Title: Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes

Authors: R. R. Cheng[1]*, O. Nordesjö[2], R. L. Hayes[3], H. Levine[1,4], S. C. Flores[2], J. N. Onuchic[1,5]*, F. Morcos[6]*

Affiliations:

[1]Center for Theoretical Biological Physics, Rice University, Houston, USA.

[2]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.

[3]Department of Biophysics, University of Michigan, Ann Arbor, USA.

[4]Department of Bioengineering, Rice University, Houston, USA.

[5]Departments of Physics & Astronomy, Chemistry, and Biosciences, Rice University, Houston, USA.

[6]Department of Biological Sciences, University of Texas at Dallas, Dallas, USA.

*To whom correspondence should be addressed: R. R. Cheng (ryan.r.cheng@gmail.com), J. N. Onuchic (jonuchic@rice.edu), F. Morcos (faruckm@utdallas.edu)

Abstract:

Two-component signaling (TCS) is the primary means by which bacteria sense and respond to the environment. TCS involves two partner proteins working in tandem, which interact to perform cellular functions while limiting interactions with non-partners (i.e., “cross-talk”). We construct a Potts model for TCS that can quantitatively predict how mutating amino acid identities affect the interaction between TCS partners and non-partners. The parameters of this model are inferred directly from protein sequence data. This approach drastically reduces the computational complexity of exploring the sequence-space of TCS proteins. As a stringent test, we compare its predictions to a recent comprehensive mutational study, which characterized the functionality of 20^4 mutational variants of the PhoQ kinase in *Escherichia coli*. We find that our best predictions accurately reproduce the amino acid combinations found in experiment, which enable functional signaling with its partner PhoP. These predictions demonstrate the evolutionary pressure to preserve the interaction between TCS partners as well as prevent unwanted “cross-talk”. Further, we calculate the mutational change in the binding affinity between PhoQ and PhoP, providing an estimate to the amount of destabilization needed to disrupt TCS.

Introduction

Early theoretical work on protein folding postulated that proteins have evolved to be minimally frustrated (Bryngelson and Wolynes 1987; Bryngelson, et al. 1995; Onuchic, et al. 1997), i.e., evolved to have favorable residue-residue interactions that facilitate folding into the native state while having minimal non-native energetic traps. The principle of minimal frustration provides intuition as to why protein sequences are not random strings of amino acids. The evolutionary constraint to fold into a particular, stable three-dimensional structure while minimizing the number of frustrated interactions greatly restricts the sequence-space of a protein (Leopold, et al. 1992; Bryngelson, et al. 1995; Onuchic, et al. 1997). Satisfaction of these constraints result in correlated amino acid identities within the sequences of a protein family. These correlated identities occur between different positions in a protein such as, for example, native contacts (Gobel, et al. 1994; Neher 1994; Shindyalov, et al. 1994). We refer to these quantifiable amino acid correlations as coevolution.

Of course, coevolution does not only arise from the constraint to fold. Proteins also fulfill cellular functions, which act as additional constraints on the sequences of proteins (Ferreiro, et al. 2014; Sikosek and Chan 2014; Wolynes 2015). In the context of signal transduction, proteins have evolved to be able to preferentially bind to a signaling partner(s) as well as catalyze the chemical reactions associated with signal transfer. An important example is two-component signaling (TCS) (Hoch 2000; Stock, et al. 2000; Laub and Goulian 2007; Casino, et al. 2010; Szurmant and Hoch 2010; Capra and Laub 2012), which serves as the primary means for bacteria to sense the environment and carry out appropriate responses. TCS consists of two partner proteins working in tandem: a histidine kinase (HK) and a response regulator (RR). Upon the detection of stimulus by an extracellular sensory domain, the HK generates a signal via autophosphorylation. Its RR partner can then transiently bind to the HK and receive the signal (i.e., phosphoryl group), thereby activating its function as a transcription factor. The HK has also evolved to catalyze the reverse signal transfer reaction (i.e., phosphatase activity), acting as a sensitive switch to turn off signal transduction. To prevent signal transfer with the wrong partner (i.e., “cross-talk”), TCS partners have mutually evolved amino acids at their respective binding interfaces that confer interaction specificity (Laub and Goulian 2007; Szurmant and Hoch 2010; Capra and Laub 2012). Thus, the collection of protein sequences of TCS partners contains quantifiable coevolution between the HK and RR sequences.

In principle, the determinants of interaction specificity for TCS can be quantified by a probabilistic model of sequence selection (Cheng, et al. 2014). Assuming that nature has sufficiently sampled the sequence-space of TCS proteins, the collection of protein sequences of TCS partners can be viewed as being selected under quasi-equilibrium from a Boltzmann distribution:

$$P(S_{\mathrm{TCS}}) = Z^{-1}\exp\left(-H(S_{\mathrm{TCS}})\right) \qquad (1)$$

where S_{TCS} is the concatenated amino acid sequence of a HK and RR protein, P is the probability of selecting S_{TCS}, and $-H$ is proportional to the additive fitness landscape that governs the evolutionary sequence selection for TCS partners. Specifically, Eq. 1 was previously derived using simple models of evolutionary biology (Sella and Hirsh 2005), where $H = -vx$ and v is the population size of the genotypes and x is the additive fitness (i.e., the negative log of the fitness); See Materials and Methods for more details. H is referred to in our work as a coevolutionary landscape.

Recently, maximum entropy-based approaches referred to as Direct Coupling Analysis (DCA) (Weigt, et al. 2009; Morcos, et al. 2011; Ekeberg, et al. 2013) have been successfully applied to infer the parameters of H (a Potts model) that governs the empirical amino acid sequence statistics. This has allowed for the direct quantification of the coevolution in protein sequence data (See Review: (de Juan, et al. 2013)). Early work using DCA to study TCS primarily focused on identifying the key coevolving residues between the HK and RR (Weigt, et al. 2009). Highly coevolving residue pairs have been used as docking constraints in a molecular dynamics simulation to predict the HK/RR signaling complex (Schug, et al. 2009), the autophosphorylation structure of a HK (Dago, et al. 2012) , and the homodimeric form (transcription factor) of the RR (dos Santos, et al. 2015). DCA has also been applied to quantify the determinants of interaction specificity between TCS proteins (Procaccini, et al. 2011; Cheng, et al. 2014), building on earlier coevolutionary approaches (Li, et al. 2003; Burger and van

Nimwegen 2008). In particular, DCA was used to predict the effect of point mutations on TCS phosphotransfer *in vitro* as well as demonstrate the reduced specificity between HK and RR domains in hybrid TCS proteins (Cheng, et al. 2014).

The experimental effort to determine the molecular origin of interaction specificity in TCS proteins (See Reviews: (Laub and Goulian 2007; Casino, et al. 2010; Szurmant and Hoch 2010; Podgornaia and Laub 2013)) precedes the recent computational efforts. Full knowledge of the binding interface between HK and RR was made possible through X-ray crystallography (Casino, et al. 2009). Scanning mutagenesis studies (Tzeng and Hoch 1997; Qin, et al. 2003; Capra, et al. 2010) provided insight on the subset of important interfacial residues that determine specificity. These key residues were mutated to enable a TCS protein to preferentially interact with a non-partner *in vitro* (Skerker, et al. 2008; Capra, et al. 2010). However, the extent of possible amino acid identities that allow TCS partners to preferentially interact *in vivo* has remained elusive until recent comprehensive work by Podgornaia and Laub (Podgornaia and Laub 2015). Their work focused on the PhoQ/PhoP TCS partners in *E. coli*, which control the response to low magnesium stress. PhoQ (HK) phosphorylates and dephosphorylates PhoP (RR) under low and high magnesium concentrations, respectively. Using exhaustive mutagenesis of 4 residues of PhoQ (20^4 = 160,000 mutational variants) at positions that form the binding interface with PhoP, Podgornaia and Laub (Podgornaia and Laub 2015) were able to characterize all mutants based on their functionality in *E. coli*. It was found that roughly 1% of all PhoQ mutants were functional, enabling *E. coli* to exhibit comparable responses to magnesium concentrations as the wild type PhoQ. This finding uncovered a broad degeneracy in the sequence-space of the HK protein that still maintained signal transfer efficiency as well as interaction specificity with its partner.

We ask whether amino acid coevolution inferred using DCA could capture the functional mutational variants observed in the comprehensive mutational study of PhoQ and if so, to what extent? Capturing this functionality requires that information gleaned from coevolution is sufficient to estimate the effect of mutations to PhoQ on its interactions with PhoP as well as on unwanted "cross-talk". Hence, our question is important to determine if coevolutionary methods can be extended from studying two interacting proteins to studying an interaction network (e.g., systems biology). Further, this question is of particular interest to those who want to engineer novel mutations in TCS proteins that can maintain or encode the interaction specificity of a TCS protein to its partner or a non-partner, respectively.

To answer this question, we first infer a Potts model, H (see Eq. 1), which forms the basis for quantifying how mutations affect the interaction between a HK and RR protein. Focusing on the parameters of H that are related to interprotein coevolution, we construct a coevolutionary landscape to quantify TCS interactions, H_{TCS}, for a given sequence of an HK and RR protein. H_{TCS} serves as a proxy for signal transfer efficiency, allowing us to quantify the effect on fitness of the interaction between any HK and RR protein. Further, we can assess how mutations affect fitness due to changes in the HK/RR interaction by computing the mutational change in H_{TCS} between the mutant sequence, $S_{\mathrm{TCS}}^{\mathrm{mutant}}$, and the wild type sequence, $S_{\mathrm{TCS}}^{\mathrm{WT}}$:

$$\Delta H_{\mathrm{TCS}} = H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) - H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{WT}}) . \qquad (2)$$

Considering the concatenated sequence of PhoQ and PhoP, we compute Eq. 2 for the 20^4 PhoQ mutational variants. We find that mutants with the most favorable ΔH_{TCS} (e.g., most negative) were classified as functional HKs by Podgornaia and Laub (Podgornaia and Laub 2015)—i.e., true positive predictions. Next, we focus on mutations predicted to be favorable by Eq. 2 that were classified as non-functional in experiment. Expanding our analysis of the PhoQ mutants beyond its interaction with PhoP, we consider how mutations affect the signal transfer efficiency, H_{TCS}, between PhoQ and all of the RR proteins in *E. coli*. We find that many of these non-functional mutants exhibit "cross-talk" interactions according to our model, accounting for their non-functionality. If we exclude these promiscuous variants, we can better isolate the true positive predictions that are functional from false positives that are non-functional. Our predictions also capture context-dependent mutational affects that were observed in experiment, i.e., epistasis. Finally, we estimate the mutational change in binding affinity in the PhoQ/PhoP bound complex using the Zone Equilibration of Mutants (ZEMU) method (Dourado and Flores 2014), a combined physics- and knowledge-based approach for free energy calculations. Consistent with what we would expect, we find that mutations that destabilize the HK/RR interaction tend to be non-functional with very high statistical significance. Non-functional mutants are on average destabilized by ~2 kcal/mol with respect to functional mutants.

The work described herein demonstrates that a coevolutionary model (i.e., additive fitness landscape) built from sequence data can directly connect molecular details at the residue-level to mutational phenotypes in bacteria. This has broad applications in systems biology, but also in synthetic biology since our computational framework can be used to select mutations that enhance or suppress interactions between TCS proteins. A more detailed description of our computational approaches can be found in the Materials and Methods section.

Results

Mutational change in coevolutionary landscape, ΔH_{TCS}*, for PhoQ/PhoP interaction*

We first focus on the parameters of the inferred Potts model (Eq. 3) that describe the coevolution between the Dimerization and Histidine phosphotransfer domain (DHp) and the Receiver (REC) domain (Fig. 1A), which form the HK/RR binding interface (Fig. 1B). The interprotein statistical couplings (Fig. 1C) of Eq. 3 are used to construct a coevolutionary landscape, H_{TCS} (Eq. 5), as a proxy for signal transfer efficiency. For each of the 1,659 functional and 158,341 non-functional PhoQ-mutational variants identified by Podgornaia and Laub (Podgornaia and Laub 2015), we compute the mutational change, ΔH_{TCS}, between PhoQ and PhoP. As an initial step, we only consider the PhoQ/PhoP sequence, i.e., we do not yet consider other RR proteins than PhoP. A histogram of ΔH_{TCS} is generated for all mutational variants (Fig. 2A). The distribution of the functional mutants tends more towards favorable ΔH_{TCS} than the distribution of non-functional mutants, but more interestingly, the most favorable predictions of our model contain mostly functional mutations. This is made clear by a plot of the Positive Predictive Value (PPV) for the top N mutational variants ranked by ΔH_{TCS} (Fig. 2B) from most favorable to most deleterious. The top 25 mutational variants ranked by ΔH_{TCS} contain 20 functional mutants and 5 non-functional mutants (i.e., PPV=0.8).

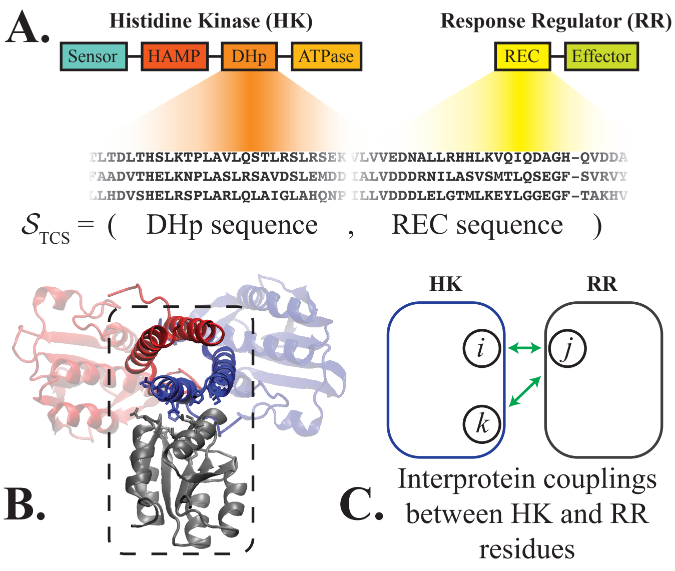
A.
Histidine Kinase (HK)
Response Regulator (RR)
Sensor
HAMP
DHp
ATPase
REC
Effector
TLTDLTHSLKTPLAVLQSTLRSLRSEK VLVVEDNALLRHHLKVQIQDAGH-QVDDA
FAADVTHELKNPLASLRSAVDSLEMDD IALVDDDRNILASVSMTLQSEGF-SVRVY
LLHDVSHELRSPLARLQLAIGLAHQNP ILLVDDDLELGTMLKEYLGGEGF-TAKHV
$\mathcal{S}_{\mathrm{TCS}}$ = (DHp sequence , REC sequence)
B.
HK
RR
i
j
k
C.
Interprotein couplings between HK and RR residues

Fig. 1. **TCS domain interactions of interest.** We focus only on HK proteins that have the following domain architecture from N to C terminus: sensor, HAMP, DHp, and ATPase. Likewise, we consider RR proteins that consist of a REC domain followed by an effector domain. (A) The interaction between the DHp and REC domains of the HK and RR proteins, respectively, form the TCS complex. Sequences of TCS partners are collected and stored as the concatenated sequence of the DHp and REC domains, S_{TCS} (See Materials and Methods). (B) A representative structure of the HK/RR TCS complex previously predicted for the KinA/Spo0F complex in *B. subtilis (Cheng, et al. 2014)*. The HK homodimer is shown in red and blue while the receiver domain of the RR is shown in gray. The dashed box highlights the DHp and REC interface. This predicted complex is consistent with the experimentally determined crystal structure of HK853/RR468 of *T. maritima* (Casino, et al. 2009) as well as another computationally predicted TCS complex (Schug, et al. 2009). (C) Our proxy for signal transfer efficiency, H_{TCS} (Eq. 5), is composed of the statistical coupling parameters that describe coevolution between interprotein residues (depicted in green). Hence, H_{TCS} naturally captures the context-dependence of mutating a residue in the HK when a residue in the RR is also mutated, or vice versa (See Materials and Methods for more details).

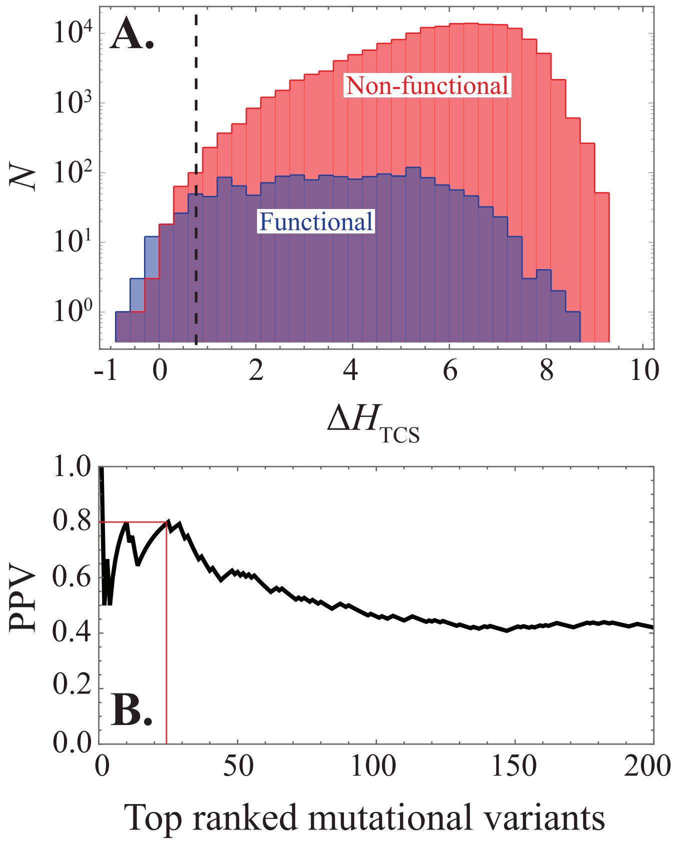

Fig. 2. **Effect of mutations on the PhoQ/PhoP interaction.** (A) Considering the concatenated sequence of PhoQ/PhoP, a histogram of ΔH_{TCS} (Eq. 5) is plotted for the functional (blue) and non-functional (red) mutational variants reported by Podgornaia and Laub (Podgornaia and Laub 2013). The color purple shows parts of the plot where the blue and red histograms overlap. The dashed line roughly partitions the 200 most favorable mutational variants given by ΔH_{TCS}, which contains more functional than non-functional mutants. By definition, $\Delta H_{\text{TCS}} = 0$ corresponds to the wild type PhoQ/PhoP and $\Delta H_{\text{TCS}} < 0$ corresponds to mutations that we predict to be more favorable to PhoQ/PhoP signaling than the wild type. (B) We plot the positive predictive value (PPV) as a function of the N mutational variants ranked by ΔH_{TCS} from the most to least favorable

for the first 200 mutants. PPV=TP/(TP+FP), where true positives (TP) and false positives (FP) refer to the fraction of mutants that are functional or non-functional, respectively, in the top N ranked variants. The thin red lines denote that the top 25 ranked mutational variants have a PPV of 0.8.

System-level analysis using ΔH_{TCS} *: functional mutants limit "cross-talk"*

Mutations that may enhance signal transfer efficiency between PhoQ and PhoP *in vitro* may still result in a non-functional PhoQ/PhoP system *in vivo*. This would occur if the mutations to PhoQ sufficiently encoded it to preferentially interact with another RR in *E. coli*. For this reason, we focused our computational analysis on the subset of mutational variants that preserve PhoQ/PhoP specificity by limiting "cross-talk" according to our coevolutionary model.

We first calculate the proxy for signal transfer efficiency, H_{TCS}, between the wild type PhoQ sequence and all of the non-hybrid RR proteins in *E. coli* (Fig. 3A). We find that for wild type PhoQ, the most favorable H_{TCS} (most negative) is with its known signaling partner, PhoP. As a consistency check, we also plot H_{TCS} for different combinations of the cognate partners TCS proteins in *E. coli* (Fig. S1). This result is consistent with previous computational predictions that used information-based quantities (Procaccini, et al. 2011; Cheng, et al. 2014) to quantify interaction specificity.

Extending upon Fig. 3A, we assess "cross-talk" in our model by calculating H_{TCS} between each PhoQ mutant and all of the non-hybrid RR in *E. coli*. We exclude all mutant-PhoQ variants that have a more favorable H_{TCS} with a non-partner RR. These excluded mutants are excellent candidates for engineering specificity in *E. coli*. Applying our exclusion criterion, we find that only 181 functional and 1,532 non-functional variants remain, i.e., 89% and 99% of the functional and non-functional variants, respectively, were removed. A histogram of the remaining (cross-talk excluded) mutants as a function of ΔH_{TCS} (Fig. 3B) shows that a filter based on interaction specificity is better able to isolate the true positive (functional) variants. Notably, the first 17 ranked variants are all functional variants. Once again, ranking the filtered variants by ΔH_{TCS} from the most favorable to the least favorable, we can plot the PPV (Fig. 3C) for the top *N* ranked variants. We find that the cross-talk excluded PPV tends to lie above the original PPV from Fig. 2B.

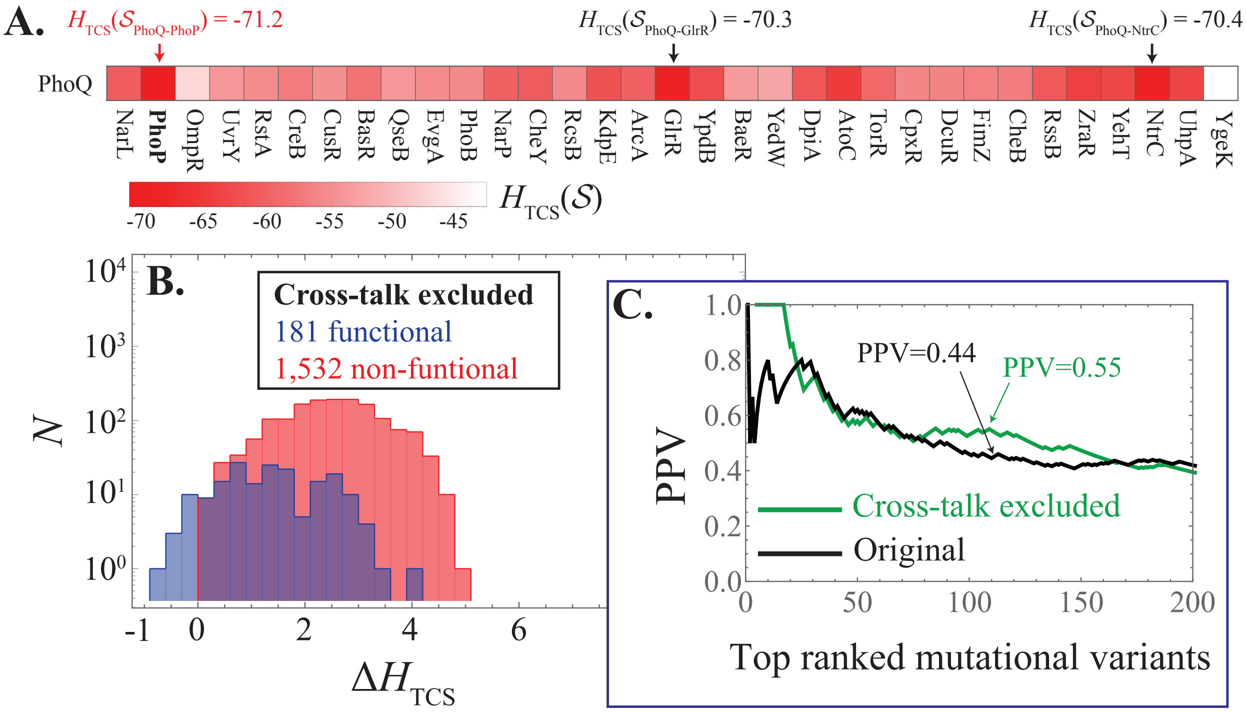

Fig. 3. Excluding mutational variants that are inferred to "cross-talk". (A) A grid plot showing H_{TCS} (Eq. 5) computed for the wild type PhoQ sequence with all of the non-hybrid RR protein sequences in *E. coli*, respectively. The most favorable interaction (most negative) given by H_{TCS} is between PhoQ and its partner PhoP. (B) We plot the

Cross-talk excluded subset (181 functional 1,532 non-functional) in a histogram as a function of the ΔH_{TCS} similar to Fig. 2A. (C) We plot the PPV as a function of the N top mutational variants ranked by ΔH_{TCS} for the first 200 mutants. The PPV for the cross-talk excluded mutational variants from Fig. 3B are plotted in green while the original PPV (Fig. 2B) is shown in black.

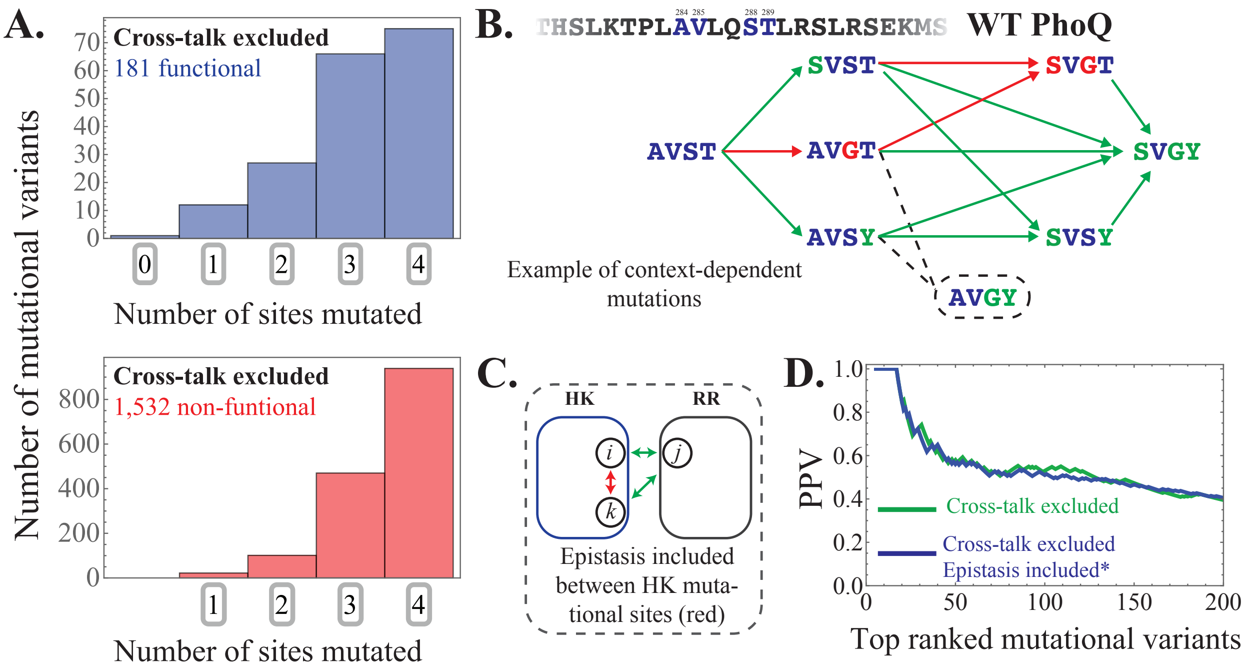

Fig. 4. Capturing epistatic mutational effects. (A) Histograms of the 0, 1, 2, 3, and 4-site mutants are plotted for the 181 functional and 1,532 non-functional mutational variants in our cross-talk excluded subset that is predicted using H_{TCS} (Eq. 5). (B) An example of mutational context-dependence predicted by H_{TCS} is shown here, considering residues 254, 255, 258, 259 of PhoQ, which have the WT amino acid configuration AVST, respectively. Green arrows drawn from AVST indicate single point mutations that

are correctly predicted to result in a functional phenotype. Successive green arrows indicate double and triple point mutations from the WT that are correctly predicted to be functional. Likewise, red arrows indicate single point mutations from an amino acid configuration that are correctly predicted to result in a non-functional phenotype. The AVGY mutation (dashed line and circle) was found to be functional in experiment but is predicted to cross-talk by H_{TCS}. (C) The schematic shows intraprotein coevolution (red) between residues *i* and *k* (of the HK) and interprotein coevolution (green) between residues *i* and *j* as well as *k* and *j*. As previous described in Fig. 1C, interprotein coevolution captures the effects of mutating the HK when the RR is also mutated (or vice versa). Epistasis between HK and RR proteins are naturally incorporated within H_{TCS}. On the other hand, epistasis between the 4-mutational sites of the Podgornaia and Laub experiment is described in our model through the statistical couplings between the 4-mutational sites (red in schematic). These additional parameters are added to H_{TCS} to obtain $H_{TCS}^{(epistasis)}$ (See Materials and Methods). (D) We plot the PPV as a function of the *N* top mutational variants ranked by ΔH_{TCS} for the first 200 mutants. The PPV predicted by H_{TCS} (Fig. 3C) is shown in green while the PPV predicted by $H_{TCS}^{(epistasis)}$ is shown in blue.

Mutational context-dependence in the coevolutionary landscape

A significant finding of the Podgornaia and Laub study was the context-dependent nature of the many of the mutations. For example, individual mutations that may result in a non-functional phenotype may result in a functional phenotype when combined. It is well known that mutations may exhibit such a context dependence, or epistasis, i.e., mutations introduced together have an effect on fitness that is not simply the combined effect of each mutation alone. Such effects would restrict the connectivity between functional mutations and act as constraints on TCS evolution (Podgornaia and Laub 2015).

We find that the functional predictions in Fig. 3 tend to be 3- and 4-point mutations, highlighting their non-trivial nature (Fig. 4A). While the effect of mutating a HK protein when its partner RR is mutated (and vice versa) is naturally captured in our model through interprotein statistical couplings (Fig. 1C), the mutational context-dependence (i.e., intraprotein couplings) of HK only mutations or RR only mutations is not explicitly contained within Eq. 5. Even with these limitations, the model is still able to distinguish between functional multi-point mutations that are composed of non-functional single-point mutations for the most favorable predictions, in accordance with experiment (Podgornaia and Laub 2015). One example is provided in Fig. 4B for the mutation of WT PhoQ from AVST at residues 284, 285, 288 and 289, respectively, to SVGY (i.e., a 3-point mutation). For single point mutations from AVST, H_{TCS} correctly predicts that SVST and AVSY are functional while AVGT is non-functional. For two point mutations, H_{TCS} correctly predicts that SVSY is functional while SVGT is non-functional. More interestingly, H_{TCS} finds that when the non-functional 2-point mutation SVGT is combined with a Tyrosine mutation to position 289 (i.e., SVGY), the functionality is recovered. It should be noted that the two-point mutation AVGY is predicted to have a favorable ΔH_{TCS} between PhoQ and PhoP, but is predicted to undergo cross-talk, and thus, is discarded despite being functional in experiment.

We next consider a model where epistasis between the 4-mutational sites explored by Podgornaia and Laub are explicitly introduced into Eq. 5 (See Eq. 6 in Materials and Methods). The epistatic effects of multiple HK mutations are captured by the statistical couplings of the

Potts model that describe the pairwise coevolution between the HK mutational sites. This is illustrated schematically in Fig. 4C. We find that such a model, $H_{TCS}^{(epistasis)}$ (Eq. 6), is consistent with the original model, H_{TCS}, in terms of its predictions and predictive quality (Fig. 4D). In particular, the prediction of functional phenotypes for the example in Fig. 4B yields identical results. Nevertheless, this model provides additional epistatic mutational effects that are not simply the added sum of the individual mutational effects.

Finally, we examine a model that only considers HK coevolution in SI Main Text. This model naturally includes the epistasis between the 4-mutational sites explored in experiment (Podgornaia and Laub 2015). From this model, we find that coevolution between HK residues alone is insufficient for capturing the functional phenotypes observed in experiment.

Mutational change in the binding affinity using a combined physics- and knowledge-based approach

We used ZEMu (See Materials and Methods) to compute the mutation-induced change in the binding affinity, $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$, between PhoQ and PhoP. The calculation converged for 42,985 mutants (702 functional and 42,283 non-functional) from a randomly selected subset of the 20^4 variants. We first examined a scatter plot of ΔH_{TCS} vs. $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ (Fig. S6) to observe whether a functional relationship can be deduced from the computational data alone. The mutational change in the coevolutionary landscape, ΔH_{TCS}, would in principle exhibit a non-linear dependence on $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$, while also depending on many other quantities including the binding affinity with other proteins. We find no obvious relationship between the two quantities in our current analysis.

A histogram of $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ is plotted for the 42,985 mutants in Fig. 5A. A histogram of ΔH_{TCS} for the same subset of mutants is shown in Fig. S7A. On the population level, functional mutations exhibit a mean $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ of 1.76 ± 0.06 kcal/mol lower than that of the non-functional mutants, with a Wilcoxon rank-sum test p-value < 2.2×10^{-16}. This indicates that destabilizing mutations of ~2 kcal/mol are sufficient for disrupting TCS. Furthermore, destabilizing mutations that are more than 2 standard deviations greater than the mean $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ for functional variants are significantly less likely to be functional, with a p-value < 10^{-6} computed from a cumulative binomial distribution (based on the 6157 mutants above this threshold, 19 of which are functional).

We next examine the potentially deleterious effect of mutations that overly stabilize the binding affinity between PhoQ and PhoP. Although we find that all 56 mutants with $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}} < -5$ kcal/mol are non-functional (Fig. 5A), this has no statistical significance (p-value ~ 0.4). Fig. 5A illustrates our point that the functional mutations tend to have a neutral affect on $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$, while high $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ is strongly associated with the loss of function. Although very low $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ may visually appear to be enriched with non-functional mutants, this is based on a small number of mutants and does not have statistical significance.

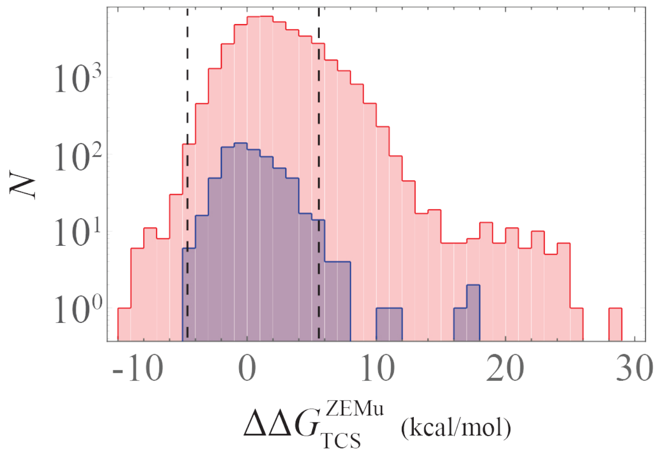

Figure 5. Mutational change in binding affinity for PhoQ/PhoP interaction. A histogram of $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ (See Materials and Methods), is plotted for the 702 functional (blue) and 42,283 non-functional (red) mutational variants analyzed in our study. The dashed lines denote ±2 standard deviations from the mean of the functional (blue) distribution.

Discussion

Treating a large collection of amino acid sequence data for TCS partner proteins as independent samples from a Boltzmann equilibrium distribution, we infer a coevolutionary landscape, H_{TCS}. Specifically, H_{TCS} is proportional to the negative of the additive fitness landscape, which captures the coevolving amino acid combinations that give rise to interaction specificity in TCS systems. In the past, we were able to predict how a point mutation to a TCS protein affects its ability to transfer signal to its partner *in vitro* (Cheng, et al. 2014). Our present work shifts the paradigm of coevolution-based analysis towards systems biology by extending our analysis to include how those mutations affect "cross-talk" in a bacterial organism. We demonstrate that our most favorable predictions for multiple site mutations can accurately capture *in vivo* TCS functionality, consistent with the comprehensive mutational study of Podgornaia and Laub (Podgornaia and Laub 2015). This is not a trivial computational task, since inferring coevolutionary information from sequence data is highly underdetermined and estimates ~10^7 parameters (for Eq. 3) from ~10^3 sequences of TCS partners. Adding to the problem complexity, it is plausible that the full functional sequence space has not yet being explored by evolutionary process (Capra and Laub 2012; Echave, et al. 2016). Despite this, the coevolutionary landscape is predictive and identifies mutational variants that are not found in nature, e.g., none of the mutational sequences are included as input data in our model. We have demonstrated the feasibility of generating predictions using coevolution and the predictive power of such an approach will only systematically improve as more sequences of TCS partners are collected.

Similar predictions to those discussed herein can readily be used to engineer novel protein-protein interactions in TCS systems. Such a strategy would potentially complement already existing strategies to match novel inputs with outputs via modular engineering (Tabor, et al. 2011; Whitaker, et al. 2012; Ganesh, et al. 2013; Schmid, et al. 2014; Hansen and Benenson 2016). The strength of our coevolutionary approach is that it makes possible an efficient search of sequence-space for mutations at arbitrary positions in either the HK or RR that desirably enhance or suppress its interaction with a RR or HK, respectively. It can also readily be applied to study the *in vivo*, system-level effect of mutating a TCS protein on insulating its interaction with a desired partner or enabling “cross-talk” with non-partners. Our study highlights an intuitive but key principle for selecting mutations to a TCS protein that encodes specificity *in vivo*: mutations must be selected to enhance protein-protein interactions with a desired partner while limiting protein-protein interactions with undesired partners. While also intuitive, we demonstrate that mutations that significantly destabilize the binding affinity result in the loss of signaling. Further, we estimate that destabilization of ~2 kcal/mol in the binding affinity between TCS partners is sufficient to disrupt TCS.

It is also important to note that coevolutionary methods described here for identifying mutational phenotypes (e.g., response to magnesium stress) is that they are not particular to TCS systems. This framework is transferable to other systems where molecular interactions coevolve to preserve function, opening the window to a large set of open problems in molecular and systems biology. Our results further extend the idea that a combination of coevolutionary based methods, molecular modeling and experiment can be used to identify the proper amino acids sites and identities that can be used to identify mutational phenotypes. Our study highlights the important role of coevolution in maintaining protein-protein interactions, as in the case of bacteria signal transduction. Statistical methods that probe coevolution not only allow us to connect molecular, residue-level details to mutational phenotypes, but also to explore the evolutionary selection mechanisms that are employed by nature to maintain interaction specificity, e.g., negative selection (Zarrinpar, et al. 2003). Further investigations of other systems that are evolutionarily constrained to maintain protein-protein interactions could elucidate the extent at which our methods can be used in alternative systems. One potential example is the toxin-antitoxin protein pairs in bacteria, which was the focus of recent experimental work (Aakre, et al. 2015) elucidating the determinants of interaction specificity.

Materials and Methods

Sequence database for HK and RR inter-protein interactions: DHp and REC

We obtain multiple sequence alignments (MSA) from Pfam (Finn, et al. 2014) (Version 28), focusing on the DHp (PF00512) and REC (PF00072) domains of the HK and RR, respectively (Fig. 1A). The first 4 positions (columns) of PF00512 were removed due to poor alignment of the PhoQ sequence at those positions. The remaining DHp MSA has a length of $L_{\mathrm{DHp}} = 60$. Each REC MSA had a default length of $L_{\mathrm{REC}} = 112$. Here, we considered HK proteins that have the same domain architecture as the PhoQ kinase from *E. coli*, i.e., DHp domain sandwiched between an N-terminal HAMP domain (PF00672) and a C-terminal ATPase domain (PF02518). The remaining HK (DHp) sequences were paired with a TCS partner RR (REC) by taking advantage of the observation that TCS partners are typically encoded adjacent to one another under the same operon (Skerker, et al. 2005; Yamamoto, et al. 2005), i.e., ordered locus numbers differ by 1. Further, we exclude all TCS pairs that are encoded adjacent to multiple HKs or RRs. Each DHp and REC sequence that was paired in this fashion was concatenated into a sequence (Fig. 1A), $S_{\mathrm{TCS}} = (A_1, A_2, ..., A_{L-1}, A_L)$ of total length L where A_i is the amino acid at position i which is indexed from 1 to $q = 21$ for the 20 amino acids and MSA gap. The DHp sequence is indexed from positions 1 to L_{DHp} and REC sequence from positions $L_{\mathrm{DHp}} + 1$ to the total length of $L = L_{\mathrm{DHp}} + L_{\mathrm{REC}} = 172$. Our remaining dataset consisted of 6,519 non-redundant concatenated sequences.

Inference of parameters of coevolutionary model

An amino acid sequence $s = (A_1, A_2, ..., A_L)$ for a protein or interacting proteins can be viewed as being selected from a Boltzmann equilibrium distribution, i.e., $P(s) = Z^{-1}\exp(-H(s))$. The Boltzmann form of P was previously derived for an evolving population in the limit where the product of the population size and mutation rate is very small (Sella and Hirsh 2005). Specifically, it was shown (Sella and Hirsh 2005) that $H(s) = -\nu x(s)$, where ν is the population size and $x(s)$ is the additive fitness landscape (i.e., log of the fitness). A high population size suggests many viable sequences that a protein can mutate to, which makes the population more robust to deleterious mutations. Related work (Halpern and Bruno 1998) modeled site-specific selection of sequences, which has been extended upon by numerous works (Tamuri, et al. 2012; Spielman and Wilke 2015; Bloom 2016).

Under certain limiting conditions, $H(s)$ appears to share a correspondence with the energetics of protein folding (See review: (Sikosek and Chan 2014)). Assuming that the sequence diversity is completely due to stability considerations, $H(s)=\beta E(s)$ where $E(s)$ is the energy of the folded protein with respect to the unfolded state and $\beta=(k_B T_{sel})^{-1}$ is the inverse of the evolutionary selection temperature from protein folding theory (Pande, et al. 1997, 2000; Morcos, et al. 2014). Several studies have reported strong linear correlation between mutational changes in $H(s)$ with mutational changes in protein stability (Lui and Tiana 2013; Morcos, et al. 2014; Contini and Tiana 2015). However, $H(s)=\beta E(s)$ may not be an appropriate approximation for proteins that have evolved with interacting partners, for which sequence selection is plausibly influenced by additional factors such as binding affinities as well as binding/unbinding rates.

Often it is of interest to solve the inverse problem of inferring an appropriate $H(s)$ when provided with an abundant number of protein sequences. Typical approaches to this problem have applied the principle of maximum entropy to infer a least biased model that is consistent with the input sequence data (Weigt, et al. 2009; Morcos, et al. 2011), e.g., the empirical single-site and pairwise amino acid probabilities, $P_i(A_i)$ and $P_{ij}(A_i, A_j)$, respectively. The solution of which is the Potts model:

$$H(\mathrm{s})=-\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} J_{ij}(A_i,A_j)-\sum_{i=1}^{L} h_i(A_i) \tag{3}$$

where A_i is the amino acid at position i for a sequence in the MSA, $J_{ij}(A_i, A_j)$ is the pairwise statistical couplings between positions i and j in the MSA with amino acids A_i and A_j, respectively, and $h_i(A_i)$ is the local field for position i. We estimate the parameters of the Potts model, $\{\mathbf{J},\mathbf{h}\}$, using the pseudo-likelihood maximization Direct Coupling Analysis (plmDCA) (See Ref: (Ekeberg, et al. 2013) for full computational details). It is important to note that the mutational context-dependence (epistasis) between residues i and j is naturally captured in the model through the statistical couplings, $J_{ij}(A_i, A_j)$.

Previous studies have applied DCA to a number of problems in structural biology. DCA has been used to identify highly coevolving pairs of residues to predict the native state conformation of a protein (Marks, et al. 2012; Sulkowska, et al. 2012; Sutto, et al. 2015), including repeat proteins (Espada, et al. 2015), as well as identify additional functionally relevant conformational states (Morcos, et al. 2013; Malinverni, et al. 2015; Sutto, et al. 2015) and multi-meric states (Schug, et al. 2009; Weigt, et al. 2009; Morcos, et al. 2011; dos Santos, et al. 2015; Malinverni, et al. 2015). Structural and coevolutionary information share complementary roles in the molecular simulations of proteins (See review: (Noel, et al. 2016)). The Potts model (Eq. 3) obtained from DCA has been related to the theory of evolutionary sequence selection (Morcos, et al. 2014) as well as mutational changes in protein stability (Lui and Tiana 2013; Morcos, et al. 2014; Contini and Tiana 2015). Additional work has applied DCA to protein folding to predict the effect of point mutations on the folding rate (Mallik, et al. 2016) as well as construct a statistical potential for native contacts in a structure-based model of a protein (Cheng, et al. 2016) to better capture the transition state ensemble.

DCA and inference methods have also been applied to study problems in systems biology, such as the identification of relevant protein-protein interactions in biological interaction networks (Procaccini, et al. 2011; Feinauer, et al. 2016). Recently, a number of studies have focused on inferring quantitative landscapes that capture the effects of mutations on biological phenotypes (Ferguson, et al. 2013; Cheng, et al. 2014; Figliuzzi, et al. 2016) by constructing models from sequence data (e.g., Eq. 3). Two separate studies, which focused on antibiotic drug resistance in *E. coli* (Figliuzzi, et al. 2016) and viral fitness of HIV-1 proteins (Ferguson, et al. 2013), respectively, inferred a Potts model (i.e., additive fitness landscape), H, and calculated mutational changes as ΔH. This approach is analogous to the approach adopted in this study. Likewise, a study examining the mutational effects on TCS phosphotransfer (Cheng, et al. 2014) constructed a mutational landscape from an information-based quantity. All of these approaches capture epistatic effects and rely on the accuracy of the inferred Potts model (Eq. 3) from sequence data.

Mutational changes in coevolutionary landscape

For the concatenated sequences of HK (DHp) and RR (REC) (Fig. 1A), we infer a Potts model (Equation 3). We focus on a subset of parameters in our model consisting of the interprotein couplings, J_{ij}, between positions in the DHp and REC domains (Fig. 1C) that are in close proximity in a representative structure of the TCS complex (Fig. 1B). All local fields terms, h_i, are included to partially capture the fitness effects that give rise to the amino acid composition observed at each site. These considerations allow us to construct coevolutionary landscape for TCS, which is a negative, additive fitness landscape:

$$H_{\mathrm{TCS}}(S_{\mathrm{TCS}}) = -\sum_{i=1}^{L_{\mathrm{DHp}}} \sum_{j=L_{\mathrm{DHp}}+1}^{L_{\mathrm{DHp}}+L_{\mathrm{REC}}} J_{ij}(A_i, A_j) \times \Theta(c - r_{ij}) - \sum_{i=1}^{L_{\mathrm{DHp}}+L_{\mathrm{REC}}} h_i(A_i) \qquad (5)$$

where S_{TCS} is the concatenated sequence of the DHp and REC domains, the double summation is taken over all interprotein statistical couplings between the DHp and REC domains, Θ is a

Heaviside step function, c is the a cutoff distance of 16Å which was determined in a previous study (Morcos, et al. 2014), and r_{ij} is the minimum distance between residues i and j in the representative structure. Mutational changes in Eq. 5 are then computed as $\Delta H_{\text{TCS}}(S_{\text{TCS}}^{\text{mutant}}) = H_{\text{TCS}}(S_{\text{TCS}}^{\text{mutant}}) - H_{\text{TCS}}(S_{\text{TCS}}^{\text{WT}})$. Hence, mutational changes in the signal transfer efficiency are approximated from mutational changes in the additive fitness.

An additional model is considered to analyze the epistatic effects of the 4-point mutations explored in experiment (Podgornaia and Laub 2015). While Eq. 5 naturally captures the epistatic effect of mutating a residue in the HK when the RR has also been mutated (or vice versa), it does not explicitly contain the statistical couplings that capture the epistatic effects of HK only mutations. Hence, the statistical couplings between the 4-mutational sites of the HK (Fig. 4C) are explicitly added to Eq. 5:

$$f_{\text{epistasis}}(S_{\text{TCS}}) = -\sum_{i,j \in \text{mutational sites}} J_{ij}(A_i, A_j) \tag{6a}$$

$$H_{\text{TCS}}^{(\text{epistasis})}(S_{\text{TCS}}) = H_{\text{TCS}}(S_{\text{TCS}}) + f_{\text{epistasis}}(S_{\text{TCS}}) \tag{6b}$$

In Eq. 6a, the summation contains the 6 statistical couplings between the 4-mutational sites explored by Podgornaia and Laub, i.e., positions 14, 15, 18 and 19 in our MSA of TCS partners, which map to positions 254, 255, 258, and 259 of PhoQ.

Zone Equilibration of Mutants (ZEMu) calculation

ZEMu consists of a multiscale minimization by dynamics, restricted to a flexibility zone of five residues about each substitution site (Dourado and Flores 2014), which is followed by a mutational change in stability using FoldX (Guerois, et al. 2002). ZEMu has been used to explain the mechanism of Parkinson's disease associated mutations in Parkin (Caulfield, et al. 2014; Fiesel, et al. 2015). The minimization is done in MacroMoleculeBuilder (MMB), a general-purpose internal coordinate mechanics code also known for RNA folding (Flores and Altman 2010), homology modeling (Flores, et al. 2010), morphing (Tek, et al. 2016), and fitting to density maps (Flores 2014).

We use the Zone Equilibration of Mutants (ZEMu) (Dourado and Flores 2014) method to predict the mutational change in binding energy between PhoQ and PhoP. ZEMu first treats mutations as small perturbations on the structure by using molecular dynamics simulations (See Ref. (Dourado and Flores 2014) for full computational details) to equilibrate the local region around mutational sites. ZEMu can then estimate the binding affinity between the mutant-PhoQ/PhoP, $\Delta G_{\text{TCS}}^{\text{ZEMu}}(\text{mutant})$, and the wild type-PhoQ/PhoP, $\Delta G_{\text{TCS}}^{\text{ZEMu}}(\text{WT})$, using the knowledge-based FoldX (Guerois, et al. 2002) potential. This allows for the calculation of the mutational change in binding affinity as: $\Delta\Delta G_{\text{TCS}}^{\text{ZEMu}} = \Delta G_{\text{TCS}}^{\text{ZEMu}}(\text{mutant}) - \Delta G_{\text{TCS}}^{\text{ZEMu}}(\text{WT})$.

ZEMu calculation was performed according to Ref: (Dourado and Flores 2014), with the following two differences. First, due to the large number of mutations we capped the computer time permitted to 3 core-hours per mutant, whereas in (Dourado and Flores 2014) the limit was 48 hours. This meant that of 122802 mutants attempted, 42923 completed within the time limit, whereas in (Dourado and Flores 2014), almost all mutants converged. The major reason for non-convergence in the current work involved mutation to bulky or constrained residues. Steric clashes produced by such residues force the error-controlled integrator (Flores, et al. 2011) to take small time steps and hence use more computer time. Exemplifying this, the amino acids F, W and Y are the most common residues for non-converging mutations at positions 285 and 288 in PhoQ. The second difference was that we permitted flexibility in the neighborhood of all four possible mutation sites, even when not all of them were mutated, whereas in (Dourado and Flores 2014) only the mutated positions were treated as flexible. This allowed us to compare all of the mutational energies to a single wild type simulation, also performed with flexibility at all four sites.

Database of TCS partners, Potts model, and code for calculating Eq. 5 can be obtained from: http://utdallas.edu/~faruckm/PublicationDatasets.html

Acknowledgments: We would like to thank Drs. Michele Di Pierro and Lena Simine for helpful comments. **Funding**: This research was supported by the NSF INSPIRE award (MCB-1241332) and the NSF-funded Center for Theoretical Biological Physics (PHY-1427654). SF and ON acknowledge funds from eSSENCE (http://essenceofescience.se/), Uppsala University, and the Swedish Foundation for International Cooperation in Research and Higher Education (STINT). We also acknowledge a generous allocation of supercomputer time from the Swedish National Infrastructure for Computing (SNIC) at Uppmax, and applications assistance from Drs. Rudberg, Karlsson, and Freyhult.

Author contributions: R.R.C., F.M., and J.N.O designed the research with the assistance of H.L., S.C.F., and O.N.; R.R.C and O.N. performed the research; R.R.C. and R.L.H. curated the protein databases that were used in our study; R.R.C., F.M., S.F. and O.N. wrote the paper.

Competing interests: The authors declare no competing interests.

References

Aakre CD, Herrou J, Phung TN, Perchuk BS, Crosson S, Laub MT. 2015. Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. Cell 163:594-606.

Bloom JD. 2016. Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. In. bioRxiv: Cold Spring Harbor Labs Journals.

Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. Proteins-Structure Function and Genetics 21:167-195.

Bryngelson JD, Wolynes PG. 1987. Spin-Glasses and the Statistical-Mechanics of Protein Folding. Proc Natl Acad Sci U S A 84:7524-7528.

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

Capra EJ, Laub MT. 2012. Evolution of two-component signal transduction systems. Annu Rev Microbiol 66:325-347.

Capra EJ, Perchuk BS, Lubin EA, Ashenberg O, Skerker JM, Laub MT. 2010. Systematic Dissection and Trajectory-Scanning Mutagenesis of the Molecular Interface That Ensures Specificity of Two-Component Signaling Pathways. PLoS Genetics 6:e1001220.

Casino P, Rubio V, Marina A. 2010. The mechanism of signal transduction by two-component systems. Current Opinion in Structural Biology 20:763-771.

Casino P, Rubio V, Marina A. 2009. Structural Insight into Partner Specificity and Phosphoryl Transfer in Two-Component Signal Transduction. Cell 139:325-336.

Caulfield TR, Fiesel FC, Moussaud-Lamodiere EL, Dourado DFAR, Flores SC, Springer W. 2014. Phosphorylation by PINK1 Releases the UBL Domain and Initializes the Conformational Opening of the E3 Ubiquitin Ligase Parkin. PLoS Comput Biol 10:e1003935.

Cheng RR, Morcos F, Levine H, Onuchic JN. 2014. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci U S A 111:E563-E571.

Cheng RR, Raghunathan M, Noel JK, Onuchic JN. 2016. Constructing sequence-dependent protein models using coevolutionary information. Protein Science 25:111-122.

Contini A, Tiana G. 2015. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. J Chem Phys 143:025103.

Dago AE, Schug A, Procaccini A, Hoch JA, Weigt M, Szurmant H. 2012. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. Proceedings of the National Academy of Sciences 109:E1733-E1742.

de Juan D, Pazos F, Valencia A. 2013. Emerging methods in protein co-evolution. Nat Rev Genet 14:249-261.

dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN. 2015. Dimeric interactions and complex formation using direct coevolutionary couplings. Scientific Reports 5:13652.

Dourado DFAR, Flores SC. 2014. A multiscale approach to predicting affinity changes in protein-protein interfaces. Proteins-Structure Function and Bioinformatics 82:2681-2690.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. Nat Rev Genet 17:109-121.

Ekeberg M, Lovkvist C, Lan YH, Weigt M, Aurell E. 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Physical Review E 87:012707.

Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. 2015. Capturing coevolutionary signals inrepeat proteins. BMC Bioinformatics 16:207.

Feinauer C, Szurmant H, Weigt M, Pagnani A. 2016. Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. PLoS One 11:e0149166.

Ferguson AL, Mann JK, Omarjee S, Ndung'u T, Walker BD, Chakraborty AK. 2013. Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. Immunity 38:606-617.

Ferreiro DU, Komives EA, Wolynes PG. 2014. Frustration in biomolecules. Quarterly Reviews of Biophysics 47:285-363.

Fiesel FC, Caulfield TR, Moussaud-Lamodiere EL, Ogaki K, Dourado DFAR, Flores SC, Ross OA, Springer W. 2015. Structural and Functional Impact of Parkinson Disease-Associated Mutations in the E3 Ubiquitin Ligase Parkin. Human Mutation 36:774-786.

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. 2016. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Molecular Biology and Evolution 33:268-280.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. Nucleic Acids Research 42:D222-D230.

Flores SC. 2014. Fast fitting to low resolution density maps: elucidating large-scale motions of the ribosome. Nucleic Acids Res 42:e9.

Flores SC, Altman RB. 2010. Turning limited experimental information into 3D models of RNA. Rna-a Publication of the Rna Society 16:1769-1778.

Flores SC, Sherman MA, Bruns CM, Eastman P, Altman RB. 2011. Fast Flexible Modeling of RNA Structure Using Internal Coordinates. Ieee-Acm Transactions on Computational Biology and Bioinformatics 8:1247-1257.

Flores SC, Wan Y, Russell R, Altman RB. 2010. Predicting RNA structure by multiple template homology modeling. Pac Symp Biocomput:216-227.

Ganesh I, Ravikumar S, Lee SH, Park SJ, Hong SH. 2013. Engineered fumarate sensing Escherichia coli based on novel chimeric two-component system. J Biotechnol 168:560-566.

Gobel U, Sander C, Schneider R, Valencia A. 1994. Correlated mutations and residue contacts in proteins. Proteins 18:309-317.

Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. Journal of Molecular Biology 320:369-387.

Halpern AL, Bruno WJ. 1998. Evolutionary distances for protein-coding sequences: Modeling site-specific residue frequencies. Molecular Biology and Evolution 15:910-917.

Hansen J, Benenson Y. 2016. Synthetic biology of cell signaling. Natural Computing 15:5-13.

Hoch JA. 2000. Two-component and phosphorelay signal transduction. Current Opinion in Microbiology 3:165-170.

Laub MT, Goulian M. 2007. Specificity in Two-Component Signal Transduction Pathways. Annual Review of Genetics 41:121-145.

Leopold PE, Montal M, Onuchic JN. 1992. Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship. Proc Natl Acad Sci U S A 89:8721-8725.

Li L, Shakhnovich EI, Mirny LA. 2003. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. Proceedings of the National Academy of Sciences 100:4463-4468.

Lui S, Tiana G. 2013. The network of stabilizing contacts in proteins studied by coevolutionary data. J Chem Phys 139:155103.

Malinverni D, Marsili S, Barducci A, De Los Rios P. 2015. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. PLoS Comput Biol 11:e1004262.

Mallik S, Das S, Kundu S. 2016. Predicting protein folding rate change upon point mutation using residue-level coevolutionary information. Proteins-Structure Function and Bioinformatics 84:3-8.

Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. Nat Biotechnol 30:1072-1080.

Morcos F, Jana B, Hwa T, Onuchic JN. 2013. Coevolutionary signals across protein lineages help capture multiple protein conformations. Proceedings of the National Academy of Sciences 110:20533-20538.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences 108:E1293-E1301.

Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. 2014. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Natl Acad Sci U S A 111:12408-12413.

Neher E. 1994. How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci U S A 91:98-102.

Noel JK, Morcos F, Onuchic JN. 2016. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. F1000Res 5.

Onuchic JN, LutheySchulten Z, Wolynes PG. 1997. Theory of protein folding: The energy landscape perspective. Annual Review of Physical Chemistry 48:545-600.

Pande VS, Grosberg AY, Tanaka T. 2000. Heteropolymer freezing and design: Towards physical models of protein folding. Reviews of Modern Physics 72:259-314.

Pande VS, Grosberg AY, Tanaka T. 1997. Statistical mechanics of simple models of protein folding and design. Biophysical Journal 73:3192-3210.

Podgornaia AI, Laub MT. 2013. Determinants of specificity in two-component signal transduction. Current Opinion in Microbiology 16:156-162.

Podgornaia AI, Laub MT. 2015. Pervasive degeneracy and epistasis in a protein-protein interface. Science 347:673-677.

Procaccini A, Lunt B, Szurmant H, Hwa T, Weigt M. 2011. Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. PLoS One 6:e19729.

Qin L, Cai S, Zhu Y, Inouye M. 2003. Cysteine-Scanning Analysis of the Dimerization Domain of EnvZ, an Osmosensing Histidine Kinase. Journal of Bacteriology 185:3429-3435.

Schmid SR, Sheth RU, Wu A, Tabor JJ. 2014. Refactoring and Optimization of Light-Switchable Escherichia coli Two-Component Systems. Acs Synthetic Biology 3:820-831.

Schug A, Weigt M, Onuchic JN, Hwa T, Szurmant H. 2009. High-resolution protein complexes from integrating genomic information with molecular simulation. Proceedings of the National Academy of Sciences 106:22124-22129.

Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A 102:9541-9546.

Shindyalov IN, Kolchanov NA, Sander C. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7:349-358.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. Journal of the Royal Society Interface 11:20140419.

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 133:1043-1054.

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 Biology 3:1770-1788.

Spielman SJ, Wilke CO. 2015. The Relationship between dN/dS and Scaled Selection Coefficients. Molecular Biology and Evolution 32:1097-1108.

Stock AM, Robinson VL, Goudreau PN. 2000. Two-component signal transduction. Annual Review of Biochemistry 69:183-215.

Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN. 2012. Genomics-aided structure prediction. Proceedings of the National Academy of Sciences 109:10340-10345.

Sutto L, Marsili S, Valencia A, Gervasio FL. 2015. From residue coevolution to protein conformational ensembles and functional dynamics. Proc Natl Acad Sci U S A 112:13567-13572.

Szurmant H, Hoch JA. 2010. Interaction fidelity in two-component signaling. Current Opinion in Microbiology 13:190-197.

Tabor JJ, Levskaya A, Voigt CA. 2011. Multichromatic control of gene expression in Escherichia coli. J Mol Biol 405:315-324.

Tamuri AU, dos Reis M, Goldstein RA. 2012. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. Genetics 190:1101-1115.

Tek A, Korostelev AA, Flores SC. 2016. MMB-GUI: a fast morphing method demonstrates a possible ribosomal tRNA translocation trajectory. Nucleic Acids Res 44:95-105.

Tzeng Y-L, Hoch JA. 1997. Molecular recognition in signal transduction: the interaction surfaces of the Spo0F response regulator with its cognate phosphorelay proteins revealed by alanine scanning mutagenesis. Journal of Molecular Biology 272:200-212.

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. 2009. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci U S A 106:67-72.

Whitaker WR, Davis SA, Arkin AP, Dueber JE. 2012. Engineering robust control of two-component system phosphotransfer using modular scaffolds. Proc Natl Acad Sci U S A 109:18090-18095.

Wolynes PG. 2015. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 119:218-230.

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. Journal of Biological Chemistry 280:1448-1456.

Zarrinpar A, Park S-H, Lim WA. 2003. Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426:676-680.