Coevolutionary landscapes of interaction specificity in two-component signaling

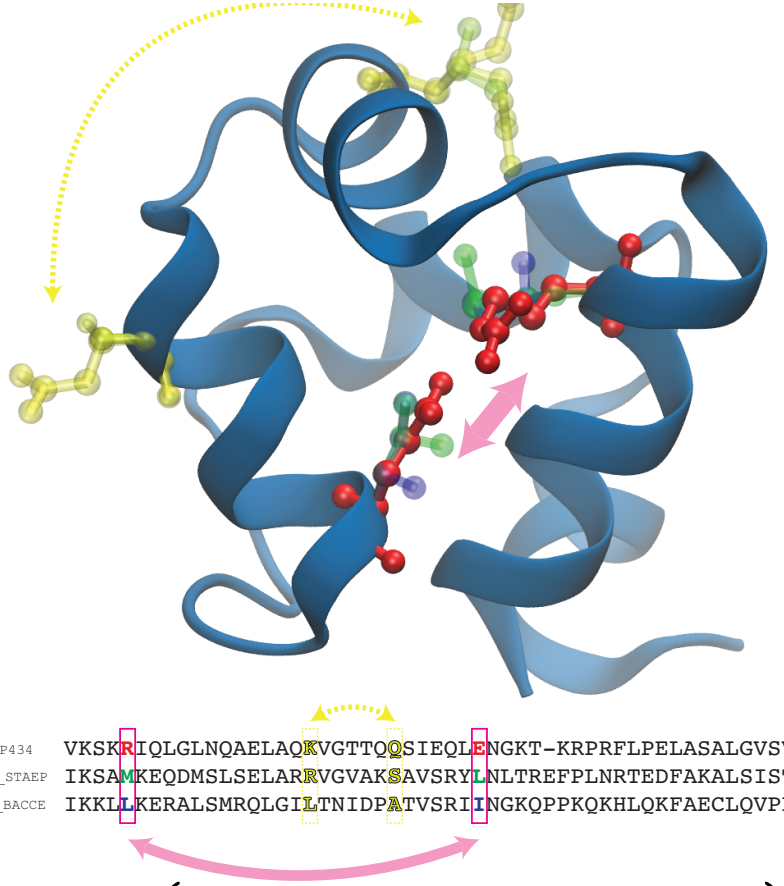
Faruck Morcos
faruckm@utdallas.edu

morcoslab.org

Department of Biological Sciences
Center for Systems Biology
University of Texas at Dallas

Coevolution in proteins and RNA, theory and experiments
Cargèse, France. April, 2016

Coevolutionary landscapes from ***Direct Coupling Analysis*** *(DCA)*

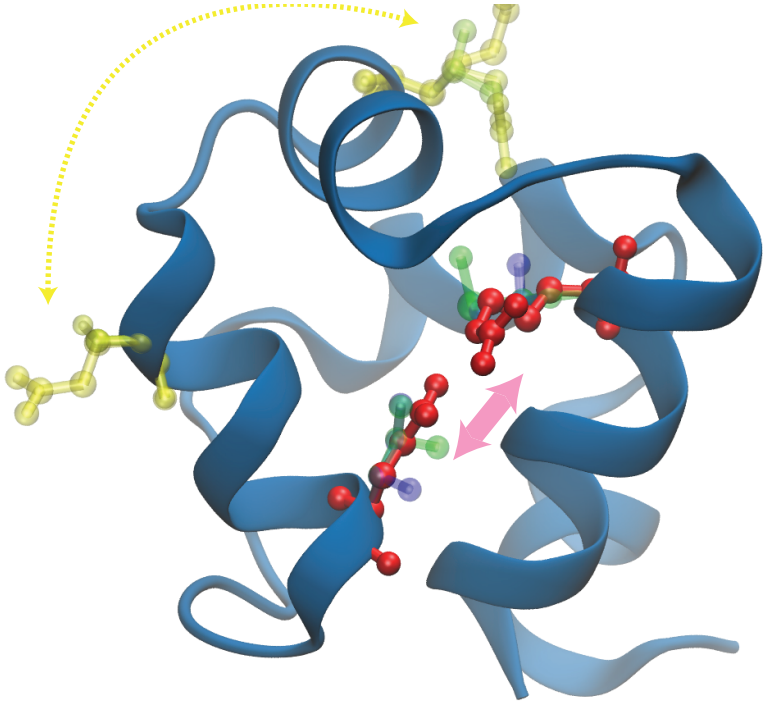


$$P_{DCA}(\{a_i,\ldots,a_L\}) = \frac{1}{Z}\exp\left\{\sum_{i<j} e_{ij}(a_i,a_j) + \sum_i h_i(a_i)\right\} = \frac{1}{Z}\exp\{H_{DCA}\}$$

Weigt et al. PNAS 2009, Morcos, Pagnani et al. PNAS 2011,
Sulkowska et al. PNAS 2012, Eckeberg et al. Phys. Rev. E. 2013

Related work by:
Aurell, Jones, Marks, Taylor, Kamichetty, Baker, etc.

Coevolutionary landscapes from ***Direct Coupling Analysis*** *(DCA)*



$$P(a_i, \ldots, a_L) = \frac{1}{Z} \exp\left(-\frac{\Delta E}{T_{sel} k_B}\right)$$

Ramanathan & Shakhnovich. Phys Rev E, 1994
Saven & Wolynes. J Phys Chem B, 1997

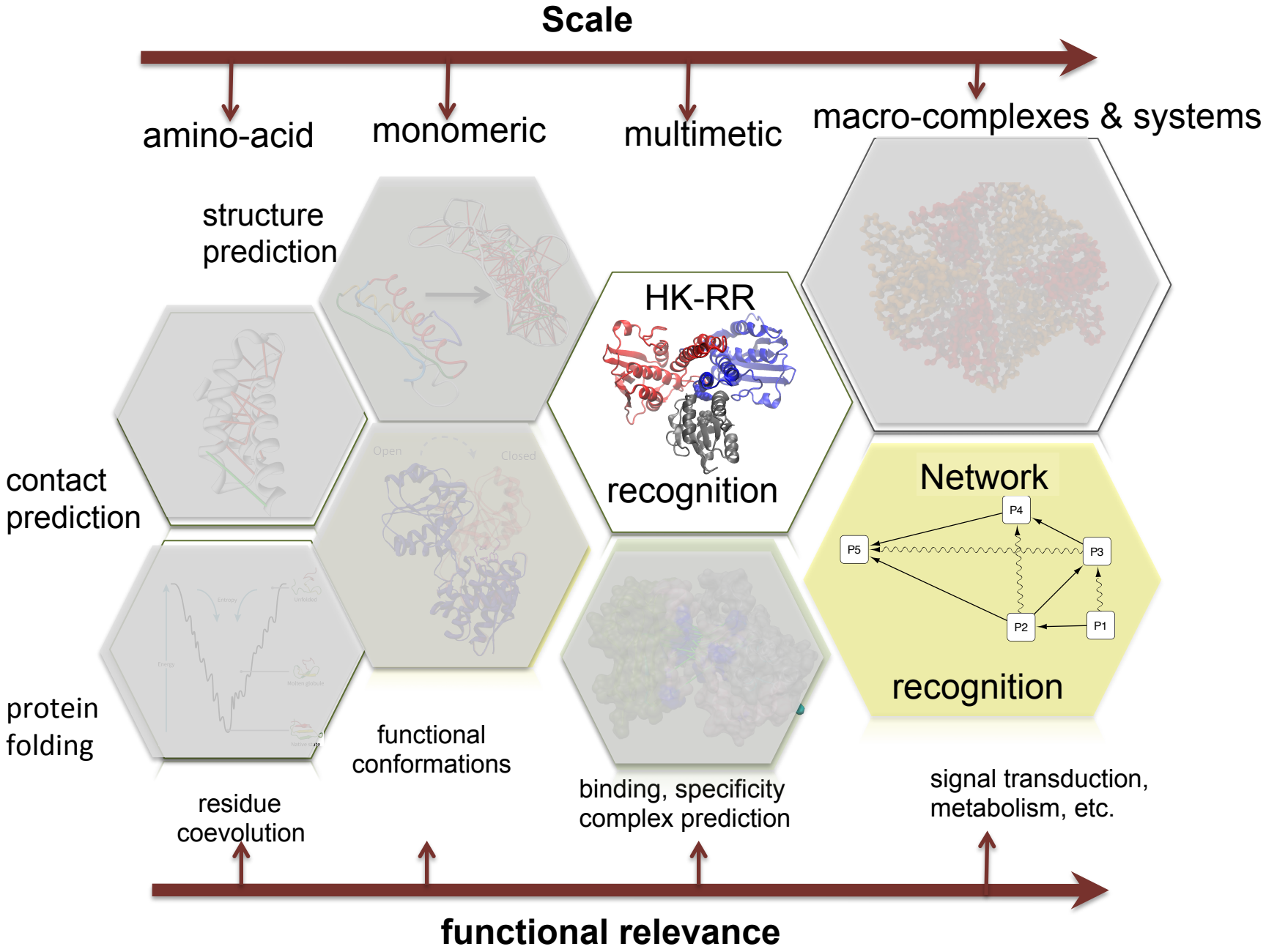
ΔE Energy gap between a folded configuration and compact misfolded configurations

T_{sel} is the apparent temperature at which sequences were selected by evolution for a particular gtein family or fold

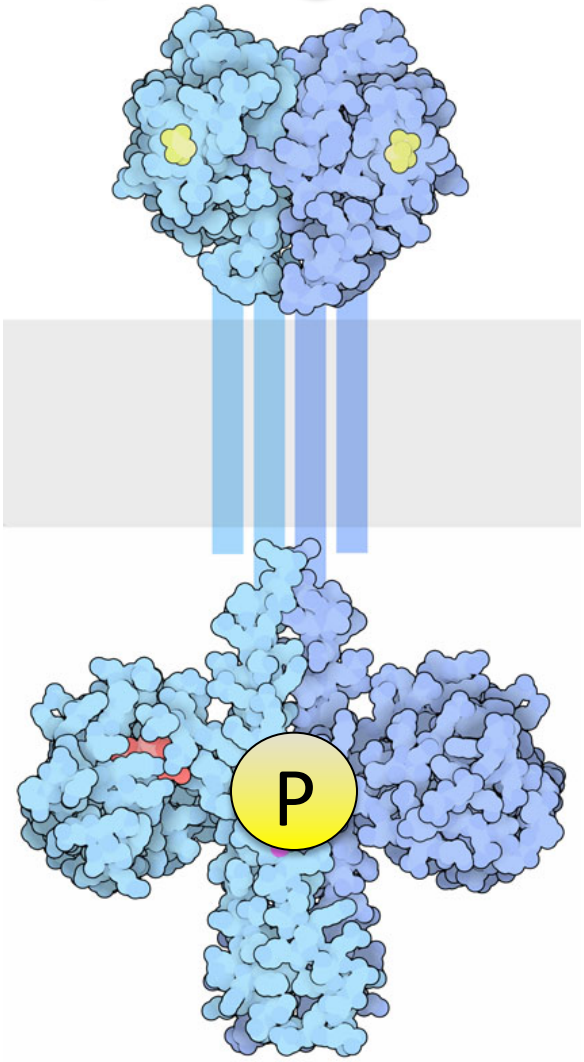
$$T_{sel} = \frac{-H^{nat}_{AWSEM} + H^{mg}_{AWSEM}}{k_B \log\left(P^{nat}_{DCA}/P^{mg}_{DCA}\right)} = \frac{-H^{nat}_{AWSEM} + H^{mg}_{AWSEM}}{k_B \left(H^{nat}_{DCA} - H^{mg}_{DCA}\right)}$$

Morcos et al. PNAS , 2014

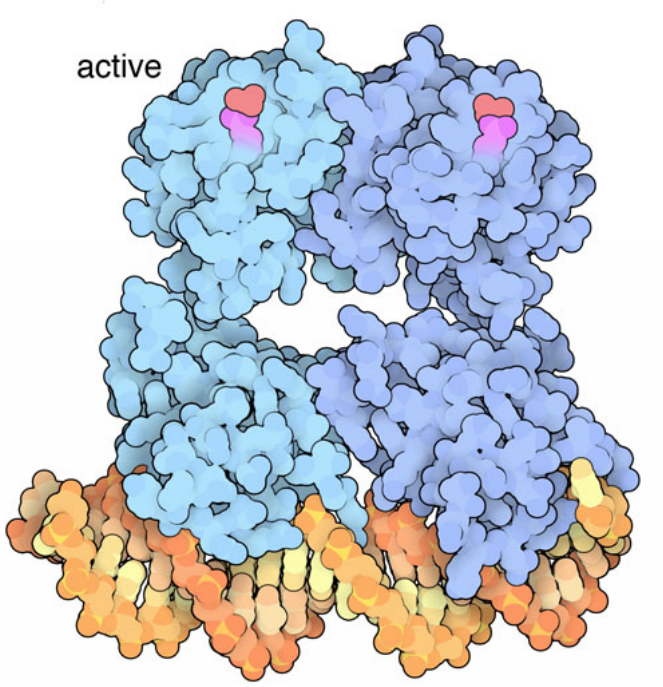
Coevolutionary analysis research has a broad scope



Two Component Systems



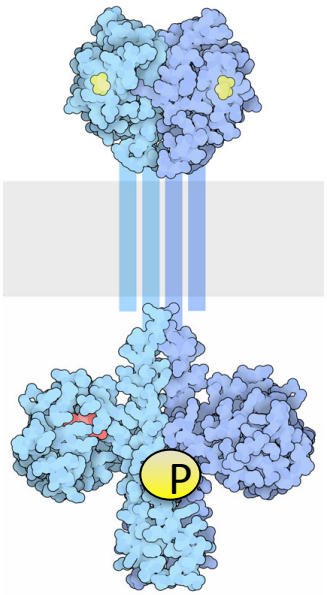
Histidine Kinase

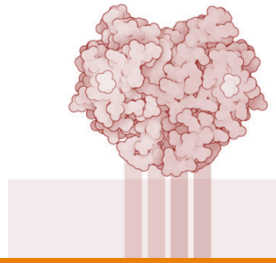


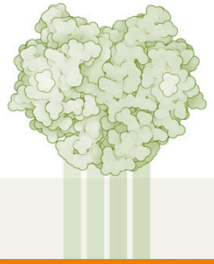
Response regulator

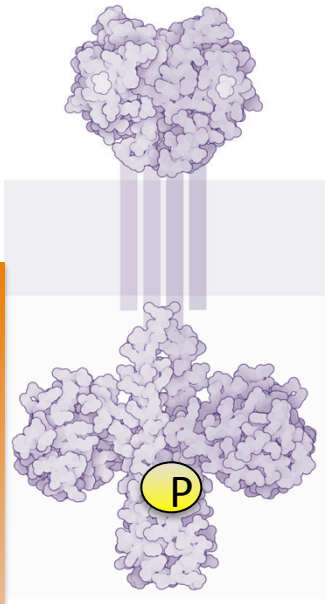
Figures by David Goodsell (PDB.org)

Organisms use multiple TCS to respond to signals



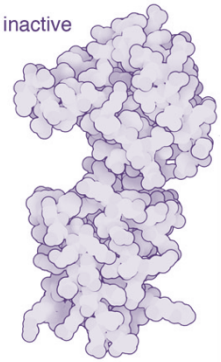


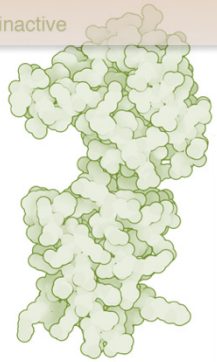


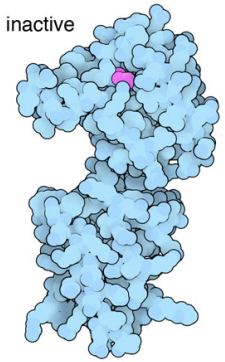


10^2-10^3 TCS partners in bacteria

How does a TCS protein stay faithful to its signaling partner? (specificity encoded through coevolved HK/RR interface)

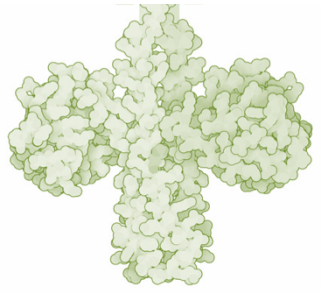






Source: David Goodsell (PDB.org)

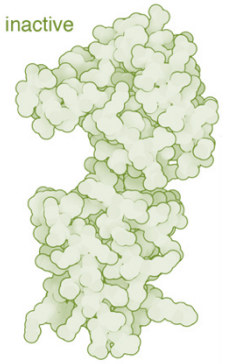
Building a coevolutionary model of TCS signaling



Histidine Kinase (HK)

Response Regulator (RR)

```
AAKAPSARGHATKPRAPKDAQHEAA
AAKAPSARGHATKPRAPKDAQHEAA
SAKEKNEKMKIVKN-LIDKGKKSGS
TELETKFTLDQVKDQLEEQGKKRSS
LAPSGNTALATAKKKKITDRTDDPV
TELETKFTLDQVKPRAEKDGKKRSS
```



HK RR

$$\text{sequence} = (A_1, A_2, \ldots, A_{N_{HK}}, A_{N_{HK}+1}, \ldots, A_{N_{HK}+N_{RR}})$$

$$\mathfrak{A}_{DCA}(\text{sequence}) = -\sum_{i<j} J_{ij}(A_i, A_j) - \sum_i h_i(A_i)$$

Key assumption: Chromosomal adjacency is used as a proxy for TCS signaling partners

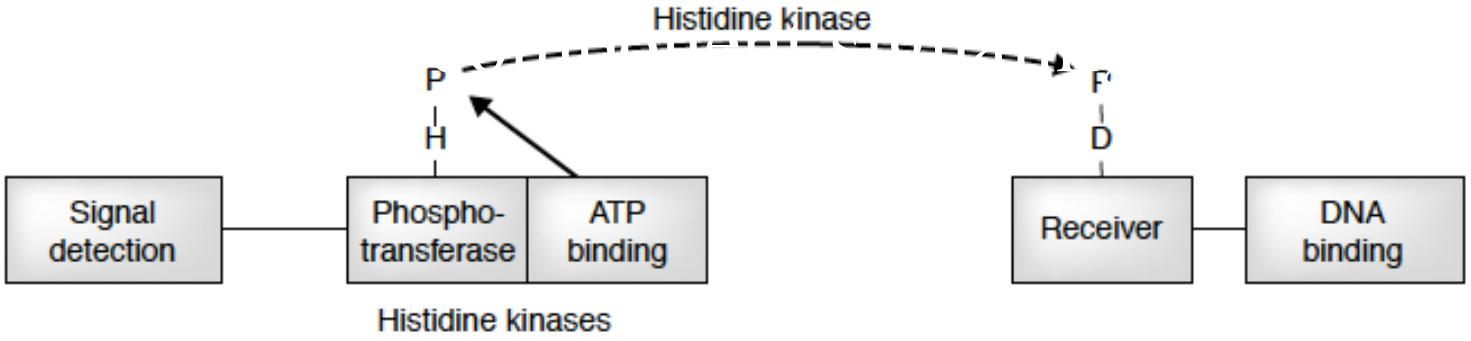
$$P(\text{sequence}) = \exp(-\mathfrak{A}_{DCA}) / Z$$

Statistical coevolutionary couplings inferred by DCA

1-body fields (related to amino acid composition at site)

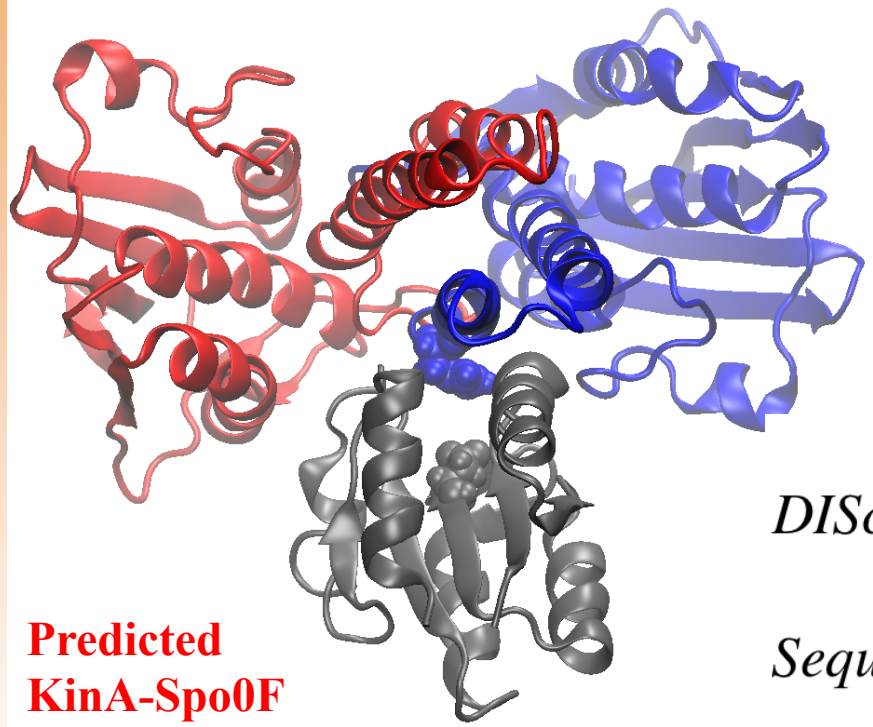
Early work: Li et al. PNAS 2003, White et al. Methods Enzymology 2007, Skerker et al. Cell 2008, Schug et al. PNAS, 2009, Procaccini et el. PLoS One, 2011, Dago et al. PNAS, 2012

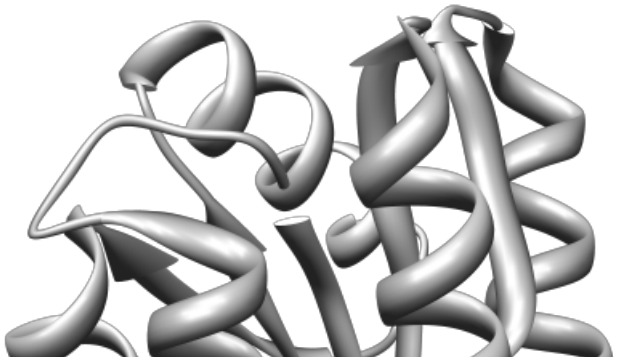
An early complex estimate for TCS



Histidine Kinase (HK)

Response Regulator (RR)





$$DIScore = \sum_{i \in HK, j \in RR} P_{ij}^{(dir)}(S_i, R_j) \ln\left(\frac{P_{ij}^{(dir)}(S_i, R_j)}{f_i(S_i) f_j(R_j)}\right)$$

$$Sequence = (S_1, \ldots, S_{N_{HK}}, R_{N_{HK}+1}, \ldots, R_{N_{HK}+N_{RR}})$$

Predicted KinA-Spo0F complex RMSD 2.5A

TM0853
(Thermotoga maritima)

Spo0F (Bacillus Subtilis)

Schug et al. PNAS 2009
Cheng et al. PNAS 2014

No cognate assumption (i.e., scramble)

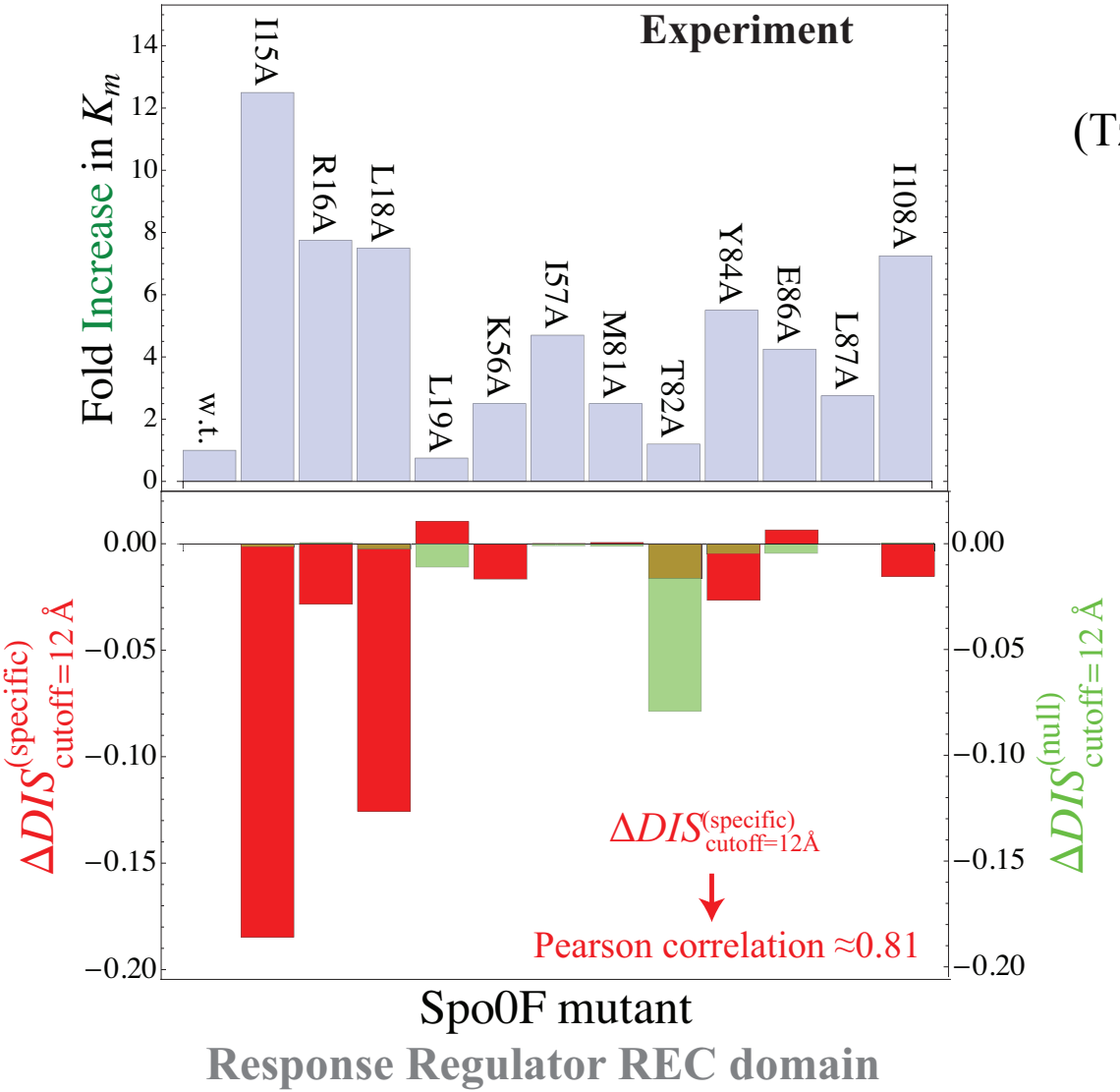
$$DIS^{(specific)} = DIS - DIS^{(null)}$$

Cognate assumption

Determinants of interaction specificity amongst cognate pairs

Contains generic, conserved features of HK/ RR pairs

$$DIS^{(specific)} = DIS - DIS^{(null)}$$



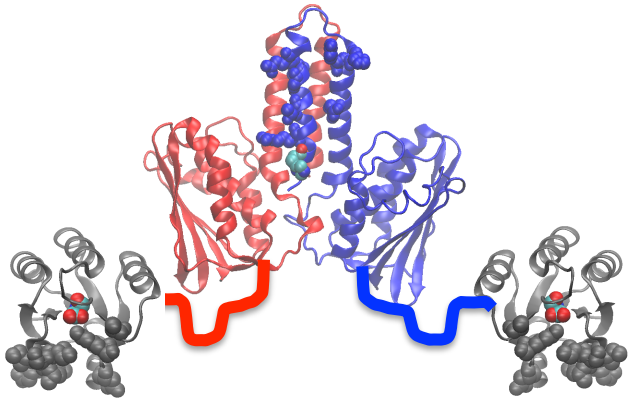
(Tzeng and Hoch, JMB 1997)

"Specific" model better predictor of binding/unbinding

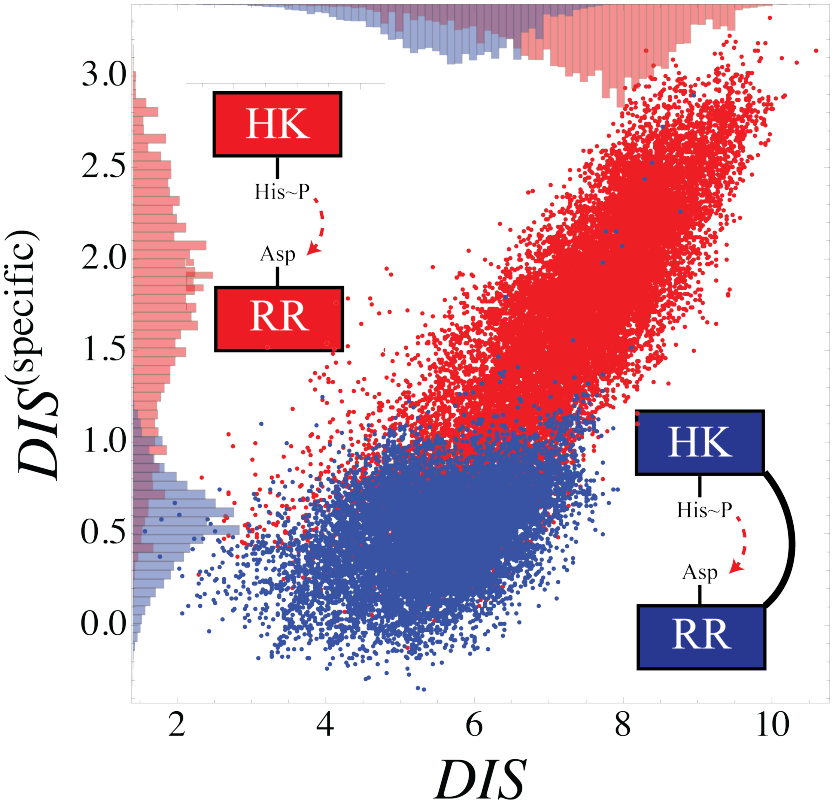
$$\frac{d[P]}{dt} = \frac{V_{max}[S]}{K_m}$$

Cheng et al. PNAS, 2014

DCA-based metric discerns regular from hybrid TCS systems



Cartoon depiction of a hybrid TCS protein



Analysis on ~17,000 Hybrid TCS

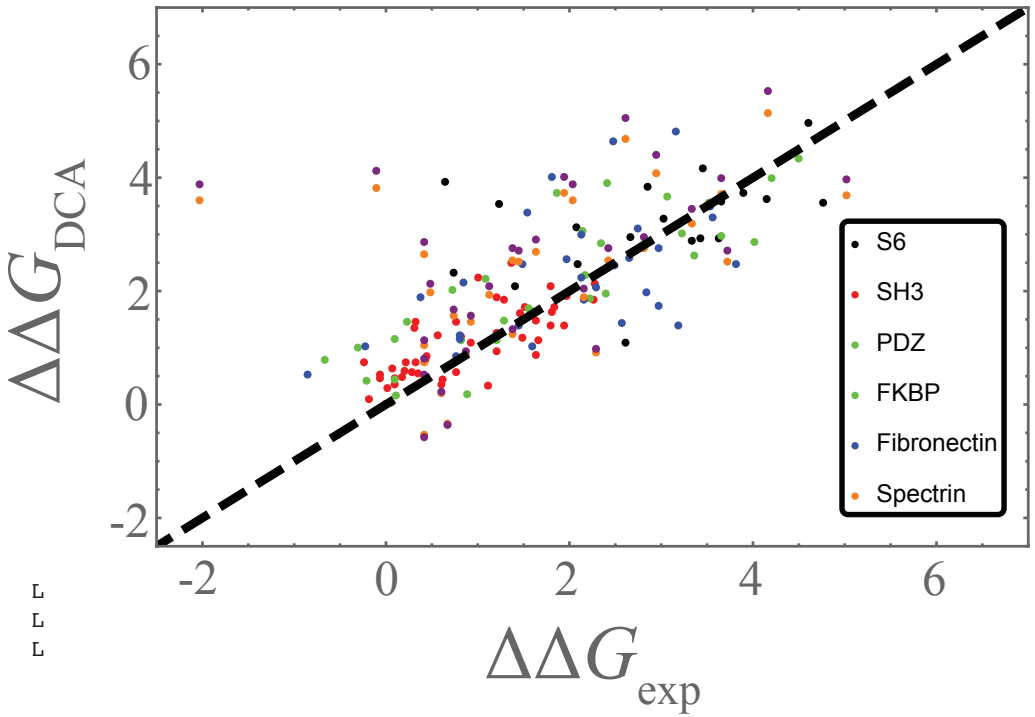
Experimental work for 10 hybrid TCS (Townsend *et al*, PNAS 2013) suggests that

Hybrid TCS proteins do not need to have a highly co-evolved recognition interface since tethering greatly increases their rate of encounter

Inferred couplings are highly correlated with mutational changes in protein stability

$$\mathfrak{A}_{DCA}(\text{sequence}) = -\sum_{(i,j)\in contacts} J_{ij}(A_i, A_j) - \sum_i h_i(A_i)$$

$$\Delta\Delta G_{DCA} = \mathfrak{A}_{DCA}(\text{mutant}) - \mathfrak{A}_{DCA}(\text{wild type})$$



Pearson correlation ~ 0.7

Morcos, Schafer, Cheng, Onuchic, Wolynes, PNAS, 2014
S. Lui and G. Tiana, *J. Chem. Phys.*, 2013
A. Contini and G. Tiana, *J. Chem. Phys*, 2015
Cheng, Raghunathan, Noel, Onuchic, *Protein Sci,* 2015
Figliuzzi, Jaquier, Schug, Tenaillon, Weigt. MBE, 2015

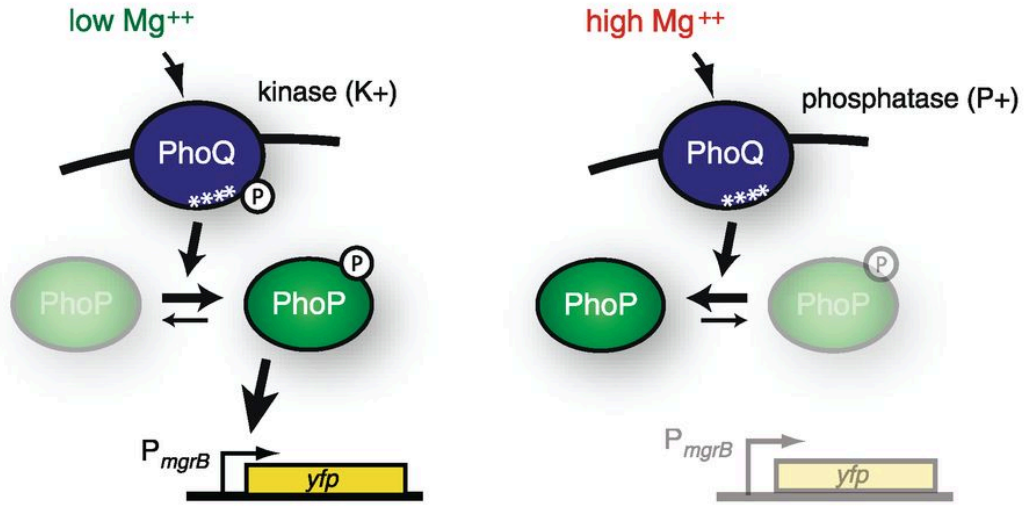
The evolutionary landscape of a TCS interface

PROTEIN EVOLUTION

Pervasive degeneracy and epistasis in a protein-protein interface

Anna I. Podgornaia[1,2]* and Michael T. Laub[2,3]†

Mapping protein sequence space is a difficult problem that necessitates the analysis of 20^N combinations for sequences of length *N*. We systematically mapped the sequence space of four key residues in the *Escherichia coli* protein kinase PhoQ that drive recognition of its substrate PhoP. We generated a library containing all 160,000 variants of PhoQ at these positions and used a two-step selection coupled to next-generation sequencing to identify 1659 functional variants. Our results reveal extensive degeneracy in the PhoQ-PhoP interface and epistasis, with the effect of individual substitutions often highly dependent on context. Together, epistasis and the genetic code create a pattern of connectivity of functional variants in sequence space that likely constrains PhoQ evolution. Consequently, the diversity of PhoQ orthologs is substantially lower than that of functional PhoQ variants.



20^4 = 160,000 total amino acid variants

1,659 functional variants

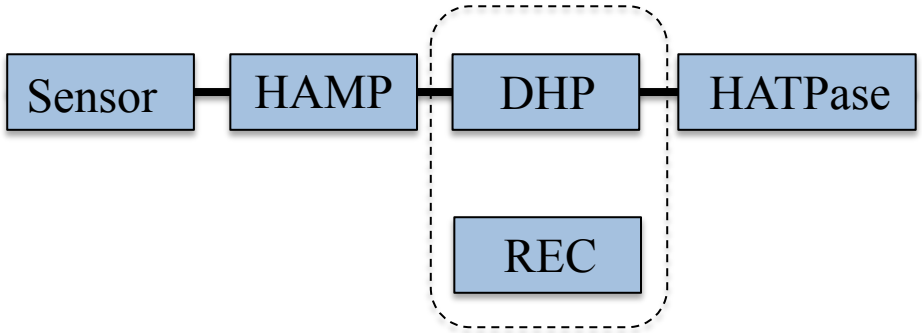
158,341 non-functional variants

Building a statistical model of the PhoQ-PhoP mutational landscape

Histidine Kinase (DHP) | Response Regulator (REC)

```
AAKAESARGHAT PRAPKDAQHEAA
AAKAESARGHAT PRAPKDAQHEAA
SAKEKNEKMKIV N-LIDKGKKSGS
TELETKFTLDQV DQLEEQGKKRSS
LAPSGNTALATA KKEITDRTDDPV
TELETKFTLDQV PRAEKDGKKRSS
```

Key assumption: Analyzing sequences that match the domain architecture



DHP | REC

$$\text{sequence} = (A_1, A_2, \ldots, A_{N_{HK}}, A_{N_{HK}+1}, \ldots, A_{N_{HK}+N_{RR}})$$

$$\mathfrak{A}_{DCA}(\text{sequence}) = -\sum_{i<j} J_{ij}(A_i, A_j) - \sum_i h_i(A_i)$$

$$P(\text{sequence}) = \exp\left(-\mathfrak{A}_{DCA}\right) / Z$$

Histidine Kinase sequence

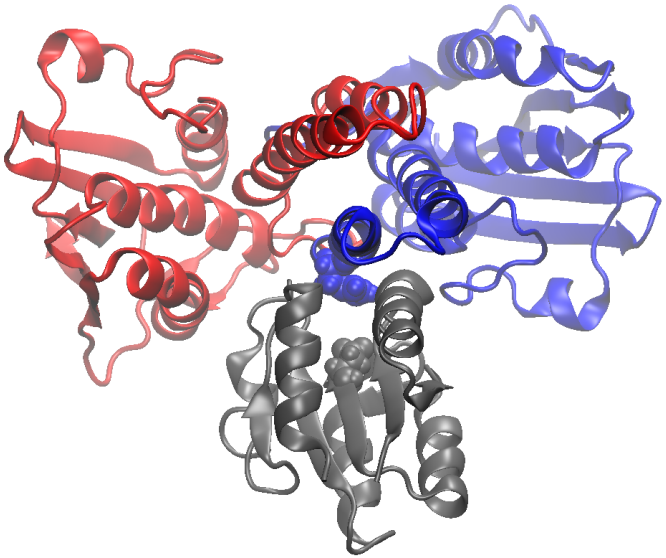
Response regulator sequence

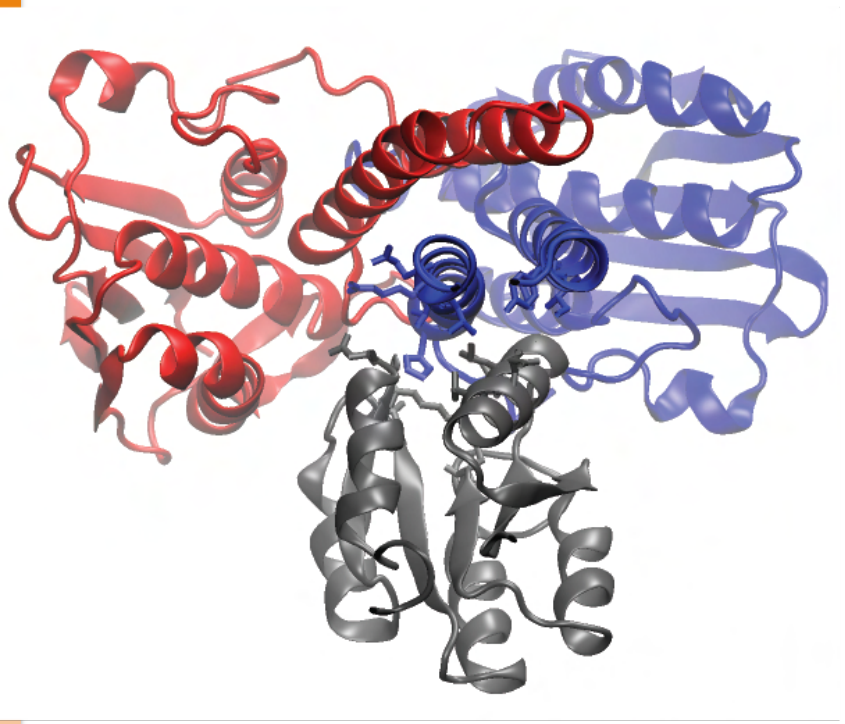
$$\text{sequence} = (A_1, A_2, \ldots, A_{N_{HK}}, A_{N_{HK}+1}, \ldots, A_{N_{HK}+N_{RR}})$$

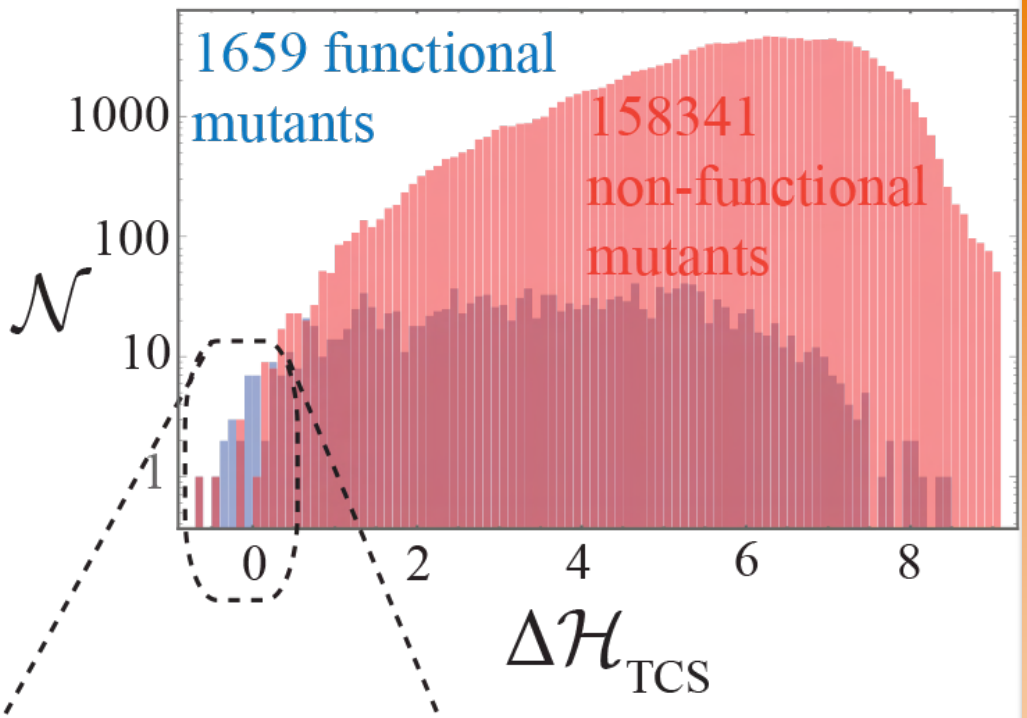
Including only interfacial contacts

$$\mathfrak{A}_{TCS}(\text{sequence}) = -\sum_{i,j \in \text{interprotein}} J_{ij}(A_i, A_j) \times \Theta(L - r_{ij}) - \sum_i h_i(A_i)$$

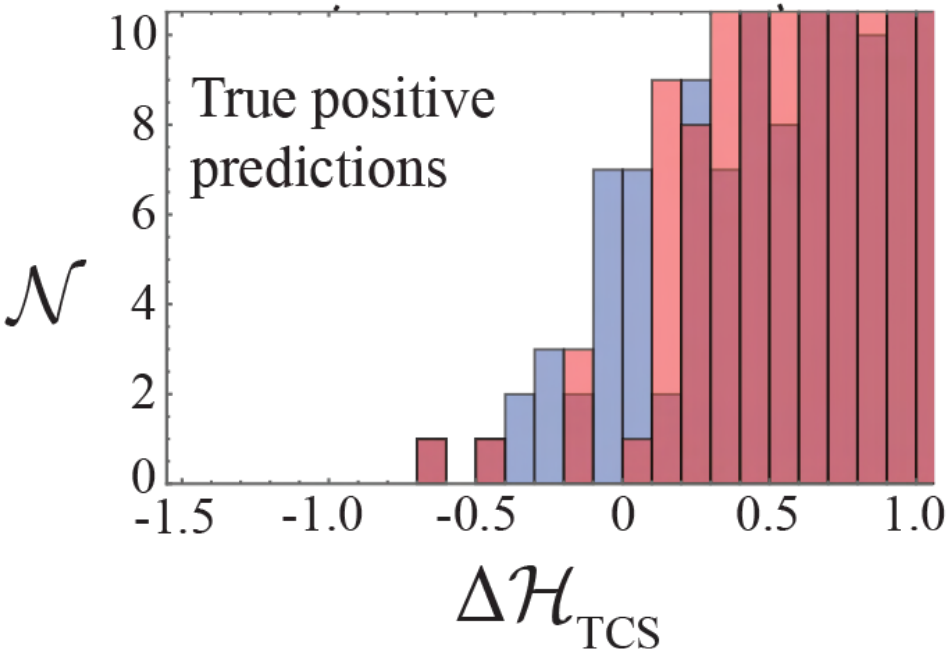
$$\Delta\mathfrak{A}_{TCS} = \mathfrak{A}_{TCS}(\text{mutant sequence}) - \mathfrak{A}_{TCS}(\text{wt sequence})$$





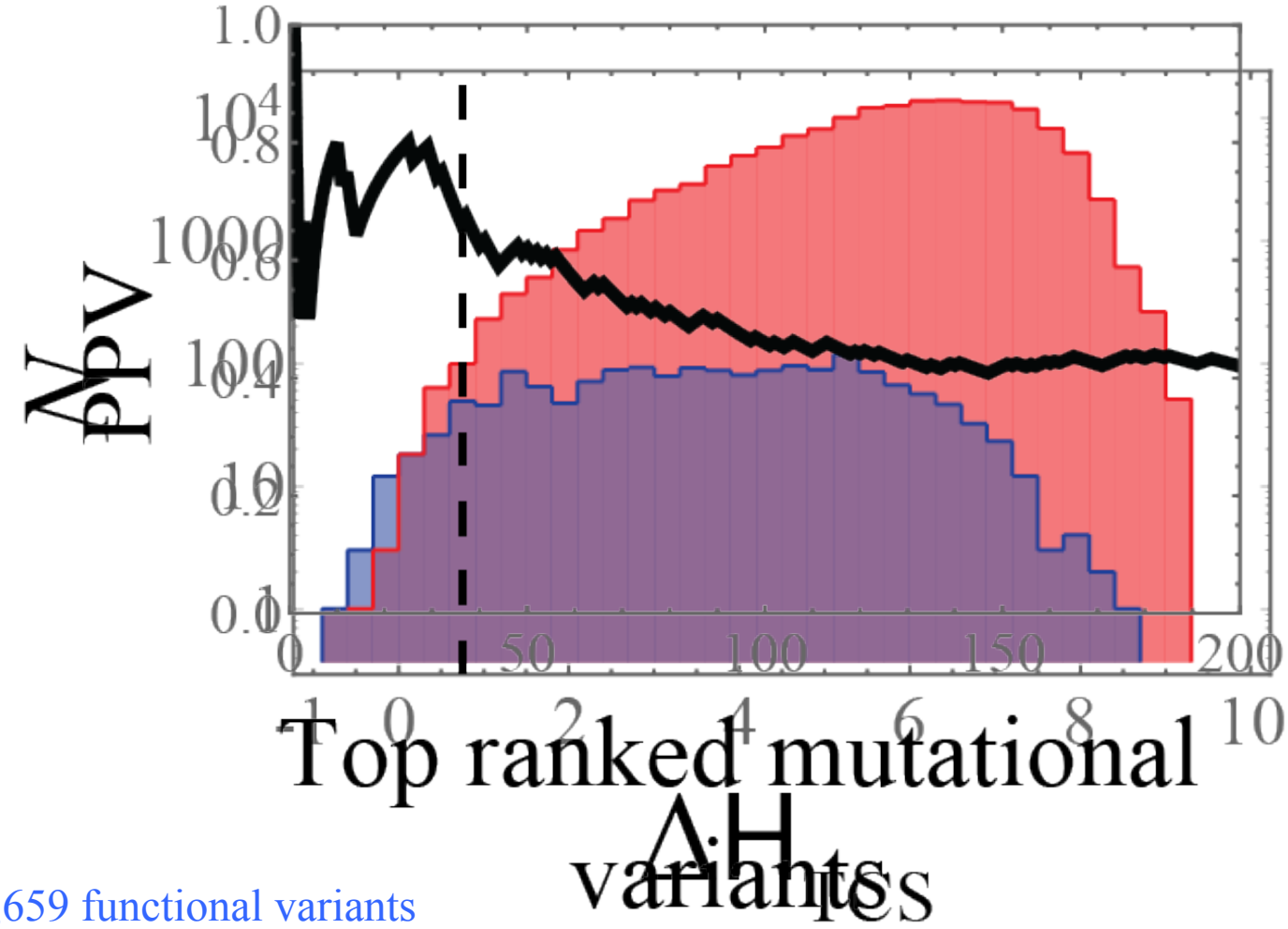


Tail of distribution mostly true positive mutational variants



Cheng et al. (submitted)

Top ranked mutational variants are a good predictors of functional mutants

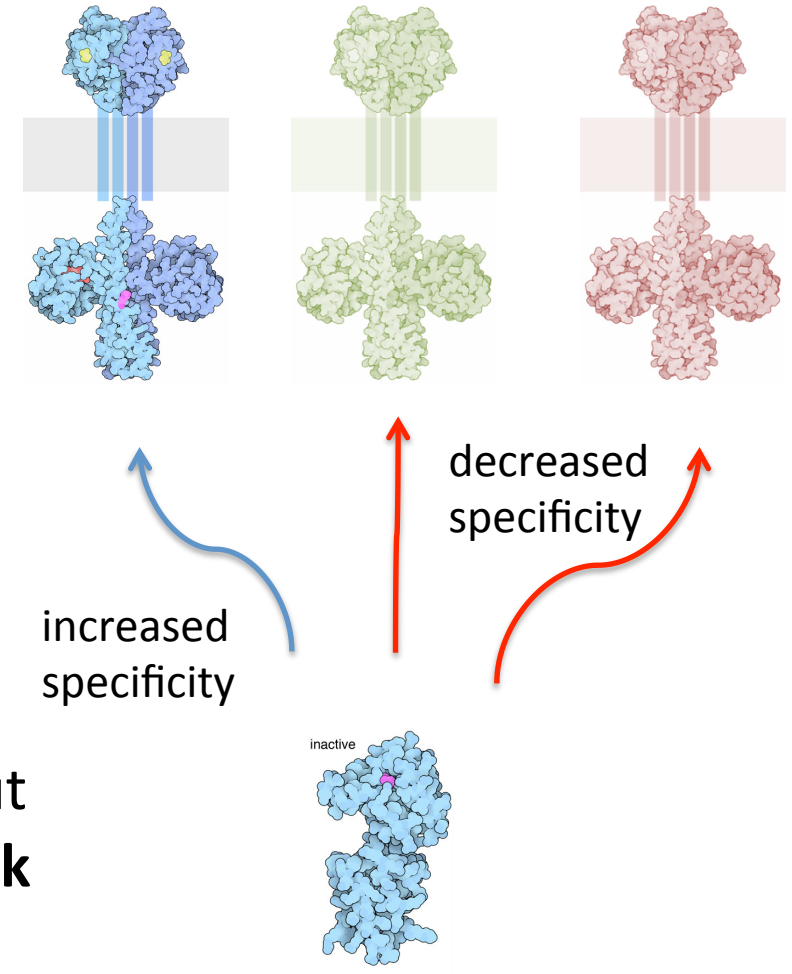


1,659 functional variants
158,341 non-functional variants

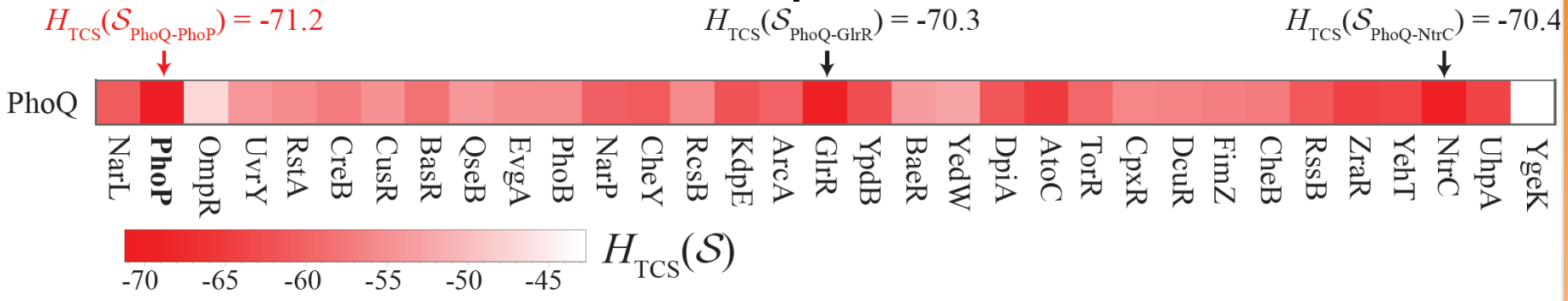
Cheng et al. (submitted)

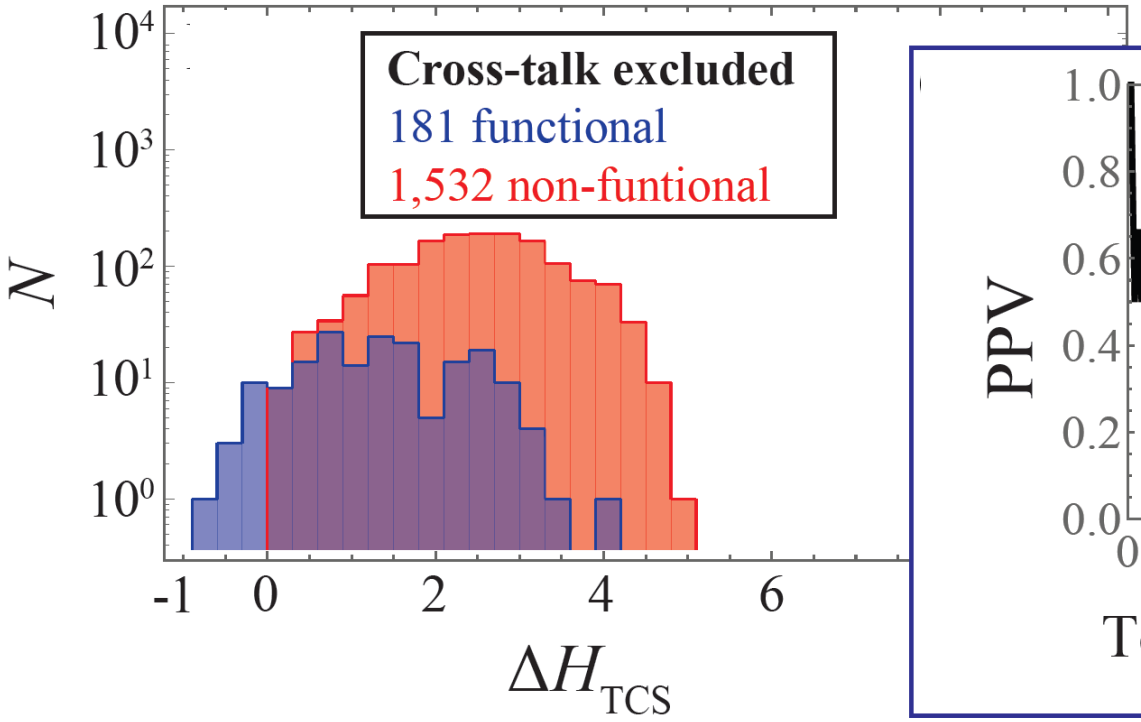
Multiple coevolution in HK-RR pairs

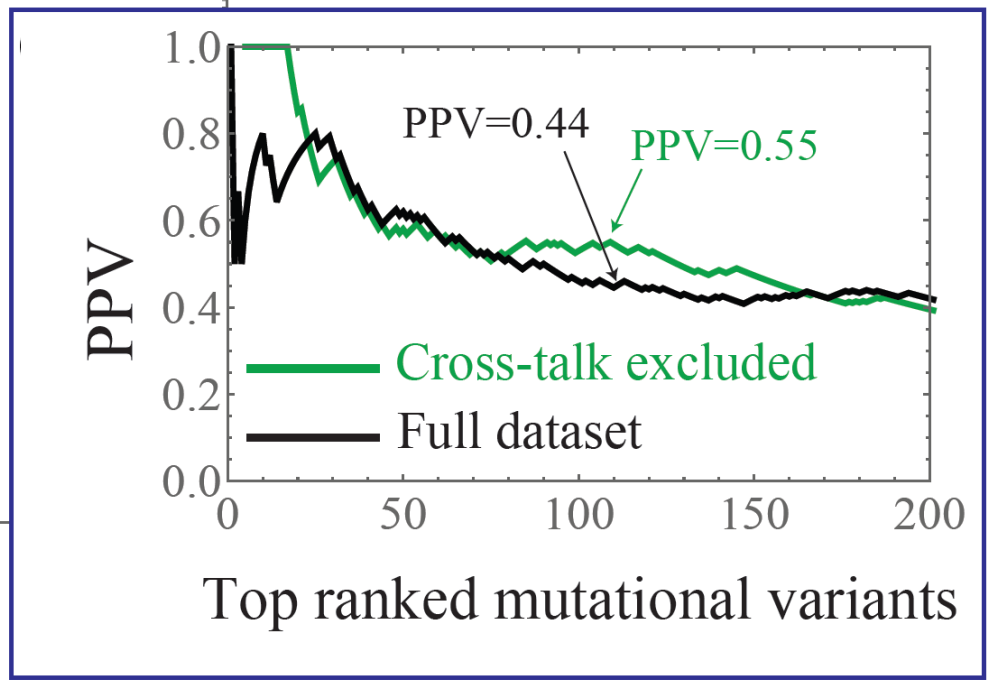
- The comprehensive experiment by Podgornalia & Laub was perfomed ***in vivo***
- Multiple Histidine Kinases and response regulators are active in *E. coli*
- Idea: A **functional** phenotype must be achieved not only by HK-RR pairs that **coevolve for specificity** but also should evolve to **avoid cross-talk** on the other potential partners



Specificity preservation is key for accurate functional predictions





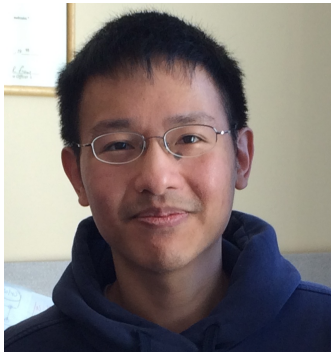
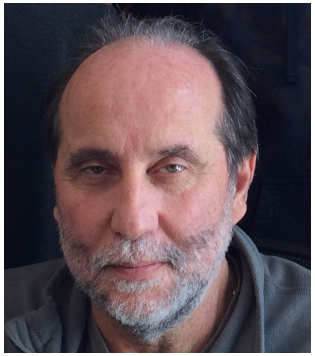


Summary of PhoQ-PhoP analysis

- A model based on coevolutionary couplings can predict the effects of mutations at a HK-RR interface
- For the top mutational variants, we were able to predict to a high degree of accuracy if a given mutated interface is functional or not
- We propose the effects of those mutations on cross talk and its role in the proper function of PhoQ-PhoP

Acknowledgments

Direct Coupling Analysis
José N. Onuchic
Terence Hwa
Martin Weigt


TCS
Ryan Cheng
Herbert Levine
Ellinor Haglund
Samuel Flores
Olle Nordesjo

Protein Interactions
Ricardo Dos Santos
Rachel Nechushtai
Patricia Jennings

Circadian TCS
Joseph Boyd
Mark Paddock
Susan Golden

DCA server
Bryan Lunt
Martin Weigt
Alex Schug

Structure & Dynamics
Biman Jana
Jeff Noel
Joanna Sulkowska

Folding & Coevolution
Ryan Cheng
Nicholas Schafer
Peter Wolynes

UT DALLAS
NSF
CTBP