AMERICAN SOCIETY FOR MICROBIOLOGY Journal of Bacteriology

Experimental Analysis of Functional Variation within Protein Families: Receiver Domain Autodephosphorylation Kinetics

Stephani C. Page,[a]* Robert M. Immormino,[b]* Thane H. Miller,[b] Robert B. Bourret[b]

Department of Biochemistry & Biophysics, University of North Carolina, Chapel Hill, North Carolina, USA[a]; Department of Microbiology & Immunology, University of North Carolina, Chapel Hill, North Carolina, USA[b]

ABSTRACT

Plants and microorganisms use two-component signal transduction systems (TCSs) to mediate responses to environmental stimuli. TCSs mediate responses through phosphotransfer from a conserved histidine on a sensor kinase to a conserved aspartate on the receiver domain of a response regulator. Typically, signal termination occurs through dephosphorylation of the receiver domain, which can catalyze its own dephosphorylation. Despite strong structural conservation between receiver domains, reported autodephosphorylation rate constants (k_{dephos}) span a millionfold range. Variable receiver domain active-site residues D + 2 and T + 2 (two amino acids C terminal to conserved phosphorylation site and Thr/Ser, respectively) influence k_{dephos} values, but the extent and mechanism of influence are unclear. We used sequence analysis of a large database of naturally occurring receiver domains to design mutant receiver domains for experimental analysis of autodephosphorylation kinetics. When combined with previous analyses, k_{dephos} values were obtained for CheY variants that contained D + 2/T + 2 pairs found in 54% of receiver domain sequences. Tested pairs of amino acids at D + 2/T + 2 generally had similar effects on k_{dephos} in CheY, PhoB$_N$, or Spo0F. Acid or amide residues at D + 2/T + 2 enhanced k_{dephos}. CheY variants altered at D + 2/T + 2 exhibited rate constants for autophosphorylation with phosphoramidates and autodephosphorylation that were inversely correlated, suggesting that D + 2/T + 2 residues interact with aspects of the ground or transition states that differ between the two reactions. k_{dephos} of CheY variants altered at D + 2/T + 2 correlated significantly with k_{dephos} of wild-type receiver domains containing the same D + 2/T + 2 pair. Additionally, particular D + 2/T + 2 pairs were enriched in different response regulator subfamilies, suggesting functional significance.

IMPORTANCE

One protein family, defined by a conserved domain, can include hundreds of thousands of known members. Characterizing conserved residues within a conserved domain can identify functions shared by all family members. However, a general strategy to assess features that differ between members of a family is lacking. Fully exploring the impact of just two variable positions within a conserved domain could require assessment of 400 (i.e., 20 × 20) variants. Instead, we created and analyzed a nonredundant database of receiver domain sequences. Five percent of D + 2/T + 2 pairs were sufficient to represent 50% of receiver domain sequences. Using protein sequence analysis to prioritize mutant choice made it experimentally feasible to extensively probe the influence of positions D + 2 and T + 2 on receiver domain autodephosphorylation kinetics.

Two-component systems (TCSs) are a prevalent means of signal transduction used by plants and microorganisms to mediate responses to stimuli (1). TCSs are present in more than 95% of sequenced bacterial genomes (2, 3), and one species can contain tens to more than a hundred TCSs. TCSs regulate a wide range of processes from cell development to virulence. Signal transduction by TCSs occurs through the transfer of phosphoryl groups between histidyl and aspartyl residues of different protein components (4). Canonically, the sensory component (the sensor kinase) is a phosphodonor to the response regulator (the response-mediating component) (4). The conserved domain in the response regulator, the receiver domain, functions as a molecular switch. The phosphorylation status of a conserved Asp on the receiver domain corresponds to turning the output response on and off. Typically, receiver domain phosphorylation initiates the output response, and dephosphorylation terminates the response. Dephosphorylation can occur with the assistance of another protein, such as a phosphatase, or by self-catalysis by the receiver domain, which is termed autodephosphorylation. Reported autodephosphorylation rate constants of receiver domains span almost 6 orders of magnitude (5) (see Fig. S1 in the supplemental material). The large variation in receiver domain autodephosphorylation kinetics is striking in light of the strong conservation of structure among receiver domains. There are now almost 300 structures of receiver domains in the RCSB Protein Data Bank (6, 7). Receiver

Received 17 October 2015 **Accepted** 28 June 2016

Accepted manuscript posted online 5 July 2016

Citation Page SC, Immormino RM, Miller TH, Bourret RB. 2016. Experimental analysis of functional variation within protein families: receiver domain autodephosphorylation kinetics. J Bacteriol 198:2483–2493. doi:10.1128/JB.00853-15.

Editor: A. M. Stock, Rutgers University-Robert Wood Johnson Medical School

Address correspondence to Stephani C. Page, stephani_page@med.unc.edu.

* Present address: Stephani C. Page, Department of Pharmacology, University of North Carolina, Chapel Hill, North Carolina, USA; Robert M. Immormino, Center for Environmental Medicine, Asthma, and Lung Biology, University of North Carolina, Chapel Hill, North Carolina, USA.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.00853-15.

domains have a conserved Rossmanoid fold structure, with a β-sheet made up of five β-strands, surrounded by five α-helices (8). Five conserved residues and a divalent metal ion are arranged in a conserved geometry and comprise the active site that catalyzes both the phosphorylation and dephosphorylation of the conserved Asp (8). Despite the conserved fold and active-site geometry, on average any two receiver domains will have only about 25% amino acid sequence identity (9), suggesting that there is a fair amount of variability from one receiver domain to the next. Variable residues located within or close to the receiver domain active site could potentially be positioned to exert influence on the catalysis of autodephosphorylation. Residues at D + 2 (located two amino acids C terminal to the site of phosphorylation on the β3-α3 loop) and T + 2 (located two amino acids C terminal to the conserved Thr/Ser on the β4-α4 loop) are located such that their side chains may potentially interact with conserved active-site residues, the phosphoryl group, and/or the attacking water nucleophile. Furthermore, mutual information analysis suggests a high degree of coevolution between the amino acids at positions D + 2 and T + 2, implying functional importance. In previous studies using a limited set of *Escherichia coli* CheY and *Bacillus subtilis* Spo0F mutants based on the wild-type sequences of fewer than 10 receiver domains, we found that the particular amino acids at positions D + 2 and T + 2 altered the autodephosphorylation rate constant by almost 2 orders of magnitude (10, 11). Due to the small data set, the full extent to which residues at D + 2 and T + 2 can influence receiver domain autodephosphorylation is unknown. Further, the mechanisms by which positions D + 2 and T + 2 influence autodephosphorylation remain unclear.

In this study, we extended investigation of the D + 2 and T + 2 residues and their influence on receiver domain autophosphorylation by significantly expanding our data set to be much more representative of response regulators. Pursuing a larger and more relevant data set would potentially mean that conclusions could be applied more broadly to receiver domains. Protein sequence analysis revealed that 20 (out of 400 possible) D + 2/T + 2 pairs account for 50% of receiver domain sequences found in a nonredundant database of naturally occurring response regulators; in essence, only 5% of possible amino acid pairs represent the majority of receiver domains. This circumstance made expansion of the data set experimentally feasible. We also expanded our experimental analysis to include another receiver domain, $\mathrm{PhoB_N}$, representative of a large family of response regulators. Combined with previous studies, the 44 D + 2/T + 2 pairs tested in CheY represent 54% of receiver domain sequences. Analysis of the expanded mutant collection showed that residues at D + 2 and T + 2 modulated autodephosphorylation rate constants (k_{dephos}) over 2 orders of magnitude, indicating that additional factors are required to account for the much larger range of k_{dephos} values reported for wild-type response regulators. Nevertheless, k_{dephos} of CheY mutants were correlated ($R^2 = 0.62$) with k_{dephos} of wild-type response regulators bearing the same amino acids at D + 2/T + 2. Furthermore, for the 11 cases tested, a particular pair of amino acids at D + 2 and T + 2 generally had similar effects on autodephosphorylation in CheY and PhoB or Spo0F. Negatively charged amino acids (and, to a lesser extent, amide residues) at D + 2 or T + 2 enhanced autodephosphorylation in all three receiver domains tested, but positively charged amino acids did not have a consistent effect on hydrolysis. The previously measured rate constants for autophosphorylation with phosphoramidate or monophosphoimidazole by CheY variants that differ at D + 2 and T + 2 (12, 13) are inversely correlated with autodephosphorylation rate constants of the same variants, suggesting that the amino acids at D + 2 and T + 2 interact with aspects of the ground or transition states that differ between the two reactions.

Finally, analysis of receiver domains from different response regulator subfamilies revealed dramatically different distributions of D + 2/T + 2 pairs between receiver domain subclasses. Typically, a few pairs composed of chemically similar amino acids and exerting similar effects on the autodephosphorylation rate constant dominated the D + 2/T + 2 pairs in each response regulator subfamily. The biased distribution strengthens the notion that amino acids at positions D + 2 and T + 2 are important for response regulator function.

MATERIALS AND METHODS

Mutagenesis and protein purification. Plasmids encoding $\mathrm{His_6}$-tagged CheY variants were made using QuikChange (Agilent) with plasmid pKC1 (14) as the template. CheA was purified as described previously (11). Each CheY variant was purified as described previously (5), and removal of the $\mathrm{His_6}$ tag by thrombin cleavage left three additional residues (GSH) on the N terminus. The additional GSH does not affect CheY autodephosphorylation kinetics (5). Plasmids encoding $\mathrm{His_6}$-tagged $\mathrm{PhoB_{1\text{-}127}}$ variants used pET28a-$\mathrm{phoB_N}$ (14), which encodes a thrombin-cleavable $\mathrm{His_6}$-tagged $\mathrm{PhoB_{1\text{-}127}}$, as a template. Thrombin cleavage leaves the N-terminal GSH. PhoB variants were purified as described for CheY. $\mathrm{His_6}$-tagged $\mathrm{PhoR_{193\text{-}431}}$ was expressed and purified as described previously (14). All protein variants were gel filtered using a Superdex 75 1660 size exclusion column (GE Biosciences).

We apply a nomenclature for discussion of CheY and $\mathrm{PhoB_N}$ mutants in which the first letter is the D + 2 residue and the second letter is the T + 2 residue. One D + 2/T + 2 pair in the top 20 was not assessed. EL (D + 2 is E; T + 2 is L), twelfth in abundance, was made in CheY, but attempts at protein purification were unsuccessful. Only six of pairs 21 to 34 were tested. The GY, KF, KH, RS, EY, HD, TY, and HS pairs (pairs 21 to 23, 28 to 31, and 33, respectively) were not made because they were not among the most prevalent pairs in our initial database analysis. Notably, analysis of Spo0F in reference 11 included three high-frequency pairs (EL, KH, and EY) missing from the CheY mutant set.

Autodephosphorylation rate constant measurement using ^{32}P. Autodephosphorylation rate constants for CheY variants were measured by following the loss of ^{32}P as described previously (5). CheY variants (8 μM) were incubated with 0.5 μM purified [^{32}P]CheA-P (15) in 100 mM Tris at pH 7.5 and 10 mM $\mathrm{MgCl_2}$. Because some substitutions resulted in diminished rates of phosphotransfer from [^{32}P]CheA-P to CheY, incubations times were varied for different CheY mutants in order to ensure that at least 95% of the ^{32}P was transferred by the first time point. At each time point, 6 μl of the reaction mixture was removed and mixed with an equal volume of 2× SDS sample buffer to quench the reaction. Components of samples taken at each time point were separated using SDS-PAGE. Loss of ^{32}P from [^{32}P]CheY-P was detected using phosphorimaging analysis of the dried gel. The signals were quantified using pixel volume analysis in which the background signal was manually subtracted. The amounts of [^{32}P]CheY-P were plotted versus time and fit to one-phase exponential decays yielding the k_{dephos}. Each time course was designed to follow loss of ^{32}P over at least 4 or 5 half-lives (~3 to 6% ^{32}P remaining on [^{32}P]CheY-P). All measurements were repeated three times.

To determine the autodephosphorylation rate constants of $\mathrm{PhoB_N}$ variants, 4 μM $\mathrm{His_6}$-tagged PhoR was incubated with 0.3 mM [γ-^{32}P]ATP in 3.5 mM $\mathrm{MgCl_2}$, 35 mM KCl, and 35 mM Tris (pH 8.0) for 30 min at room temperature to generate [^{32}P]PhoR-P. The reaction mixture containing [^{32}P]PhoR-P was pipetted onto a 0.22-μm polyvinylidene difluoride (PVDF) centrifugal filter column (Millipore) containing ~200 μl of a nickel-nitrilotriacetic acid (Ni-NTA)–agarose (Qiagen) slurry that was equilibrated in 35 mM Tris (pH 8.0), 3.5 mM $MgCl_2$, and 35 mM KCl buffer. After centrifugation, the column was washed multiple times with equilibration buffer to remove ATP. $PhoB_N$ (90 μM in 60 μl of equilibration buffer) was added directly to the slurry, mixed, and then incubated at room temperature for 5 min on the column to allow for sufficient phosphotransfer from [^{32}P]PhoR-P to $PhoB_N$. The column was centrifuged to elute [^{32}P]$PhoB_N$-P. Time courses were completed and phosphorimaging was analyzed as described above for CheY.

Mutual information analysis. To examine coevolution of residues within receiver domains, in October 2015 we performed mutual information analysis of the Pfam Response_reg RP15 database (16) using the MISTIC server (http://mistic.leloir.org.ar/index.php) (17).

Receiver domain sequence analysis. To analyze the frequency of amino acids that naturally occur at various positions within receiver domains, we created a searchable database of nonredundant receiver domain amino acid sequences using a combination of publicly available web-based utilities and databases along with custom Perl (v5.16) scripts. The choice of D + 2/T + 2 combinations for experimental work was guided by preliminary analyses made several years ago (data not shown). The up-to-date analysis used to generate amino acid frequencies given in this report is described below.

The sequences of 250,546 proteins containing receiver domains, identified using the Agfam signaling domain library of the Microbial Signal Transduction (MiST2.2) database (18), were provided by the curators on 11 February 2015 (see Database S2 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm). To reduce bias and redundancy from multiple closely related genomes, one genome (the first encountered in the list) was chosen for each genus and proteins from all other genomes in the same genus were discarded. This process left 39,132 sequences. Random reordering of the original list of sequences using a Fisher-Yates shuffle resulted in different genome sequences being in the final database but did not significantly change reported results.

To facilitate sequence alignment, protein sequences were trimmed to retain only the receiver domain(s). First, a unique AseqID (19) was assigned to each protein sequence database entry. The AseqIDs were then used to query the Agfam annotation in SeqDepot (19) to retrieve the location of the first and last residues of each receiver domain. Amino acids outside the receiver domain(s) were discarded. Because some proteins contain more than one receiver domain, the number of database entries increased to 41,370. USEARCH v5.1.221 (20) was used to group the receiver domain sequences into clumps of 1,000 (a manageable size for alignment) based on sequence similarity. The sequences in each clump were then aligned using MUSCLE v3.8.31 (21). To further reduce the redundancy of the database, sequences within each clump were compared in a pairwise manner. Sequences with ≥90% identity to the amino acids (gap positions were not tested) of a query sequence were discarded. This step removed 2,989 sequences (7.2% of the total) and did not significantly affect the amino acid frequencies at positions D + 2 and T + 2 in the final database. Because the clumps were created based on sequence similarity and response regulators are on average only ~25% identical (9), the screen against high sequence identity was not applied between clumps.

The positions of the five conserved active-site residues that are critical for receiver domain function (8) were then located. Each clump of aligned sequences was searched to identify the positions that had the highest number of matches to "DD" (two adjacent Asp and/or Glu residues), "D" (Asp only), "T" (Ser or Thr), and "K" (Lys only), with the constraint that the identified positions must occur in the listed order from N to C terminal. The 6,772 sequences that did not initially appear to contain all five conserved residues were cycled through the clumping, alignment, and conserved residue identification procedures again. On the second filtering attempt, 1,643 sequences met the criterion of containing all five conserved residues. Most of the sequences that passed upon rescreening came from the same few clumps as in the first attempt and appeared to have initially failed due to sequence misalignment. Sequences that passed the first and second attempts were pooled to yield a collection of 33,252 receiver domain sequences. The 5,129 sequences (13.4%) that failed to pass the second screen for the presence of all five conserved residues were presumed to be pseudo-receiver domains (22) and were not considered further.

Receiver domains share a three-dimensional structure of alternating β-strands and α-helices, with the conserved residues on the loops at the C-terminal ends of β-strands 1, 3, 4, and 5 (8). The nine loops connecting the five β-strands and five α-helices often differ in length between different receiver domains. Therefore, a header indicating the position of the DD, D, T, and K landmarks was added to each receiver domain sequence and gaps introduced during multiple-sequence alignment were removed. This left a primary receiver domain database (see Database S3 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm) in which the amino acid at a given position with respect to a landmark (e.g., D + 2, where + signifies in the C-terminal direction and − signifies in the N-terminal direction) can be identified for each entry in spite of differences in length. Sequences that match a specified search criterion (e.g., Met at position D + 2) can also be retrieved.

Finally, response regulators can be categorized based on their output domains (23). Therefore, the primary receiver domain database was subdivided into secondary databases representing the major classes of response regulators. The AseqID was used to query Seqdepot for the Pfam (16) domain annotation associated with each protein in the primary database. The presence of specific domains led to assignment to subfamilies as follows: PF00486 (Trans_reg_C) domain, OmpR subfamily; PF00196 (GerE) domain, NarL/FixJ subfamily; both PF00158 (Sigma54_activat) and PF02954 (HTH_8) domains, NtrC subfamily; PF04397 (LytTr) domain, LytR subfamily; and PF02518 (HATPase_c) domain, hybrid kinase subfamily. Proteins that had exactly one receiver domain in Agfam annotation and no domains other than PF00072 (Response_reg) in Pfam were assigned to the single receiver domain subfamily. A total of 188 proteins (~0.6%) were assigned to more than one family; 6,697 proteins were not assigned. The sorting process resulted in seven secondary databases (the six families listed above plus the proteins without assignment) that can be individually searched (see Databases S4 to S10 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm).

RESULTS

Amino acids at receiver domain positions D + 2 and T + 2 have high mutual information content. If two variable positions within a protein participate in a functionally important interaction, then the amino acids at the two positions would be expected to covary during evolution, as a change at one position leads to selection of a compensating change at the other position (17). When the mutual information content of all 5,995 possible pairwise comparisons within receiver domains was determined, most (10/13) of the pairs with Z-scores over 100 involved interactions that appeared to be important for protein structure. Five were between the β-sheet in the core of the receiver domain and an adjacent α-helix, three were within the β-sheet, one was between the α1- and α2-helices, and one was between positions K + 1 (Pro in 82% of receiver domains) and K + 2 (data not shown). Positions D + 2 and T + 2, located on loops in the active site, stood out as having the highest mutual information content of the remaining top pairs (Z-score of 163, sixth highest overall). The other two high-scoring pairs involved positions D + 5 and D + 6 in the β3-α3 loop interacting with residues at the C-terminal end of α2. The high mutual information content of variable positions D + 2 and T + 2, combined with their location in the active site, suggests that the amino acids at D + 2 and T + 2 are potentially important for receiver domain function and good candidates for experimental investigation.

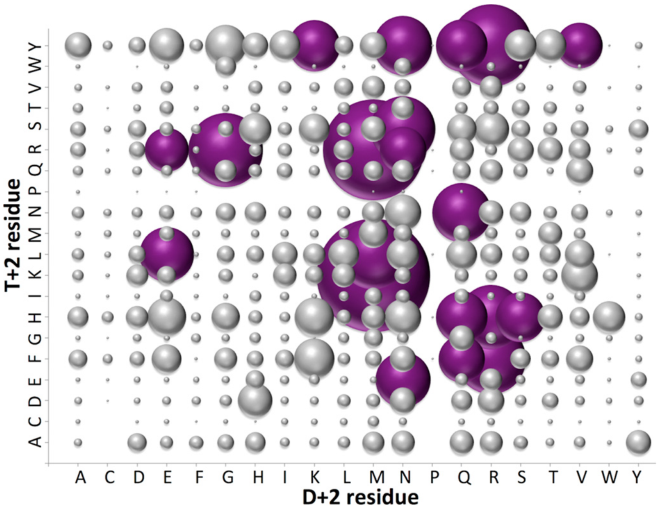


FIG 1 Bubble plot showing distribution of D + 2/T + 2 pairs found in a nonredundant receiver domain database (see Database S3 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm). The position of the bubble corresponds to a certain D + 2 (*x* axis)/T + 2 (*y* axis) pair. The area of the bubble reflects the frequency of that pair in the database. In purple are the 5% of pairs that represent 50% of receiver domain sequences.

Twenty D + 2/T + 2 pairs account for half of receiver domain sequences. Previous analysis of the influence of D + 2 and T + 2 residue pairs on receiver domain autodephosphorylation (11) was limited to pairs that mimicked fewer than 10 receiver domains. Protein sequence analysis was employed to guide experimental design for assessing the impact of D + 2/T + 2 residues in a manner that represented receiver domains more broadly and that was experimentally feasible. After accessing receiver domain sequences from the MiST2 database (18), a nonredundant database of 33,252 receiver domain sequences was created by reducing sequences to one genome per genus and then removing sequences with ≥90% sequence identity (see Database S3 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm). Out of 400 possible amino acid combinations at D + 2 and T + 2, 364 pairs were present (Fig. 1). Twenty D + 2/T + 2 pairs (only 5% of the possible combinations) accounted for 50% of natural receiver domain sequences. The pair frequencies are displayed in Fig. 1 such that the bubble area reflects the total number of sequences in our database that contain a specific D + 2/T + 2 pair (see Table S11 in the supplemental material). Some pairs, such as MK (the first letter is the D + 2 residue and the second letter is the T + 2 residue) and MR, were quite abundant, each accounting for more than 5% of sequences analyzed. CheY mutants were made with the goal of completing a set of the most common D + 2/T + 2 pairs. In the end, the collection included 19 of the top 20 and 25 of the 34 most frequent D + 2/T + 2 pairs (the wild-type CheY pair NE was pair 13). Numerous single mutants representing less abundant D + 2/T + 2 pairs were also made and tested en route to making desired D + 2/T + 2 pairs in CheY.

Substitutions at D + 2 and T + 2 in CheY modulated autodephosphorylation rate constants over 2 orders of magnitude. Autodephosphorylation rate constants were measured for 26 CheY variants. With consolidated CheY data from references 10, 11, and 13, the expanded data set (Table 1) contained 12 D + 2 single mutants, 10 T + 2 single mutants, and 21 double mutants in CheY that, together (including the NE pair found in wild-type CheY), represent 54% of receiver domains. At T + 2, amino acids L, K, R, H, and Y were assessed in the study described in reference 11 and expanded to include Q, S, A, N, and F in this study. All T + 2 single substitutions removed the wild-type Glu in CheY and resulted in diminished autodephosphorylation rate constants (Table 1). Overall, single substitutions at T + 2 modulated rates constants over a 10-fold range, similar to what was reported in reference 11. Rate constants for T + 2 single mutants appeared to decrease as residue size and hydrophobicity increased; aromatics (H, Y, and F) made up the slowest T + 2 single mutants.

In contrast, single substitutions at D + 2 (changing the Asn) resulted in both enhanced and diminished rate constants. D, M, E, K, and L were assessed at D + 2 in the study described in reference 11, which was extended to include S, Q, V, A, R, G, and W in this study. Overall, single substitutions at D + 2 resulted in an ~30-fold range in rate constants. Almost all (11/12) D + 2 single mutants resulted in rate constants that were within 3-fold that of wild-type CheY. One D + 2 single mutant (Trp), which slowed the reaction 11-fold, was responsible for the added order of magnitude in the range of CheY D + 2 single mutants.

Double mutants DL, EH, MR, MK, and KY were analyzed in the study described in reference 11, RH and ML were analyzed in the study described in reference 10, and QS was analyzed in the study described in reference 13. This study expanded analysis of CheY double mutants to include ER, QN, SH, QH, GR, AA, QF, QY, VK, WH, RF, RY, and VY. An ~30-fold range in autodephosphorylation rate constants was observed for the consolidated CheY double mutants with simultaneous substitutions at D + 2 and T + 2. All of the CheY D + 2/T + 2 double mutants exhibited diminished autodephosphorylation rate constants compared to that of the wild type. Altogether, single and double substitutions at D + 2 and T + 2 in CheY resulted in a range in autodephosphor ylation rate constants that spans 2 orders of magnitude.

The previously studied data set included 16 CheY variants (accounting for 23% of receiver domain sequences) and resulted in an ~90-fold range in autodephosphorylation rate constants (11). Given the small sample size, it was unknown whether the previously observed range of k_{dephos} values represented an upper bound or a lower bound for the effect of D + 2 and T + 2 residues on autodephosphorylation. Increasing the data set to 44 CheY variants (54% of receiver domain sequences) expanded the range of CheY autodephosphorylation rate constants only ~30% to 120-fold. The expanded data set is more representative of naturally occurring D + 2/T + 2 pairs than the original and includes D + 2/T + 2 pairs found in wild-type response regulators exhibiting a range in autodephosphorylation rate constants of 4 orders of magnitude. Therefore, we can reasonably expect to attribute only about 2 orders of magnitude of the range in wild-type response regulator autodephosphorylation rate constants to the particular amino acids at positions D + 2 and T + 2 and conclude that additional factors must contribute substantially to variation in the rate of autodephosphorylation between receiver domains.

Substitution mimicking CheY at T + 2 in PhoB$_N$ enhanced autodephosphorylation. *E. coli* PhoB$_N$, a receiver domain with a relatively low autodephosphorylation rate constant (14), was used to analyze substitutions that could potentially enhance autodephosphorylation. This expands a previous approach using the *B. subtilis* Spo0F response regulator (11). PhoB was chosen for mul-

TABLE 1 CheY autodephosphorylation rate constants

	Amino acid at:			
CheY variant(s)	D + 2	T + 2	k_{dephos} (min^{-1})[a]	Fold change from wild type
Wild type	N	E	2.2 ± 0.2	NA[e]
T + 2 single mutants	N	Q	2.1 ± 0.07	−1.0
	N	L	1.6 ± 0.1[b]	−1.4
	N	S	1.2 ± 0.04	−1.8
	N	A	1.2 ± 0.1	−1.8
	N	K	1.2 ± 0.3[b]	−1.8
	N	N	0.88 ± 0.04	−2.5
	N	R	0.69 ± 0.04[b]	−3.2
	N	H	0.55 ± 0.04[b]	−4.1
	N	Y	0.26 ± 0.01[b]	−8.5
	N	F	0.20 ± 0.04	−11
D + 2 single mutants	D	E	5.5 ± 0.4[b]	+2.5
	M	E	4.6 ± 0.4[b]	+2.1
	E	E	4.5 ± 0.2[b]	+2.0
	S	E	3.3 ± 0.2	+1.5
	Q	E	3.2 ± 0.4	+1.4
	K	E	3.0 ± 0.1[b]	+1.4
	V	E	2.2 ± 0.2	−1.0
	A	E	1.8 ± 0.2	−1.2
	R	E	1.6 ± 0.04	−1.4
	L	E	1.3 ± 0.1[b]	−1.7
	G	E	1.0 ± 0.1	−2.2
	W	E	0.20 ± 0.02	−11
Double mutants	E	R	1.3 ± 0.1	−1.7
	Q	S	1.1 ± 0.2[c]	−2.0
	Q	N	0.91 ± 0.04	−2.4
	D	L	0.57 ± 0.08[b]	−3.9
	E	H	0.50 ± 0.03[b]	−4.4
	S	H	0.39 ± 0.05	−5.6
	Q	H	0.30 ± 0.03	−7.2
	G	R	0.30 ± 0.04	−7.2
	A	A	0.30 ± 0.02	−7.3
	Q	F	0.28 ± 0.04	−7.8
	Q	Y	0.22 ± 0.02	−9.9
	V	K	0.10 ± 0.006	−21
	W	H	0.098 ± 0.01	−22
	M	R	0.094 ± 0.012[b]	−23
	M	K	0.078 ± 0.004[b]	−28
	K	Y	0.062 ± 0.006[b]	−35
	R	H	0.060 ± 0.008[d]	−37
	M	L	0.060 ± 0.006[d]	−37
	R	F	0.056 ± 0.003	−39
	R	Y	0.048 ± 0.004	−46
	V	Y	0.045 ± 0.008	−49

[a] Values are means ± standard deviations.
[b] Autodephosphorylation rate constant from reference 11 determined by loss of ^{32}P.
[c] Autodephosphorylation rate constant from reference 13 determined by fluorescence, which typically gives slightly (<2×) higher values than measurement by loss of ^{32}P (5).
[d] Autodephosphorylation rate constant from reference 10 determined by loss of ^{32}P. The corresponding mutant also contains Glu rather than Phe at position 14 (DD + 1), which has little (<2×) effect on k_{dephos} (11).
[e] NA, not applicable.

TABLE 2 PhoB$_N$ autodephosphorylation rate constants

	Amino acid at:			
PhoB$_N$ variant(s)	D + 2	T + 2	k_{dephos} (min^{-1})[a]	Fold change from wild type
Wild type	M	R	0.015 ± 0.002	NA
T + 2 single mutants	M	E	0.67 ± 0.06	+45
	M	A	0.024 ± 0.003	+1.6
D + 2 single mutants	N	R	0.024 ± 0.003	+1.6
	A	R	0.018 ± 0.003	+1.2
Double mutants	N	E	0.27 ± 0.03	+18
	A	A	0.024 ± 0.004	+1.6

[a] Values are means ± standard deviations.

tiple reasons. While CheY and Spo0F are single-domain response regulators, PhoB has an output domain and is representative of the large OmpR class of response regulators. Although the presence or absence of the output domain has little effect on PhoB autodephosphorylation (14), receiver and output domains within response regulators have coevolved (24, 25), and D + 2/T + 2 residues conceivably could have different effects in PhoB$_N$ than in a single-domain response regulator. Another contrast with CheY and Spo0F is that PhoB$_N$ forms dimers (26). To the best of our knowledge, PhoB$_N$ is the only response regulator for which autodephosphorylation rate constants have been measured in different multimeric states. The autodephosphorylation rate constants for monomeric PhoB$_N$-P, the PhoB$_N$·PhoB$_N$-P heterodimer, and the PhoB$_N$-P·PhoB$_N$-P homodimer are indistinguishable (14), removing a potential complication from interpretation of data.

Single and double substitutions at D + 2 and T + 2 were made in PhoB$_N$ (changing Met and Arg at D + 2 and T + 2, respectively) that mimic CheY (Table 2). Mutants with Glu substitutions at T + 2 resulted in the largest increases compared to the wild-type PhoB$_N$ autodephosphorylation rate constant (PhoB$_N$ NE and PhoB$_N$ ME autodephosphorylated 18- and 45-fold faster, respectively, than wild-type PhoB$_N$).

The PhoB$_N$ mimics of CheY were compared to single and double Ala mutants of PhoB$_N$. Steric occlusion, i.e., hindrance of the in-line attack of nucleophilic water by large residues, was previously hypothesized as one mechanism by which D + 2 and T + 2 residues might influence receiver domain autodephosphorylation (10). Based on this hypothesis, removal of the large Arg and/or Met residues by substitution with Ala should result in enhanced autodephosphorylation. The autodephosphorylation rate constant of PhoB$_N$ AR was unchanged from that of wild-type PhoB$_N$, whereas rate constants for PhoB$_N$ MA and PhoB$_N$ AA were both 2-fold faster than that of wild-type PhoB$_N$. The relatively small effects of single and double Ala substitutions in PhoB$_N$ suggest that steric obstruction is not the key mechanism of influence by D + 2/T + 2 residues. Because introducing a Glu at T + 2 had a much larger effect than an Ala, it is clear that the key influence in the enhanced autodephosphorylation rate constant of PhoB$_N$ ME was from adding a beneficial feature of Glu, as opposed to removing a detrimental feature of Arg.

Autodephosphorylation rate constants appear to be influenced by negative charge at positions D + 2 and/or T + 2 but not by positive charge. The magnitudes of effects between the same D + 2/T + 2 pairs in CheY and PhoB$_N$ were similar (Fig. 2). Generally, similar effects of D + 2/T + 2 pairs were also previously observed between CheY and Spo0F (11) (Fig. 2). The observation that the effects were similar regardless of backbone suggests that there may be some generality to the effect of a particular D + 2/

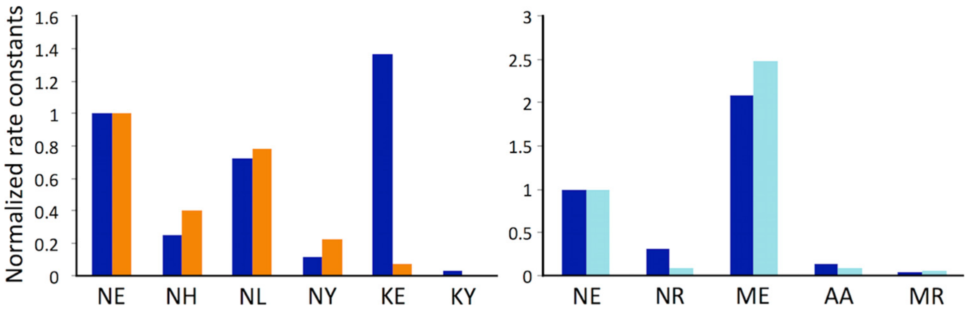


FIG 2 Normalized autodephosphorylation rate constants for the same D + 2/T + 2 pairs in different receiver domains. Rate constants for variants containing a specific D + 2/T + 2 pair are normalized using the NE variant for each receiver domain. Variants are identified by single letter amino acid abbreviations, where the first letter is the D + 2 residue and the second letter is the T + 2 residue. Dark blue represents CheY variants (wild-type NE). Orange represents Spo0F variants (wild-type KY). Cyan represents $PhoB_N$ variants (wild-type MR). *y* axis scales are different between each data set.

T + 2 pair. Kinetic data were analyzed to probe for mechanistic insight that could be applied to all three receiver domain backbones. Data were analyzed to determine whether kinetic effects correlated with features of the amino acids (such as size, solvent accessibility, and hydrophobic surface area) present at D + 2 and T + 2. Though relationships were not observed for other features, a plot of autodephosphorylation rates constants against net charge at D + 2 and T + 2 revealed that more negatively charged active sites weakly correlated with higher autodephosphorylation rate constants (best fit $R^2 = 0.44$) (Fig. 3). Deeper inspection provides additional evidence that negative charge had a role in modulating autodephosphorylation. There are 13 CheY variants with Glu at T + 2 (Table 1). Because the D + 2 residue was varied experimentally, for purposes of discussion, this class of variants is designated XE. Of the XE variants, 12 have one or more related double mutants in which the Glu is replaced with another amino acid, designated XZ. In all 31 of the CheY XE-versus-XZ comparisons, the autodephosphorylation rate constant is higher for the XE variant (Fig. 4A). Because Glu is the wild-type residue at T + 2 in CheY, the result could be explained as loss of wild-type function. However, in both $PhoB_N$ (3/3) and Spo0F (6/6) (Fig. 4A), in which Glu is not the wild-type residue at T + 2, all of the XE mutants tested resulted in a gain of function compared to the corresponding XZ mutants.

Acidic residues at D + 2 similarly correlated with higher autodephosphorylation rate constants. The CheY variants with an acidic residue at D + 2 are designated EX, where E represents either Asp or Glu. There are 34 comparisons between EX pairs in CheY and variants that have the D + 2 residue replaced, designated ZX (Fig. 4B). In 30/34 comparisons, the EX variant exhibited a higher k_{dephos} than the ZX variant. Two of the four exceptions have a negatively charged Glu at the "X" (T + 2) position and the other two exceptions have the wild-type amide (Asn) at D + 2. In Spo0F, acidic substitutions at D + 2 also resulted in faster autodephosphorylation (11). There are two Spo0F EX variants that can be compared to ZX variants with the same T + 2 residue. In all four Spo0F comparisons, the EX variants autodephosphor ylated faster than the corresponding ZX variants (Fig. 4B). There were no $PhoB_N$ mutants in this study that contain an acidic residue at D + 2.

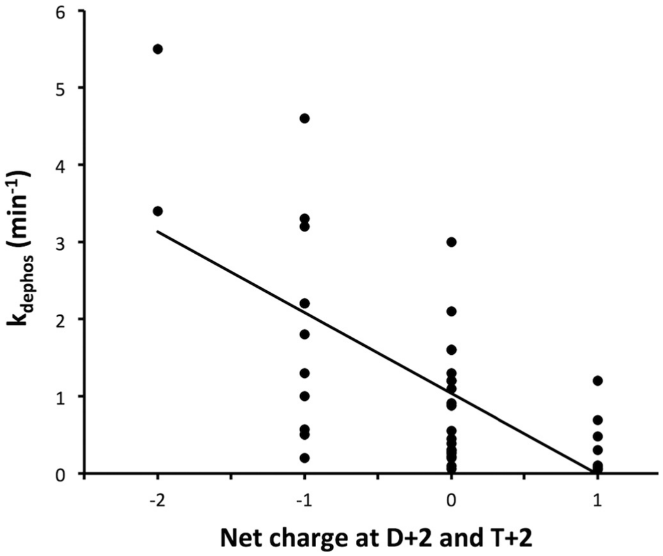
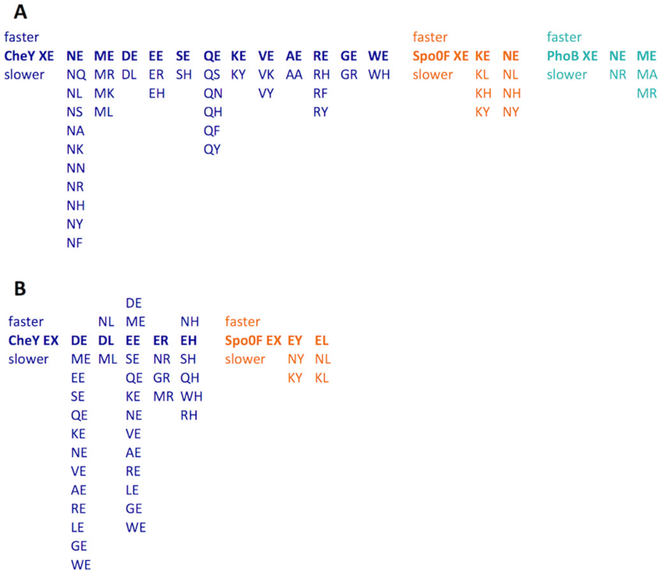


FIG 3 Correlation between net charge at positions D + 2 and T + 2 and CheY autodephosphorylation rate constants. His was considered neutral for calculations of net charge. Best fit $R^2 = 0.44$.



A

faster												
CheY XE	**NE**	**ME**	**DE**	**EE**	**SE**	**QE**	**KE**	**VE**	**AE**	**RE**	**GE**	**WE**
slower	NQ	MR	DL	ER	SH	QS	KY	VK	AA	RH	GR	WH
	NL	MK		EH		QN		VY		RF		
	NS	ML				QH				RY		
	NA					QF						
	NK					QY						
	NN											
	NR											
	NH											
	NY											
	NF											

faster		
Spo0F XE	**KE**	**NE**
slower	KL	NL
	KH	NH
	KY	NY

faster		
PhoB XE	**NE**	**ME**
slower	NR	MA
		MR

B

			DE		
faster		NL	ME		NH
CheY EX	**DE**	**DL**	**EE**	**ER**	**EH**
slower	ME	ML	SE	NR	SH
	EE		QE	GR	QH
	SE		KE	MR	WH
	QE		NE		RH
	KE		VE		
	NE		AE		
	VE		RE		
	AE		LE		
	RE		GE		
	LE		WE		
	GE				
	WE				

faster		
Spo0F EX	**EY**	**EL**
slower	NY	NL
	KY	KL

FIG 4 Comparisons of effects on k_{dephos} due to removal of acidic residue at T + 2 (XE in panel A) or D + 2 (EX in panel B) in CheY, Spo0F, and $PhoB_N$.

If negatively charged amino acids at positions D + 2 or T + 2 enhance autodephosphorylation, then positively charged amino acids might be expected to diminish autodephosphorylation. This general trend was observed in Fig. 3. However, pairwise comparisons of CheY variants with or without positively charged residues analogous to that shown in Fig. 4 did not yield similar outcomes. When looking among groupings of CheY variants that included some with a positively charged residue (Arg or Lys) at D + 2 or T + 2, we did not discern a consistent relationship between kinetic data and removal or addition of a positive charge. In 36 of 56 comparisons (data not shown), the CheY variant containing Arg or Lys exhibited a lower value for k_{dephos} than corresponding variants without Arg or Lys. Further, because His has a pK_a within a reasonable pH range for autodephosphorylation experiments, we were able to directly assess pH dependence of CheY variants with a His at T + 2. Changing the pH will alter the charge of a His between neutral and positive. Autodephosphorylation rate constants of the CheY QH, RH, and WH variants were not affected by pH (data not shown), suggesting that positive charge at T + 2 may not influence CheY autodephosphorylation. In contrast to CheY, tested $PhoB_N$ and Spo0F variants containing Arg or Lys always supported slower autodephosphorylation than corresponding proteins without the positively charged residues, but the available data set is much smaller (only four comparisons for $PhoB_N$ and five for Spo0F). Arg, His, and Lys contain large hydrophobic surface areas, so hydrophobic surface area or size, rather than positive charge, may be the means by which these residues influence autodephosphorylation.

D + 2/T + 2 substitutions inversely affect CheY autodephosphorylation with water and autophosphorylation with phosphoramidate or monophosphoimidazole. In addition to autodephosphorylation, receiver domains self-catalyze phosphorylation with small-molecule phosphodonors, such as phosphoramidate (PAM), monophosphoimidazole (MPI), or acetyl phosphate (AcP). Analysis in CheY showed that the amino acids at D + 2 and T + 2 influence autophosphorylation with small-molecule phosphodonors (12). Of the CheY variants for which autodephosphorylation rate constants are available (Table 1), there were 20 CheY variants for which PAM and AcP autophosphorylation rate constants (k_{phos}/K_s) were both known and 7 for which MPI autophosphorylation rate constants were available (12, 13). For the CheY variants, there were significant correlations between k_{dephos} and k_{phos}/K_s with PAM ($R^2 = 0.51$) or MPI ($R^2 = 0.87$) but not with AcP ($R^2 = 0.18$). For both PAM and MPI, plotting k_{dephos} versus k_{phos}/K_s revealed inverse relationships between the rate constants for the two reactions (Fig. 5), as exemplified by the negative slopes of the best fit lines on log-log plots of the data (Fig. 5, insets).

The plots of k_{dephos} versus k_{phos}/K_S showed informative correlations between the amino acids found at T + 2 and the rate constants for each reaction. Pairs containing residues with large hydrophobic surface areas typically had higher PAM/MPI autophosphorylation rate constants (12) and lower autodephosphorylation rate constants (Fig. 5). Pairs containing acid or amide residues at T + 2 were typically faster for autodephosphorylation and slower for PAM/MPI autophosphorylation, with acid residues having the greatest impact.

Certain D + 2/T + 2 pairs are enriched in receiver domains from different classes of response regulators. Because response regulator subfamilies with different output domains have different functions, we wanted to determine whether particular D + 2/T + 2 pairs were associated with specific response regulator subfamilies and, potentially, with specific response regulator functions. The primary nonredundant database of receiver domain sequences (see Database S3 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm) was divided into seven secondary databases (see Databases S4 to S10 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm) based on association with major classes of response regulators as described in Materials and Methods. Subfamilies represented the following fractions of the nonredundant database: hybrid kinases, 24%; OmpR, 20%; single receiver domains, 18%; NarL/FixJ, 12%; NtrC, 6%; and LytR, 4%. Sixteen percent of sequences were not assigned to a response regulator subfamily in our analysis. The smaller databases revealed that the distributions of D + 2/T + 2 pairs were strikingly different between response regulator subfamilies, as well as between response regulator subfamilies and response regulators taken as a whole (Table 3). The functional consequences of this circumstance are considered in Discussion.

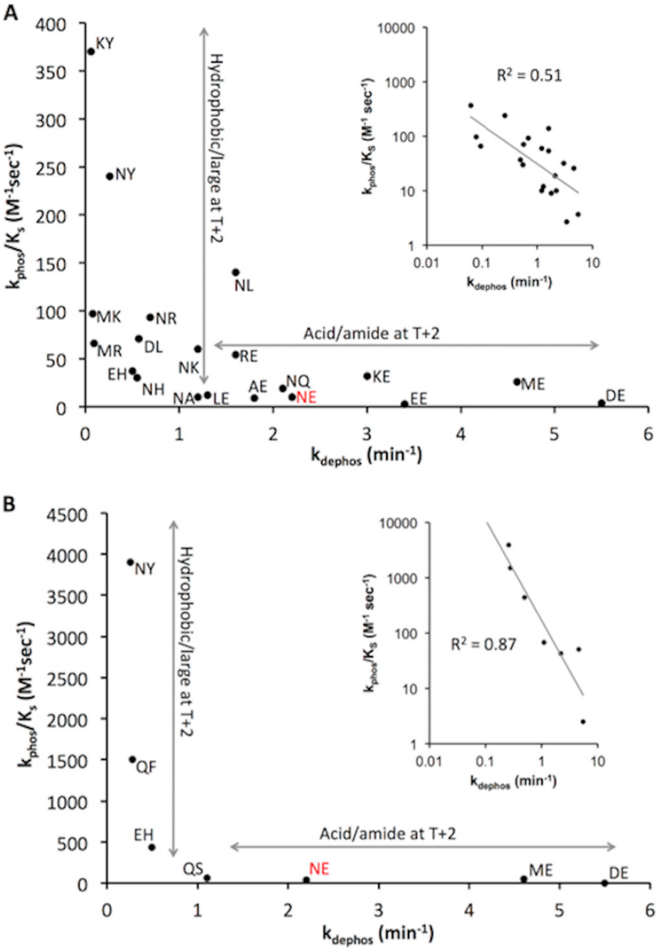


FIG 5 Inverse relationship between k_{dephos} and k_{phos}/K_S of CheY D + 2/T + 2 variants. Wild-type CheY is indicated in red. (A) Autophosphorylation with PAM. (B) Autophosphorylation with MPI. Note the difference in *y* axis scales for panels A and B. k_{phos}/K_S values are from references 12 and 13. k_{dephos} values are from Table 1. Insets show the same data replotted in log-log form. The best-fit lines through the data have the equations $k_{phos}/K_S = 31(k_{dephos})^{-0.71}$ ($R^2 = 0.51$) for PAM and $k_{phos}/K_S = 170(k_{dephos})^{-1.8}$ ($R^2 = 0.87$) for MPI.

TABLE 3 Top 10 D + 2/T + 2 pairs for response regulator subfamilies

All receivers		Hybrid kinase		OmpR		Single domain		NarL/FixJ		NtrC		LytR		Nonassigned	
D + 2/T + 2[a]	%[b]	D + 2/T + 2	%	D + 2/T + 2	%	D + 2/T + 2	%	D + 2/T + 2	%	D + 2/T + 2	%	D + 2/T + 2	%	D + 2/T + 2	%
MK	7.7	**QN**	8.0	**MK**	24	**NS**	10	**RF**	16	**RY**	16	QY	19	EL	9.2
MR	6.0	**MK**	5.4	**MR**	24	NE	7.8	**RY**	8.1	**RH**	13	EY	11	**RY**	6.5
RY	4.0	QH	4.3	**GR**	15	**MK**	4.0	**RH**	8.0	WH	7.8	**NY**	8.0	**ML**	4.6
GR	3.2	VY	4.0	**ML**	5.9	VK	3.0	SH	6.2	**RF**	7.0	QF	7.5	**MK**	3.1
RF	2.9	**MR**	3.8	NR	5.1	**RY**	2.5	NH	3.1	KY	6.0	**RY**	6.8	KY	2.9
ML	2.9	GY	3.6	MM	1.4	KS	2.4	EF	2.8	**NY**	5.2	HY	4.1	**NY**	2.8
NS	2.5	ER	3.4	TR	1.2	**NY**	2.3	**NS**	2.8	KF	4.2	EH	2.8	KF	2.5
RH	2.5	**ML**	2.5	SR	1.2	KY	1.7	**NY**	2.6	KH	3.4	KY	2.3	KH	2.1
QN	2.1	TY	2.2	NL	0.93	SH	1.7	QF	2.5	NW	2.6	QH	2.3	RH	2.0
NY	2.1	IY	1.7	DA	0.78	QF	1.7	HS	2.1	QY	2.1	DH	1.8	NE	1.7
Σ[c]	36		39		80		37		54		67		66		37

[a] Amino acids at positions D + 2 and T + 2 in the indicated subfamily. Bold font indicates D + 2/T + 2 pairs found in the top 10 of all receiver domains in the nonredundant Database S3 at http://www.unc.edu/~bourret/PageSupplementalDatabaseFiles.htm.

[b] Percentage of subfamily members containing the indicated D + 2/T + 2 pair.

[c] Percentage of subfamily members represented by the top 10 D + 2/T + 2 pairs for that subfamily.

DISCUSSION

Acids and amides at positions D + 2 and/or T + 2 enhance autodephosphorylation. The weak correlation between net charge of the amino acids at D + 2/T + 2 and the autodephosphorylation rate constant of CheY variants (Fig. 3) resolved upon closer inspection into a consistent correlation between negative charge and enhanced k_{dephos} in three different receiver domains (Fig. 4) but no correlation between positive charge and diminished k_{dephos}. This discrepancy suggests that charge is not the fundamental basis for the underlying mechanism by which amino acids at D + 2/T + 2 enhance autodephosphorylation. It may be relevant that negatively charged amino acids are hydrogen bond acceptors, whereas positively charged amino acids are hydrogen bond donors. This suggests that negatively charged side chains at D + 2/T + 2 might enhance the reaction by interacting with the attacking water molecule. Because amide residues can act as hydrogen bond acceptors, this hypothesis predicts that amides should also enhance autodephosphorylation. Our mutant collection contains few examples with an amide residue at T + 2, but we have measured k_{dephos} for many CheY variants with an amide at D + 2 (Table 1) and can analyze the data as was done in Fig. 4 for acid residues. In 33 pairwise comparisons between CheY NX and ZX variants (matched at T + 2 and differing at D + 2), there were only three cases in which k_{dephos} for the CheY ZX mutant was faster and the "Z" (D + 2) residue was not an acid or amide (data not shown). Similarly, there were only three cases among 25 matched pairs in which k_{dephos} was greater for CheY QX than for CheY ZX without an acid or amide residue occupying D + 2 (data not shown). For PhoB$_N$ (Table 2), in two of three NX-versus-ZX comparisons, k_{dephos} was larger for the amide-containing variant (data not shown). For Spo0F (11), in four of five comparisons, the NX variant exhibited faster autodephosphorylation than the ZX variant, and in the one exception an acid residue at D + 2 resulted in a greater k_{dephos} than with an amide (data not shown). Although the difference in k_{dephos} is small in many pairwise comparisons, a preponderance of evidence indicates that amide residues at position D + 2 enhance autodephosphorylation in multiple receiver domains, although not to the same extent as acid residues.

The conclusions that (i) acid and amide residues at D + 2/T + 2 enhance receiver domain autodephosphorylation and (ii) the mechanism likely involves interaction with the attacking water molecule are supported by a previous study of CheY mimics of haloacid dehalogenase phosphatases (27). The active sites of HAD phosphatases are similar to the active sites of receiver domains and catalyze similar chemistries (28). Although position D + 2 is variable in receiver domains and rarely (2%) occupied by an Asp, HAD phosphatases contain a conserved Asp at D + 2 that accelerates phosphorylation and dephosphorylation by acid/base catalysis. The amino acid at T + 2 in HAD phosphatases often helps position the Asp at D + 2. We previously determined k_{dephos} values in five CheY variants containing D + 2/T + 2 pairs (DR, DK, DQ, DY, and DT) that are rare in receiver domains but common in HAD phosphatases (27). Although the CheY DX variants do not utilize acid/base catalysis, they support faster autodephosphorylation in all (12/12) possible comparisons with corresponding CheY ZX variants (matched at T + 2 and differing at D + 2) reported in Table 1. Furthermore, structural evidence was obtained for interactions between an Asp at D + 2 and an attacking water molecule (27).

Mechanistic insights from relationships between k_{dephos} and k_{phos}/K_s for CheY D + 2/T + 2 variants. Receiver domains self-catalyze both phosphorylation with small-molecule phosphodonors and dephosphorylation with water. Both reactions are substitution reactions at the phosphoryl atom. The proposed transition states for autodephosphorylation and autophosphorylation share multiple features, including a partially formed Asp-P bond and a planar PO_3^{2-} group coordinated by conserved active-site residues and a divalent cation (10, 12). If residues at D + 2 and T + 2 affected aspects of the two reactions that are similar, it would be reasonable to expect a direct correlation between the corresponding rate constants. However, direct correlation of k_{dephos} for CheY D + 2/T + 2 variants was not observed with the rate constants for autophosphorylation with PAM (Fig. 5A), MPI (Fig. 5B), or AcP (data not shown). The lack of a direct correlation suggests that residues at D + 2 and T + 2 affected features of autophosphorylation and autodephosphorylation that are not similar. Obvious differences between the reactions include the ground states (CheY-P for autodephosphorylation versus CheY for autophosphorylation) and the regions of the transition states near the attacking water for autodephosphorylation or the phosphodonor for autophosphorylation. Notably, the residues at D + 2/T + 2 are within appropriate distance to affect the regions around the water or phosphodonor. A previous study concluded on different grounds that the amino acids at D + 2/T + 2 affect CheY autophosphorylation kinetics by interacting with the leaving group (12).

Instead of a direct correlation, k_{dephos} of CheY variants differing at D + 2/T + 2 varied inversely with k_{phos}/K_S for PAM or MPI (Fig. 5) but not AcP. This distinction is consistent with a previous report that the kinetic determinants for autophosphorylation of CheY with phosphoramidates (PAM and MPI) or acyl phosphates (AcP) are different (12). In particular, autophosphorylation with PAM is more strongly influenced by the hydrophobic surface area of the amino acids at D + 2 and T + 2, whereas autophosphorylation with AcP is more strongly affected by the charge of the D + 2/T + 2 residues. The inverse relationship between k_{dephos} and k_{phos}/K_S for PAM/MPI displayed in Fig. 5 may arise because the two reactions are primarily influenced by mutually exclusive properties of the amino acids at D + 2/T + 2. Specifically, hydrogen bond accepting ability appears to promote autodephosphorylation (presumably by interacting with water as discussed above), whereas a hydrophobic surface area appears to stimulate autophosphorylation through direct interaction with the imidazole ring or related portions of the phosphodonor (12, 13). Thus, D + 2/T + 2 variants of CheY that are adept at catalyzing both autophosphorylation with PAM/MPI and autodephosphorylation with water were not observed.

Autodephosphorylation rate constants of CheY variants and wild-type response regulators with the same D + 2/T + 2 pair are correlated. In many instances, autodephosphorylation rate constants for CheY, $PhoB_N$, and Spo0F variants appeared to be consistent with the effects of substitutions at D + 2 and T + 2 anticipated from rate constants for wild-type response regulators containing the same D + 2/T + 2 pairs. For example, wild-type PrrA is at the slower end of the range of autodephosphorylation rate constants for wild-type response regulators and contains RY at D + 2/T + 2. In CheY, the RY variant is one of the slowest of the CheY variants (Table 1). There was a direct correlation (R^2 = 0.62) between k_{dephos} for wild-type response regulators (range, 4 orders of magnitude) and k_{dephos} for CheY variants with the same D + 2/T + 2 pair (range, 2 orders of magnitude) when plotted in log-log form (Fig. 6). The correlation between k_{dephos} for wild-type response regulators and corresponding CheY variants generally holds even though the effects of changing amino acids at D + 2 and T + 2 in CheY are sufficient to account for <1% of the range in autodephosphorylation rate constants observed among wild-type response regulators. The correlation may indicate that other factors affecting k_{dephos} often evolved to reinforce the direction set by the amino acids at D + 2 and T + 2.

D + 2/T + 2 amino acid pairs enriched in response regulator subfamilies correlate with functionally relevant autodephosphorylation kinetics. The enrichment of particular pairs of amino acids at positions D + 2 and T + 2 in various response regulator subfamilies (Table 3) is substantial. For example, whereas the MK and MR D + 2/T + 2 pairs are found in 14% of receiver domain sequences overall, they are present in 49% of receiver domain sequences in the OmpR subfamily (Table 3). Furthermore, 72% of all sequences containing MK or MR pairs belong to the OmpR subfamily. OmpR subfamily receiver domains typically function to regulate transcription, a relatively slow process, and receiver domains containing MK or MR at D + 2/T + 2 have autodephosphorylation rate constants at the low end of the range of wild-type response regulators. Changing D + 2/T + 2 in CheY to MK or MR made CheY autodephosphorylation rate constants at least an order of magnitude lower than for the wild type (Table 1). Similarly, replacing the natural MR pair in $PhoB_N$ with the NE pair found in CheY increased autodephosphorylation by more than an order of magnitude.

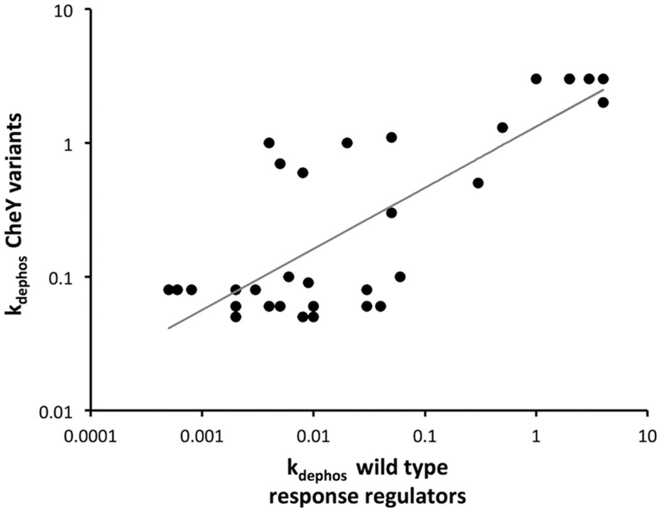


FIG 6 Correlation between k_{dephos} for wild-type response regulators (sources cited in the legend to Fig. S1 in the supplemental material) and CheY variants (Table 1) with same residues at positions D + 2 and T + 2. Best fit R^2 = 0.62. The point (0.00003, 0.4) corresponding to the relationship between RedF and CheY SH appears to be an outlier (not shown) and therefore was not included in the analysis.

A central result reported here is that negatively charged amino acids at D + 2 or T + 2 enhance autodephosphorylation (Fig. 4). There are no D + 2/T + 2 pairs with a negatively charged residue in the top 10 pairs of all receiver domains, and there are only three in the top 20 (see Fig. S11 in the supplemental material). However, the distribution among response regulator subfamilies of the three most common D + 2/T + 2 pairs containing a negative charge is informative (Table 3). Prevalent in chemotaxis, CheYs consist of a single receiver domain and typically contain an NE pair. While NE represents less than 2% of receiver domain sequences overall, the pair makes up 8% of single receiver domain sequences. A total of 78% of sequences overall that contained an NE pair are found in the single receiver domain database. While ER is found in 1.1% of receiver domain sequence, this D + 2/T + 2 pair represents 3.4% of receiver domains in hybrid kinases. A total of 75% of receiver domains that contain an ER pair are found in the hybrid kinase database. CheB-type response regulators contain a methylesterase domain (23) and are not part of the other response regulator subfamilies chosen for analysis in Table 3. CheB receiver domains typically contain the EL D + 2/T + 2 pair. The "nonassigned" database contained 78% of the EL-representative receiver domain sequences overall. It is striking that the NE, EL, and ER combinations of amino acids at D + 2 and T + 2 are strongly enriched in response regulators that require rapid autodephosphorylation for function. Signal transduction in chemotaxis occurs in a fraction of a second (29), and the receiver domain in hybrid kinases often functions as a phosphate sink to drain phosphoryl groups out of the system through autodephosphorylation (30–33).

The enrichment of specific D + 2/T + 2 pairs with particular effects on autodephosphorylation kinetics in different response regulator subfamilies strengthens the correlation between which D + 2/T + 2 pairs are present in a receiver domain and the biological function of that particular receiver domain. This correlation is further reinforced by the observations that (i) D + 2/T + 2 pairs with chemically similar amino acids (e.g., a basic residue at D + 2 paired with a hydrophobic residue at T + 2) often cluster together in response regulator subfamilies (Table 3) and (ii) D + 2/T + 2 pairs with chemically similar amino acids have very similar (within 2-fold) effects on CheY autodephosphorylation kinetics (Table 1).

The frequency of D + 2/T + 2 pairs in receiver domains from a particular microorganism likely reflects the distribution of response regulator subfamilies in that organism rather than phylogeny. A central part of the research strategy described here is to identify the combinations of amino acids at variable positions D + 2 and T + 2 most commonly found in naturally occurring receiver domains and then use this information to prioritize subjects of experimental investigation. However, phylogenetic groups are not uniformly represented in sequence databases, which as a result are inevitably biased by what has been sequenced. It would be useful to assess possible sources of database bias and corresponding means to mitigate the impact of bias. When constructing our receiver domain databases, we attempted to minimize the effects of overrepresentation bias by including information from only one genome per genus and excluding sequences that were more than 90% identical to other sequences. This addressed problems resulting from closely related organisms that have been sequenced multiple times but not biases due to the absence of sequences from other organisms.

The observation that a large majority of occurrences of a given D + 2/T + 2 pair were often found in a single response regulator subfamily (Table 3) provides a way to assess the potential impact of underrepresentation bias. Individual species typically encode response regulators from many different subfamilies, and there is variation in the abundance of different response regulator subfamilies across phylogenetic groups (23). Thus, the abundance of particular pairs of amino acids at D + 2/T + 2 might more directly reflect the distribution of response regulator subfamilies across species than the phylogeny of sequenced genomes. To test this idea, we compared the abundance of D + 2/T + 2 pairs in sample species from various phylogenetic groups (see Table S12 in the supplemental material) with the abundance of response regulator subfamilies in the same organisms (see Table S13 in the supplemental material). For example, more than three-quarters of response regulators in the *Actinobacteria* sample belonged to the OmpR or NarL/FixJ subfamilies (see Table S13), and six of the seven most abundant D + 2/T + 2 pairs in *Acinetobacteria* (see Table S12) were highly enriched in the OmpR and NarL/FixJ subfamilies (Table 3). Similarly, >70% of receiver domains in the *Cyanobacteria* sample belonged to the hybrid kinase or single-domain subfamily, and the seven most common D + 2/T + 2 pairs in the *Cyanobacteria* sample were characteristic of these two subfamilies. Additional examples are evident. Overall, the data in Tables S12 and S13 are consistent with the abundance of response regulator subfamilies being a primary determinant of the most frequent D + 2/T + 2 pairs found in a given species.

If the distribution of response regulator families is very different in organisms that have not been sequenced compared to those that have, then the most frequent D + 2/T + 2 pairs in our database (see Table S11 in the supplemental material) may not be the most abundant in nature. Nevertheless, the 20 overall most common D + 2/T + 2 pairs, which were the focus of our experimental investigation, account for two-thirds of the 70 pairs (10 most common D + 2/T + 2 pairs in each of seven receiver domain subfamilies) listed in Table 3. Representation of naturally occurring receiver domains could be further enhanced by experimental characterization of the effects of the D + 2/T + 2 pairs that are most abundant in each response regulator subfamily, rather than across all receiver domains.

Experimental design guided by protein sequences has broad implications for strategies to assess variation within protein families. With rapid advancements in genome sequencing, the known sizes of protein families have dramatically increased. While conserved domains identify the family to which a protein belongs, questions of functionality remain. Assessment of functionality within a protein family has historically focused on conserved residues within the conserved domain. Typically, protein sequences are aligned to determine the conserved residues, the conserved residues are changed to alanines or other amino acids, and the mutant proteins are assessed using functional assays. However, when considering the different functionalities within a family of proteins, it is often the variable features that distinguish one protein in a family from another. Candidates for functionally important variable positions can be identified by mutual information analysis, inspection of protein structures, or alanine-scanning mutagenesis. Protein sequence analysis dramatically enhanced the feasibility of studying variable features of a large protein family, i.e., response regulators. Not only were we able to assess receiver domain autodephosphorylation kinetics of a mutant set that reflects what is found in nature, but also we were able to gain more confidence in the mechanistic insights elucidated by an expanded data set. While previous work indicated steric occlusion as a key means of influence by D + 2/T + 2 pairs on autodephosphorylation, the larger data set reported here suggests with more confidence that interaction with the attacking water is a key means of influence.

Protein sequence analysis also revealed distinctions between response regulator subfamilies. The D + 2/T + 2 pairs enriched in the response regulator subfamily databases (Table 3) modulated autodephosphorylation in ways that are consistent with the biological functions associated with response regulators in those subfamilies. Grouping protein sequences into subcategories (such as the response regulator subfamilies used here) combined with the knowledge of influence from variable features (such as results from changing D + 2/T + 2 reported here) could provide further mechanistic insight into functional variation with protein families.

ACKNOWLEDGMENTS

We acknowledge Igor Zhulin and Ogun Adebali for providing receiver domain sequences and for advice on protein sequence analysis. We acknowledge Ruth Silversmith, Rachel Creager-Allen, and Chrystal Starbird for helpful discussions regarding this work. Suggestions from anonymous peer reviewers strengthened the manuscript.

FUNDING INFORMATION

This work, including the efforts of Bob Bourret, was funded by HHS | NIH | National Institute of General Medical Sciences (NIGMS) (GM050860). This work, including the efforts of Stephani Page, was funded by HHS | NIH | National Institute of General Medical Sciences (NIGMS) (GM008570).

REFERENCES

1. **Stock AM, Robinson VL, Goudreau PN.** 2000. Two-component signal transduction. Annu Rev Biochem **69:**183–215. http://dx.doi.org/10.1146/annurev.biochem.69.1.183.
2. **Wuichet K, Cantwell BJ, Zhulin IB.** 2010. Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol **13:**219–225. http://dx.doi.org/10.1016/j.mib.2009.12.011.
3. **Capra EJ, Laub MT.** 2012. Evolution of two-component signal transduction systems. Annu Rev Microbiol **66:**325–347. http://dx.doi.org/10.1146/annurev-micro-092611-150039.
4. **West AH, Stock AM.** 2001. Histidine kinases and response regulator proteins in two-component signaling systems. Trends Biochem Sci **26:**369–376. http://dx.doi.org/10.1016/S0968-0004(01)01852-7.
5. **Bourret RB, Thomas SA, Page SC, Creager-Allen RL, Moore AM, Silversmith RE.** 2010. Measurement of response regulator autodephosphorylation rates spanning six orders of magnitude. Methods Enzymol **471:**89–114. http://dx.doi.org/10.1016/S0076-6879(10)71006-5.
6. **Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE.** 2000. The Protein Data Bank. Nucleic Acids Res **28:**235–242. http://dx.doi.org/10.1093/nar/28.1.235.
7. **Rose PW, Prlic A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK.** 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res **43:**D345–D356. http://dx.doi.org/10.1093/nar/gku1214.
8. **Bourret RB.** 2010. Receiver domain structure and function in response regulator proteins. Curr Opin Microbiol **13:**142–149. http://dx.doi.org/10.1016/j.mib.2010.01.015.
9. **Volz K.** 1993. Structural conservation in the CheY superfamily. Biochemistry **32:**11741–11753. http://dx.doi.org/10.1021/bi00095a001.
10. **Pazy Y, Wollish AC, Thomas SA, Miller PJ, Collins EJ, Bourret RB, Silversmith RE.** 2009. Matching biochemical reaction kinetics to the timescales of life: structural determinants that influence the autodephosphorylation rate of response regulator proteins. J Mol Biol **392:**1205–1220. http://dx.doi.org/10.1016/j.jmb.2009.07.064.
11. **Thomas SA, Brewster JA, Bourret RB.** 2008. Two variable active site residues modulate response regulator phosphoryl group stability. Mol Microbiol **69:**453–465. http://dx.doi.org/10.1111/j.1365-2958.2008.06296.x.
12. **Thomas SA, Immormino RM, Bourret RB, Silversmith RE.** 2013. Nonconserved active site residues modulate CheY autophosphorylation kinetics and phosphodonor preference. Biochemistry **52:**2262–2273. http://dx.doi.org/10.1021/bi301654m.
13. **Page SC, Silversmith RE, Collins EJ, Bourret RB.** 2015. Imidazole as a small molecule analogue in two-component signal transduction. Biochemistry **54:**7248–7260. http://dx.doi.org/10.1021/acs.biochem.5b01082.
14. **Creager-Allen RL, Silversmith RE, Bourret RB.** 2013. A link between dimerization and autophosphorylation of the response regulator PhoB. J Biol Chem **288:**21755–21769. http://dx.doi.org/10.1074/jbc.M113.471763.
15. **Silversmith RE, Appleby JL, Bourret RB.** 1997. Catalytic mechanism of phosphorylation and dephosphorylation of CheY: kinetic characterization of imidazole phosphates as phosphodonors and the role of acid catalysis. Biochemistry **36:**14965–14974. http://dx.doi.org/10.1021/bi9715573.
16. **Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M.** 2014. Pfam: the protein families database. Nucleic Acids Res **42:**D222–D230. http://dx.doi.org/10.1093/nar/gkt1223.
17. **Simonetti FL, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C.** 2013. MISTIC: mutual information server to infer coevolution. Nucleic Acids Res **41:**W8–W14. http://dx.doi.org/10.1093/nar/gkt427.
18. **Ulrich LE, Zhulin IB.** 2010. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. Nucleic Acids Res **38:**D401–D407. http://dx.doi.org/10.1093/nar/gkp940.
19. **Ulrich LE, Zhulin IB.** 2014. SeqDepot: streamlined database of biological sequences and precomputed features. Bioinformatics **30:**295–297. http://dx.doi.org/10.1093/bioinformatics/btt658.
20. **Edgar RC.** 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.
21. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
22. **Maule AF, Wright DP, Weiner JJ, Han L, Peterson FC, Volkman BF, Silvaggi NR, Ulijasz AT.** 2015. The aspartate-less receiver (ALR) domains: distribution, structure and function. PLoS Pathog **11:**e1004795. http://dx.doi.org/10.1371/journal.ppat.1004795.
23. **Galperin MY.** 2010. Diversity of structure and function of response regulator output domains. Curr Opin Microbiol **13:**150–159. http://dx.doi.org/10.1016/j.mib.2010.01.005.
24. **Pao GM, Saier MH, Jr.** 1995. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J Mol Evol **40:**136–154. http://dx.doi.org/10.1007/BF00167109.
25. **Grebe TW, Stock JB.** 1999. The histidine protein kinase superfamily. Adv Microb Physiol **41:**139–227. http://dx.doi.org/10.1016/S0065-2911(08)60167-8.
26. **Mack TR, Gao R, Stock AM.** 2009. Probing the roles of the two different dimers mediated by the receiver domain of the response regulator PhoB. J Mol Biol **389:**349–364. http://dx.doi.org/10.1016/j.jmb.2009.04.014.
27. **Immormino RM, Starbird CA, Silversmith RE, Bourret RB.** 2015. Probing mechanistic similarities between response regulator signaling proteins and haloacid dehalogenase phosphatases. Biochemistry **54:**3514–3527. http://dx.doi.org/10.1021/acs.biochem.5b00286.
28. **Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L.** 2006. Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol **361:**1003–1034. http://dx.doi.org/10.1016/j.jmb.2006.06.049.
29. **Segall JE, Manson MD, Berg HC.** 1982. Signal processing times in bacterial chemotaxis. Nature **296:**855–857. http://dx.doi.org/10.1038/296855a0.
30. **Sourjik V, Schmitt R.** 1998. Phosphotransfer between CheA, CheY1, and CheY2 in the chemotaxis signal transduction chain of *Rhizobium meliloti*. Biochemistry **37:**2327–2335. http://dx.doi.org/10.1021/bi972330a.
31. **Freeman JA, Bassler BL.** 1999. Sequence and function of LuxU: a two-component phosphorelay protein that regulates quorum sensing in *Vibrio harveyi*. J Bacteriol **181:**899–906.
32. **Janiak-Spens F, Sparling DP, West AH.** 2000. Novel role for an HPt domain in stabilizing the phosphorylated state of a response regulator domain. J Bacteriol **182:**6673–6678. http://dx.doi.org/10.1128/JB.182.23.6673-6678.2000.
33. **Perraud AL, Kimmel B, Weiss V, Gross R.** 1998. Specificity of the BvgAS and EvgAS phosphorelay is mediated by the C-terminal HPt domains of the sensor proteins. Mol Microbiol **27:**875–887. http://dx.doi.org/10.1046/j.1365-2958.1998.00716.x.