Statistical modeling of biological sequences

Martin Weigt

Laboratoire de Génomique des Microorganismes

Université Pierre et Marie Curie

Hillerød

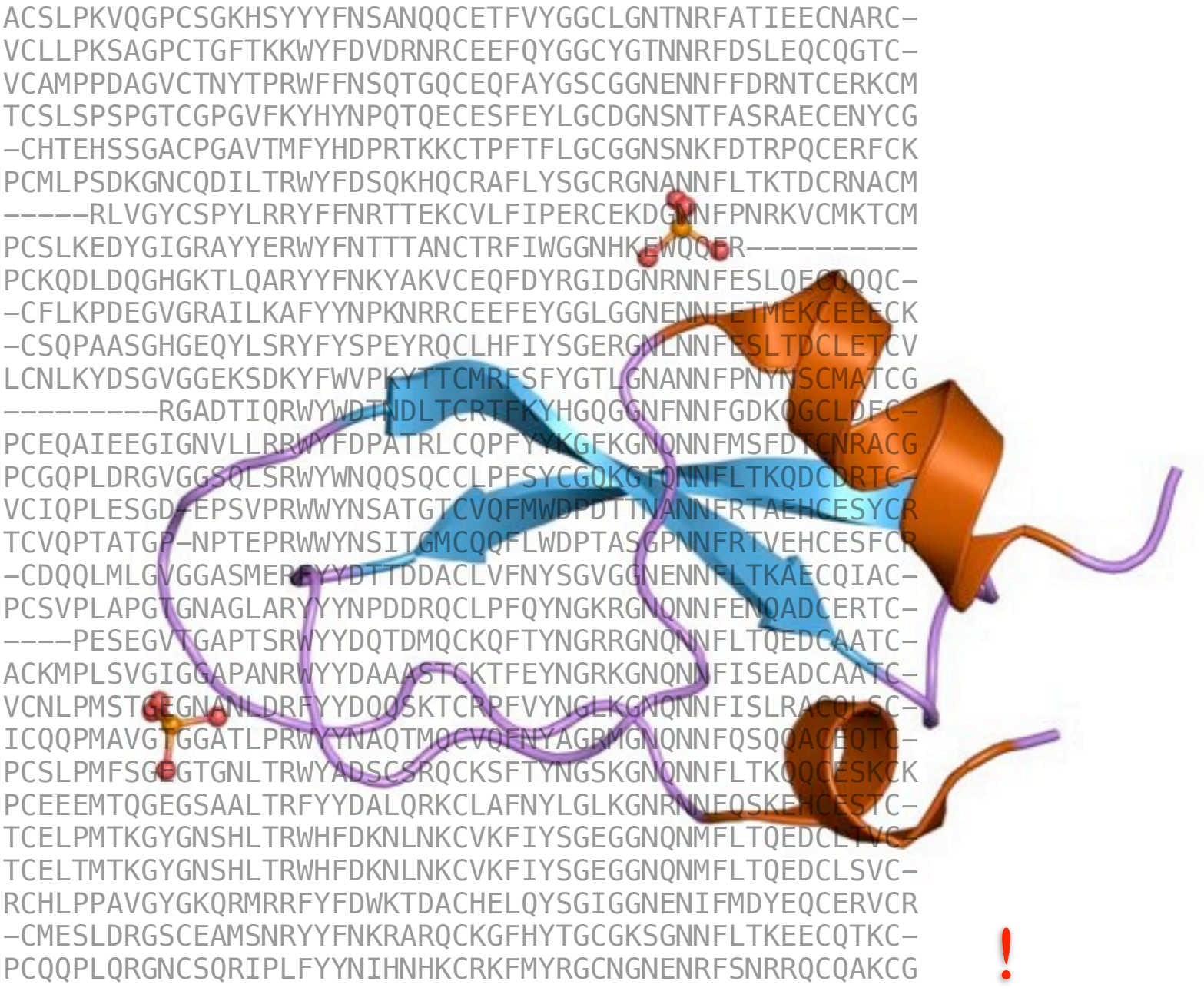
September 2013

Is there 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
```

?

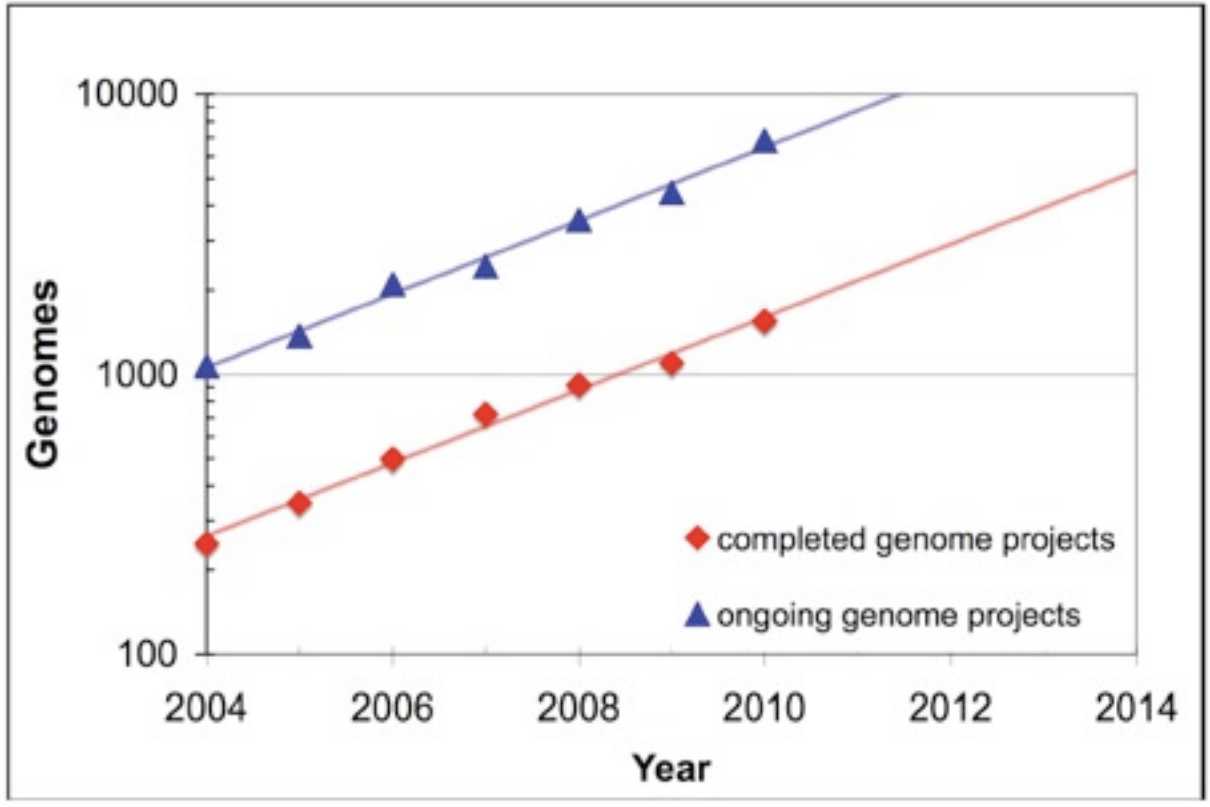
There is information in



!

There are many data...

- >6,800 completed genome sequencing projects
- doubling every 2-3 years (with increasing rate thanks to new technology)



GOLD data base

- equivalent biological functions based on diverged genomic sequences

➡ statistical analysis allows to discover conserved signals in sequences

Plan of the lectures

1. DNA sequence motifs, transcription-factor binding sites and position-specific weight matrices

   [van Nimwegen, BMC Bioinformatics (2007)]

2. Direct-coupling analysis: From residue co-evolution in proteins to protein-structure prediction

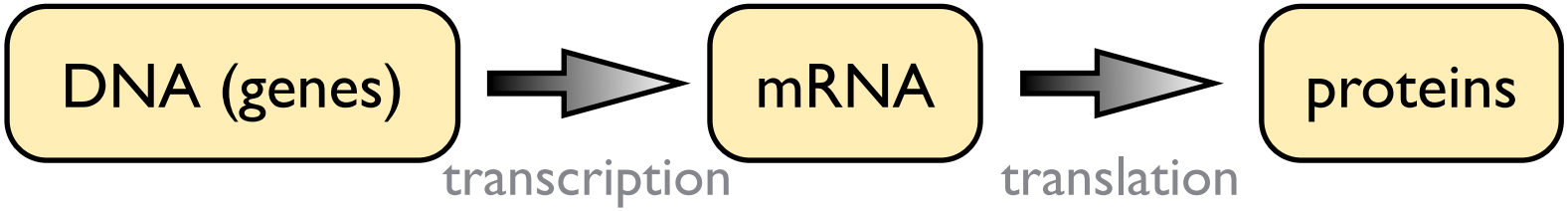
   [Morcos et al., PNAS (2011); Juan, Pazos, Valencia, Nature Rev Gen (2013)]

3. Aligning biological sequences and detecting sequence similarity

   [Durbin, Eddy, Krogh, Mitchison, Biological Sequence Analysis, CUP 1998]

Gene regulation

Central dogma of molecular biology: directed information flow

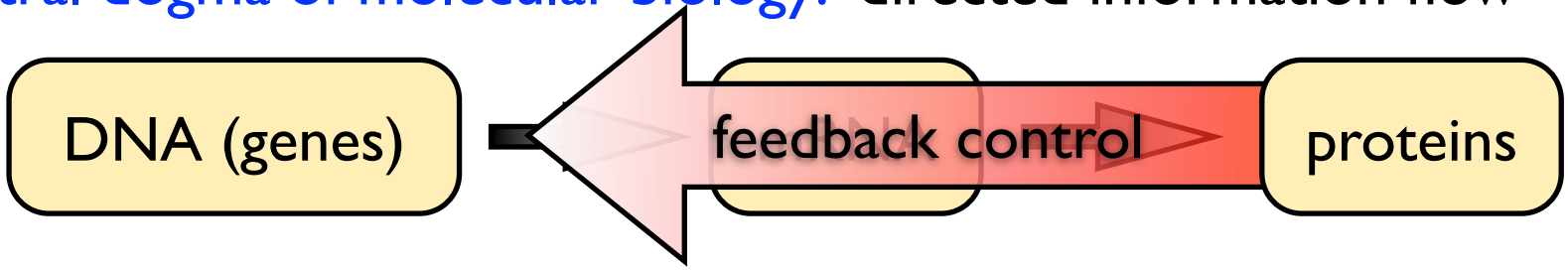


BUT

- different cell types from same genome
  - differential gene expression
- precise timing of gene expression during cell cycle
- response to external signals, nutrient availability etc.

Gene regulation

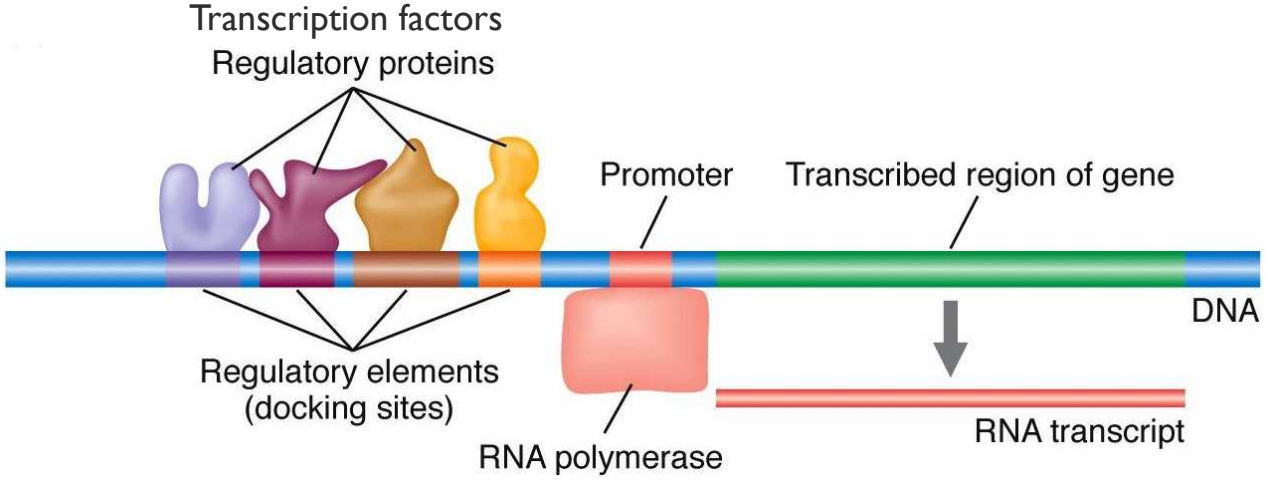
Central dogma of molecular biology: directed information flow



BUT

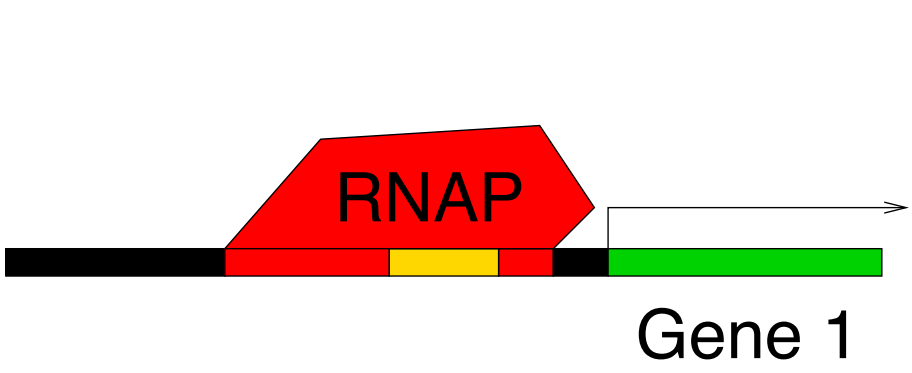
- different cell types from same genome
  - differential gene expression
- precise timing of gene expression during cell cycle
- response to external signals, nutrient availability etc.

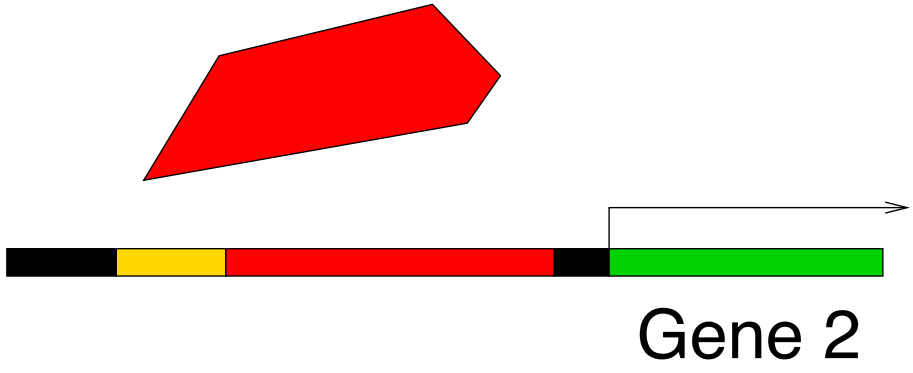
Gene regulation = fundamental process for differential gene expression

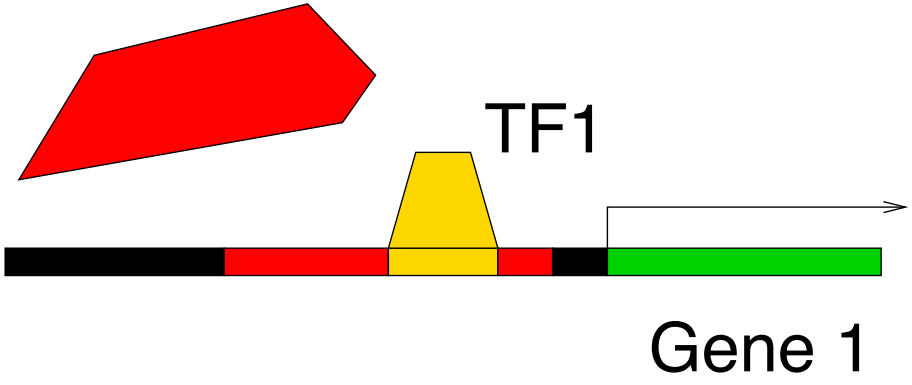


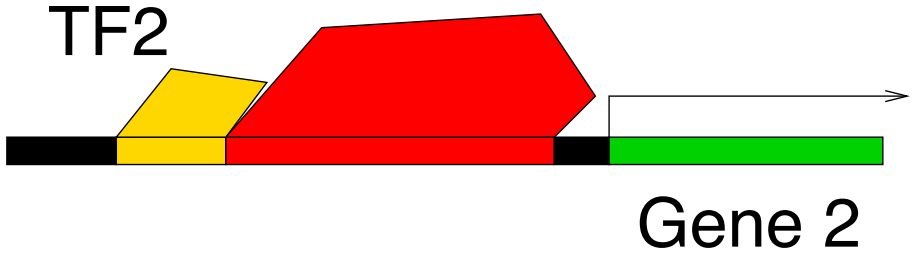
Transcriptional repression vs. activation

Simplest regulatory functions:







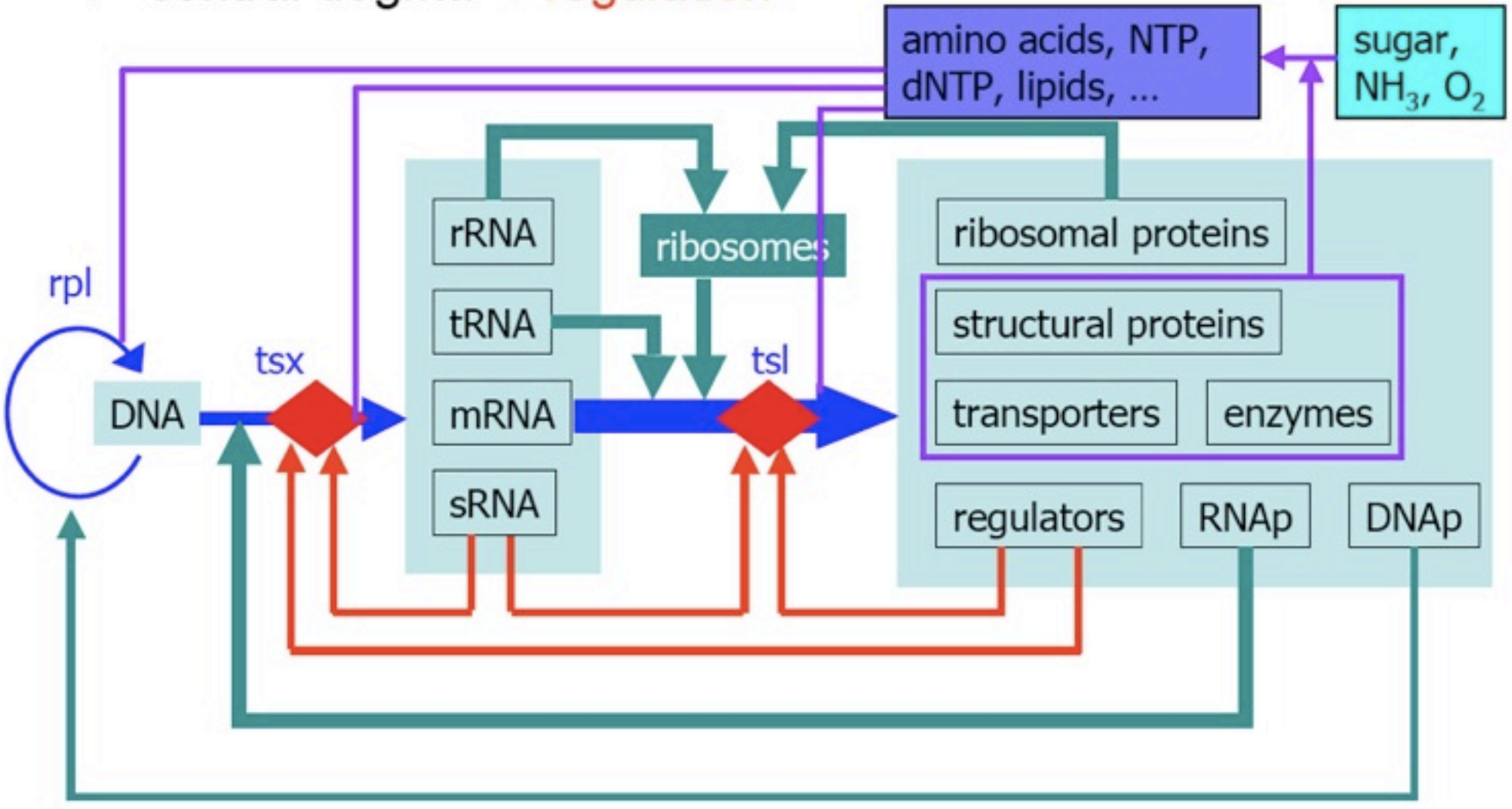


repression
by competitive binding

activation
by collective binding

Gene regulation

❖ central dogma + regulation

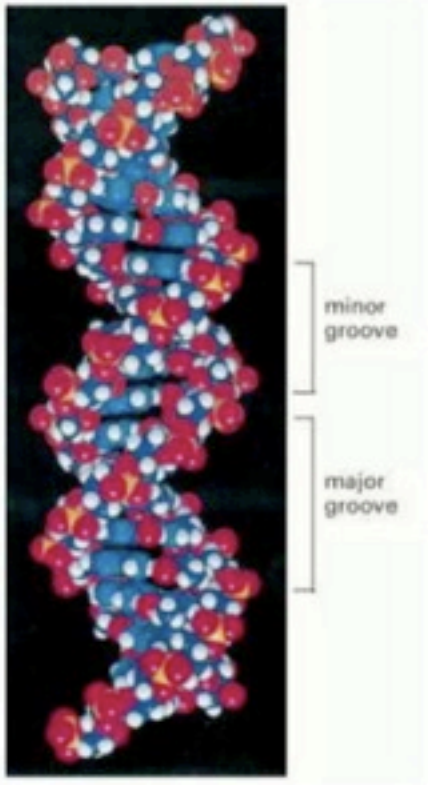


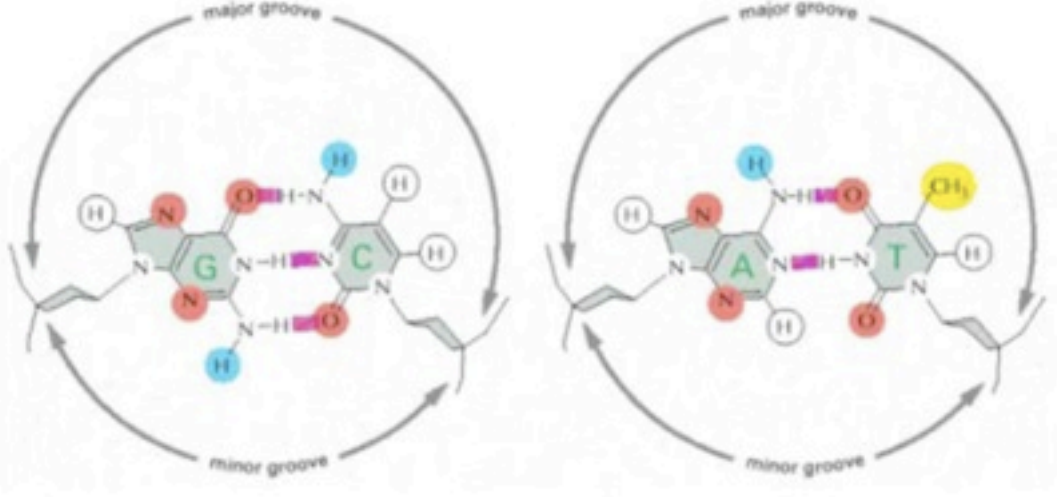
➡ concentrate on transcriptional regulation

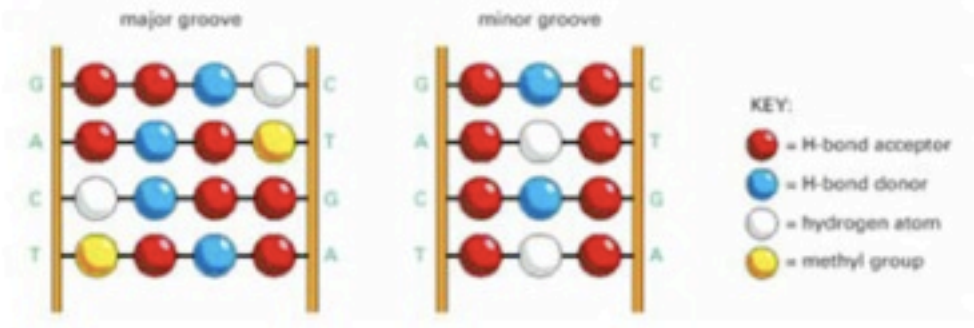
Protein-DNA interactions

A. Empirical facts

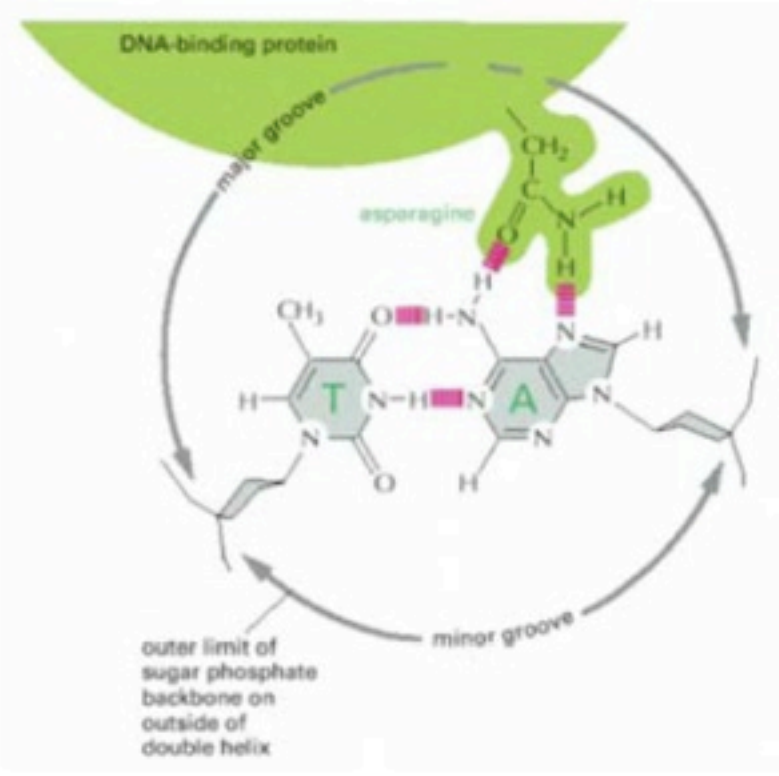
1. Transcription Factors
   - size: ~5nm (10-20 bp)
   - molecular basis of sequence recognition

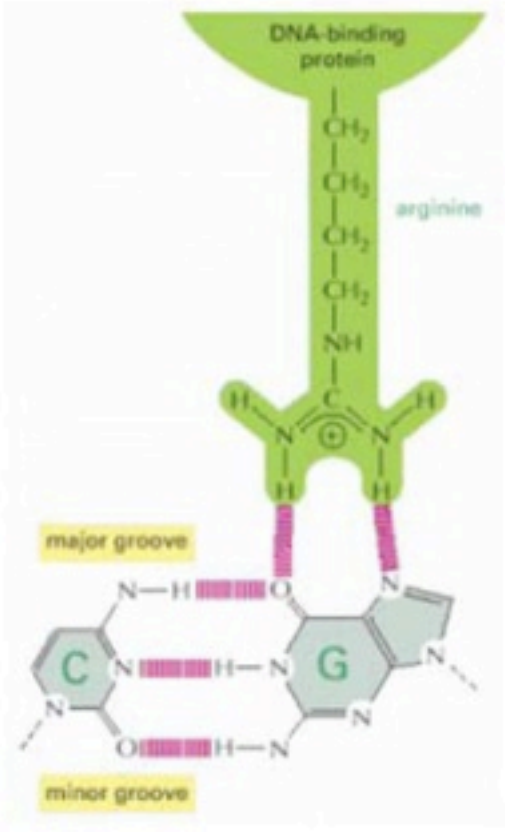






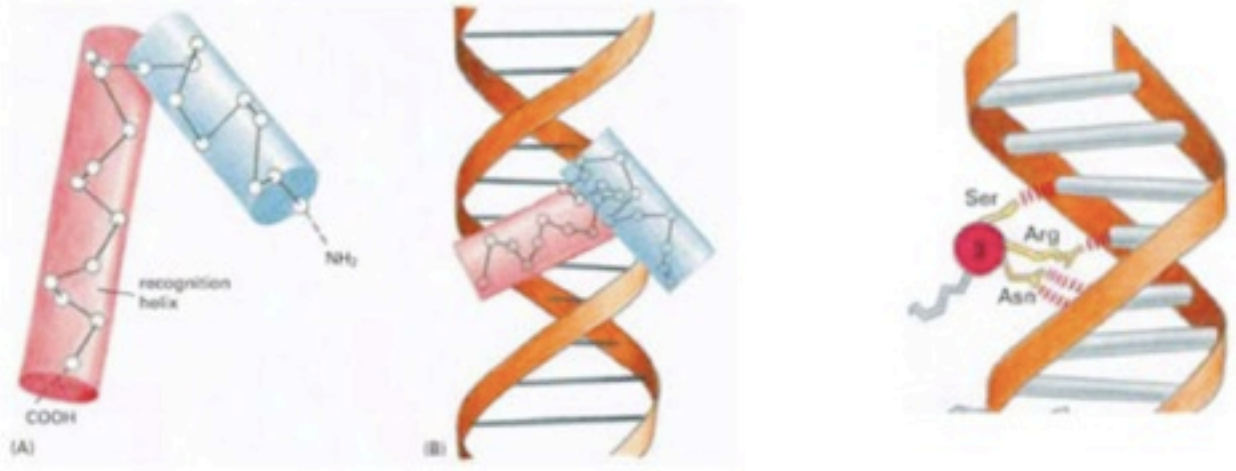
- contact between TF and DNA



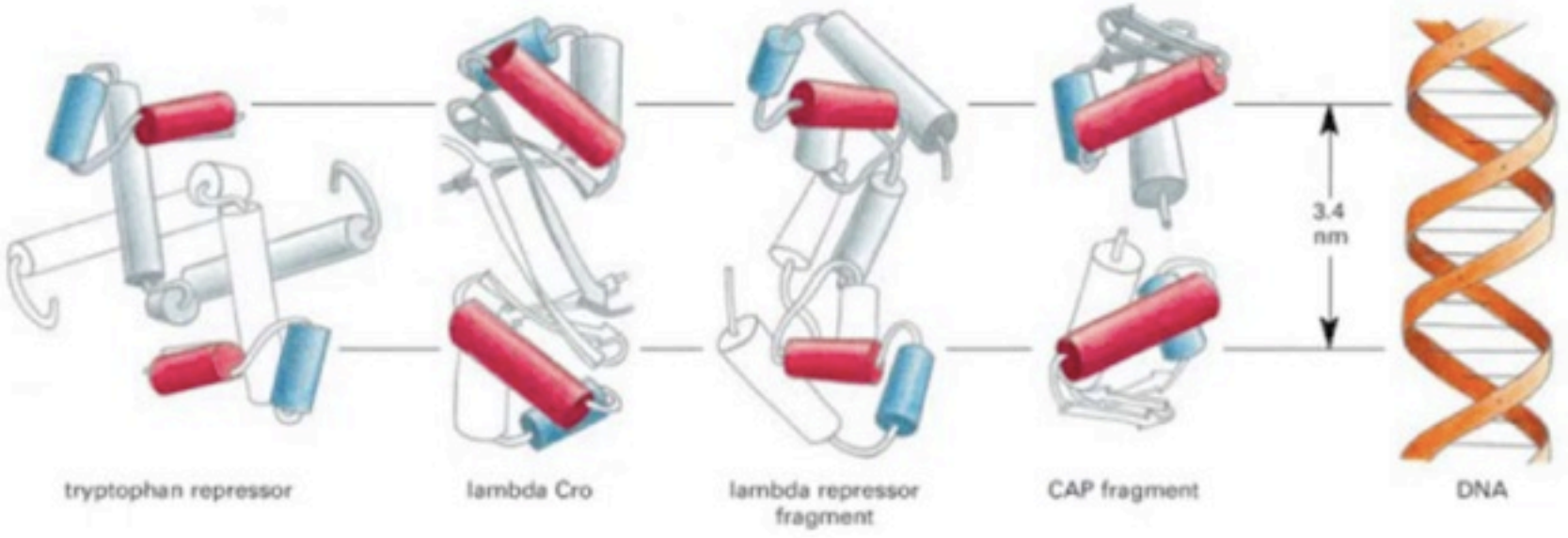


➔ structure of a TF must place the appropriate amino acids next to the base pairs they contact

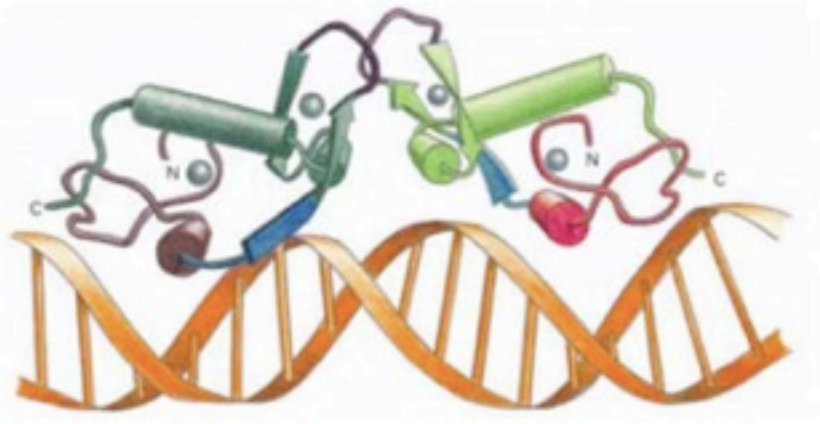
- various molecular strategies
  - Helix-Turn-Helix



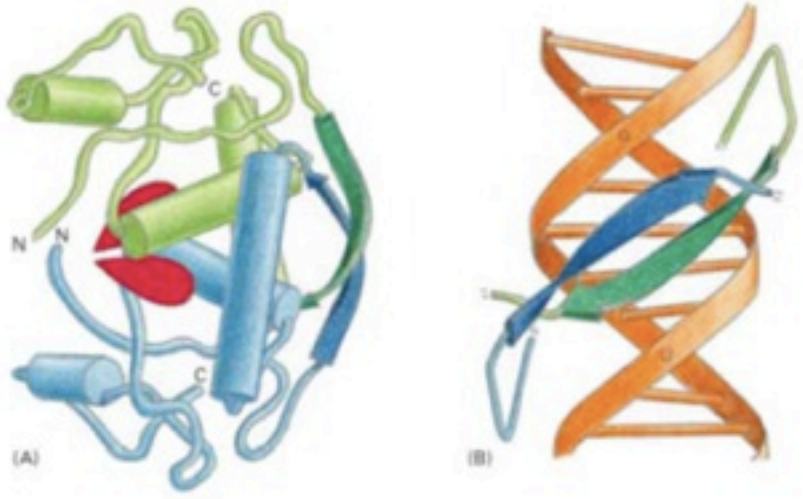
well-known examples in bacteria (note: homodimers)



– zinc-finger domain

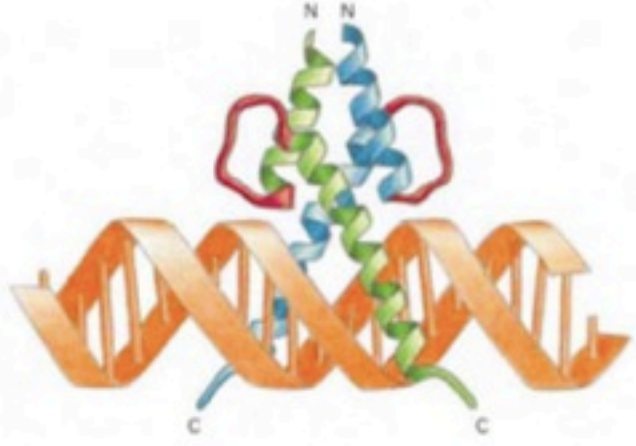


– beta-sheets



– leucine zipper

– helix-loop-helix



2. DNA binding sequences

- typically 10-20 bp in bacteria

protein	target sequence
lac repressor	5' AATTGTGAGCGGATAACAATT 3' TTAACACTCGCCTATTGTTAA
CRP	TGTGAGTTAGCTCACT ACACTCAATCGAGTGA
λ repressor	TATCACCGCCAGAGGTA ATAGTGGCGGTCTCCAT

- lots of sequence variants
- consensus sequence often palindromic
- common to have 2~3 mismatches from the core consensus sequence -- "fuzzy" binding motif

3. TF-DNA interaction

- passive (no energy consumption)
- strong electrostatic attraction indept of binding seq

  e.g., $[TF-DNA] > 10 \times [TF]_{free}$ for LacI in 0.1M salt

  ➔non-specific binding: $G_{ns} - G_{cyto} \simeq -15kT$

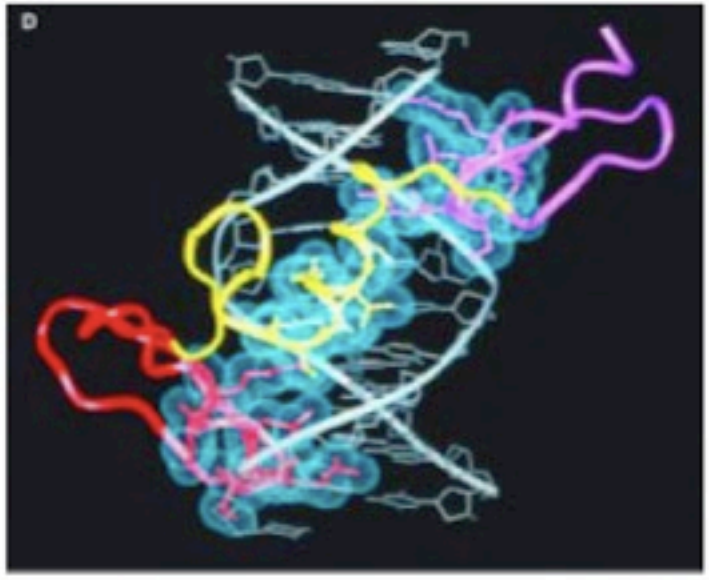
  ($kT \approx$ 0.62 kcal/mole at 37C)

- additional energy gained from hydrogen bonds to preferred sequences

  strongest binder: $G^* - G_{ns} \simeq -15kT$

- graded increase in binding energy for sequences with partial match to the preferred sequence

- relative binding affinity for Mnt



binding energy matrix

(in unit of kT ≈ 0.6 kcal/mole)

pos.	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(D.S. Fields, Y. He, A. Al-Uzri & G. Stormo, 1997)

(from competitive binding expts)

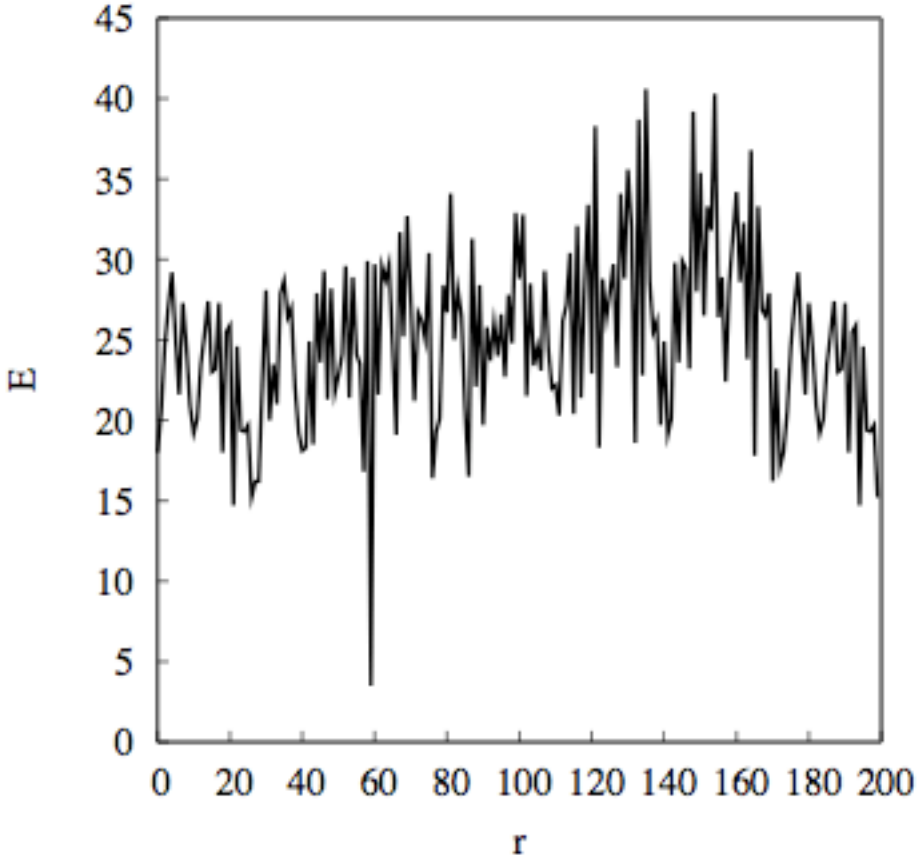
➔ weak energetic preference -- weak specificity

➔ similar results for other TFs studied (e.g., LacI, λ-CI, λ-Cro)

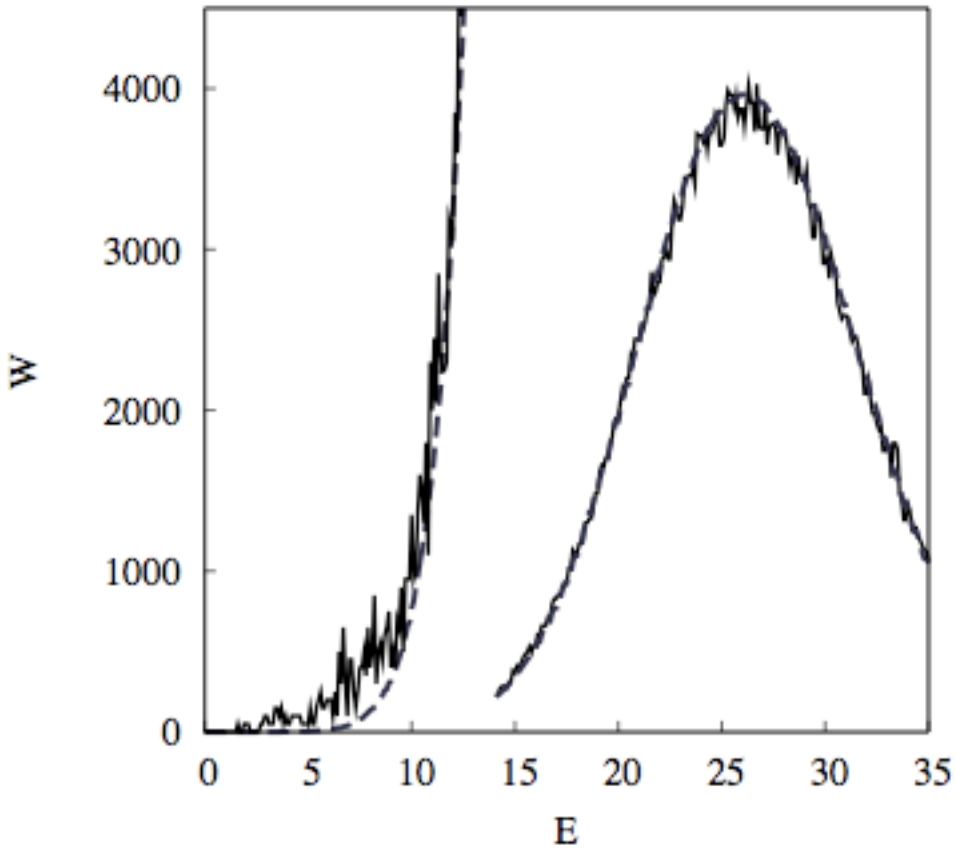
- double mutation: binding energy approx additive

➔ Can we say something generic about the design of TF-DNA interaction from these facts/data?

CRP / E. coli vs. random genome



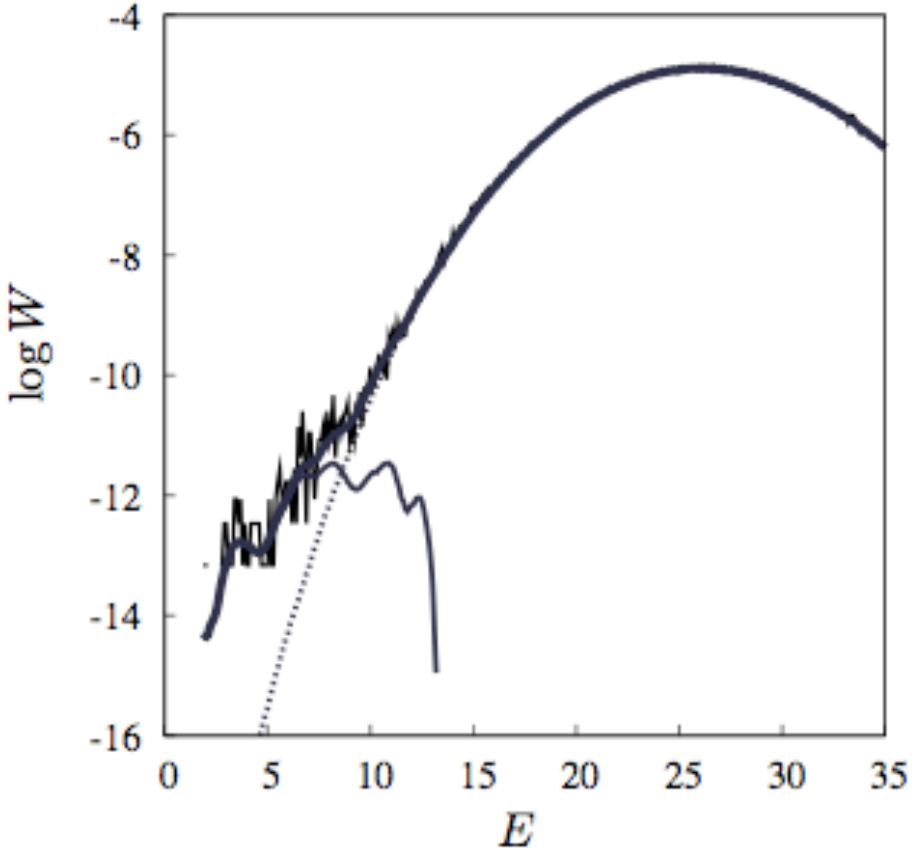
one intergenic region with TFBS in position 59

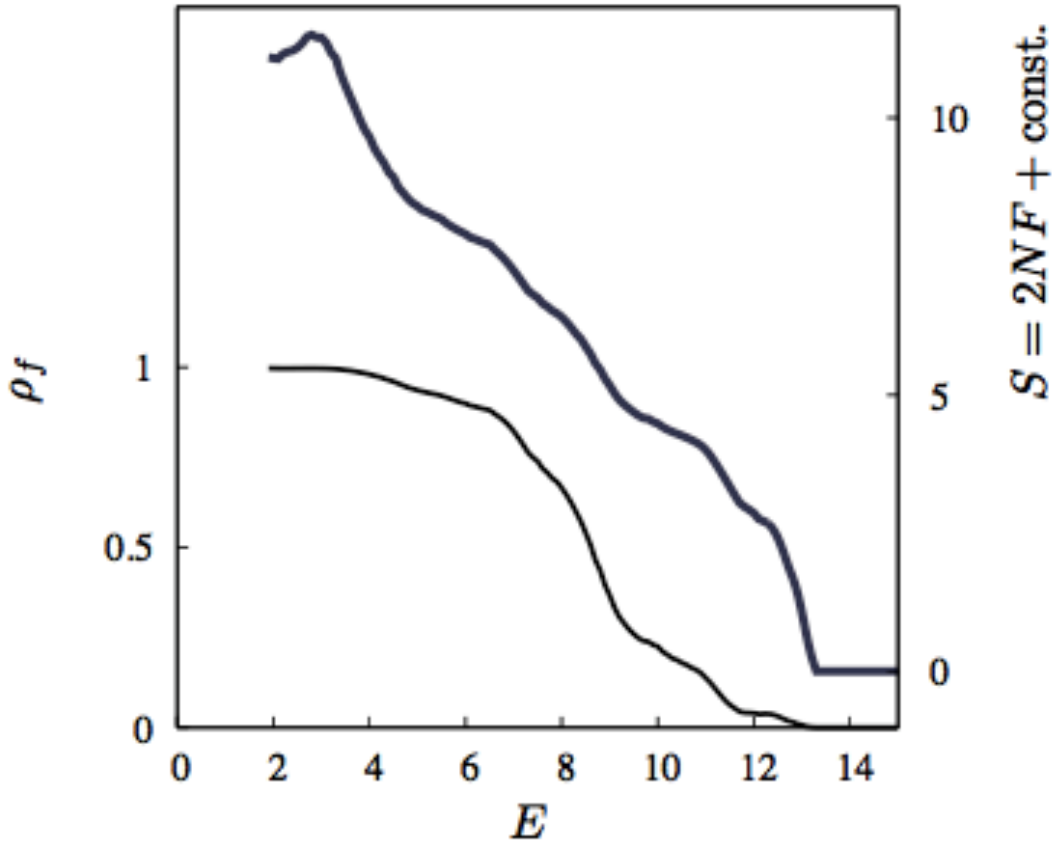


Frequency count: E. coli intergenic vs. randomized genome

for transcription factor CRP

CRP / E. coli vs. random genome





From networks of residue coevolution to protein (complex) structure prediction

Martin Weigt

Laboratoire de Génomique des Microorganismes

Université Pierre et Marie Curie

Is there 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 is this interesting?

High-throughput sequencing technology:

- >6,800 fully sequenced genomes
- >21,000 incomplete genome sequencing projects
- >42,000,000 protein sequences (500,000 with annotation)
- exponential growth of databases

Why is this interesting?

High-throughput sequencing technology:

- >6,800 fully sequenced genomes
- >21,000 incomplete genome sequencing projects
- >42,000,000 protein sequences (500,000 with annotation)
- exponential growth of databases

Structural databases:

- ~93,000 known protein structures
- linear growth of databases (no high-throughput technology)

...but function relies on structure!

Why is this interesting?

High-throughput sequencing technology:

- >6,800 fully sequenced genomes
- >21,000 incomplete genome sequencing projects
- >42,000,000 protein sequences (500,000 with annotation)
- exponential growth of databases

Structural databases:

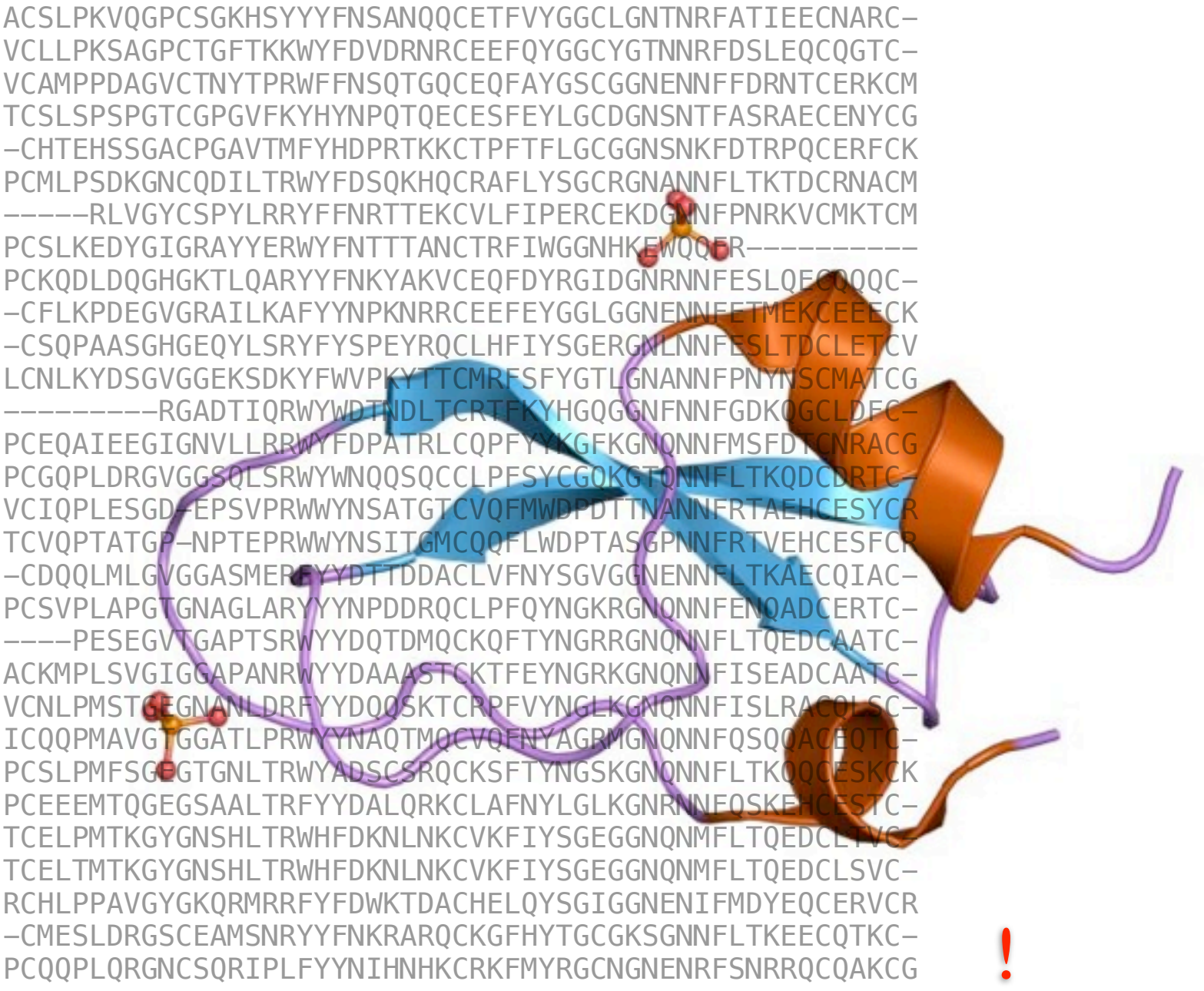
- ~93,000 known protein structures
- linear growth of databases (no high-throughput technology)

...but function relies on structure!

How to close the gap:

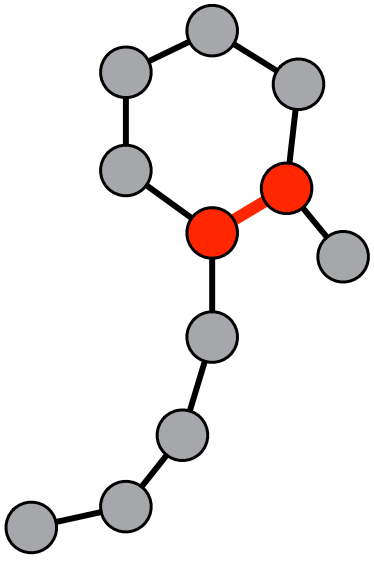
- sequences classified into ~14,000 families of *homologous proteins*
  - common evolutionary origin
  - conserved structure / function but variable sequence
- >4,000 families with >1,000 sequences
- >1,000 of these families without structural representatives

There is information in



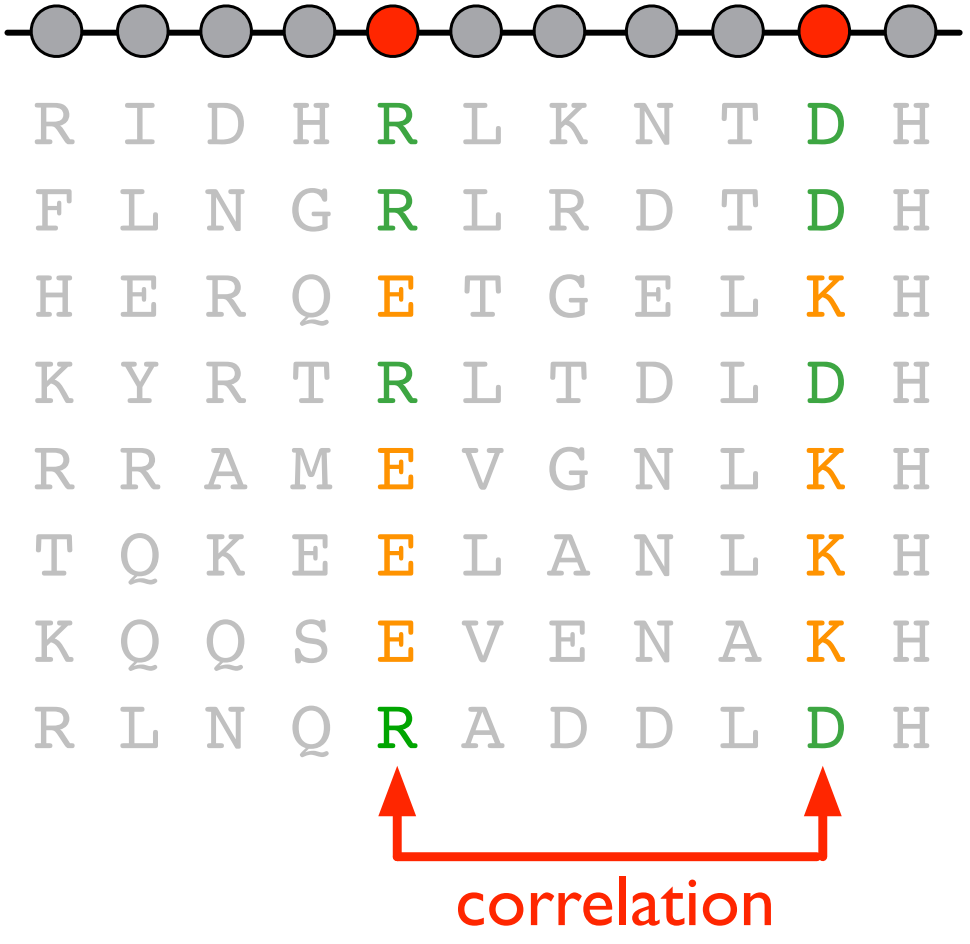
!

Residue contacts induce residue co-evolution

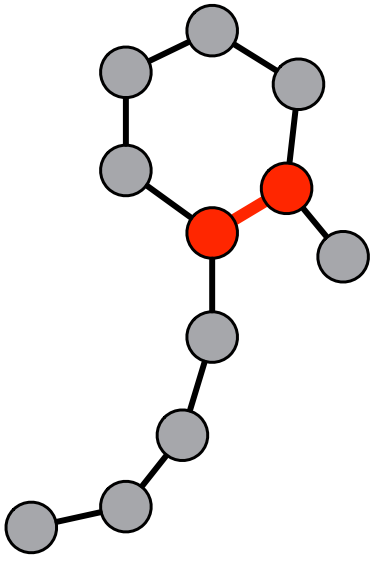


contact in 3D

co-evolution



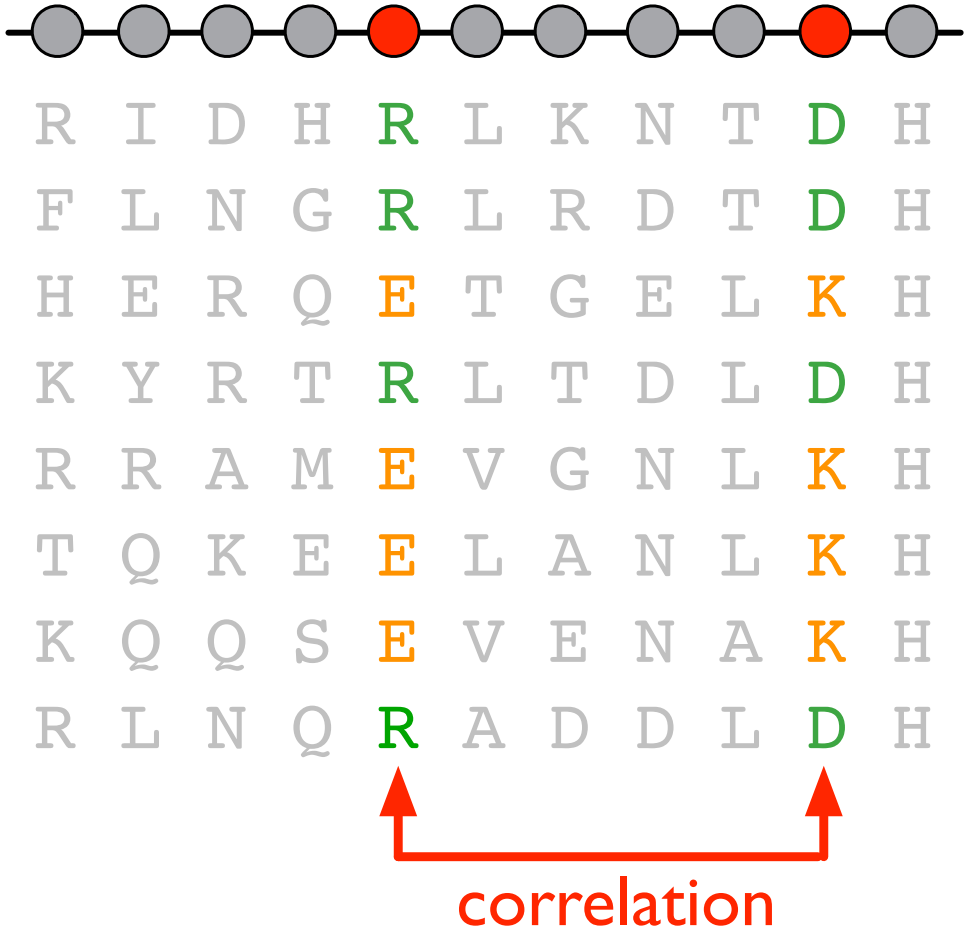
Residue contacts induce residue co-evolution



contact in 3D

co-evolution

statistical analysis



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

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

i j

$f_i(A)$ $f_j(B)$

$f_{ij}(A, B)$

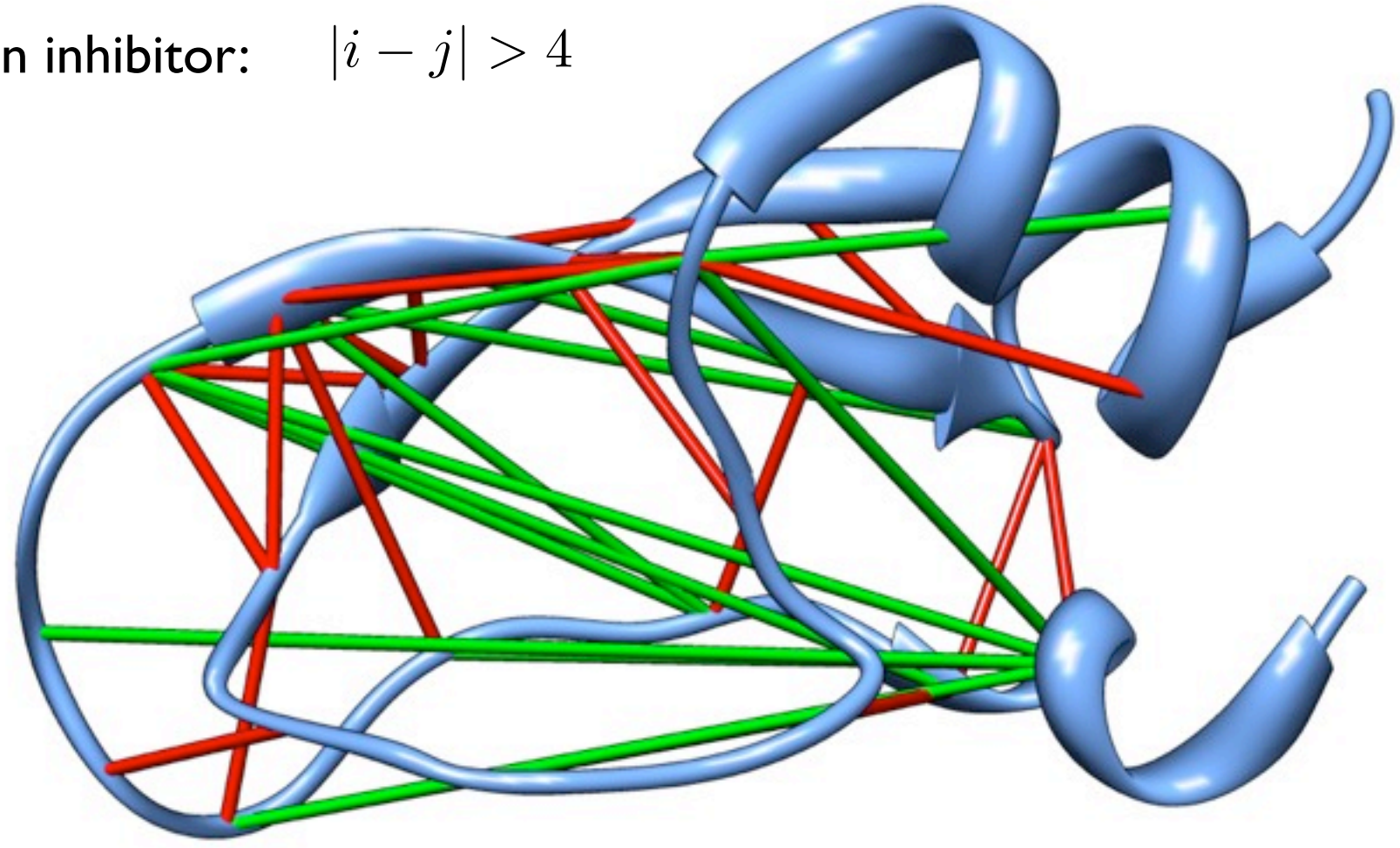
Mutual information measures pair correlation

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

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

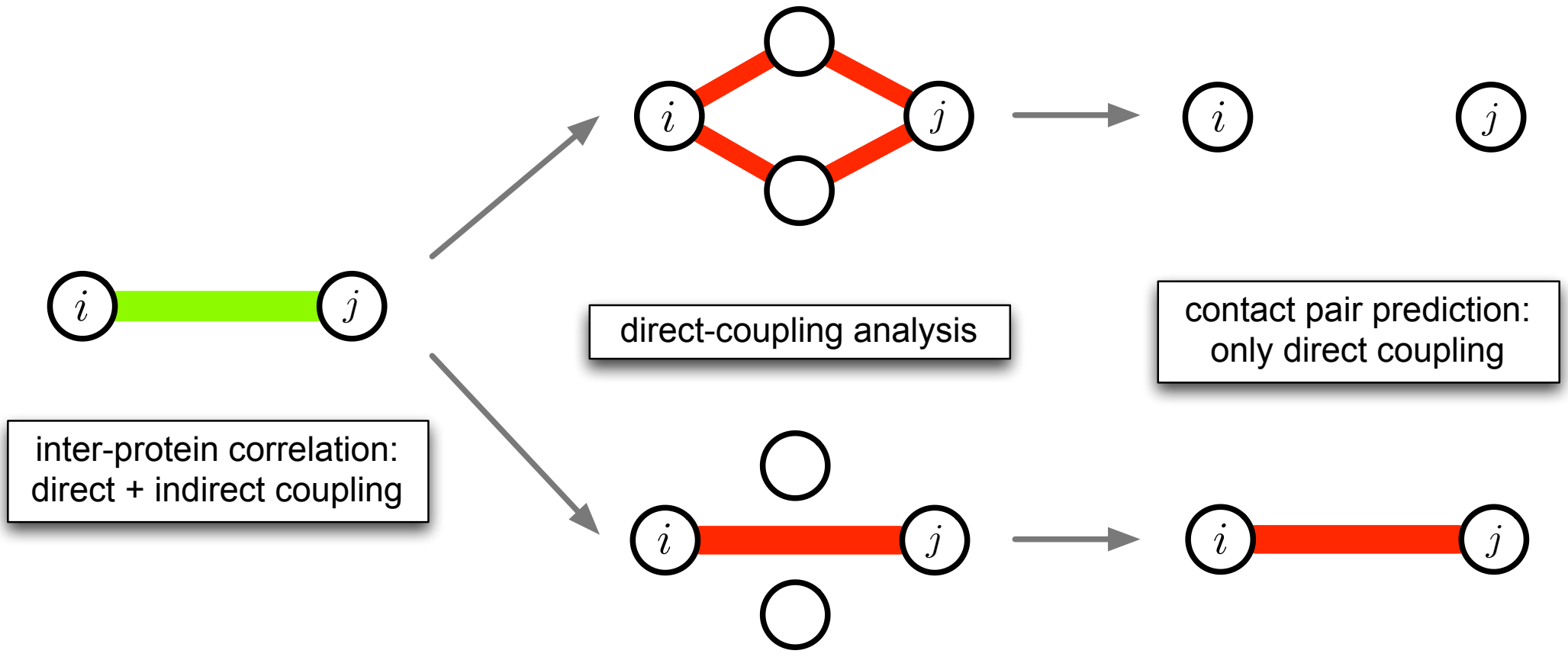
Correlations vs. residue contacts

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



- contact
- no contact

Correlation is not coupling



- correlations are generated by network of direct couplings
- 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

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

- 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

- 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

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

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

  - fluctuation-dissipation relation

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

  leads to explicit equation for couplings

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

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

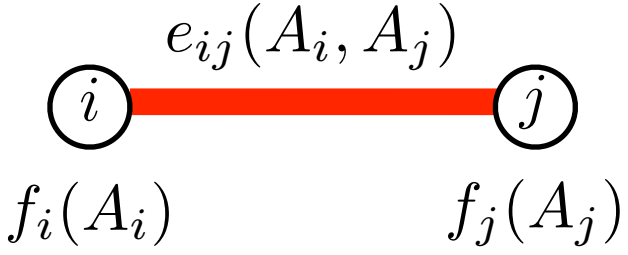
➡ more complicated approximations (Bethe-Peierls, Thouless-Anderson-Palmer) do not improve performance on biological sequence data

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

Interaction strength and direct information

How to quantify direct interaction by scalar quantity:

➡ consider isolated two-spin system

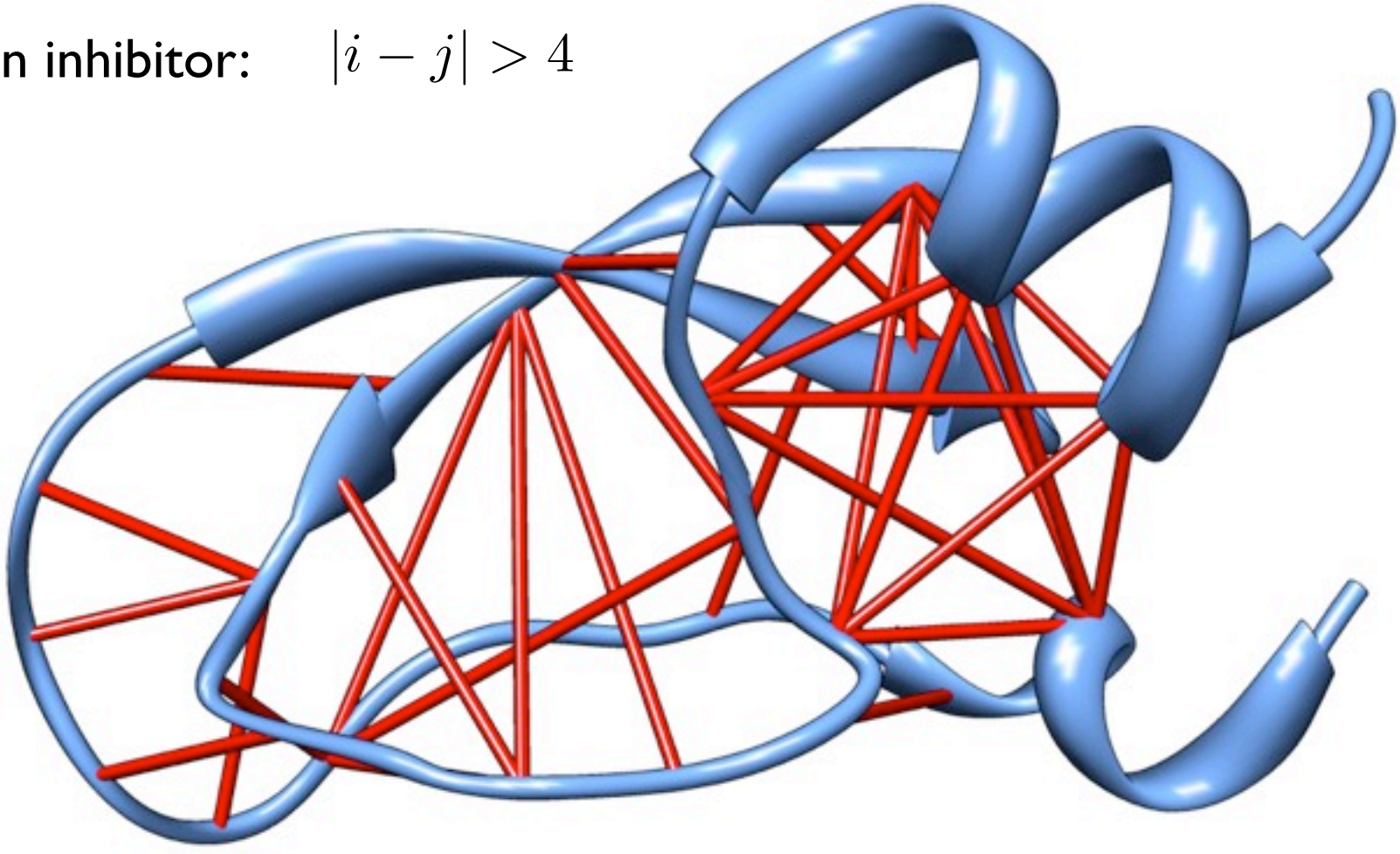


➡ direct information = mutual information due to direct coupling

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

Couplings vs. residue contacts

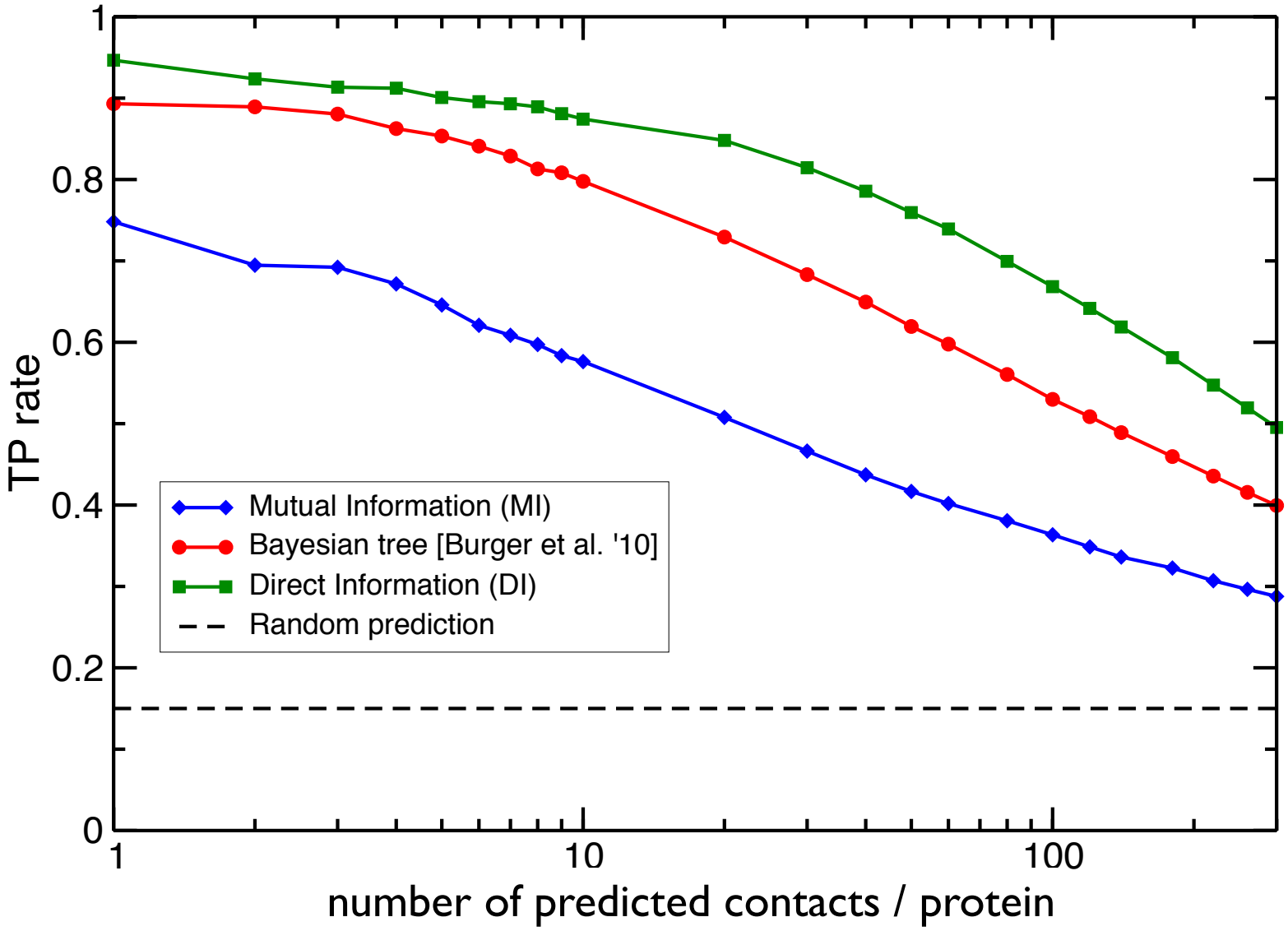
Trypsin inhibitor: $|i - j| > 4$



- contact
- no contact

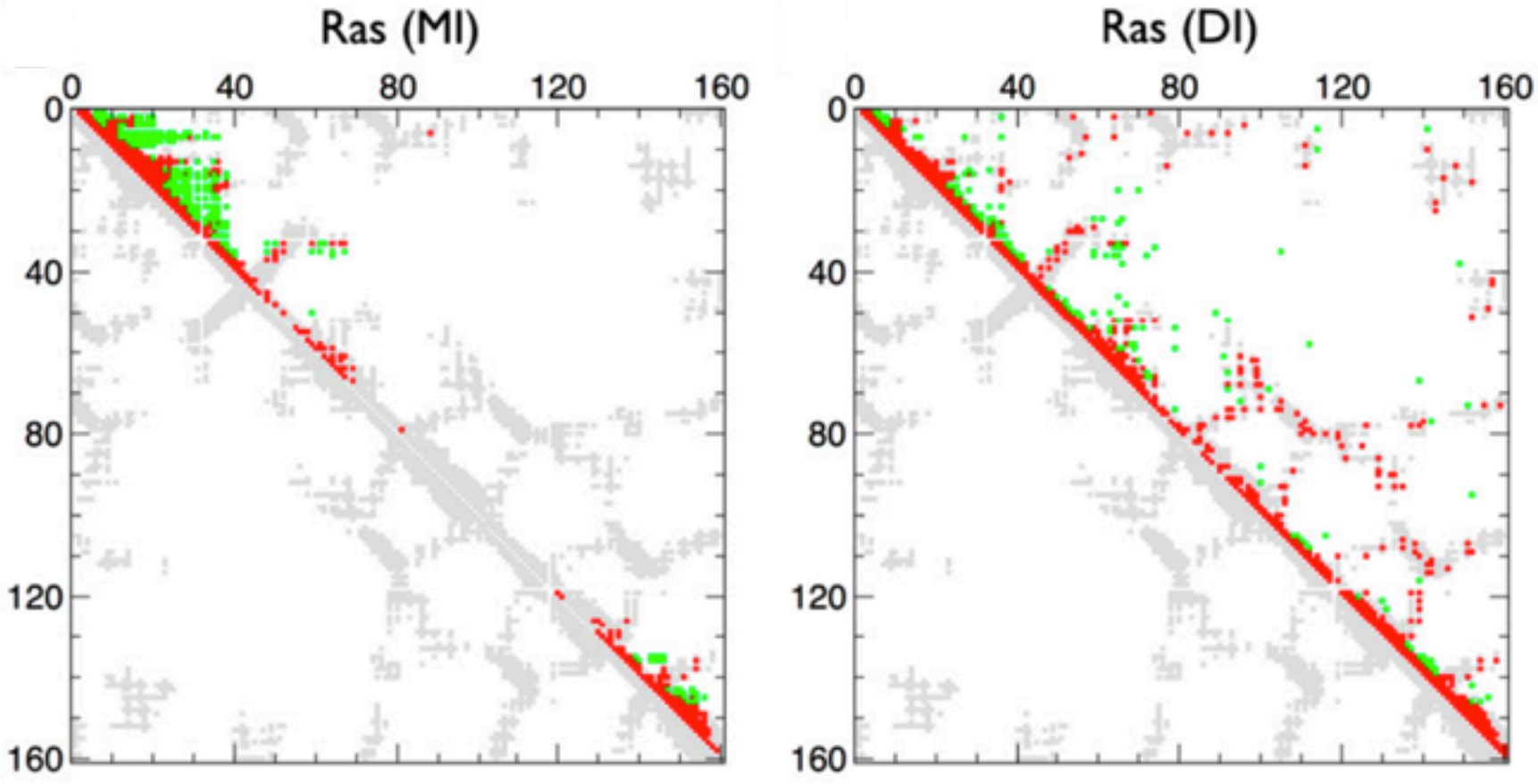
Couplings vs. residue contacts

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



DCA strongly improves contact prediction!

Not all contacts co-vary, but...



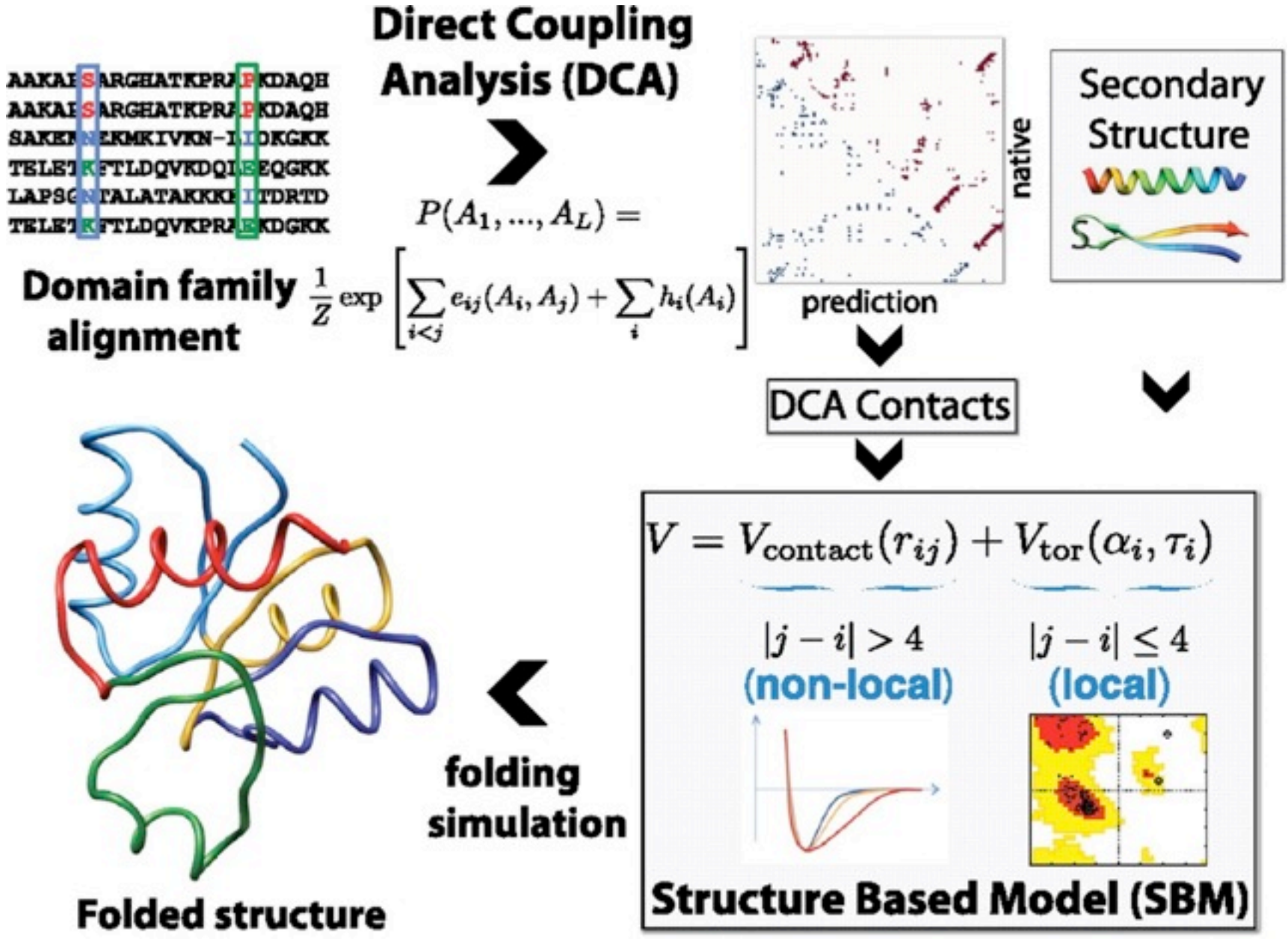
..can guide protein complex assembly

[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



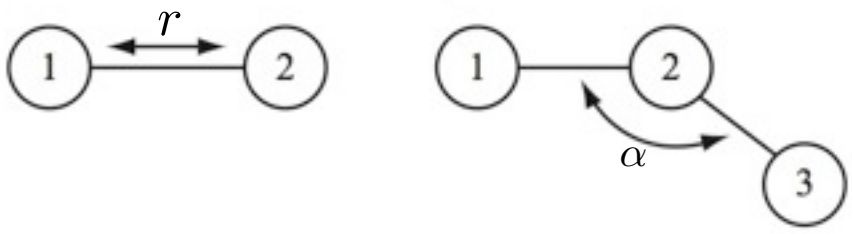
[Sulkowska, Morcos, MW, Hwa, Onuchic, PNAS '12]

...global protein structure defined

ab initio protein folding simulations:

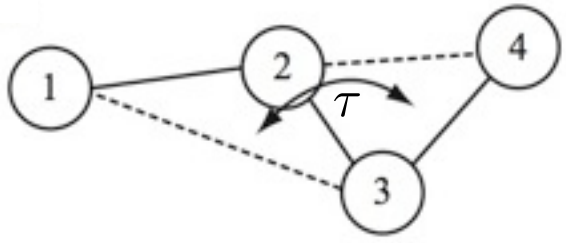
- molecular-dynamics simulations of structure-based models (Go-models):

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



with

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



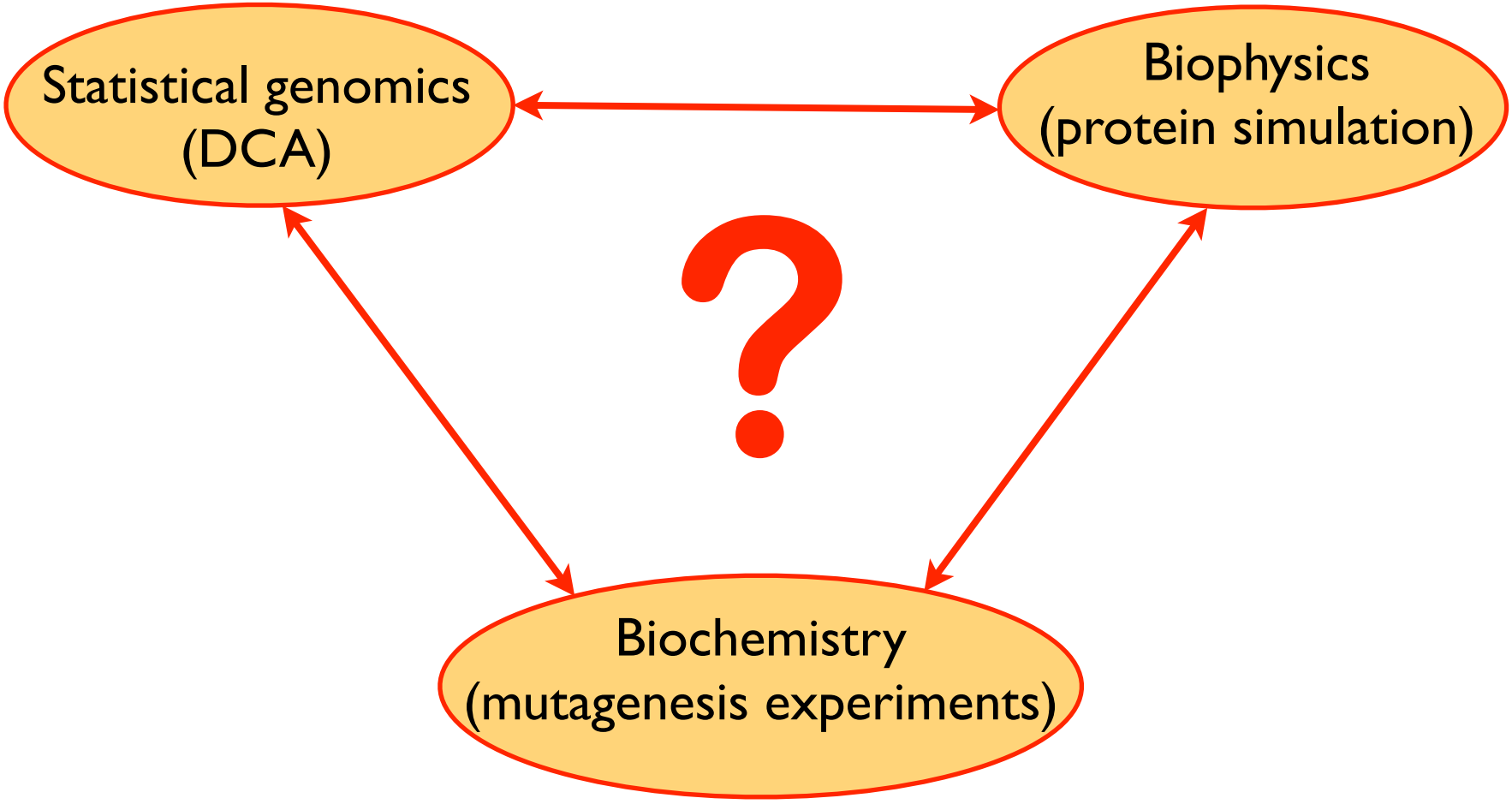
$$V_{torsion} = k_a \sum_{angles} (\alpha - \alpha_0)^2 + k_d \sum_{dihedral} [1 - \cos(\tau - \tau_0)] + \frac{1}{2}[1 - \cos 3(\tau - \tau_0)]$$

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

use only DCA contacts

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

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

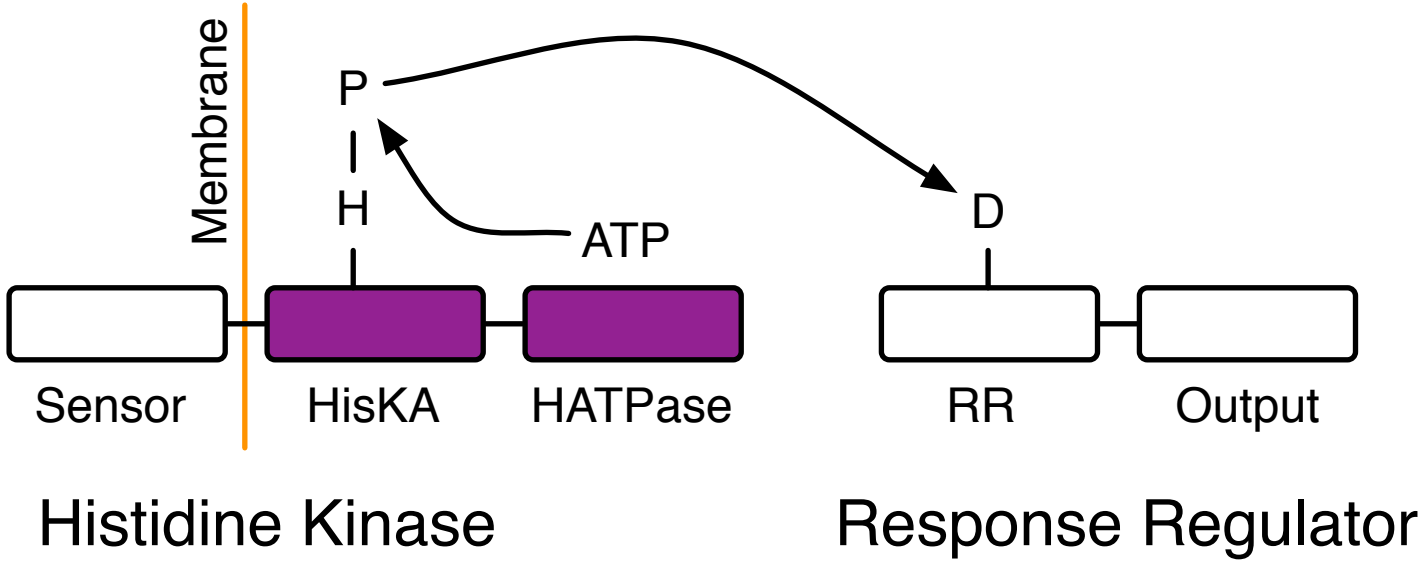


[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS '12]

Histidine-kinase auto-phosphorylation complex

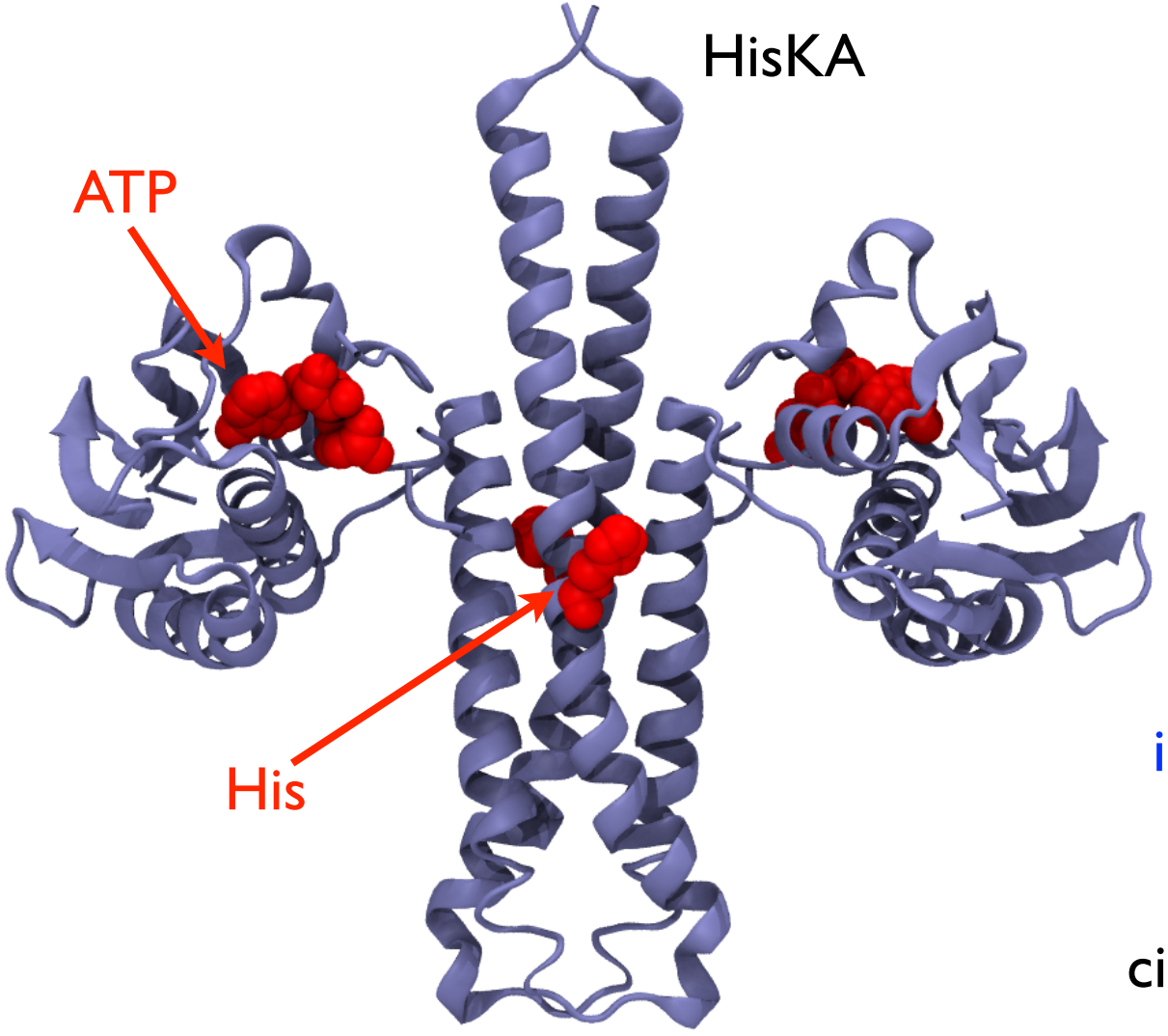
Two-component signaling system

- 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



inactive form of kinase
ATP distant from His

cis- vs. trans-interaction?

HK853
(*Thermotoga maritima*)

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

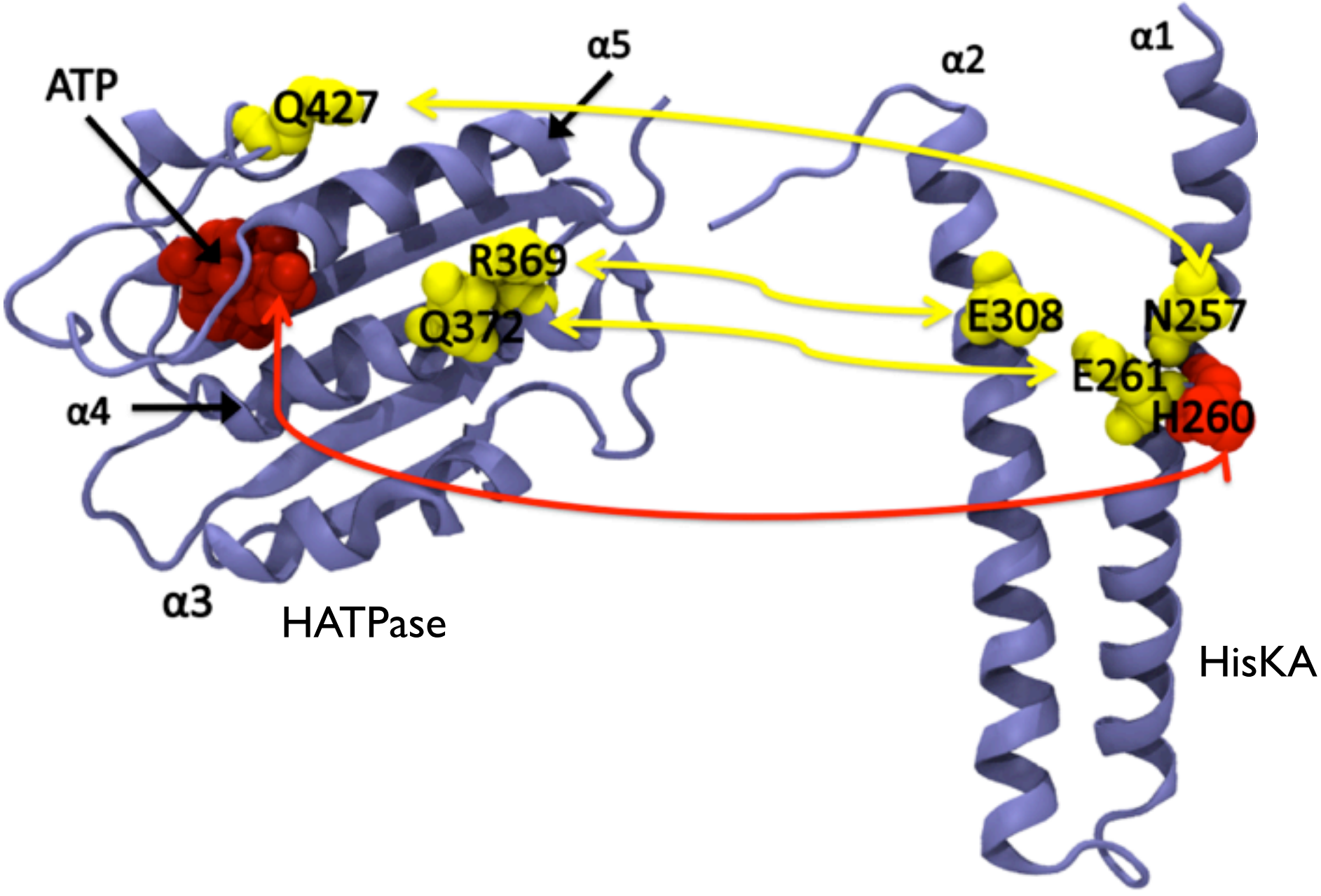
DCA results

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

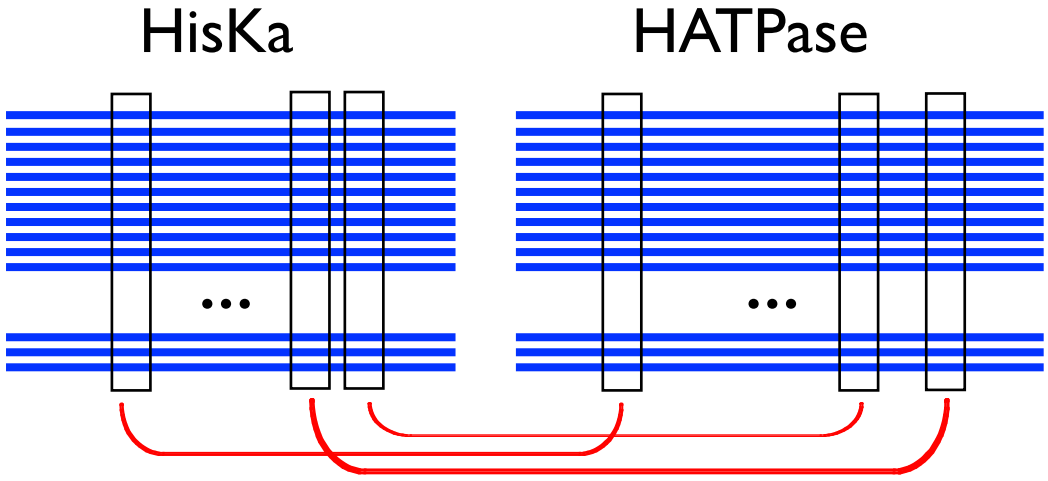
First 36 DI-ranking pairs:

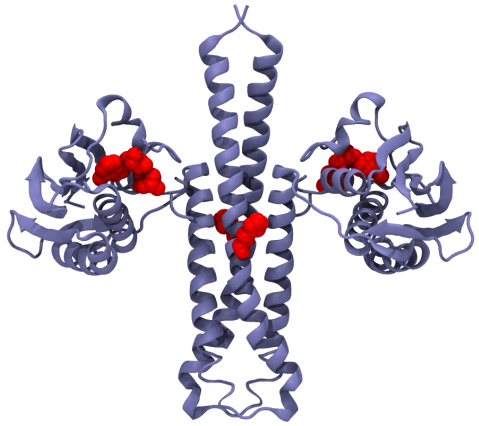
- 31 intra-domain pairs
  - 28 in contact
  - 3 distant
    - >90% TP rate
- 5 inter-domain pairs
  - 2 in contact
  - 3 distant
    - predicted contacts in active structure

DCA predicts 3 inter-domain contacts for auto-phosphorylation complex



DCA-guided molecular dynamics simulations

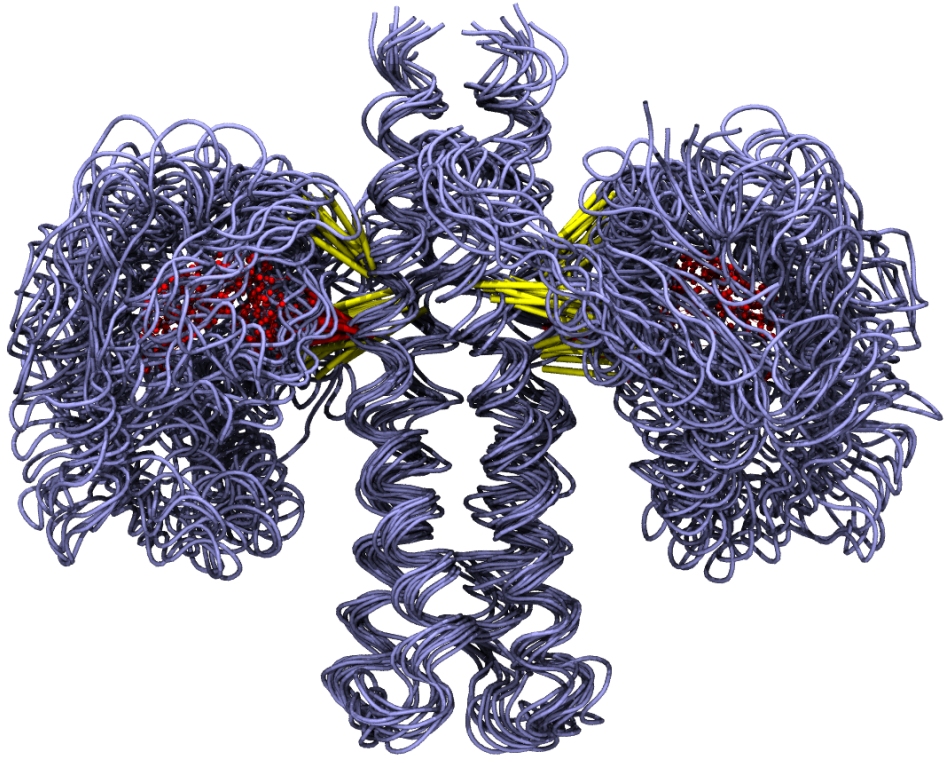




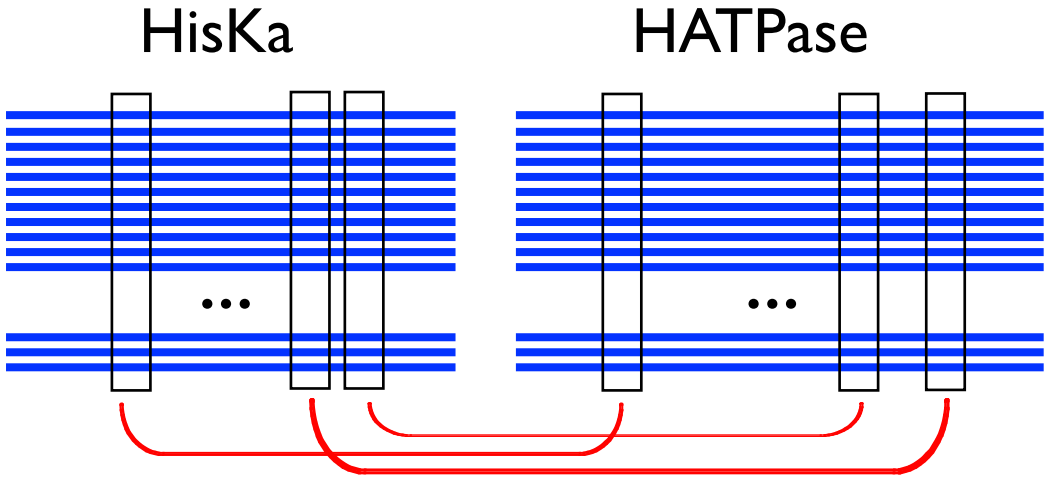
DCA-predicted inter-domain contacts

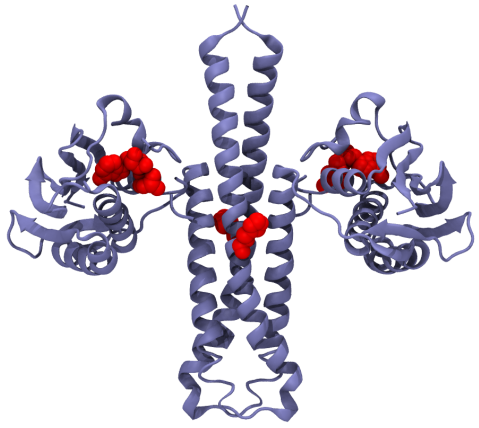
inactive protein structure

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



DCA-guided molecular dynamics simulations



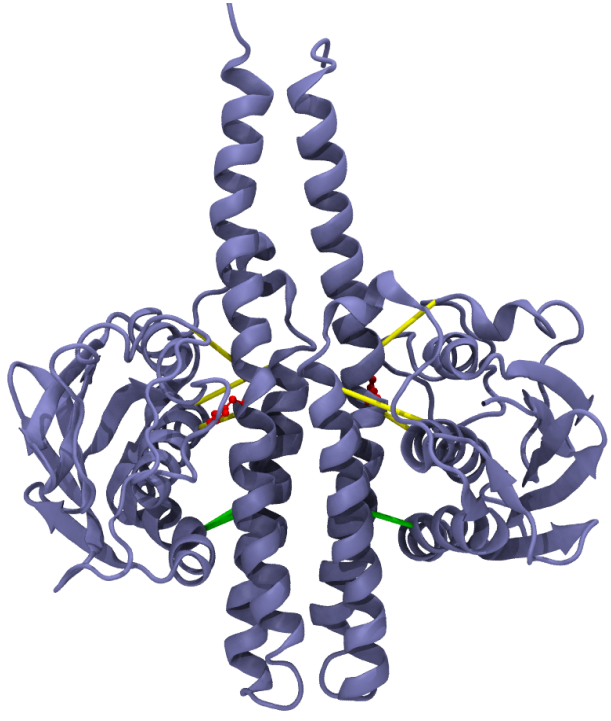


DCA-predicted inter-domain contacts

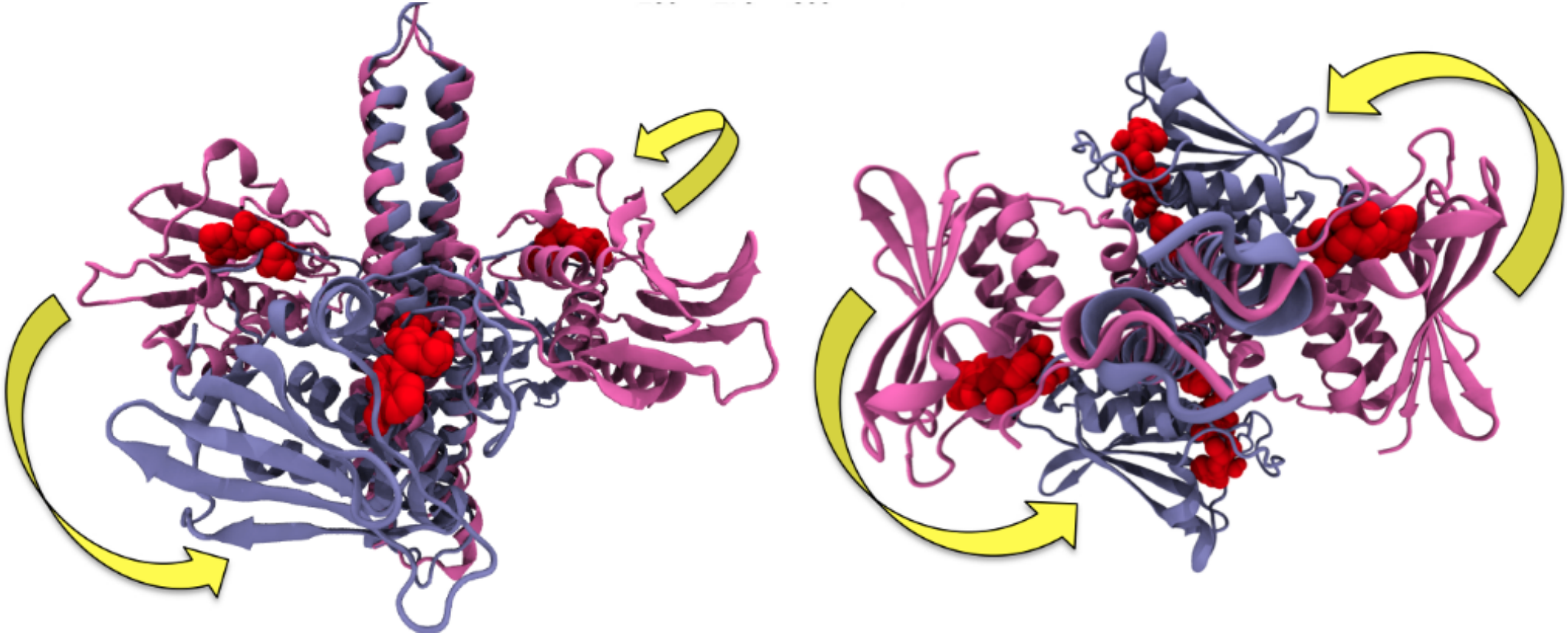
inactive protein structure

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

MD in realistic force
field
(Amber, Gromos)

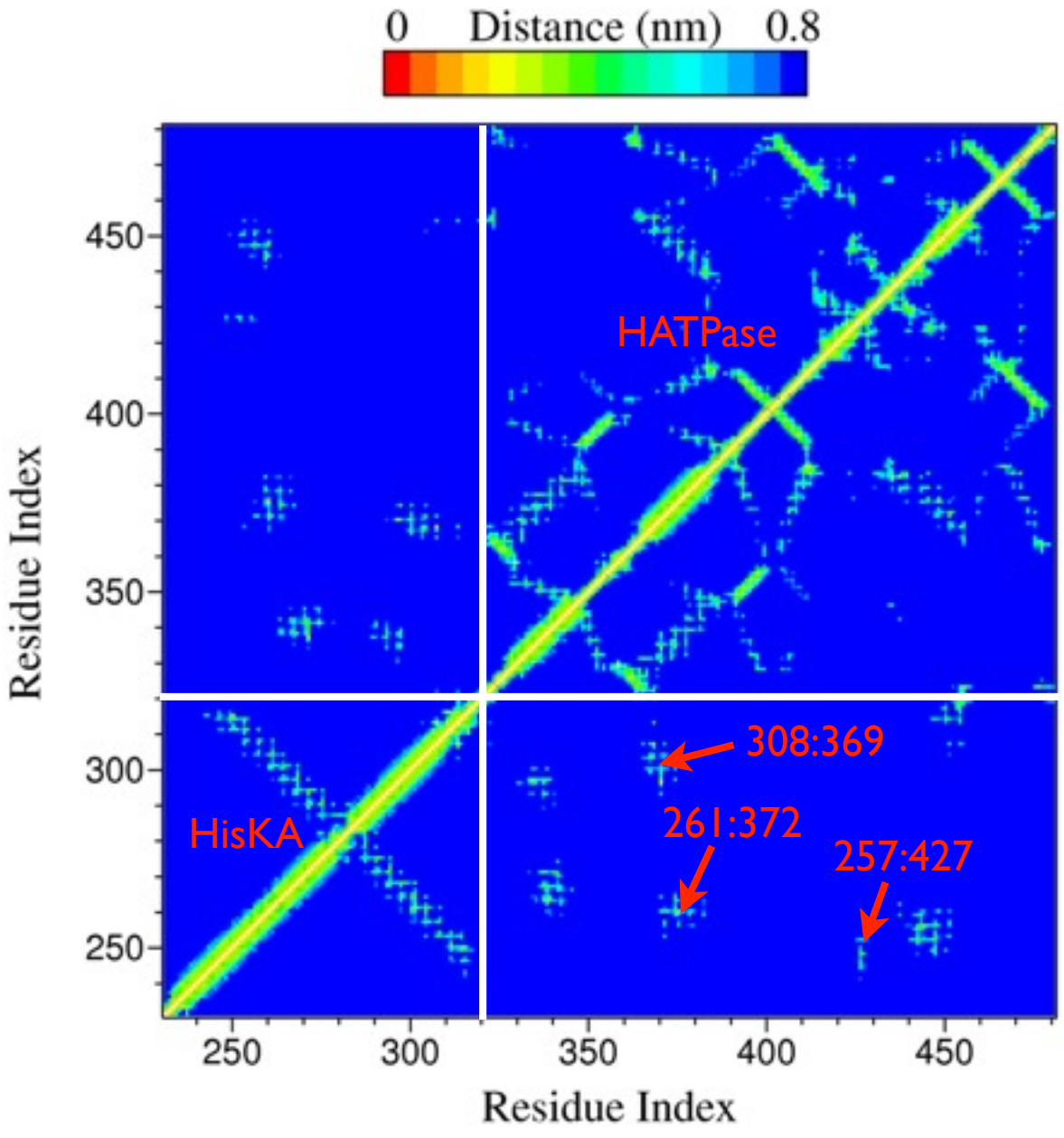


Comparison of the active / inactive structure



➡ major conformational change:
ATP close to Histidine residue

Predicted contact map



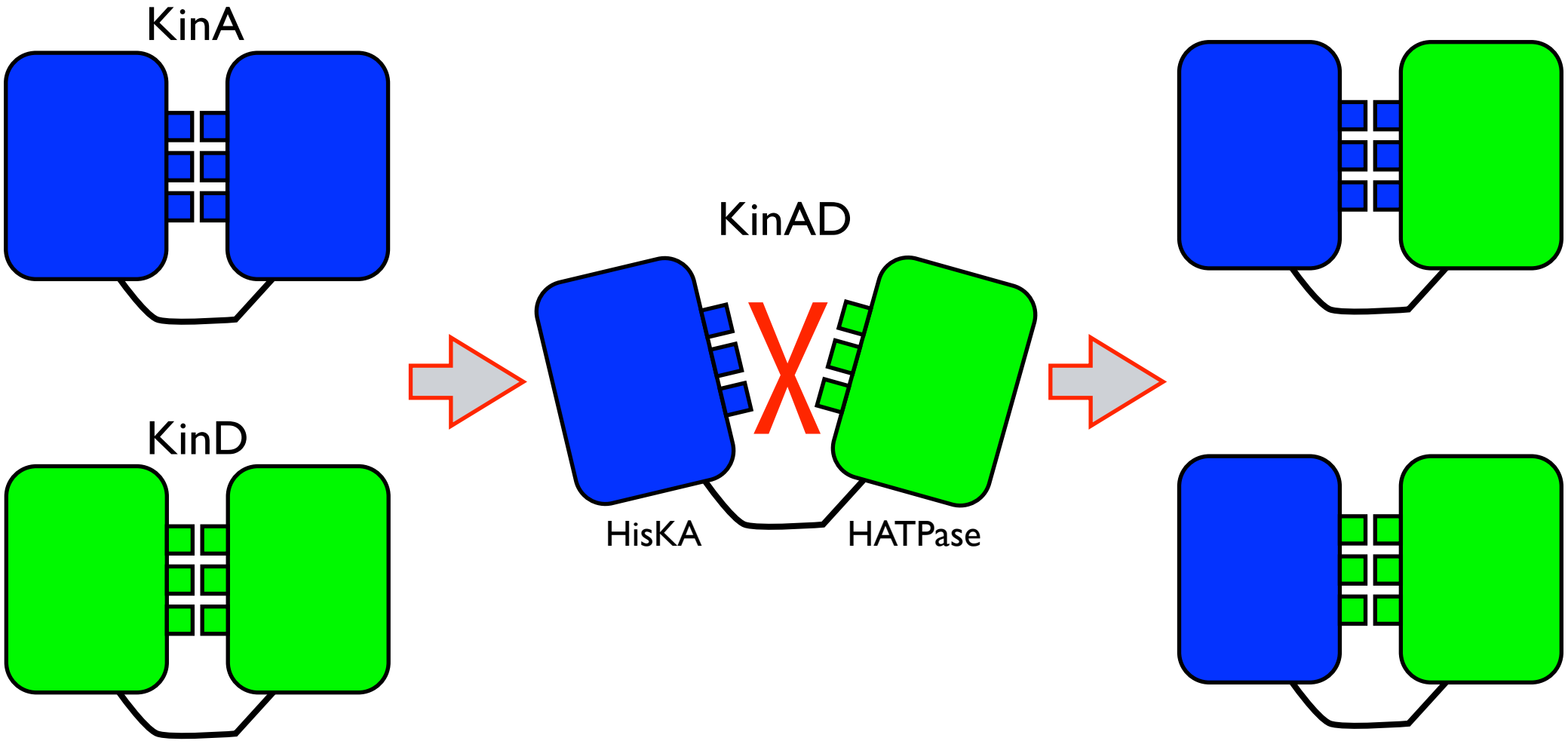
- DCA-predicted pairs in stable contact
- represent clusters of contacts
- MD predicts clusters of contacts with helix 3
  - not seen by DCA

Experiment:

verify contacts with helix 3!

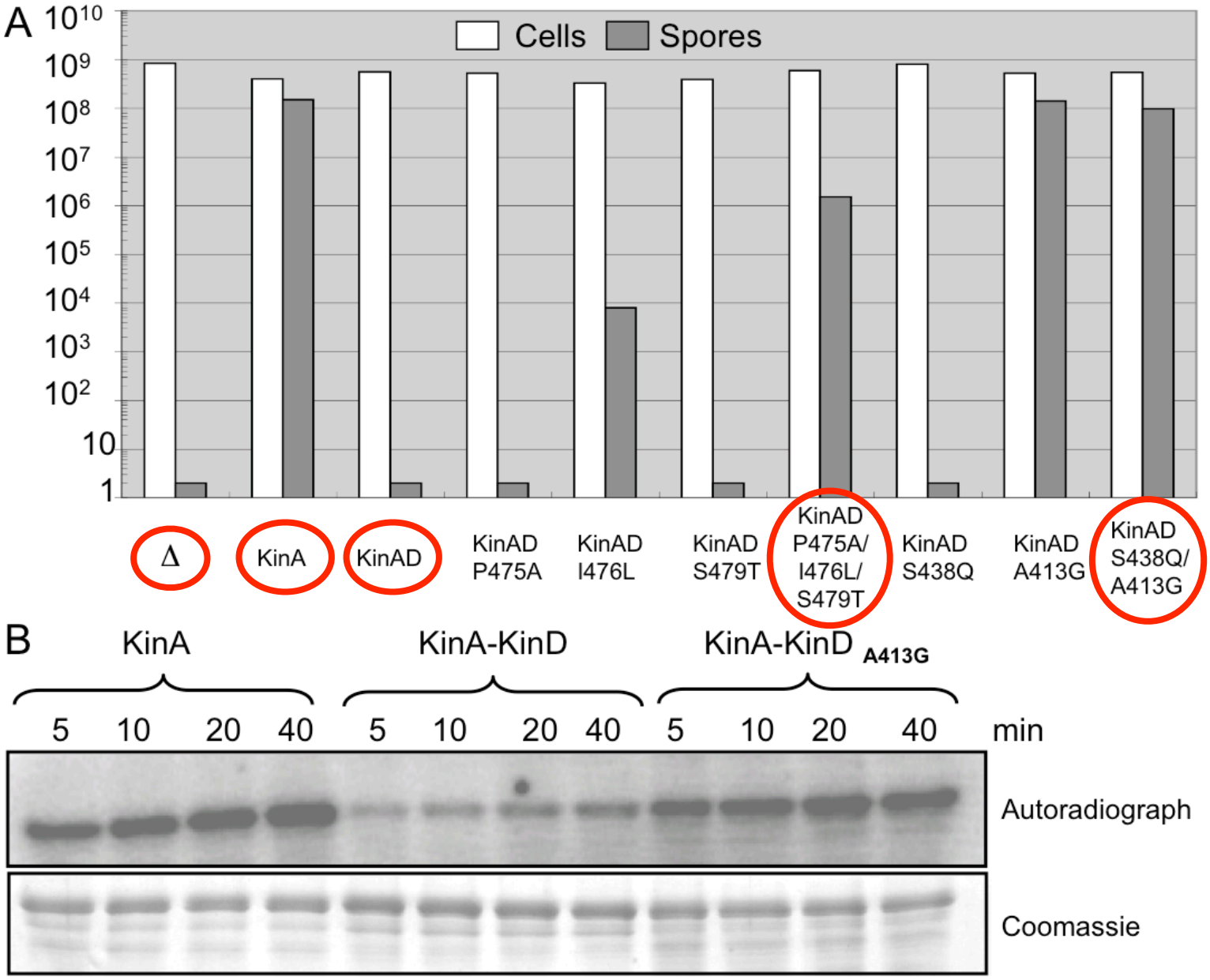
Repairing hybrid kinases

Bacillus subtilis sporulation kinase: KinA as experimental test system

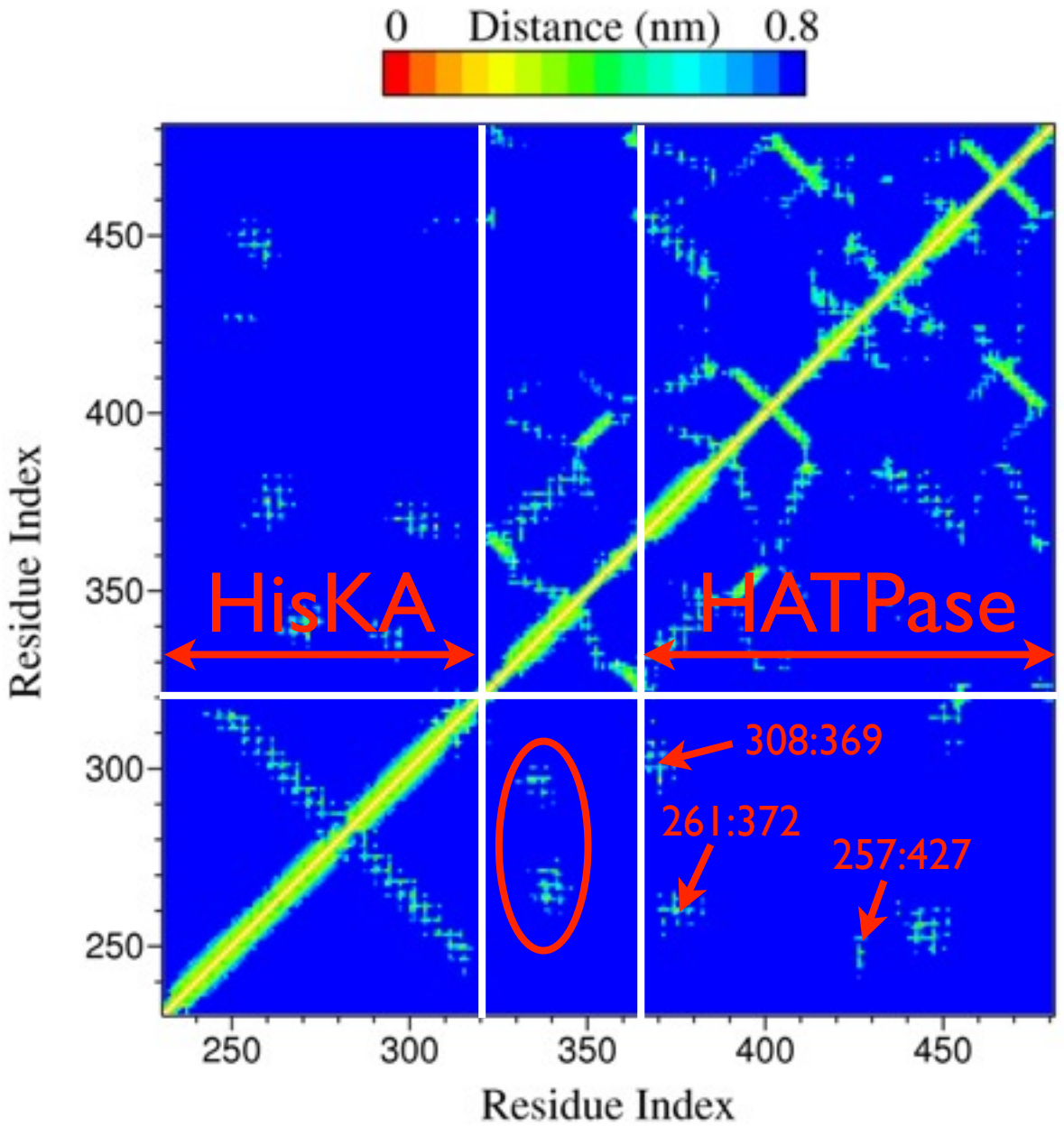


➡ exchange contact residues in one domain by those in cognate domain

Substituting contacts helix 3 - HisKA



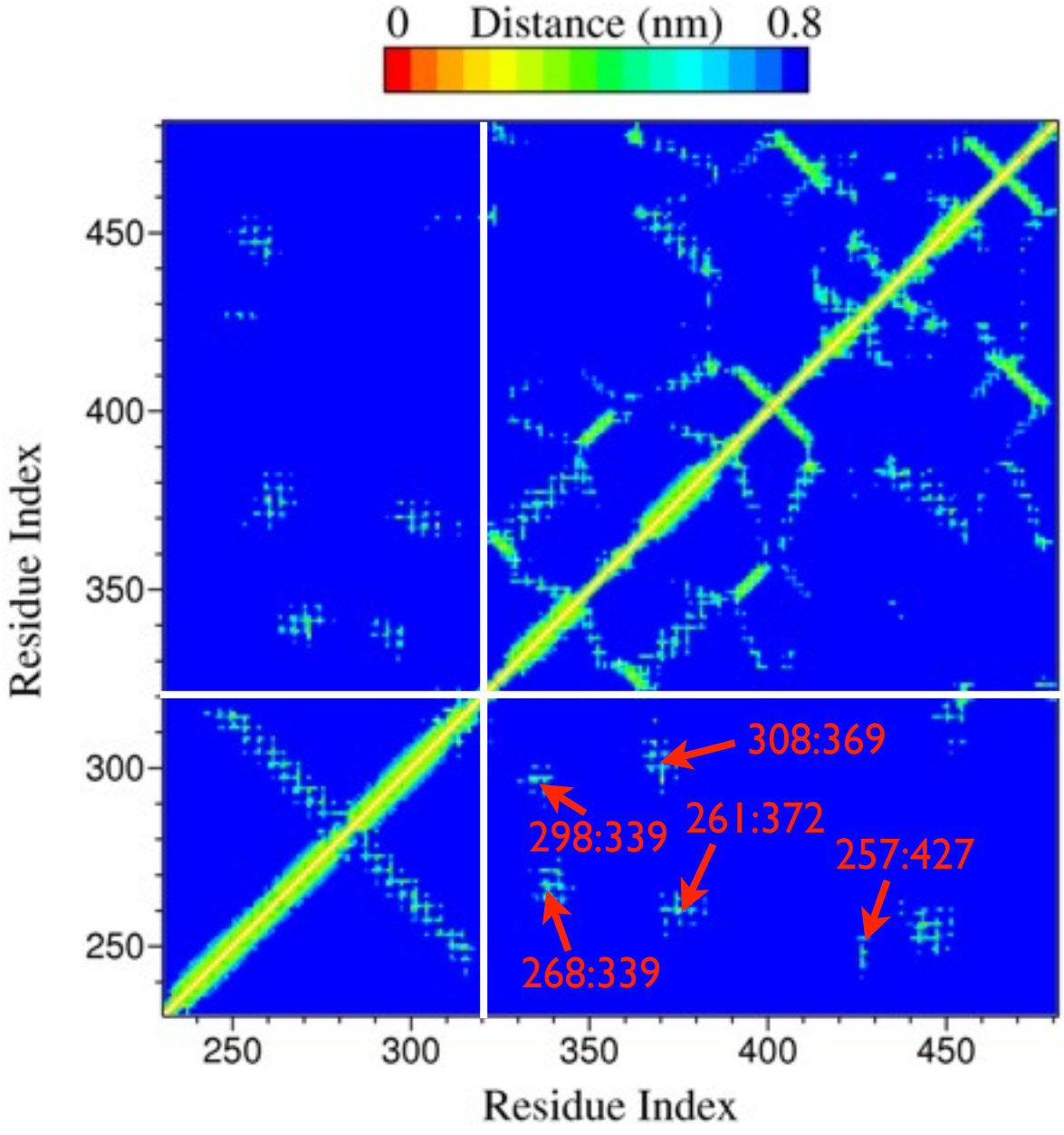
Why don't we see these residues in DCA?

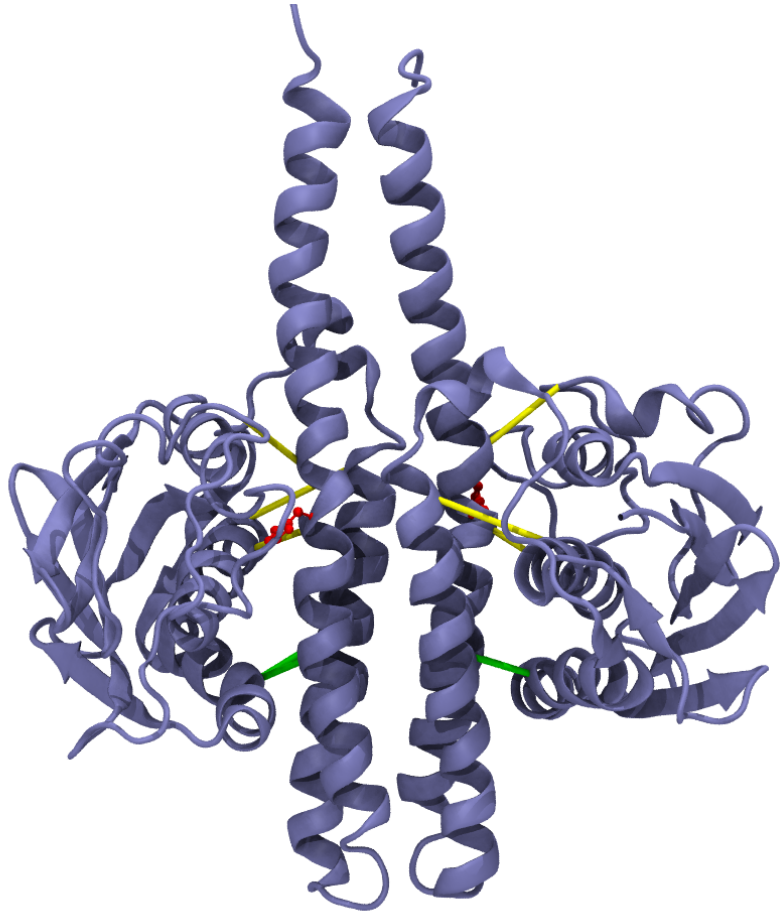


Pfam domains did not cover helix 3!

Improved contact prediction

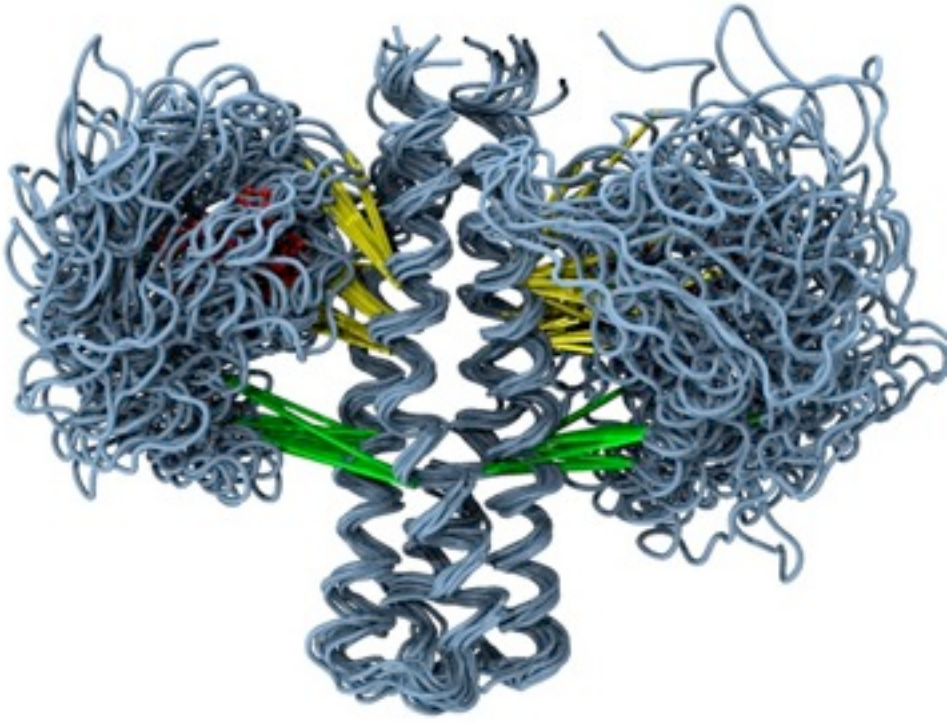
by using a full-length alignment



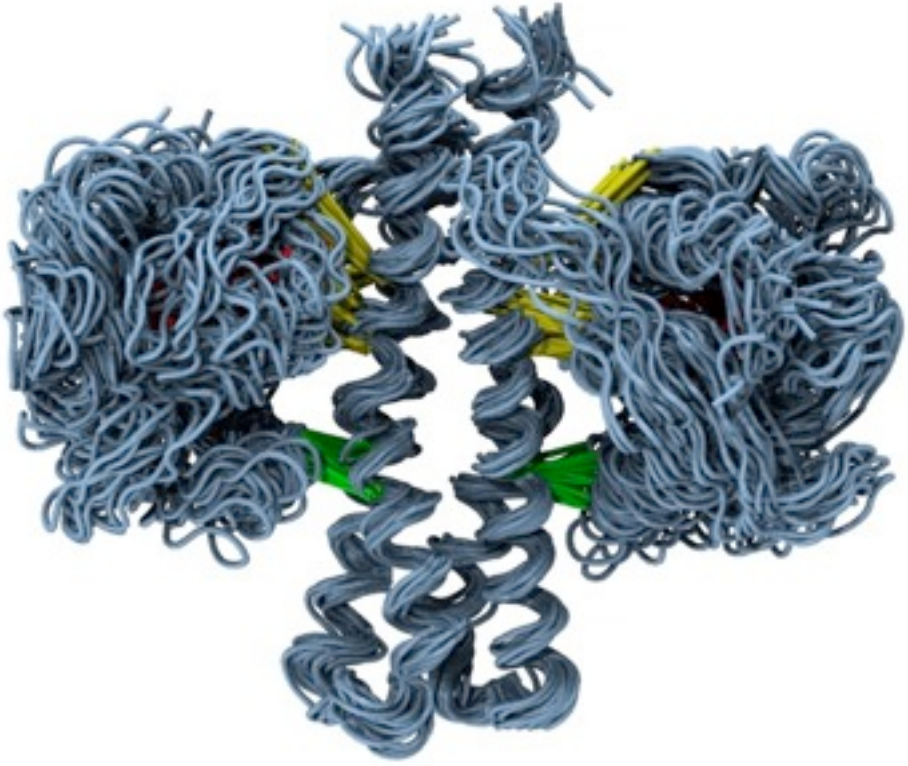


[Dago, Schug, Procaccini, Hoch, MW, Szurmant, PNAS '12]

Improved Go-model structure



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

Thanks to

HuGeF Torino:
Andrea Pagnani
Andrea Procaccini

UC San Diego:
Terry Hwa
Bryan Lunt

Rice University:
Jose Onuchic
Faruck Morcos

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

Karlsruhe Institute of Technology
Alexander Schug

Ecole Normale Supérieure
Rémi Monasson
Simona Cocco