Inter-protein residue contacts and interaction specificity in bacterial signal transduction

Martin Weigt

KITP Santa Barbara

27 Jan 2011

Structural conservation vs. sequence variability in families of homologous proteins

```
RIDHFLKNTTHELNTPMSALVLSLKTLED
FLNGFIRDTTHEINTPLSVILMSIEMFXT
HERQLTGELSHELRTPLSRIIAELDWWQT
KYRTTLTDLTHSLKTPLAVLQSTLRSLRS
RARMQVGNLAHSLKTPIAVLLNEARVLEK
TQKEFLANLSHELKTPLAVVMNTLETLLD
KQQSFVENASHELRTPLAVLQNRLETLFR
RQERFSADASHQLKTPLAVLKTQAAVALA
RLNQFADDLAHELRTPVNILLGKNQVMLS
RLSQFSSNLAHDMRTPLTNLLAEAQVALS
KLSRFSADLAHDLRTPLNNLIGHAEVALS
KLSDFSSDIAHELRTPVSNLMMQTQFALA
RLSNFSADIAHELRTPISNLRTHTEVILA
RQSNFSADIAHEIRTPITNLITQTEIALS
```

conserved structure and function across species

↓

constrained evolution

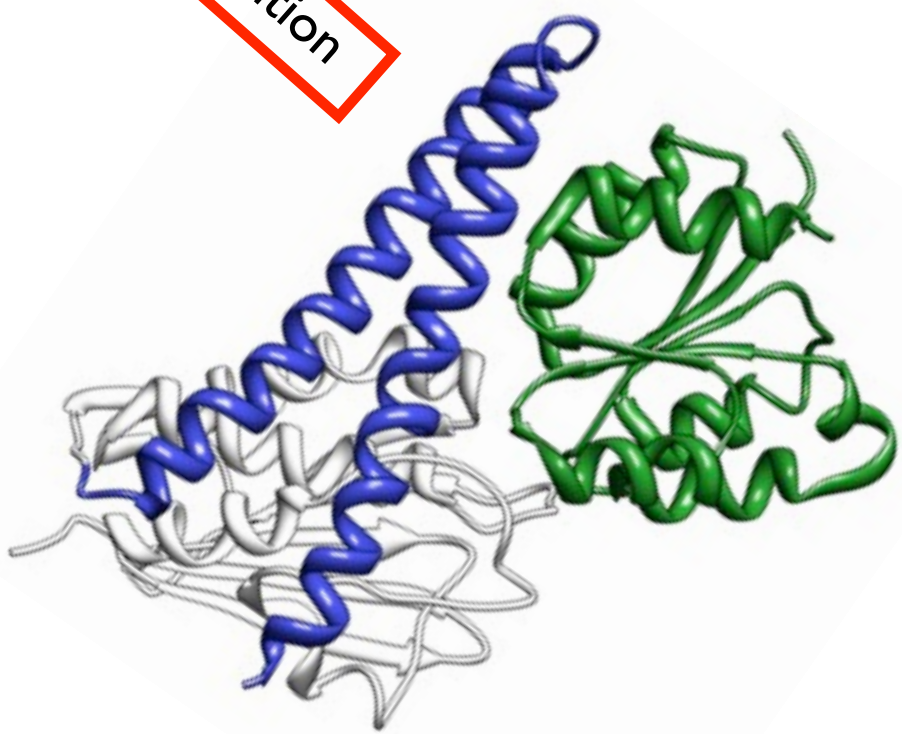
protein evolution

statistical inference

sequence variability across species

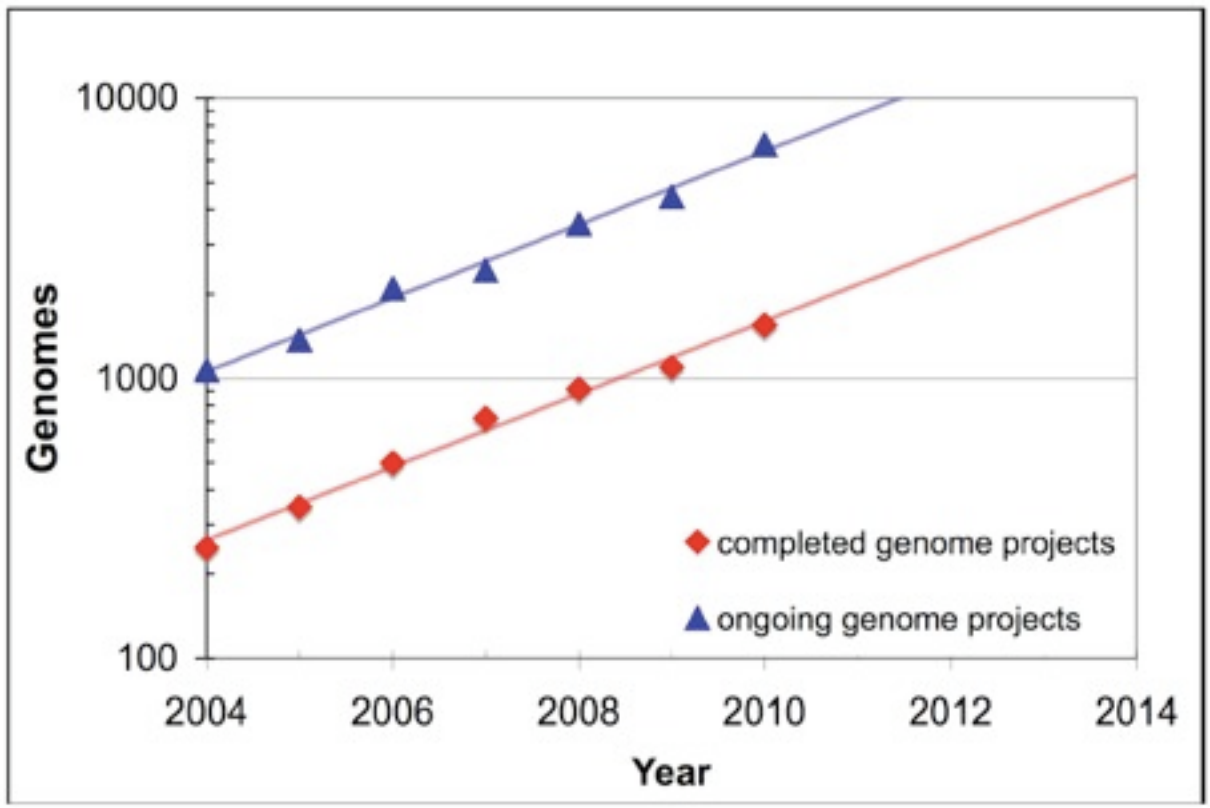
↓

informs structural and functional prediction



Data

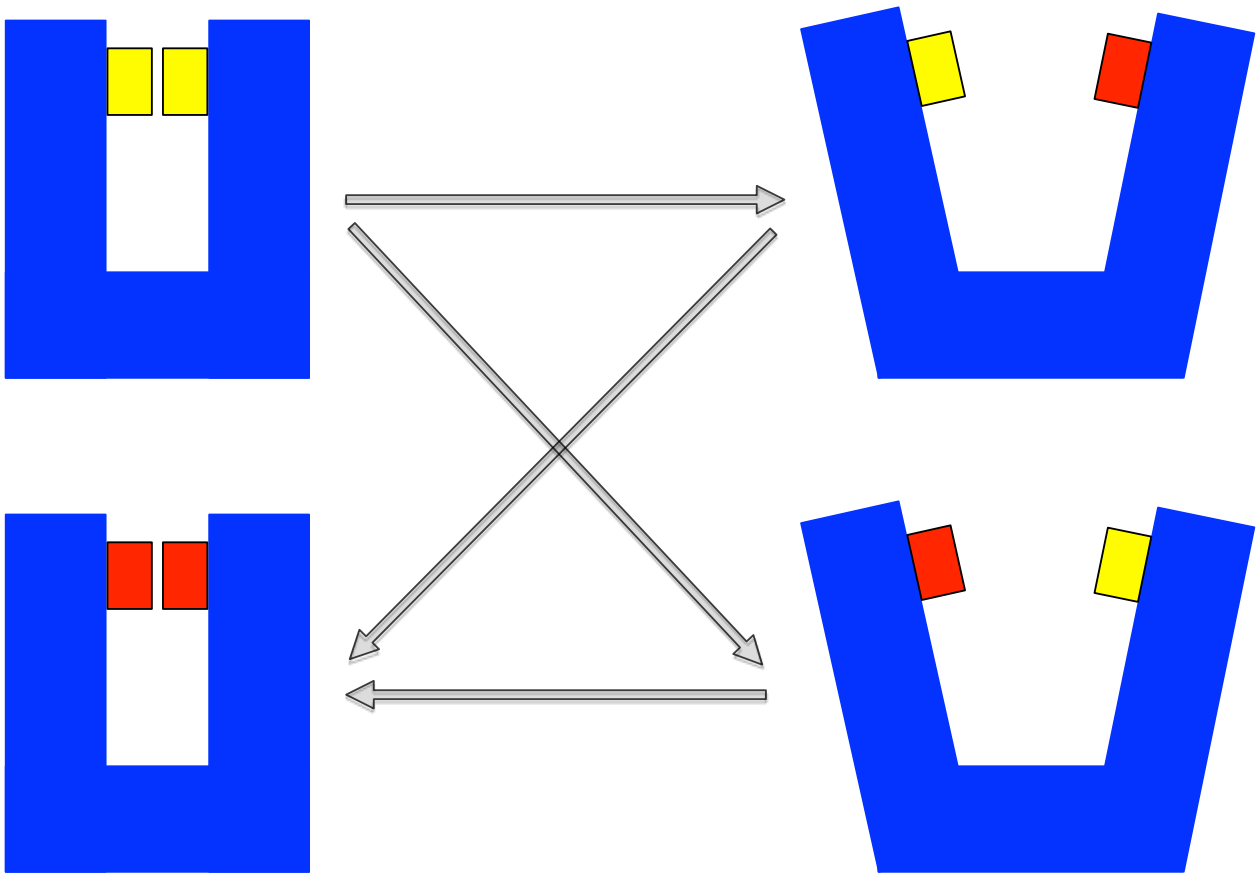
- >1700 completed genome sequencing projects
- doubling every 2-3 years



GOLD data base

- abundant protein domain families: 1,000 - 100,000 sequences
- homologous proteins from distant species: ~20-40% sequence identity

Residue contacts induce sequence correlations



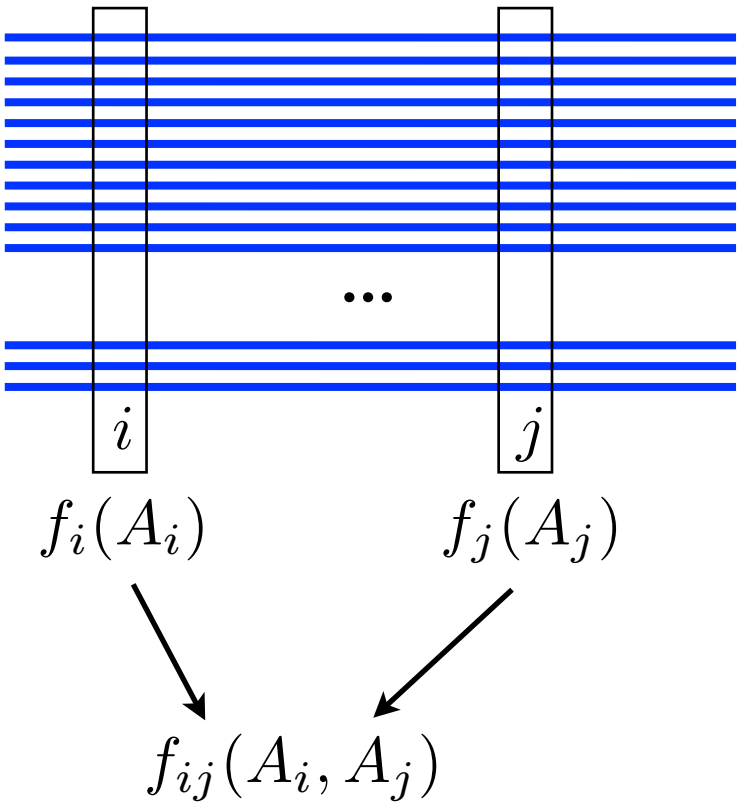
Inverse question:

- Are sequence correlations indicative for inter-protein residue contacts?

[Gobel et al. '94, Neher '94, Ranganathan et al. '99]

Sequence statistics and correlations

Multiple sequence alignment (MSA): $\{A_i^a \,|\, i = 1, ..., L; a = 1, ..., M\}$



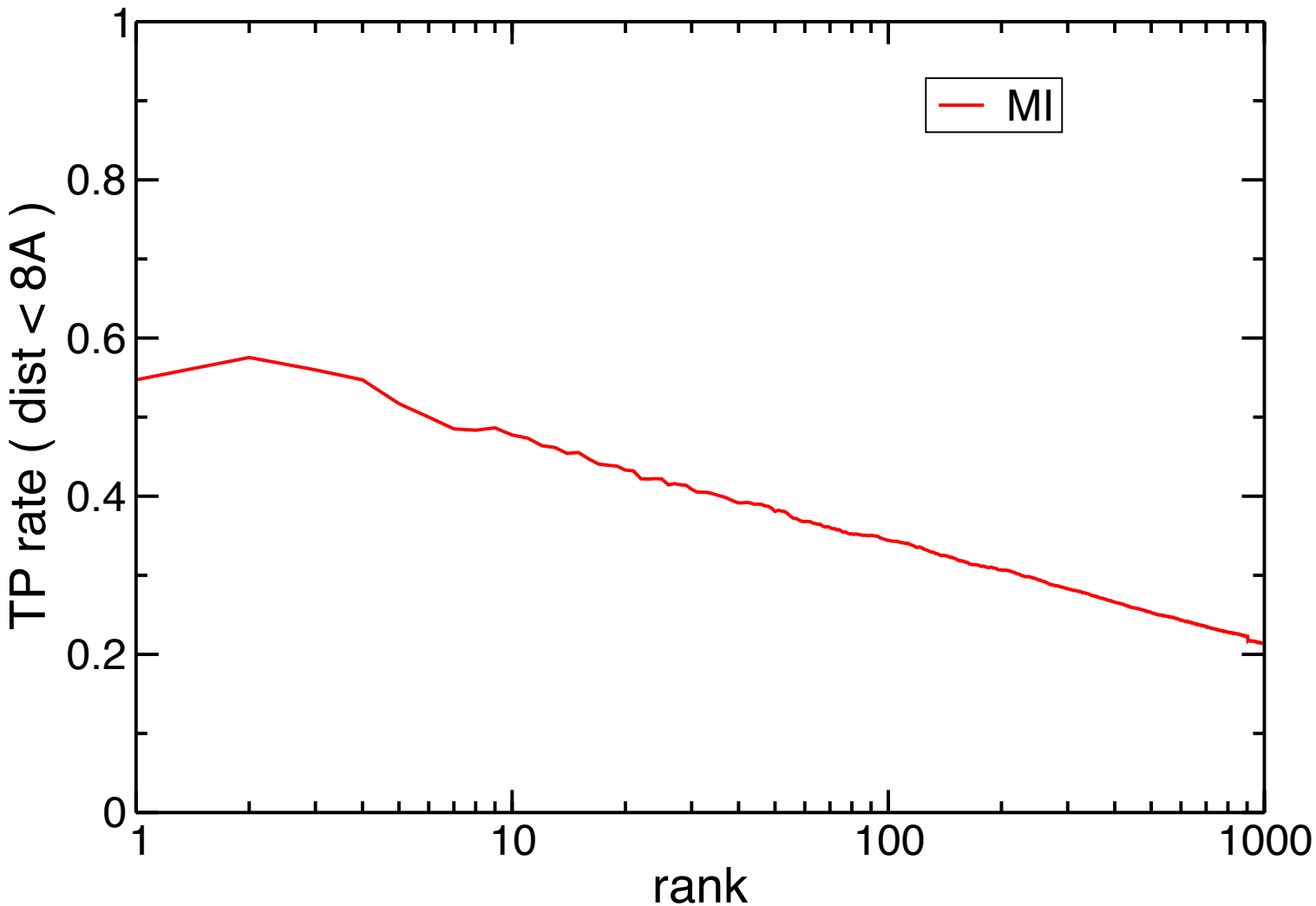
Mutual information measures pair correlation

$$MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A)\, f_j(B)}$$

Compare to 3D protein structure: Are correlated column pairs in contact?

Correlations and residue contacts

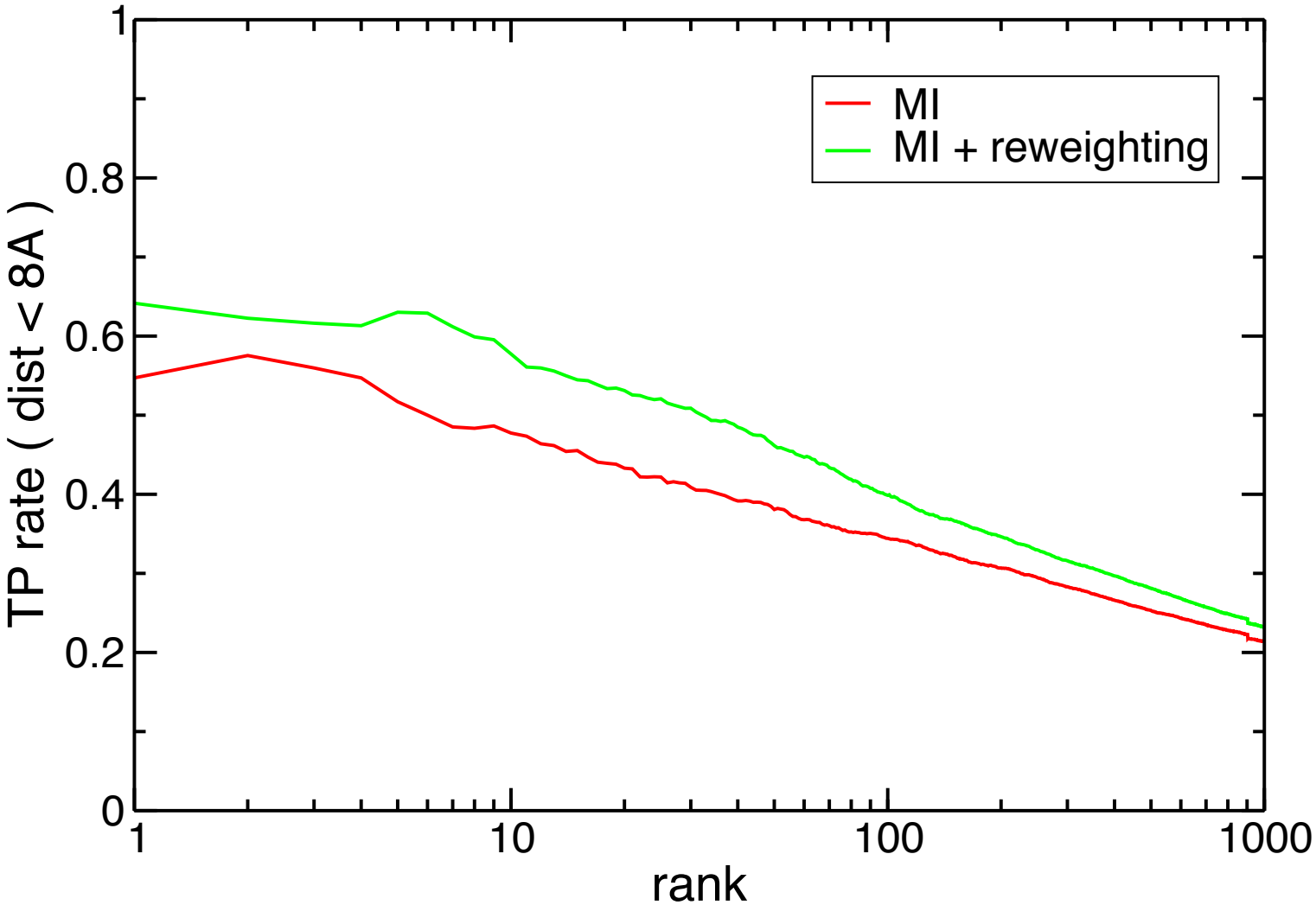
Comparison for 53 abundant protein families: $|i - j| \geq 5$



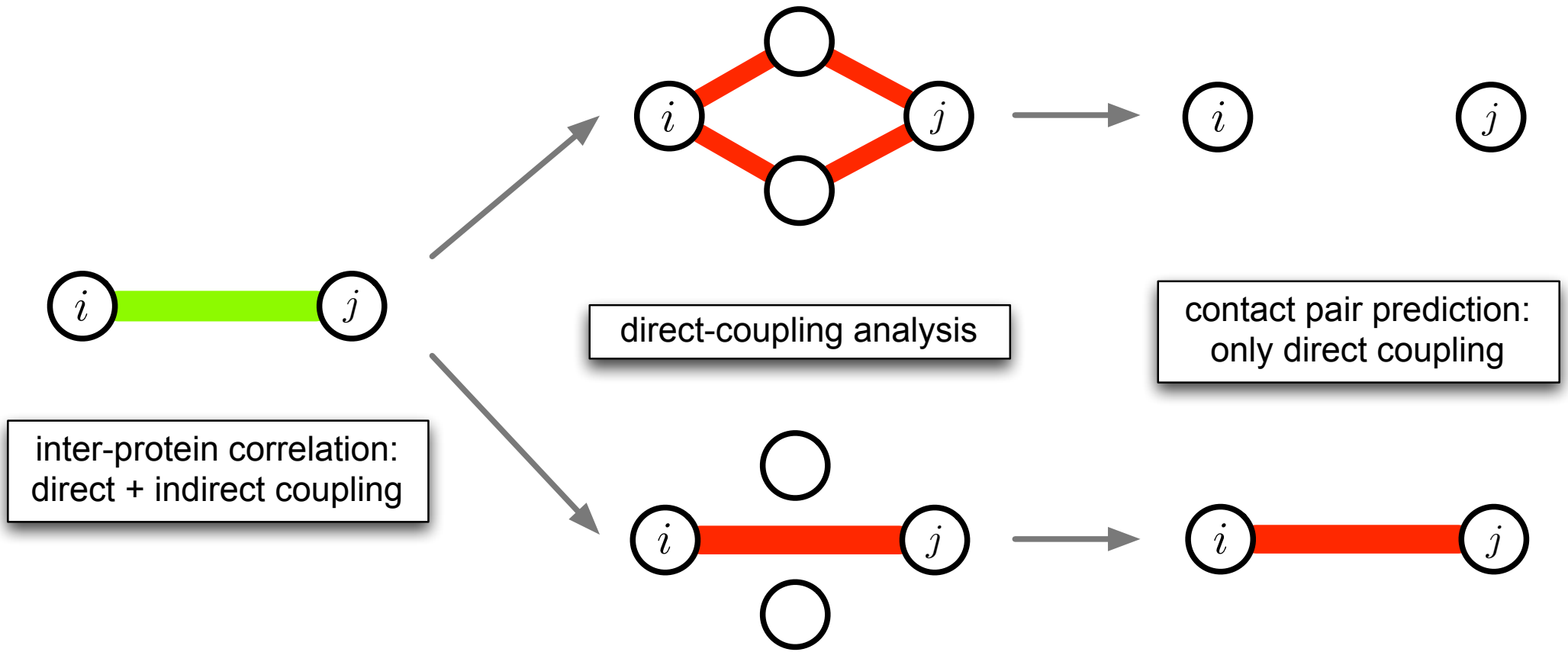
enriched in contacts, but many false positives

Sampling bias

Uneven sampling due to phylogeny, multiple strains...



Correlation results from direct and indirect coupling



- correlations not sufficient to identify residue contacts
- disentangle direct and indirect couplings: $P(A_1, ..., A_L)$
- statistical-physics inspired direct coupling analysis (DCA)

[MW, White, Szurmant, Hoch, Hwa, PNAS '09]

Direct coupling analysis

- model data via global distribution $P(A_1, ..., A_L)$ such that

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i,j\}} P(A_1, ..., A_L) \stackrel{!}{=} f_{ij}(A_i, A_j)$$

Direct coupling analysis

- model data via global distribution $P(A_1, ..., A_L)$ such that

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i,j\}} P(A_1, ..., A_L) \stackrel{!}{=} f_{ij}(A_i, A_j)$$

- maximum-entropy model:

$$-\sum_{\{A_i\}} P(A_1, ..., A_L) \ln P(A_1, ..., A_L) \rightarrow \max$$

➡ disordered 21-states Potts model

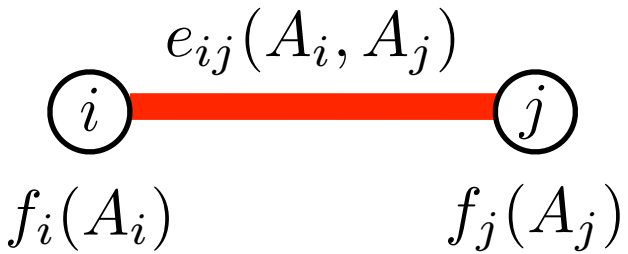
$$P(A_1, ..., A_L) \sim \exp\left\{ + \sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right\}$$

direct coupling of residues *i* and *j*

Interaction strength and direct information

How to quantify direct interaction by scalar quantity:

➡ consider isolated two-spin system



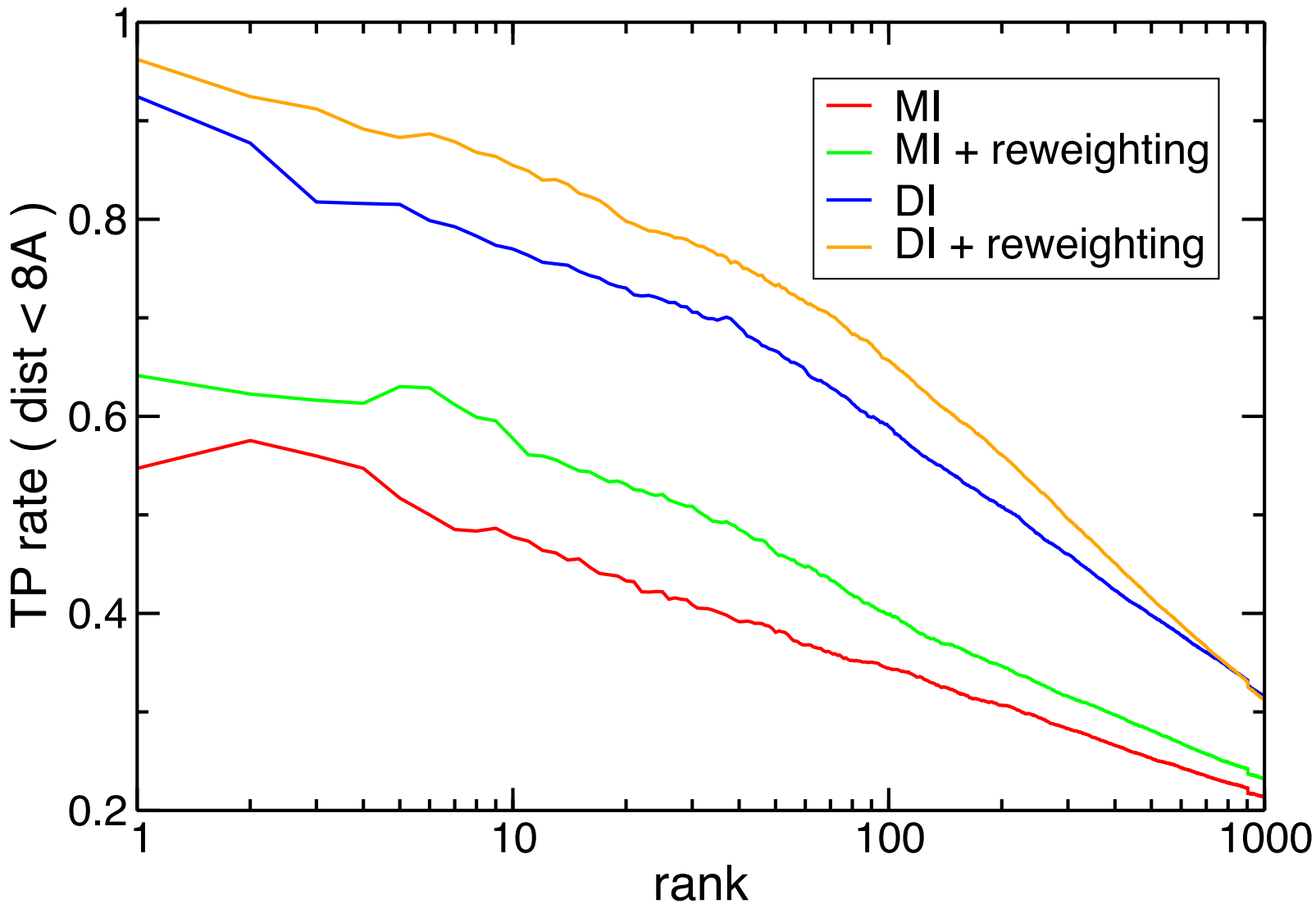
➡ direct information = mutual information due to direct coupling

$$DI_{ij} = \sum_{A_i, A_j} P_{ij}^{(dir)}(A_i, A_j) \log \frac{P_{ij}^{(dir)}(A_i, A_j)}{f_i(A_i) f_j(A_j)}$$

➡ multi-information in Bethe-Peierls approximation

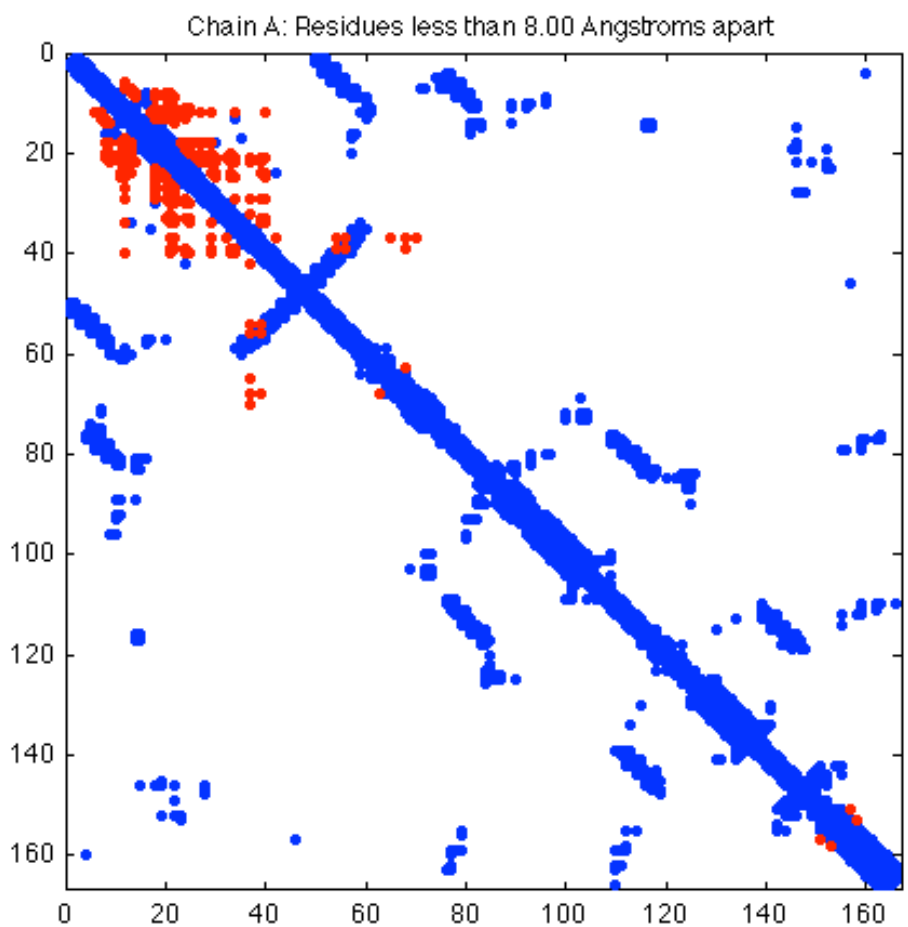
$$-\sum_{\{A_i\}} P(A_1, ..., A_L) \ln \frac{P(A_1, ..., A_L)}{\prod_i f_i(A_i)} \simeq \sum_{i<j} DI_{ij}$$

Direct information vs. residue distance

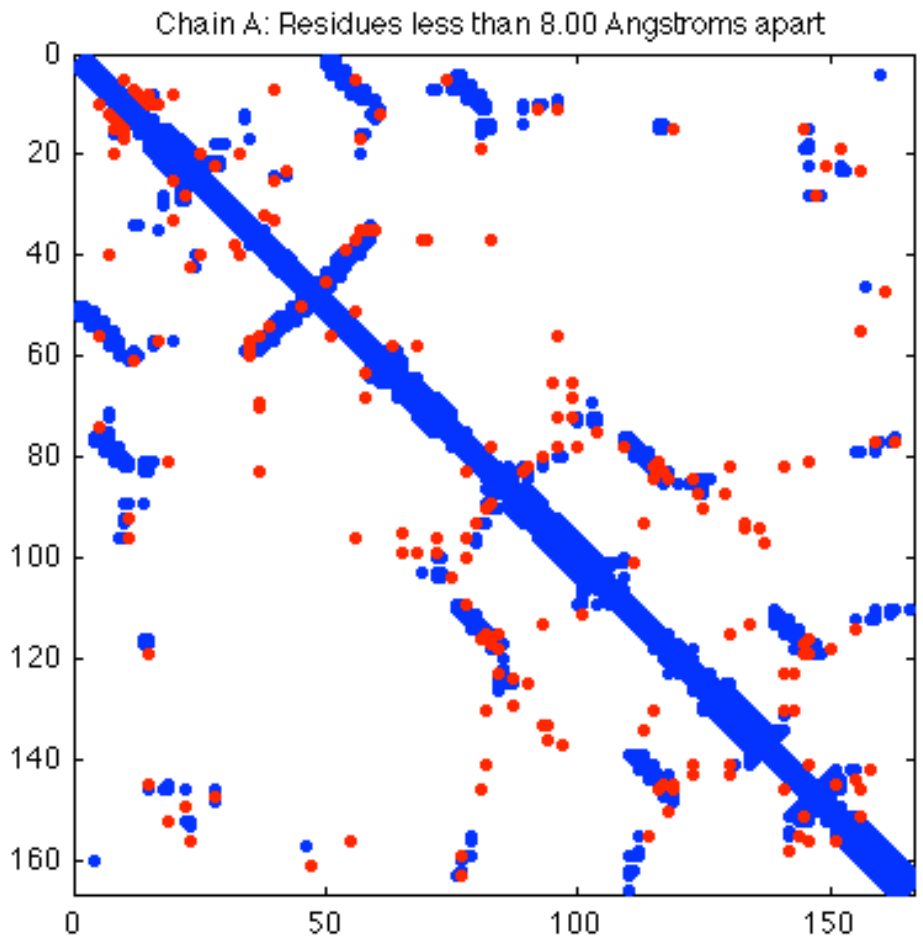


Not all contacts co-vary, but...

Ras contact map as example:

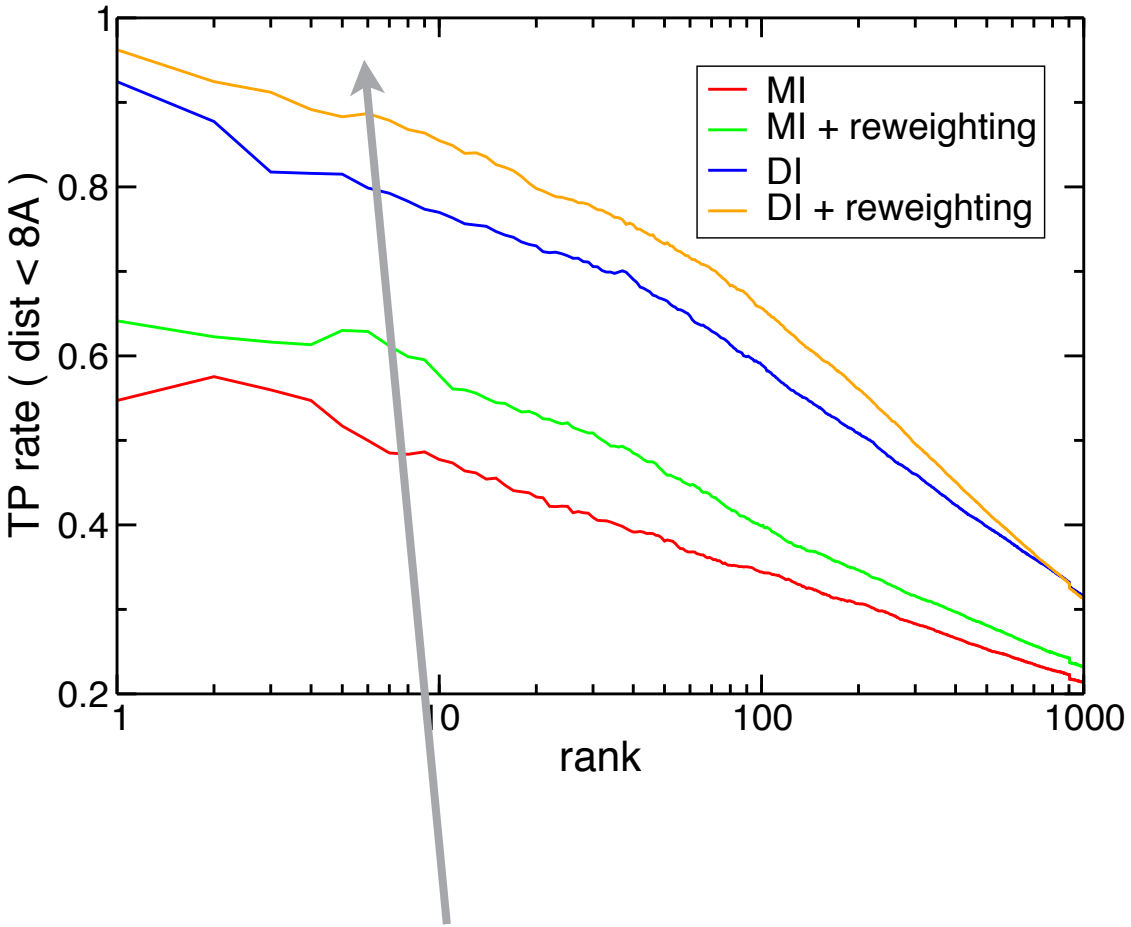


contacts (blue) vs. MI (red)



contacts (blue) vs. DI (red)

Signal beyond residue distance

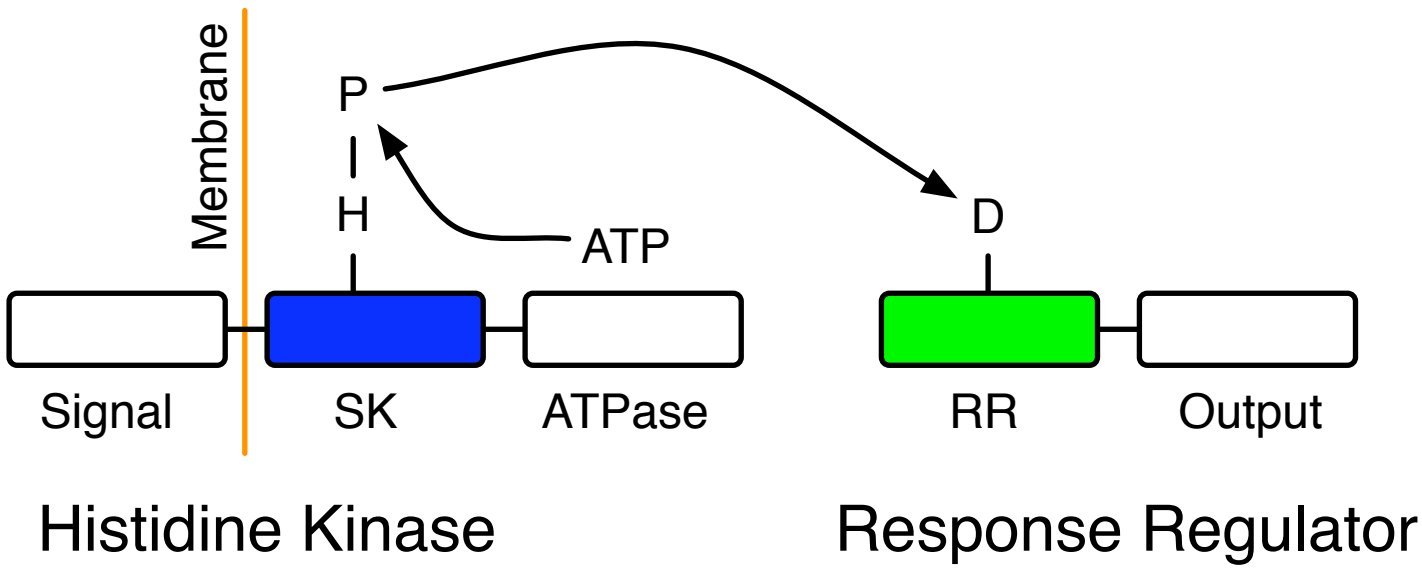


may contain sensible information beyond intra-domain contacts:

- contacts specific to active / inactive protein conformation
- inter-domain / inter-protein contacts
- ligand mediated correlations
- allosteric long-range correlation (?)

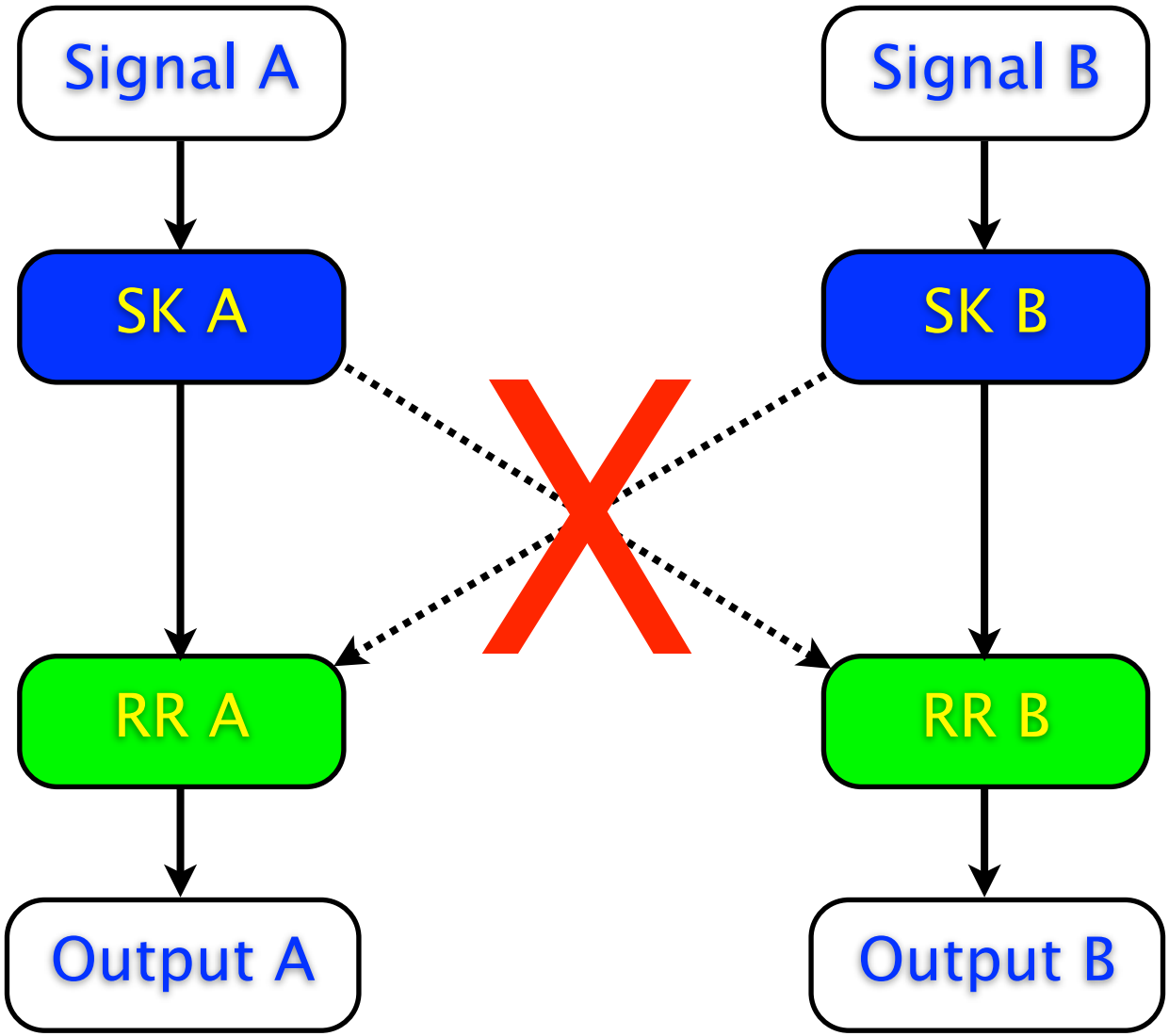
Two-component signaling

- most common signaling system in bacteria



- conservation: most SK, RR belong to two Pfam domain families
- amplification: ~*O*(10) interacting pairs per genome
- specificity of interaction: little cross-talk between signaling pathways

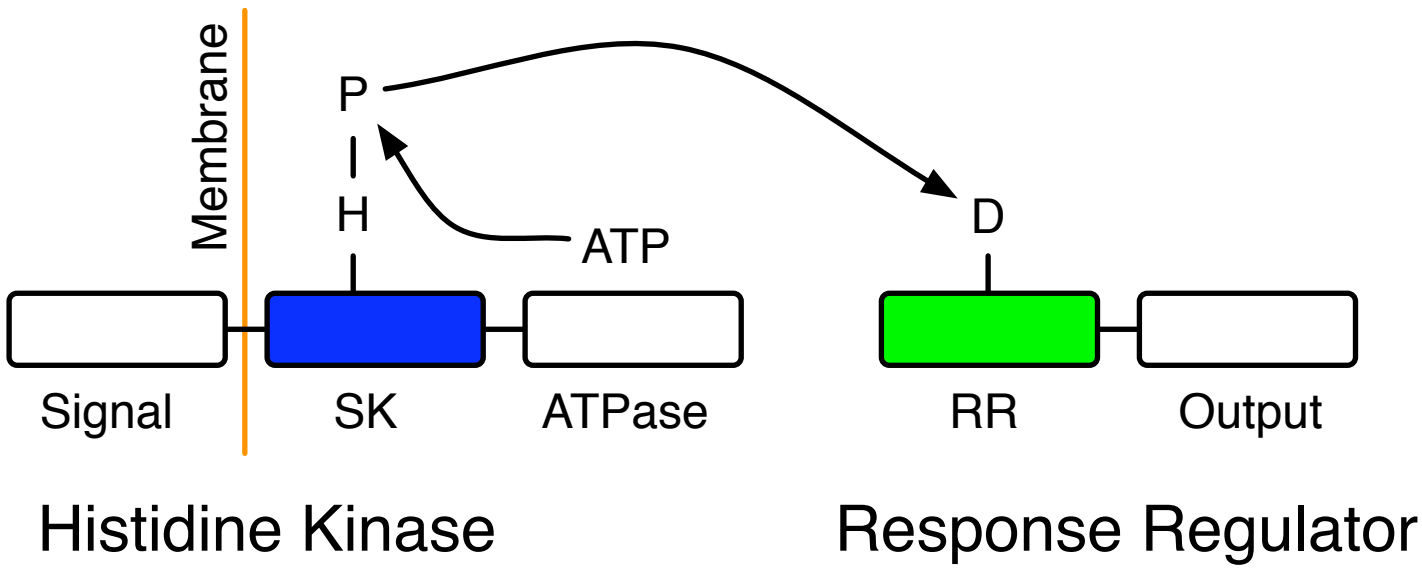
Specificity vs. crosstalk of signaling pathways



Specific interaction but conserved structure!

Inter-protein contacts: Two-component signaling

- most common signaling system in bacteria



- conservation: most SK, RR belong to two Pfam domain families
- amplification: ~*O*(10) interacting pairs per genome
- specificity of interaction: little cross-talk between signaling pathways
- operon organization: partner SK/RR genes frequently co-localized on DNA

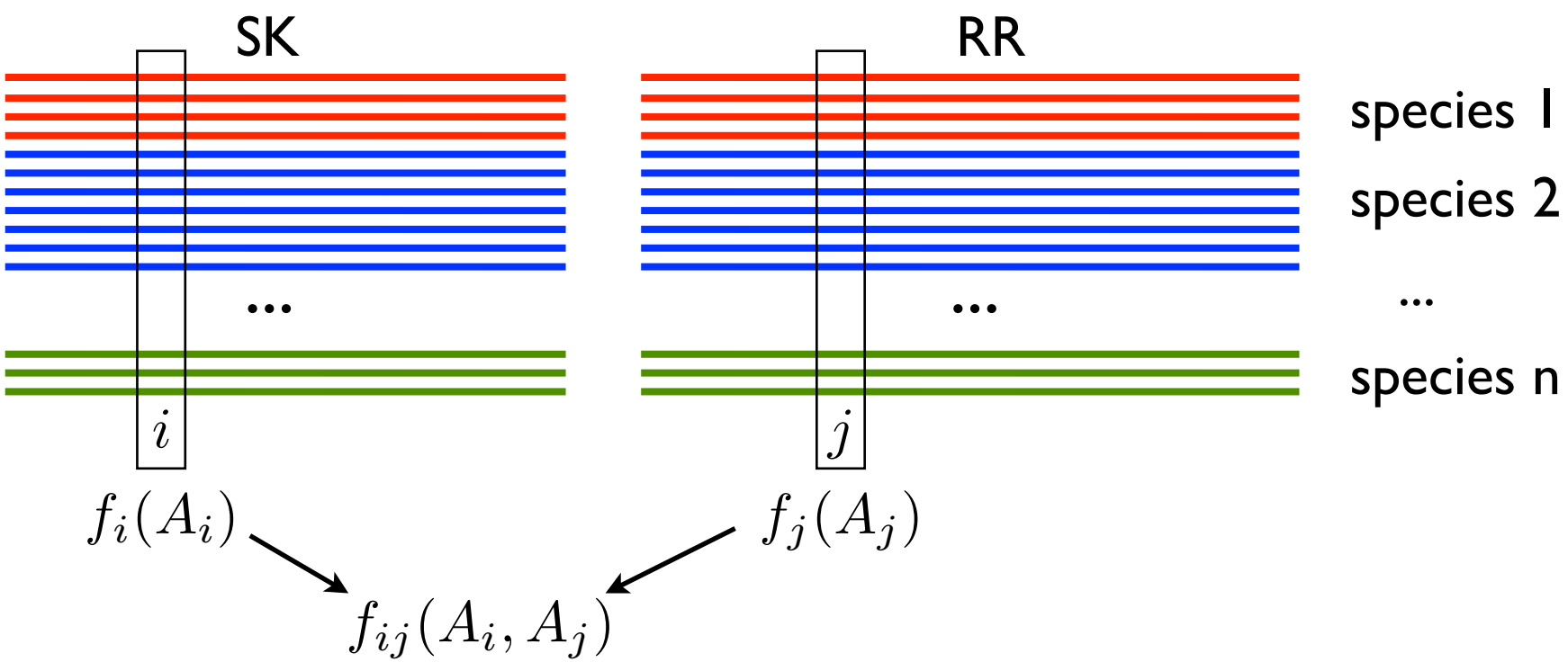
Multiple-sequence alignments for TCS

- ca. 750 bacterial genomes

  ➡ multiple-sequence alignment: $L_{SK} = 87,\ L_{RR} = 117$

  ➡ *M* ~ 9000 cognate SK-RR pairs in same operon,

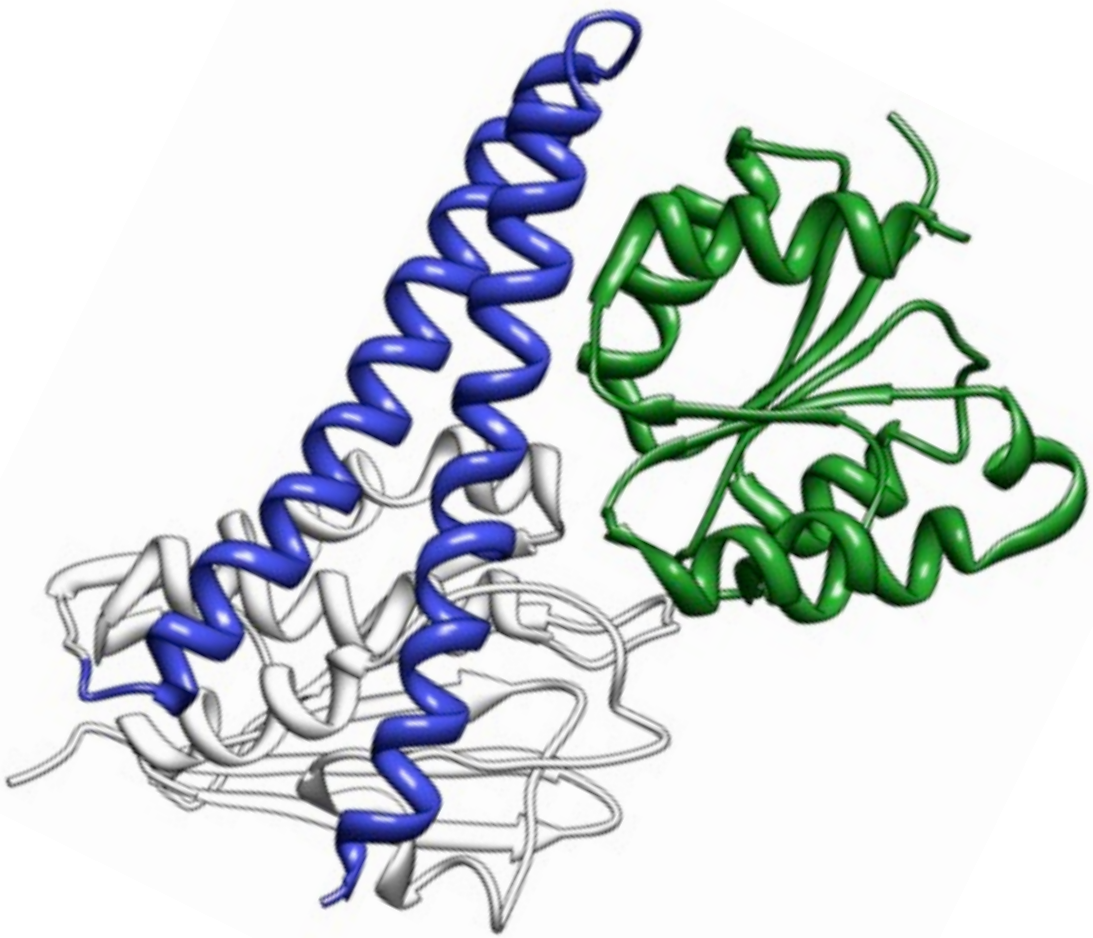
  ca. 3800 orphan SK, ca. 9000 orphan RR



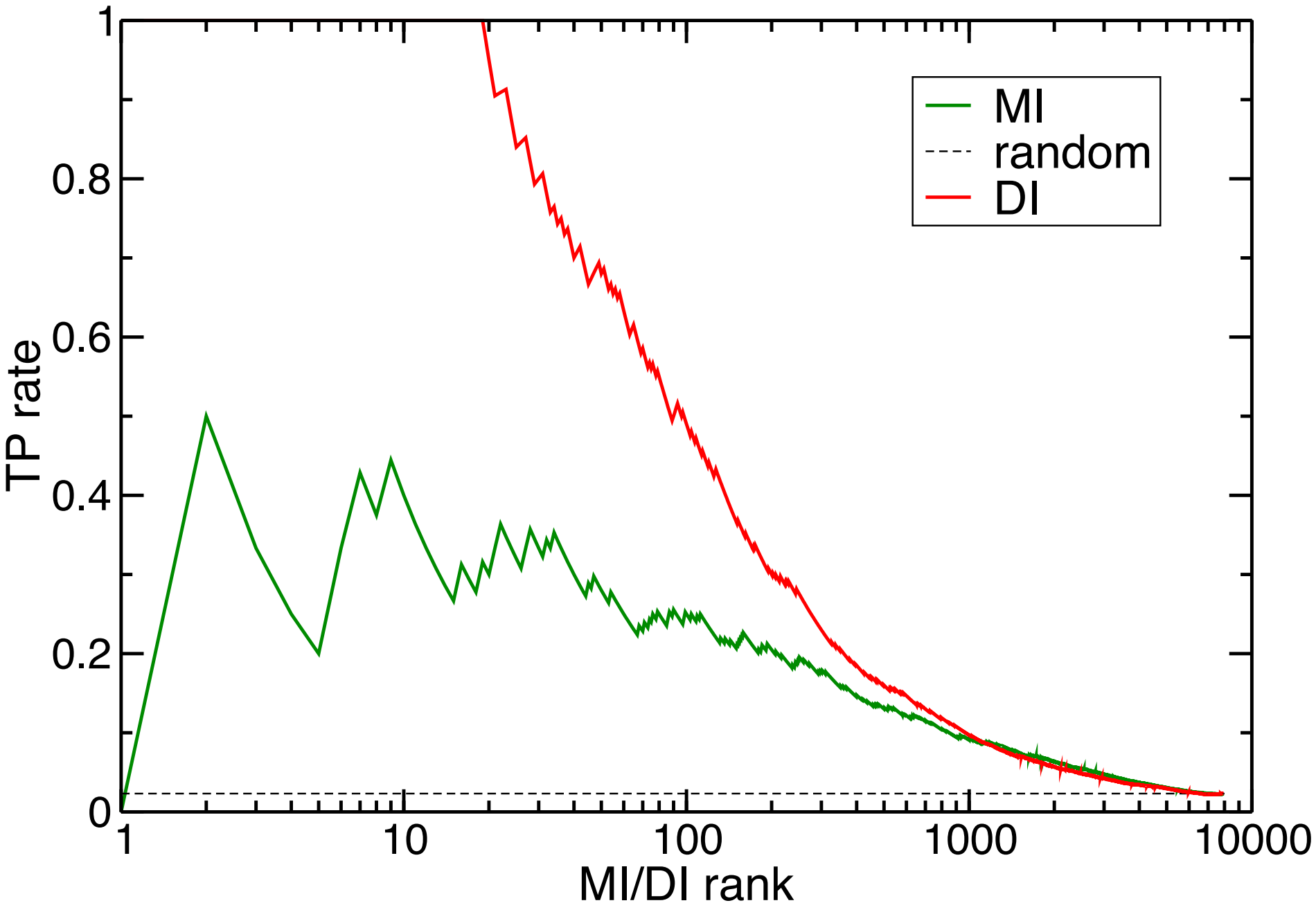
How to test the results?

- one similar co-crystal structure [Zapf et al., *Structure* 2000]
  - sporulation pathway in *Bacillus subtilis*
- two SK/RR structures published in Oct. 2009

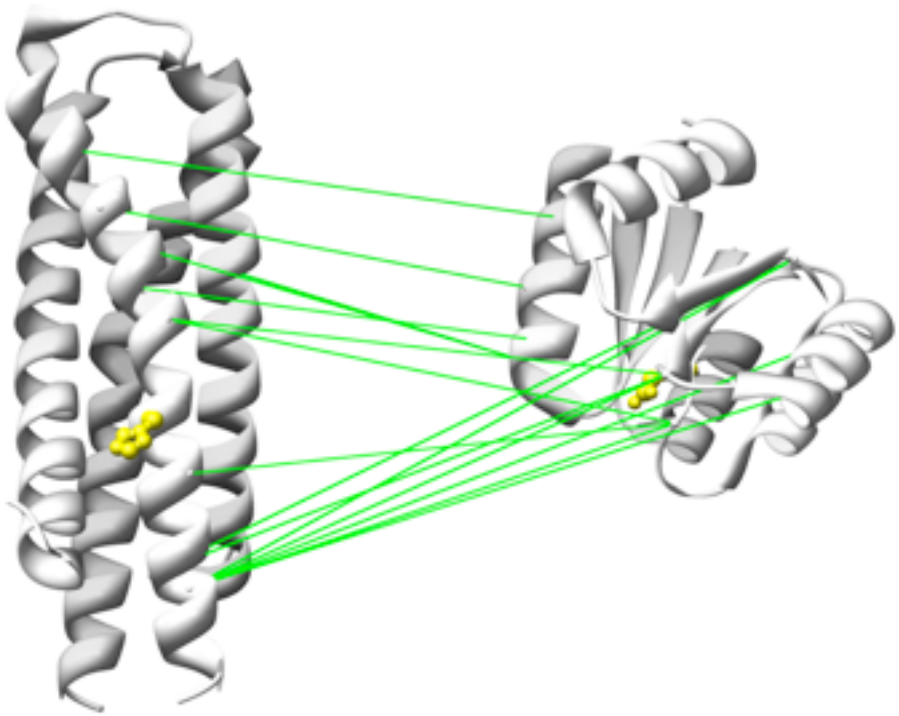
  [Yamada et al., *Structure* 2009; Casino et al., *Cell* 2009]



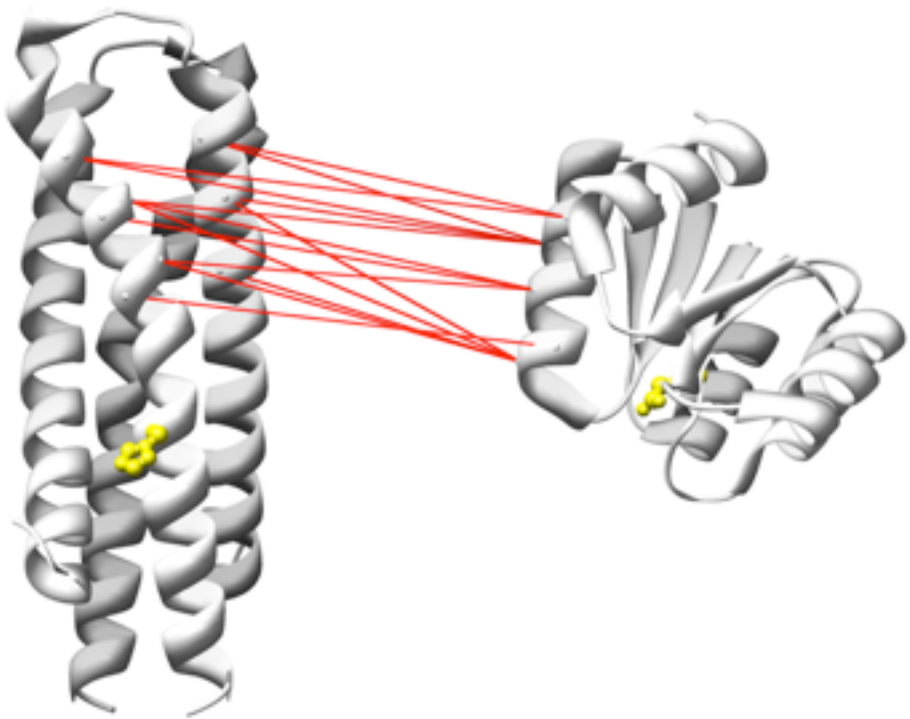
Inter-protein contacts: Two-component signaling



Inter-protein contacts: Two-component signaling

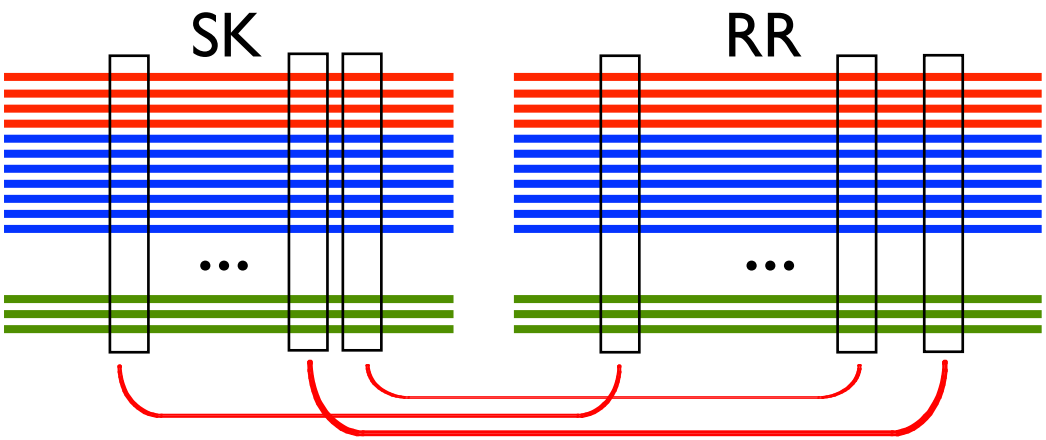


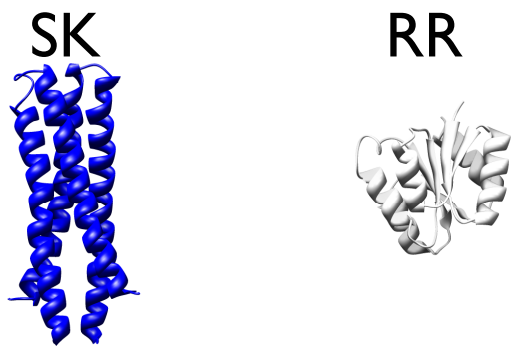
strongest correlations



strongest direct couplings

In silico prediction of high-resolution structures of transient protein complexes

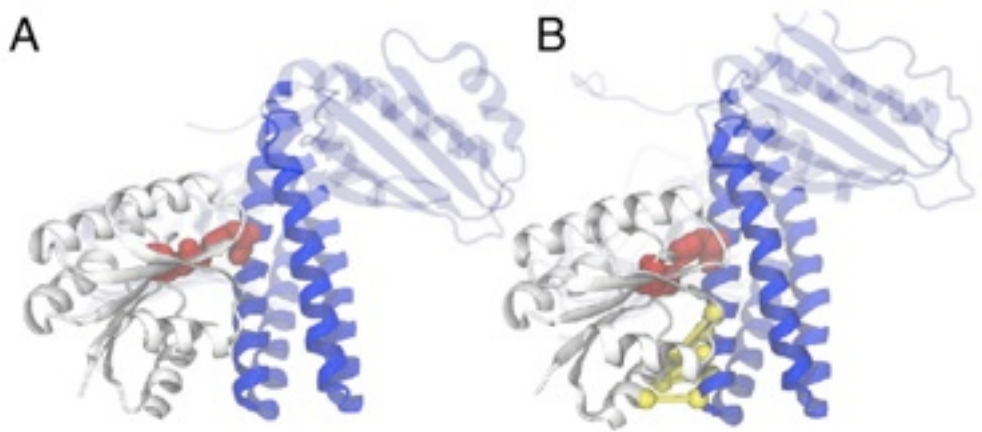




DCA identifies residue contacts

protein monomer structures

guided molecular dynamics simulations →

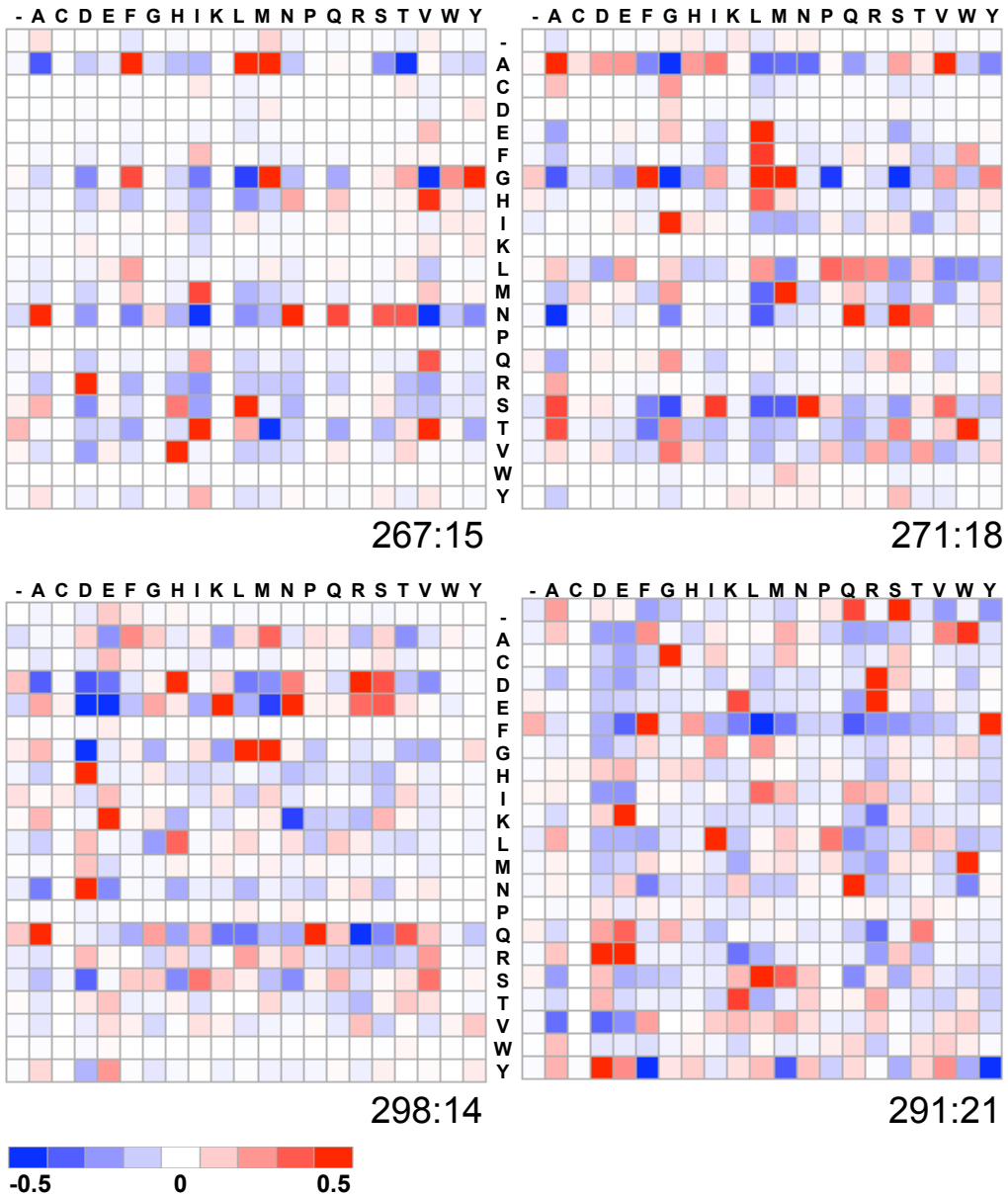


Spo0B/0F: co-crystal [Zapf *et al.* (2000)] vs. our model

[Schug, MW, Onuchic, Hwa, Szurmant, PNAS '09]

Specificity and molecular recognition

Examples for direct-coupling matrix $e_{ij}(A_i, A_j)$



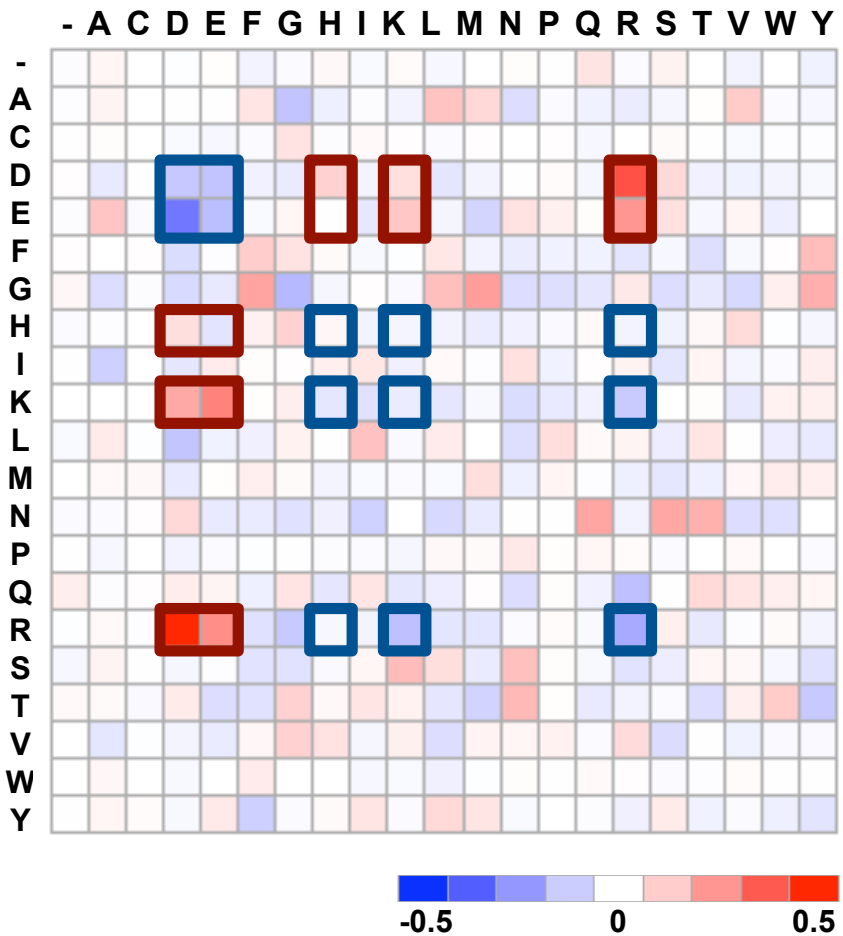
Identification of site-specific (un)favorable AA combinations

Questions:

- physical interaction mechanisms
- scoring of SK/RR pairs and interaction partner prediction
  - crosstalk between cognate TCS
  - orphan partner prediction

Physical interaction mechanisms

Average top ten direct-interaction matrices $e_{ij}(A_i, A_j)$



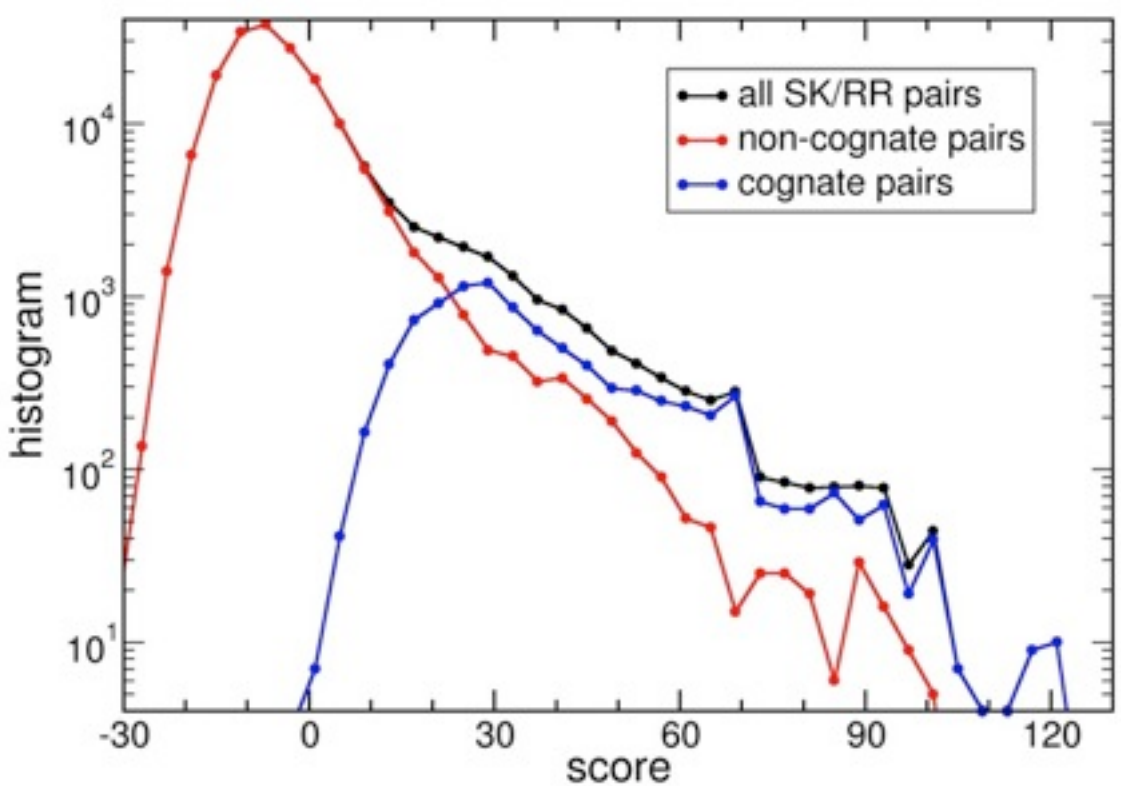
- almost symmetric
- strongest entries explainable by electrostatic interaction (p-value 3e-16)
- sub-dominant contribution: hydrophilic interaction (p-value 5e-4)
- physical mechanisms unveiled by statistical analysis

Scoring SK/RR pairs

Log-likelihood score for arbitrary SK and RR sequences

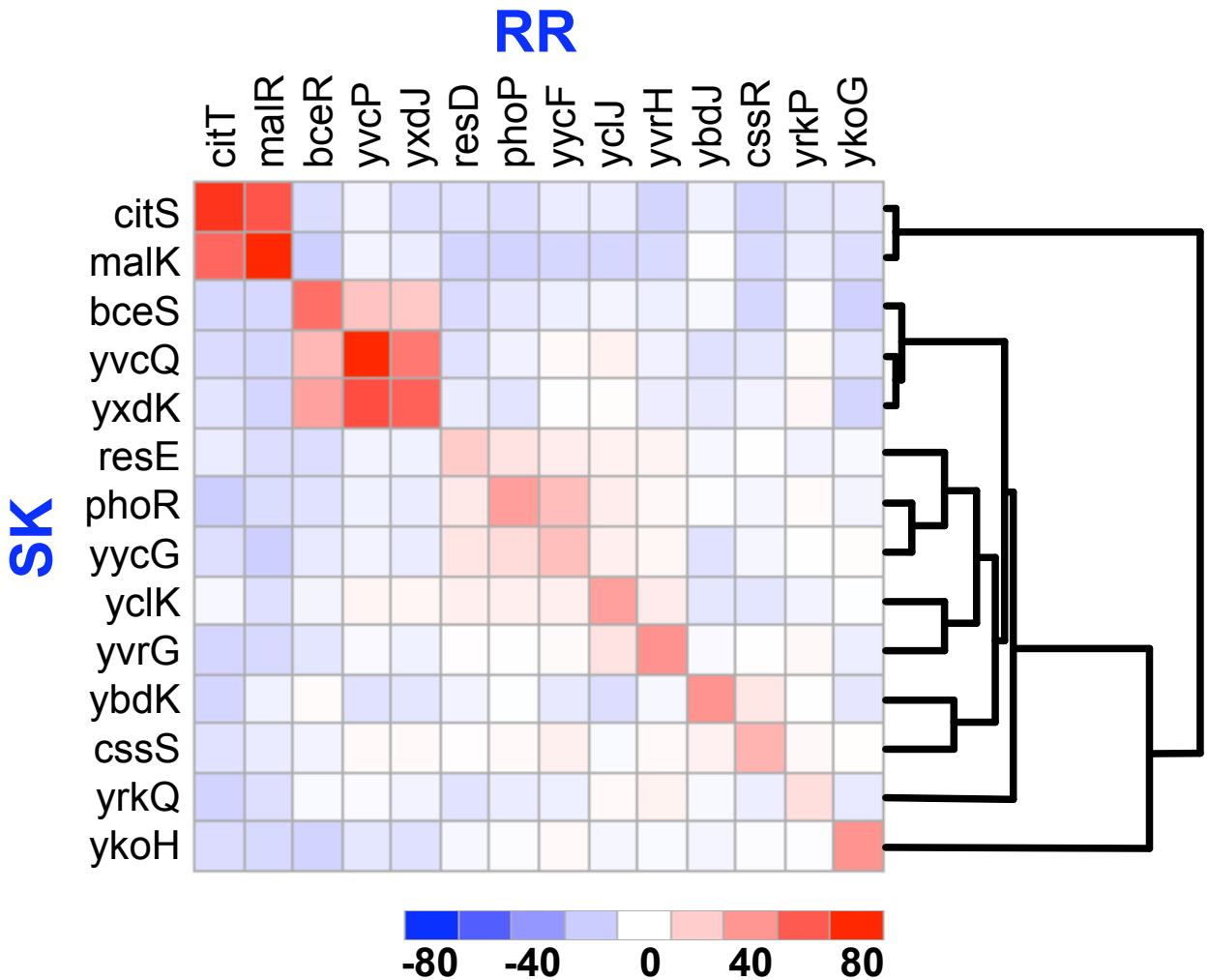
$$Score(SK, RR) = \sum_{i \in SK, j \in RR} \log \frac{P_{ij}^{(dir)}(A_i, A_j)}{f_i(A_i) f_j(A_j)}$$

- statistical model against null model of independent proteins
- test scoring all SK and RR from cognate TCS (intra species)



Crosstalk between cognate TCS

Crosstalk in *Bacillus subtilis:*

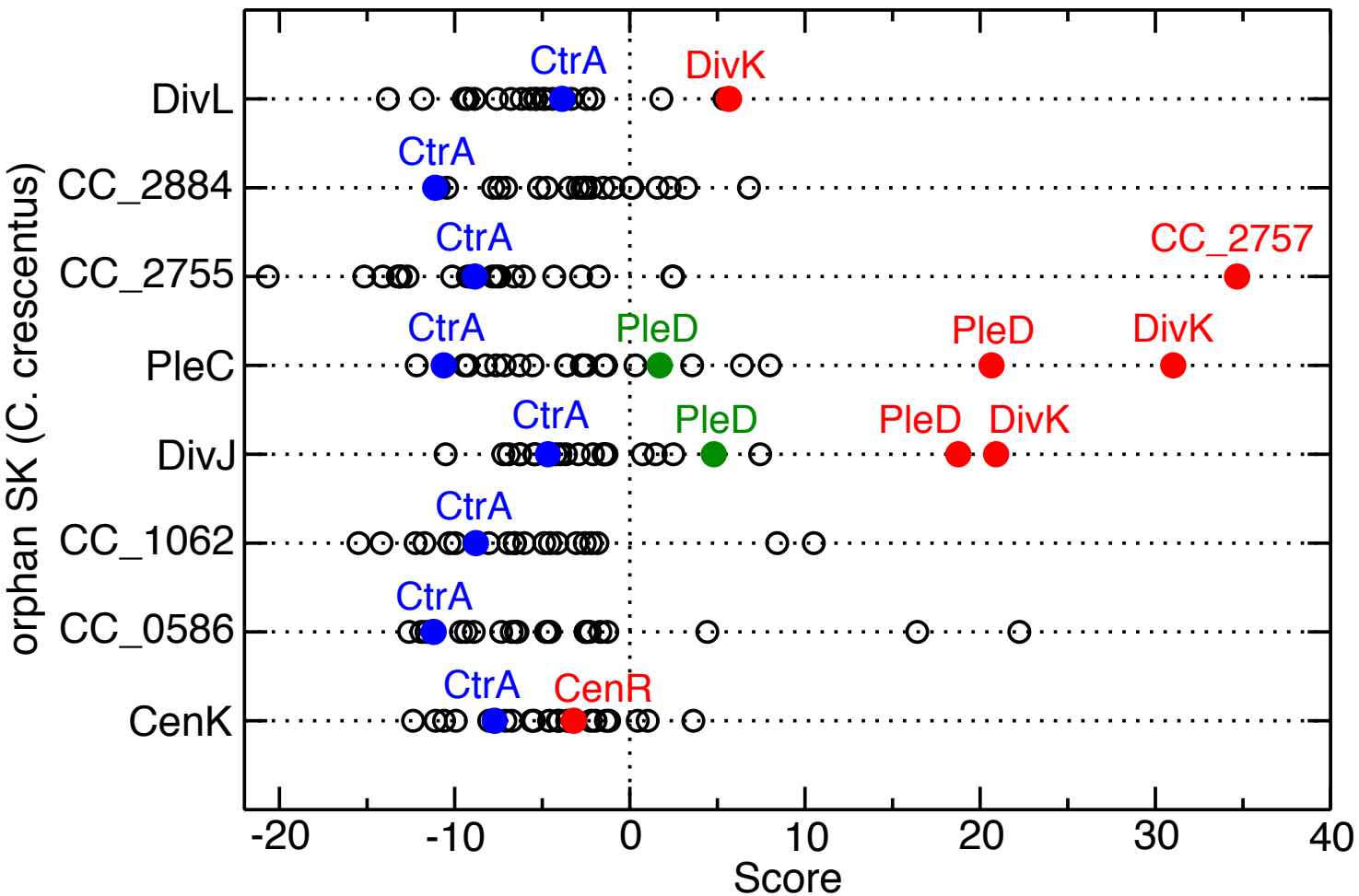


experimental evidence for crosstalk in BceRS/YvcPQ/YxdJK [Rietkoetter et al. (2008)] and PhoPR/YycFG [Howell et al. (2006)]

Orphan prediction in *C. crescentus*

Orphans = SK or RR being isolated on genome

- no obvious identification of interaction partners
- experimental results [Ohta et al. (2003), Skerker et al. (2005)]



- toward computational reconstruction of bacterial signaling networks

Outlook

- Algorithmic development
  - fast approaches for large-scale analysis
  - phylogenetic effects, finite-sample effects, sparse inference
  - integration of biological / biophysical knowledge
  - extraction of alternative co-evolutionary signals

- Residue contacts in single proteins
  - structural role of co-evolving sites
  - co-evolution without direct contact
  - structure prediction

- More complex protein-protein interactions
  - HisKA-HATPase autophosphorylation complex
  - enzymatic pathways / multi-protein complexes
  - filament formation (FtsZ, Tubulin)
  - chemotaxis receptor arrays (CheW, CheA, Mcp)

Thanks to

Torino:
Andrea Procaccini
Arianna Bertolino
Andrea Pagnani

UC San Diego:
Terry Hwa
Jose Onuchic
Bryan Lunt
Faruck Morcos

Scripps Research La Jolla:
Hendrik Szurmant
James A. Hoch
Angel Ernesto Dago

Umea University:
Alexander Schug