The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964–2014

IAEA

2584-8

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

Direct Coupling Analysis of residue coevolution in proteins

Martin Weigt
Université Pierre et Marie Curie
Paris

Direct Coupling Analysis of residue coevolution in proteins

Martin Weigt

Laboratoire de Biologie Computationelle et Quantitative

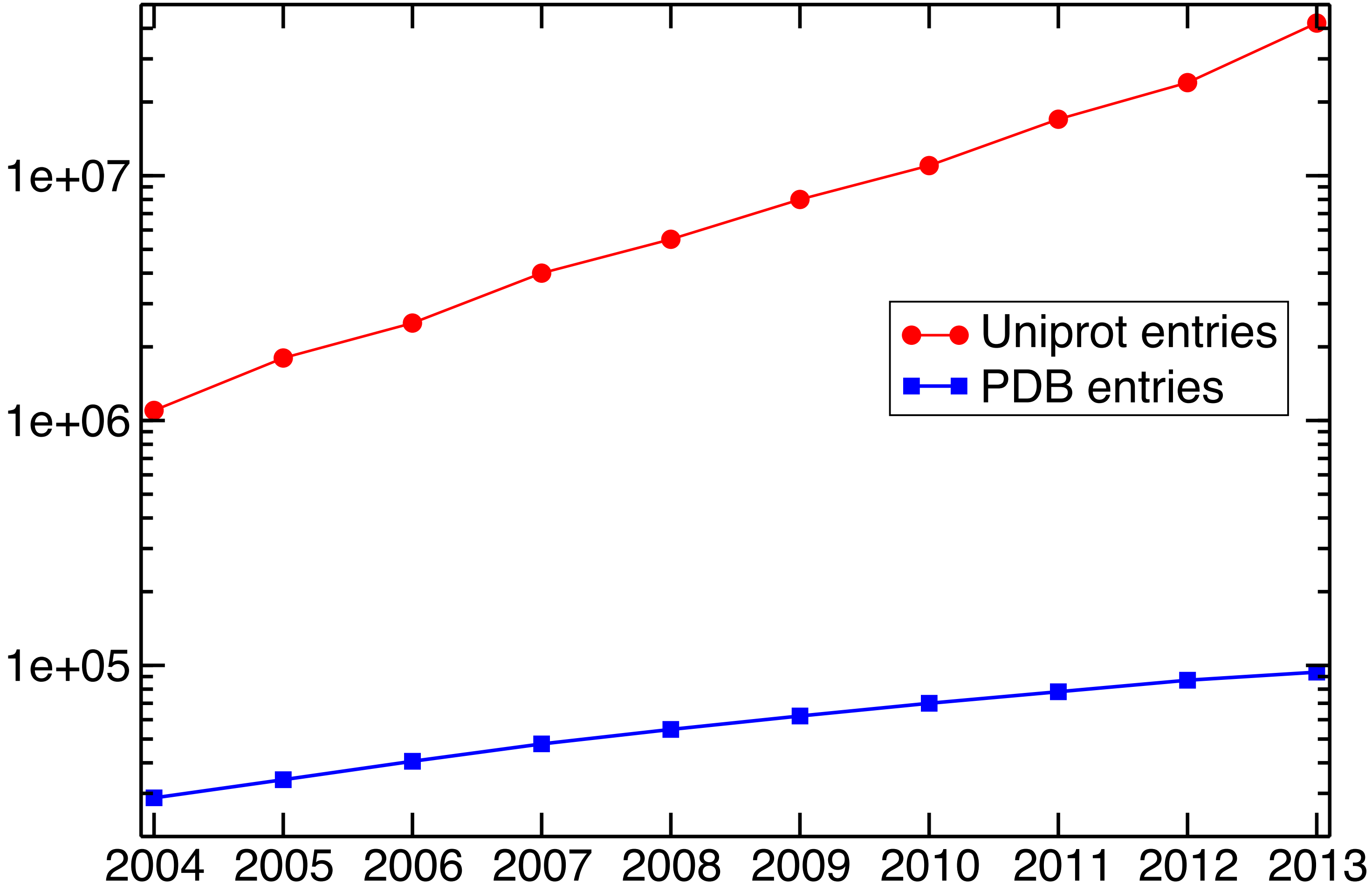
Université Pierre et Marie Curie

What is the information in

```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
-----RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQFR----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
---------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPTEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERFYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
----PESEGVTGAPTSRWYYDQTDMQCKQFTYNGRRGNQNNFLTQEDCAATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```

?

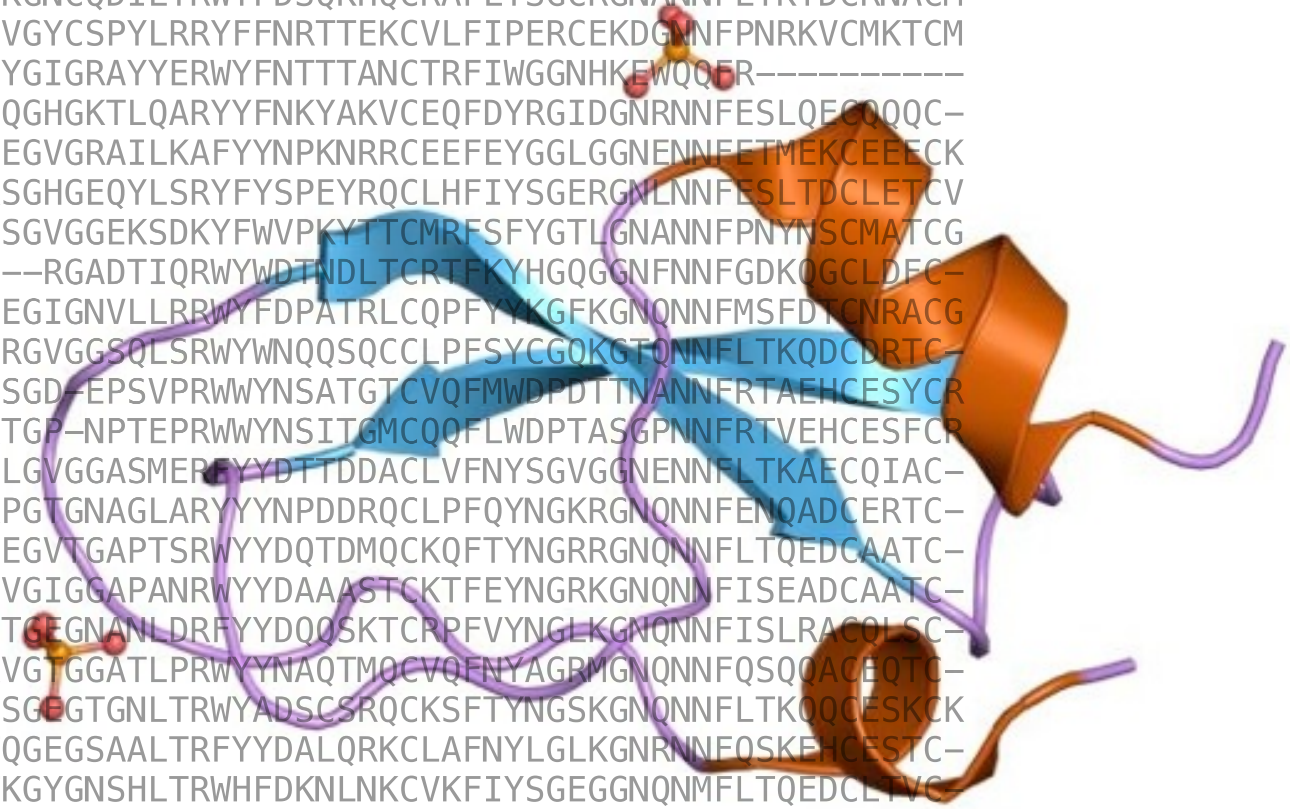
Why?



...but function relies on structure!

There is information in

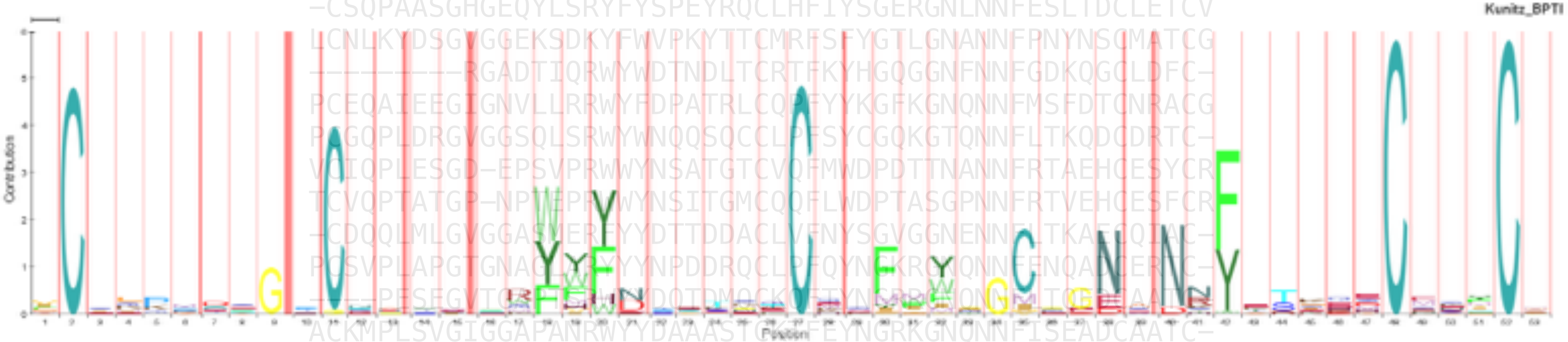
```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
-----RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQFR----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
---------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPTEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERFYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
----PESEGVTGAPTSRWYYDQTDMQCKQFTYNGRRGNQNNFLTQEDCAATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKTFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```



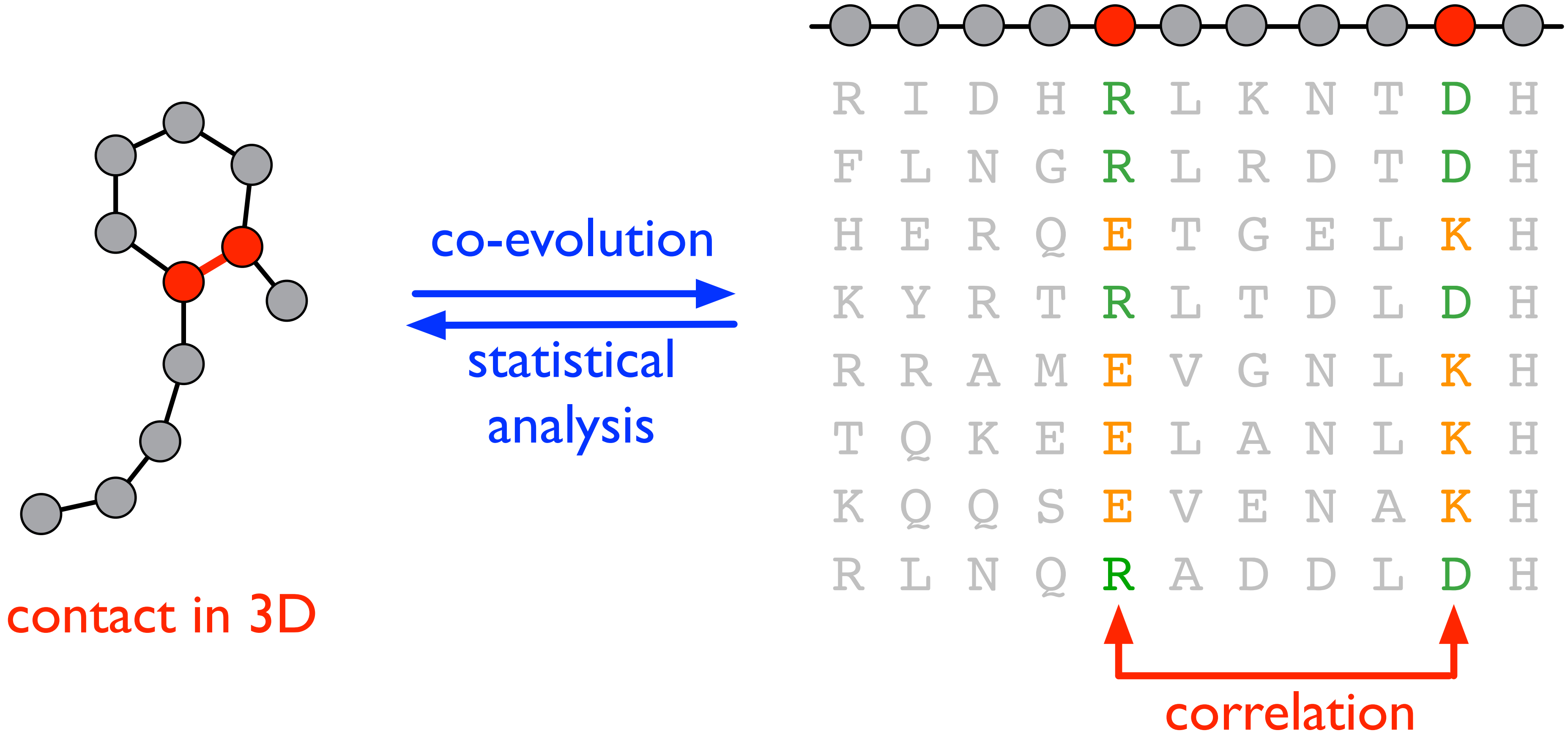
!

What we know to do so far…

```
ACSLPKVQGPCSGKHSYYYFNSANQQCETFVYGGCLGNTNRFATIEECNARC-
VCLLPKSAGPCTGFTKKWYFDVDRNRCEEFQYGGCYGTNNRFDSLEQCQGTC-
VCAMPPDAGVCTNYTPRWFFNSQTGQCEQFAYGSCGGNENNFFDRNTCERKCM
TCSLSPSPGTCGPGVFKYHYNPQTQECESFEYLGCDGNSNTFASRAECENYCG
-CHTEHSSGACPGAVTMFYHDPRTKKCTPFTFLGCGGNSNKFDTRPQCERFCK
PCMLPSDKGNCQDILTRWYFDSQKHQCRAFLYSGCRGNANNFLTKTDCRNACM
-----RLVGYCSPYLRRYFFNRTTEKCVLFIPERCEKDGNNFPNRKVCMKTCM
PCSLKEDYGIGRAYYERWYFNTTTANCTRFIWGGNHKEWQQFR----------
PCKQDLDQGHGKTLQARYYFNKYAKVCEQFDYRGIDGNRNNFESLQECQQQC-
-CFLKPDEGVGRAILKAFYYNPKNRRCEEFEYGGLGGNENNFETMEKCEEECK
-CSQPAASGHGEQYLSRYFYSPEYRQCLHFIYSGERGNLNNFESLTDCLETCV
LCNLKYDSGVGGEKSDKYFWVPKYTTCMRFSFYGTLGNANNFPNYNSCMATCG
---------RGADTIQRWYWDTNDLTCRTFKYHGQGGNFNNFGDKQGCLDFC-
PCEQAIEEGIGNVLLRRWYFDPATRLCQPFYYKGFKGNQNNFMSFDTCNRACG
PCGQPLDRGVGGSQLSRWYWNQQSQCCLPFSYCGQKGTQNNFLTKQDCDRTC-
VCIQPLESGD-EPSVPRWWYNSATGTCVQFMWDPDTTNANNFRTAEHCESYCR
TCVQPTATGP-NPWEPRWWYNSITGMCQQFLWDPTASGPNNFRTVEHCESFCR
-CDQQLMLGVGGASMERFYYDTTDDACLVFNYSGVGGNENNFLTKAECQIAC-
PCSVPLAPGTGNAGLARYYYNPDDRQCLPFQYNGKRGNQNNFENQADCERTC-
----PESFGVTGR[illegible]DQTDMQCKQFIYNGRRGNQNN[illegible]AATC-
ACKMPLSVGIGGAPANRWYYDAAASTCKPFEYNGRKGNQNNFISEADCAATC-
VCNLPMSTGEGNANLDRFYYDQQSKTCRPFVYNGLKGNQNNFISLRACQLSC-
ICQQPMAVGTGGATLPRWYYNAQTMQCVQFNYAGRMGNQNNFQSQQACEQTC-
PCSLPMFSGEGTGNLTRWYADSCSRQCKSFTYNGSKGNQNNFLTKQQCESKCK
PCEEEMTQGEGSAALTRFYYDALQRKCLAFNYLGLKGNRNNFQSKEHCESTC-
TCELPMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLTVC-
TCELTMTKGYGNSHLTRWHFDKNLNKCVKFIYSGEGGNQNMFLTQEDCLSVC-
RCHLPPAVGYGKQRMRRFYFDWKTDACHELQYSGIGGNENIFMDYEQCERVCR
-CMESLDRGSCEAMSNRYYFNKRARQCKGFHYTGCGKSGNNFLTKEECQTKC-
PCQQPLQRGNCSQRIPLFYYNIHNHKCRKFMYRGCNGNENRFSNRRQCQAKCG
```



Residue contacts induce residue co-evolution



R	I	D	H	R	L	K	N	T	D	H
F	L	N	G	R	L	R	D	T	D	H
H	E	R	Q	E	T	G	E	L	K	H
K	Y	R	T	R	L	T	D	L	D	H
R	R	A	M	E	V	G	N	L	K	H
T	Q	K	E	E	L	A	N	L	K	H
K	Q	Q	S	E	V	E	N	A	K	H
R	L	N	Q	R	A	D	D	L	D	H

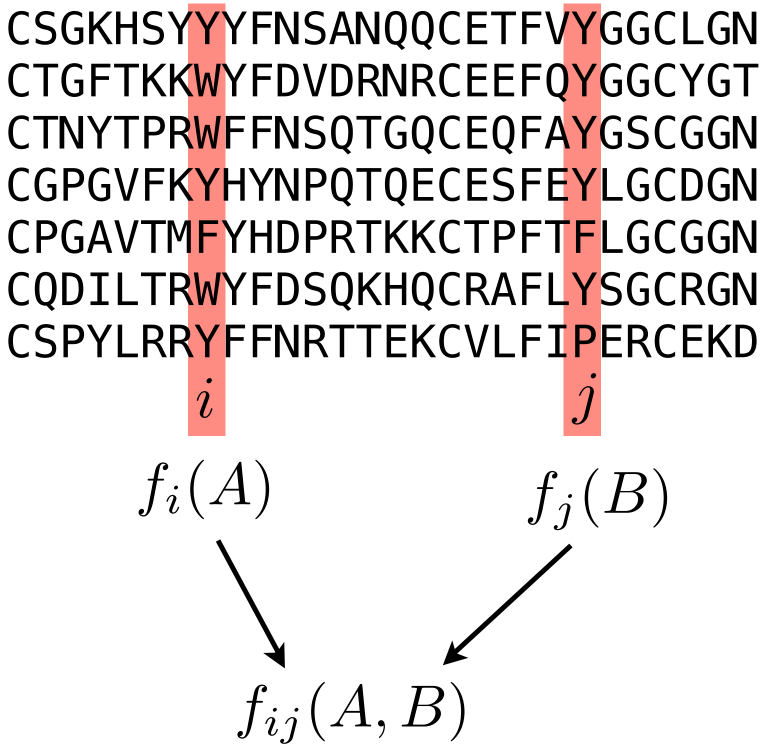
Inverse question:

- 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): $D = \{A_i^m \,|\, i = 1, \ldots, L; m = 1, \ldots, M\}$



```
CSGKHSYYYFNSANQQCETFVYGGCLGN
CTGFTKKWYFDVDRNRCEEFQYGGCYGT
CTNYTPRWFFNSQTGQCEQFAYGSCGGN
CGPGVFKYHYNPQTQECESFEYLGCDGN
CPGAVTMFYHDPRTKKCTPFTFLGCGGN
CQDILTRWYFDSQKHQCRAFLYSGCRGN
CSPYLRRYFFNRTTEKCVLFIPERCEKD
```

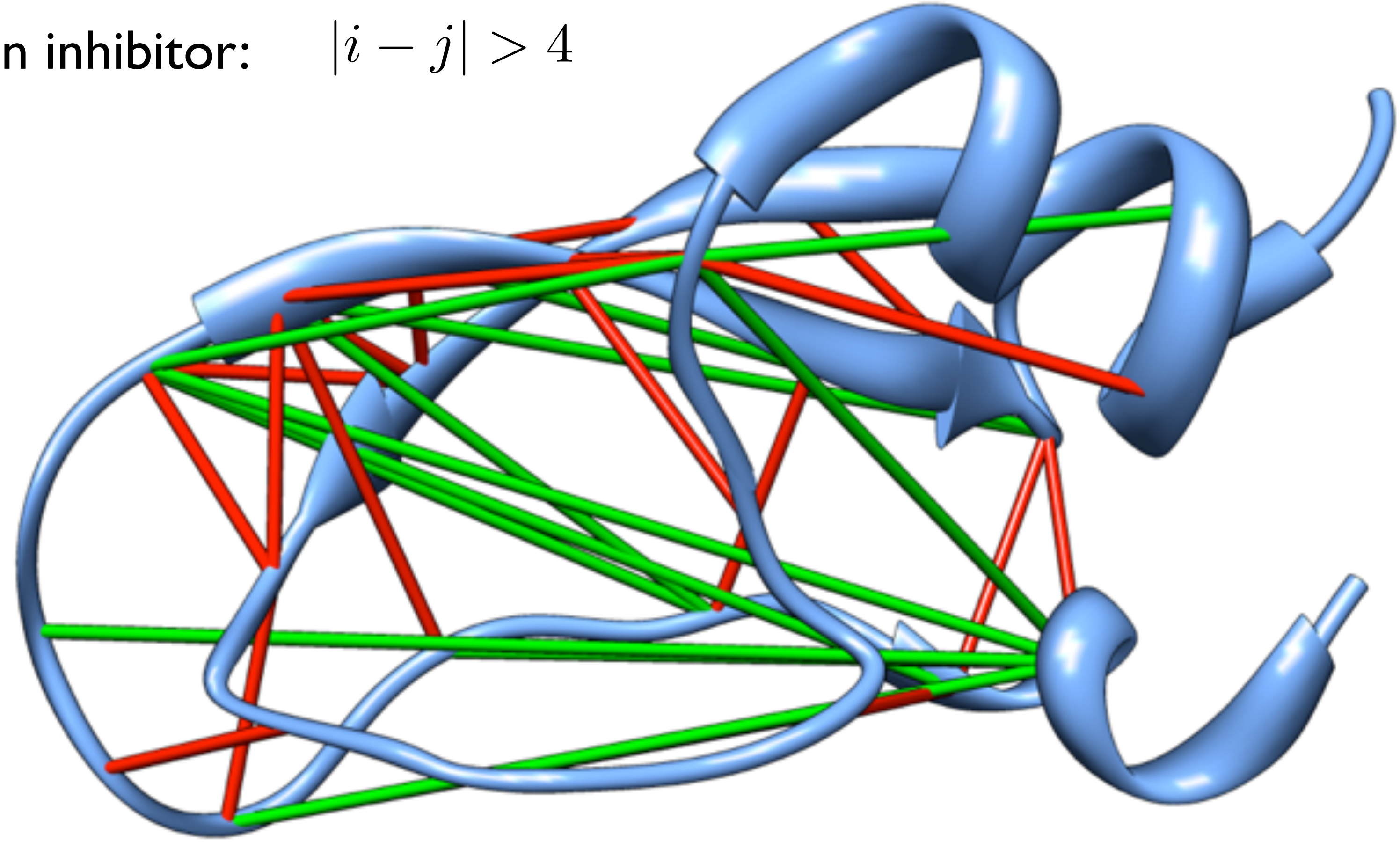
Mutual information measures pair correlation

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

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

Correlations vs. residue contacts

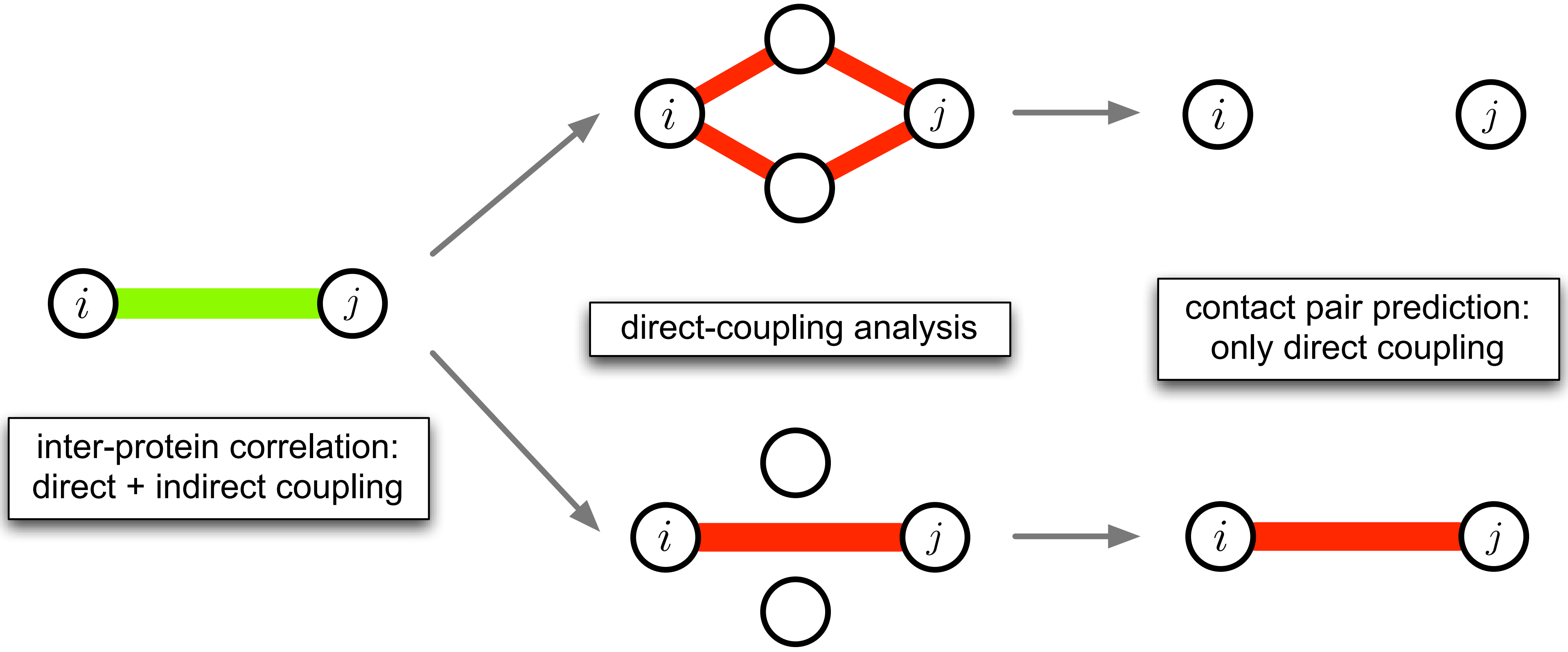
Trypsin inhibitor: $|i-j| > 4$



🟥 contact

🟩 no contact

Correlation is not coupling



inter-protein correlation:
direct + indirect coupling

direct-coupling analysis

contact pair prediction:
only direct coupling

- correlations are generated by network of direct couplings
- disentangle direct and indirect couplings: $P(A_1, \ldots, A_L)$
- 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

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

- maximum-entropy model:

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

➡ disordered 21-states Potts model / Markov random field

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

direct coupling of residues *i* and *j*

[MW, White, Szurmant, Hoch, Hwa, PNAS '09]
[Burger, van Nimwegen, PLoS Comp Biol '10]
[Morcos, Pagnani,..., MW, PNAS '11]
[Balakrishnan et al., Proteins '11]
[Jones et al., Bioinformatics '12]

Direct coupling analysis (DCA)

- minimal (maximum-entropy) modeling of sequence statistics including residue covariation = Markov random field / disordered Potts model

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

direct coupling of residues *i* and *j*

- our approximations
  - the first: loopy belief propagation [Weigt et al, PNAS '09]
  - the fastest: naive mean-field [Morcos et al, PNAS '11]
  - the most accurate: pseudo-likelihood max [Ekeberg et al, Phys Rev E '13]
  - less overfitting: dimensional reduction [Cocco et al, PLoS CB '13]
- and by others
  - MCMC sampling [Lapedes et al, LANL preprint '02]
  - Bayesian networks [Burger et al, PLoS Comp Biol '10]
  - pseudo-likelihood maximization [Balakrishnan et al., Proteins '11]
  - sparse inverse covariance (PSICOV) [Jones et al., Bioinformatics '12]
  - meta classification [Skwark et al., Bioinformatics '13]

Direct coupling analysis

- Boltzmann-machine learning:
  - start with initialized fields/couplings
  - calculate

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

  - update couplings

$$\Delta e_{ij}(A, B) = \varepsilon \left[f_{ij}(A, B) - P_{ij}(A, B) \right]$$

  - iterate until sufficiently precise fitting

➡ exact calculation requires exponential time ~ 21^L

➡ approximations needed

Direct coupling analysis

need to estimate marginals

$$
\begin{aligned}
P_1(A_1) &= \sum_{\{A_j|j>1\}} P(A_1, ..., A_L) \\
&= \sum_{\{A_j|j>1\}} P(A_1|A_2, ..., A_L)P(A_2, ..., A_L) \\
&\sim \sum_{\{A_j|j>1\}} \exp\left\{h_1(A_1) + \sum_{j=2}^{M} e_{1j}(A_1, A_j)\right\} P(A_2, ..., A_L)
\end{aligned}
$$

- MCMC - slow (?)
- pseudo-likelihood maximization - average over data

$$
P_1(A_1) = \frac{1}{M} \sum_{m=1}^{M} P(A_1|A_2^m, ..., A_L^m)
$$

[Balakrishnan et al., Proteins '11]
[Ekeberg et al., Phys Rev E '13]

Direct coupling analysis

- Mean-field approximation:
  - mean-field equation for single-site marginal probabilities

$$P_i(A) \sim \exp\left\{ h_i(A) + \sum_{j\neq i} \sum_B e_{ij}(A,B) P_j(B) \right\}$$

  - correlations from linear response

$$\frac{\partial P_i(A)}{\partial h_j(B)} = C_{ij}(A,B) = P_{ij}(A,B) - P_i(A)P_j(B)$$

lead to explicit equation for couplings

$$e_{ij}(A,B) = \left[C^{-1}\right]_{ij}(A,B)$$

➡ couplings estimated in time $\mathcal{O}(21^3 N^3)$

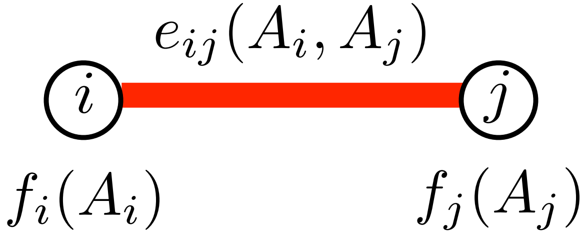
➡ more complicated approximations (Bethe-Peierls, Thouless-Anderson-Palmer) do not improve performance on biological sequence data

[Morcos, Pagnani,..., MW PNAS '11]

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

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

➡ average-product corrected Frobenius norm

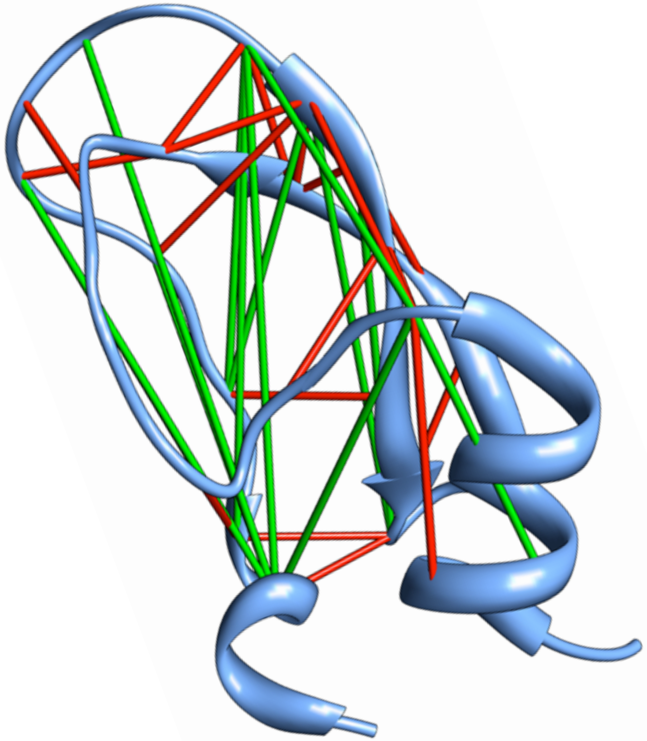
$$F_{ij} = \sum_{AB} |e_{ij}(A, B)|^2$$

$$\tilde{F}_{ij} = F_{ij} - \frac{F_{\cdot j} F_{i \cdot}}{F_{\cdot\cdot}}$$

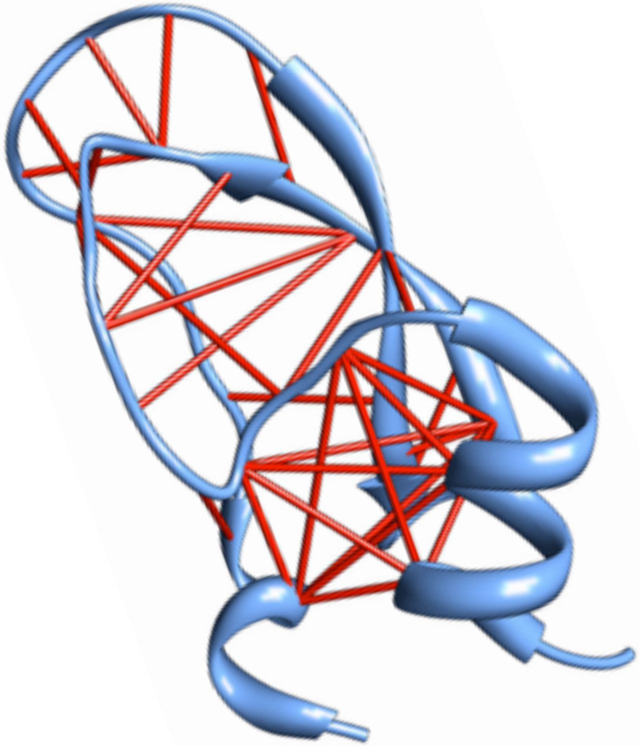
DCA strongly improves contact prediction

Trypsin inhibitor: $|i-j|>4$

strongest correlations



strongest direct couplings

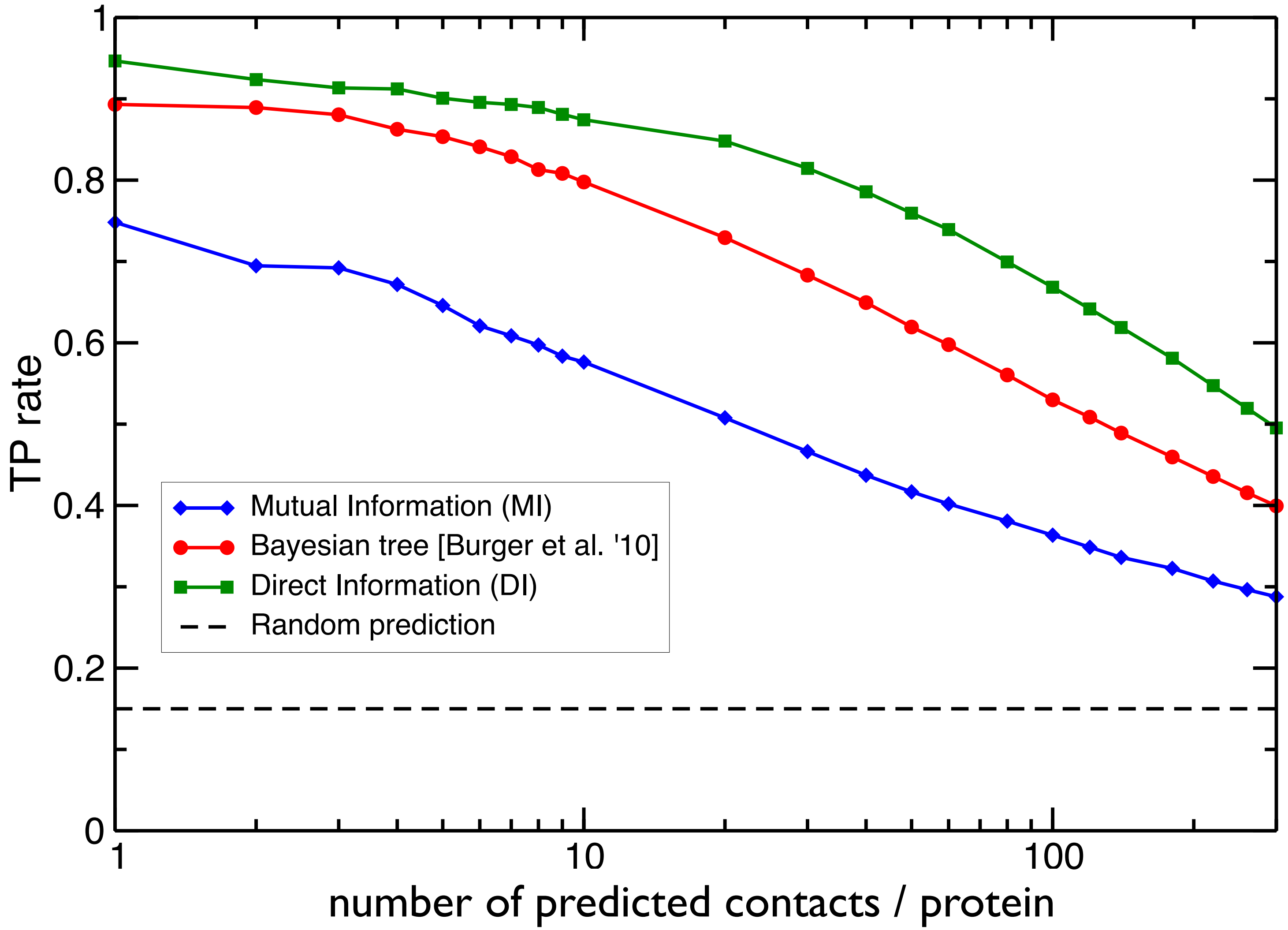


🟥 contact

🟩 no contact

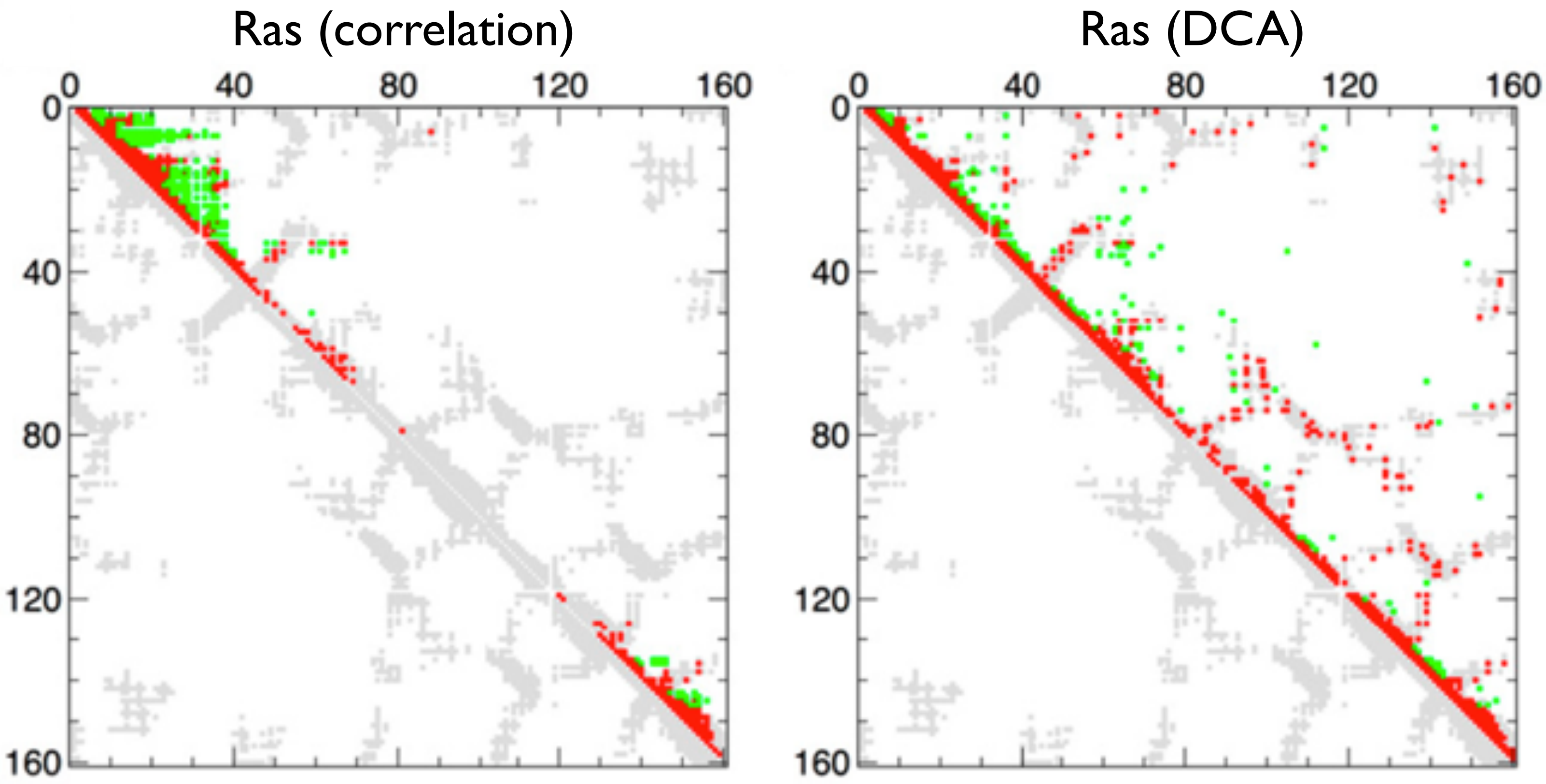
Couplings vs. residue contacts

Comparison for 131 abundant protein families: $|i-j| \geq 5$



DCA strongly improves contact prediction!

Not all contacts co-vary, but...



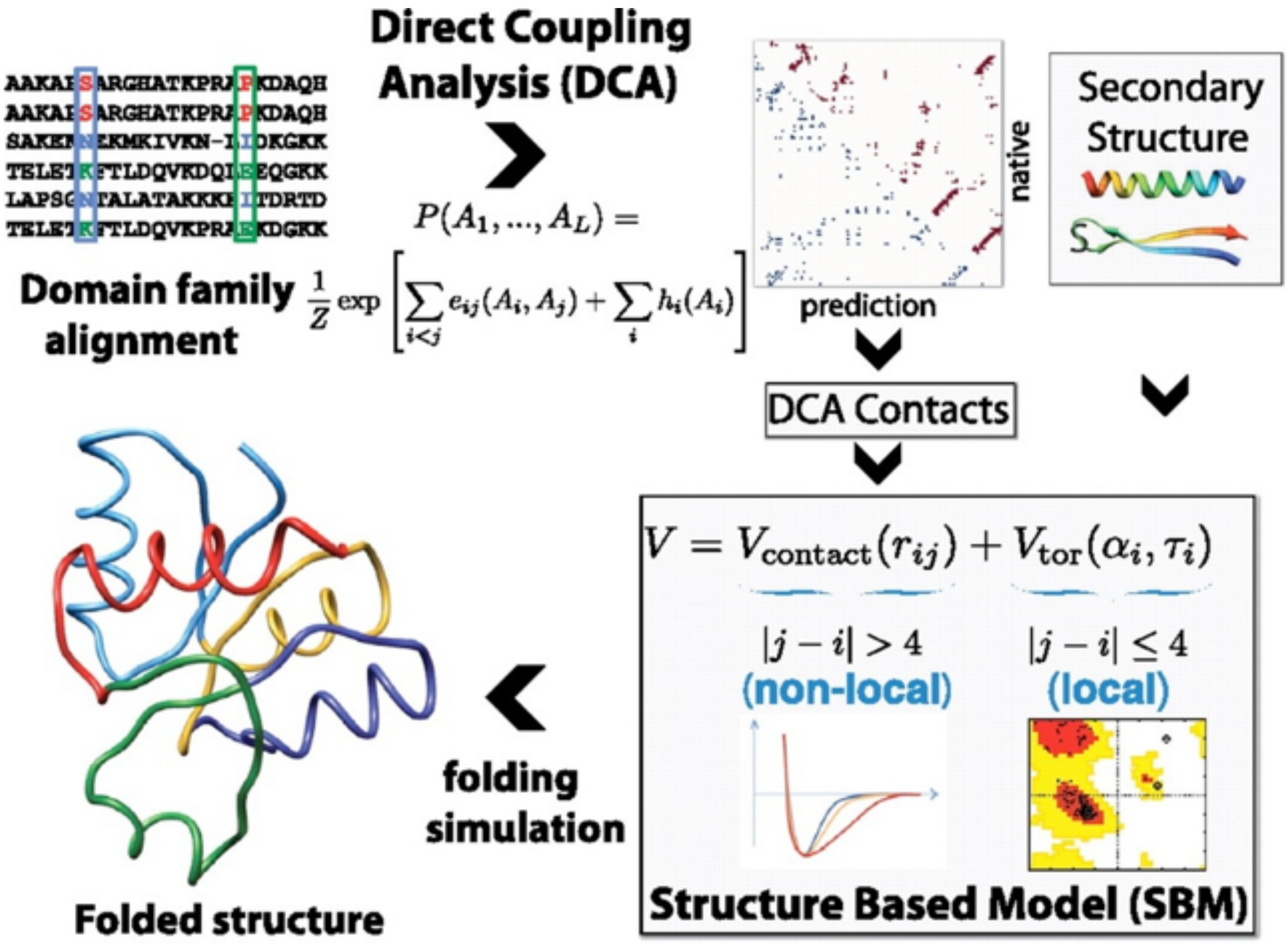
...coevolution can guide complex assembly

[Schug, MW, Onuchic, Hwa, Szurmant, PNAS '09]
[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS '12]

and protein structure prediction

[Marks et al., PLoS ONE '11]
[Sadowski et al., Comp Biol Chem '11]
[Sulkowska, Morcos, MW, Hwa, Onuchic, PNAS '12]
[Hopf et al., Cell '12]
[Nugent, Jones, PNAS '12]

From contacts to 3D structure



$$P(A_1, ..., A_L) = \frac{1}{Z} \exp\left[\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\right]$$

$$V = V_{\text{contact}}(r_{ij}) + V_{\text{tor}}(\alpha_i, \tau_i)$$

$|j-i| > 4$ (non-local)

$|j-i| \leq 4$ (local)

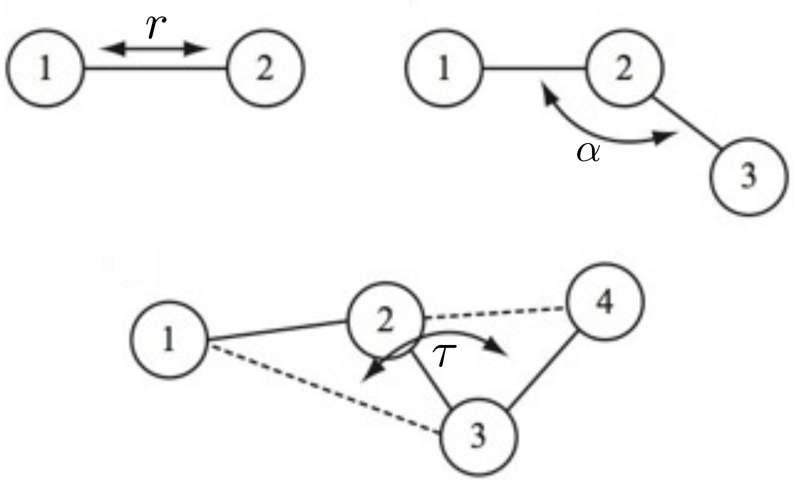
[Sulkowska, Morcos, MW, Hwa, Onuchic, PNAS ’12]

From contacts to 3D structure

ab initio protein folding simulations:

- molecular-dynamics simulations of structure-based models (Go-models):

$$V = V_{bond} + V_{torsion} + V_{contact}$$



with

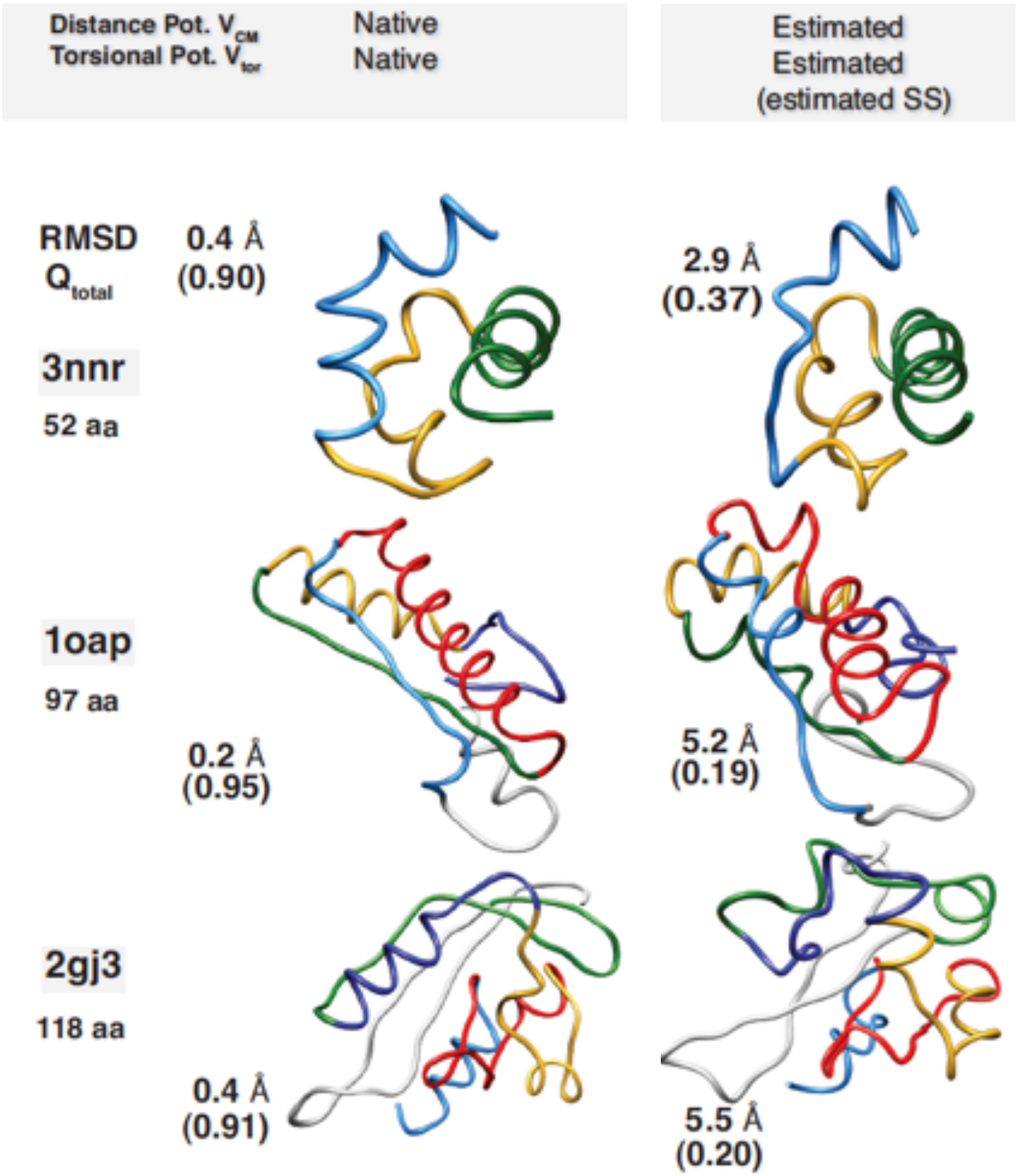
$$V_{bond} = k_b \sum_{bonds} (r - r_0)^2$$

$$V_{torsion} = k_a \sum_{angles} (\alpha - \alpha_0)^2 + k_d \sum_{dihedral} [1 - \cos(\tau - \tau_0)] + \frac{1}{2}[1 - \cos 3(\tau - \tau_0)]$$

$$V_{contact} = \varepsilon_c \sum_{contacts} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

use only DCA contacts

...global protein structure defined



Left:

- DCA contacts
- native distances
- native torsional angles

▸ lower RMSD bound

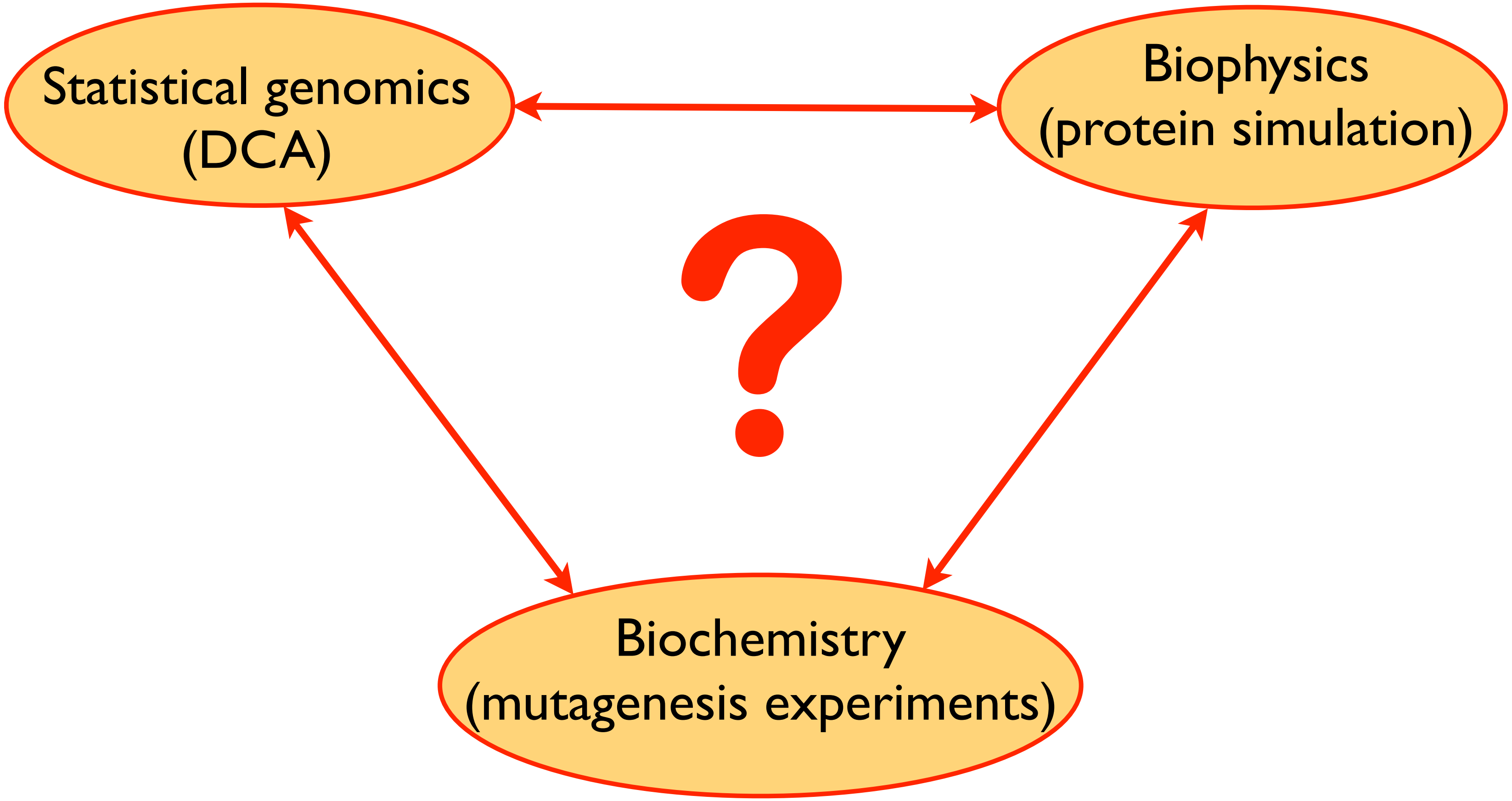
Right:

- DCA contacts
- statistical distance potential
- statistical torsional potential

▸ upper RMSD bound

[Sulkowska, Morcos, MW, Hwa, Onuchic, PNAS '12]

Can we use this to actually predict and test **unknown protein structures?**

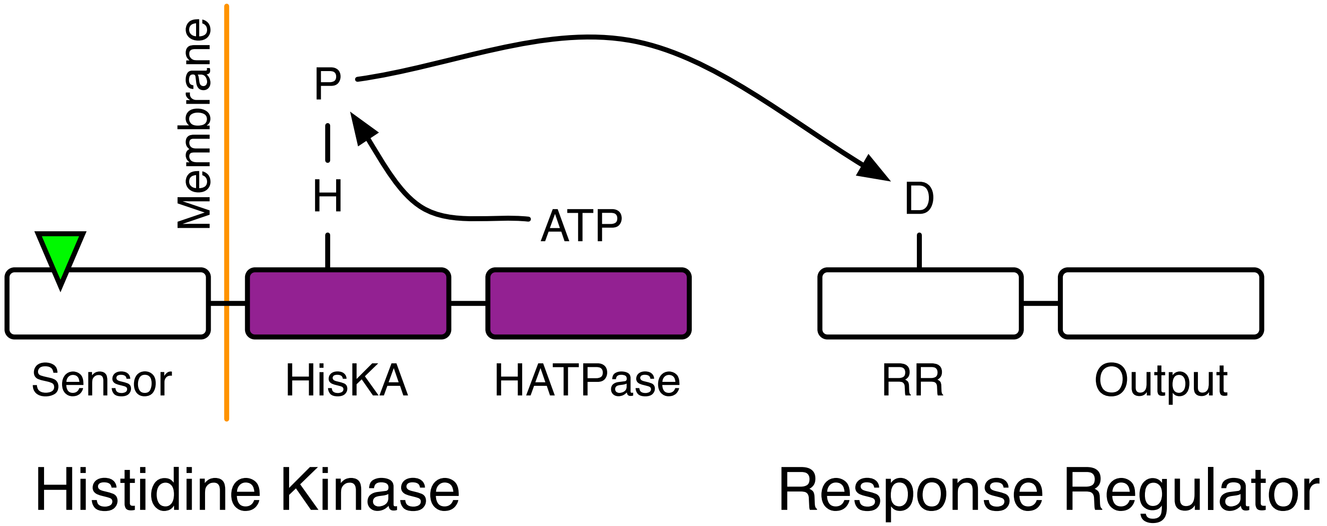


[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS ’12]

Histidine-kinase auto-phosphorylation complex

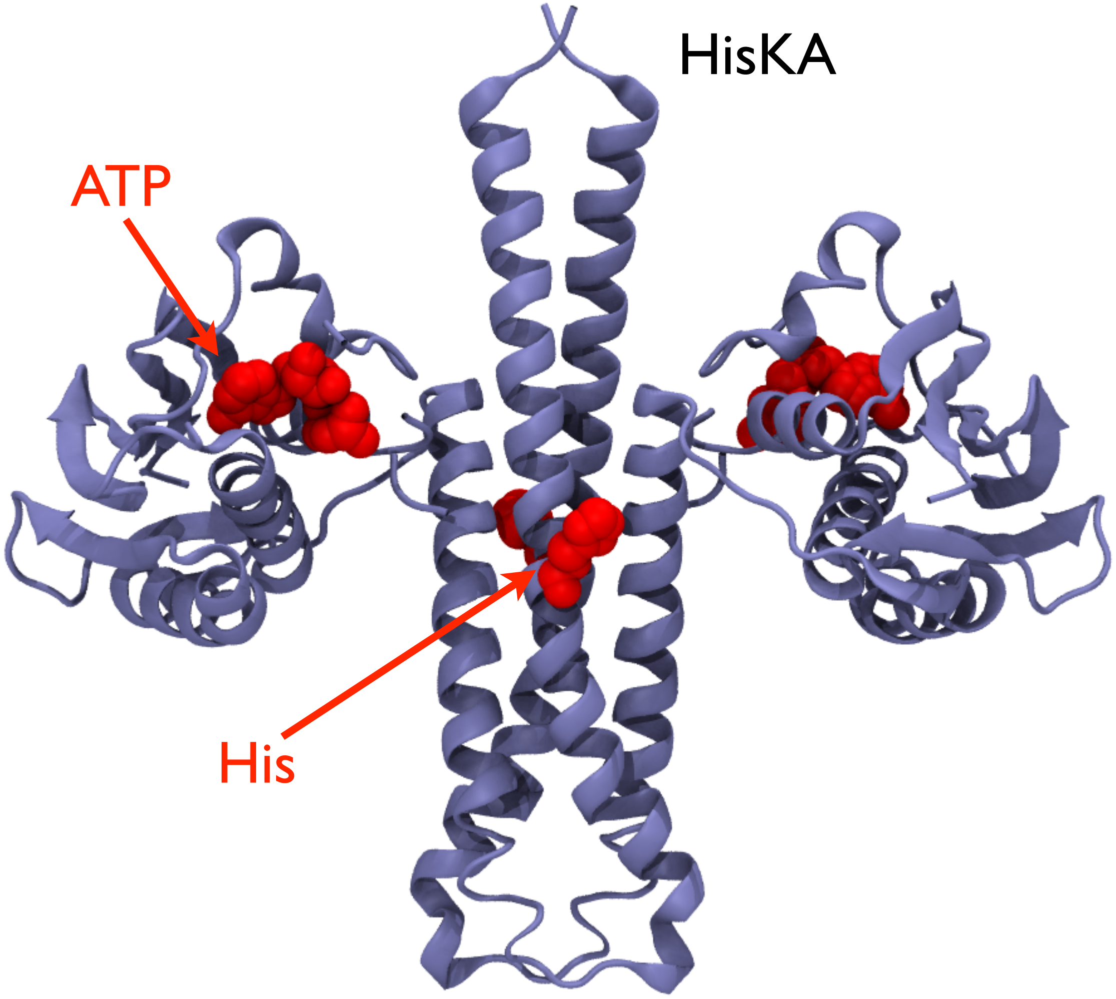
Two-component signaling system

- most common signaling system in bacteria



- on average ~20 TCS / bacterial genome
- >13,000 sequences of proteins with HisKA/HATPase domains (back in 2008)

Histidine-kinase auto-phosphorylation complex



HK853
(*Thermotoga maritima*)

inactive configuration:
ATP distant from His

[Marina, Waldburger, Hendrickson, *EMBO J.* (2005)]
[Casino, Rubio, Marina, *Cell* (2009)]
[Bick et al., *J. Mol. Biol.* (2009)]

DCA results

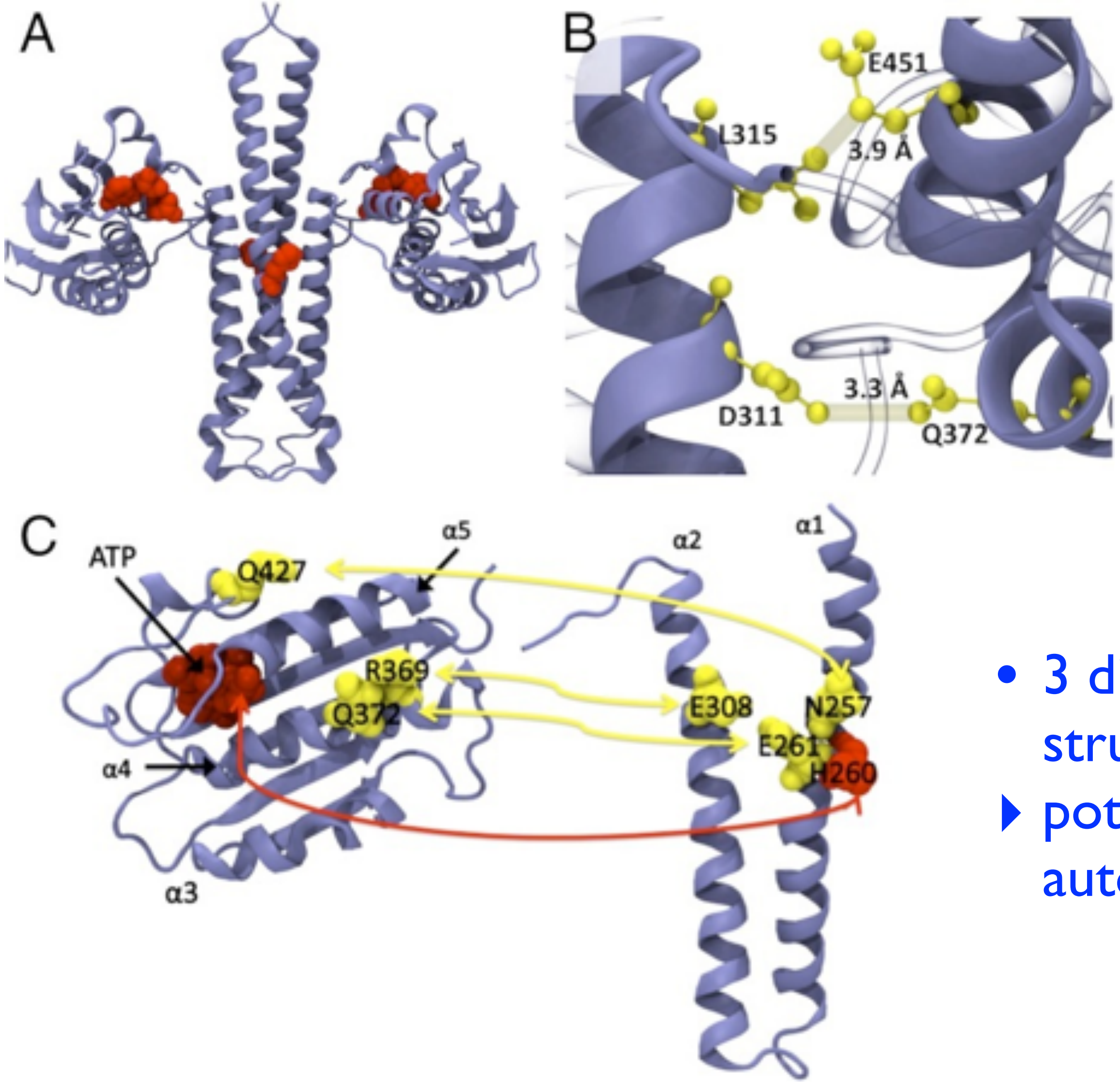
Rank	Res 1	Res 2	d/Å	Domain
1	388	392	4.6	22
2	268	272	3.2	11
3	268	298	3.2	11
4	365	456	3.7	22
5	385	392	3.9	22
6	310	311	1.3	11
7	311	312	1.3	11
8	303	307	3.0	11
9	261	372	14.5	12
10	420	421	1.3	22
11	272	298	6.9	11
12	369	372	2.9	22
13	375	379	2.7	22
14	310	312	3.2	11
15	429	431	3.9	22
16	251	255	2.9	11
17	257	272	20.5	11
18	379	383	2.8	22
19	420	429	3.7	22
20	431	432	1.3	22
21	385	388	6.4	22
22	251	252	1.3	11
23	250	251	1.3	11
24	308	369	8.0	12
25	298	310	14.8	11
26	369	455	7.0	22
27	383	384	1.3	22
28	426	429	3.1	22
29	420	431	3.8	22
30	451	455	2.9	22
31	251	268	23.6	11
32	315	451	3.6	12
33	257	427	12.7	12
34	372	375	3.4	22
35	369	456	4.7	22
36	311	372	3.3	12

First 36 DCA predictions

- 31 intra-domain pairs
  - 28 in contact
  - 3 distant
  - ▸ >90% TP rate
- 5 inter-domain pairs
  - 2 in contact
  - 3 distant
  - ▸ predicted contacts in active structure

DCA inter-domain contact prediction

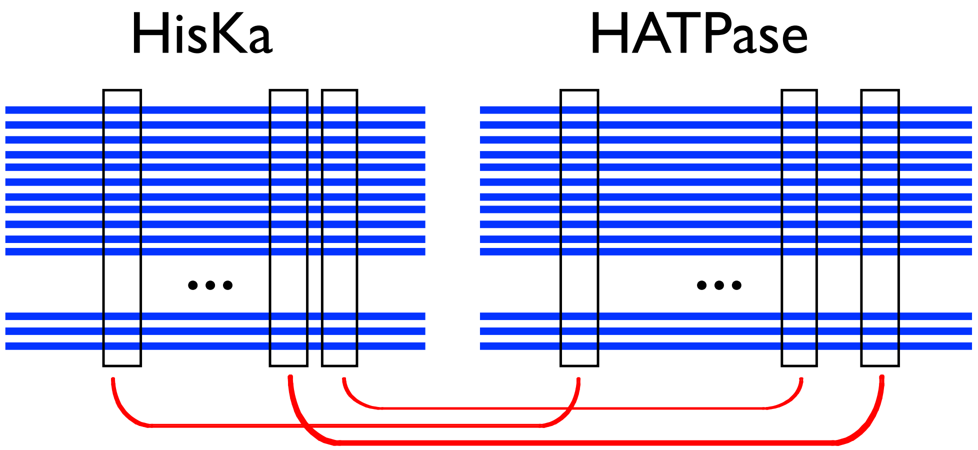
DCA predicts 5 inter-domain pairs



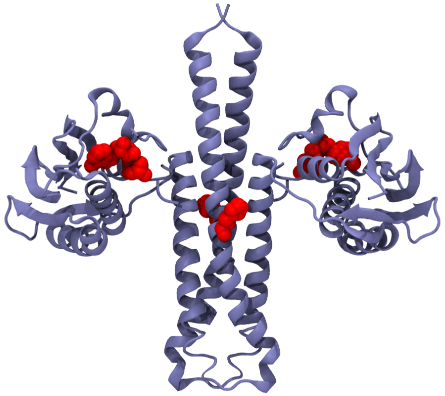
- 2 contacts in the inactive structure

- 3 distant pairs in the inactive structure
  - potential contacts in the auto-phosphorylation complex

DCA-guided molecular dynamics simulations

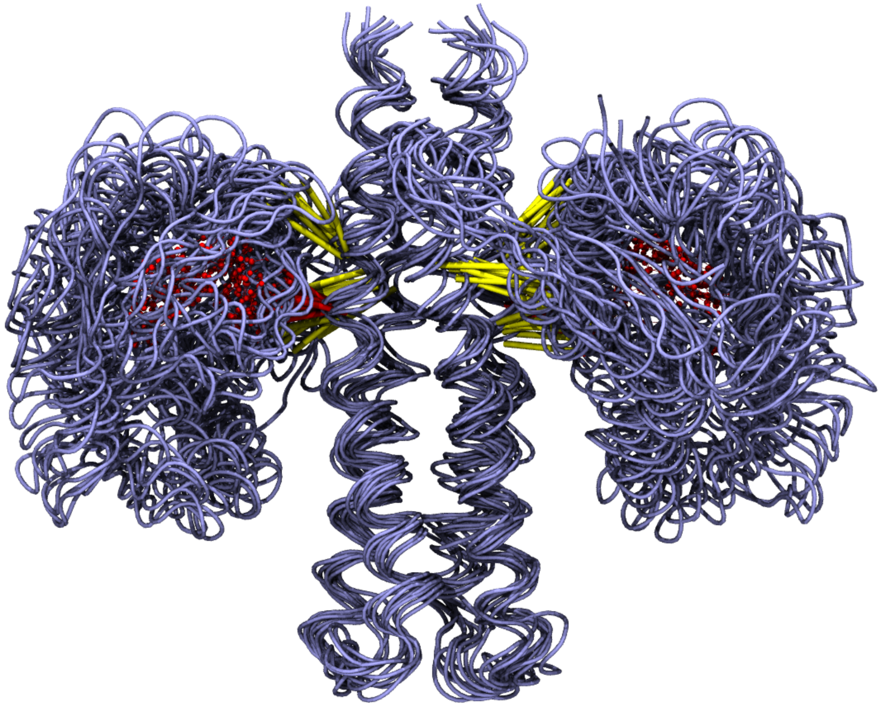


DCA-predicted inter-domain contacts

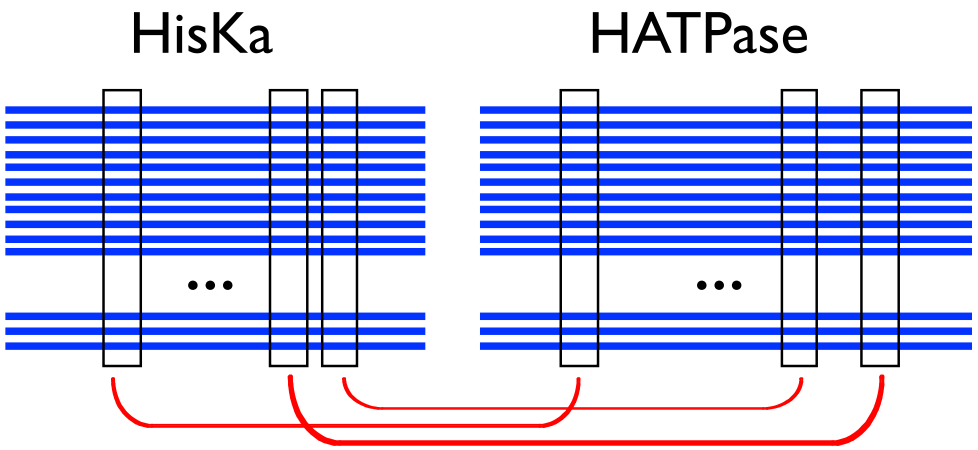


inactive protein structure

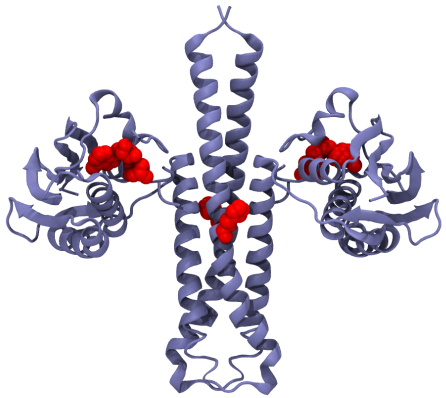
guided MD simulations
of coarse-grained model
(Go model)



DCA-guided molecular dynamics simulations



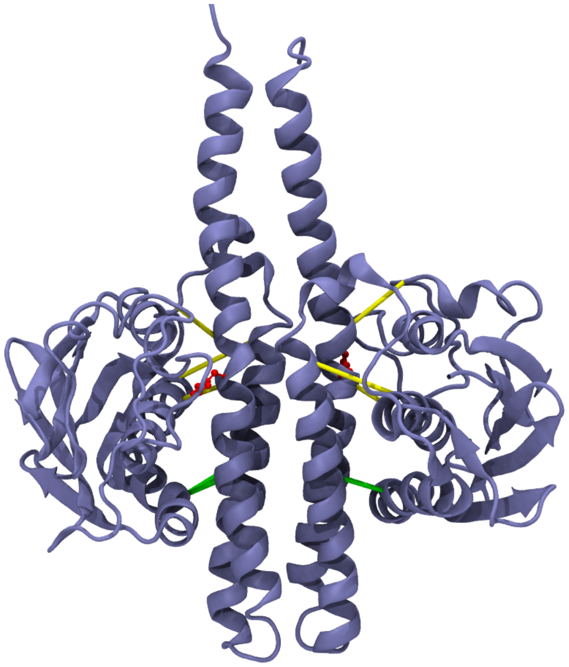
DCA-predicted inter-domain contacts



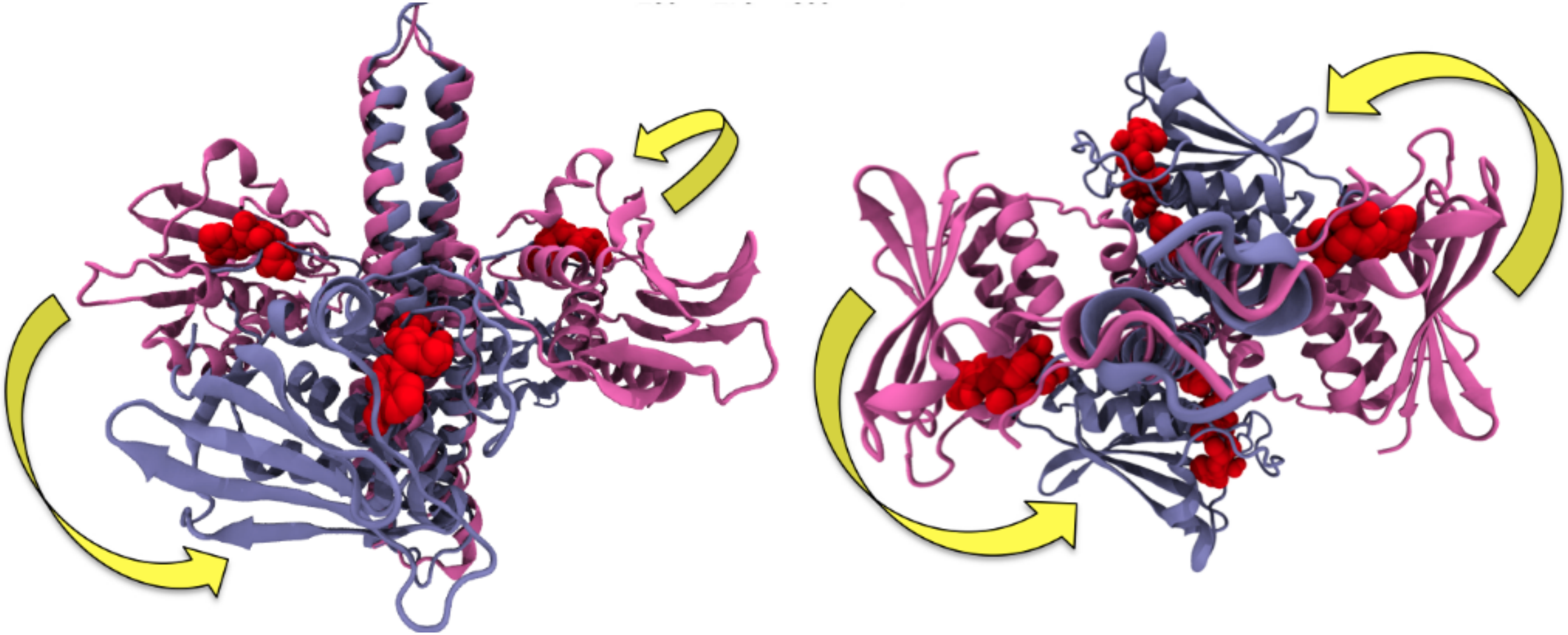
inactive protein structure

guided MD simulations
of coarse-grained model
(Go model)

MD in realistic force
field
(Amber, Gromos)
no use of DCA pairings

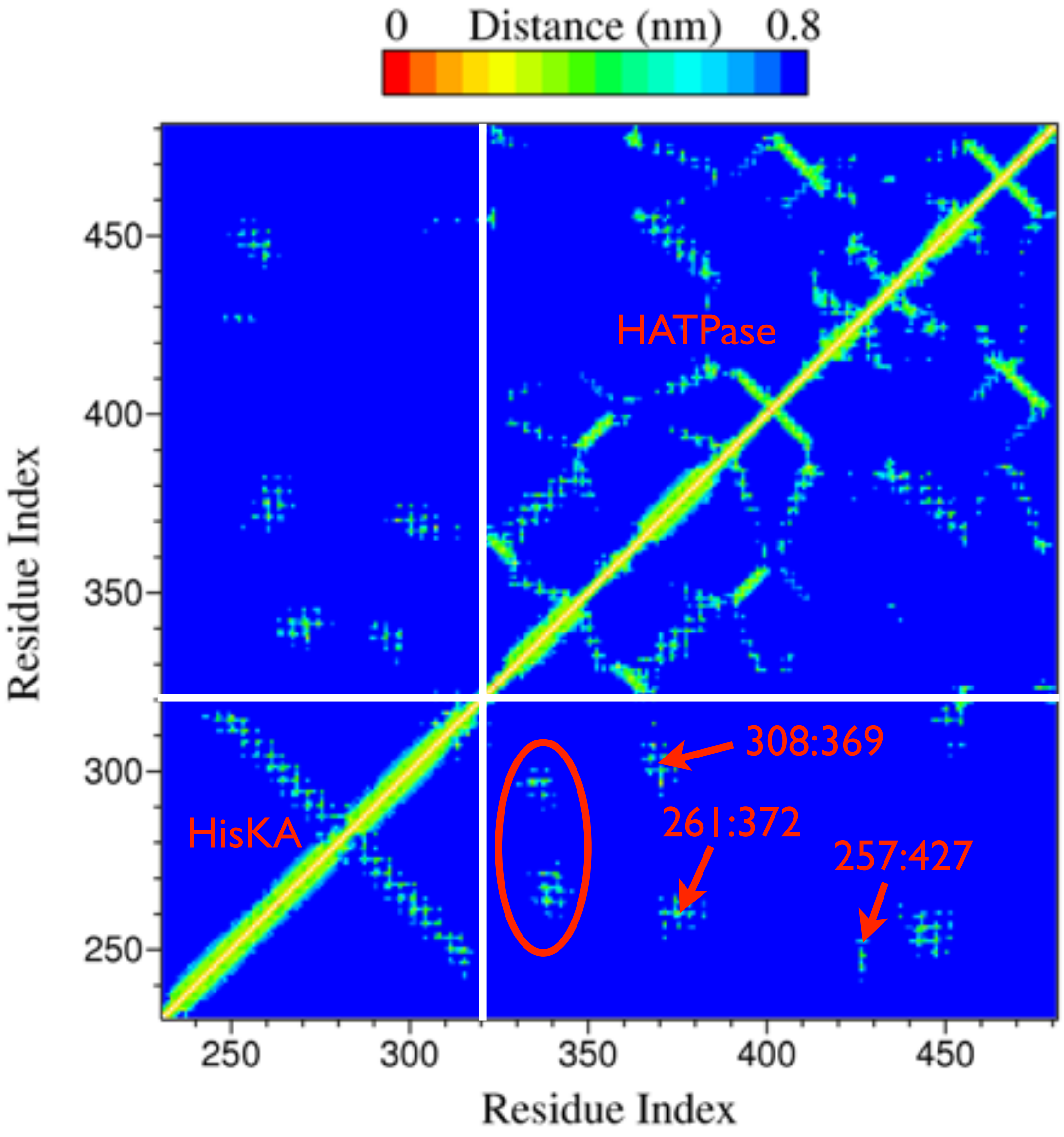


Comparison of the active / inactive structure



➡ major conformational change:
ATP close to Histidine residue

Predicted contact map



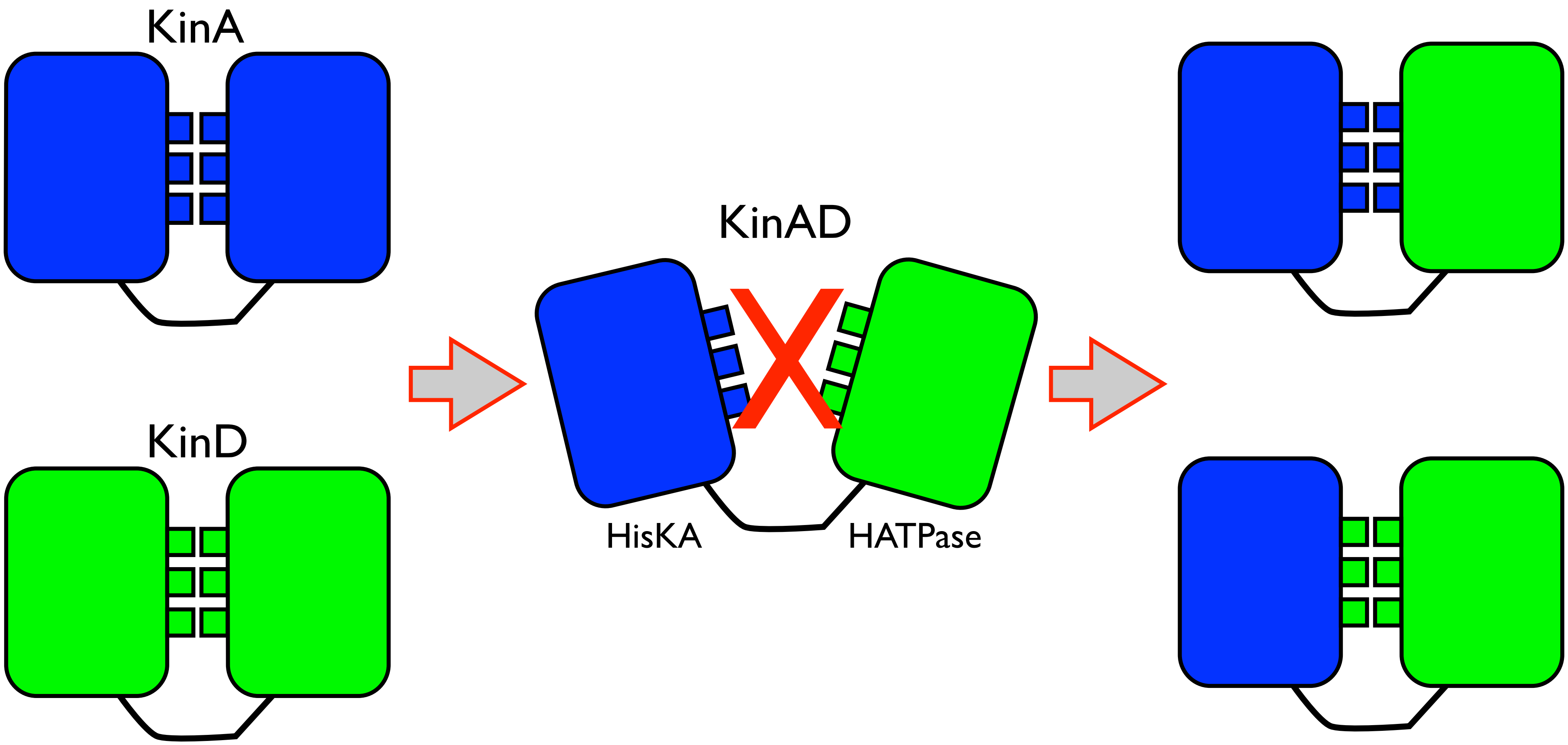
- DCA-predicted pairs in stable contact
- represent clusters of contacts
- MD predicts clusters of contacts with helix 3
  - ▸ not seen by DCA

Experiment:

verify contacts with helix 3!

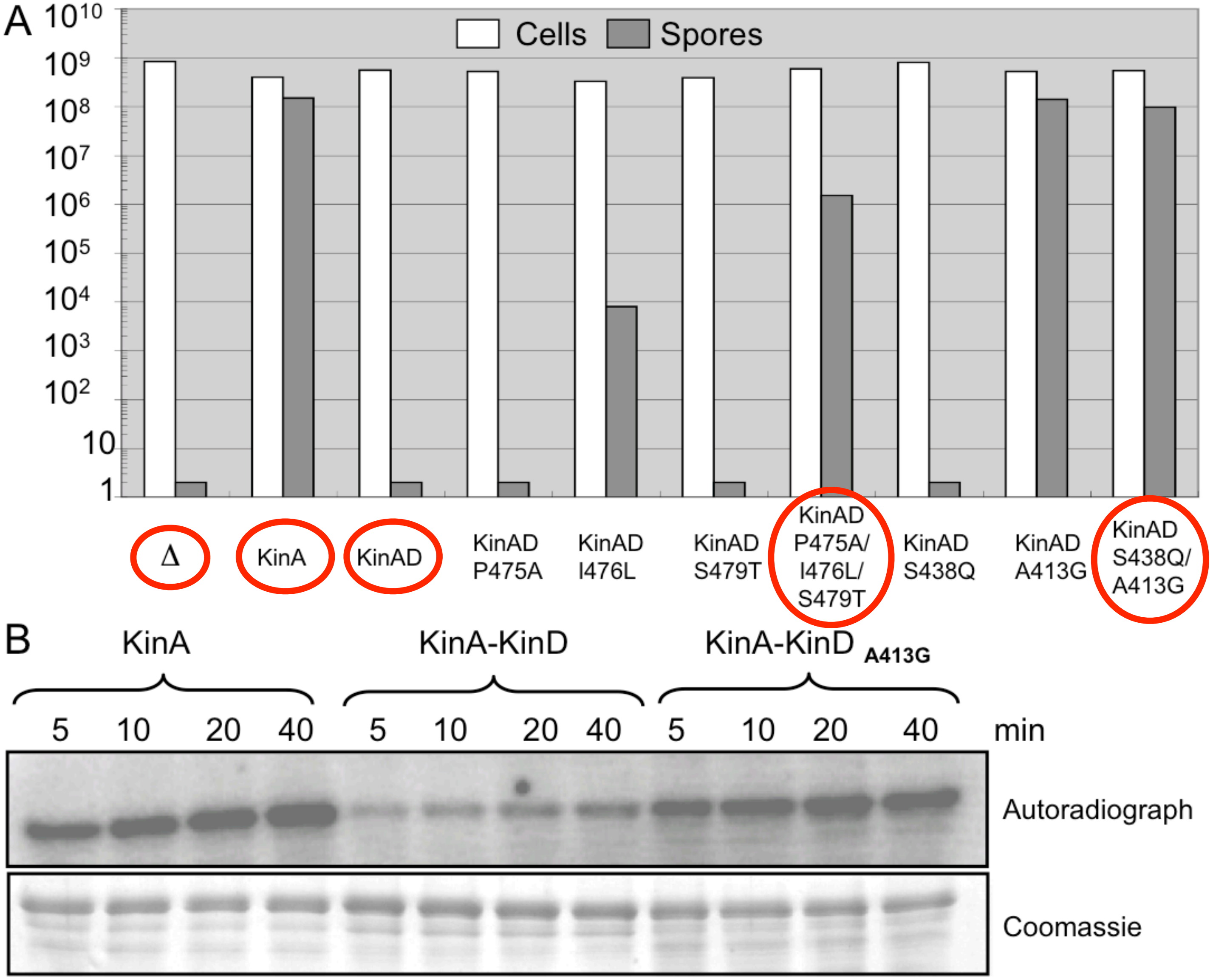
Repairing hybrid kinases

Bacillus subtilis sporulation kinase: KinA as experimental test system

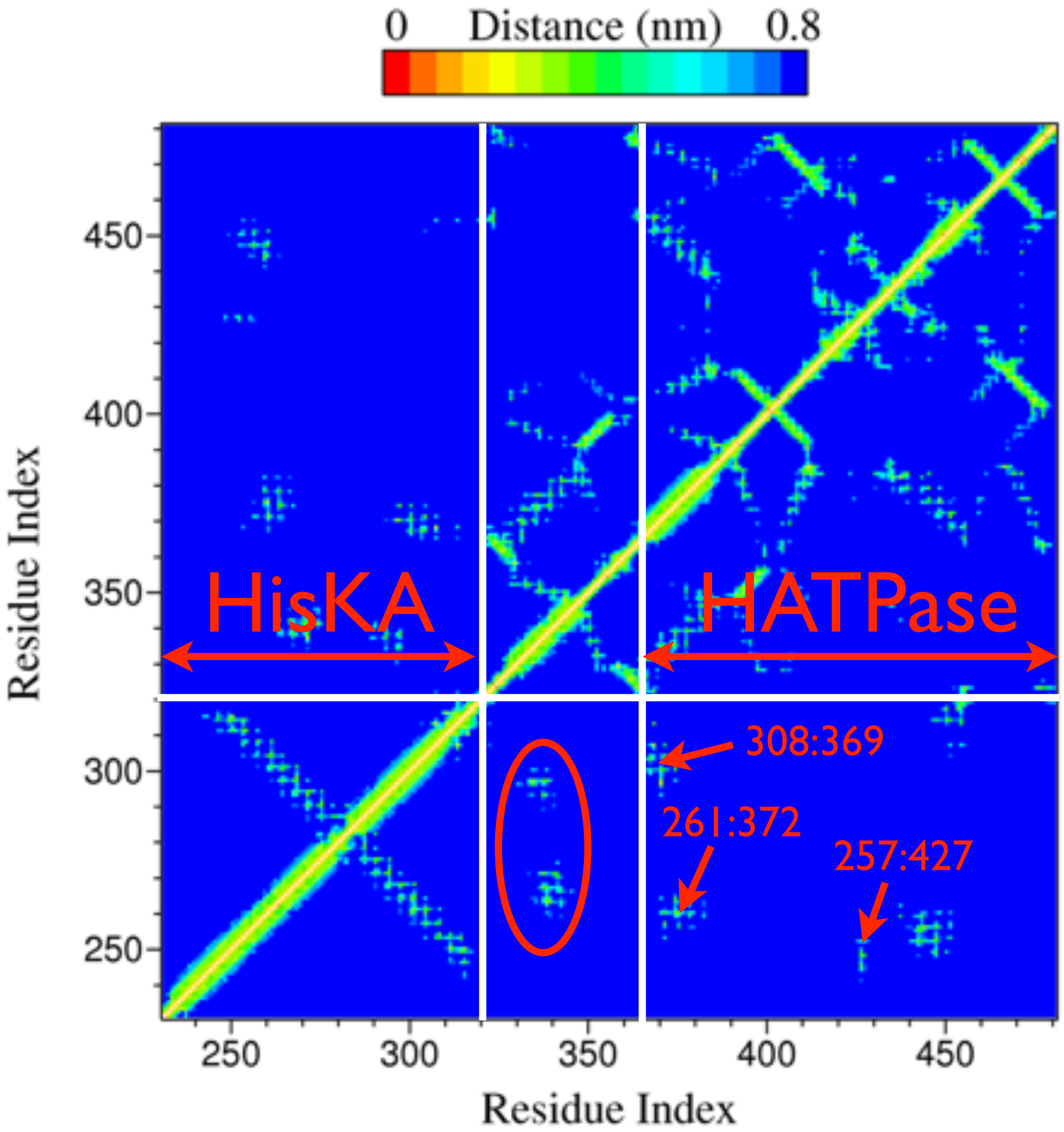


➡ exchange contact residues in one domain by those in cognate domain

Substituting contacts helix 3 - HisKA



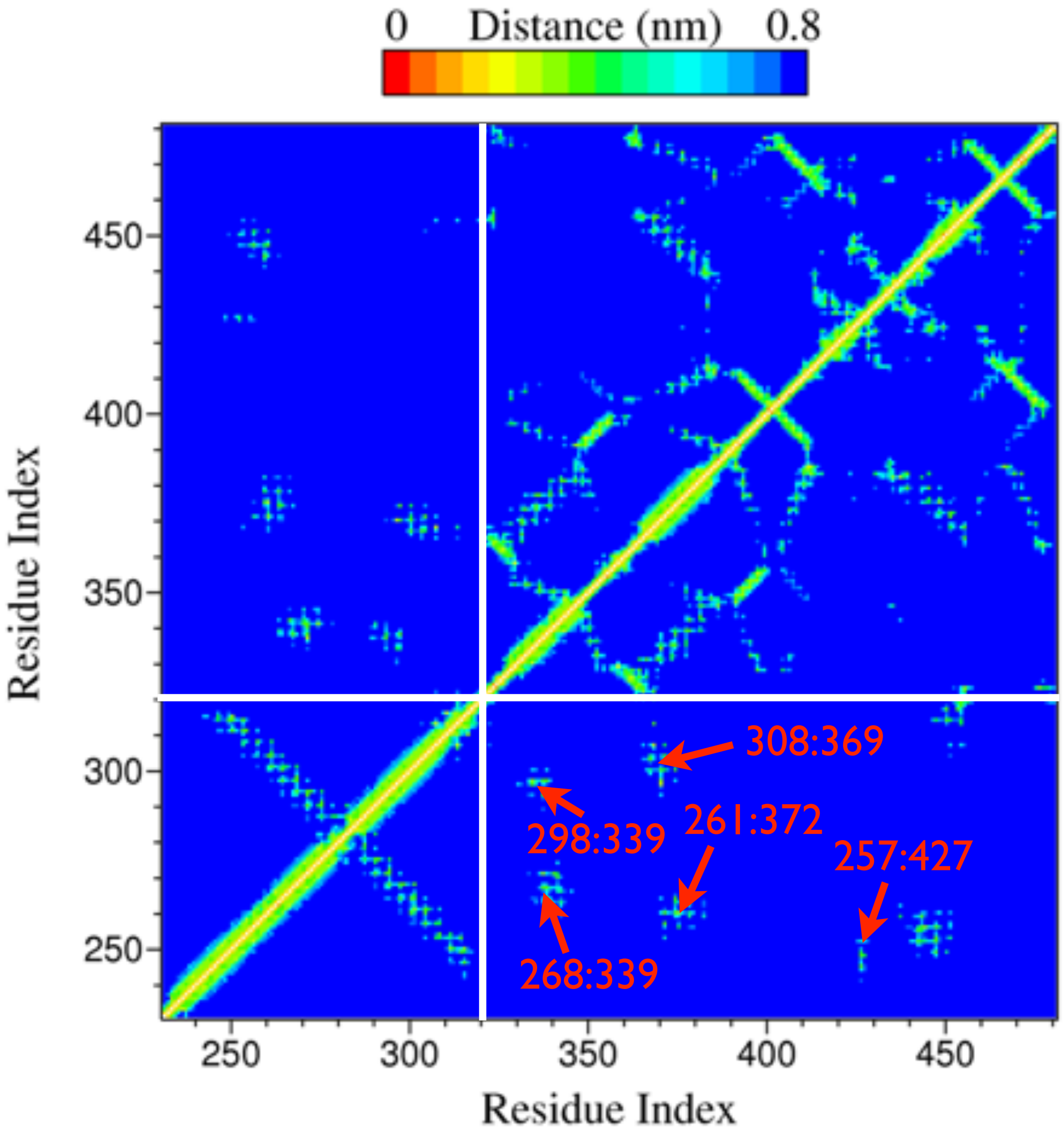
Why don't we see these residues in DCA?

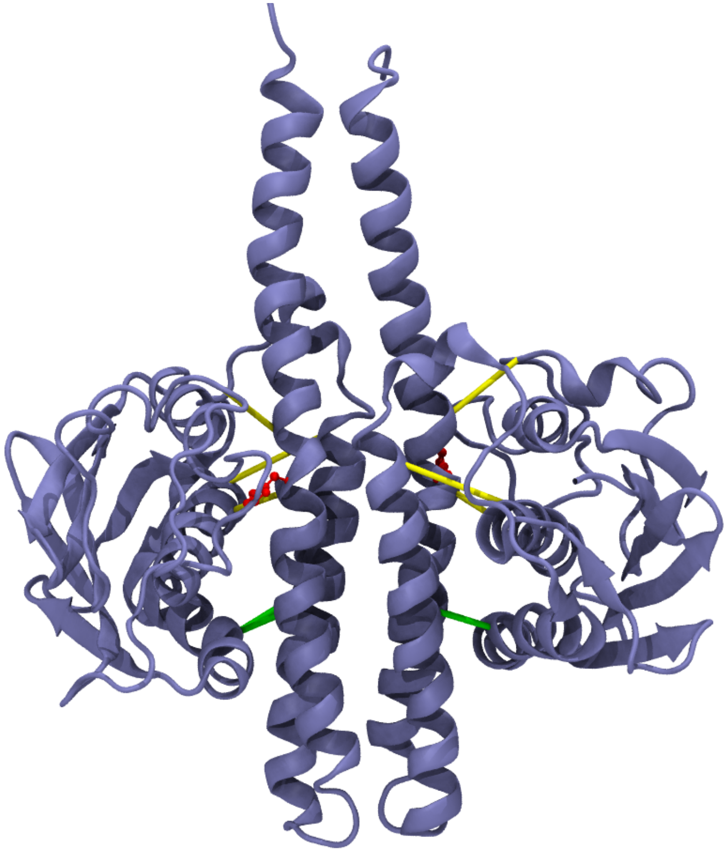


Pfam domains did not cover helix 3!

Improved contact prediction

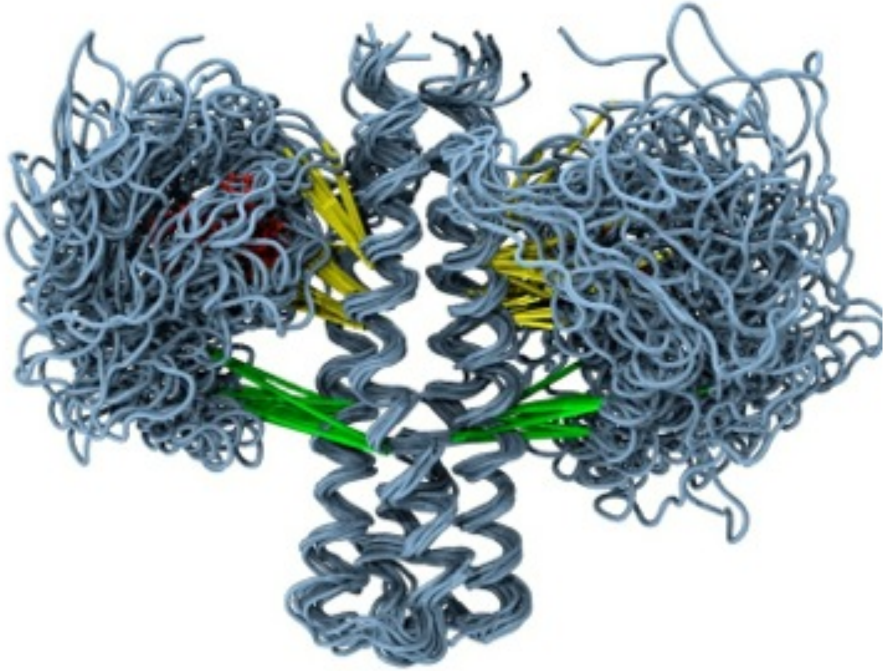
by using a full-length alignment



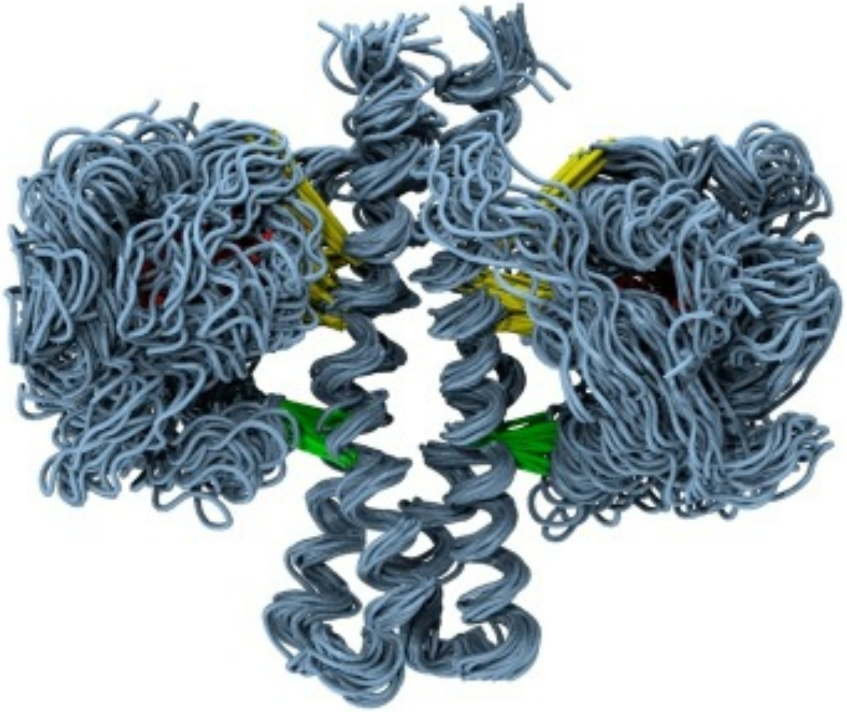


[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS '12]

Improved ensemble of Go-model results



Prediction with 3 contacts



Prediction with 5 contacts

[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS '12]

...and we were just in time

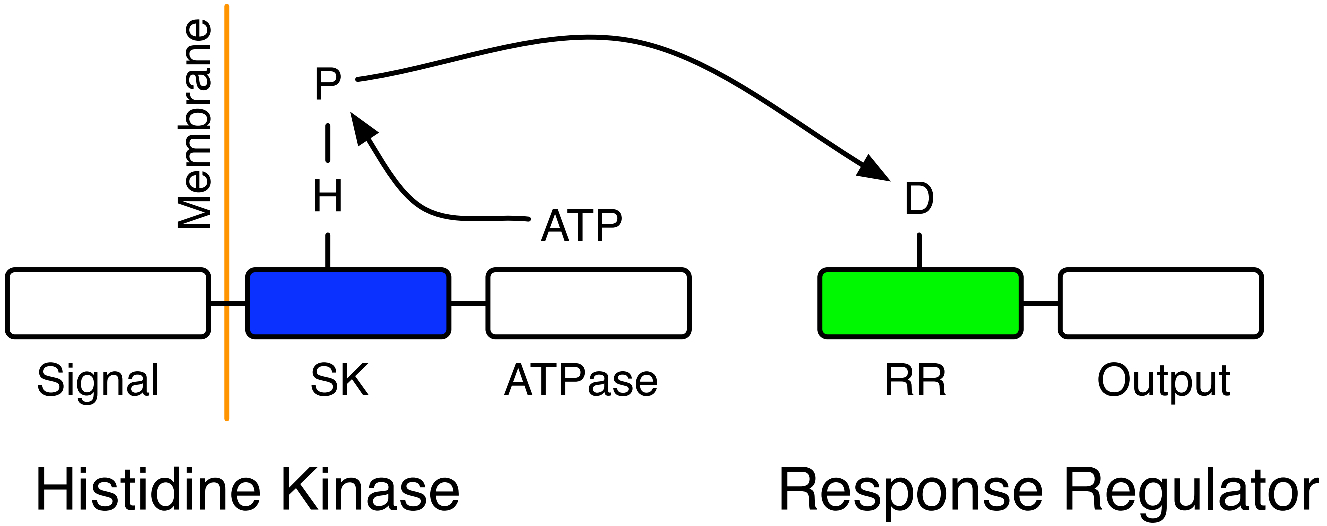
[Wang et al., PLoS Biology ’13]:

- crystal structure of His kinase VicK from *Streptococcus mutans*
- homodimer with

- one monomer in inactive conformation:
  2 inactive DCA predictions at 3.5 – 3.7 Å
- one monomer in active conformation:
  5 active DCA predictions at 2.6 – 5.4 Å

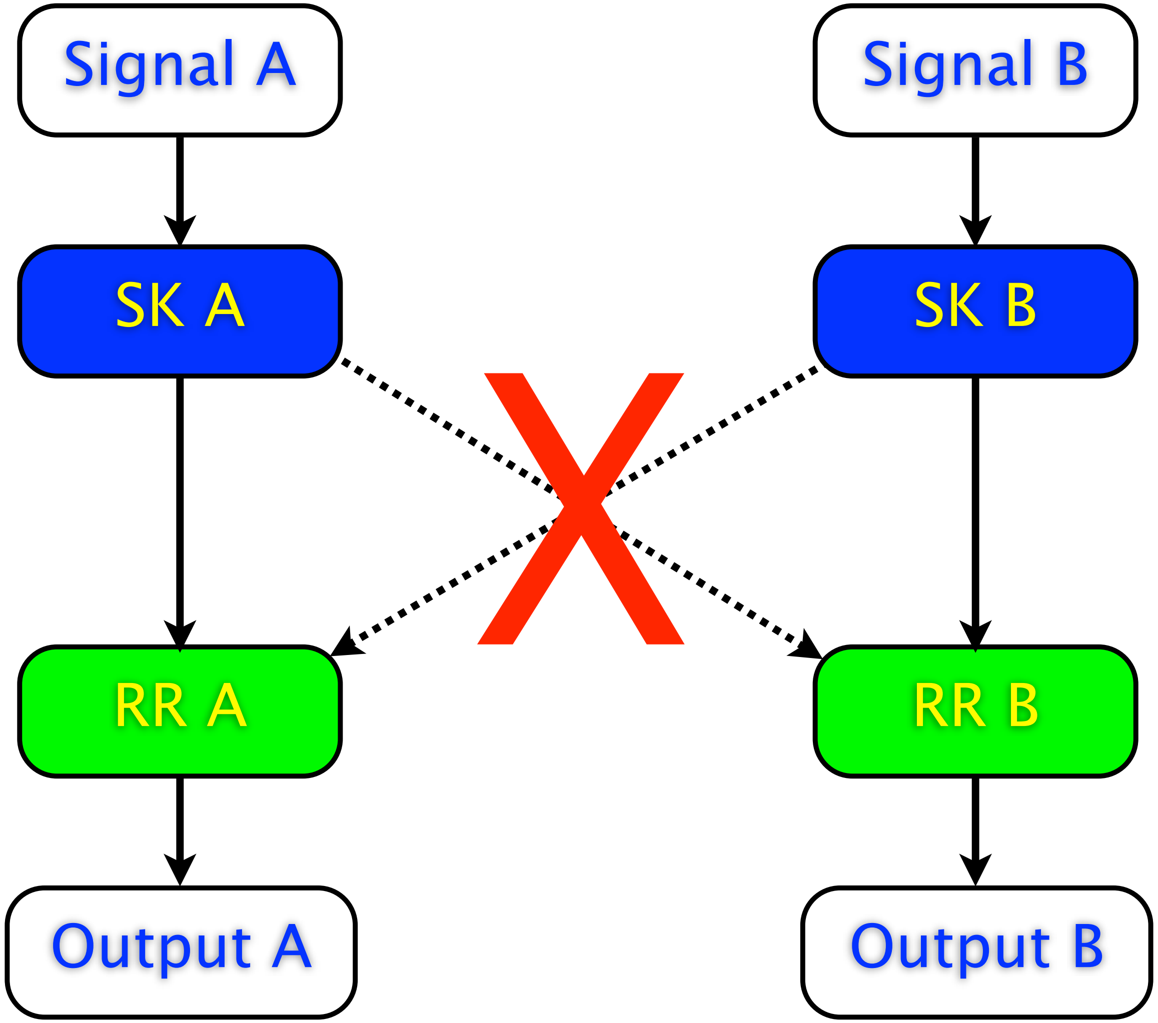
Interaction specificity in two-component signaling

- most common signaling system in bacteria



- amplification: ~*O*(10) homologous SK/RR pairs per genome
- operon organization: partner SK/RR genes frequently co-localized on DNA
- specificity of interaction: little cross-talk between signaling pathways

Specificity vs. crosstalk of signaling pathways



Specific interaction but conserved structure!

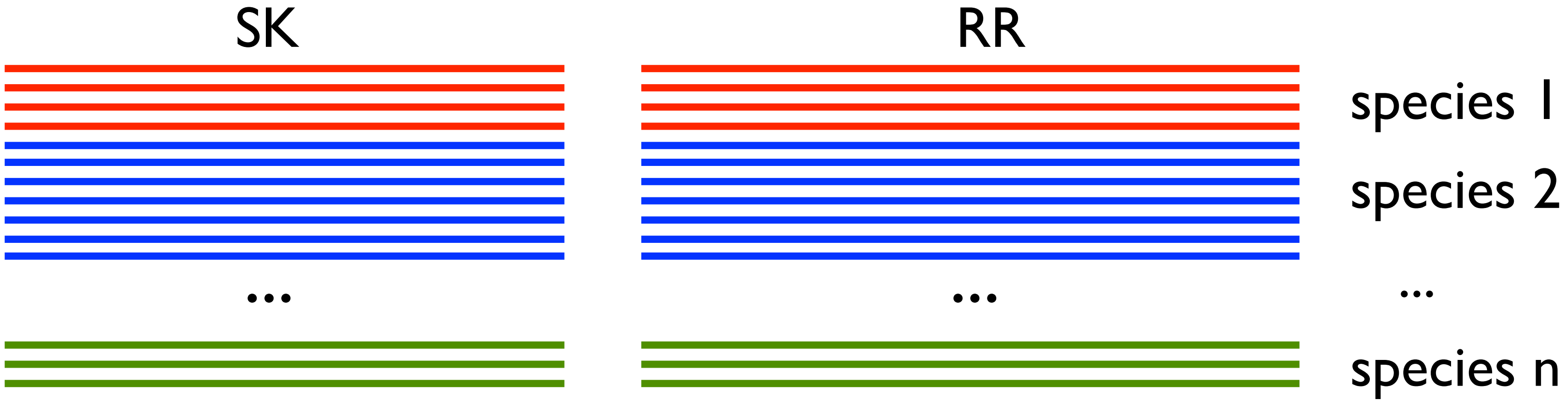
Sequence data

- ca. 750 bacterial genomes

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

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

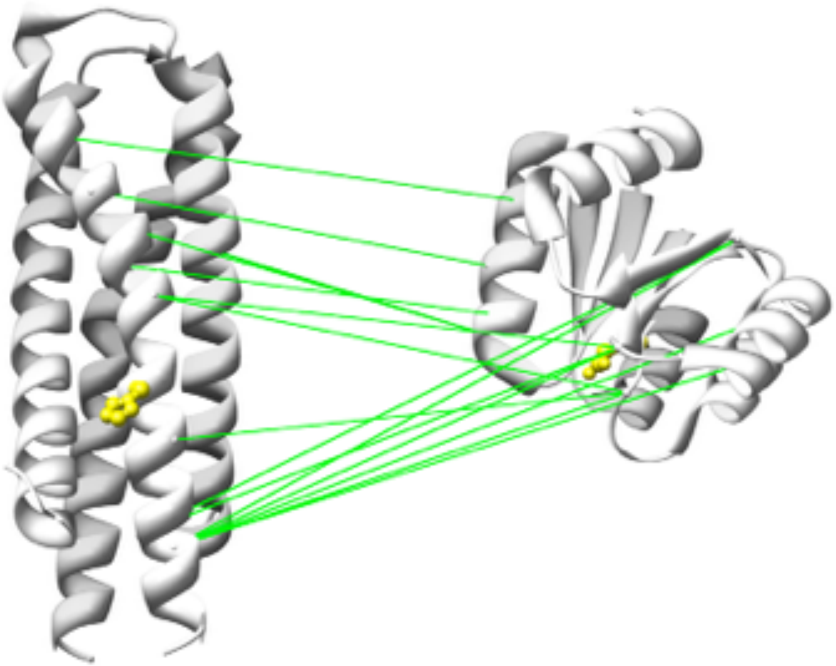
ca. 3800 orphan SK, ca. 9000 orphan RR



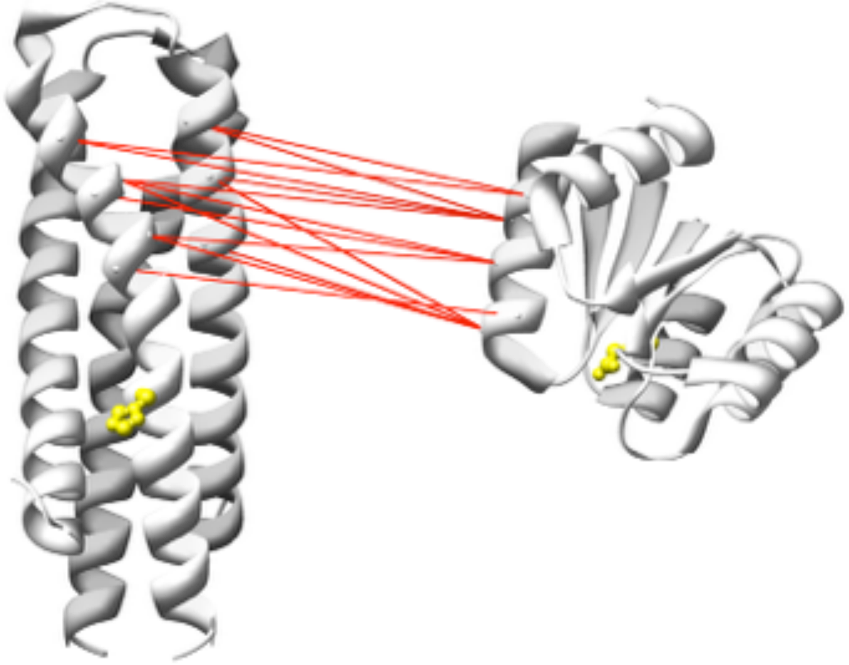
➡ joint statistical model $P(SK, RR)$

[Procaccini, Lunt, Szurmant, Hwa, Weigt, PLoS ONE 2011]

Inter-protein contacts: Two-component signaling



strongest correlations



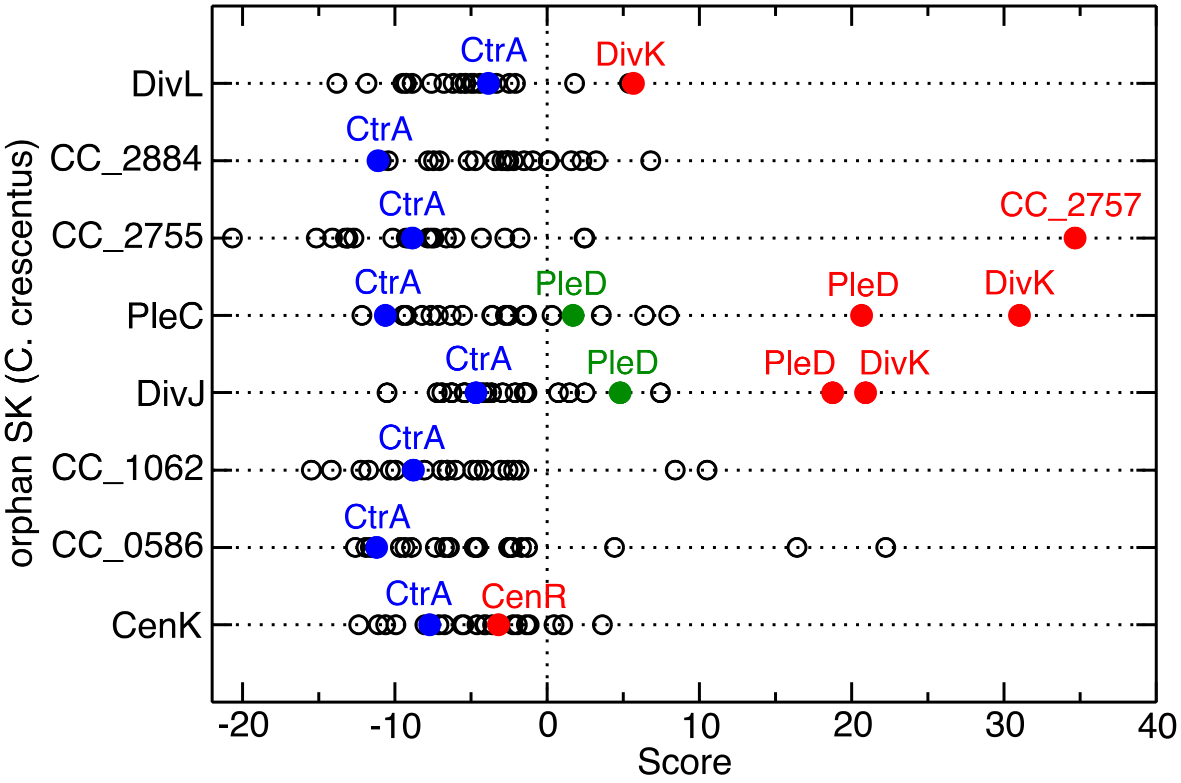
strongest direct couplings

Scoring SK/RR pairs

Log-likelihood score for arbitrary SK and RR sequences

$$S(SK, RR) = \log \frac{P(SK, RR)}{P(SK)P(RR)}$$

- joint statistical model against null model of independent proteins
- score all orphan SK against all orphan RR



[Procaccini, Lunt, Szurmant, Hwa, Weigt, PLoS ONE 2011]

Thanks to:

The group in Paris:

Eleonora de Leonardis
Alice Coucke
Matteo Figliuzzi
Guido Uguzzoni

Rémi Monasson (ENS)
Simona Cocco (ENS)

Collaborations:

Terry Hwa
Hendrik Szurmant
Alexander Schug
Andrea Pagnani
James A. Hoch
Jose Onuchic
Andrea Procaccini
Bryan Lunt
Faruck Morcos
Angel E. Dago
Joanna Sulkowska
Erik Aurell
Magnus Ekeberg
Benjamin Lutz