Proteins: sequences and physics

Anne-Florence Bitbol

LABORATOIRE Jean Perrin

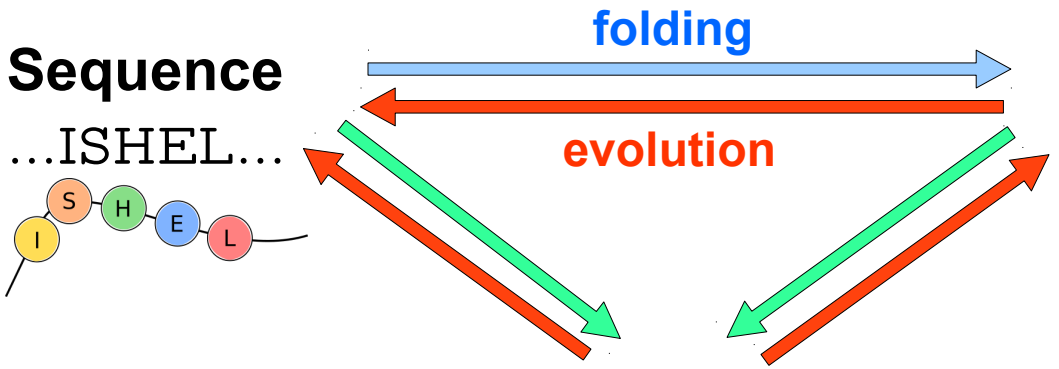
cnrs

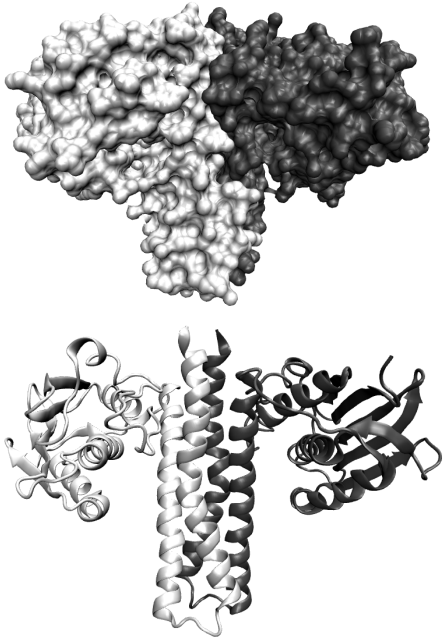
SORBONNE UNIVERSITÉ

39èmes Journées de Physique Statistique
January 31, 2019

Introduction

- **Understanding proteins**

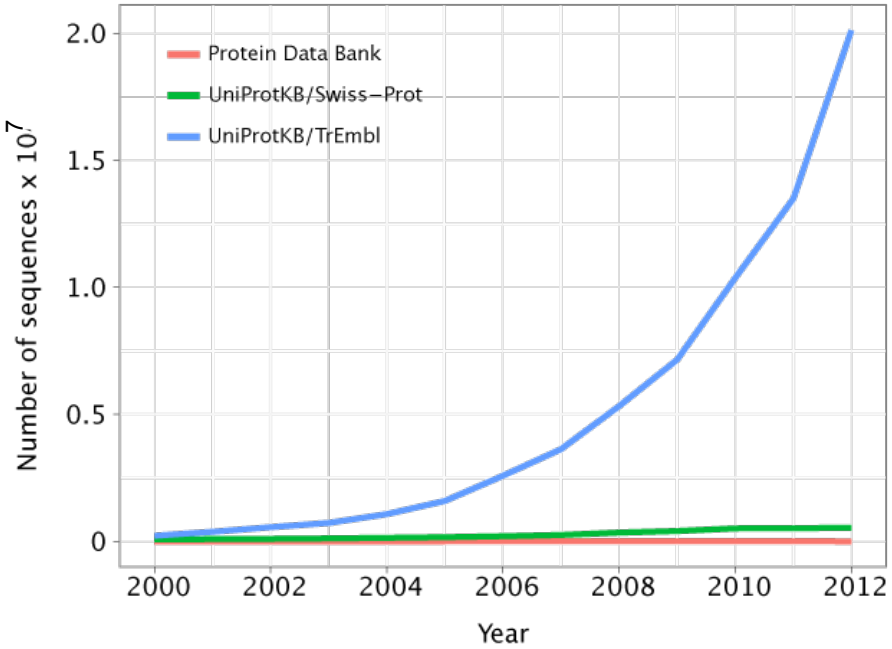




Function (including interactions)

- Enzyme
- Molecular motor
- Receptor...

- **A growing amount of data; mostly unannotated sequences**



Currently: more than 100 million sequences in Uniprot

I. Inferring interaction partners from protein sequences

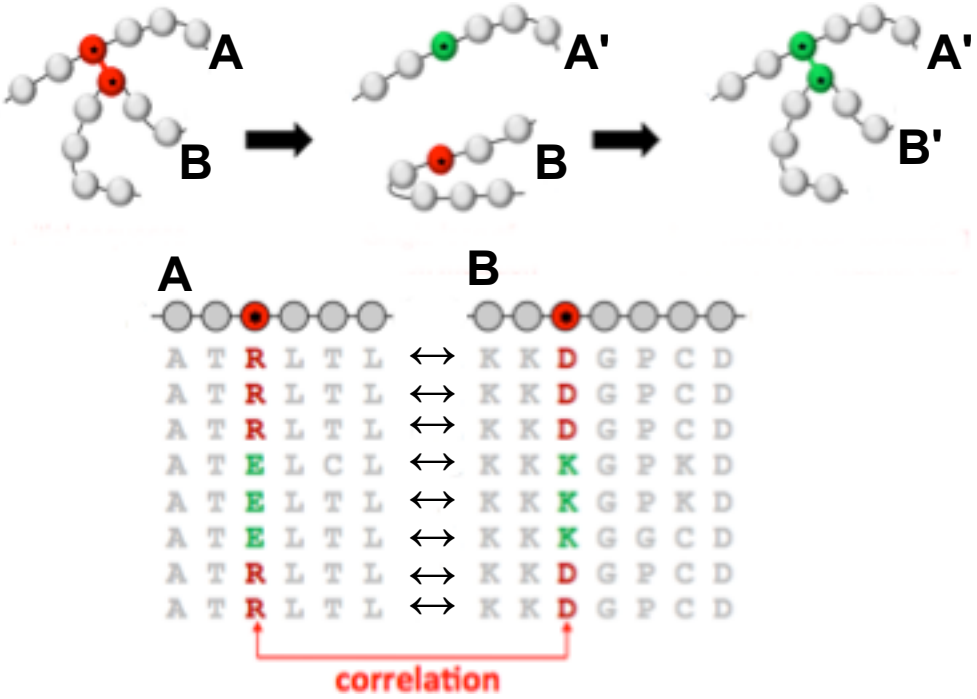
with Ned S. Wingreen, Lucy J. Colwell, Rob S. Dwyer

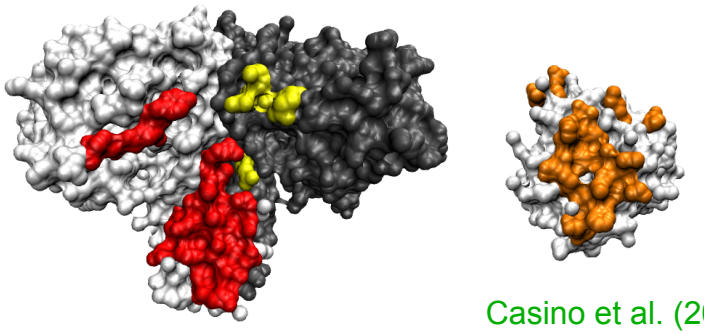
II. A physical interpretation of sectors of collectively correlated amino acids

with Ned S. Wingreen & Shou-Wen Wang

Introduction

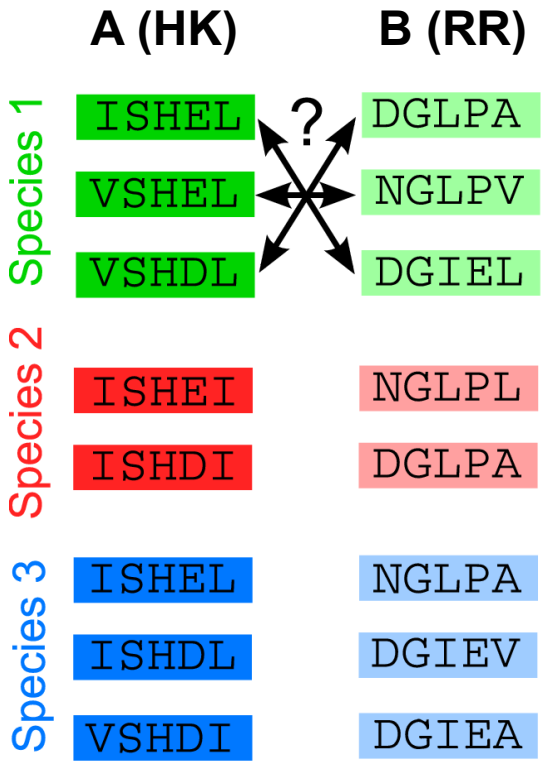
- **Co-evolution and correlations between interacting partners**





Casino et al. (2009)

Often, several paralogs in each species

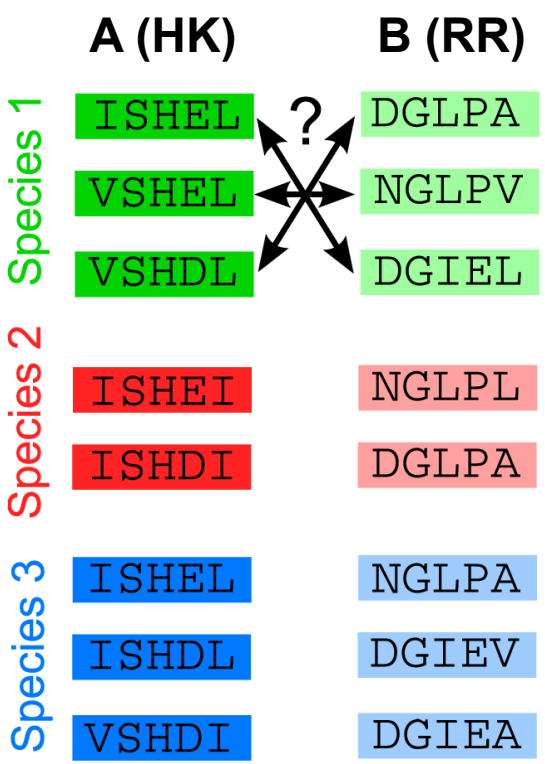


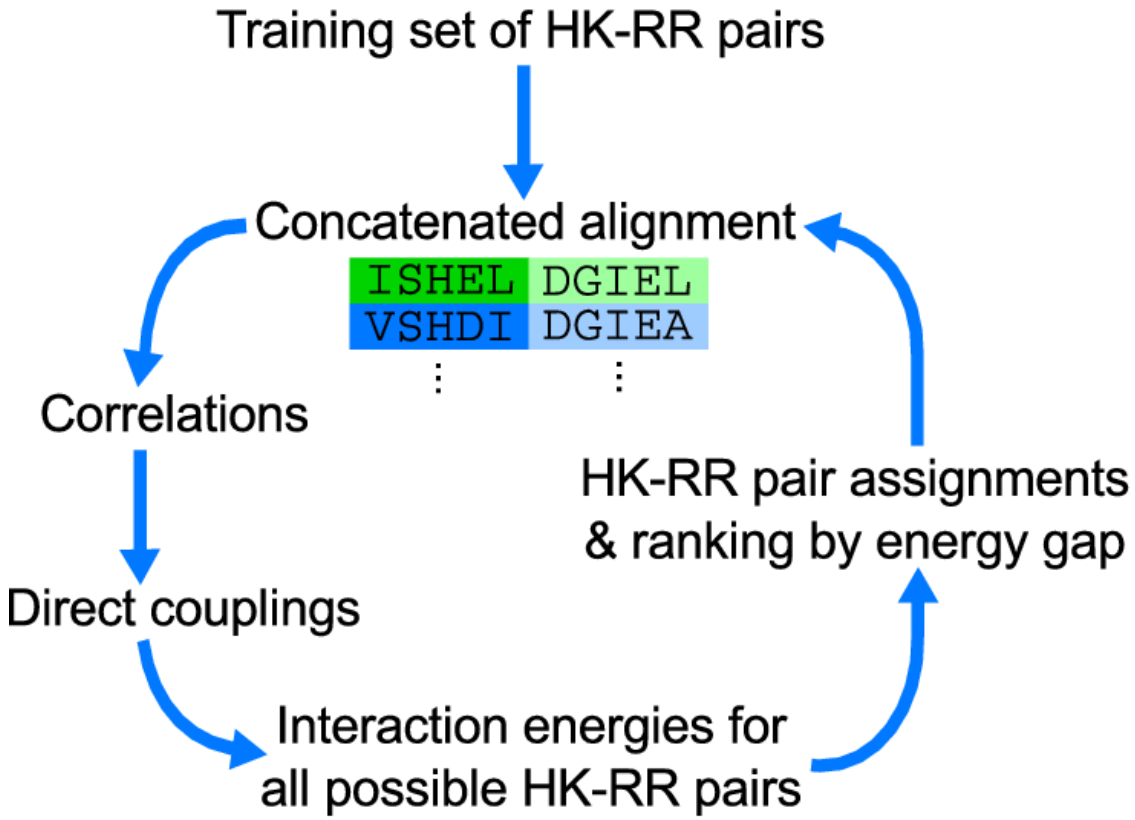
→ **Can we use these patterns of correlations to infer specific interaction partners?**

(1) Do protein families A and B interact or not?
(2) Within a species, which A interacts with which B?

DCA-based method

- **Iterative pairing algorithm (IPA)**





Approximately minimizes effective interaction energies between partners

DCA-based method

- **Correlations, direct couplings and interaction energies**

ISHEL	DGLPA
VSHDI	DGIEA
⋮	⋮

$$\rightarrow \begin{cases} f_i(\alpha) & i \in \{1,..,L\} \\ f_{ij}(\alpha,\beta) & \alpha \in \{A_1,..,A_{20},A_{21}=-\} \\ C_{ij}(\alpha,\beta) = f_{ij}(\alpha,\beta) - f_i(\alpha) f_j(\beta) \end{cases}$$

Pairwise maximum entropy model:

$$P(\alpha_1,...,\alpha_L) = \frac{1}{Z} \exp\left\{ -\left[\sum_{i=1}^{L} h_i(\alpha_i) + \sum_{i<j} e_{ij}(\alpha_i,\alpha_j) \right] \right\}$$

Inverse statistical physics
Cocco et al. (2018)

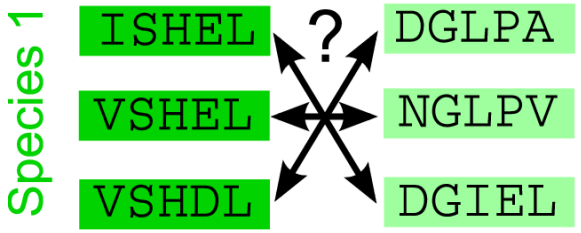
Mean-field approximation: $e_{ij}(\alpha,\beta) = C_{ij}^{-1}(\alpha,\beta)$
(20 *L* x 20 *L* matrix)

Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)

$e_{ij}(\alpha,\beta)$ much better predictor of 3D contact than $C_{ij}(\alpha,\beta)$

Weigt et al. (2009)
Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)

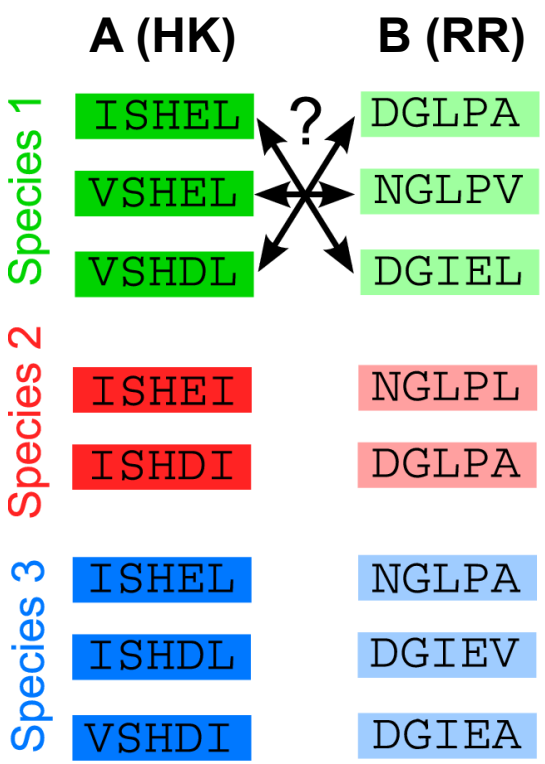
Interaction energies for all possible A-B (HK-RR) pairs in each species:

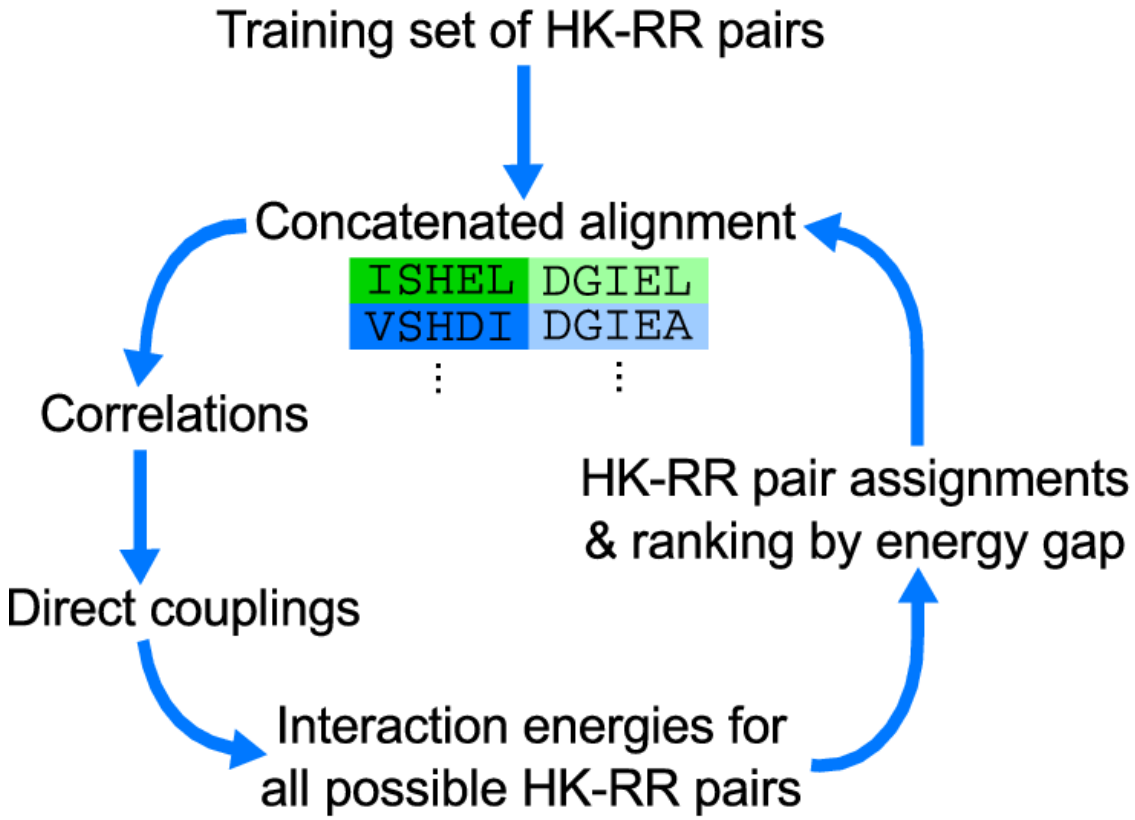


$$E(\alpha_1,...,\alpha_{L_A},\alpha_{L_A+1},...,\alpha_L) = \sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L} e_{ij}(\alpha_i,\alpha_j)$$

DCA-based method

- **Iterative pairing algorithm (IPA)**



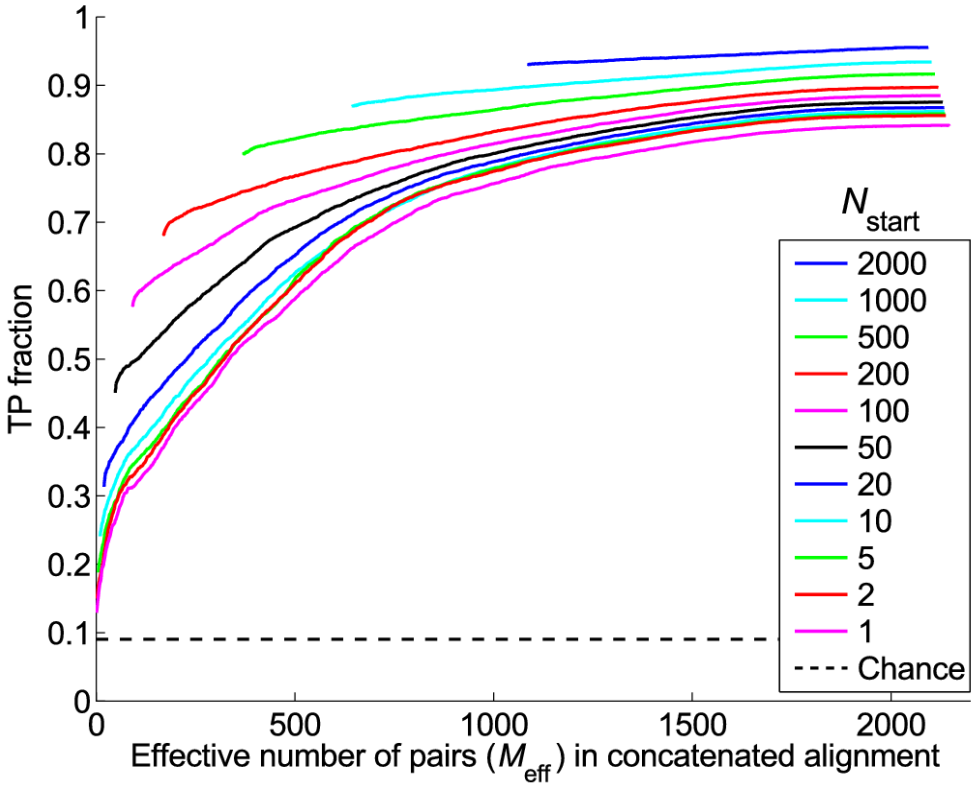


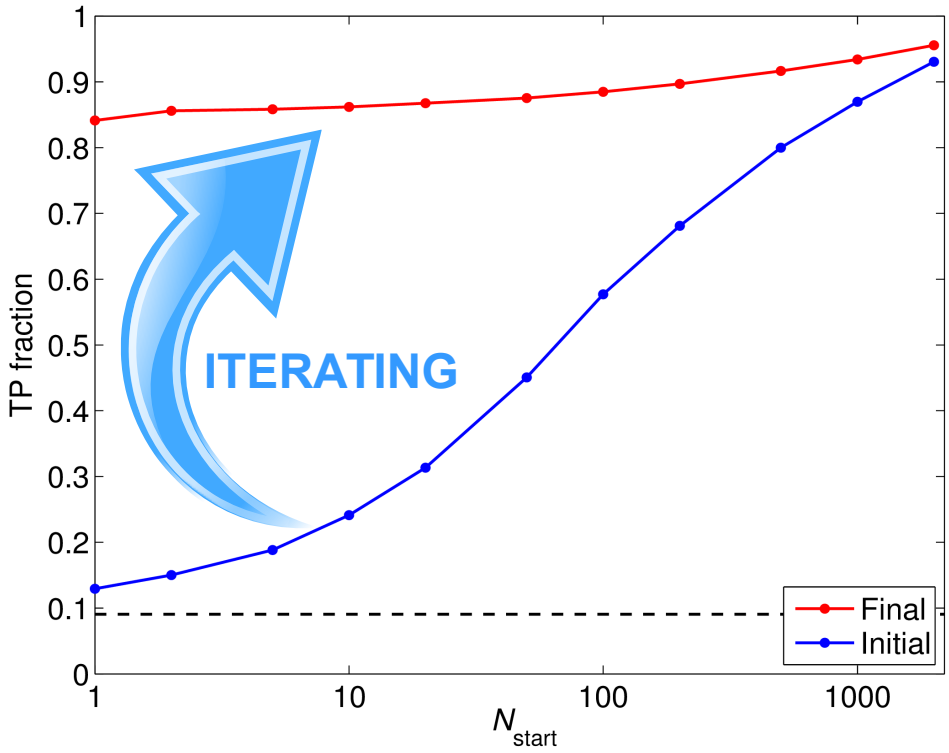
Approximately minimizes effective interaction energies between partners

Performance on real data

- **Prediction of interacting pairs among HK and RR proteins**

Dataset: **5064** pairs, mean **11.0 /species**; Meff=2091 (from full dataset with 23,424 pairs)
Nincrement=6; different Nstart (number of training HK-RR pairs)
Results averaged over 50 replicates, with different random choices of training pairs

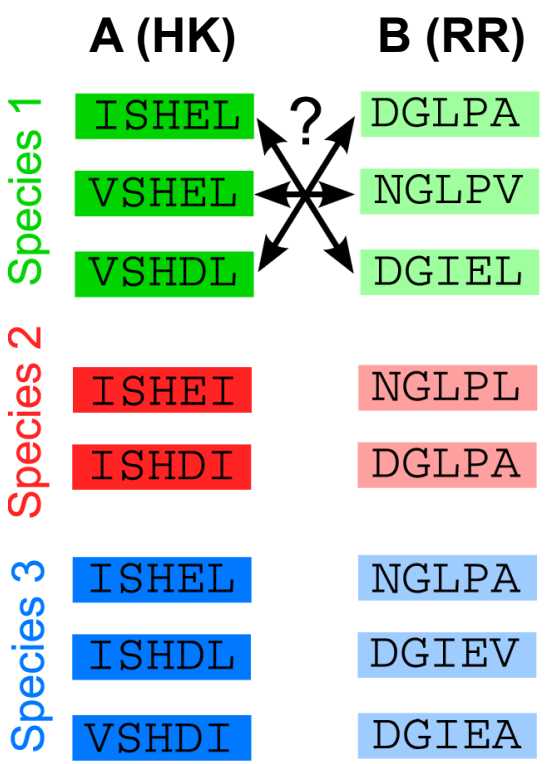


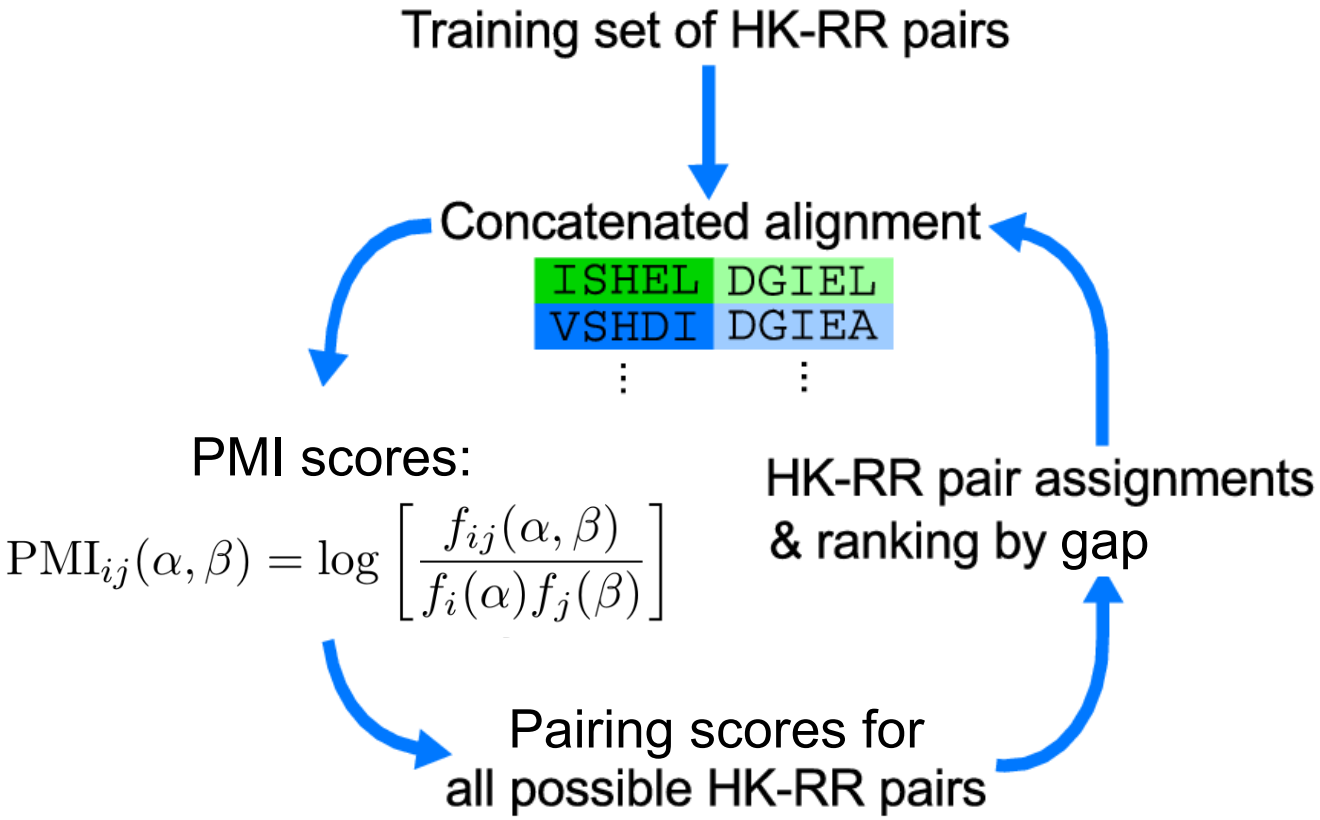


With **no training set**, TP fraction **0.84**

A mutual information (MI) based IPA

- **MI based iterative pairing algorithm (MI-IPA)**



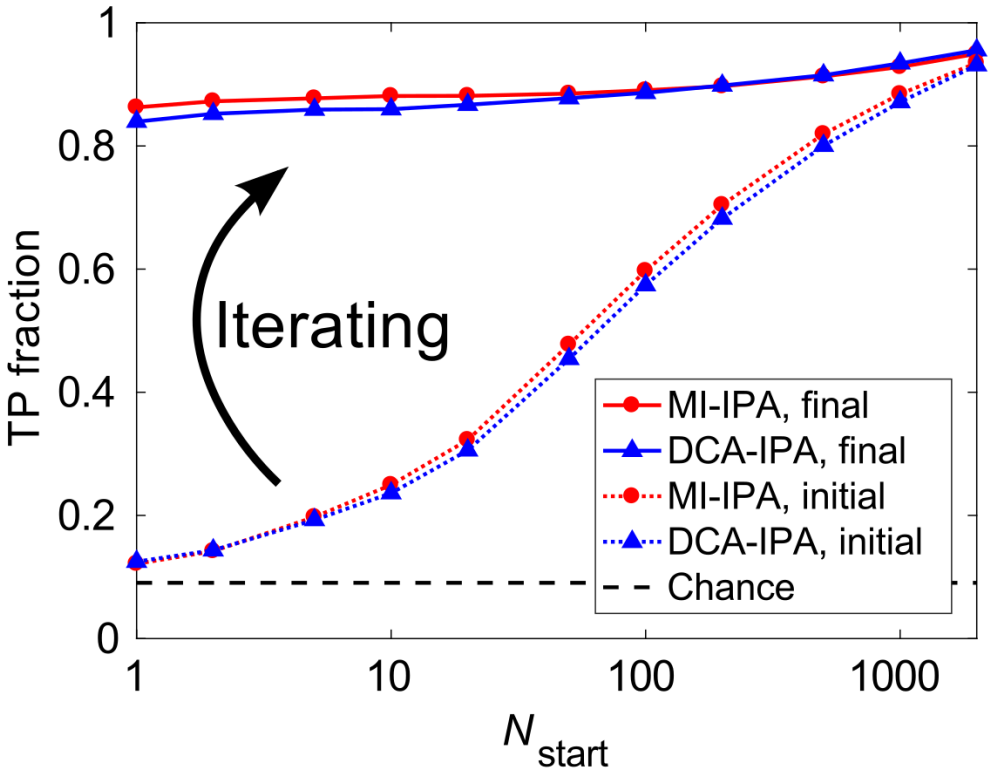


Approximately maximizes pairwise mutual information between partners

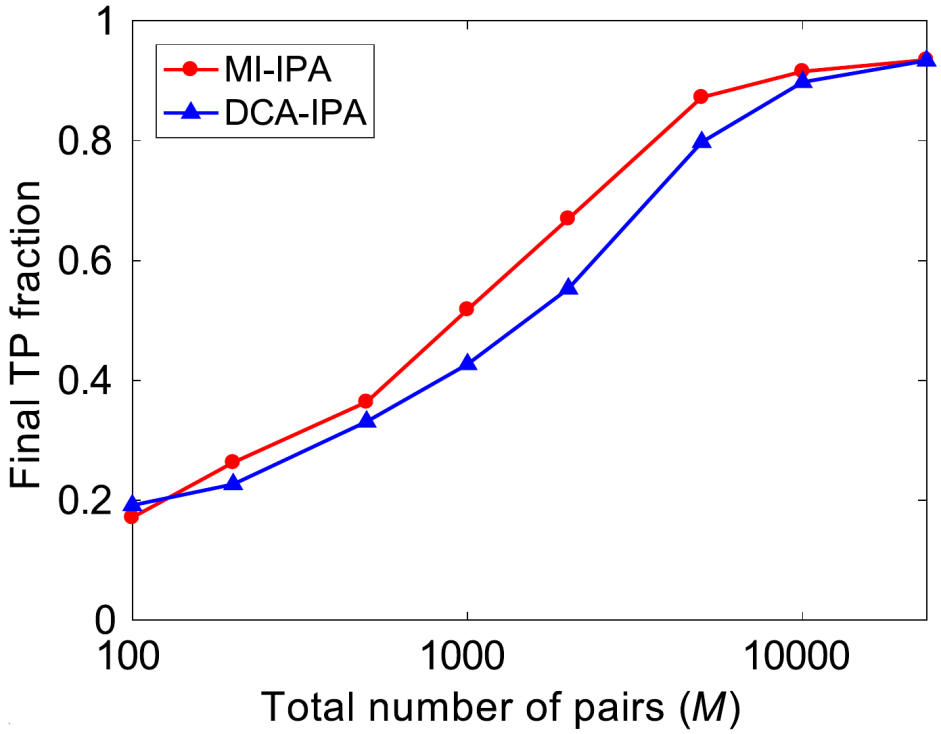
MI-IPA vs. DCA-IPA

- **Prediction of interacting pairs among HK and RR proteins**

Dataset of 5064 pairs, mean 11.0 /species
Nincrement=6; different Nstart (number of training HK-RR pairs)



No initial training set
Total dataset: 23,424 pairs



- Good performance even without a training set
- MI does as well and sometimes better than DCA (vs. contact prediction)

- **Potential signatures of the existence of an interaction between 2 protein families**

I. Inferring interaction partners from protein sequences

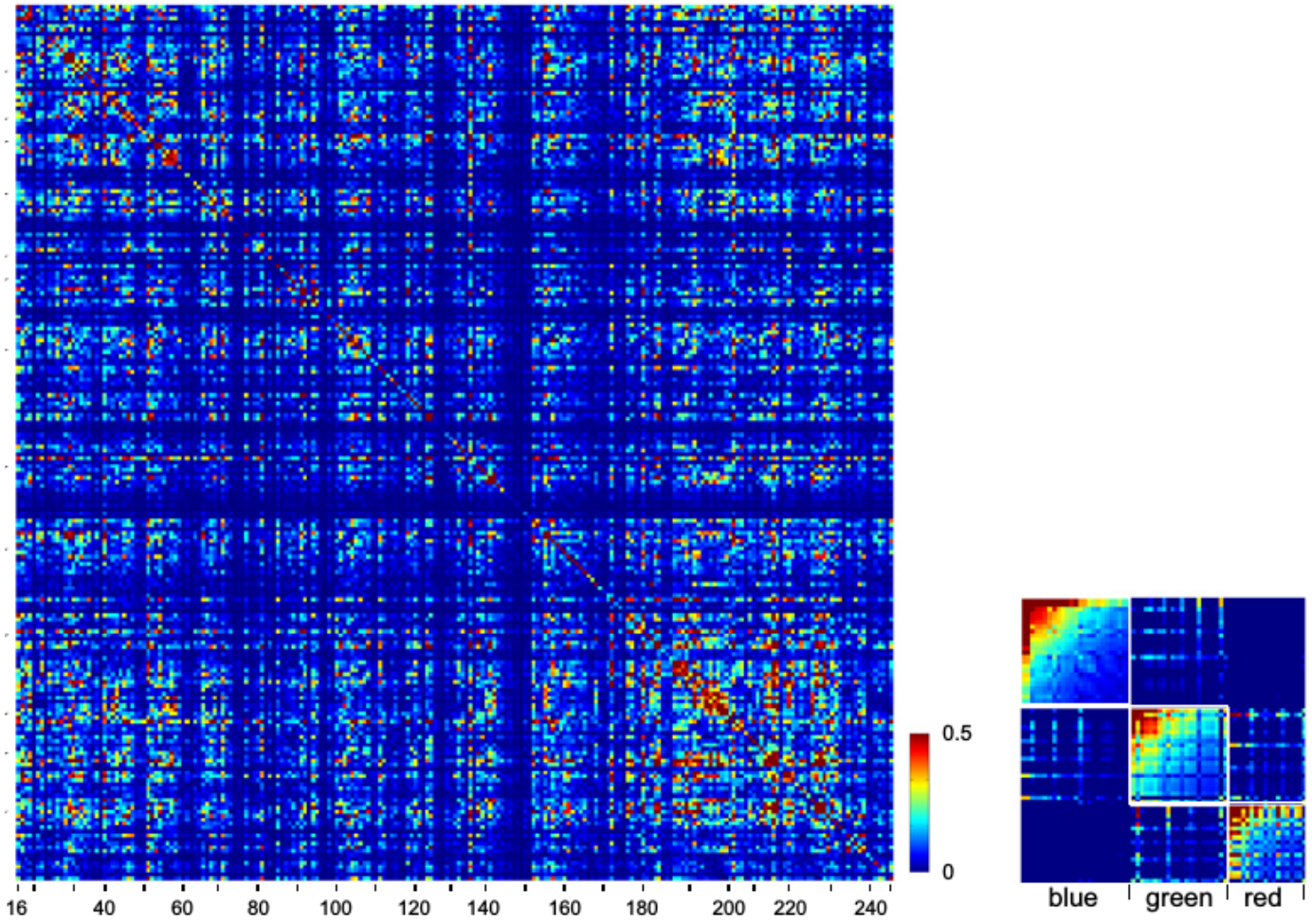
with Ned S. Wingreen, Lucy J. Colwell, Rob S. Dwyer

II. A physical interpretation of sectors of collectively correlated amino acids

with Ned S. Wingreen & Shou-Wen Wang

Introduction

- **Sectors: Halabi, Rivoire, Leibler & Ranganathan, 2009 (S1A serine protease)**

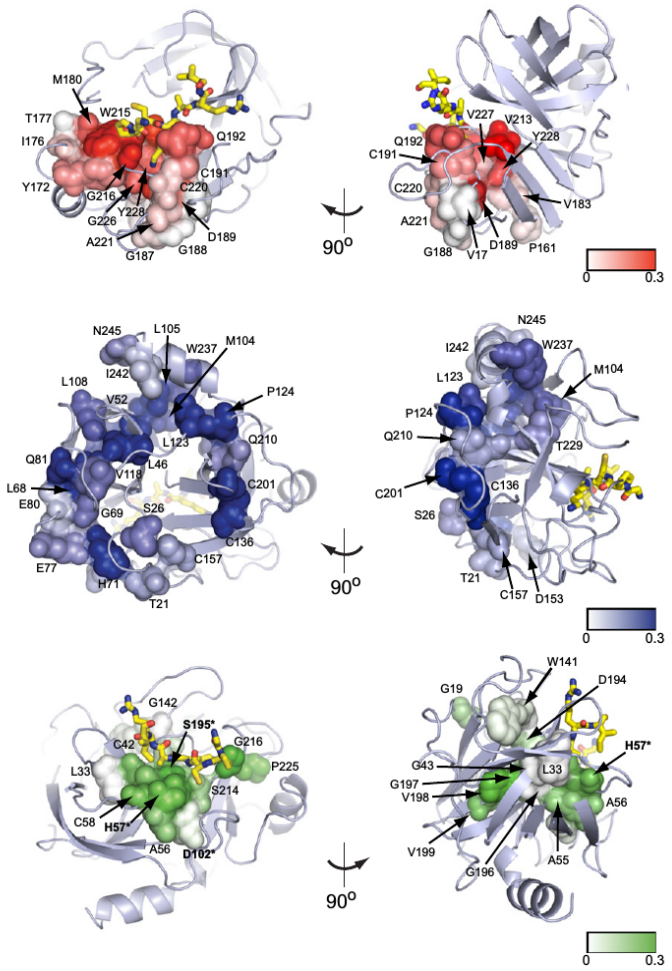


- Covariance matrix weighted by conservation reveals groups of collectively coevolving amino acids: “sectors”

- Sectors are obtained from the top modes of the weighted covariance matrix

Introduction

- **Sectors: Halabi, Rivoire, Leibler & Ranganathan, 2009 (S1A serine protease)**



Sectors are connected in 3D

Each is associated to different characteristics (mutagenesis + analysis of sequence divergence in each sector):

- primary catalytic specificity (substrate recognition) → function
- organism type → *phylogeny*
- whether they are catalytic or not → function

→ What is the physical origin of sectors?
→ Can we identify sectors from sequence data in a principled way?

A physical model for sectors

- **Additive traits and sector definition**

$T(\vec{\alpha}) = \sum_{l=1}^{L} \Delta_l(\alpha_l)$ where:

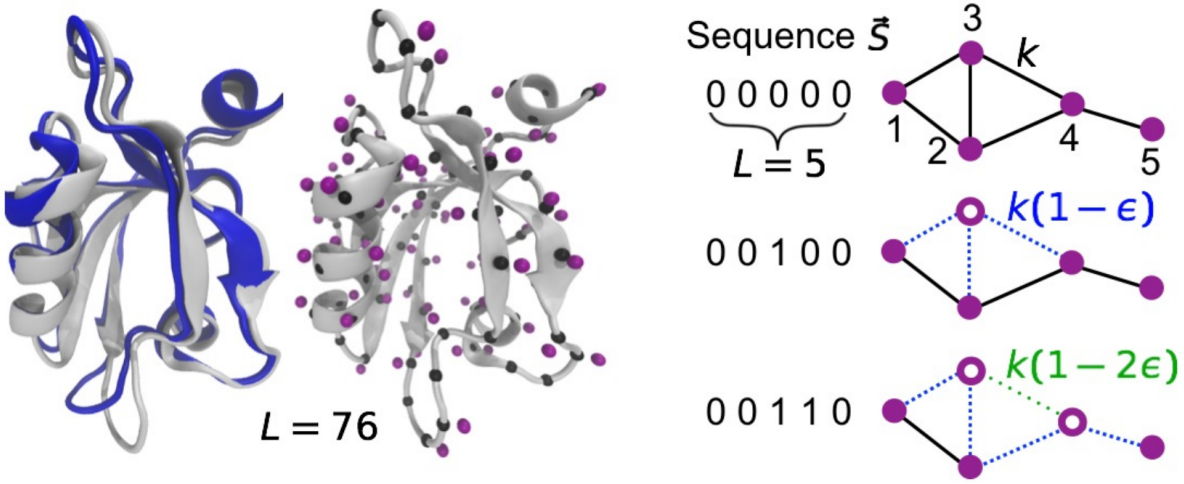
- $\vec{\alpha} = (\alpha_1, \ldots, \alpha_L)$: amino-acid sequence
- $\Delta_l(\alpha_l)$: mutational effect on T of a mutation to α_l at site l

Thermal stability De Pristo et al., 2005 Wylie & Shakhnovich, 2011 ; nonlinear selection on additive traits Otwinowski et al., 2018

Sector: set of sites with dominant mutational effects on a trait under selection

- **A "toy model" additive trait based on a concrete physical example**

- Coarse-grained elastic-networks → good description of many protein properties Bahar et al., 2010 Zheng et al., 2010 Yan et al., 2017
- Elastic-network model with sequence dependence (PDZ domain):



- Small deformations: $E = \frac{1}{2} \sum_{i,j} (\boldsymbol{r}_i - \boldsymbol{r}_i^0) M_{ij} (\boldsymbol{r}_j - \boldsymbol{r}_j^0) = \frac{1}{2} \delta \boldsymbol{r}^T M \delta \boldsymbol{r}$ M: Hessian matrix

- First-order perturbation analysis (in ε): $\delta E = E - E^{(0)} = \sum_{l=1}^{L} S_l \Delta_l$ Δ_l = effect of a mutation at site l

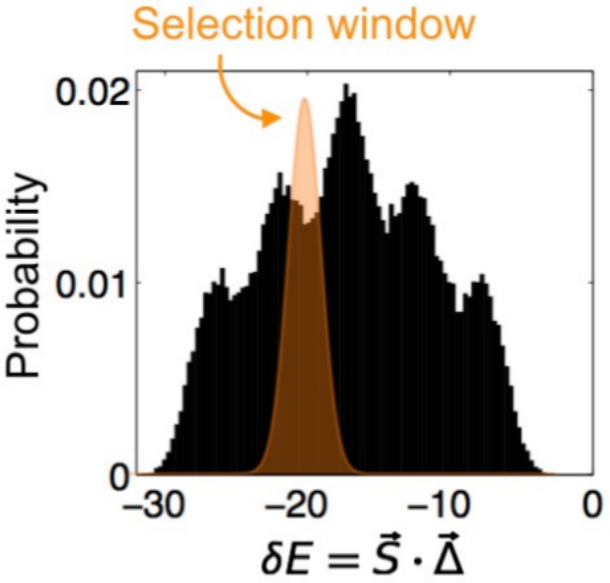
A physical model for sectors

- **Signature of a physical sector**

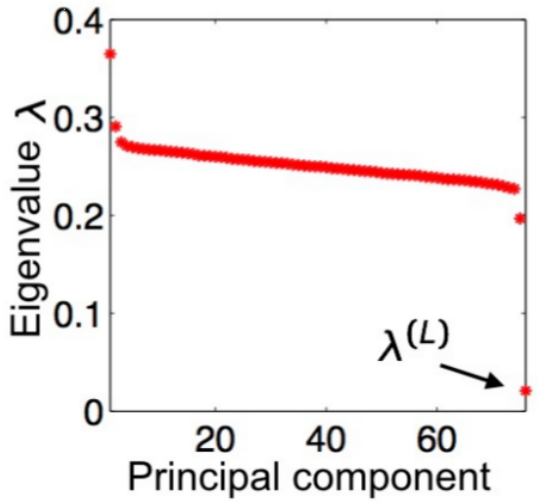
  - Selection on $\delta E = E - E^{(0)} = \sum_{l=1}^{L} S_l \Delta_l$

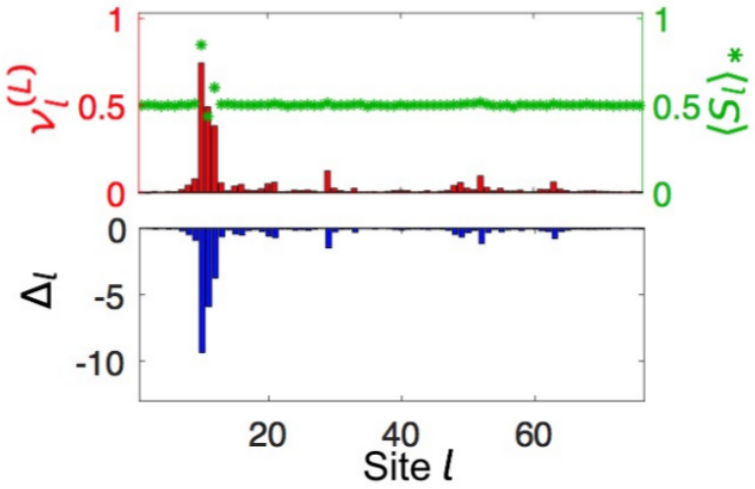
    Fitness $w(\vec{S}) = -\frac{\kappa}{2}\left(\sum_{l=1}^{L} \Delta_l S_l - \delta E^*\right)^2$

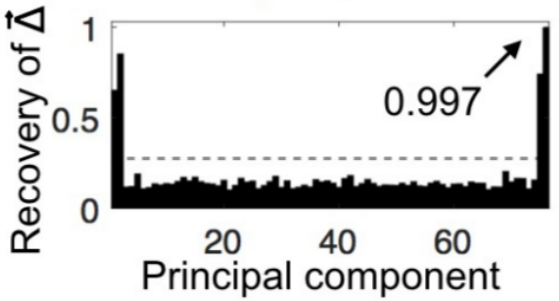
    + Boltzmann distribution
    → Gaussian selection window (selective weighting) $P(\vec{S}) = \frac{\exp(w(\vec{S}))}{\sum_{\vec{S}} \exp(w(\vec{S}))}$



- Eigendecomposition of the covariance matrix of selected sequences (PCA)





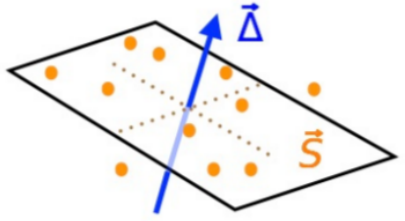


$$\text{Recovery} = \frac{\sum_l |\nu_l \Delta_l|}{\sqrt{\sum_l \nu_l^2}\sqrt{\sum_l \Delta_l^2}}$$

Selected sequences satisfy $\sum_l S_l \Delta_l = \vec{S} \cdot \vec{\Delta} \approx \delta E^*$

→ $\vec{\Delta}$ is a direction of particularly low variance (repulsive pattern in a generalized Hopfield model + field)

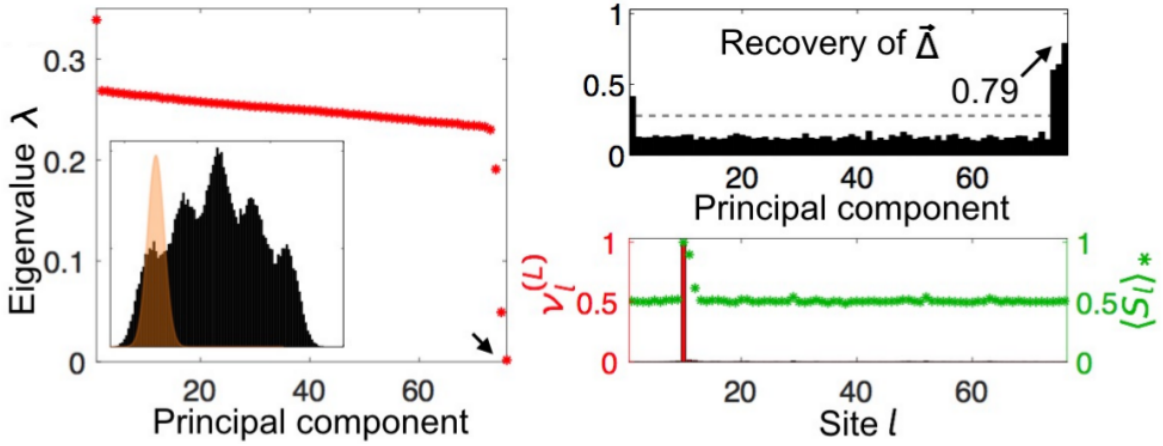
Cocco et al., 2011 & 2013



Detecting sectors from sequence data

- **Other small-variance directions can exist**

  Conservation → other small-variance directions (example: sites with $\langle S_l \rangle_* \approx 1$)



Strongly-biased selection

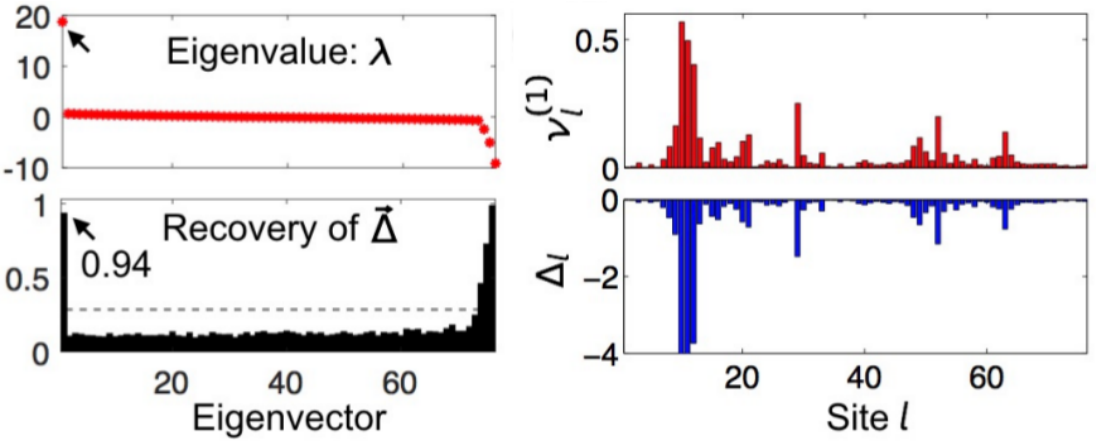
Components of the last eigenvector
Probability that a site is mutated (conservation)

- **Introducing a more robust method: ICOD**

  Inverse covariance matrix → mean-field approximation of couplings (cf. DCA)

$$C^{-1}_{ll'} \approx (1-\delta_{ll'})\ \kappa \Delta_l \Delta_{l'} + \delta_{ll'} \left(\frac{1}{P_l} + \frac{1}{1-P_l} \right)$$

Setting the diagonal to zero: $\tilde{C}^{-1}_{ll'} \approx (1-\delta_{ll'})\kappa \Delta_l \Delta_{l'}$
$\rightarrow \vec{\Delta} \otimes \vec{\Delta}$



→ robust to biased selection

Conclusion

Summary

- Sequence covariation → structure & protein-protein interactions & functional sectors
- Methods to predict PPI from sequences
- Selection on any relevant physical property of a protein → sector

Perspectives

- PPI: roles of correlations due to phylogeny and to interactions - with Martin Weigt
- Predicting new PPI; improving complex structure prediction

Acknowledgments

Ned S. Wingreen, Princeton U.
Lucy Colwell, Cambridge U.
Rob Dwyer, Princeton U.
Shou-Wen Wang, Tsinghua U. (now Harvard U.)

Aspen Center for Physics
Mohamed Barakat & Philippe Ortet
(CEA Cadarache)

References

Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS, **PNAS** 113(43): 12180-12185 (2016)
Bitbol AF, **PLOS Comput. Biol.** 14(11): e1006401 (2018)
Wang SW*, Bitbol AF* and Wingreen NS, ArXiv:1808.07149 (under review)

- **Other projects: evolution at the population scale**
  In particular: evolution of antimicrobial resistance

  → Loïc Marrec (earlier today)
  → Claude Loverdo (tomorrow afternoon)

LABORATOIRE
Jean Perrin

cnrs

Thanks!