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]*

One Sentence Summary: A statistical energy function for bacterial signaling captures mutational phenotypes.

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:
We construct a statistical model of bacterial two-component signaling (TCS) proteins for predicting amino acid configurations that allow signaling partners to preferentially interact. This model is applied to a recent exhaustive mutational experiment of 4 interfacial residues on the histidine kinase of the magnesium response TCS system in *E. coli*. We demonstrate that our top mutational predictions can accurately capture experimentally observed mutational variants that preserve interaction specificity between TCS partners. Interestingly, we can isolate the true positive predictions by focusing on mutations that we predict to limit signal transfer with non-partner TCS proteins (i.e., "cross-talk"). This demonstrates that our model also captures the amino acid configurations that lead to "cross-talk" between non-partner TCS proteins, which can be used to engineer specificity. We further supplement our analysis by calculating the mutational change in the binding affinity between TCS partners, supporting the intuitive concept that overly destabilizing mutations disrupt TCS.

Introduction

Early theoretical work on protein folding postulated that proteins have evolved to be minimally frustrated (*1-3*), i.e., evolved to have favorable residue-residue interactions that facilitate folding into the native state while having minimal non-native energetic traps. Frustration provides intuition as to why protein sequences are far from random strings of amino acids. The evolutionary constraint to be able to fold into a particular, stable three-dimensional structure while minimizing the number of frustrated interactions greatly restricts the sequence-space of a protein (*1, 3, 4*). Close observation of the collection of amino acid sequences for a particular protein family would reveal that the satisfaction of these evolutionary constraints manifests itself in the sequences as correlated mutational patterns between different positions in the protein such as, for example, positions that form native contacts (*5-7*). We refer to these quantifiable correlated amino acid identities between different positions in a protein 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 (*8-10*). 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 of cell signaling that can be found in bacteria is two-component signaling (TCS) (*11-16*), 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 intracellular HK generates a signal via autophosphorylation. Its intracellular partner RR can then transiently bind to it and receive the signal (i.e., phosphoryl group), thereby activating its ability to homodimerize and act as a transcription factor that can up- or down-regulate genes. Furthermore, 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 a TCS protein from transferring signal to/from the wrong partner (i.e., "crosstalk"), the HK and RR proteins have mutually evolved amino acids at their binding interface to give rise to interaction specificity (*14-16*). Thus, the collection of protein sequences of TCS partners contains quantifiable coevolution between the HK and RR sequences.

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:

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

where S_{TCS} is the concatenated amino acid sequence of a HK and RR protein, P is the probability of selecting S_{TCS}, Z is the normalization (partition function), T_{Sel} is the evolutionary selection temperature (*17*), and H is an appropriate energetic function in units of $k_B T_{Sel}$. Recently, maximum entropy-based approaches referred to as Direct Coupling Analysis (DCA) (*18-20*) have been successfully applied to infer the parameters of H (a Potts model) that governs the empirical amino acid sequence statistics, allowing for the direct quantification of the coevolution in protein sequence data (See Review: (*21*)). Early work using these types of models to study TCS primarily focused on identifying the key coevolving residues between the HK and RR (*20*) and subsequently using them as docking constraints in a molecular dynamics simulation to predict the HK/RR signaling complex (*22*). This work was extended to predict the autophosphorylation structure of a HK (*23*) as well the homodimeric form (transcription factor) of the RR (*24*). Recently, DCA-based approaches have been applied to quantify the determinants of interaction specificity between TCS partners (*25, 26*), building on earlier coevolutionary

approaches (*27, 28*). 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 (*26*).

The experimental effort to determine the molecular origin of interaction specificity in TCS proteins (See Reviews: (*13-15, 29*)) precedes the recent parallel computational efforts. Full knowledge of the binding interface between HK and RR was made possible through X-ray crystallography (*30*). Scanning mutagenesis studies (*31-33*) offered insight on the subset of interfacial residues that were critical for determining specificity. These sites were used as mutational sites to engineer a TCS protein to bind and transfer signal to a non-partner (*31, 34*). However, the extent of the functional sequence-space of a TCS protein that gives rise to interaction specificity remained unanswered until recent comprehensive work by Podgornaia and Laub (*35*) on the magnesium response TCS system in *E. coli,* PhoQ/PhoP. Under low magnesium concentrations, PhoQ (HK) autophosphorylates and transfers a phosphoryl group to its partner, PhoP (RR), regulating the response to low magnesium stress. Using exhaustive mutagenesis of 4 residues of PhoQ (20^4 = 160,000 mutational variants) at positions that form the binding interface with PhoP, it was found that roughly 1% of all mutational variants were fully functional HKs—i.e., capable of phosphotransfer as well as phosphatase activity with its partner RR. 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 coevolutionary information obtained with DCA could capture the functional mutational variants observed in the exhaustive mutational study of PhoQ and if so, to what extent? 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. This question is also important to assess the potential predictive power of coevolutionary models. Current DCA algorithms (*18, 19*) are computationally inexpensive and can infer an energy function (see Eq. 1) that forms the basis for determining mutations to the WT sequence that are more favorable to maintain specificity. We answer this question by first constructing a coevolutionary statistical energy, H_{TCS}, as a proxy for signal transfer efficiency. We then assess how mutations affect the HK/RR interaction by computing the mutational change in our energy function between the mutant sequence, $S_{\mathrm{TCS}}^{\mathrm{mutant}}$, and the wild type sequence, $S_{\mathrm{TCS}}^{\mathrm{WT}}$:

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

It is important to note that the mutational study of Podgornaia and Laub has a dichotomous result (i.e., mutations result in either functional or non-functional variants) and thus, mutations that are weakly unfavorable to signaling can still result in a functional PhoQ kinase. We apply Equation 2 to assess the mutational change in energy for the 20^4 variants explored in the exhaustive study of PhoQ and find that the most favorable mutational variants of our model correspond mostly to functional mutational variants identified in experiment—i.e., true positive predictions. Further, we find that many of the non-functional variants predicted to be favorable mutations by our proxy for signal transfer efficiency, H_{TCS}, appear to exhibit "cross-talk" with non-partner RR when our analysis is extended to include the interaction of PhoQ with all RR proteins in *E. coli*. If we exclude these promiscuous variants, we can better isolate the true positive predictions that are functional mutants from false positives that are non-functional mutants. We next constructed a coevolutionary energy function for the HK alone, H_{HK}, to assess the extent to which a model that only considers the HK can capture the experimentally observed mutational phenotypes.

While such a model is not better at identifying functional mutations than H_{TCS}, we find the most favorable mutations that preserve autophosphorylation are also the most favorable for preserving the HK/RR interaction, demonstrating the evolutionary pressure to simultaneously preserve both monomeric function and complex formation. Finally, we estimate the mutational change in binding affinity in the PhoQ/PhoP bound complex for a substantive, randomly selected subset of the mutational variants using the Zone Equilibration of Mutants (ZEMU) method (*36*), 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. Further, we provide inconclusive support for the view that the mutations that overly strengthen the binding affinity may also be deleterious towards TCS. A more detailed description of our computational approaches can be found in the Materials and Methods section.

The work described herein demonstrates that a coevolutionary model 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.

Results

Mutational change in coevolutionary energy, ΔH_{TCS}, between *PhoQ and PhoP*

Considering the 1,659 functional and 158,341 non-functional PhoQ-mutational variants, identified by Podgornaia and Laub (*35*), we first quantify the effect of mutating PhoQ on its interaction with its partner, PhoP, using our energy function, H_{TCS}, as a proxy for signal transfer efficiency (Eq. 5). Focusing on the Dimerization and Histidine phosphotransfer domain (DHp) and the Receiver (REC) domain (Fig. 1A) which form the HK/RR binding interface (Fig. 1C), we compute the mutational change in our coevolutionary energy, ΔH_{TCS}, for the experimentally determined functional and non-functional mutants. A histogram of ΔH_{TCS} is generated for all mutational variants (Fig. 2A), where $\Delta H_{\mathrm{TCS}} = 0$ corresponds to the WT by definition and $\Delta H_{\mathrm{TCS}} < 0$ corresponds to mutations that we predict to be more favorable to PhoQ/PhoP signaling than the WT. The distribution of the functional mutants tends more towards favorable ΔH_{TCS} than the distribution of non-functional mutants by our model, 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).

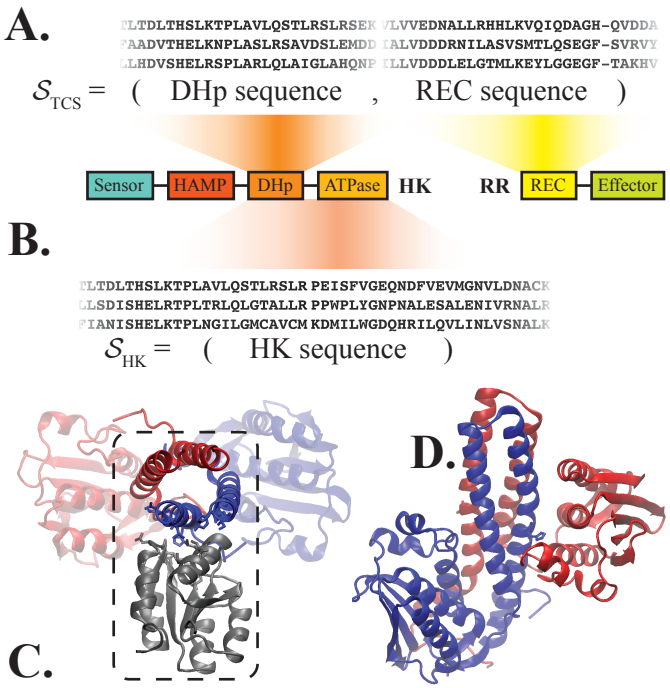

Fig. 1. **TCS domain interactions of interest.** We focus only on HK proteins that have the following domain architecture from N to C terminus: sensor, HAMP, DHp, and ATPase. Likewise, we consider RR proteins that consist of a REC domain followed by an effector domain. (A) The interaction between the 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 for more details). (B) We also consider a model of the individual HK, while focusing on the segment of each HK sequence that contains only the DHp and ATPase domains, S_{HK}. (C) A representative structure of the HK/RR TCS complex previously predicted for the KinA/Spo0F complex in *B. subtilis (26)*. 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. (D) The crystal structure of a representative autophosphorylation state is shown for the HAMP-containing HK CpxA in *E. coli (37)*. The ATPase domain of one HK protein (red) is bound to the DHp domain of the other HK protein (blue).

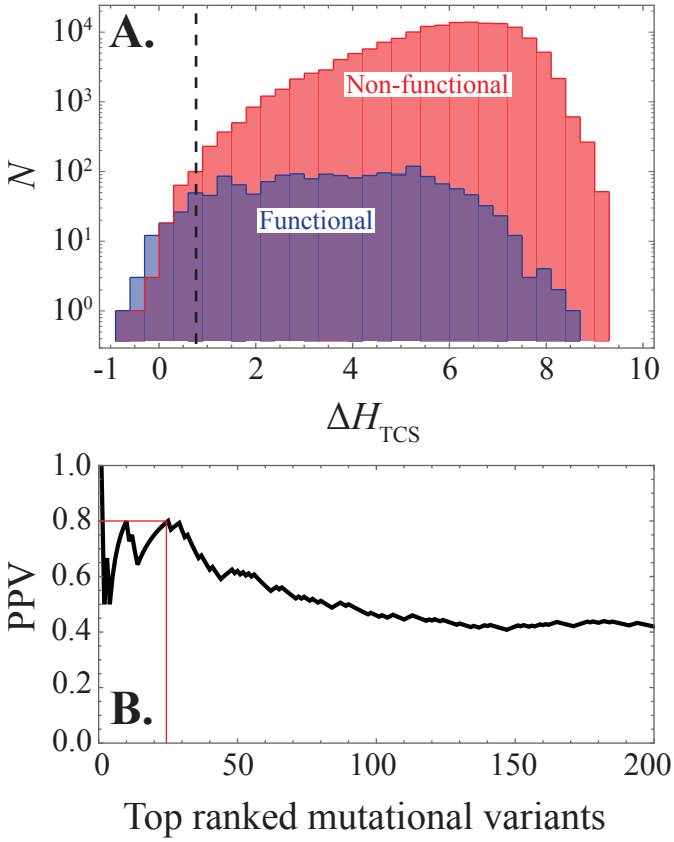

Fig. 2. **Effect of mutations on the PhoQ/PhoP interaction.** (A) A histogram of the mutational change in our coevolutionary energy, ΔH_{TCS} (Eq. 5), is plotted for the functional (blue) and non-functional (red) mutational variants reported by Podgornaia and Laub (*29*). The color purple shows parts 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, $\Delta H_{TCS} = 0$ corresponds to the mutational change in energy with respect to the WT PhoQ and its interaction with PhoP. (B) We plot the positive predictive value, PPV=TP/(TP+FP), as a function of the N mutational variants ranked by ΔH_{TCS} from the most to least favorable for the first 200 mutants. Here, 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.

Most favorable TCS predictions limit "cross-talk"

Mutations that may enhance binding and phosphotransfer ability 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".

We first calculate the coevolutionary energy, H_{TCS}, between the WT PhoQ sequence and all of the non-hybrid RR proteins in *E. coli* (Fig. 3A). We find that for WT PhoQ, the most favorable H_{TCS} (most negative) is with its known signaling partner, PhoP. This result is consistent with previous computational predictions that used information-based quantities (*25, 26*) as a proxy for interaction specificity. Extending upon this idea, we exclude all mutational variants that exhibit "cross-talk" with a non-partner RR, i.e., we exclude all mutant-PhoQ variants that have a more favorable H_{TCS} with any other RR in *E. coli* other than its partner PhoP. Applying this 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 over the first 200 ranked mutational variants.

Construction of a coevolutionary energy to assess mutational effect on autophosphorylation

We further extend our analysis of the PhoQ mutants by examining whether a model based solely on HK (intraprotein) coevolution can identify the functional and non-functional mutational variants of PhoQ. We construct a coevolutionary energy of HK sequence selection, H_{HK} (Eq. 6) while focusing on the DHp and ATPase domains of the HK proteins (Fig. 1B), which form a binding interface during autophosphorylation (Fig. 1C). To test the fidelity of our inferred statistical model, we plot the top coevolving residue pairs in the HK using the metric Direct Information (DI) (*18, 20*) (Fig. S1). We find that the top coevolving pairs are contacts in the experimentally determined autophosphorylation state of the HK (*37*).

Applying ΔH_{HK} to the full set of the functional and non-functional mutational variants, we can generate a 2D histogram of ΔH_{HK} and ΔH_{TCS} (Fig. 4A). Note that H_{HK} (Eq. 5) and H_{TCS} (Eq. 6) are models of the intraprotein (HK only) and interprotein (HK/RR) coevolution, respectively. As in the case of ΔH_{TCS}, $\Delta H_{HK} = 0$ corresponds to the WT and $\Delta H_{HK} < 0$ corresponds to mutations that are more favorable than WT. Interestingly, the mutational variants as a function of ΔH_{HK} and ΔH_{TCS} are highly correlated with one another with Pearson correlations of 0.72 and 0.75 for the functional and non-functional mutants, respectively (Fig. 4A). This reflects the evolutionary constraint to satisfy both HK/RR phosphotransfer and HK autophosphorylation simultaneously, restricting the sequence-space of the HK to the same subset of amino acid residues on the binding region of the DHp domain. However, we find that ΔH_{HK} is not better at classifying the functional mutants from the non-functional than ΔH_{TCS} which is quantified by the plot of PPV versus the N top mutational variants ranked by ΔH_{HK} (Fig. S2).

Next, we examine the subset mutational variants that limit "cross-talk" (detailed in previous subsection) in a 2D histogram of ΔH_{HK} and ΔH_{TCS} (Fig. 4B). We find that while the remaining functional mutants are clustered around more favorable values of ΔH_{TCS}, they cover a

wide range of deleterious mutational changes in ΔH_{HK}. This would suggest that the HK is perhaps more tolerant of mutations that may reduce autophosphorylation but more sensitive to mutations that reduce interaction specificity for phosphotransfer or phosphatase activity. Interestingly, it was reported that the response regulator PhoP, is capable of receiving a phosphoryl group from acetyl-phosphate (*35*) and thus, mutations that minimally reduce PhoQ autophosphorylation activity but preserving interaction specificity with PhoP may still be identified as functional in experiment.

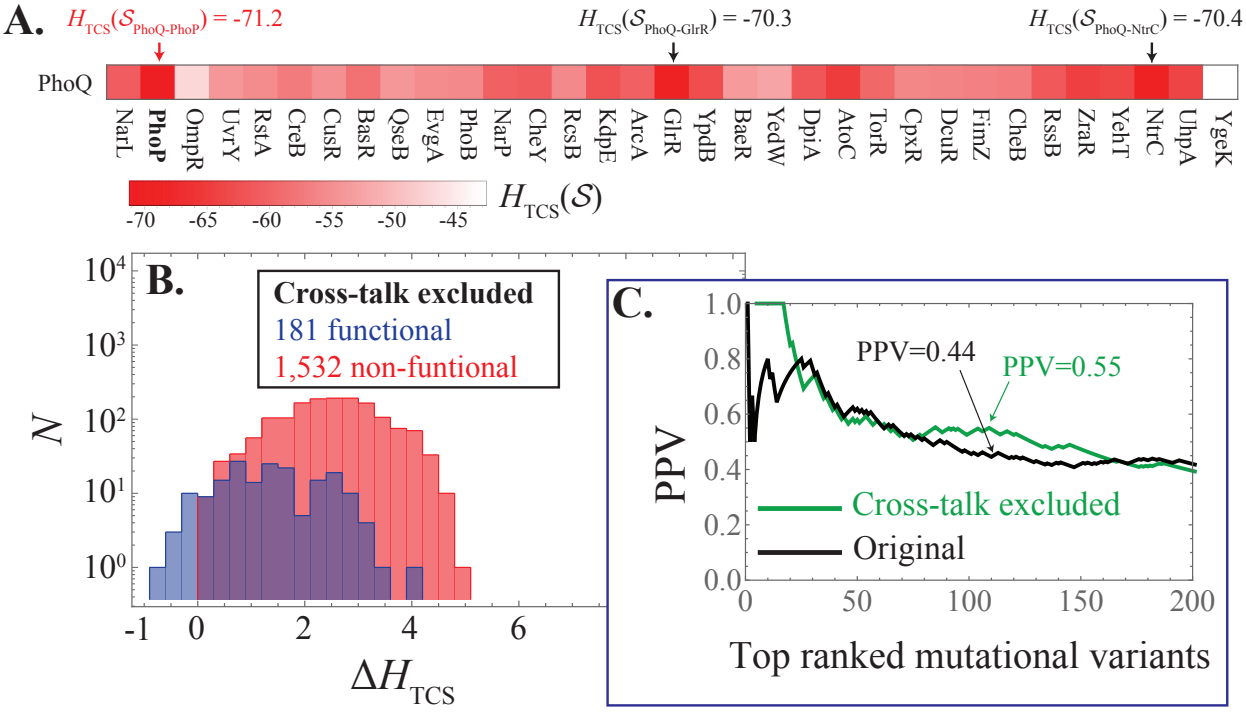

Fig. 3. Excluding mutational variants that are inferred to "cross-talk". (A) A grid plot showing the coevolutionary energy, H_{TCS}, (Eq. 5) computed for the WT PhoQ sequence with all of the non-hybrid RR protein sequences in *E. coli*, respectively. The most favorable energy (most negative) is between PhoQ and its partner PhoP. (B) For each of the 20^4 mutational variants of PhoQ, we remove all mutants for which H_{TCS} is not the most favorable between the PhoQ-mutant sequence and PhoP sequence. We plot the remaining 1,713 mutants (181 functional 1,532 non-functional) in a histogram as a function of the mutational change in our coevolutionary energy, 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 Figure 3B are plotted in green while the original PPV (Fig. 2B) is shown in black.

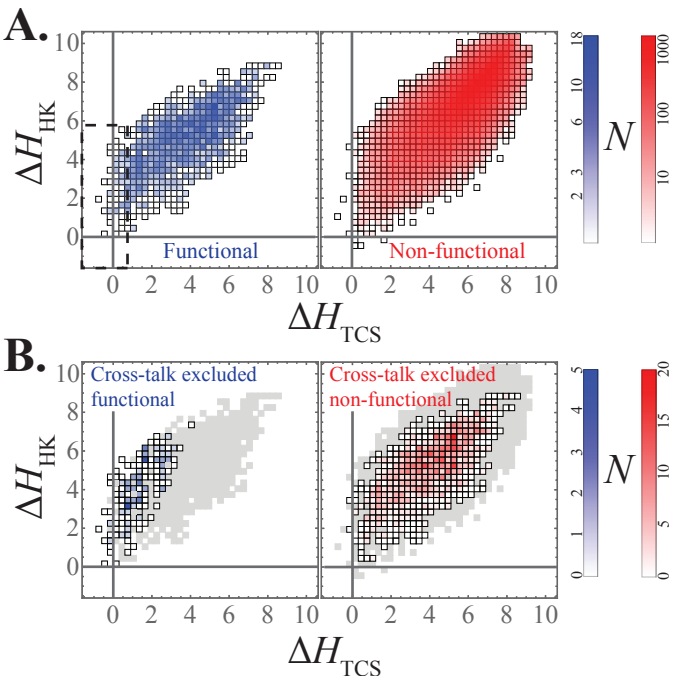

Fig. 4. Effect of mutation on PhoQ autophosphorylation. (A) Considering a 2D histogram of ΔH_{TCS} and ΔH_{HK} for the functional variants (blue) and non-functional variants (red). The two solid lines corresponding to $\Delta H_{TCS} = 0$ and $\Delta H_{HK} = 0$ represent the mutational change with respect to the WT PhoQ. The number of variants in each bin is shown on the logarithmic scale by the shade of blue or red, respectively. We observe an enrichment of functional mutants for lower values of both ΔH_{TCS} and ΔH_{HK}. (B) The 2D histogram is plotted for the subset of mutational variants that limit "cross-talk" by preserving the specificity of the PhoQ/PhoP interaction (labeled Cross-talk excluded). The gray squares denote the parts of the original 2D histogram (panel A) that have been removed.

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 between PhoQ and its partner PhoP, $\Delta\Delta G_{TCS}^{ZEMu}$. The calculation converged for 42,985 mutants (702 functional and 42,283 non-functional) from a randomly selected subset of the 20^4 variants. A histogram of $\Delta\Delta G_{TCS}^{ZEMu}$ is plotted for the 42,283 variants in Fig. 5A. A histogram of ΔH_{TCS} for the same subset of mutants is shown in Fig. S3A. On the population level, functional mutations exhibit a mean $\Delta\Delta G_{TCS}^{ZEMu}$ of 1.76 ± 0.06 kcal/mol lower than that of the non-functional mutants, indicating that destroying affinity disrupts TCS. The difference in means between the functional and non-functional mutational variants is statistically significant with a Wilcoxon rank-sum test p-value $< 2.2 \times 10^{-16}$. Furthermore, destabilizing mutations that are more than 2 standard deviations greater than the mean $\Delta\Delta G_{TCS}^{ZEMu}$ 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).

Confirming that mutations that significantly destabilize $\Delta\Delta G_{TCS}^{ZEMu}$ can be associated with a decrease in functionality, 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 $\Delta\Delta G_{TCS}^{ZEMu} < -5$ kcal/mol are non-functional (Fig. 5A), this has no statistical significance (p-value ~ 0.4). However, when we further examine the subset of mutants explored using ZEMu while excluding the mutants that exhibited "cross-talk" based on our coevolutionary energy, ΔH_{TCS}, in the previous subsection, we find that the remaining subset of mutants falls within roughly ±5 kcal/mol from the WT PhoQ/PhoP binding affinity (Fig. 5B). Selecting this subset of 363 mutational variants (92 functional and 272 non-functional) from the population of 42,985 variants is statistically significant with a p-value < 10^{-81} from a cumulative hypergeometric distribution for selecting 92 or more functional mutations in a random selection of 363 mutational variants from the population. A histogram of ΔH_{TCS} for this subset of mutational variants is shown for consistency in Fig. S3B. Fig. 5B shows that a combination of our ZEMu calculation and coevolutionary energy, ΔH_{TCS}, can provide support (albeit inconclusive) for the existence of an upper limit to the binding affinity for TCS proteins that would disrupt the transient nature of TCS signaling, since both the strong and weak binding mutants exhibited diminished signal transfer efficiency (i.e., ΔH_{TCS}) with PhoP respective of the other RR proteins in *E. coli*. However, further statistical analysis would be necessary to firmly establish such an upper limit for TCS.

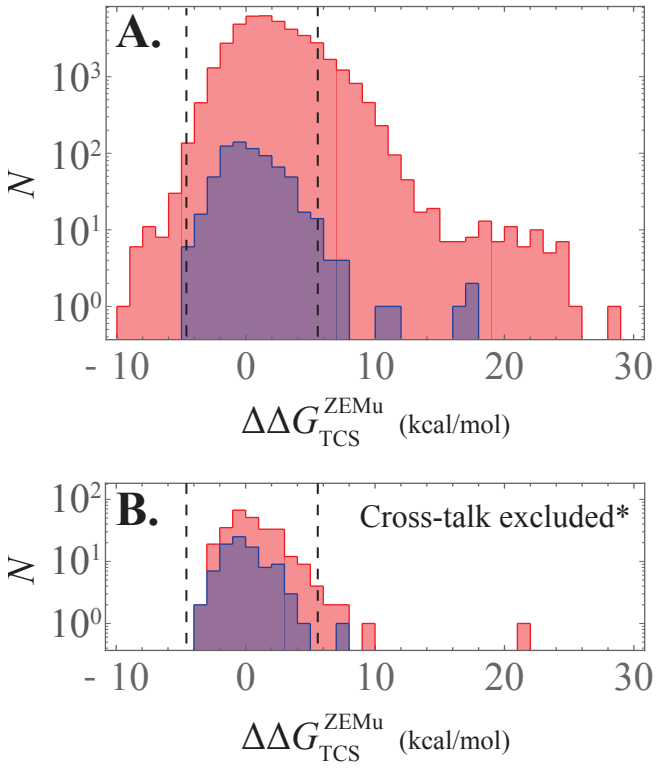

Figure 5. Mutational change in binding affinity for PhoQ/PhoP interaction. (A) A histogram of the mutational change in the binding affinity predicted by ZEMu(*36*), $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ (See Materials and Methods), is plotted for the 702 functional (blue) and 42,283 non-functional (red) mutational variants analyzed in our study. The dashed lines denote ±2 standard deviations from the mean of the functional (blue) distribution. (B) A histogram of $\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}$ is plotted for the subset of mutants in panel A that overlap with the subset of mutants that limited "cross-talk" based on the coevolutionary energy, ΔH_{TCS} (i.e., Cross-talk excluded subset in Fig. 3B). The resultant subset consists of 92 functional and 271 non-functional mutational variants, respectively, and is denoted with an asterisk in the label.

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 statistical energy function, H_{TCS}. This energy function captures the coevolving amino acid combinations that give rise to interaction specificity in TCS systems. In the past, we were able to predict with accuracy the deleterious effect of single point mutations on TCS phosphotransfer *in vitro* (*26*). Our present results show that the most favorable mutational variants predicted by our energy function accurately captures the functional mutational variants determined in the 4-position exhaustive mutational study of Podgornaia and Laub (*35*). Our key finding is that if we focus on the subset of mutational variants of PhoQ that preferentially interact with its partner, PhoP, based on our proxy for signal transfer efficiency H_{TCS}, we are better able to discriminate between functional and non-functional mutants. This highlights an intuitive but key design 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 beyond that of protein-protein interactions with undesired partners. While intuitive, we also demonstrate that mutations that significantly destabilize the binding affinity result in the loss of signaling. Further analysis would be necessary to establish a possible deleterious effect of over stabilizing the binding affinity between TCS partners, disrupting their transient interaction.

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 undesirable partners. Firstly, the construction of the coevolutionary energy described in our study relies on a large collection of sequence data having sufficiently sampled the sequence-space of TCS partners. 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. Secondly, this approach relies on the accurate inference of parameters for the energy landscape of the probability of sequence selection (Eq. 1) using DCA. The application of DCA is a crucial step in generating our statistical energy function because it is able to disentangle direct correlations between amino acids related to coevolution from indirect correlations (*21*).

While we were not able to predict the entire degenerate space (i.e., 1,659 functional

mutants) determined by Podgornaia and Laub (*35*) because the entire sequence-space has plausibly yet to be explored by evolutionary processes (*16, 38*), our method can accurately predict functional mutations amongst our most favorable predictions. These predictions can readily be used to engineer novel protein-protein interactions in TCS systems that could serve as another strategy amongst the already existing methods to match novel inputs with outputs via modular engineering (*39-42*). Further, the predictive power of our approach will systematically improve as more sequences of TCS partners are collected.

Another advantage of the statistical methods described here for identifying mutational phenotypes 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 and those of other members of the field (*43-45*) 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 (*46*). Further investigations of other systems that are evolutionarily constrained to maintain protein-protein interactions could elucidate the extent at which coevolutionary 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 (*47*) elucidating the determinants of interaction specificity.

Materials and Methods

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

We first obtain multiple sequence alignments (MSA) from Pfam (*48*) (Version 28), focusing on the DHp (PF00512) and REC (PF00072) domains of the HK and RR, respectively (Fig. 1A). The only notable change from the default MSAs of PF00512 is that the first 4 positions (columns) of the MSA were removed due to poor coverage of the PhoQ sequence such that each DHp MSA had 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). All hybrid TCS proteins were excluded from our study due to the relaxed specificity between HK and RR domains in a hybrid protein (*26*). Further, HKs with multiple HAMP or ATPase domains were also excluded. 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 (*49, 50*), i.e., HK and RR are inferred to be partners if their respective ordered locus numbers differ by exactly 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 (External Databases S1) consisted of 6,519 non-redundant concatenated sequences.

Sequence database for HK intra-protein interactions: DHp and ATPase

We searched the Representative Proteomes (RP55) (*51*) database for HK proteins using the Jackhmmer algorithm on the HMMER web server (*52*) with default parameters. Once again, we restricted our curated set of HK proteins to ones having the PhoQ domain architecture. We restricted our MSA to a length (number of columns) of $L_{\mathrm{HK}} = 222$ by only including the DHp and ATPase domains. Our remaining dataset (External Databases S2) consisted of 4,483 non-redundant HK sequences of the form: $S_{\mathrm{HK}} = (A_1, A_2, ..., A_{L-1}, A_L)$ with a total length $L = L_{\mathrm{HK}} = 222$.

Inference of parameters of coevolutionary model

The collection of protein sequences for a protein family or coevolved families can be viewed as being selected from a Boltzmann equilibrium distribution, i.e., $P(s) = Z^{-1}\exp(-\beta H(s))$, where $s = (A_1, A_2, ..., A_L)$ is a sequence, $\beta = (k_B T_{sel})^{-1}$ is the inverse of the evolutionary selection temperature (*17*), and $H(s)$ is an appropriate energetic function in units of $k_B T_{sel}$. However, more appropriate for our interests is 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 (See Review: (*53*)) to infer a least biased model that is consistent with the input sequence data (*18, 20*), 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: (*19*) for full computational details).

Inference problems of this nature exhibit a known gauge freedom associated with being able to add energy to the inferred fields that can be subtracted from the couplings to maintain a constant *H*. We fix the gauge by adopting the Ising condition, which can be obtained for a Potts model in any particular gauge choice, $\{\hat{\mathbf{J}}, \hat{\mathbf{h}}\}$, using the transformation:

$$\begin{aligned} J_{ij}(a,b) &= \hat{J}_{ij}(a,b) - \hat{J}_{ij}(:,b) - \hat{J}_{ij}(a,:) + \hat{J}_{ij}(:,:) \\ h_i(a) &= \hat{h}_i(a) - \hat{h}_i(:) + \sum_{\substack{j=1 \\ j\neq i}}^{L}\left(\hat{J}_{ij}(a,:) - \hat{J}_{ij}(:,:)\right) \end{aligned} \tag{4}$$

where the colon symbol (:) denotes an average over all amino acids identities at its respective position, i.e., $\hat{J}_{ij}(:,b)$ is the average statistical coupling between positions *i* and *j* where position *j* has amino acid *b* and the average is taken over all possible amino acid combinations at position *i*. The Ising condition has the following property: $\sum_{a=1}^{q} J_{ij}(a,b) = \sum_{b=1}^{q} J_{ij}(a,b) = \sum_{a=1}^{q} h_i(a) = 0$, where $q = 21$ represents the 20 amino acids and MSA gap. This gauge condition ensures that the ensemble of random sequences has a mean energy of 0.

Previous studies have used DCA to identify highly coevolving pairs of residues to predict the native state conformation of a protein (*54-56*) as well as identify additional functionally relevant conformational states (*55, 57, 58*) and multi-meric states (*18, 20, 22, 24, 58*). The Potts model (Eq. 3) obtained from DCA has been related to the theory of evolutionary sequence selection (*17*) as well as mutational changes in protein stability (*17, 59, 60*). Additional work has applied DCA to protein folding to predict the effect of point mutations on the folding rate (*61*) as well as construct a statistical potential for native contacts in a structure-based model of a protein (*62*) to better capture the transition state ensemble. Finally, DCA has been used to identify relevant protein-protein interactions in biological interaction networks (*25, 63*) as well as identify high fitness variants for a number of proteins by relating mutational changes in stability to organism fitness (*44, 45*) or viral fitness (*43*).

Mutational changes in coevolutionary energy

For the concatenated sequences of HK (DHp) and RR (REC) (Fig. 1A), we infer an energy function in the form of a Potts model (Equation 3). To focus on quantifying how mutations to the DHp domain of the HK affect its binding and phosphotransfer with the REC domain of the RR, 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 that are in close proximity in a representative structure of the TCS complex (Fig. 1C). All local fields terms, h_i, are included to partially capture the affect of mutations on domain stability. These considerations allow us to construct our energy function, H_{TCS}:

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

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 (*17*), and r_{ij} is the minimum distance between residues *i* and *j* in the representative structure. Mutational changes in the energy are then computed using Equation 2, i.e., $\Delta H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) = H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{mutant}}) - H_{\mathrm{TCS}}(S_{\mathrm{TCS}}^{\mathrm{WT}})$.

Likewise, for the HK model (Fig. 1B) we infer an energy function with the form of Equation 3 and focus on the statistical couplings between positions in the HK that are in close proximity in a representative structure of the HK autophosphorylation state (Fig. 1D):

$$H_{\mathrm{HK}}(S_{\mathrm{HK}}) = -\sum_{i=1}^{L_{\mathrm{HK}}-1} \sum_{j=i+1}^{L_{\mathrm{HK}}} J_{ij}(A_i, A_j) \times \Theta(c - r_{ij}) - \sum_{i=1}^{L_{\mathrm{HK}}} h_i(A_i) \qquad (6)$$

where Θ is a Heaviside step function, c is the a cutoff distance of 16Å, and r_{ij} is the minimum distance between residues i and j in the representative structure. The mutational changes in energy are then computed using $\Delta H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{mutant}}) = H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{mutant}}) - H_{\mathrm{HK}}(S_{\mathrm{HK}}^{\mathrm{WT}})$.

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 (*36*), which is followed by a mutational change in stability using FoldX (*64*). ZEMu has been used to explain the mechanism of Parkinson's disease associated mutations in Parkin (*65, 66*). The minimization is done in MacroMoleculeBuilder (MMB), a general-purpose internal coordinate mechanics code also known for RNA folding (*67*), homology modeling (*68*), morphing (*69*), and fitting to density maps (*70*).

We use the Zone Equilibration of Mutants (ZEMu) (*36*) 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. (*36*) for full computational details) to equilibrate the local region around mutational sites. ZEMu can then estimate the binding affinity between the mutant-PhoQ/PhoP, $\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{mutant})$, and the WT-PhoQ/PhoP, $\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{WT})$, using the knowledge-based FoldX (*64*) potential. This allows for the calculation of the mutational change in binding affinity as:
$\Delta\Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}} = \Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{mutant}) - \Delta G_{\mathrm{TCS}}^{\mathrm{ZEMu}}(\mathrm{WT})$.

ZEMu calculation was performed according to Ref: (*36*), 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 (*36*) the limit was 48 hours. This meant that of 122802 mutants attempted, 42923 completed within the time limit, whereas in (*36*), 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 (*71*) 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 (*36*) 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.

Supplementary Materials

Supplementary Text

Fig. S1. High coevolving pairs of residues in HK correspond to contacts in autophosphorylation state.

Fig. S2. Effect of mutation on the PhoQ autophosphorylation: 1D histogram.

Fig. S3. Histogram of mutational change in coevolutionary energy, ΔH_{TCS}, for subset of mutational variants explored by ZEMu calculation.

External Database S1. Collection of partnered sequences of DHp and REC.

External Database S2. Collection of HAMP-containing HK sequences.

Acknowledgments: We would like to thank Dr. 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 and Notes

1. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, P. G. Wolynes, Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis. *Proteins-Structure Function and Genetics* **21**, 167-195 (1995).
2. J. D. Bryngelson, P. G. Wolynes, Spin-Glasses and the Statistical-Mechanics of Protein Folding. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 7524-7528 (1987).
3. J. N. Onuchic, Z. LutheySchulten, P. G. Wolynes, Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem* **48**, 545-600 (1997).
4. P. E. Leopold, M. Montal, J. N. Onuchic, Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 8721-8725 (1992).
5. U. Gobel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309-317 (1994).
6. I. N. Shindyalov, N. A. Kolchanov, C. Sander, Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng* **7**, 349-358 (1994).
7. E. Neher, How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America* **91**, 98-102 (1994).
8. D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration in biomolecules. *Q Rev Biophys* **47**, 285-363 (2014).
9. P. G. Wolynes, Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **119**, 218-230 (2015).
10. T. Sikosek, H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface* **11**, 20140419 (2014).
11. A. M. Stock, V. L. Robinson, P. N. Goudreau, Two-component signal transduction. *Annual Review of Biochemistry* **69**, 183-215 (2000).
12. J. A. Hoch, Two-component and phosphorelay signal transduction. *Current Opinion in Microbiology* **3**, 165-170 (2000).
13. P. Casino, V. Rubio, A. Marina, The mechanism of signal transduction by two-component systems. *Current Opinion in Structural Biology* **20**, 763-771 (2010).
14. H. Szurmant, J. A. Hoch, Interaction fidelity in two-component signaling. *Current Opinion in Microbiology* **13**, 190-197 (2010).
15. M. T. Laub, M. Goulian, Specificity in Two-Component Signal Transduction Pathways. *Annual Review of Genetics* **41**, 121-145 (2007).
16. E. J. Capra, M. T. Laub, Evolution of two-component signal transduction systems. *Annu Rev Microbiol* **66**, 325-347 (2012).
17. F. Morcos, N. P. Schafer, R. R. Cheng, J. N. Onuchic, P. G. Wolynes, Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 12408-12413 (2014).
18. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures

native contacts across many protein families. *Proceedings of the National Academy of Sciences* **108**, E1293-E1301 (2011).

19. M. Ekeberg, C. Lovkvist, Y. H. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys Rev E* **87**, (2013).
20. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 67-72 (2009).
21. D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein co-evolution. *Nature reviews. Genetics* **14**, 249-261 (2013).
22. A. Schug, M. Weigt, J. N. Onuchic, T. Hwa, H. Szurmant, High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences* **106**, 22124-22129 (2009).
23. A. E. Dago, A. Schug, A. Procaccini, J. A. Hoch, M. Weigt, H. Szurmant, Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences* **109**, E1733-E1742 (2012).
24. R. N. dos Santos, F. Morcos, B. Jana, A. D. Andricopulo, J. N. Onuchic, Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci Rep-Uk* **5**, (2015).
25. A. Procaccini, B. Lunt, H. Szurmant, T. Hwa, M. Weigt, Dissecting the Specificity of Protein-Protein Interaction in Bacterial Two-Component Signaling: Orphans and Crosstalks. *PloS one* **6**, e19729 (2011).
26. R. R. Cheng, F. Morcos, H. Levine, J. N. Onuchic, Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proceedings of the National Academy of Sciences*, (2014).
27. L. Li, E. I. Shakhnovich, L. A. Mirny, Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proceedings of the National Academy of Sciences* **100**, 4463-4468 (2003).
28. L. Burger, E. van Nimwegen, Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4**, (2008).
29. A. I. Podgornaia, M. T. Laub, Determinants of specificity in two-component signal transduction. *Current Opinion in Microbiology* **16**, 156-162 (2013).
30. P. Casino, V. Rubio, A. Marina, Structural Insight into Partner Specificity and Phosphoryl Transfer in Two-Component Signal Transduction. *Cell* **139**, 325-336 (2009).
31. E. J. Capra, B. S. Perchuk, E. A. Lubin, O. Ashenberg, J. M. Skerker, M. T. Laub, Systematic Dissection and Trajectory-Scanning Mutagenesis of the Molecular Interface That Ensures Specificity of Two-Component Signaling Pathways. *PLoS Genetics* **6**, (2010).
32. Y.-L. Tzeng, J. A. Hoch, 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 (1997).
33. L. Qin, S. Cai, Y. Zhu, M. Inouye, Cysteine-Scanning Analysis of the Dimerization Domain of EnvZ, an Osmosensing Histidine Kinase. *Journal of Bacteriology* **185**, 3429-3435 (2003).

34. J. M. Skerker, B. S. Perchuk, A. Siryaporn, E. A. Lubin, O. Ashenberg, M. Goulian, M. T. Laub, Rewiring the Specificity of Two-Component Signal Transduction Systems. *Cell* **133**, 1043-1054 (2008).
35. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673-677 (2015).
36. D. F. A. R. Dourado, S. C. Flores, A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins-Structure Function and Bioinformatics* **82**, 2681-2690 (2014).
37. A. E. Mechaly, N. Sassoon, J. M. Betton, P. M. Alzari, Segmental Helical Motions and Dynamical Asymmetry Modulate Histidine Kinase Autophosphorylation. *Plos Biol* **12**, (2014).
38. J. Echave, S. J. Spielman, C. O. Wilke, Causes of evolutionary rate variation among protein sites. *Nature reviews. Genetics* **17**, 109-121 (2016).
39. S. R. Schmid, R. U. Sheth, A. Wu, J. J. Tabor, Refactoring and Optimization of Light-Switchable Escherichia coli Two-Component Systems. *Acs Synth Biol* **3**, 820-831 (2014).
40. W. R. Whitaker, S. A. Davis, A. P. Arkin, J. E. Dueber, Engineering robust control of two-component system phosphotransfer using modular scaffolds. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 18090-18095 (2012).
41. I. Ganesh, S. Ravikumar, S. H. Lee, S. J. Park, S. H. Hong, Engineered fumarate sensing Escherichia coli based on novel chimeric two-component system. *J Biotechnol* **168**, 560-566 (2013).
42. J. J. Tabor, A. Levskaya, C. A. Voigt, Multichromatic control of gene expression in Escherichia coli. *J Mol Biol* **405**, 315-324 (2011).
43. A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, A. K. Chakraborty, Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38**, 606-617 (2013).
44. M. Figliuzzi, H. Jacquier, A. Schug, O. Tenaillon, M. Weigt, Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution* **33**, 268-280 (2016).
45. T. A. I. Hopf, John B.; Poelwijk, Frank J.; Springer, Michael; Sander, Chris; Marks, Debora S., Quantification of the effect of mutations using a global probability model of natural sequence variation. *arXiv*, (2015).
46. A. Zarrinpar, S.-H. Park, W. A. Lim, Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676-680 (2003).
47. C. D. Aakre, J. Herrou, T. N. Phung, B. S. Perchuk, S. Crosson, M. T. Laub, Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **163**, 594-606 (2015).
48. R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database. *Nucleic Acids Research* **42**, D222-D230 (2014).
49. K. Yamamoto, K. Hirao, T. Oshima, H. Aiba, R. Utsumi, A. Ishihama, Functional characterization in vitro of all two-component signal transduction systems from Escherichia coli. *Journal of Biological Chemistry* **280**, 1448-1456 (2005).
50. J. M. Skerker, M. S. Prasol, B. S. Perchuk, E. G. Biondi, M. T. Laub, Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: A system-level analysis. *Plos Biol* **3**, 1770-1788 (2005).

51. C. M. Chen, D. A. Natale, R. D. Finn, H. Z. Huang, J. Zhang, C. H. Wu, R. Mazumder, Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation. *PloS one* **6**, (2011).
52. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29-W37 (2011).
53. S. Presse, K. Ghosh, J. Lee, K. A. Dill, Principles of maximum entropy and maximum caliber in statistical physics. *Rev Mod Phys* **85**, 1115-1141 (2013).
54. J. I. Sulkowska, F. Morcos, M. Weigt, T. Hwa, J. N. Onuchic, Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences* **109**, 10340-10345 (2012).
55. L. Sutto, S. Marsili, A. Valencia, F. L. Gervasio, From residue coevolution to protein conformational ensembles and functional dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 13567-13572 (2015).
56. D. S. Marks, T. A. Hopf, C. Sander, Protein structure prediction from sequence variation. *Nature biotechnology* **30**, 1072-1080 (2012).
57. F. Morcos, B. Jana, T. Hwa, J. N. Onuchic, Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, (2013).
58. D. Malinverni, S. Marsili, A. Barducci, P. De Los Rios, Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. *PLoS computational biology* **11**, (2015).
59. S. Lui, G. Tiana, The network of stabilizing contacts in proteins studied by coevolutionary data. *J Chem Phys* **139**, 155103 (2013).
60. A. Contini, G. Tiana, A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J Chem Phys* **143**, 025103 (2015).
61. S. Mallik, S. Das, S. Kundu, Predicting protein folding rate change upon point mutation using residue-level coevolutionary information. *Proteins-Structure Function and Bioinformatics* **84**, 3-8 (2016).
62. R. R. Cheng, M. Raghunathan, J. K. Noel, J. N. Onuchic, Constructing sequence-dependent protein models using coevolutionary information. *Protein Sci* **25**, 111-122 (2016).
63. C. Feinauer, H. Szurmant, M. Weigt, A. Pagnani, Inter-Protein Sequence Co-Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp Operon. *PloS one* **11**, e0149166 (2016).
64. R. Guerois, J. E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *Journal of Molecular Biology* **320**, 369-387 (2002).
65. T. R. Caulfield, F. C. Fiesel, E. L. Moussaud-Lamodiere, D. F. A. R. Dourado, S. C. Flores, W. Springer, Phosphorylation by PINK1 Releases the UBL Domain and Initializes the Conformational Opening of the E3 Ubiquitin Ligase Parkin. *PLoS computational biology* **10**, (2014).
66. F. C. Fiesel, T. R. Caulfield, E. L. Moussaud-Lamodiere, K. Ogaki, D. F. A. R. Dourado, S. C. Flores, O. A. Ross, W. Springer, Structural and Functional Impact of Parkinson Disease-Associated Mutations in the E3 Ubiquitin Ligase Parkin. *Hum Mutat* **36**, 774-786 (2015).
67. S. C. Flores, R. B. Altman, Turning limited experimental information into 3D models of RNA. *Rna* **16**, 1769-1778 (2010).

68. S. C. Flores, Y. Wan, R. Russell, R. B. Altman, Predicting RNA structure by multiple template homology modeling. *Pac Symp Biocomput*, 216-227 (2010).
69. A. Tek, A. A. Korostelev, S. C. Flores, MMB-GUI: a fast morphing method demonstrates a possible ribosomal tRNA translocation trajectory. *Nucleic Acids Res* **44**, 95-105 (2016).
70. S. C. Flores, Fast fitting to low resolution density maps: elucidating large-scale motions of the ribosome. *Nucleic Acids Res* **42**, e9 (2014).
71. S. C. Flores, M. A. Sherman, C. M. Bruns, P. Eastman, R. B. Altman, Fast Flexible Modeling of RNA Structure Using Internal Coordinates. *Ieee Acm T Comput Bi* **8**, 1247-1257 (2011).