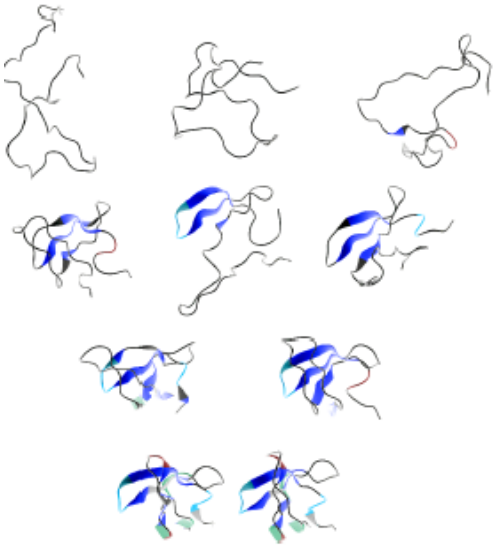
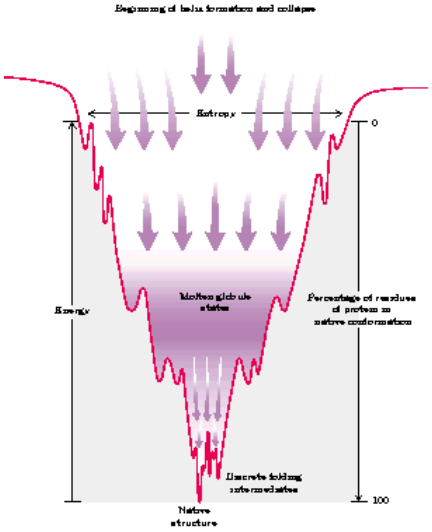
From structure to function in proteins: the convergence of structure based models and co-evolutionary information



AAKAPSARGHATKPRAPKDAQHEAA
AAKAPSARGHATKPRAPKDAQHEAA
SAKEKNEKMKIVKN-LIDKGKKSGS
TELETKFTLDQVKDQLEEQGKKRSS
LAPSGNTALATAKKKEITDRTDDPV
TELETKFTLDQVKPRAEKDGKKRSS



School on Physics Application in Biology
ICTP SAIFR

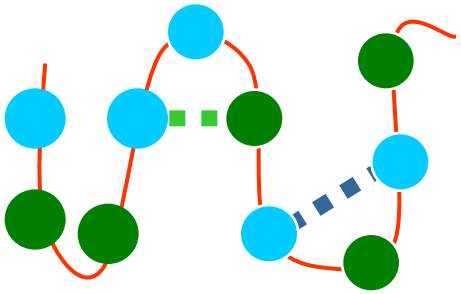
São Paulo, Brasil
22 January 2016

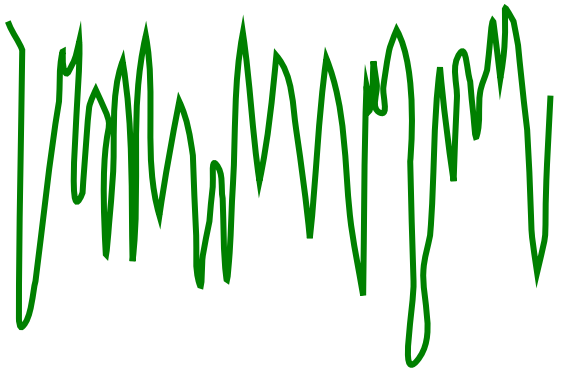
José N. Onuchic
Center for Theoretical Biological Physics
Rice University
ctbp.rice.edu

Energy Landscape Idea

- Rough Landscape

  Random Heteropolymer



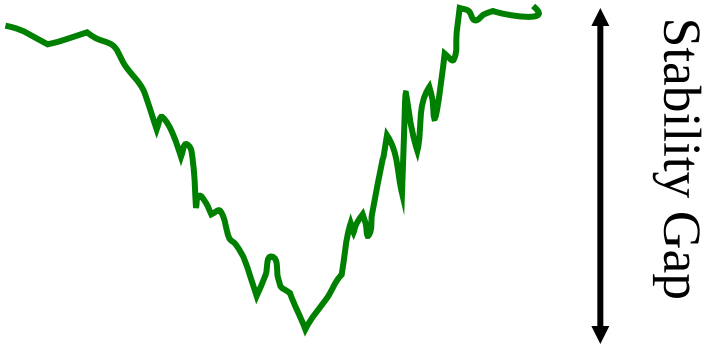


Frustration!!

- Proteins

  Try to minimize frustration!

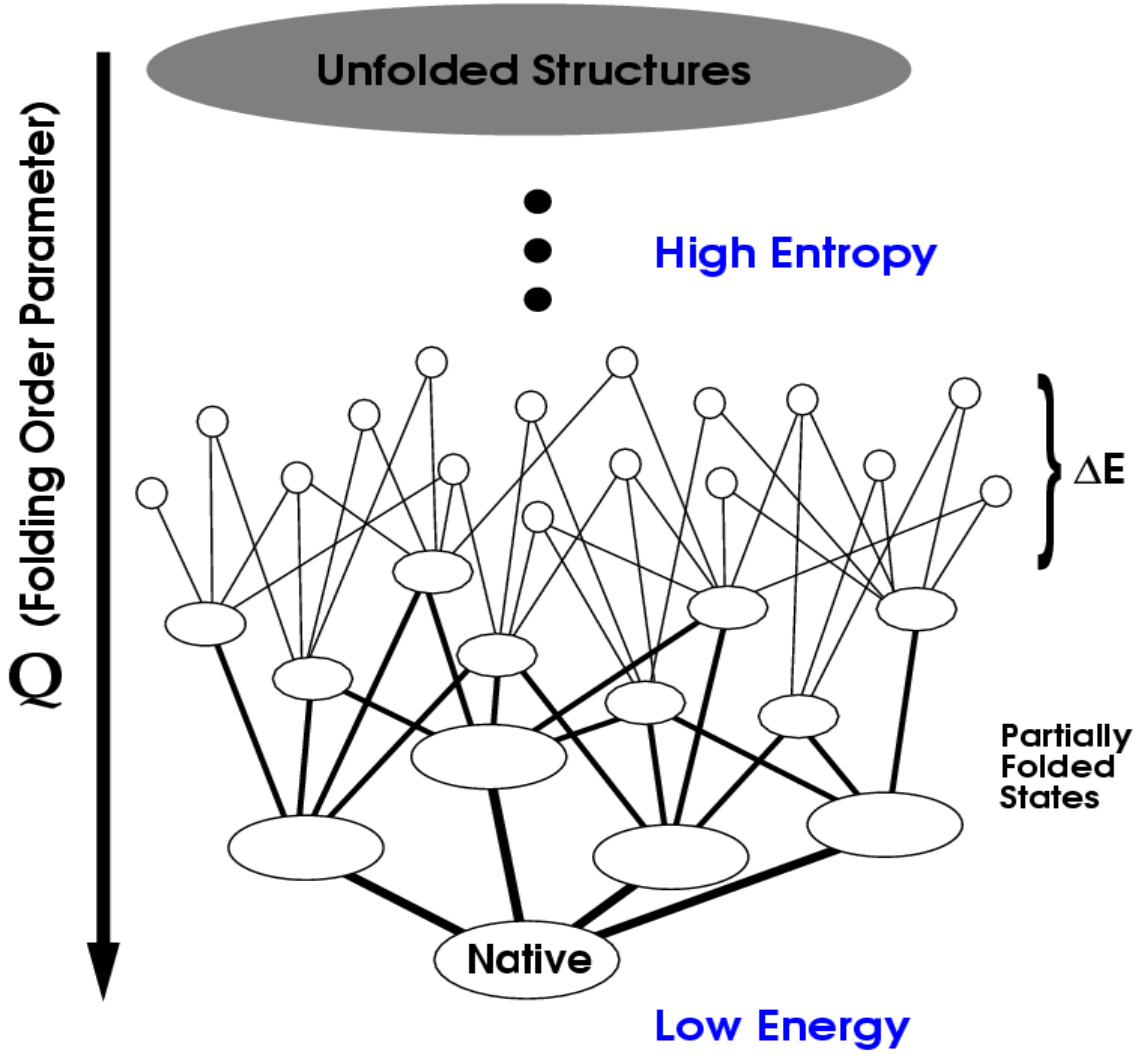
$$\frac{Tf}{Tg} \propto \frac{Stability}{Roughness}$$

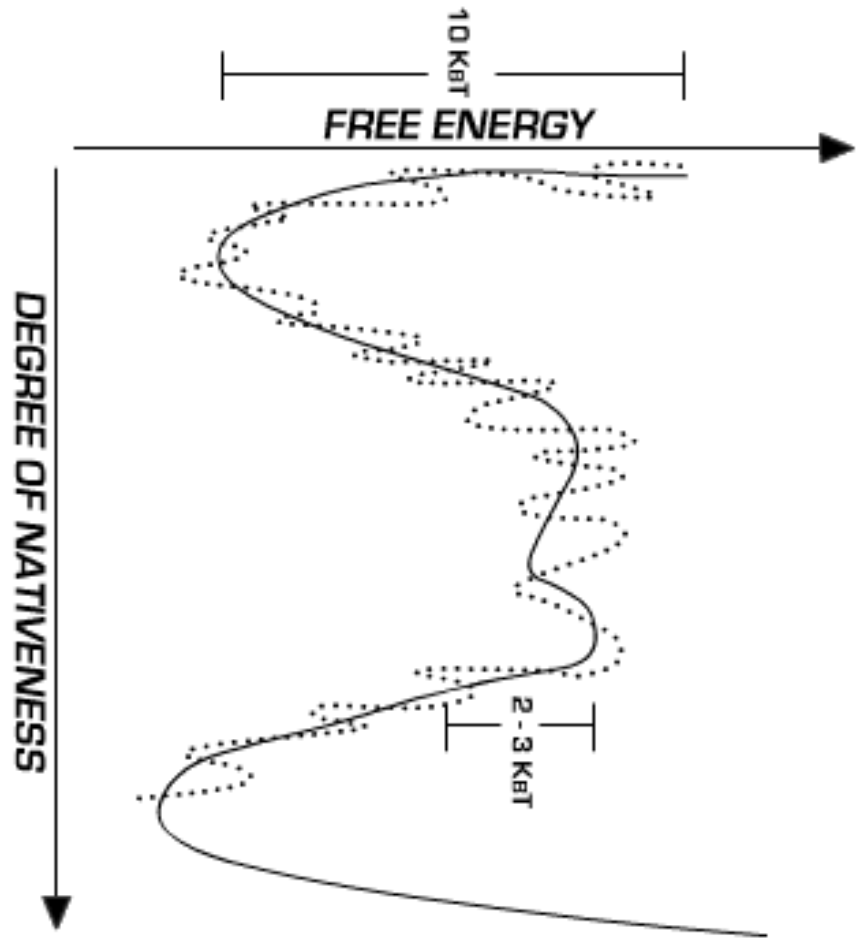


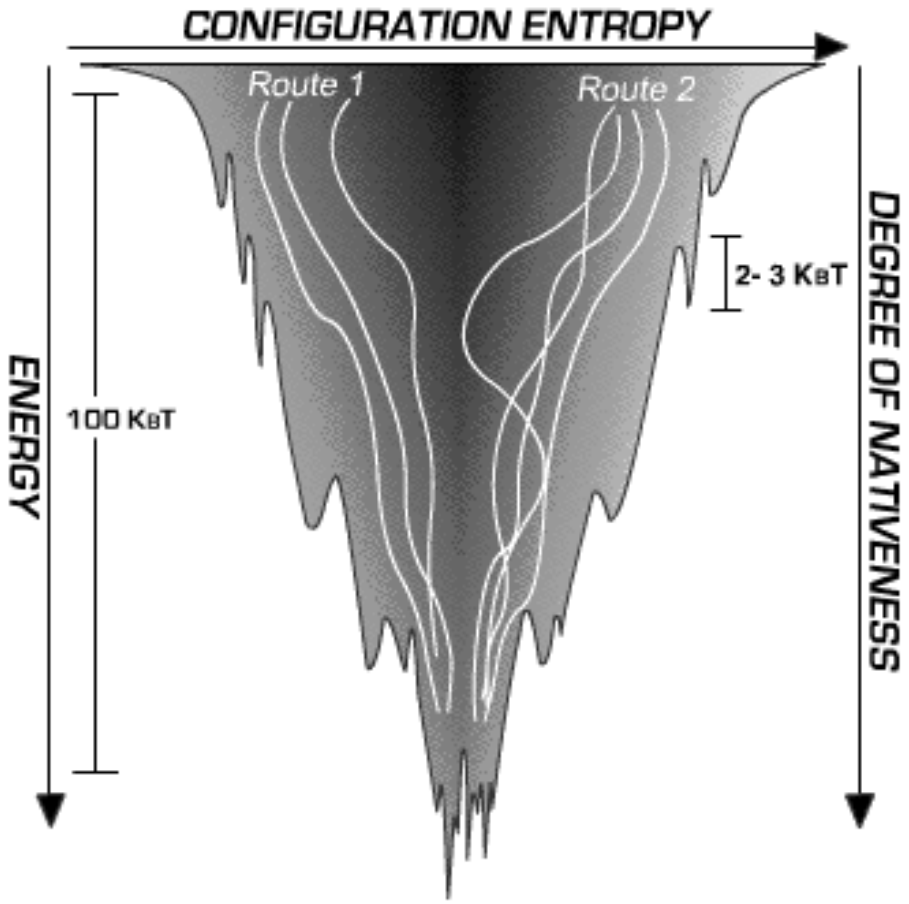
Principle of Minimal Frustration
(Bryngelson/Wolynes, Gσ)

⇒ *Realization of Minimal Frustration in Funnels*

Good Funnel: Roughness is small compared to stabilizing free energy – E linear in Q

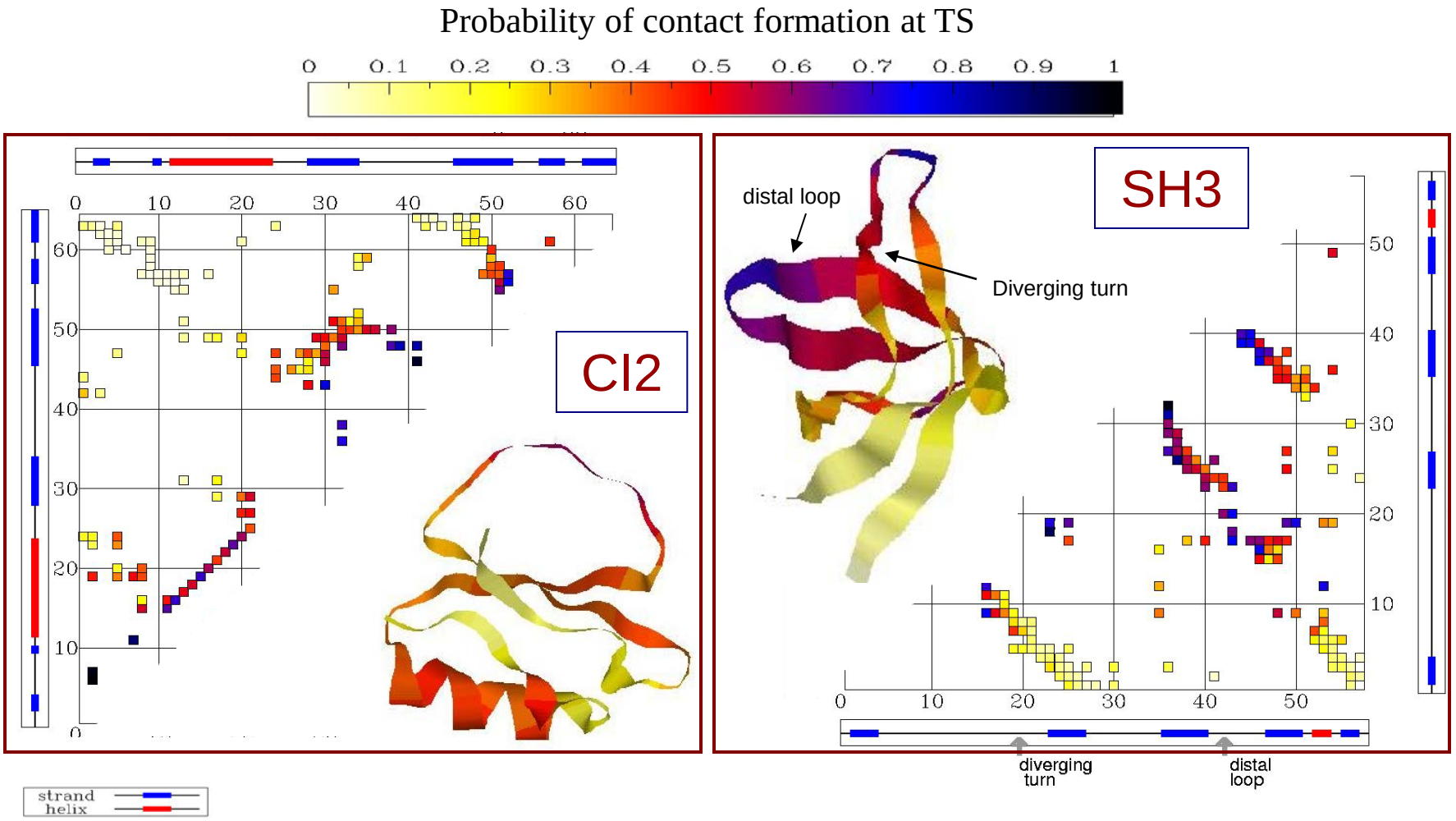






$$\Phi_i = \frac{\Delta\Delta G_i^{\ddagger}}{\Delta\Delta G_i^0} \approx -\frac{RT\ln(\mathbf{k_i}/\mathbf{k_{wt}})}{\Delta\Delta G_i^0}$$

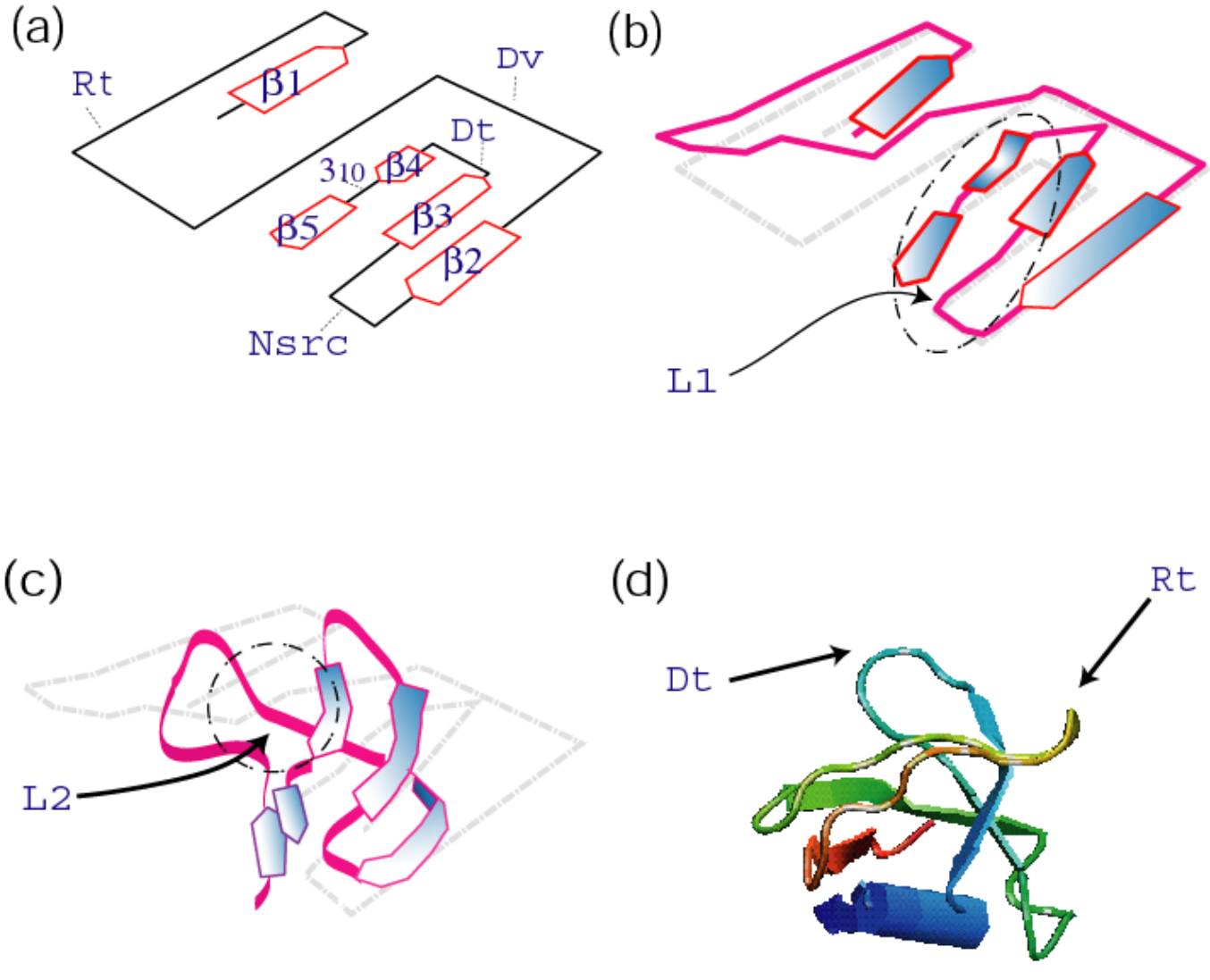
Analysis of two-state folders: Transition State structure for CI2 and SH3

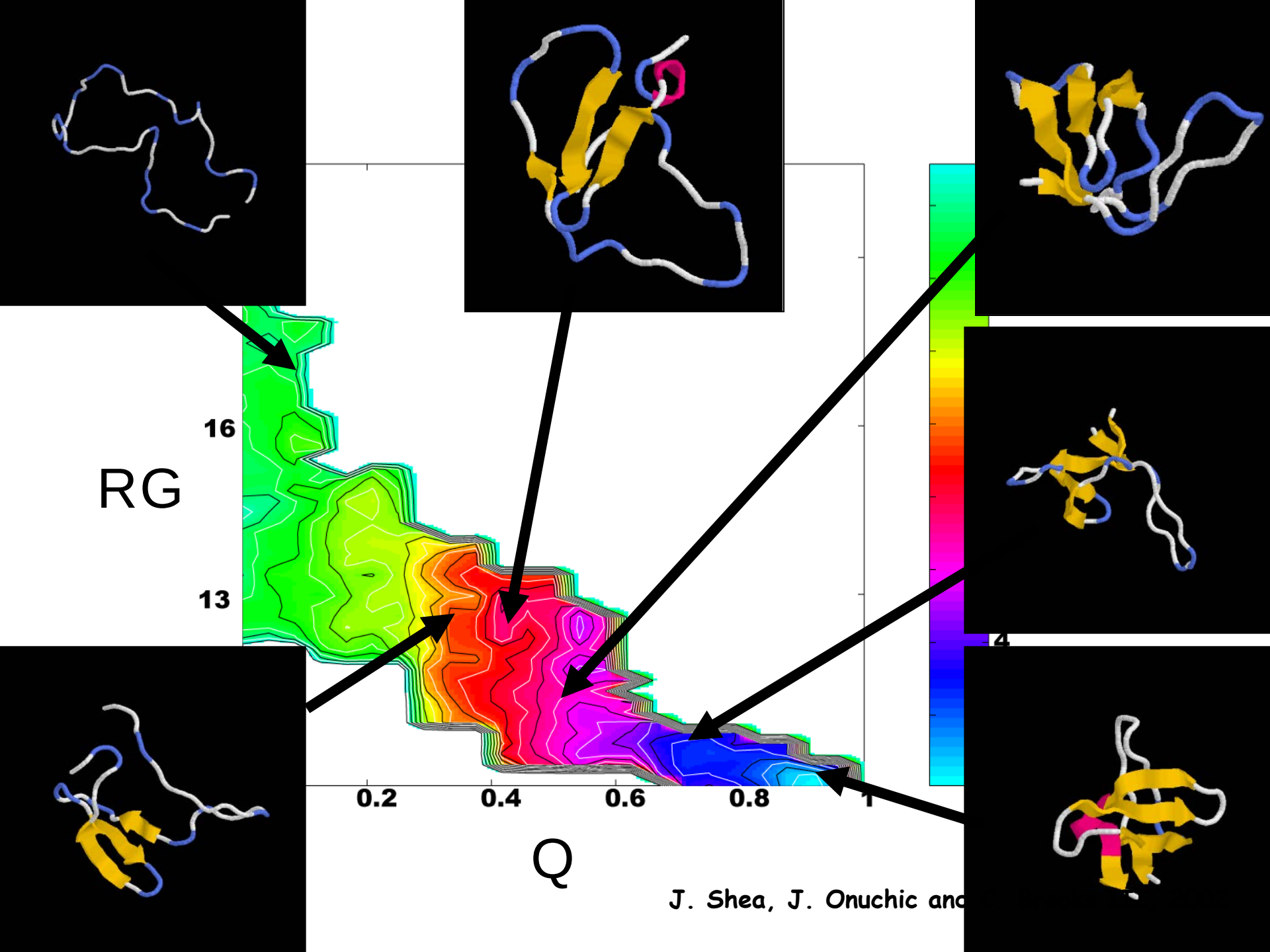


These descriptions are in good agreement with experimental results (Jackson & Fersht 1991, Grantcharova et al. 1998).

C.Clementi, H.Nymeyer, J.Onuchic, JMB 2000

What are the folding routes for SH3?




RG
16
13
0.2
0.4
0.6
0.8
1
Q
J. Shea, J. Onuchic and

Structure-based models of proteins

- Theory $\Longrightarrow$ Reduced Models
  - Native interactions are on average more stabilizing than non-native.
  - "Perfect funnel" structure-based models are the limiting case, force field completely specific to a native configuration
  - Baseline model can be extended by including non-native interactions, e.g. Debye-Huckel electrostatics or transferable backbones

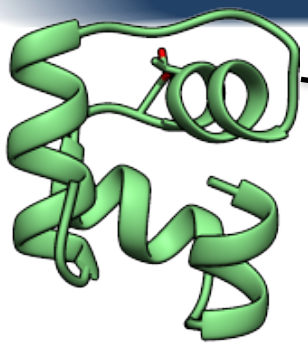
Structure Based All-Atom Model - all available at smog-server.org

Whitford, Noel, Gosavi, Schug, Sanbonmatsu & Onuchic (2009) *Proteins,* 75, 430-441. (Protein forcefield)

Whitford, Schug, Saunders,Hennelly, Onuchic & Sanbonmatsu (2009)*Biophys.J.* 96,L7-9 (RNA forcefield)

Extending the funnel ideas towards situations that we have limited information….

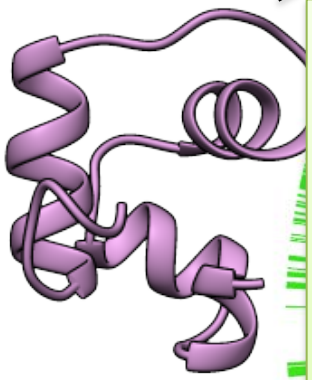
Mutations throughout history provide **examples** on **structure** conservation



Organism:
B. fragilis

Helix-Turn-Helix Phylogeny (HTH_3)

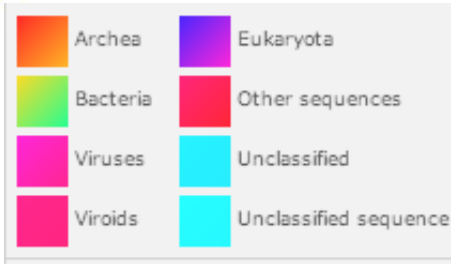
Organism:
Phage 434



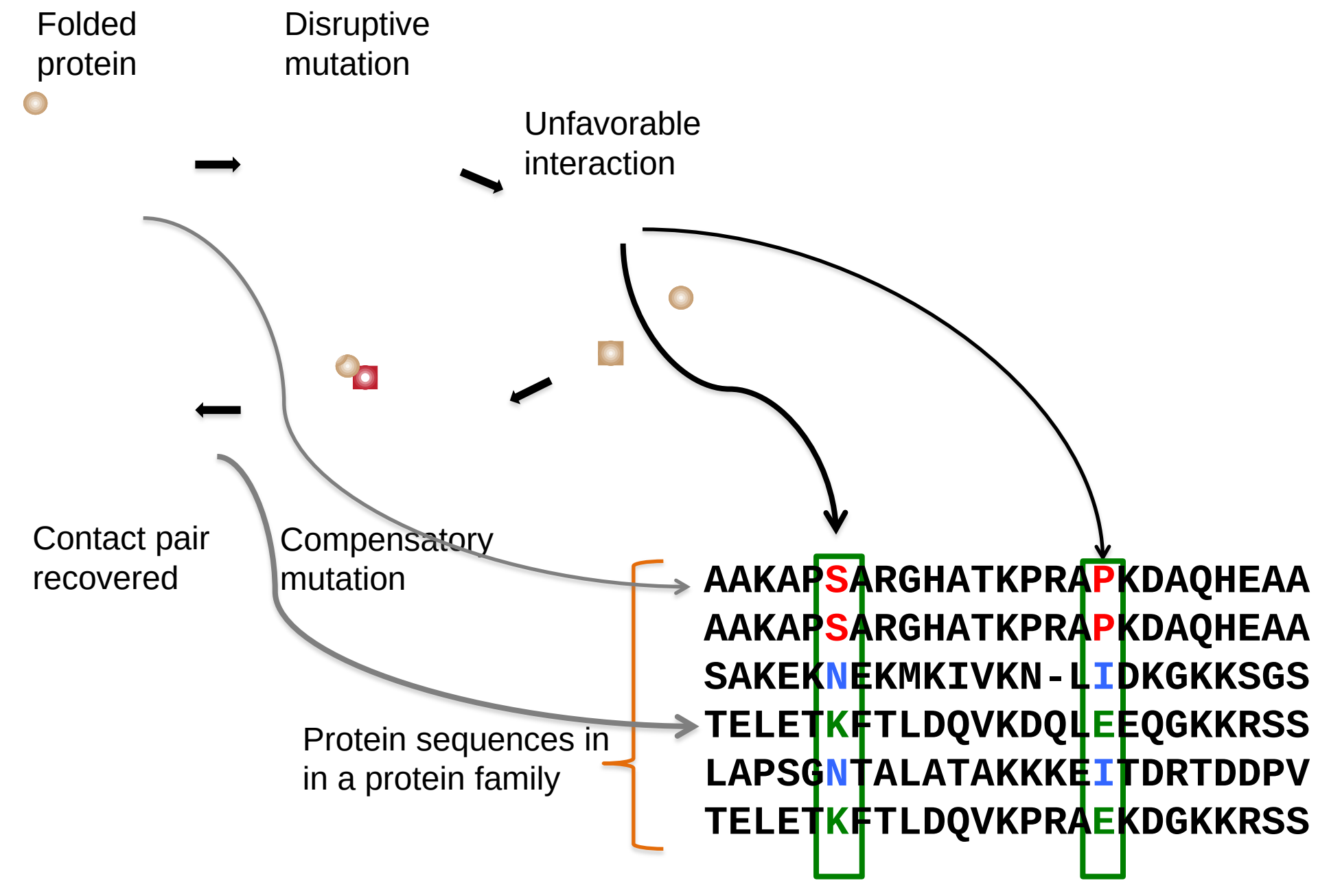
Organism:
N. gonorrhoeae

```
AAKAPSARGHATKPRAPKDAQHEAA
AAKAPSARGHATKPRAPKDAQHEAA
SAKEKNEKMKIVKN-LIDKGKKSGS
TELETKFTLDQVKDQLEEQGKKRSS
LAPSGNTALATAKKKEITDRTDDPV
TELETKFTLDQVKPRAEKDGKKRSS
```

Organism:
C. difficile 630



Residue-residue coevolution maintains protein structure



Identification of directly co-evolving residues is a challenge

Analysis of Correlated Residue Pairs

- The **problem of directly coupled residue pairs** has been a long standing challenge due to several factors:
  1. Correlations can be **direct** (e.g. physical contacts) or **indirect** (chain effects)
     - The use of **local statistical models**

       e.g. Mutual Information

  2. **Noisy or incomplete data**

- **Recently the panorama has changed:**

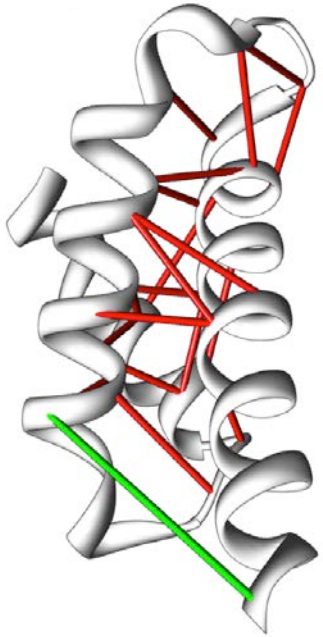
  Increase of sequence information

  - A formal concept of a protein family (e.g. Pfam)

  - **Global statistical** models (traditionally intractable) can be attacked with **novel approximate solutions**

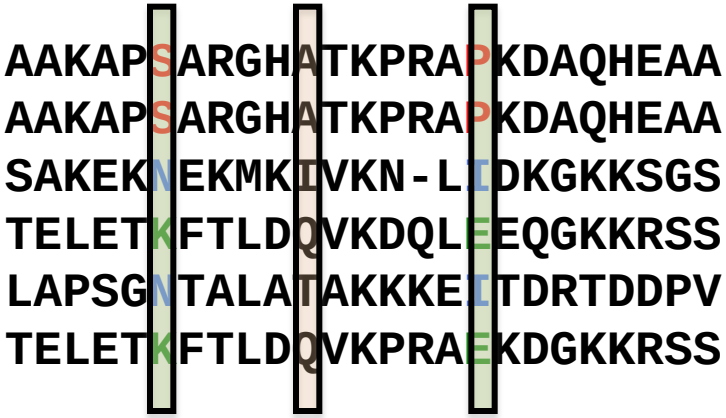
Direct Coupling Analysis predicts 3D contacts in proteins

To disentangle **direct** from **indirect** correlations we developed a statistical inference method called Direct Coupling Analysis (**DCA**).



Available – dca.rice.edu

Sequence probability distribution depends on pairwise and single site parameters



$i = 6$ $j = 17$

$$f_{6,17}(\mathrm{S}_6, \mathrm{P}_{17}) = 2/6$$

Input Data :

$$P_i(A_i) \equiv f_i(A_i)$$

$$P_{ij}(A_i, A_j) \equiv f_{ij}(A_i, A_j)$$

Using maximum entropy principle to model the joint probability distribution

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

$$e_{ij}(A, B) \approx -(C^{-1})_{ij}(A, B)$$

Disentangling direct and indirect correlations

Using maximum entropy principle to model the joint probability distribution

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

An expansion of the energy summations yields

$$e_{ij}(A, B) = -(C^{-1})_{ij}(A, B)$$

This relates pairwise energies and single and pairwise frequency counts

$$C_{ij}(A, B) = f_{ij}(A, B) - f_i(A) f_j(B)$$

Direct Information Metric

Direct Probabilities are defined as:

$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z}\exp\{e_{ij}(A,B) + \hat{h}_i(A) + \hat{h}_i(B)\}$$

The probabilities for residue couplets are ranked using Direct Information

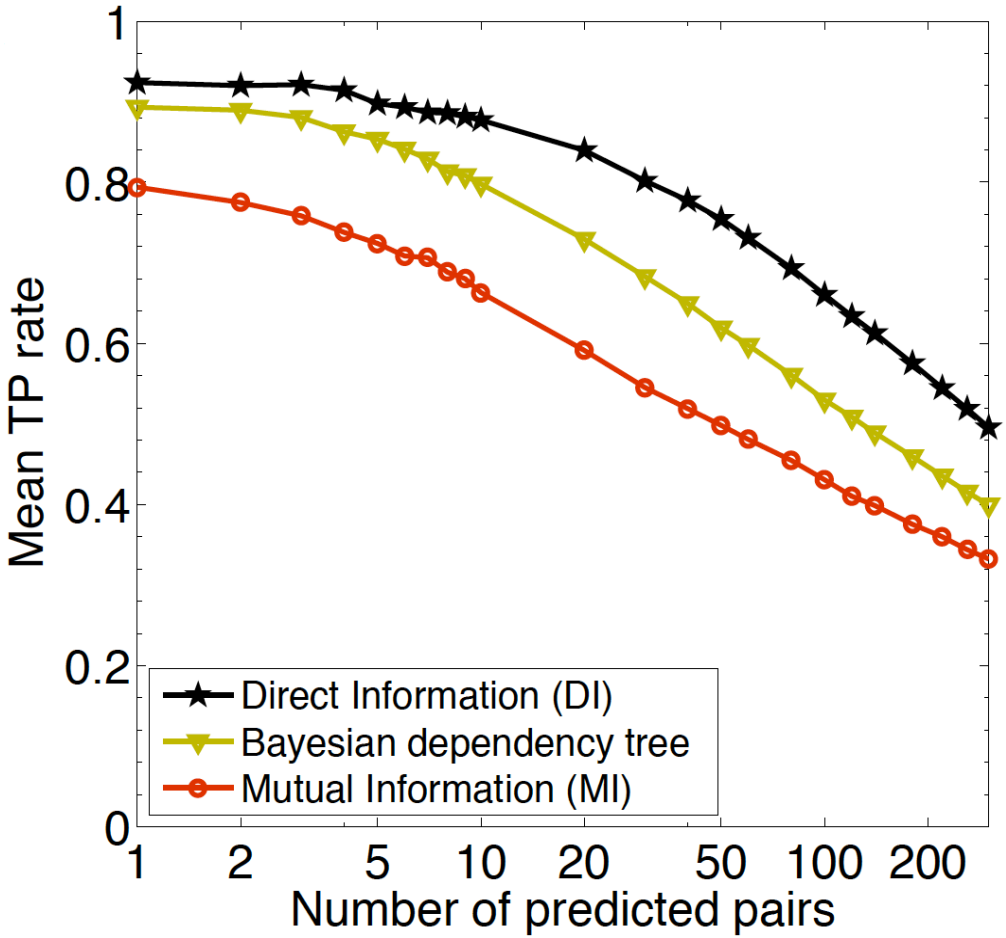
$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B)\ln\frac{P_{ij}^{(dir)}(A,B)}{f_i(A)\,f_j(B)}$$

True positive contacts (<8A) are evaluated from top couplets

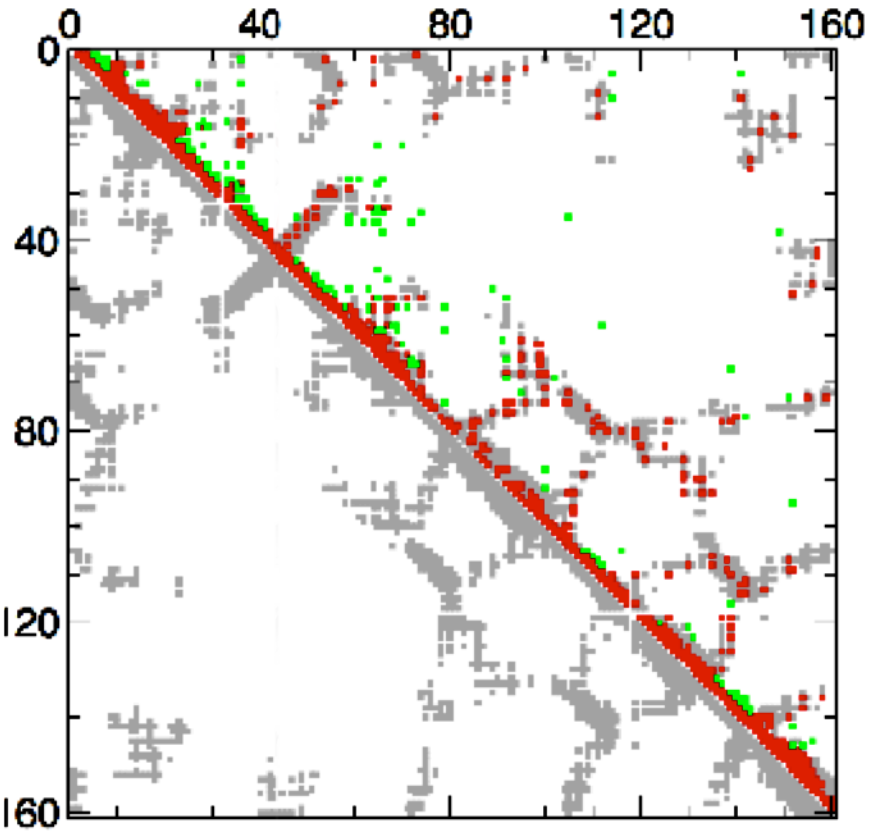
True Positive (TP) rates

DCA accurately infers contacts in protein families

- **DCA** infers high quality contacts for a large number of domain families:



Ras (DI)

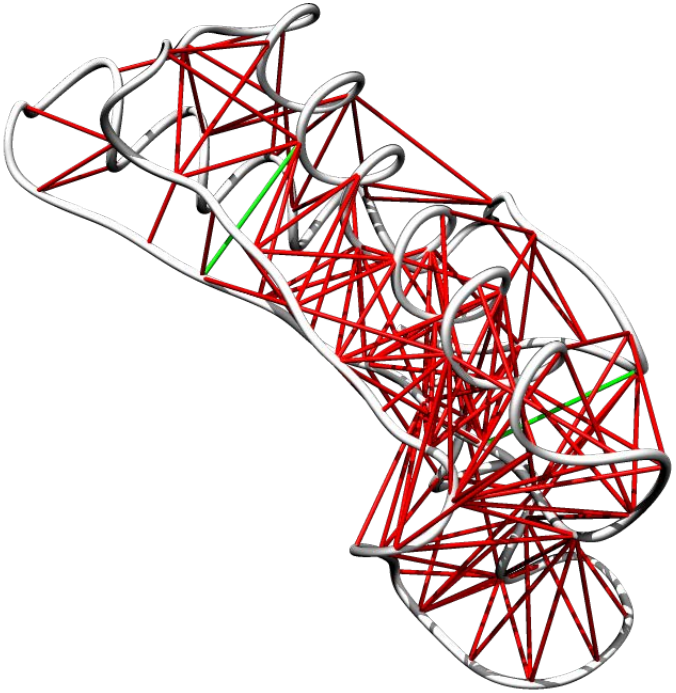


contact < 8Å

[Morcos et al. PNAS. 2011],
see also review by De Juan & Valencia. *Nature Rev. Gen.* 2013

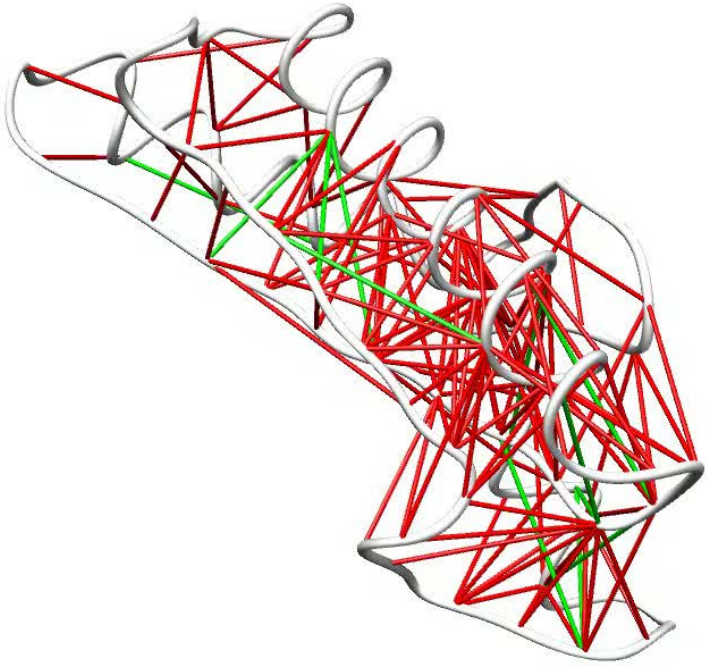
DCAfold predicts Peptidoglycan-Associated lipoprotein

RMSD: **1.5 Å**



PAL Native structure

PDB:1oap



Non-local information: DCA

Local information: Known

Red link (DCA True Positive < 8Å)

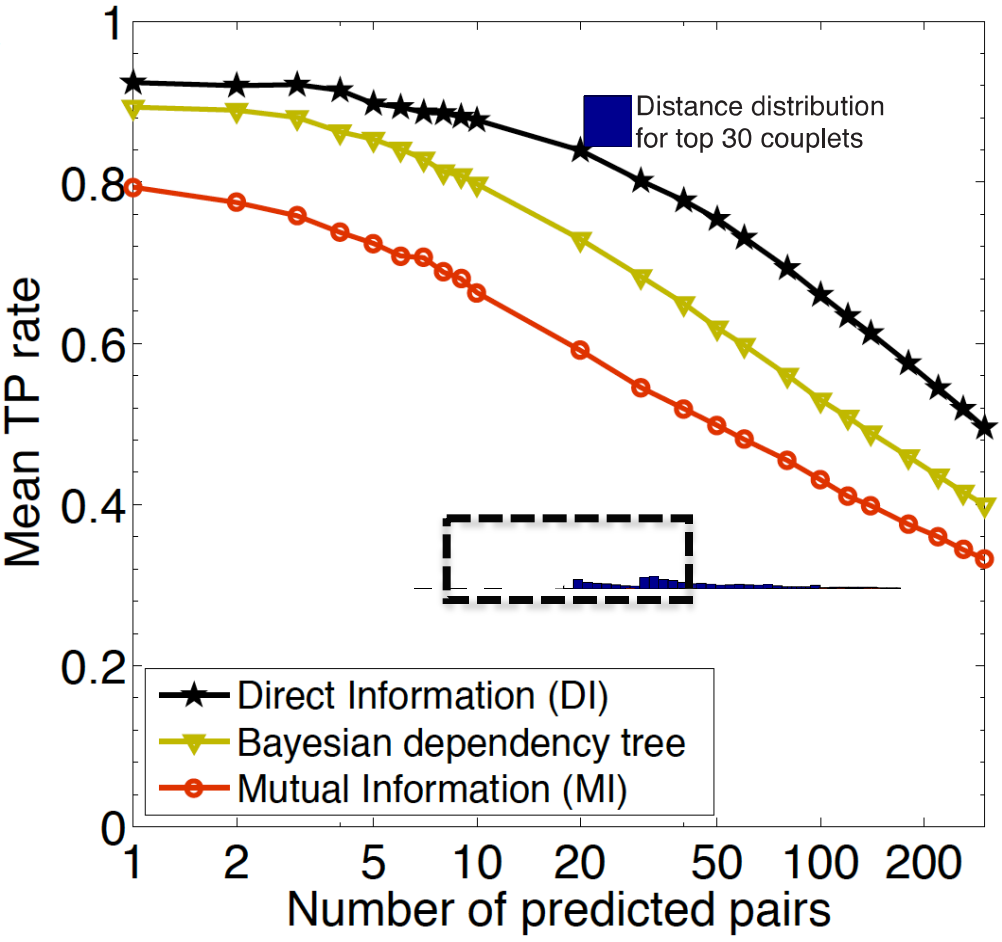
Green link (DCA False Positive > 8Å)

[Sulkowska, Morcos et al. PNAS 2012]

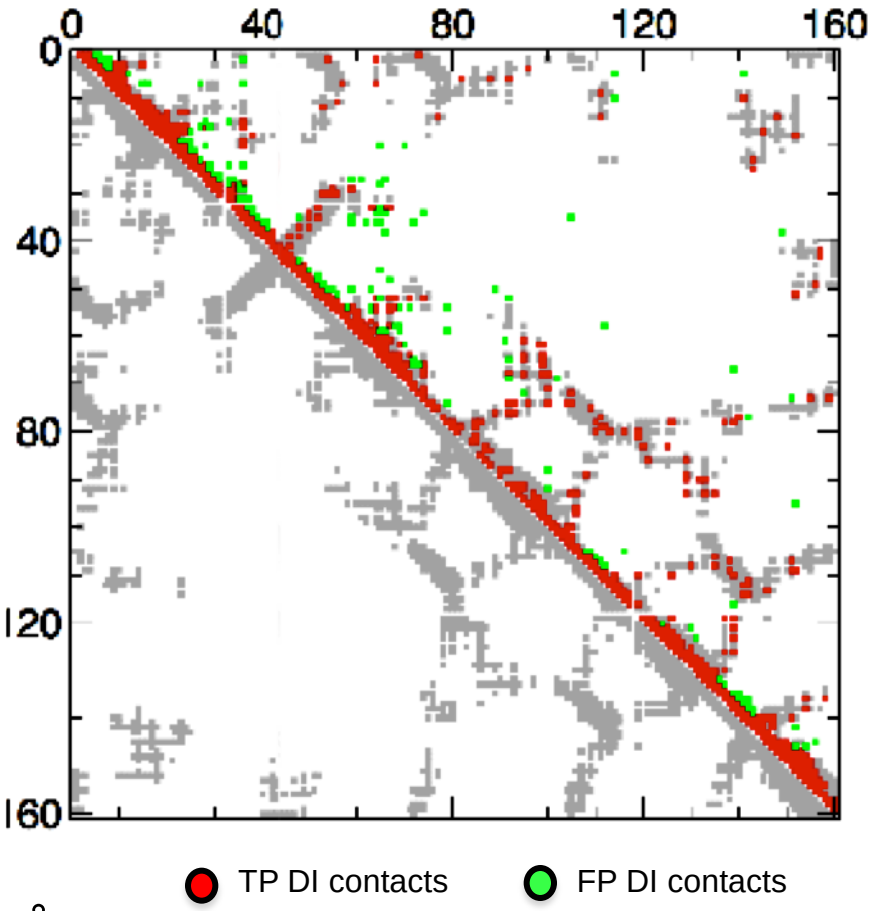
Review: Marks et al. *Nature Biotechnology* 2012

DCA accurately infers contacts in protein families

- **DCA** infers high quality contacts for a large number of domain families:



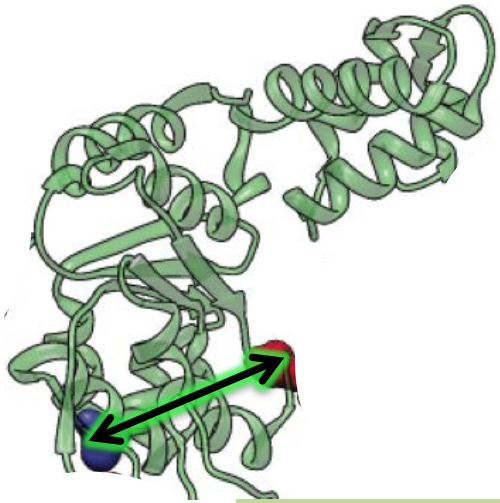
Ras (DI)



contact < 8Å

[Morcos et al. PNAS. 2011],
see also review by De Juan & Valencia. *Nature Rev. Gen.* 2013

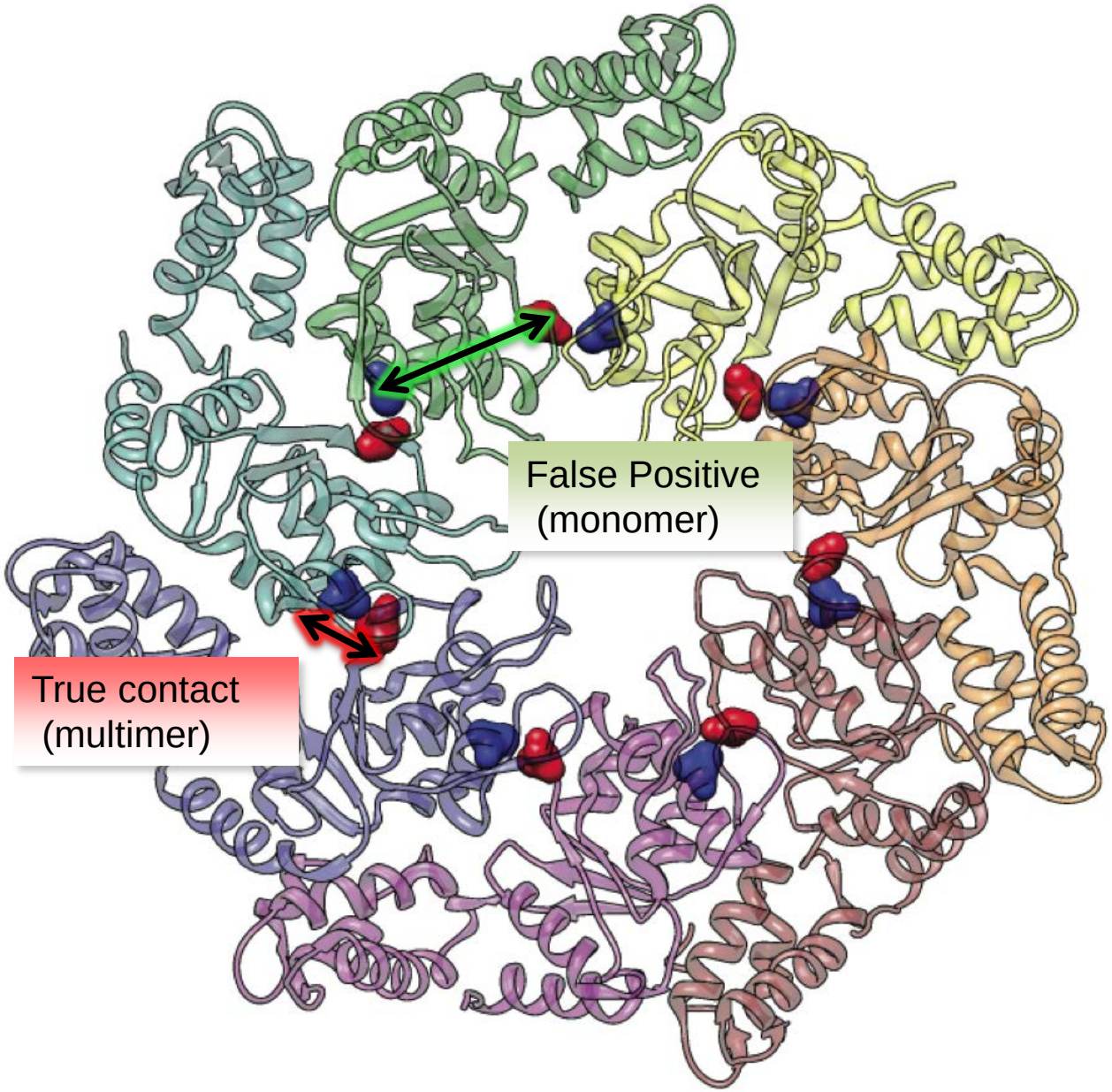
Multimerization contacs have high DI rank



False Positive
(monomer)

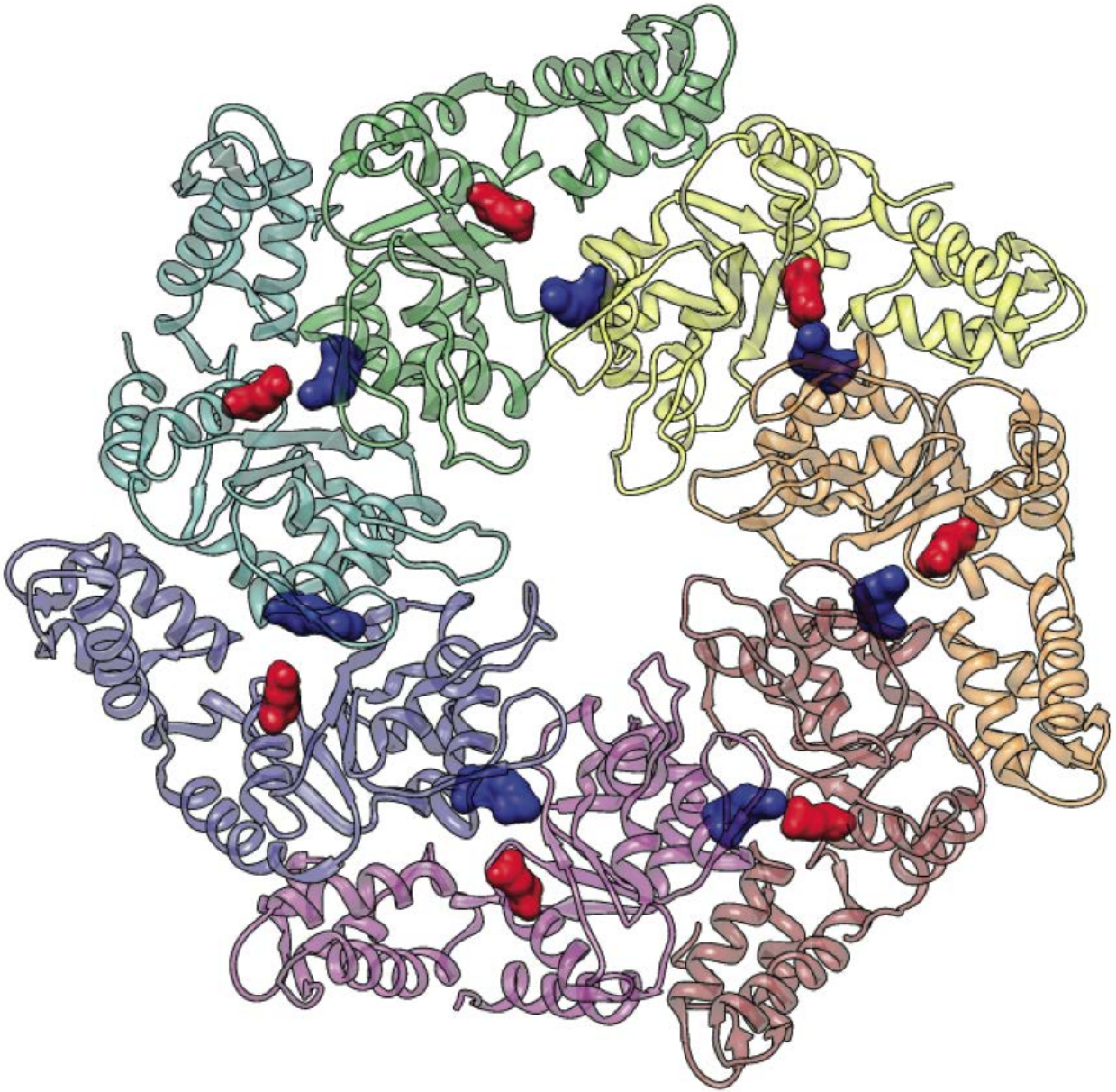
Multimerization contacs have high DI rank

- A monomeric False positive is in fact a multimerization contact



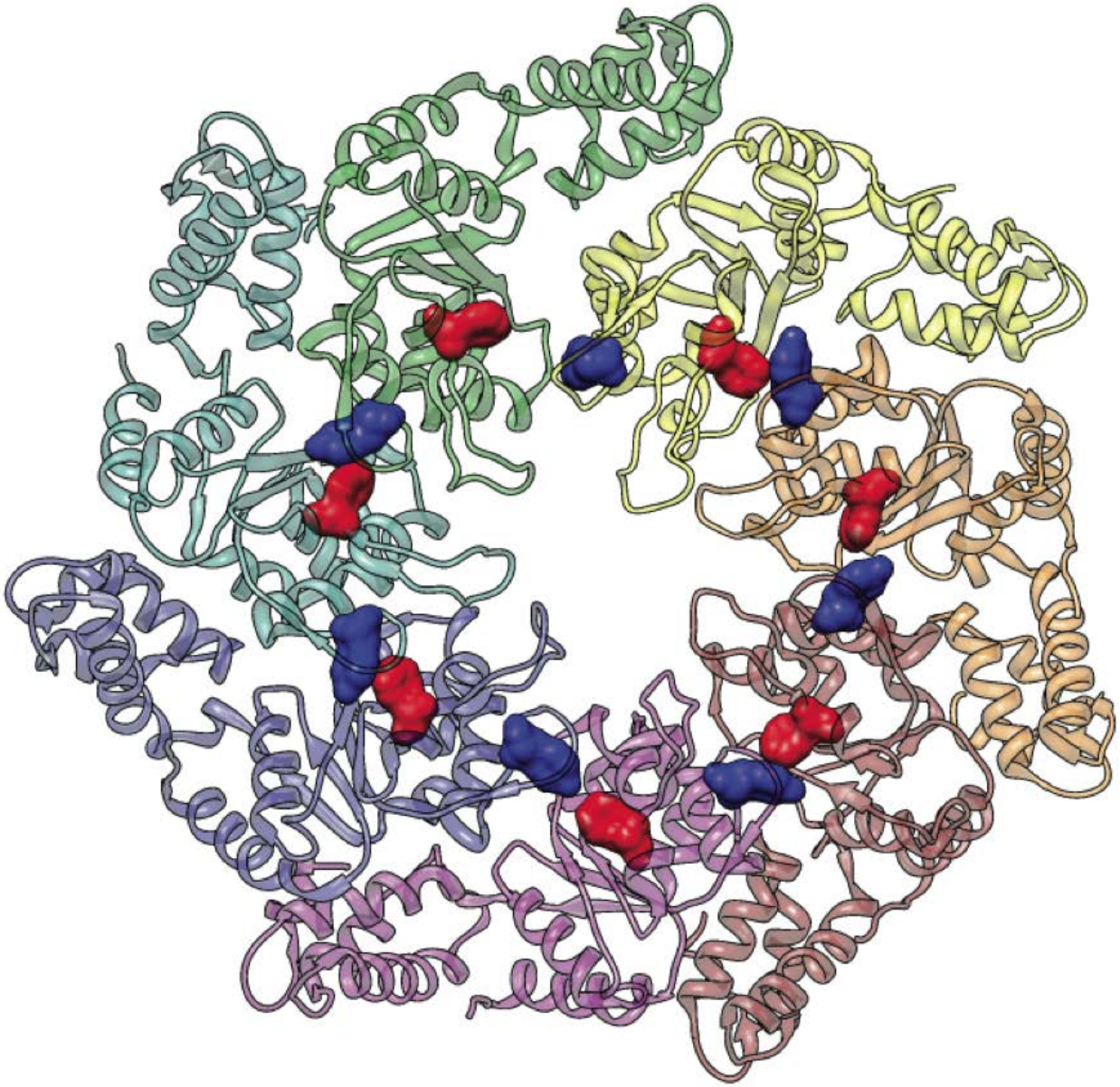
Multimerization contacs have high DI rank

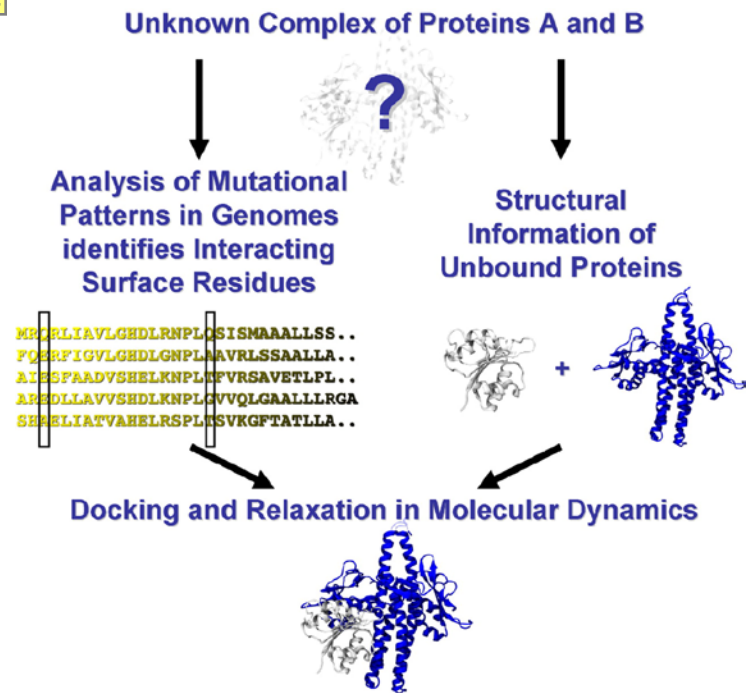
- A monomeric False positive is in fact a multimerization contact



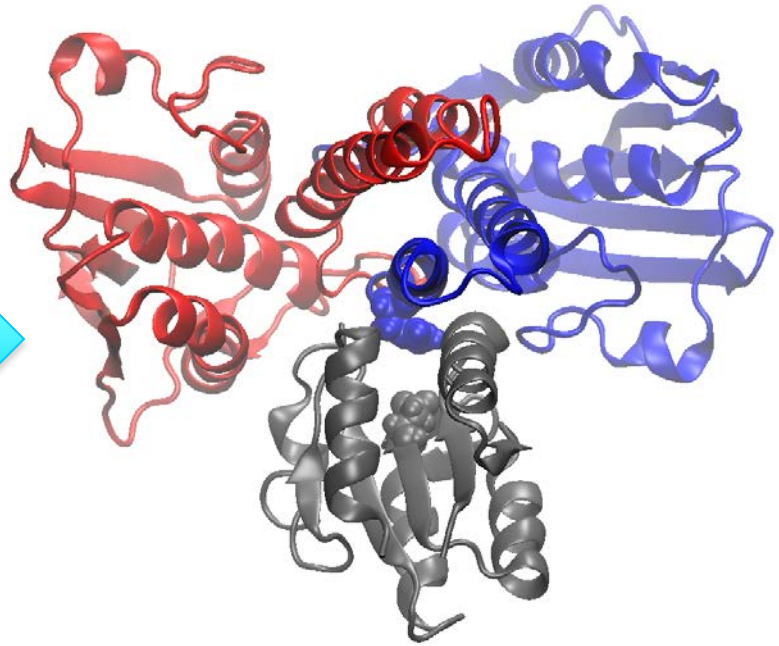
Multimerization Contacs have high DI rank

- All 3 monomeric False positives are in fact multimerization contacts



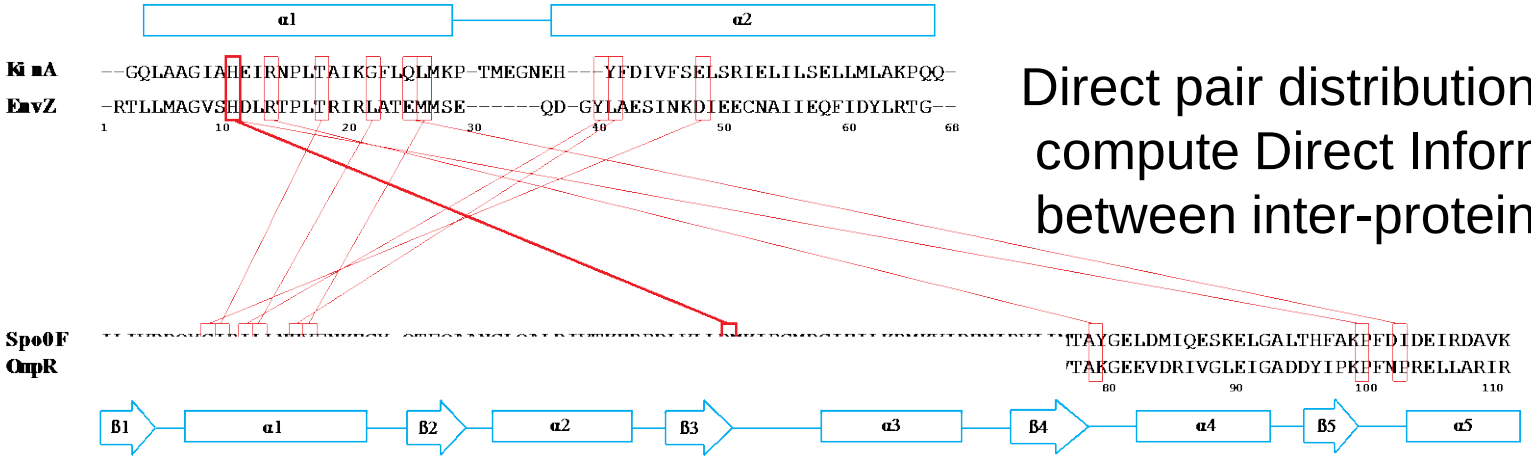


Predicted KinA-Spo0F complex



$$DI_{ij} = \sum_{A,B} P_{ij}^{(dir)}(A,B) \ln\left(\frac{P_{ij}^{(dir)}(A,B)}{f_i(A) f_j(B)}\right)$$

Schug, Weigt, Szurmant et al



Direct pair distribution used to compute Direct Information (DI) between inter-protein pairs *i* and *j*

Highest DI values reflect interprotein contacts between HK and RR

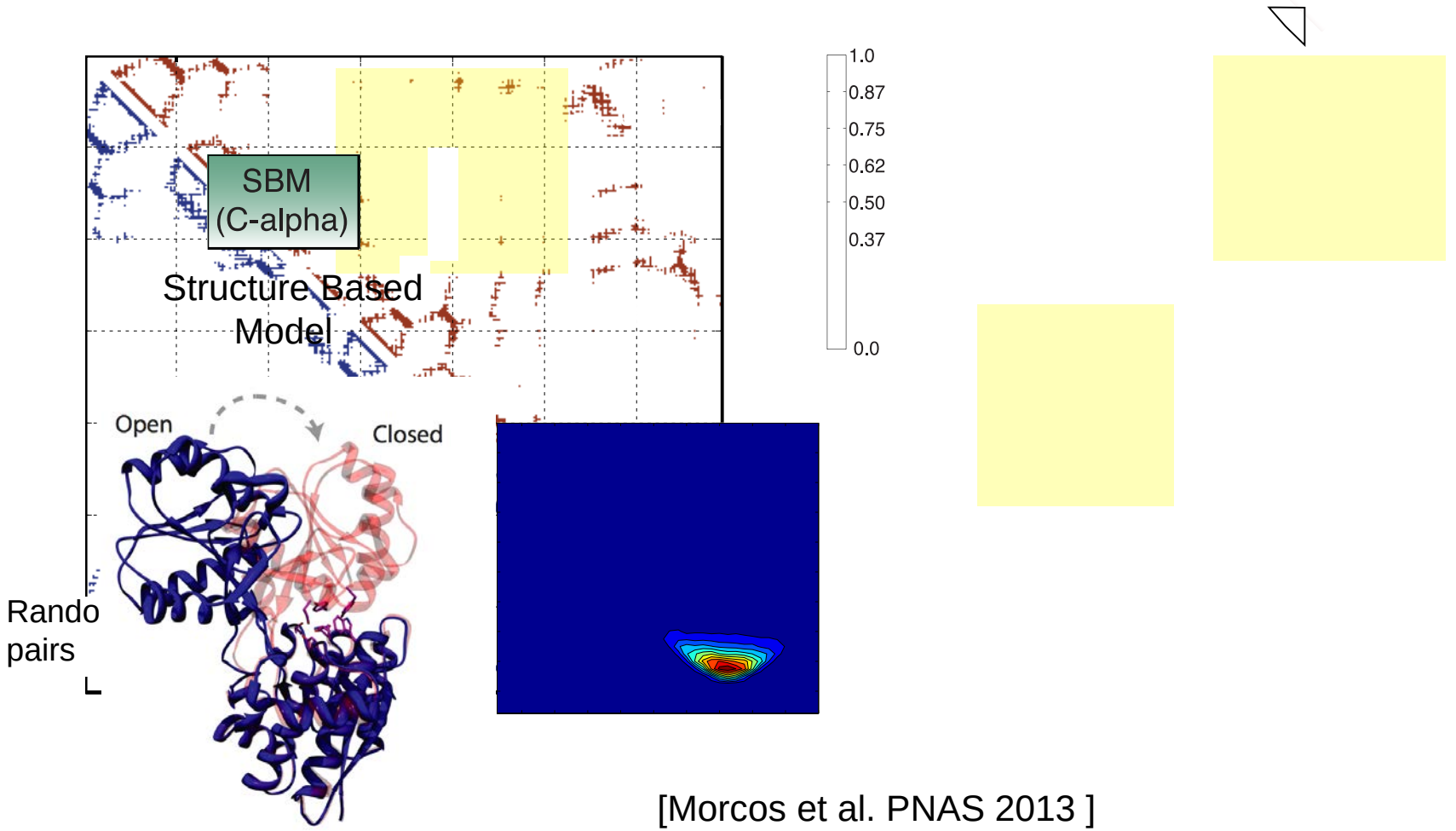
Conformational plasticity in ligand bound proteins

Name	Family	Sequences M	Effective Seq. Meff	PDB open/closed	Protein Length
L-leucine binding protein	Peripla_BP_4 Solute Binding Periplasmic	7K	3.3K	1usg/1usi	345

Ligand:
L-Leucine

Conformational plasticity in ligand bound proteins

Name	Family	Sequences M	Effective Seq. Meff	PDB open/closed	Protein Length
L-leucine binding protein	Peripla_BP_4 Solute Binding Periplasmic	7K	3.3K	1usg/1usi	345



[Morcos et al. PNAS 2013]

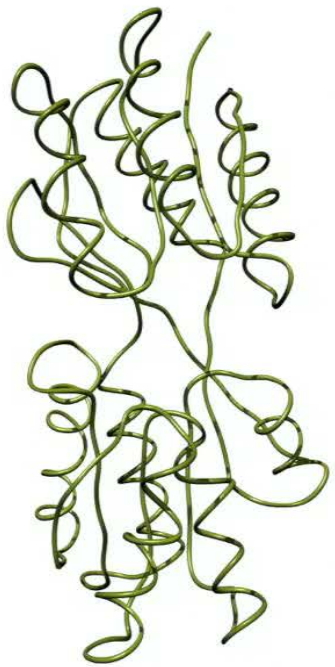
SBM+DCA finds hidden conformations

D-Ribose binding protein

RMSD open (Å)

[Morcos et al. PNAS. 2013]

Evidence of hidden conformations in sugar binding proteins



- D-Ribose intermediate state could facilitate ribose transfer in the permease complex.
  Ravindranathan et al. J. Mol. Biol. (2005)

- Twisted state suggested for D-Glucose binding protein, based on disulfite-trapping and fluorescence spectroscopy.
  C. L. Careaga et al. Biochemistry. (1995)

- Accelerated MD suggested a semi-closed state for Maltose binding protein
  D. Bucher et al. PLoS Comput Biol. (2011)

Coevolutionary Information in Protein-Protein Interactions

Ricardo Nascimento dos Santos and Faruck Morcos

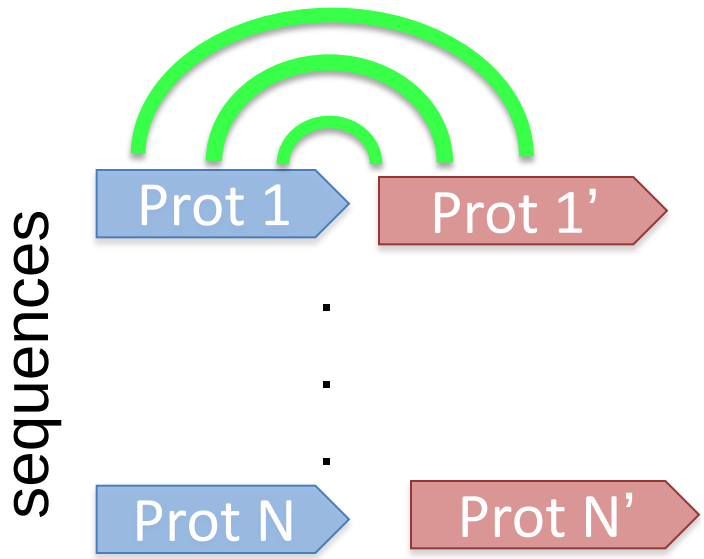
Scientific Reports, in press

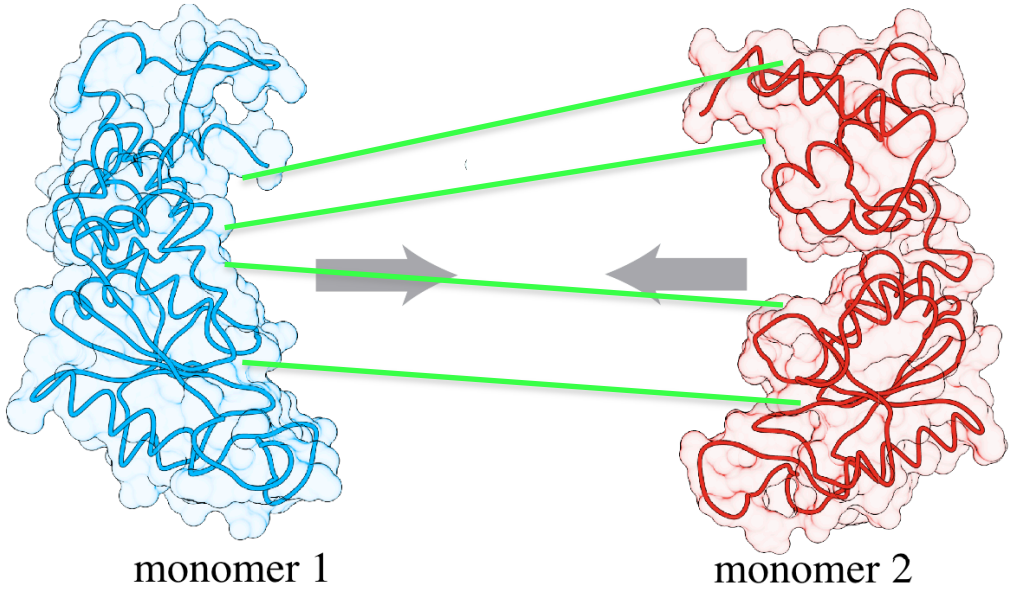
DCA of paired sequences provide interprotein constraints

- Direct Coupling Analysis can be easily extrapolated to protein pairs

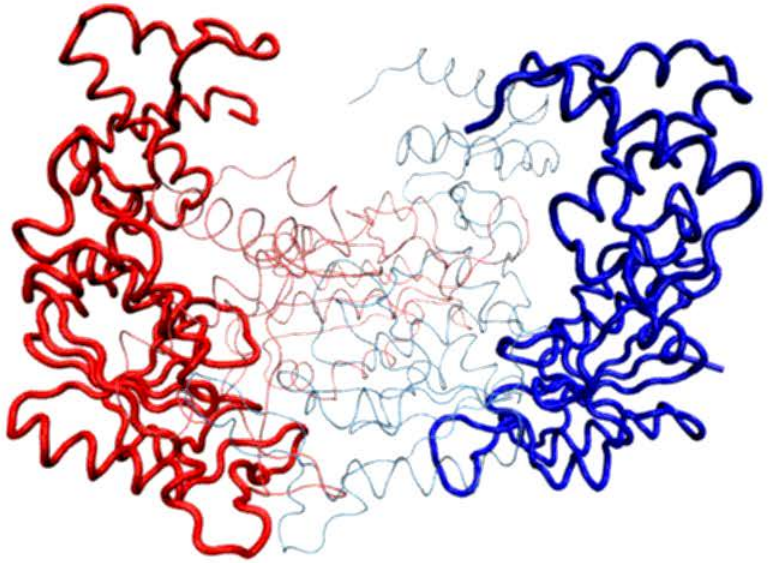
- Protein interfaces are preserved by coevolving residues

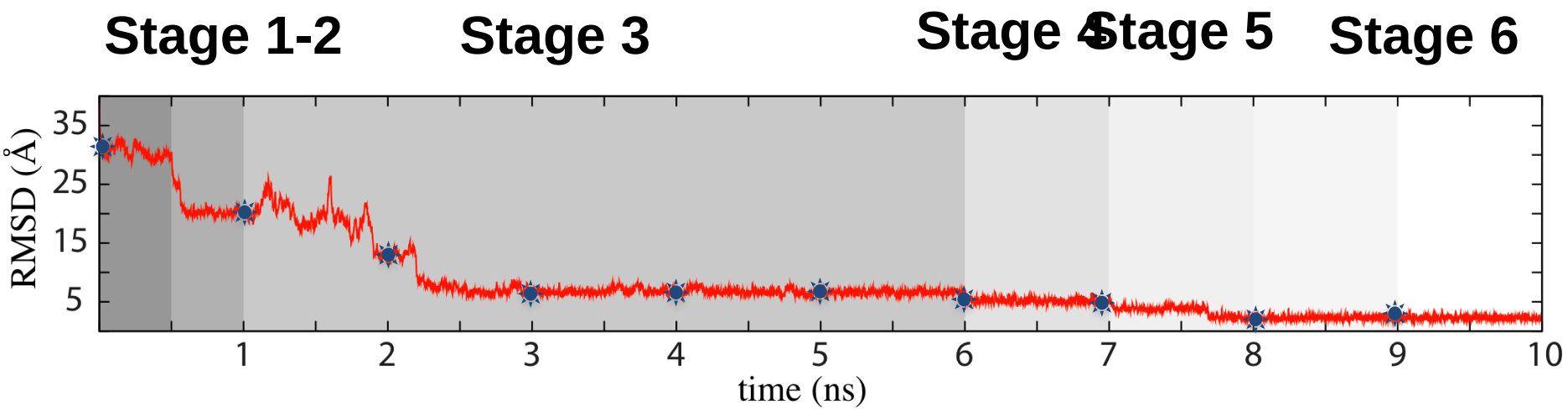
- Sequence pairing can be done by genomic adjacency, annotation or single copies per organism





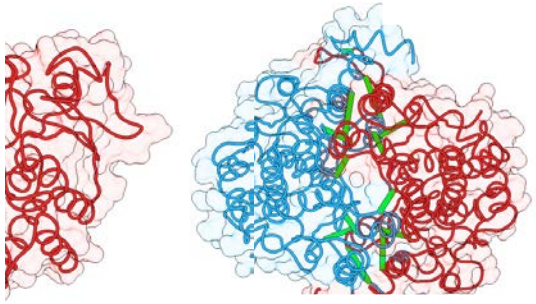
Methyltransferase PDB (1UAL)





Complex formation using coevolving residues

[Dos Santos, Morcos et al. *2015*, *in press*]



Protein association predicted in cancer proteins

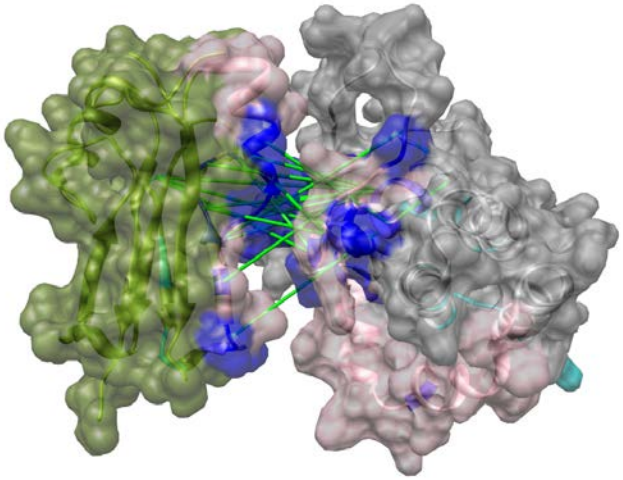
PNAS

Integrated strategy reveals the protein interface between cancer targets Bcl-2 and NAF-1

Sagi Tamir[a], Shahar Rotem-Bamberger[b], Chen Katz[b], Faruck Morcos[c], Kendra L. Hailey[d], John A. Zuris[d], Charles Wang[d], Andrea R. Conlan[d], Colin H. Lipper[d], Mark L. Paddock[d], Ron Mittler[e], José N. Onuchic[c,f,g,h,1], Patricia A. Jennings[d,1], Assaf Friedler[b,1], and Rachel Nechushtai[a,1]

[a]The Alexander Silberman Institute of Life Science and [b]Institute of Chemistry, Hebrew University of Jerusalem, Edmond J. Safra Campus at Givat Ram, Jerusalem 91904, Israel; [c]Center for Theoretical Biological Physics and Departments of [f]Physics and Astronomy, [g]Chemistry, and [h]Biochemistry and Cell Biology, Rice University, Houston, TX 77050; [d]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093; and [e]Department of Biological Sciences, University of North Texas, Denton, TX 76203

A

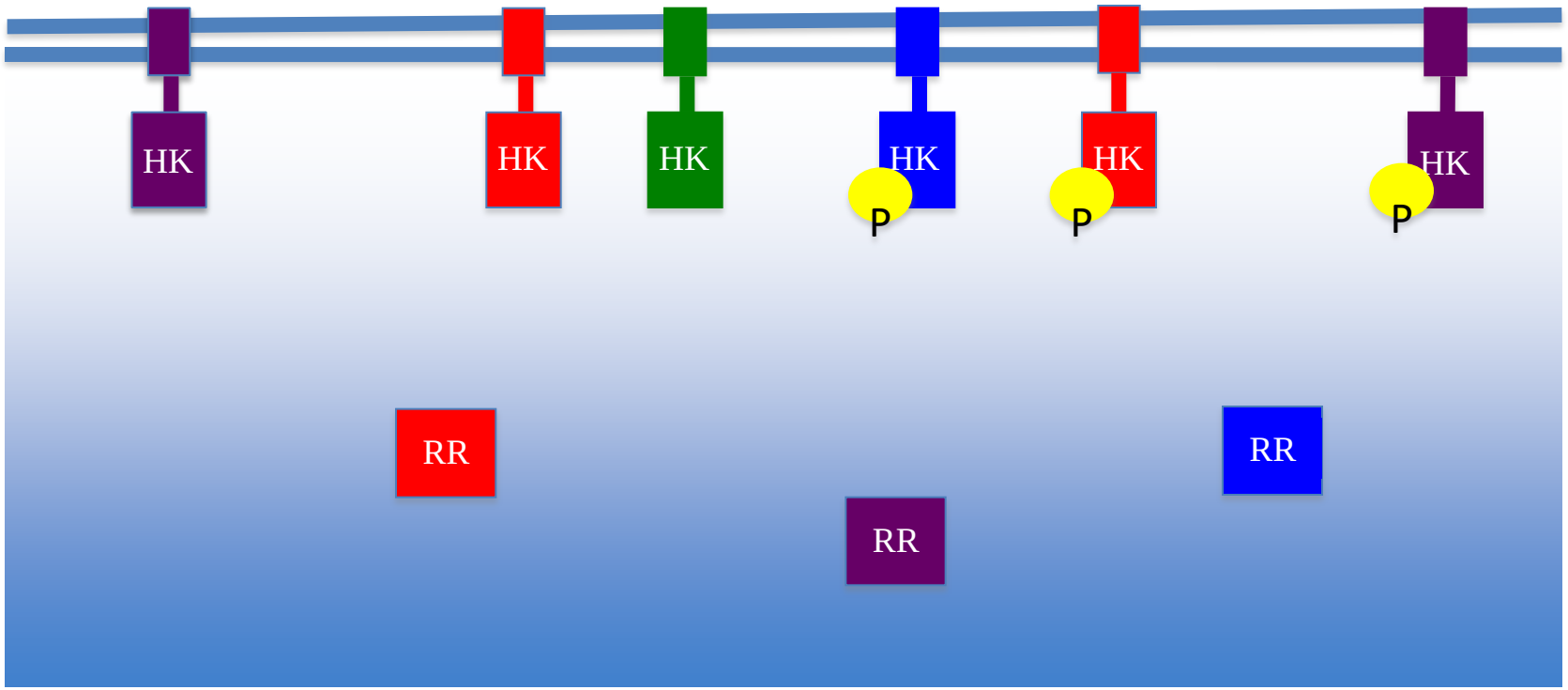


B

BCL-2
NAF-1
DXMS
DCA
Overlap DCA/ Experiment

[Tamir et al. PNAS 2014]

Background on Two-component signaling



10^2-10^3 TCS partners in bacteria

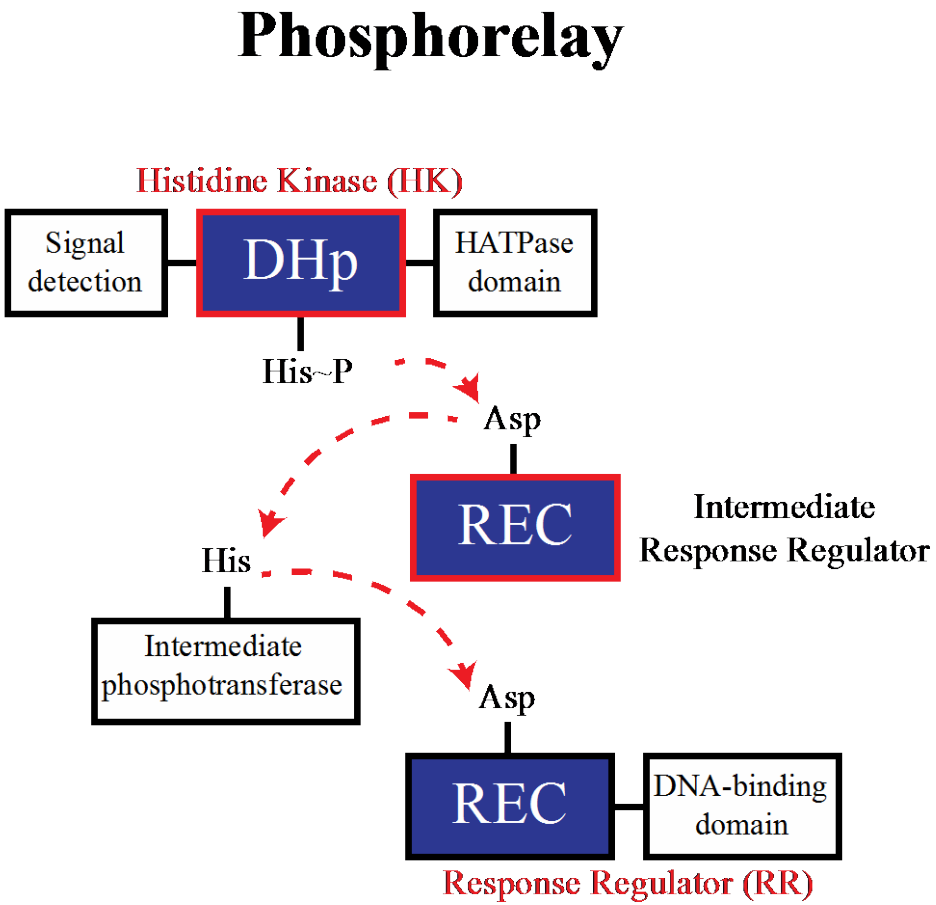
How does a TCS protein coevolved to stay faithful to its signaling partner?

Can we identify the molecular determinants of *interaction recognition* from abundant sequence data and limited structural data?

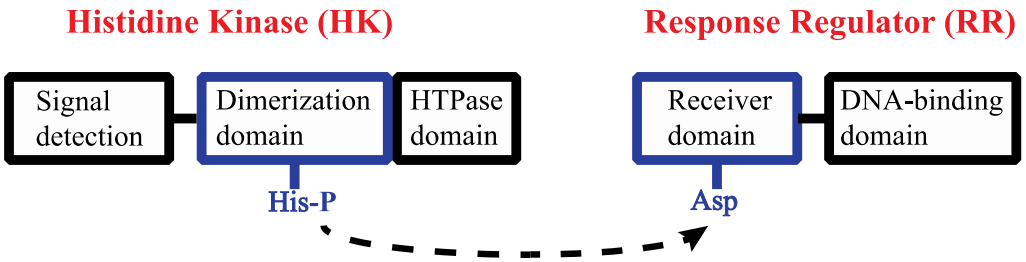
[Laub, Goulian. Annu Rev Genet. 2007]

Mutagenesis of a response regulator

Alanine-scanning mutagenesis of Spo0F (Tzeng and Hoch, JMB 1997)



DCA-based recognition metric between interacting protein partners



Related to Direct Information (DI):

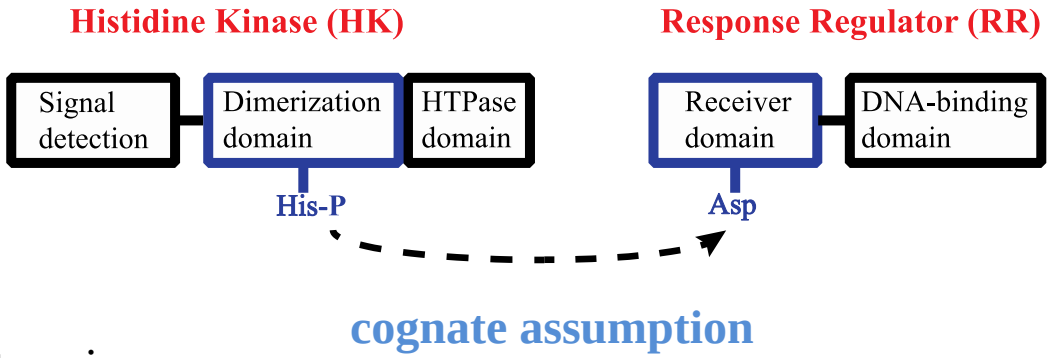
$$DIScore = \sum_{i \in HK, j \in RR} P_{ij}^{(dir)}(S_i, R_j) \ln\left(\frac{P_{ij}^{(dir)}(S_i, R_j)}{f_i(S_i) f_j(R_j)}\right)$$

Testing predictive power in capturing interaction preference between different sporulation proteins of the HK and RR families

Sporulation kinase	Sporulation response regulator Spo0F	Spo0A
KinA	**5.91**	5.31
KinB	**5.44**	5.06
KinC	**5.54**	5.05
KinD	**5.91**	5.38
KinE	**5.44**	4.96

Higher values appears to reflect preference of sporulation kinase to Spo0F but meaning of magnitude of this metric or relative differences between sporulation kinases is still being understood

Protein Recognition in TCS can be characterized with DCA



1) **Construct database:** multiple sequence alignments of known interacting partners, i.e., cognate pairs (30,623 sequences)

Key assumption: HK and RR that are adjacent on operon are cognate pairs

2) **DCA:** computation of direct couplings between interprotein residues

Using our **metric** to **infer mutational** effects on the functional **interaction** (i.e., phosphotransfer) between TCS proteins

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

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

[**Early work**: Li et al. PNAS 2003, White et al. Methods Enzymology 2007, Skerker et al. Cell 2008]

Specificity score requires substraction of generic features

No cognate assumption (i.e., scramble)

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

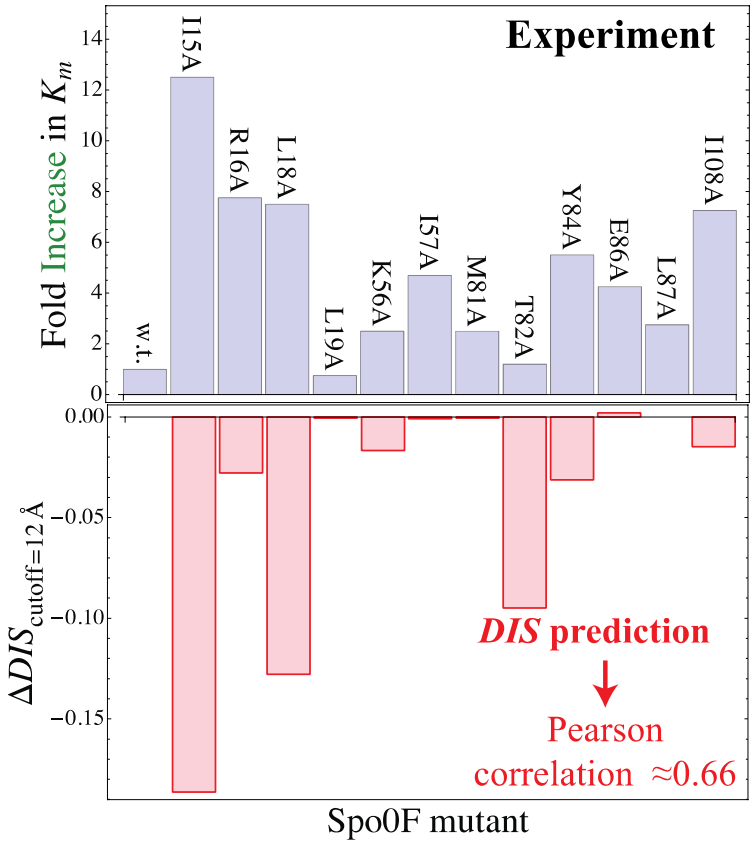
Cognate assumption

Determinants of interaction specificity amongst cognate pairs

Contains generic, conserved features of HK/RR pairs

Specificity score requires substraction of generic features

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



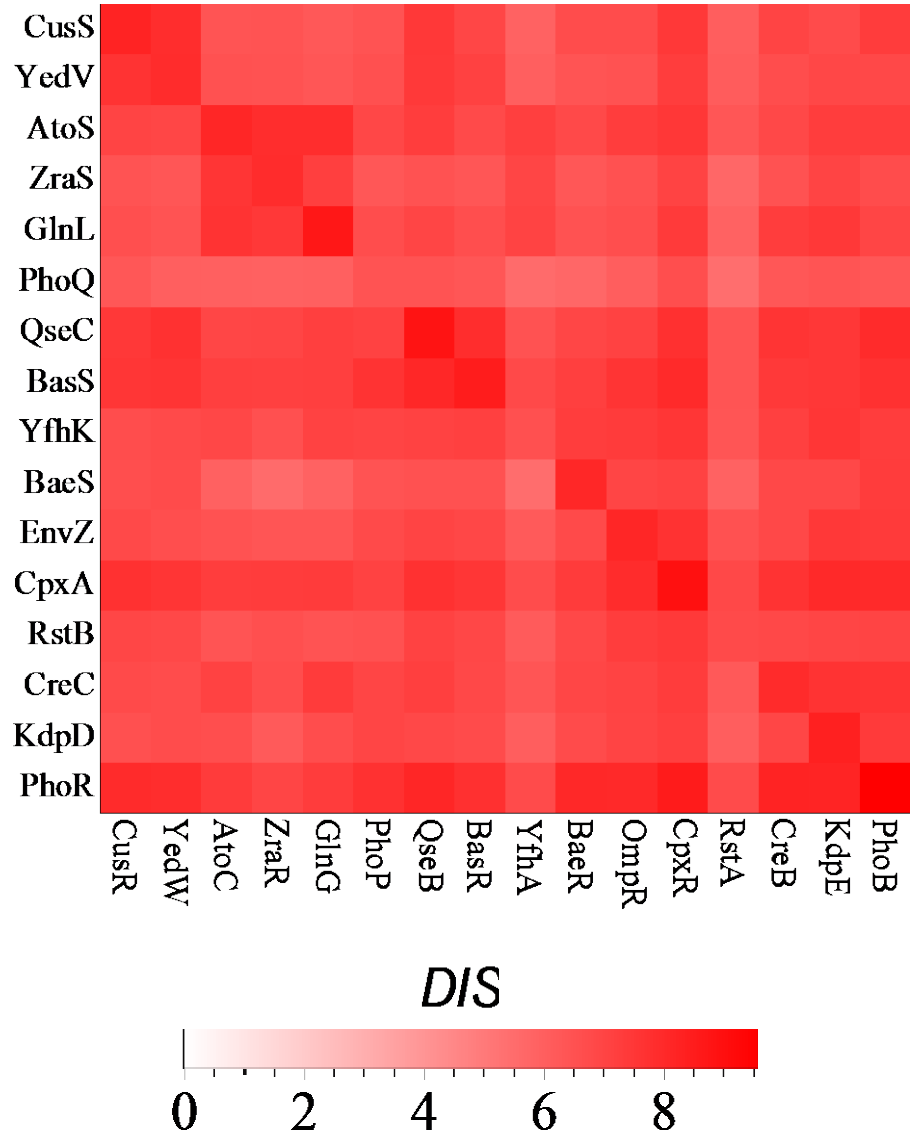
ation ≈0.81

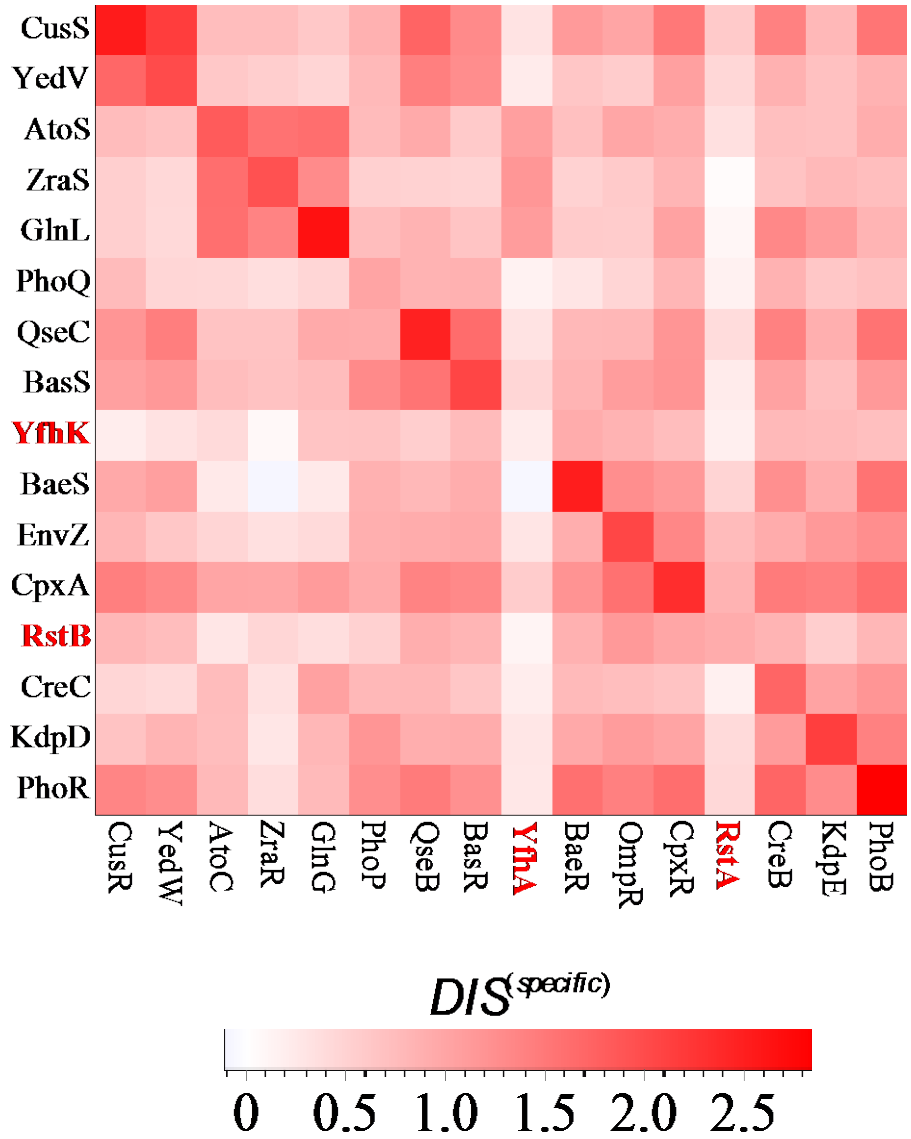
$$\frac{d[P]}{dt} \approx \frac{V_{max}[S]}{K_m}$$

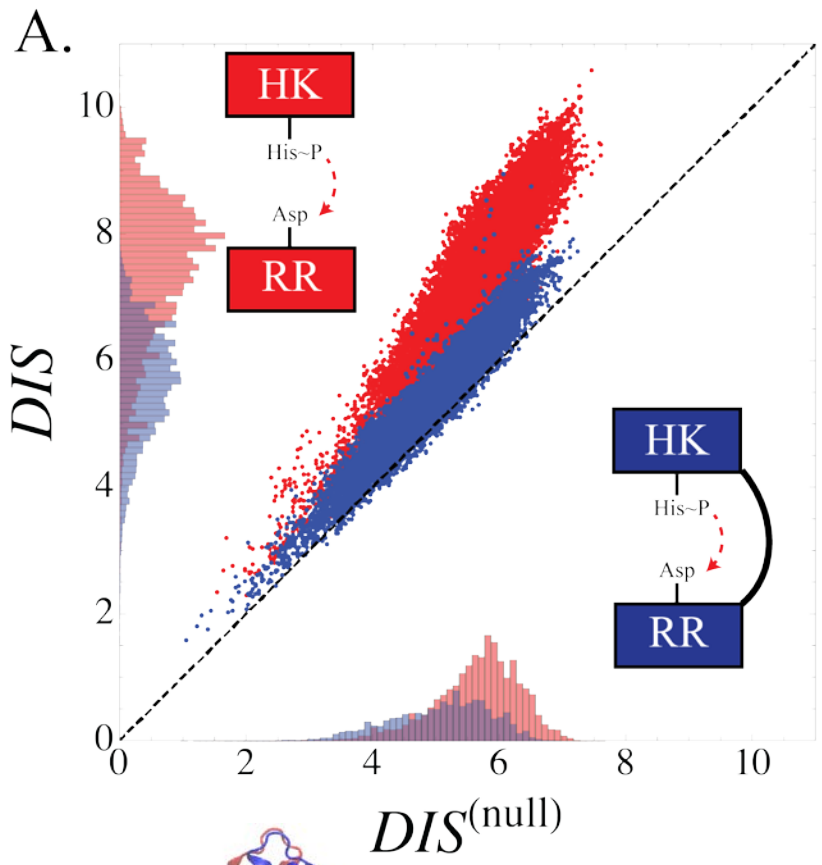
Response Regulator REC domain

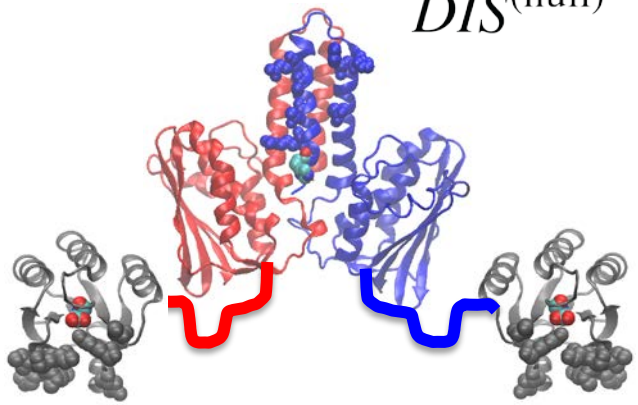
Alanine-scanning mutagenesis of Spo0F (Tzeng and Hoch, JMB 1997)

Prediction of cognate pairs in *E. Coli*

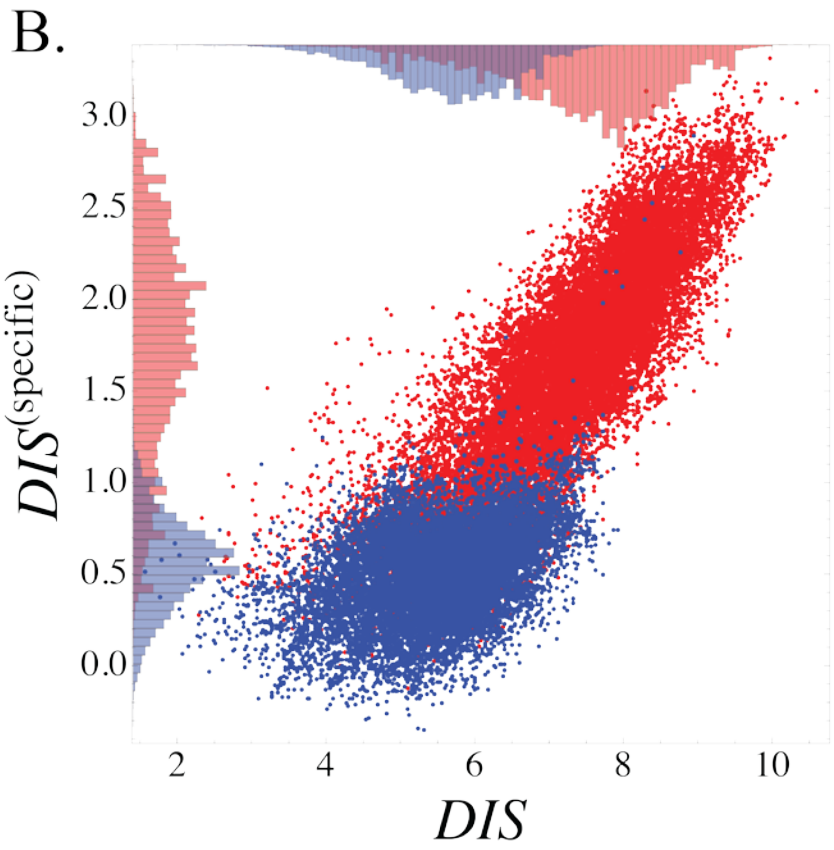








Cartoon depiction of a hybrid TCS protein



Hybrid TCS proteins (~17,000) do not need to have a highly co-evolved recognition interface since tethering greatly increases their rate of encounter.

(Consistent with Townsend *et al*, PNAS 2013)

P.F. Collaborators

Postdocs:

Faruck Morcos, Heiko Lammert, Ellinor Haglund, Jeff Noel, Ryan Cheng, Mingyang Lu, Fang Bai, Michele Di Pierro

Alex Schug - Karlsruher Institut für Technologie (KIT)

Biman Jana – Indian Assoc. Cultivation of Science - India

Joanna Sulkowska – University of Poland

Koby Levy - Weizmann Institute

Changbong Hyeon - KIAS, Korea

Cecilia Clementi – Rice University

Joan-Emma Shea – UCSB

Adelia Aquino – Texas Tech

Steve Plotkin – UBC, Canada

Osamu Miyashita – RIKEN, Japan

John Finke - University of Washington, Tacoma

Antitsa Stoycheva - Gilead Sciences, Inc.

Nick Socci – Sloan Katering

Yoko Suzuki - Meisei University - Tokyo

Jorge Chahine - Unesp, Brazil

Chigusa Kobayashi - RIKEN, Japan

Eric Nelson – UTSWMC, Texas

Jian Liu - NIH

Shachi Gosavi - NCBS – India

Daniel Schultz – Harvard

Leandro Oliveira - CTBE, Brazil

Marcio de Mello Cardoso – UFRJ, Brasil

Students

Ricardo dos Santos, Ryan Hayes, Li Sun, Nathan Eddy, Xingcheng Lin, Mohit Raghunathan, Bin Huang, Kareen Mehrabiani, Marcelo Boareto

Paul Whitford – Northeastern University

Hugh Nymeyer –Florida State University

Marcos Betancourt - IUPUI

Vitor Leite - Unesp, Brazil

Leslie Chavez - LANL

Margaret Cheung - University of Houston

Sichun Yang – Case Western University

Peter Leopold – BioAnalyte Inc.

Collaborators

Peter Wolynes, Herbie Levine and Eshel Ben-Jacob

Jianpeng Ma – Baylor College of Medicine

Angel Garcia - RPI

Pat Jennings and Terry Hwa– UCSD

Martin Weigt – Univ. Paris, France

Zan Luthey-Schulten – UIUC

Karissa Sanbonmatsu - LANL

Hualiang Jiang – SIMM Chins

Charles Brooks – U. Michigan

Stefan Klumpp – MPIKG, Germany

Shoji Takada - Kobe, Japan

Ashok Deniz – TSRI

Mikael Oliverberg – Stockholm

Antonio Francisco Pereira de Araujo – Brasilia

Yoshitaka Tanimura - Kyoto, Japan

Yuko Okamoto – Nagoya University, Japan

Ulrich Hansmann – University of Arkansas

Rachel Nechushtai – Hebrew University

Ron Mittler – Univ. North Texas

National Science Foundation
WHERE DISCOVERIES BEGIN

CPRIT
CANCER PREVENTION & RESEARCH INSTITUTE OF TEXAS

THE Welch FOUNDATION