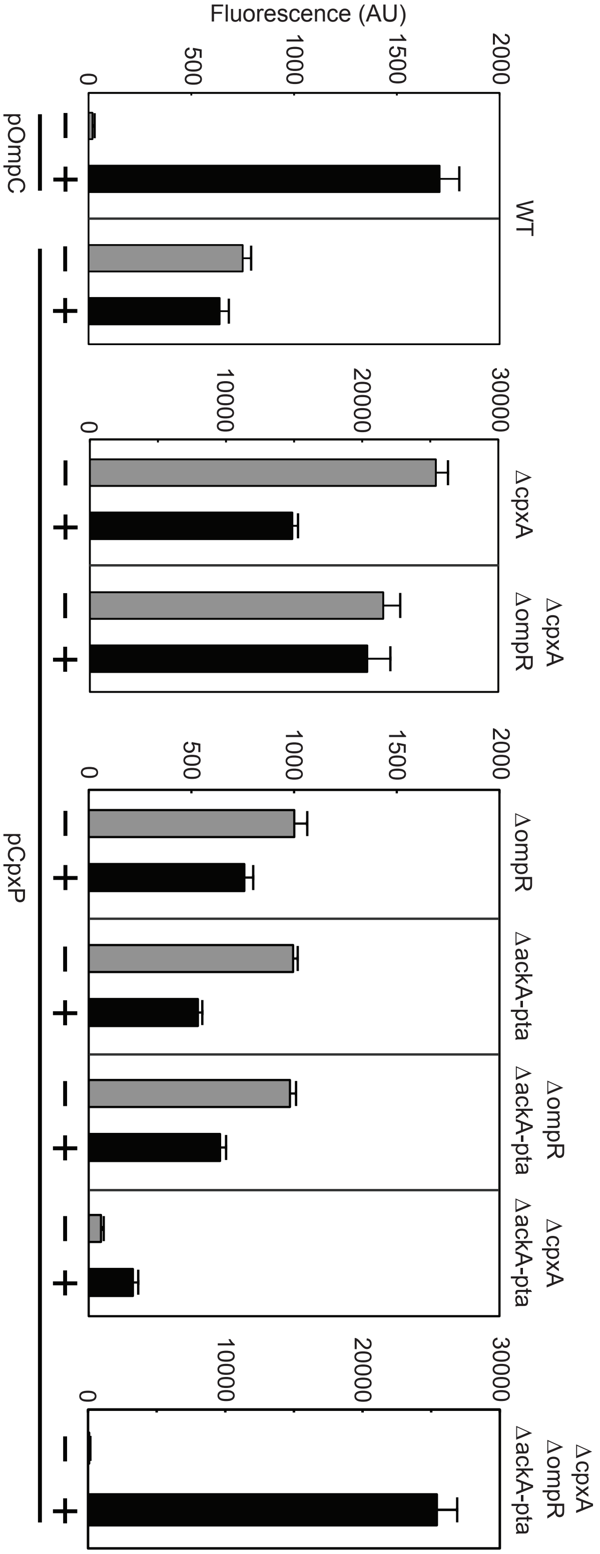
Chapter 6: Figure 3. Kinetic model of CpxR phosphorylation *in vivo*.

The effect of knocking out various components of the pathways is shown. The simulation tracks the fraction of CpxR that is phosphorylated after EnvZ is activated. (A) Various time series are shown demonstrating the evolution of CpxR~P from zero to 50 minutes post induction. The simulations are allowed to reach steady-state and then EnvZ is activated at the ten minute time point (black line). The grey line shows the time trajectory in the absence of EnvZ. (B) The bar graphs show the final steady-state fraction of CpxR~P in the presence of active EnvZ (black) and in the absence of EnvZ (gray).


A)
pOmpC
BamHI
NotI
GFP
pAC581-pOmpC-GFP
EcoRI
p15A
CmR
B)
pCpxP
BamHI
NotI
GFP
pAC581-pCpxP-GFP
EcoRI
p15A
CmR
C)
plpp
TAZ
pTJ003
ColE1
AmpR
D)
ASP
TAZ
OmpR
pOmpC
E)
Counts
100
80
60
40
20
0
10^1
10^2
10^3
10^4
Fluorescence (AU)
F)
Flourescence (AU)
2000
1500
1000
500
0
WT
ΔompR
−
+
−
+
G)
ASP
TAZ
CpxA
CpxR
pCpxP
H)
Counts
100
80
60
40
20
0
10^1
10^2
10^3
10^4
Fluorescence (AU)
I)
Flourescence (AU)
2000
1500
1000
500
0
WT
ΔcpxR
−
+
−
+

Chapter 6: Figure 4. *In vivo* measurement system for response regulator phosphorylation.

Plasmid maps are shown for the reporter systems: (A) pAC581-pOmpC-GFP and (B) pAC581-pCpxP-GFP, which contain the *ompC* and *cpxP* promoters driving GFP expression. (C) The reporters are co-transformed with a second plasmid pTJ003, which contains the TAZ protein under control of the constitutive lpp promoter. (D) TAZ phosphorylates OmpR in the presence of 5 mM asparate. (E) Gated cytometry data is shown for cells without TAZ (gray line) and with TAZ but without asparate (black line). Addition of 5 mM aspartate induces the system (black line, right). (F) The *ompC* reporter is not induced in the absence of TAZ (-), but is strongly induced when both TAZ and aspartate is present (+). When ompR is knocked out, TAZ has no affect on the system. (G) The ability of TAZ to induce CpxR was also measured in wild-type cells. (H) The cytometry distribution shows no difference between cells with TAZ and aspartate (black line) and cells without TAZ (gray line). There is a basal level of activity from the *cpxP* promoter. (I) The activity of the *cpxP* promoter is the same in the absence (-) and presence (+) of TAZ and inducer. The basal activity is eliminated by knocking out CpxR. For both parts F and I, the mean of three experiments performed on different days is shown and the error bars are one standard deviation.


Fluorescence (AU)
0
500
1000
1500
2000
0
10000
20000
30000
0
500
1000
1500
2000
0
10000
20000
30000
WT
ΔcpxA
ΔcpxA ΔompR
ΔompR
ΔackA-pta
ΔompR ΔackA-pta
ΔcpxA ΔackA-pta
ΔcpxA ΔompR ΔackA-pta
–
+
–
+
–
+
–
+
–
+
–
+
–
+
–
+
–
+
pOmpC
pCpxP

Chapter 6: Figure 5. *In vivo* experiments measuring crosstalk when various component genes are knocked out.

The knockouts are shown at the top of each panel and the promoter used as a reporter is shown at the bottom. For each knockout, the fluorescence in the absence of EnvZ (-) and presence of active EnvZ (TAZ and aspartate) (+) are shown. Panel 1 shows selective activation of the *ompC* promoter in the presence of TAZ and aspartate (black bars) but not in the absence of TAZ (gray bars). Also, TAZ is unable to activation the *cpxP* promoter. Different combinations of knock outs do not produce cross talk between EnvZ and CpxR, as shown from the inability of this kinase to activate the *cpxP* promoter. Only a triple knockout, Δ*ackA-pta* Δ*cpxA* Δ*omp*R, displays cross talk. Bars represent the mean of three different experiments and the error bars are one standard deviation from the mean.

Figure S1: Phosphate transfer from EnvZ to OmpR using Rapid Quench Flow System.

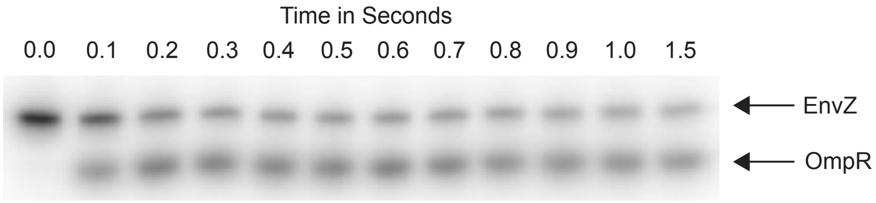


Figure S2: Phosphate transfer from CpxA to OmpR.

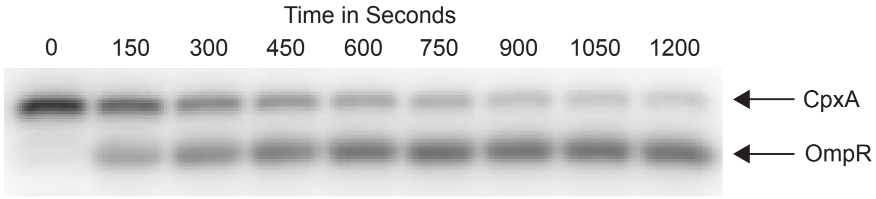


Figure S3: Dephosphorylation of OmpR by EnvZ.

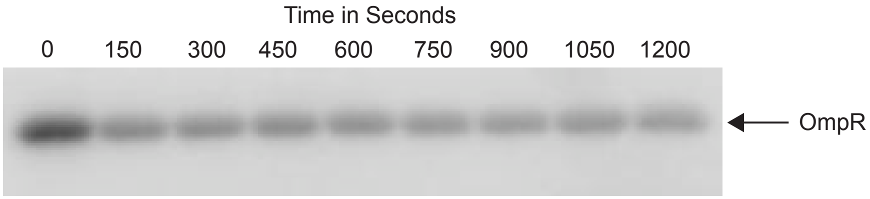


Figure S4: Dephosphorylation of OmpR by CpxA.

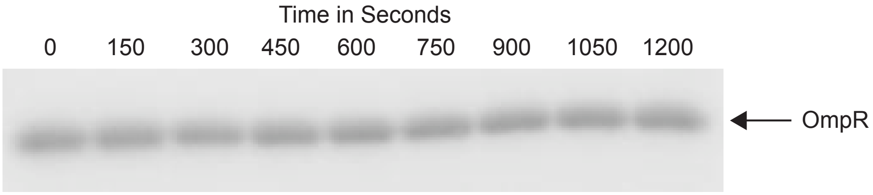


Figure S5: Phosphorylation of OmpR by ^{32}P acetyl phosphate.

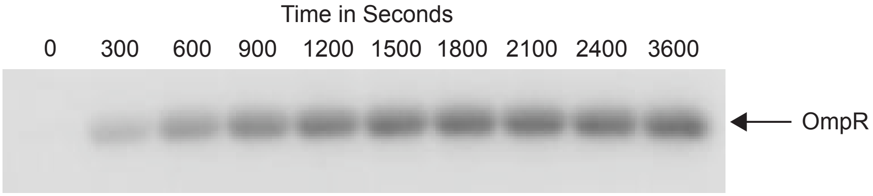


Figure S6: Phosphate transfer from CpxA to CpxR using Rapid Quench Flow System.

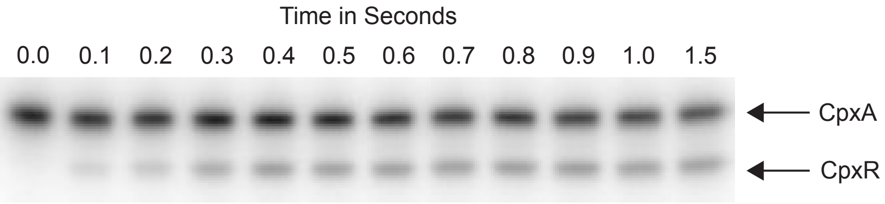


Figure S7: Phosphate transfer from EnvZ to CpxR.

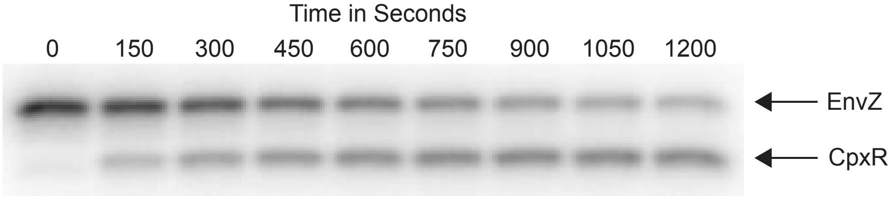


Figure S8: Dephosphorylation of CpxR by CpxA.

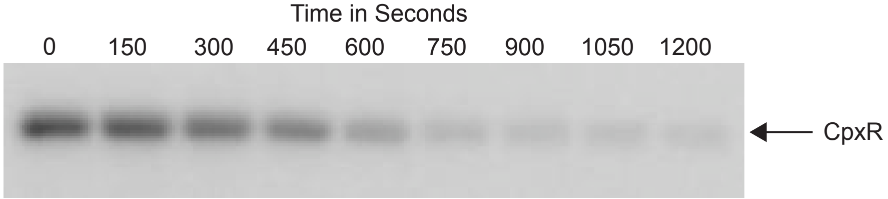


Figure S9: Dephosphorylation of CpxR by EnvZ.

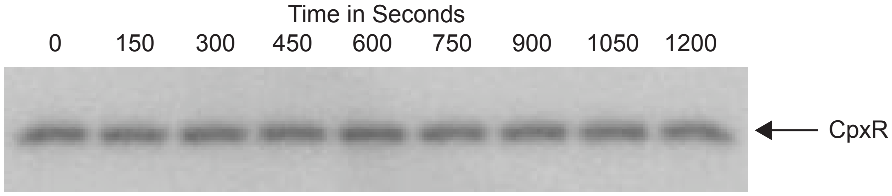
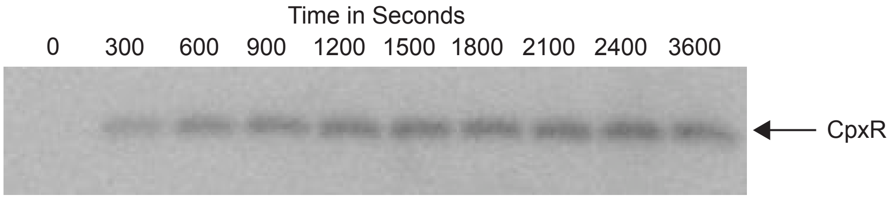


Figure S10: Phosphorylation of CpxR by ^{32}P acetyl phosphate.



Chapter 7: Remote Control of Bacterial Chemotaxis

Elizabeth J. Clarke[1,Φ], Eli S. Groban[1,Φ], Ryan M. Clark[4], Matthew Eames[1], Tanja Kortemme[3], and Christopher A. Voigt[1,2]

[1] Graduate Group in Biophysics

[2] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA.

[3] Department of Biopharmaceutical Sciences, University of California, San Francisco, San Francisco, CA 94158, USA.

[4] University of Kansas, Lawrence, KS 66045,USA

[Φ]These authors contributed equally to this work.

Abstract

Bacteria sense and respond to their environment through sensors linked to a set of regulatory networks that function to control cellular responses. One of these regulatory networks is the well-studied chemotaxis system, which allows the cell to move towards nutrients and away from toxins. Bacteria contain multiple external sensors that sense an array of different small molecules. These sensors all feed into the same central pathway that controls cellular movement. Since many different signals all converge into one signaling pathway, the cell integrates signals from multiple receptors without having the ability to distinguish between different input signals, other than deciphering a difference between food and toxins. We reprogrammed *Escherichia coli* to distinguish between two stimuli and move towards one or the other depending on the absence or presence of an outside signaling molecule. This required us to engineer two orthogonal signaling systems and to replace the native chemotaxis signaling machinery with the DNA for our novel system. Finally, we inserted an externally controlled DNA switch that recognizes a small molecule signal and directs cell movement towards one stimulus or another. This DNA encoded switch also contains a memory function, allowing the progeny of each bacterium to remember which direction to move. In summary, we successfully implemented synthetic biology techniques to first characterize a series of parts and then to combine these parts in a living organism to confer to a bacterium a novel behavior, which is entirely under human remote control.

Introduction

E. coli relies on a network of sensors linked to propulsion machinery to move towards nutrients and away from toxins (8, 9). This process is referred to as bacterial chemotaxis and evolved to be non-specific with respect to different food sources or toxins. In this work, we engineered specificity into this network, creating a strain of *E. coli* that senses its environment and chooses to chemotax towards one of two stimuli, not both. Engineering this new behavior required three different elements: A genetic switch with trans-generational memory, two orthogonal signaling pathways, and a strain compatible with both the new systems and the memory switch. All components used in this new design were taken from either the parts registry or from various biology laboratories and optimized to serve a different function. Placement of pre-existing parts into a new genetic system allowed for fast design and construction of a 12 kB piece of synthetic DNA.

Wild type *E. coli* contains five different chemotaxis receptors that sense and respond to the surrounding environment. The two most studied receptors are the transmembrane aspartate receptor (TAR) and the transmembrane serine receptor (TSR), which respond to aspartate and serine respectively (11). These receptors interface with flagella apparatus, which provides locomotive force for the bacterium, by feeding signal into a two component histidine kinase, CheA. In the active state, CheA phosphorylates the CheY response regulator, a cytoplasmic protein that binds to the base of the flagella in the phosphorylated state (3, 14). CheA, however, relies on a scaffolding protein, CheW to sense the activation state of each receptor. CheW interacts with all five chemotaxis receptors and serves as a signal integration device in bacterial sensing.

In wild type *E. coli*, the transmembrane aspartate receptor (TAR) and transmembrane serine receptor (TSR) both elicit the same CheW dependent cellular response to the presence of aspartate and serine respectively. In 1991, Liu and Parkinson showed that a mutation in either CheW or the cytoplasmic portion of the TSR receptor protein abolished chemotaxis towards serine (12). Certain pairs of complementary mutations in CheW (V108M, CheW*) and TSR (E402A, TSR*), however, restored chemotaxis, showing that CheW interacts with both TAR and TSR to initiate chemotactic signaling. In this study, however, the TAR protein senses aspartate and signals through CheW while the TSR* protein senses serine and signals through the CheW* protein.

While Parkinson used his system to investigate biology, in this work the system proves critical for design. We created a design that integrated Parkinson's system with a DNA based memory function under human control. Moreover, we also constructed a bacterial strain that would be compatible with these new pathways and the new switch. Overall, we created a novel behavior in an existing organism entirely under human control. To engineer this system we took parts and devices from both synthetic biology and basic biology labs and merged them into a larger device that controls motility in cells.

Results and Discussion

Chemotactic response of Wild Type *E. coli*

Although it has been studied many times before, we wanted to first recreate the wild type response of *E. coli* to both aspartate and serine. We did this both to see whether the assay would work in our hands and to have a control conducted in the same exact conditions. First, we plated bacteria on a semi-solid agar plate in the absence of any type of attractant. Under this condition, the cells migrate equally in all directions, forming a clearly observable circle (Figure 1). Upon the addition of aspartate, the cells respond to the presence of food by preferentially migrating, or "swimming" towards it. We use the term "swimming" in this work to represent the overall behavior of a bacterial population, not the behavior of a single bacterium, which does not swim towards attractant. In the presence of serine instead of aspartate, the cells show an equivalent response. This experiment not only reaffirms that E. coli migrate towards both aspartate and serine, but also provides a baseline as we hope to recover wild type behavior from our engineered system.

Overall system design

A unique design was required to create a bacterium that migrates towards aspartate or serine and makes this decision based on a remotely controlled signal. The design layout is in Figure 2. First, we needed an orthogonal signaling system that is able to sense aspartate and serine independently. Second, we needed a DNA encoded switch to control whether the aspartate sensing system or the serine sensing system would be functional. Finally, we needed to design a strain of *E. coli* that would be compatible with both the new signaling pathways and the new switch. We completed all of these tasks in order to develop remote control of bacterial chemotaxis.

Wild type *E. coli* is not an appropriate chassis for our chemotaxis system as it contains all chemotaxis proteins under native regulation. Instead, for our host organism, we started from strain BW28357 (4). In order to use this cell line with our system, we knocked out CheW, TSR, and TAR. We also removed all other chemoreceptors as we were worried that these receptors might interfere with our engineered system. Moreover, our switch relies on the FimE inversion system originally developed in the Arkin laboratory (7). We will describe this system in more detail later, but in order for it to function effectively we also needed to remove the FimE and FimB recombinase proteins, which control fimbriae induction in wild type *E. coli* (2, 10, 13). Leaving any of these proteins in the genome would have contributed a certain amount of noise, preventing total control over the bacterial population. By optimizing the genome of our host, we made the system robust, controllable and, therefore, reliable.

Testing and Validation of the pFIP switch

The pFIP switch is a recombinase-based system that relies on the FimE protein to reverse a segment of DNA between two recognition sites. Arkin and co-workers used the P_{BAD} promoter to express the FimE recombinase (7). Upon activation, they showed that this recombinase flipped an inverted P_{trc} promoter, lacking the LacI binding sites, that previously pointed towards two terminators (7). Although their work suggested that the switch was functional, we decided to characterize it a bit more before incorporating it into our system. First, we placed a green fluorescent protein on one side of the P_{trc} promoter so that it is constituently expressed (Figure 3). We also placed a red fluorescent protein on the other side of the P_{trc} promoter. As a simple proof of concept, we grew a culture of our genetically modified cells, split the culture, and plated on a plate containing no arabinose and a plate containing arabinose. As expected, the cells on the plate that did not have arabinose remained green, suggesting that the P_{trc} promoter did not change orientations. The cells on the arabinose plate, however, turned red, showing that the presence of arabinose fully switches the P_{trc} promoter. In this case, it no longer faces the green fluorescent protein and instead the cells produce only red fluorescent protein.

In order to measure the transfer function for this switch we grew cells in culture and exposed them to different levels of arabinose. For these measurements, all experiments started with an overnight culture of cells containing our switch. The cells were diluted into media containing different levels of arabinose ranging from micromolar to millimolar. After reaching equilibrium, which we assumed happened around eight hours after dilution, samples were analyzed by flow cytometry in order to obtain single cell measurements. As expected, the amount of cells expressing the red fluorescent protein increases as the amount of arabinose in the culture increases

(Figure 3). The induction curve for our switch is quite similar to the one Guzmann and co-workers observed in 1995, suggesting that P_{BAD} controlled expression of the FimE protein leads to a flip in the P_{trc} promoter (6).

Construction and testing of the Remotely Controlled Chemotaxis Pathway

The pathway controlling remote control of chemotaxis incorporated our DNA switch combined with the two orthogonal signaling systems (Schematics in Figures 4 and 5). These elements were placed into a modified version of the pFIP plasmid developed by Ham and co-workers (7). This plasmid contains the ampicillin resistance gene, the P_{BAD} promoter driving expression of the FimE recombinase, and the AraC protein constitutively expressed, to repress the P_{BAD} promoter in the absence of arabinose. We placed CheW* and ribosome binding site C058 (in house RBS library, manuscript submitted) on the non-expressing side of the P_{trc} promoter, which was flanked by FimE recombinase recognition sites. For the other side of the promoter, we placed a piece of synthetic DNA containing multiple elements (15). First, it has the CheW protein behind RBS C058 and in front of the T3 terminator. After this gene, we placed TSR* and TAR behind the constitutive PJ23113 promoter (1). The C055 RBS optimized expression of these proteins and the T4 terminator ended translation. We refer to the modified pFIP plasmid containing our genetic system as the chemotaxis plasmid.

Our genetic system switches which CheW the bacterium expresses depending on presence of a chemical signal, in this case, arabinose. The chemical receptors, TSR and TAR, are expressed at all times, allowing for fast switching between aspartate and serine sensitivity. In the absence of arabinose, the P_{trc} promoter points towards the wild type CheW protein and, therefore, the cells express wild type CheW. These cells should only sense and respond to the presence of aspartate in the medium. In the absence of any attractant, these bacteria form a circle, migrating outwards in all directions from a central location, with behavior that is almost identical to wild type cells (Figure 4). When aspartate is added to one side of the plate, the cells swim towards this attractant, forming a smear that is biased towards the nutrient. This also mimics the behavior of the wild type cells. In the presence of serine, however, the cells show no response, responding in the same fashion as the cells that were not exposed to any attractant.

Next, the chemotaxis strain containing the chemotaxis plasmid was exposed to arabinose. They behaved in an identical fashion to the strain that was not exposed to arabinose, and the wild type cells, in the absence of any attractant, forming a circular pattern (Figure 5). Addition of aspartate to the plate has no effect on the chemotaxis strain, showing that these cells are now unable to migrate towards this nutrient. Exposing the cells to serine, however, causes them to migrate towards this nutrient with a stronger phenotype than wild type cells.

Conclusion and Future Directions

These experiments show that we successfully created a strain of *E. coli* that is unable to respond to either serine or aspartate unless it is directly to do so by an outside, human controlled signal. While the development of a remotely controlled strain of *E. coli* is an interesting result, we realize this is only a toy problem without many downstream functions for our new strain. This work, however, produced a modular switch with infinite hysteresis, capable of controlling almost any expression system. Our chemical signal activates a promoter, which is can be exchanged at any time. Although we used arabinose in our system as a proof of principle, it is reasonable to believe that any signal including light, various chemicals, or temperature could be employed to steer this engineered strain. Also, during the construction of this system we developed a tightly inducible system with a memory function that propagates its own state from one generation to the next without the need for a constant signal or accessory proteins. Although these experiments produced very nice behavior from a model system, there are many long-term applications of both the technology developed and the lessons learned during this process.

Methods

Genomic knock outs: Strain BW28357 was obtained from the Coli Genetic Stock Center at Yale (CGSG) and all genes were knocked out according to the method of Datsenko and Wanner (4). After confirming each knock out with genomic sequencing, P1 phage was used to transfer each knock out to a recombinase free strain of BW28357. After knocking out the gene of interest we removed the kanamycin antibiotic resistance marker using plasmid pcp20 according to the method of Datsenko and Wanner. All strains were confirmed to be sensitive to kanamycin.

Chemotaxis Mobility Assay: The chemotaxis assay is a modified version of that used by Goulian et al (5). 72 hours prior to starting the assay, the chemotaxis plasmid was transformed into our chemotaxis strain and plated on ampicillin selective LB-Agar. Cells were incubated overnight after which a colony was picked into M9 Media and then grown overnight at 37° C. 25 mL of Minimal A Media Agar plus antibiotic was poured into 13 cm X 13 cm plates either with or without 100 mM arabinose. After drying, attractant was added by placing 10 uL drops every half centimeter down the middle of the plate and allowing attractant to dry before plates were placed at 4° Celsius for 12 – 16 hours. The starter M9 culture was spun down, then washed twice in 1X MinA salts, re-suspended in Minimal A Media, and diluted an OD of 0.1. Following that, cells were grown at 37° Celsius to an OD between 0.5 and 0.8 before again being spun down, re-suspended in Minimal A Salts and diluted to a OD of 0.25. 10 uL of culture was deposited onto the Minimal A Agar plate at a distance of 2 cm from the line of attractant. Plates were then air dried and incubated at 30° Celsius for 36 hours. Bacteria on the plates were imaged using ultraviolet light and a camera.

Strains: TOP10 (DH10B, Invitrogen) cells were used for all cloning. MG1655 was used to optimize chemotaxis assay and for genomic DNA preparations to obtain template for PCR.

DNA Synthesis: All DNA synthesis for this project was conducted by DNA 2.0 (Villalobos *et al*, 2000).

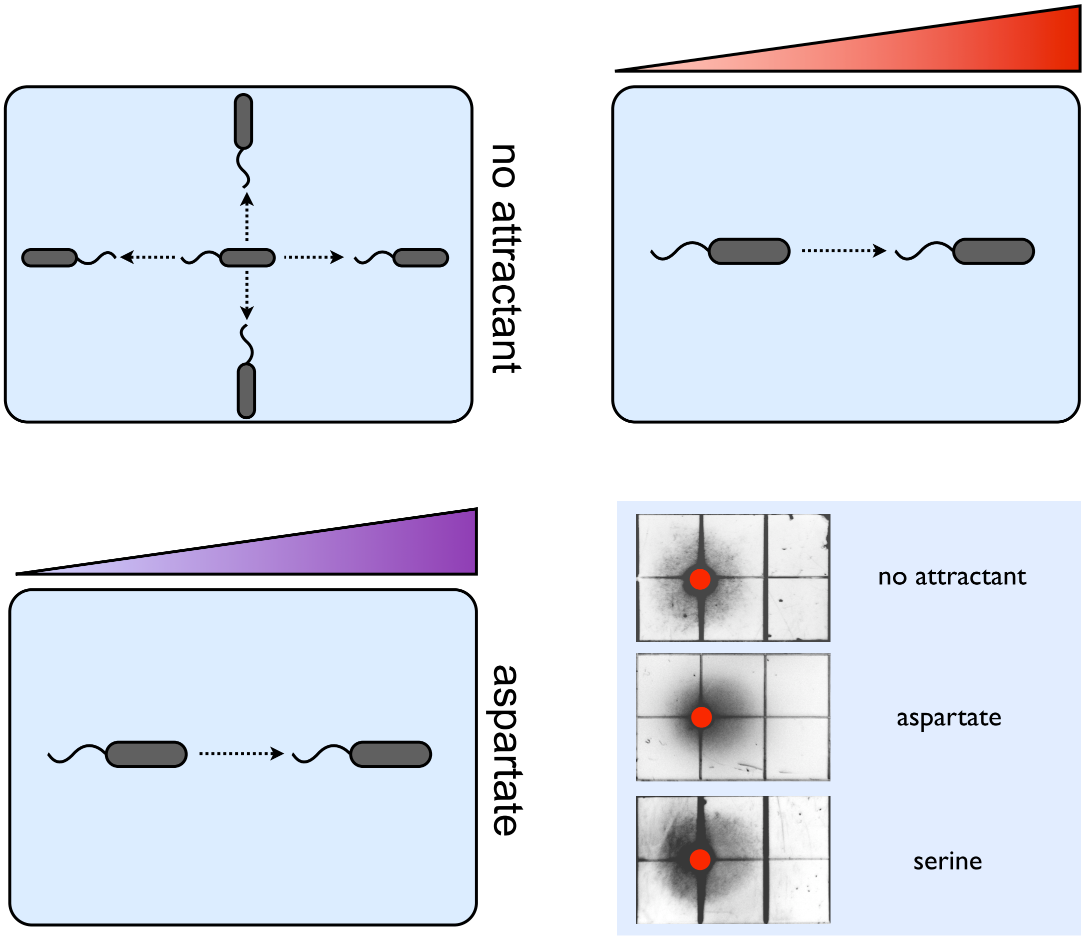
DNA Sequencing: All DNA sequencing reactions were performed by Quintara Biosciences.

Solutions/Media: All cloning was performed in 2YT liquid media (31 g 2YT per 1 liter, Teknova). Luria-Bertani Agar (LB-Agar; 10.0 g Tryptone, 5.0 g Yeast Extract, 10.0 g NaCl, 17.0 g Bacto Agar per 1 L). M9 Media (48 mM Na_2HPO_4, 22 mM KH_2PO_4, 9 mM NaCl, 19 mM NH_4Cl, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.4% glucose, 3.8 mM Thiamine, 134 mM Methionine, 129 mM Histidine, 153 mM Leucine, 168 mM Threonine, pH 7.4). Minimal A Media (33 mM KH_2PO_4, 60 mM K_2HPO_4, 8 mM $(NH_4)_2SO_4$, 2 mM Sodium Citrate, 1 mM $MgSO_4$, 0.5 mM $CaCl_2$, 0.25 mM $ZnSO_4$, 0.5% glycerol, 3.8 mM Thiamine, 134 mM Methionine, 129 mM Histidine, 153 mM Leucine, 168 mM Threonine). Minimal A Media Agar (Minimal A Media supplemented with 3.15 g Bacto Agar per liter). Ampicillin used at concentration of 100 ug/mL. Kanomycin used at concentration of 25 ug/mL.

Reference

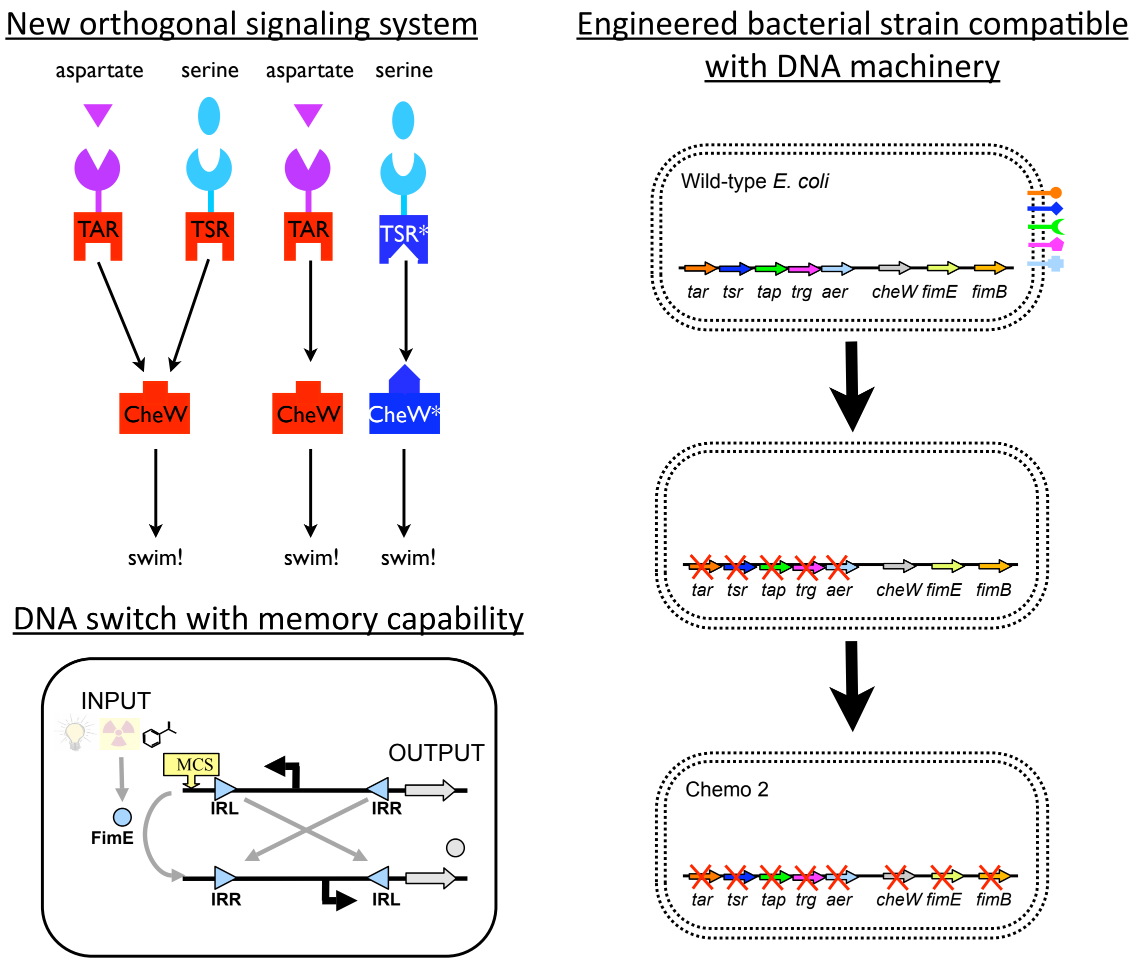
1. **Anderson, J.** 2006, posting date. [Online.]
2. **Blomfield, I. C., M. S. McClain, J. A. Princ, P. J. Calie, and B. I. Eisenstein.** 1991. Type 1 fimbriation and fimE mutants of Escherichia coli K-12. J Bacteriol **173:**5298-307.
3. **Clegg, D. O., and D. E. Koshland, Jr.** 1984. The role of a signaling protein in bacterial sensing: behavioral effects of increased gene expression. Proc Natl Acad Sci U S A **81:**5056-60.
4. **Datsenko, K. A., and B. L. Wanner.** 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A **97:**6640-5.
5. **Derr, P., E. Boder, and M. Goulian.** 2006. Changing the specificity of a bacterial chemoreceptor. J Mol Biol **355:**923-32.
6. **Guzman, L. M., D. Belin, M. J. Carson, and J. Beckwith.** 1995. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J Bacteriol **177:**4121-30.
7. **Ham, T. S., S. K. Lee, J. D. Keasling, and A. P. Arkin.** 2006. A tightly regulated inducible expression system utilizing the fim inversion recombination switch. Biotechnol Bioeng **94:**1-4.
8. **Hazelbauer, G. L., H. C. Berg, and P. Matsumura.** 1993. Bacterial motility and signal transduction. Cell **73:**15-22.
9. **Hazelbauer, G. L., J. J. Falke, and J. S. Parkinson.** 2008. Bacterial chemoreceptors: high-performance signaling in networked arrays. Trends Biochem Sci **33:**9-19.
10. **Klemm, P.** 1986. Two regulatory fim genes, fimB and fimE, control the phase variation of type 1 fimbriae in Escherichia coli. Embo J **5:**1389-93.
11. **Krikos, A., N. Mutoh, A. Boyd, and M. I. Simon.** 1983. Sensory transducers of E. coli are composed of discrete structural and functional domains. Cell **33:**615-22.
12. **Liu, J. D., and J. S. Parkinson.** 1991. Genetic evidence for interaction between the CheW and Tsr proteins during chemoreceptor signaling by Escherichia coli. J Bacteriol **173:**4941-51.
13. **Pallesen, L., O. Madsen, and P. Klemm.** 1989. Regulation of the phase switch controlling expression of type 1 fimbriae in Escherichia coli. Mol Microbiol **3:**925-31.
14. **Parkinson, J. S., S. R. Parker, P. B. Talbert, and S. E. Houts.** 1983. Interactions between chemotaxis genes and flagellar genes in Escherichia coli. J Bacteriol **155:**265-74.
15. **Villalobos, A., J. E. Ness, C. Gustafsson, J. Minshull, and S. Govindarajan.** 2006. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. BMC Bioinformatics **7:**285.

Figures



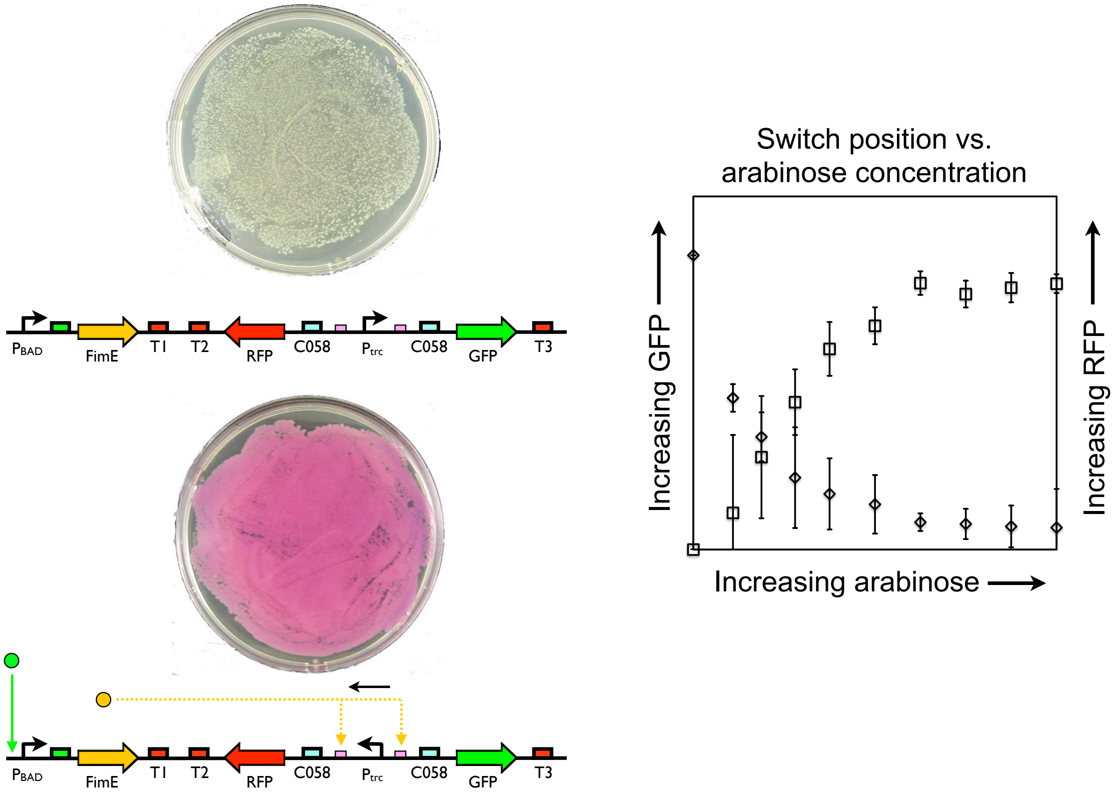
Chapter 7: Figure 1. Chemotaxis of Wild Type MG1655 *E. coli*

Top Left: Schematic diagram of MG1655 cells under conditions of no attractant. This causes the cells to move in random directions which end up in a circular pattern on a swarm plate. Bottom Left: Schematic diagram of MG1655 cells in the presence of the amino acid aspartate. The cells chemotax towards the attractant in this situation, moving up a concentration gradient. Top Right: Schematic diagram of MG1655 cells in the presence of the amino acid serine. Under this condition, the cells chemotax towards the side of the plate with serine, moving up a concentration gradient. Bottom Right: Experiments showing MG1655 responding to different environments on semi-solid agar. In the top panel the cells are placed on a plate without any attractant and form a circle. On the middle and bottom panels the cells are placed on a plate with either serine or aspartate and they move towards the attractant, creating a swarm in that direction.


New orthogonal signaling system
aspartate
serine
aspartate
serine
TAR
TSR
TAR
TSR*
CheW
CheW
CheW*
swim!
swim!
swim!
DNA switch with memory capability
INPUT
MCS
OUTPUT
IRL
IRR
FimE
IRR
IRL
Engineered bacterial strain compatible with DNA machinery
Wild-type E. coli
tar tsr tap trg aer cheW fimE fimB
tar tsr tap trg aer cheW fimE fimB
Chemo 2
tar tsr tap trg aer cheW fimE fimB

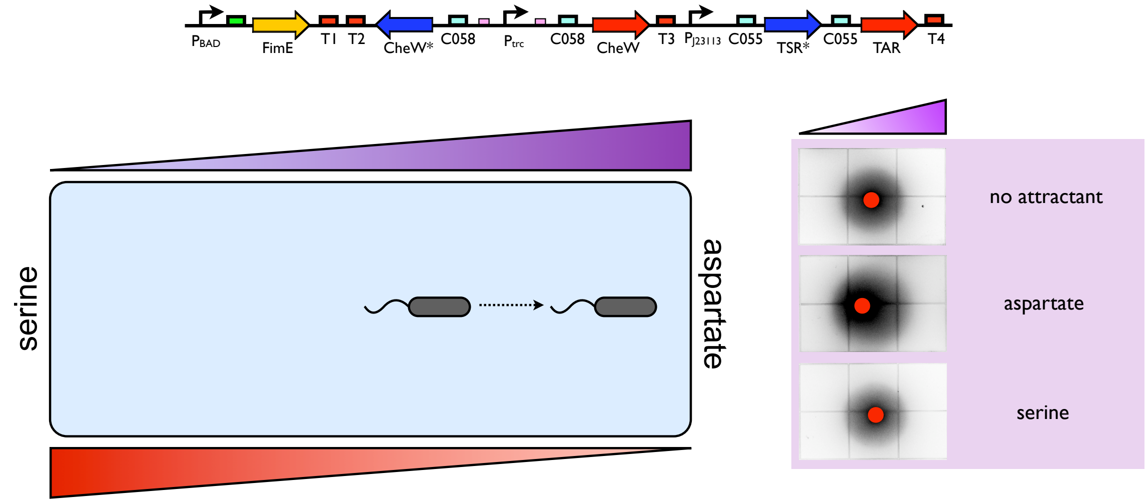
Chapter 7: Figure 2. Necessary components for Human Remote Control of Bacterial Chemotaxis.

In order to engineer remote control of bacterial chemotaxis we needed three different parts. First, we engineered an orthogonal signaling system that responds to both aspartate and serine, but can distinguish between the two molecules (Upper Left). Second, we needed a DNA switch with trans-generational memory. The pFIP switch used here is easily controllable, robust, and will maintain its position from one generation to the next. Last, we took wild type E. coli and made seven knockouts to ensure that the strain we used in our experiments would not conflict with either our switch or new chemotaxis machinery.



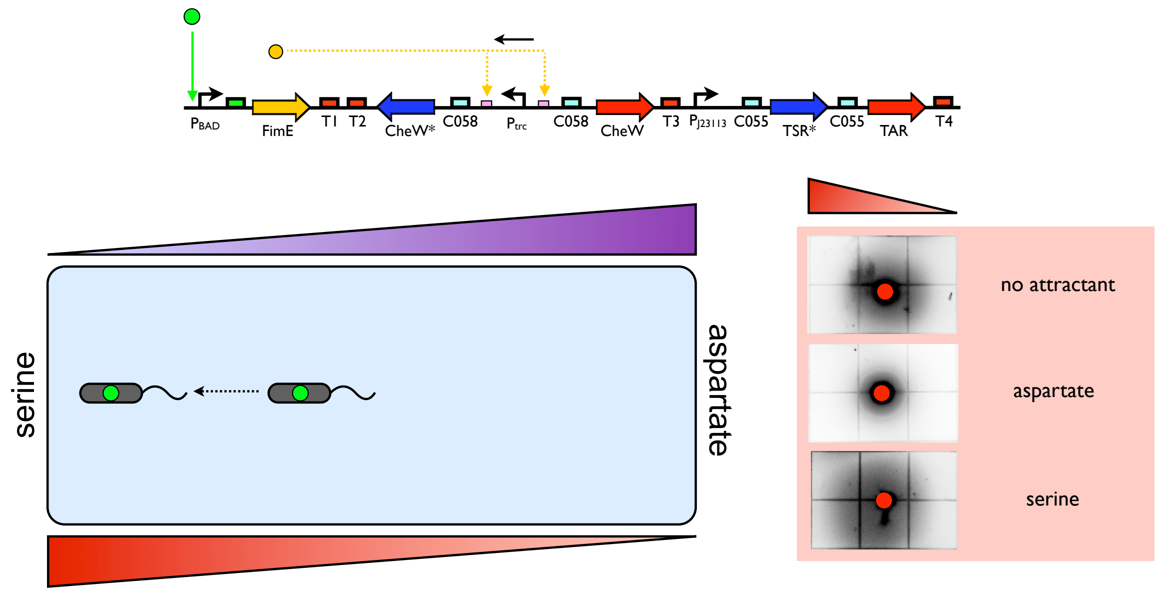
Chapter 7: Figure 3. Characterizing the pFIP DNA switch.

We placed GFP and RFP on either side of the pFIP DNA based switch. In the absence of signal, the cells express GFP and turn green (top left). In the presence of signal, in this case arabinose, the cells express RFP and turn red (bottom left). As the concentration of inducer increases, the cells produce less GFP and more RFP (right panel). This data were collected using flow cytometry and single cell analysis.



Chapter 7: Figure 4. The engineered chemotaxis system exposed to asparate and serine

The top panel is an overview of the remote control chemotaxis system. The TSR* and TAR genes are constitutively expressed in front of a PJ23113 promoter. CheW and CheW* flank a P_{trc} promoter with basal activity. In the absence of arabinose, P_{trc} drives CheW expression, which interacts with TAR. The bottom left panel shows the ideal response of our engineered cells to aspartate. The bottom right panel shows that we recover wild type behavior in the absence of any attractant and the presence of aspartate. However, in the presence of serine, the cells do not chemotax, but respond in a fashion similar to the wild type cells in the absence of attractant.



Chapter 7: Figure 5. The engineered chemotaxis strain exposed to arabinose and then asparate and serine.

The top panel shows the genetic system in the presence of arabinose. Arabinose drives the P_{BAD} promoter and allows expression of the FimE recombinase. FimE acts on recognition sites that flank the Ptrc promoter, thereby inverting it. In this situation, the cells produce CheW* instead of CheW, which interacts with TSR*. The bottom left panel shows the ideal response of our cells in the presence of arabinose and two chemical attractants. The bottom right panel shows the actual behavior of our cells in this situation. In the absence of attractant, or the presence of aspartate, the cells react similar to the wild type cells in the absence of attractant. In the presence of serine, however, the cells chemotax towards the serine amino acid.

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

____________________________ 12/22/08

Author Signature Date