Physical and evolutionary constraints *at the molecular scale*

Anne-Florence Bitbol

LABORATOIRE Jean Perrin

cnrs

SORBONNE UNIVERSITÉ

ICTS Program on “Living Matter” - Bangalore, India
April 2018

Introduction

Understanding proteins

- Heteropolymers made of 20 types of amino-acids (monomers) → ~20^{100} possible proteins
- A given natural protein folds into a compact and (almost) unique 3D **structure**
- It has specific **interactions** with other molecules → **function**

- Experiment: random proteins do not fold properly Socolich et al. (2005)
- Theory: for a random protein, interactions between monomers are random (the potential depends on the amino-acids involved) → spin-glass like: frustration
  → many locally stable low energy states

Bialek (2012)

→ Natural proteins are special

Introduction

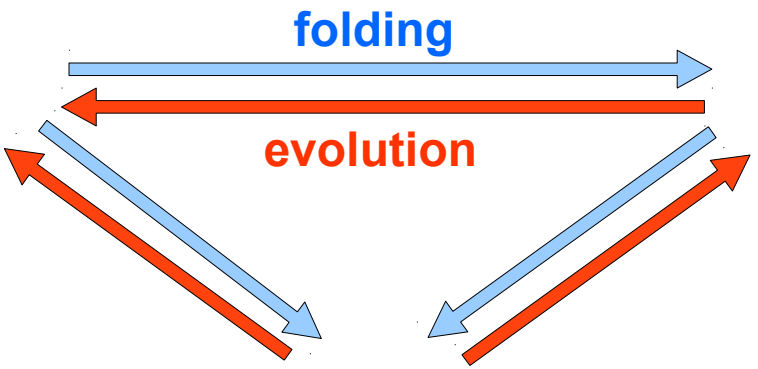
- **Understanding proteins**
  - Heteropolymers made of 20 types of amino-acids (monomers) → ~20^{100} possible proteins
  - A given natural protein folds into a compact and (almost) unique 3D **structure**
  - It has specific **interactions** with other molecules → **function**

  - Experiment: random proteins do not fold properly Socolich et al. (2005)
  - Theory: for a random protein, interactions between monomers are random (the potential depends on the amino-acids involved) → spin-glass like: frustration
    → many locally stable low energy states
    Bialek (2012)

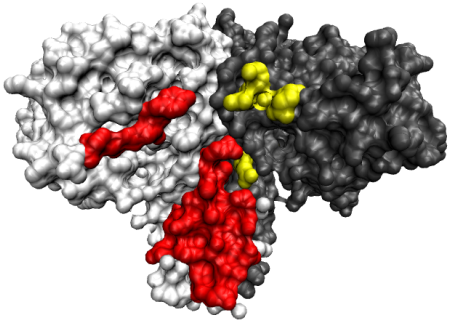
  → Natural proteins are special

Sequence

…ISHEL…



Structure

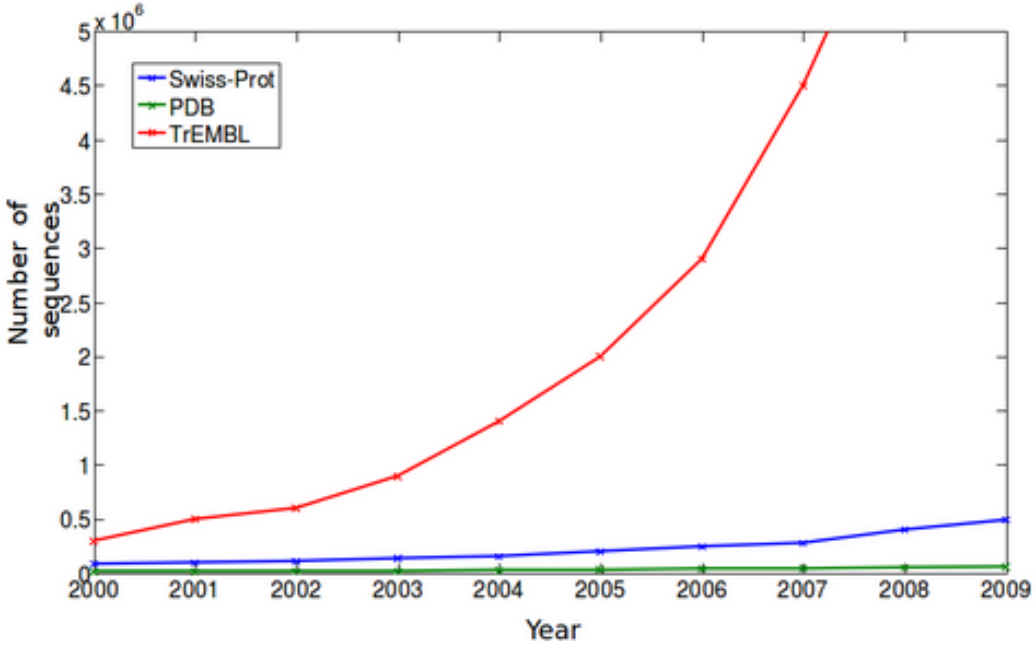


Function (including interactions)

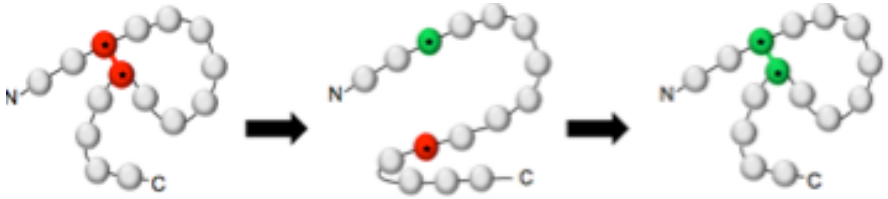
- Enzyme (histidine kinase)
- Phosphorylates another protein (response regulator)

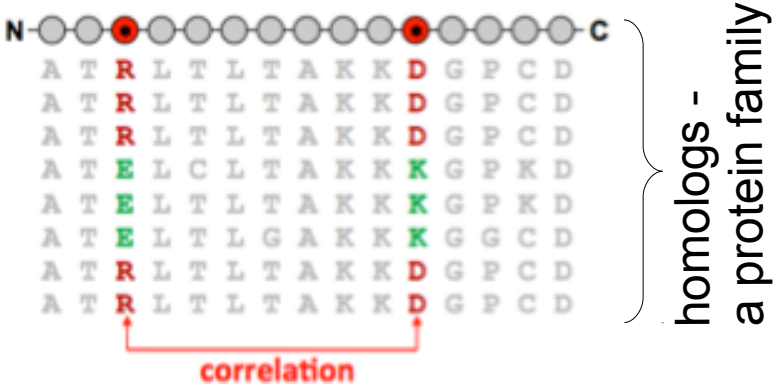
Introduction

- **Exploiting sequence data to understand natural proteins**



Recent data-driven approaches to infer structure and function from sequences





Evolutionary coupling between interacting residues
→ correlations in homolog sequence data inform us about structure
BUT... observed correlations can be indirect A ↔ B ↔ C

Outline

I. **Predicting protein structure from sequence data**
   Direct coupling analysis (DCA)

II. **Inferring interaction partners from protein sequences**
   Iterative pairing algorithm (IPA)

Direct coupling analysis (DCA)

Predicting protein structure from sequence data

Direct coupling analysis (DCA) *Weigt, White et al. (2009)*

- **Statistical inference method (cf. tutorial)**
  - **Goal: construct a global model for the protein family**

    L-site probability distribution (probability of observing a given sequence in the protein family considered): $P(\alpha_1, \alpha_2, \ldots, \alpha_L)$

  - **Construct it from the data (data-driven approach)**

    Observations retained: one- and two-body frequencies (choice)

    ```
    ...ISHEL...
    ...VSHDI...
    ...VSHEL...
          ⋮
    ```

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

    Multiple choices are consistent with these observations...

  - **Maximum entropy principle**

    Maximize $S = -\sum_{\{\alpha_1, \ldots, \alpha_L\}} P(\alpha_1, \ldots, \alpha_L) \log\left[P(\alpha_1, \ldots, \alpha_L)\right]$ (Shannon entropy) + constraints

    Yields the least-structured model consistent with the observations

  - **Resulting global model**

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

    → Potts model

    one-body terms - fields

    two-body terms - (direct) couplings

Direct coupling analysis (DCA)

- **Statistical inference method (cf. tutorial)**

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

Pairwise maximum entropy model and direct couplings:

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

One needs to determine the fields and couplings consistent with the observations

$$\sum_{\alpha_k, k\neq i} P(\alpha_1,...,\alpha_L) = f_i(\alpha_i)\,,$$

$$\sum_{\alpha_k, k\notin\{i,j\}} P(\alpha_1,...,\alpha_L) = f_{ij}(\alpha_i,\alpha_j)\,.$$

→ very hard problem! (inverse problem - general)
→ many approximation methods
Cocco et al. (2017) - in the context of proteins

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

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

- Simplest approximation, can be derived through a small-coupling expansion
- Has proved rather good in the case of proteins

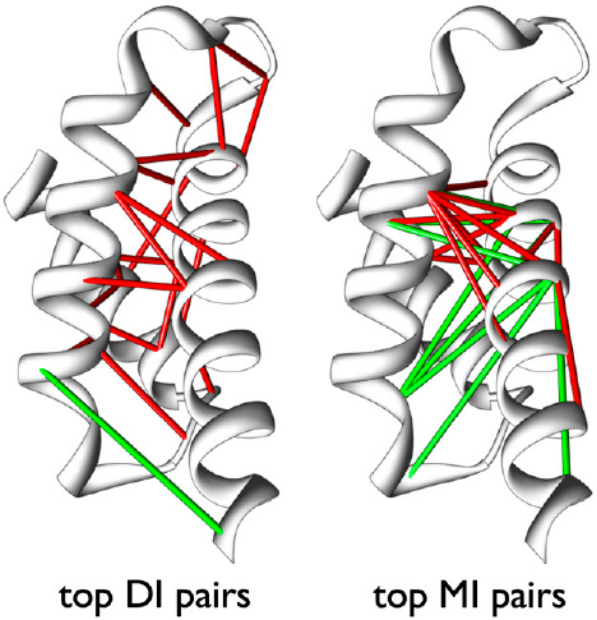
Direct coupling analysis (DCA)

- **Performance**

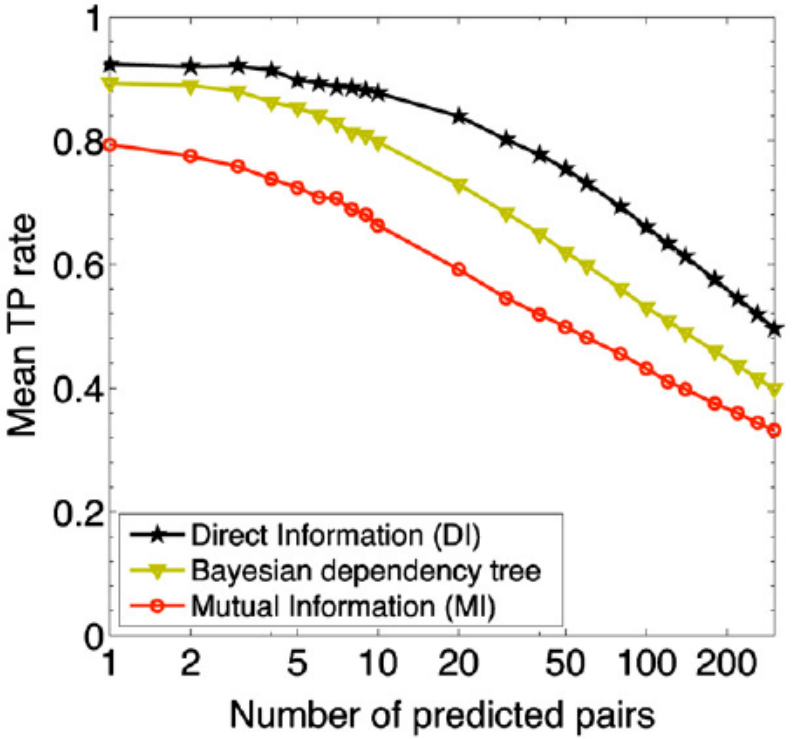
$e_{ij}(\alpha, \beta)$ much better predictor of 3D contact than $C_{ij}(\alpha, \beta)$ | Mutual Information

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

Morcos, Pagnani et al. (2011):



Bacterial Sigma factor region 2.
Top 20 DI / MI predictions
(distance along the backbone > 4).
Red: distance <8 Å; green: others.



Mean TP rate for 131 domain families
vs. number of top-ranked contacts

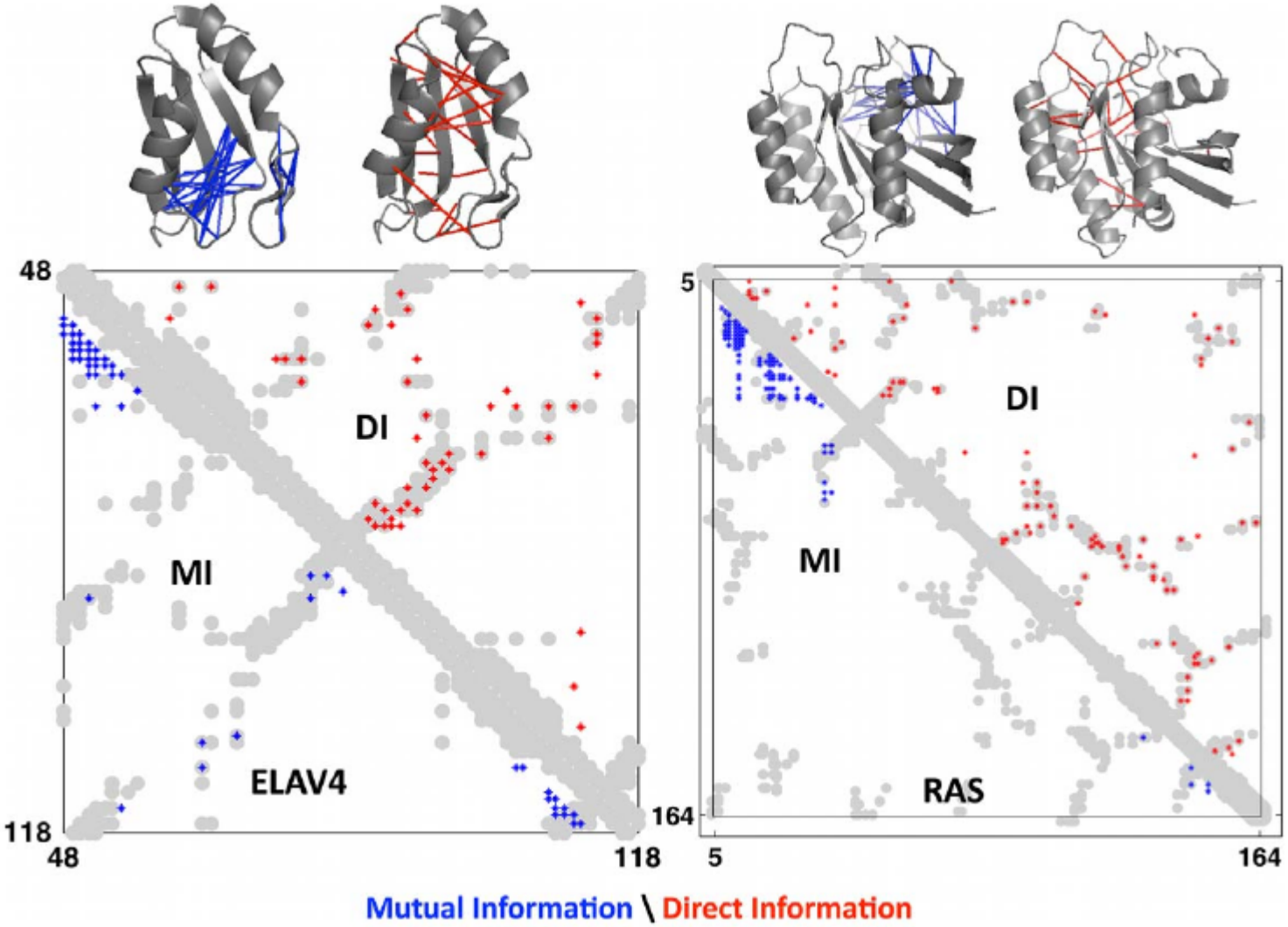
Direct coupling analysis (DCA)

- **Performance**

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

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

Marks, Colwell et al. (2011):



Predicted contacts for DI (red) overlap more accurately with the contacts in the experimentally observed structure (grey), than those for MI (blue).

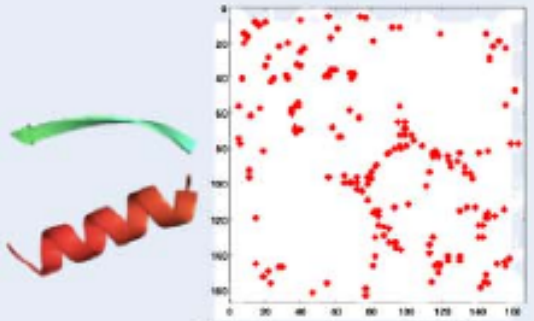
Direct coupling analysis (DCA)

- **Full prediction of protein 3D structure from sequence data**

Marks, Colwell et al. (2011):

Analyze the highest scoring pairs to produce ranked list of residue pairs which we predict to be close in 3D space. Use these pairs as predicted close "evolutionary inferred contacts" , EICs, in folding calculations

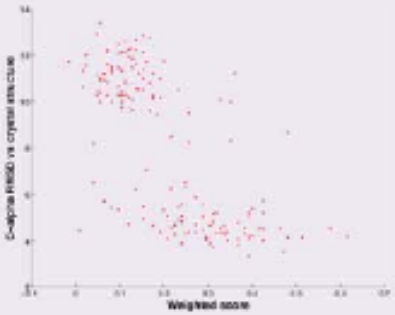
```
assign (resid 143 and name CA) (resid 123 and name CA) 4 4 3
assign (resid 16 and name CA) (resid 10 and name CA) 4 4 3
assign (resid 141 and name CA) (resid 82 and name CA) 4 4 3
assign (resid 129 and name CA) (resid 87 and name CA) 4 4 3
assign (resid 92 and name CA) (resid 11 and name CA) 4 4 3
assign (resid 116 and name CA) (resid 81 and name CA) 4 4 3
```



predicted contacts (EICs)

Start with extended structure
use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain

Rank predicted structures using quality measure of backbone alpha torsion and beta sheet twist



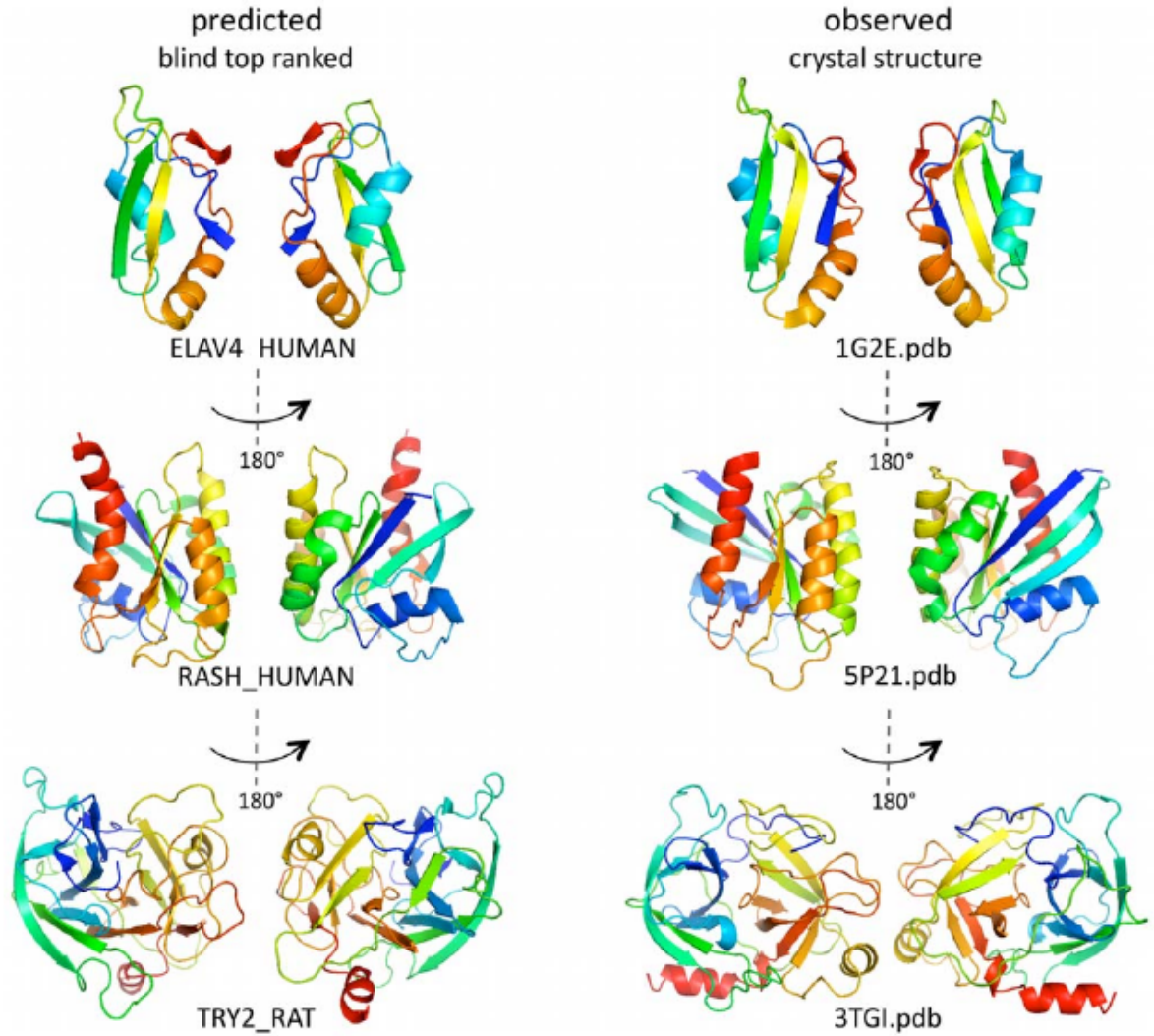
good scores

bad scores

Direct coupling analysis (DCA)

- **Full prediction of protein 3D structure from sequence data**

Marks, Colwell et al. (2011):



Results for 3 proteins:
- predicted top ranked 3D structure (left)
- experimentally observed structure (right)
Each structure in front and back view

- **Limitations**
  - DCA requires large alignments of homologous proteins (~ a few hundreds)
  - DCA requires a high diversity within these alignments

Inferring interaction partners from protein sequences

Anne-Florence Bitbol

with Robert S. Dwyer, Lucy J. Colwell and Ned S. Wingreen

LABORATOIRE Jean Perrin

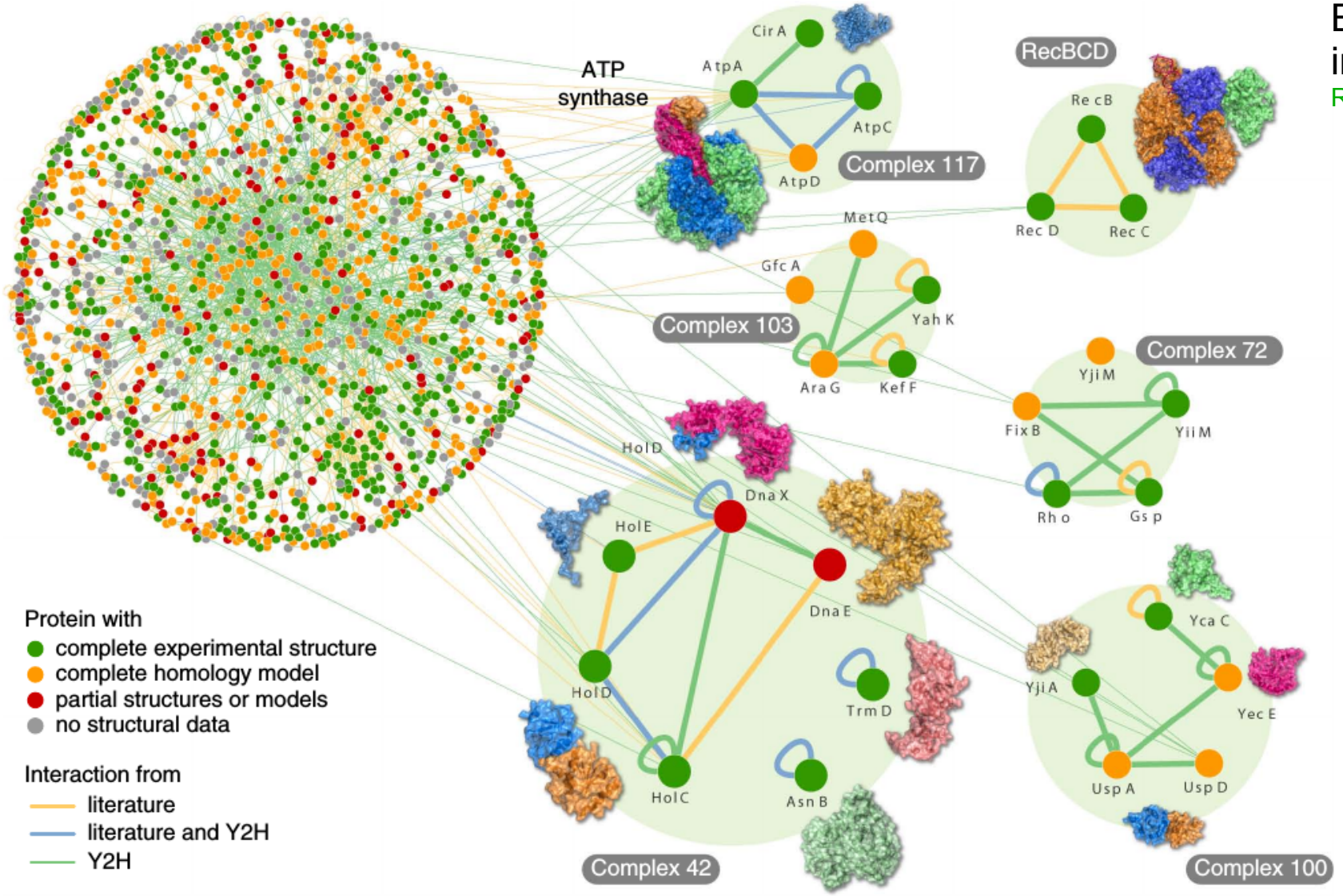
cnrs

SORBONNE UNIVERSITÉ

PRINCETON UNIVERSITY

Introduction

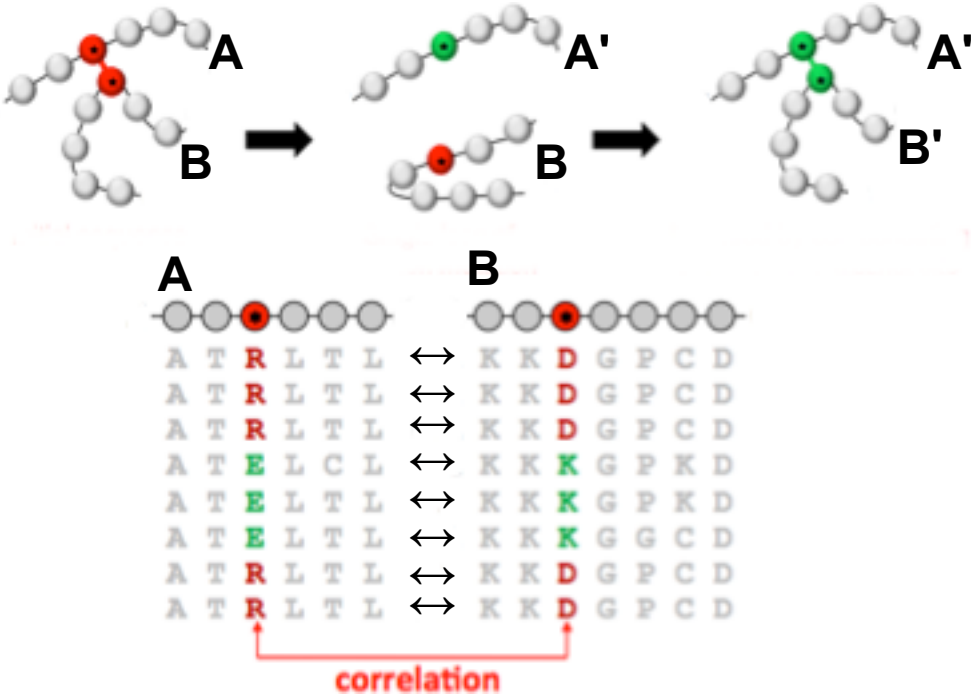
- **Protein-protein interactions**
  - Crucial for functional mutiprotein complexes, signaling pathways etc.
  - Systematic experimental determination is tedious

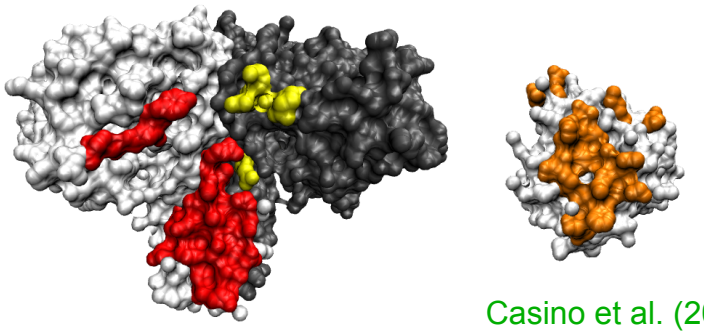


Binary PPI in *E. coli*
Rajagopala et al. (2014)

Introduction

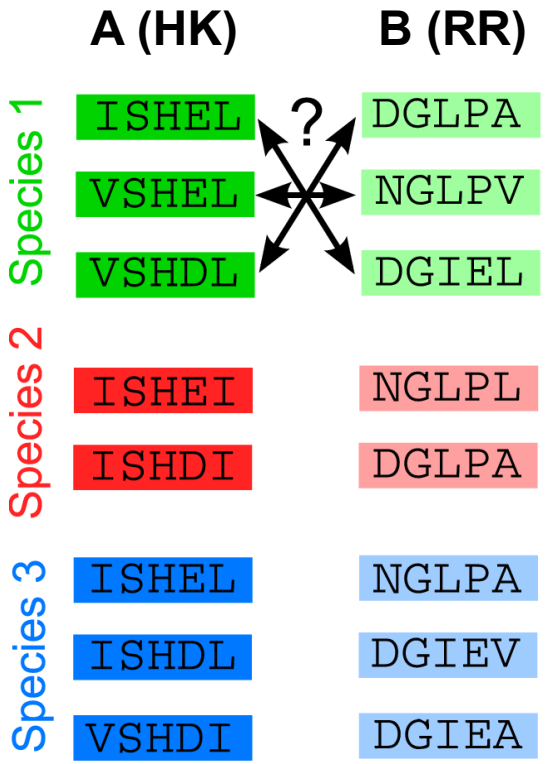
- **Co-evolution and correlations *between interacting partners***





Casino et al. (2009)

Often, several paralogs in each species



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

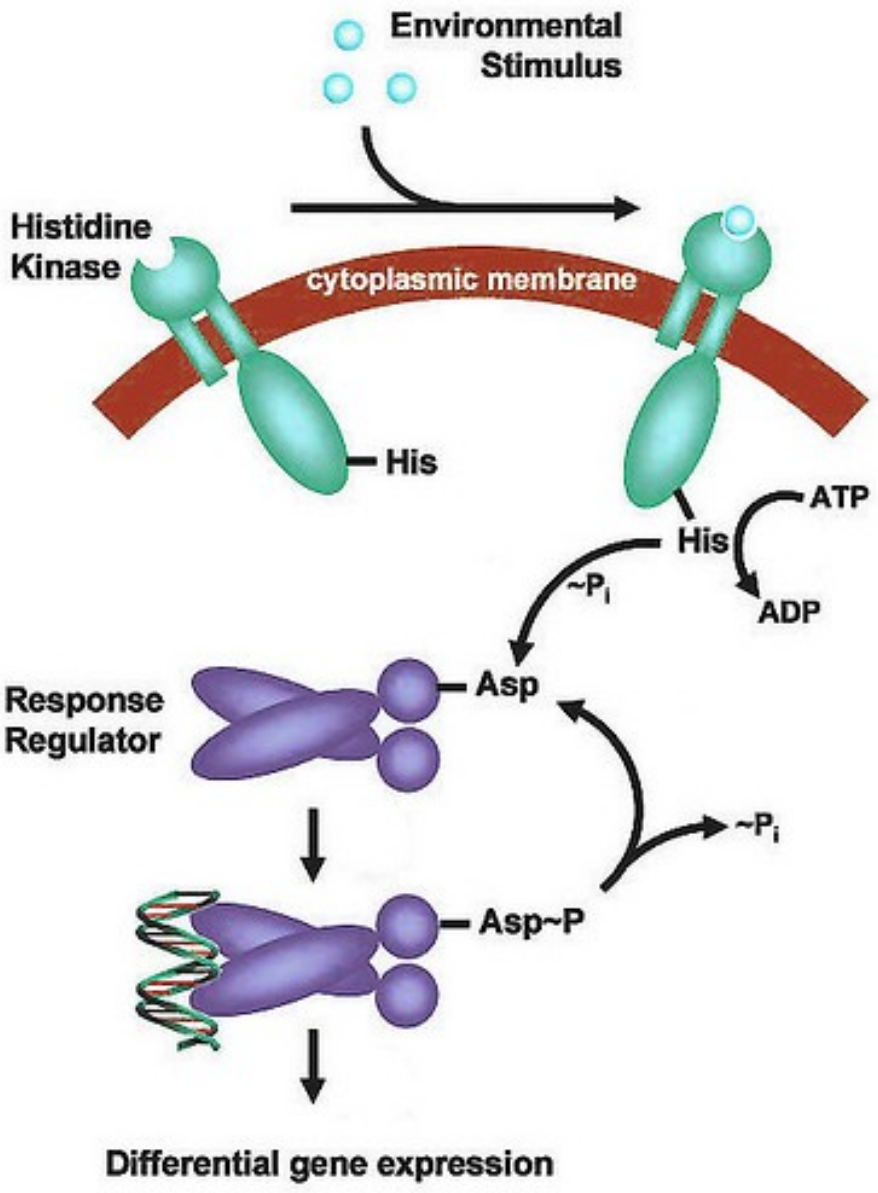
(1) Do protein families A and B interact or not?
(2) Within a species, which A interacts with which B?

(1) Do protein families A and B interact or not?

(2) Within a species, which A interacts with which B?

Dataset

- **Bacterial two-component systems:**



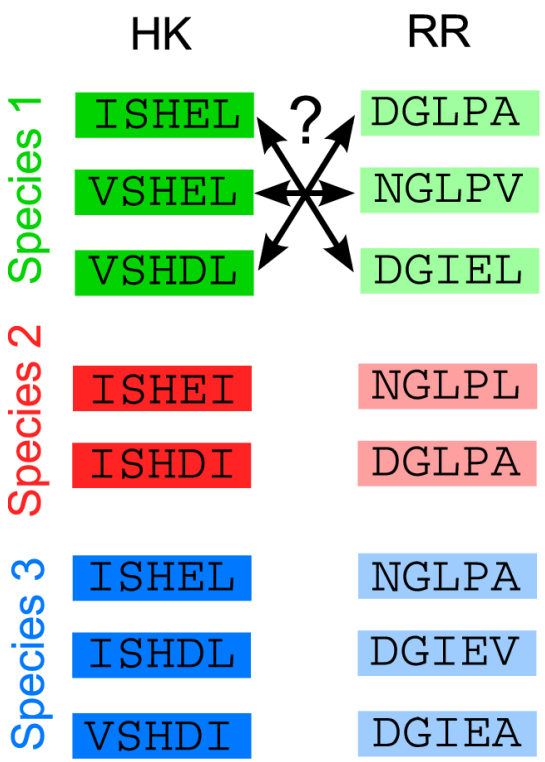
- Histidine kinase (HK)
- Response regulator (RR)

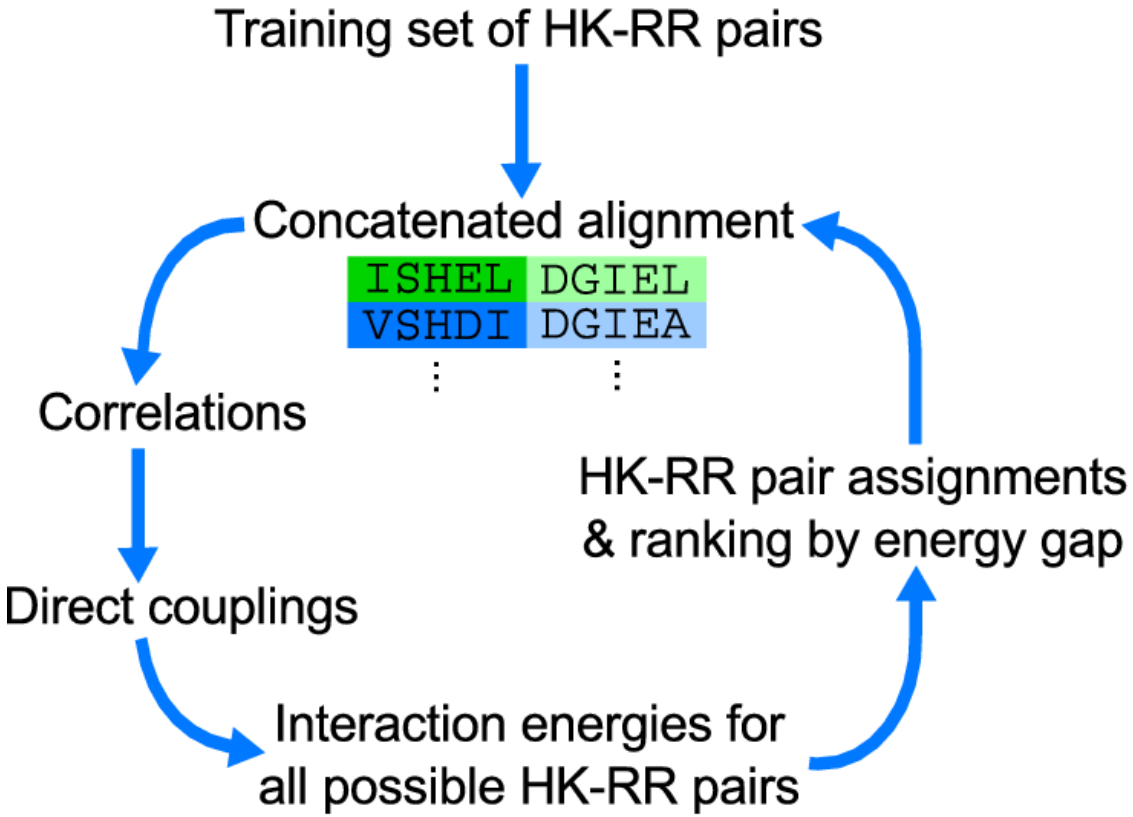
- Many fully-sequenced genomes (2,758 here)
- Lots of known interaction partners
- **Many paralogs** per species

→ a **great benchmark**

Method

- **Iterative pairing algorithm**





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 and direct couplings:

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

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

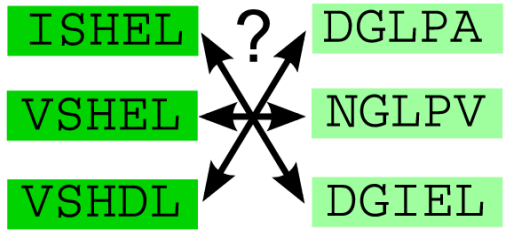
Morcos, Pagnani et al. (2011)
Marks, Colwell et al. (2011)

$e_{ij}(\alpha,\beta)$ 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 HK-RR pairs in each species:

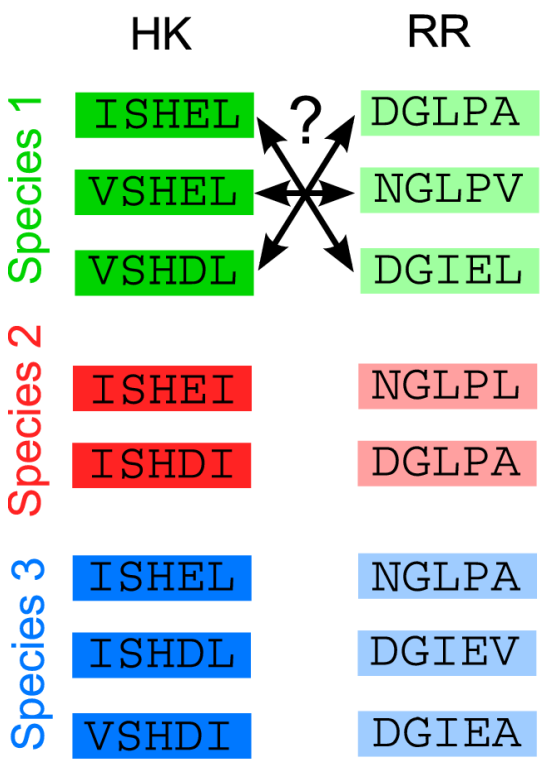
Species 1

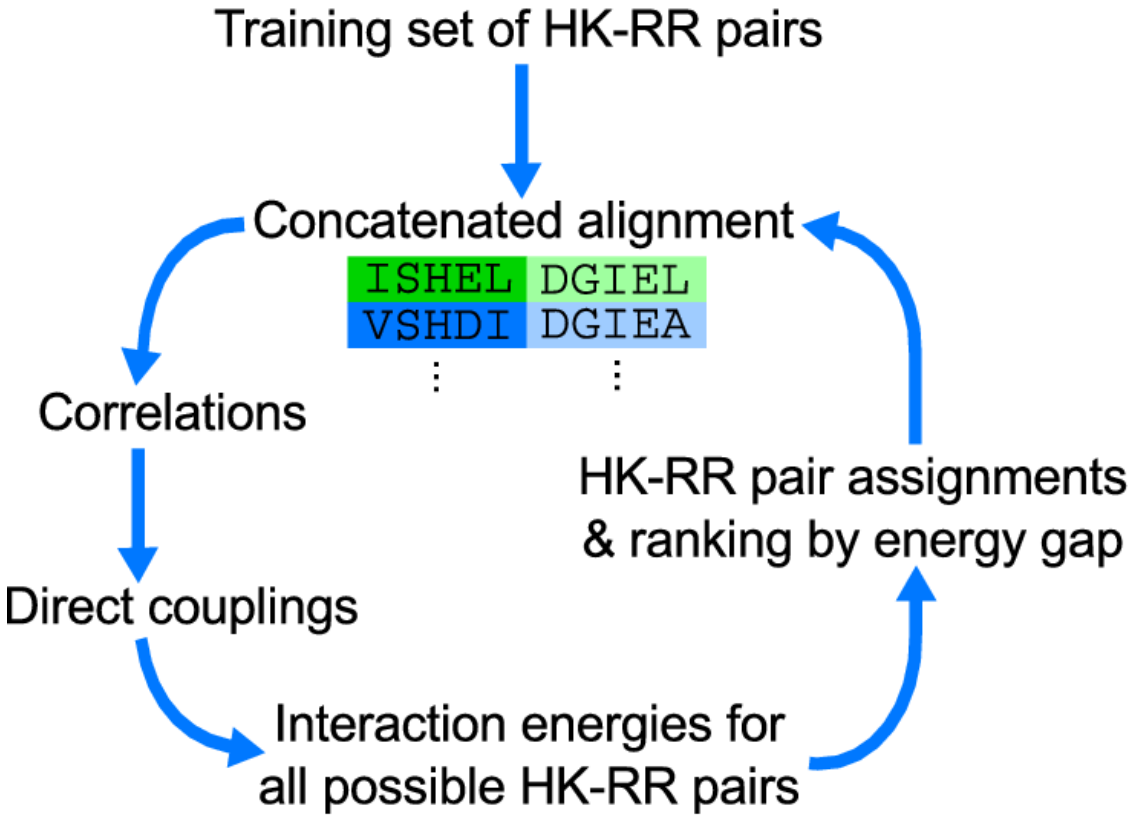


$$E(\alpha_1,...,\alpha_{L_A},\alpha_{L_A+1},...,\alpha_L) = \sum_{i=1}^{L_A} \sum_{j=L_A+1}^{L} e_{ij}(\alpha_i,\alpha_j)$$

Method

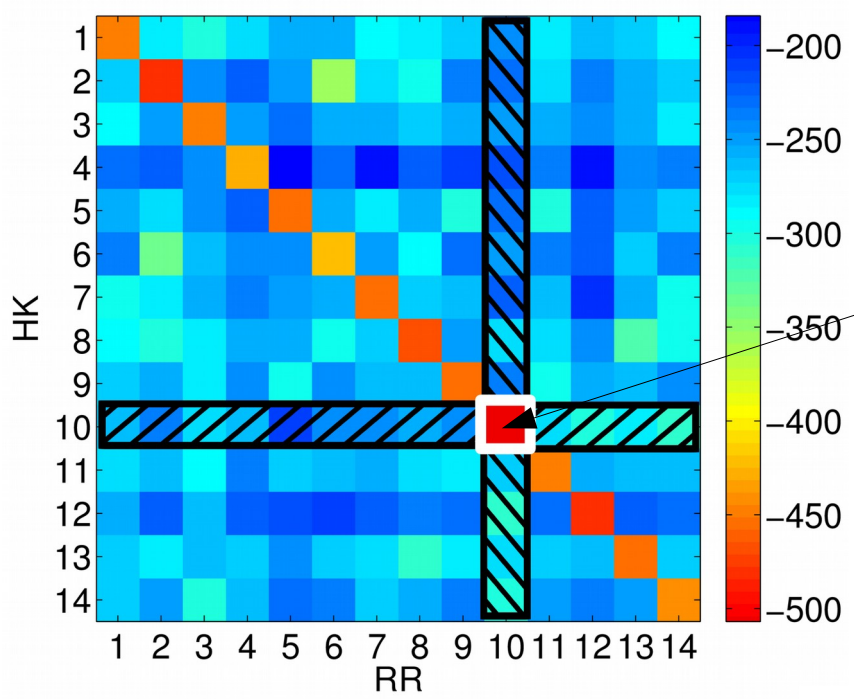
- **Iterative pairing algorithm**





Method

- **HK-RR pair assignments and ranking by gap**



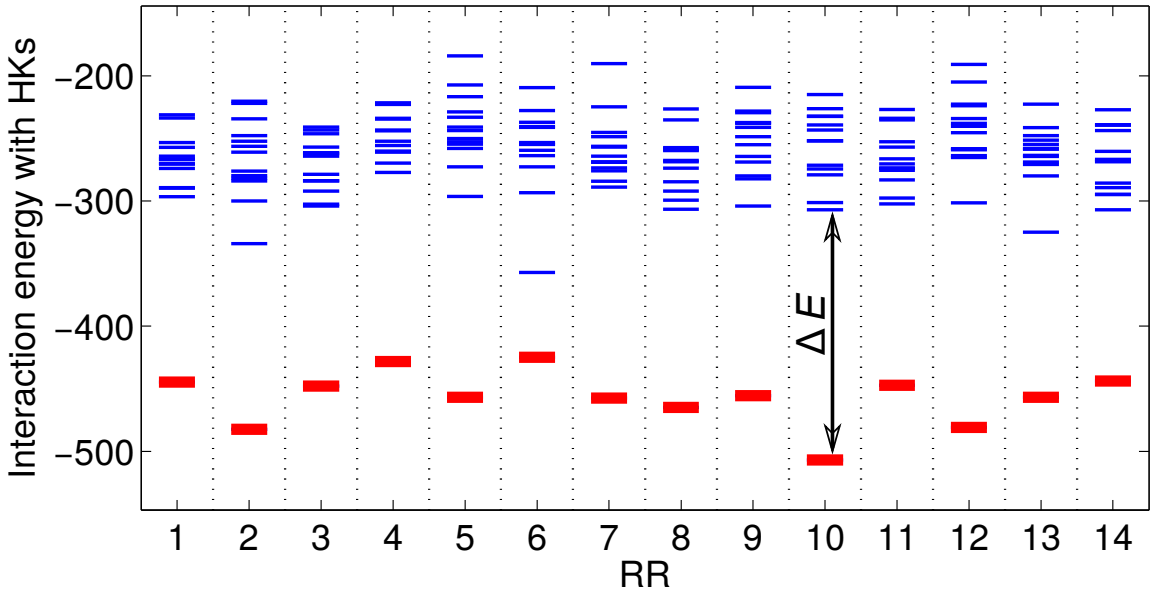
- Interaction energies between HK and RR from *E. coli* K-12 MG1655

Lowest energy
→ make this pair

- Once a pair is made, suppress this HK and RR from further consideration (1 to 1 interactions)

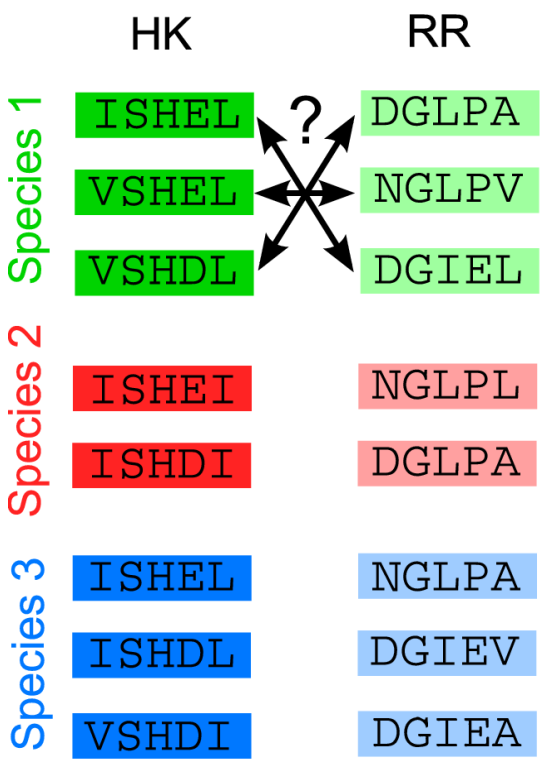
- Energy gap → confidence score used to rank pairs

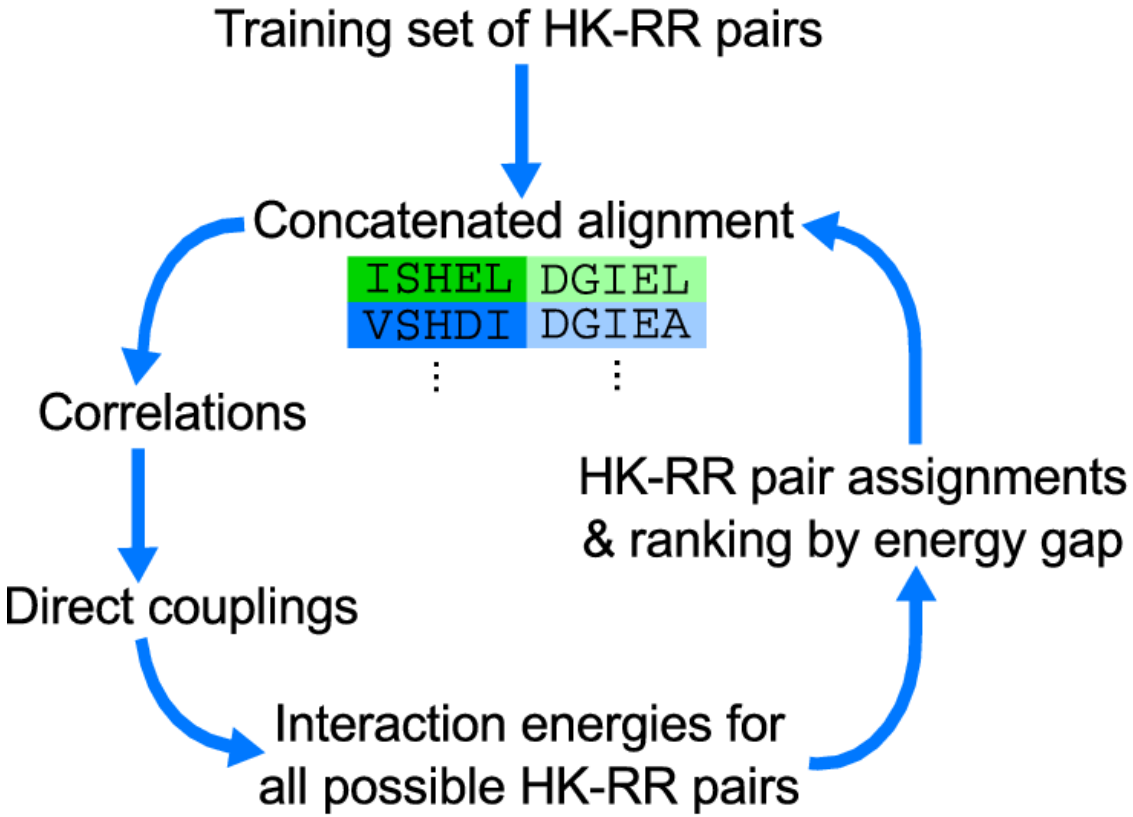
- Those with largest score are **included in the concatenated alignment (training set) at the next iteration**



Method

- **Iterative pairing algorithm**

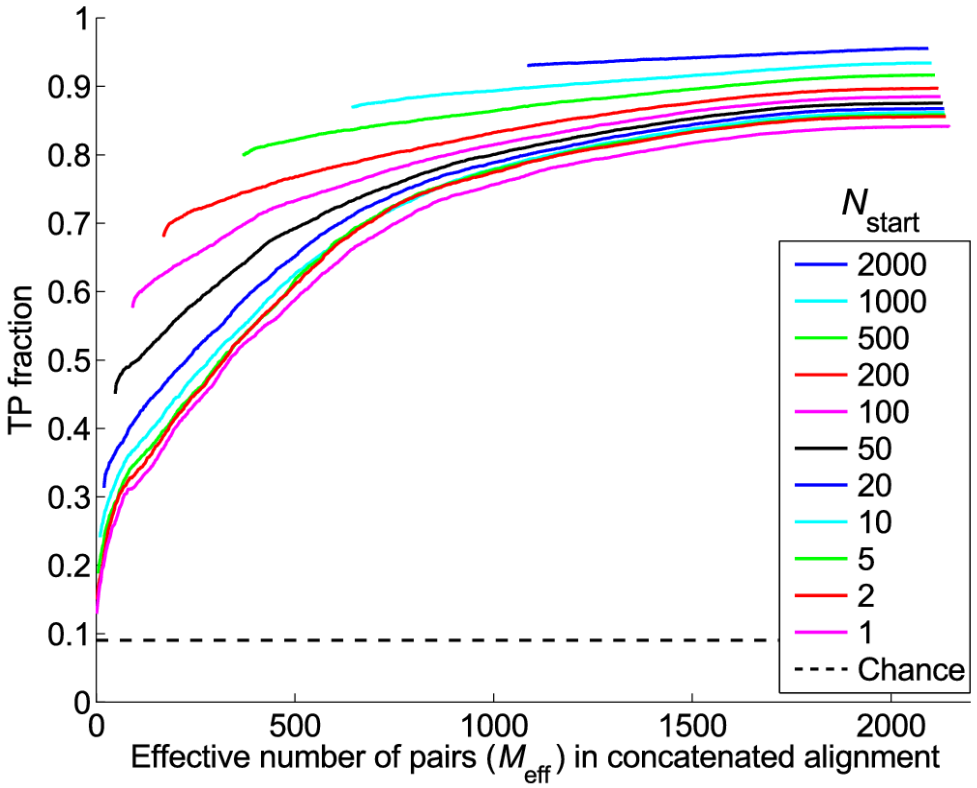


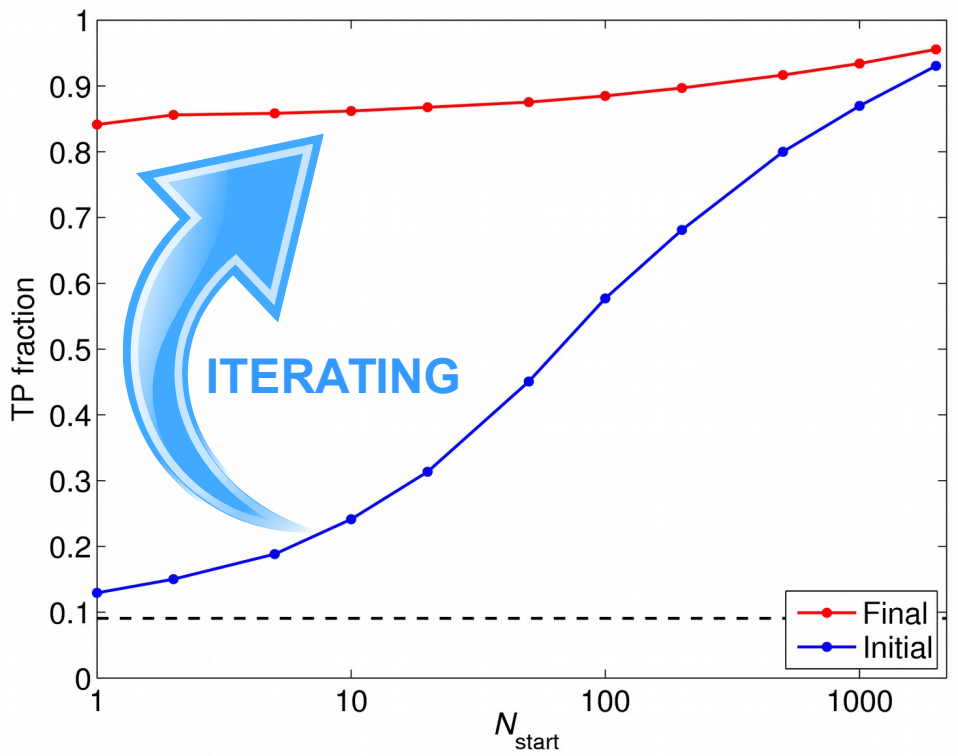


Effect of training set size (Nstart)

- **Progression of TP fraction and final value vs. Nstart**

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





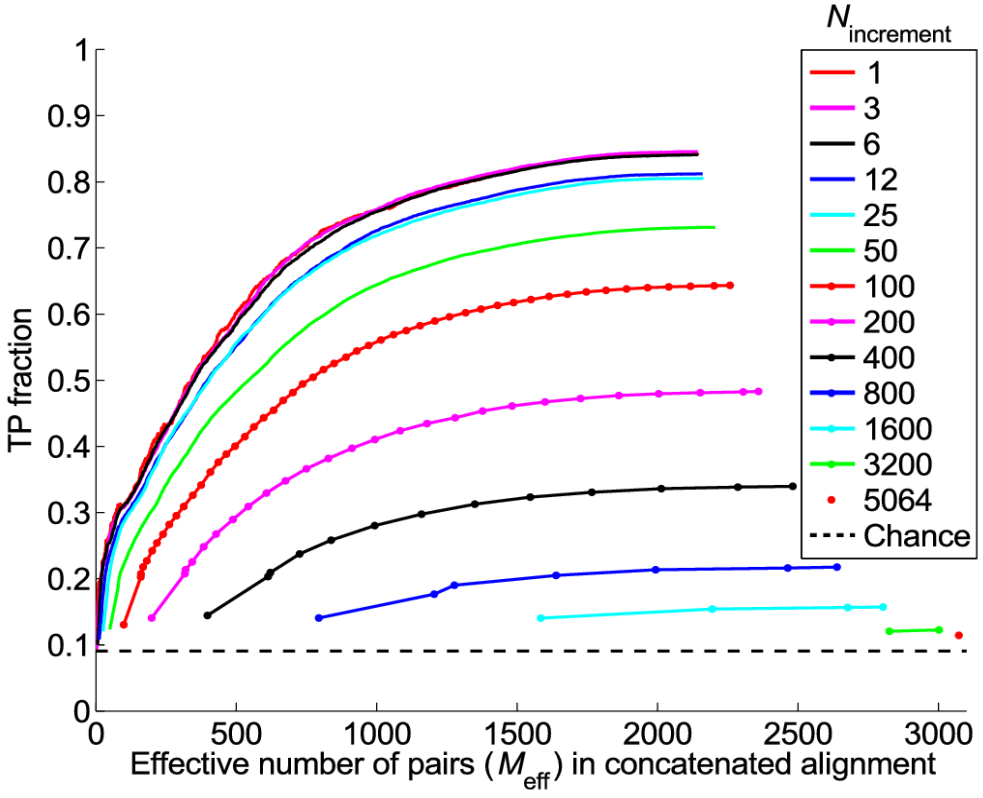
- High final TP fractions thanks to iterating
- Weak dependence of the final TP fraction on Nstart

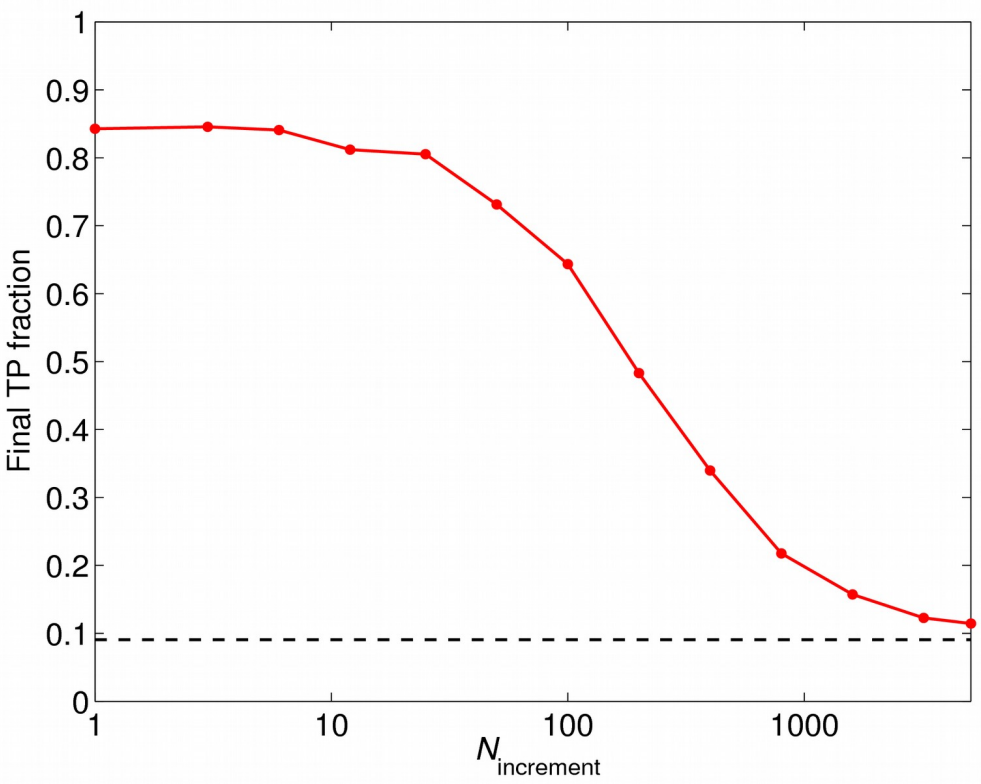
→ **Can we do without a training set?**

Starting from random pairings

- **Progression of TP fraction and final value vs. Nincrement**

Starting from random within-species pairings; different Nincrement
Results averaged over 50 replicates, with different initial random pairings





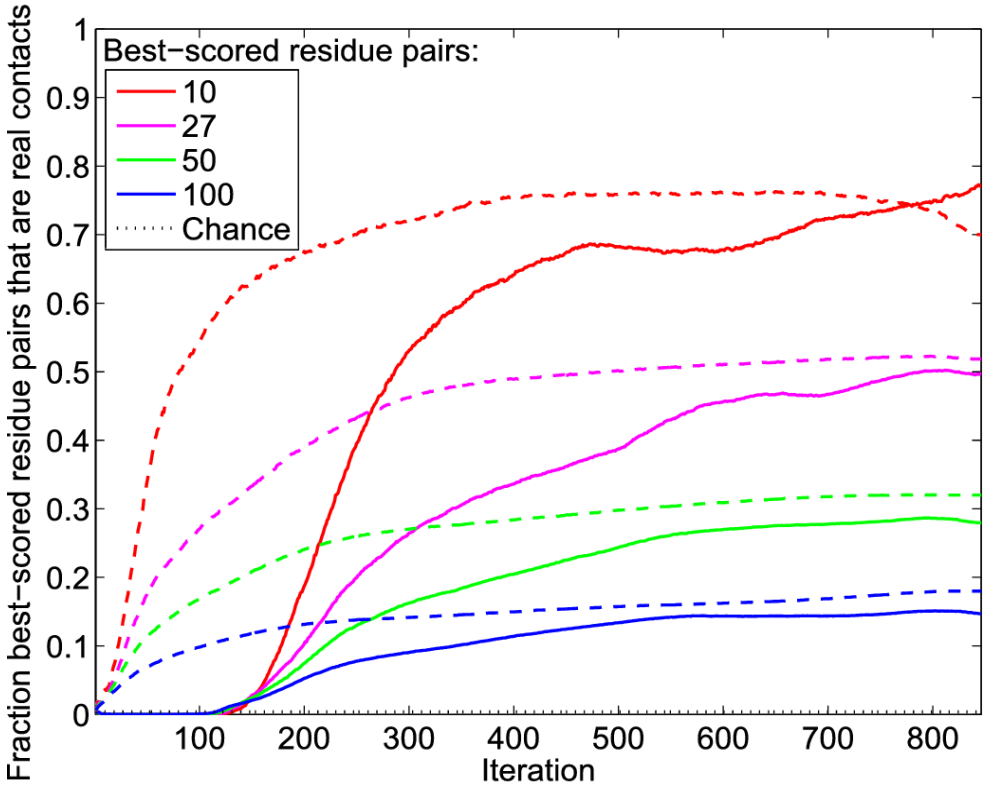
- Performance increases as Nincrement decreases
- With **no training set**, TP fraction **0.84** for Nincrement=6 and lower; **robust**: std 0.04

- **How does the TP fraction increase at early stages (in the concatenated alignment)?**

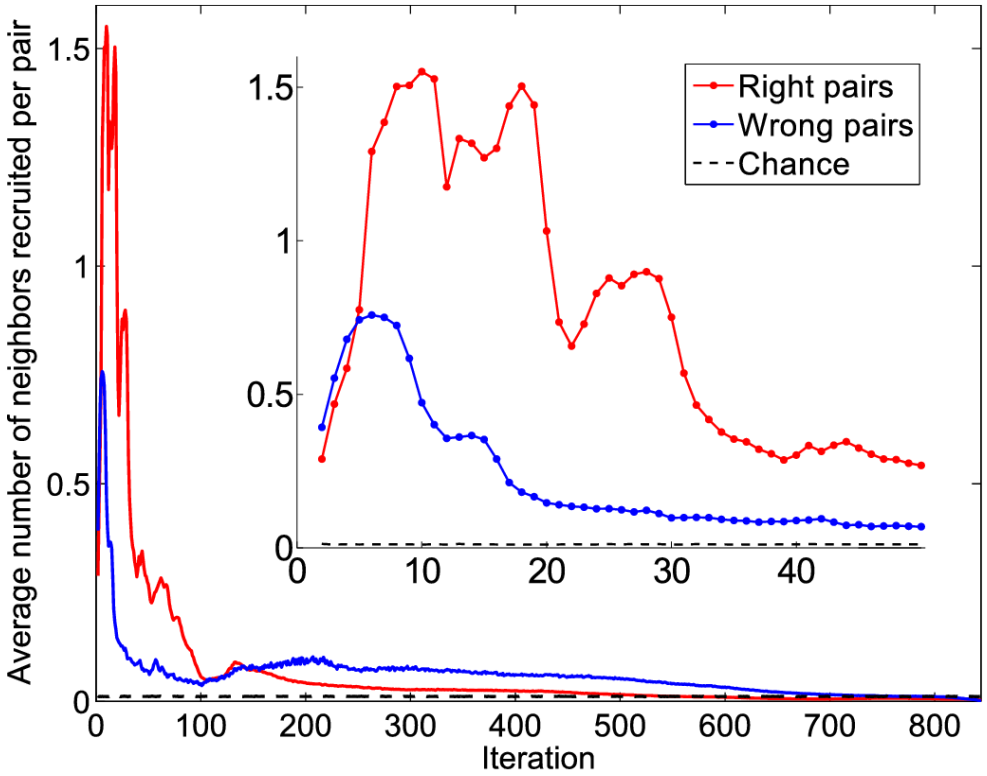
Training process

- **Evolution of the couplings and of the concatenated alignment**

HK-RR residue pairs with highest Frobenius norm vs. actual contacts Casino et al. (2009)



Impact of sequence similarity in recruitment into the concatenated alignment



- Initially, models are no better than chance, but they improve a lot upon iterating

- Initially, **sequence similarity** is crucial to recruitment into the concatenated alignment

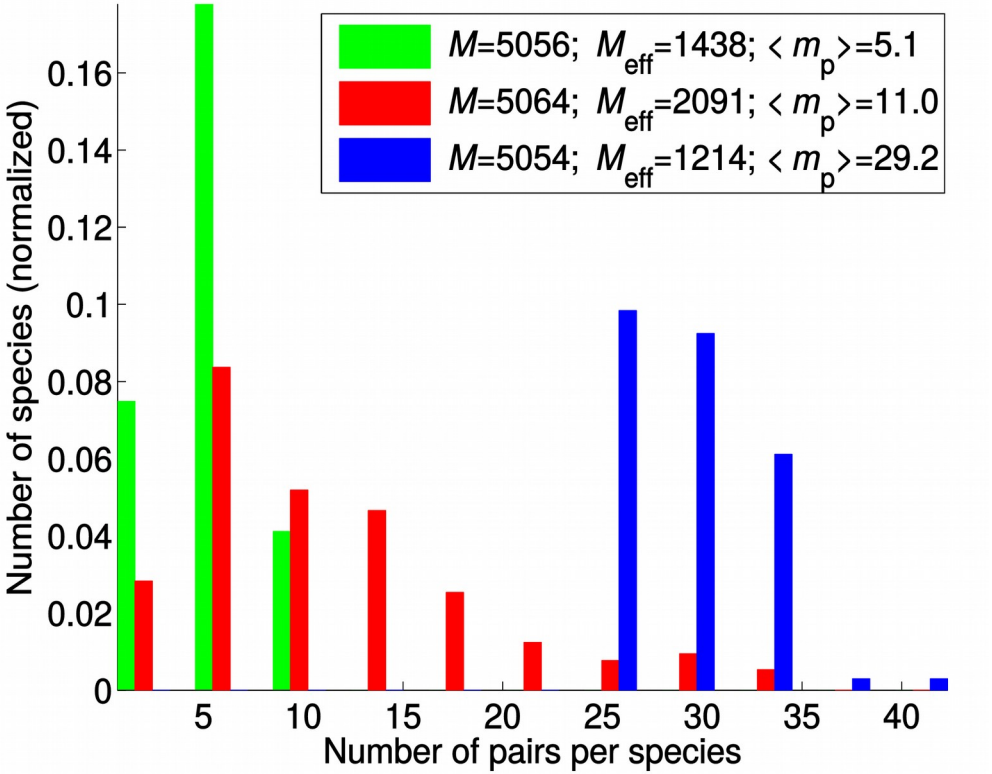
→ **Favors correct pairs, which have ~2x more neighbors**

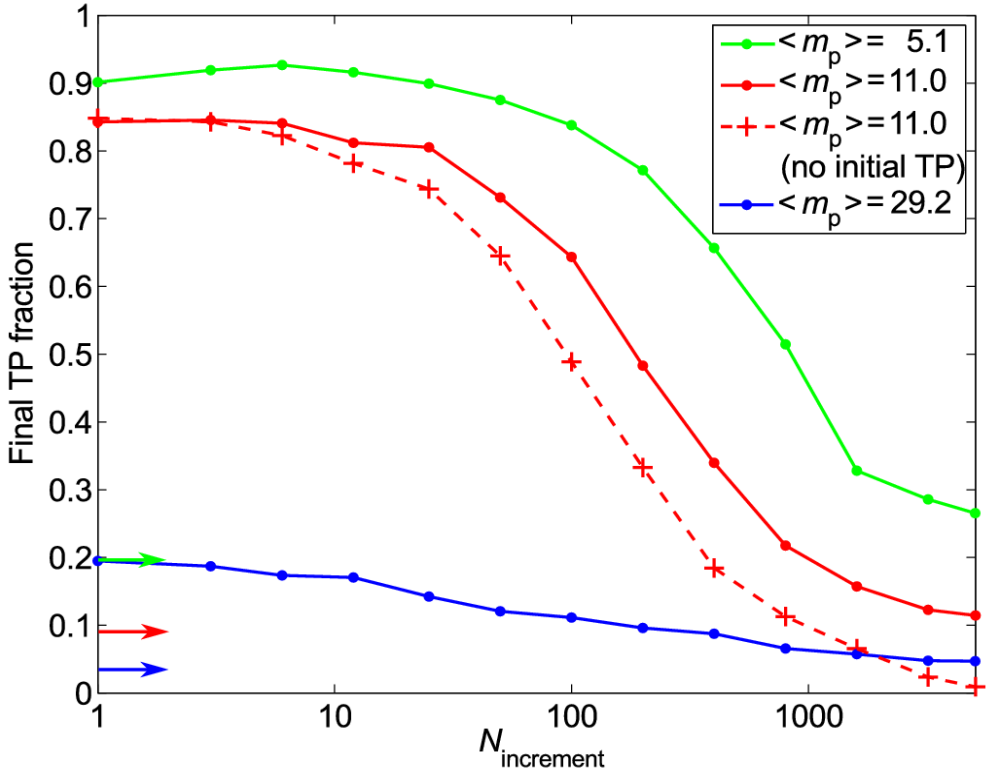
Impact of the number of pairs per species

- **Consider 3 datasets with the same number of sequences**

  - Standard (random) extract from the full dataset
  - Extracts with fewer / more pairs per species

→ Starting from random pairings:
final TP fraction vs. Nincrement





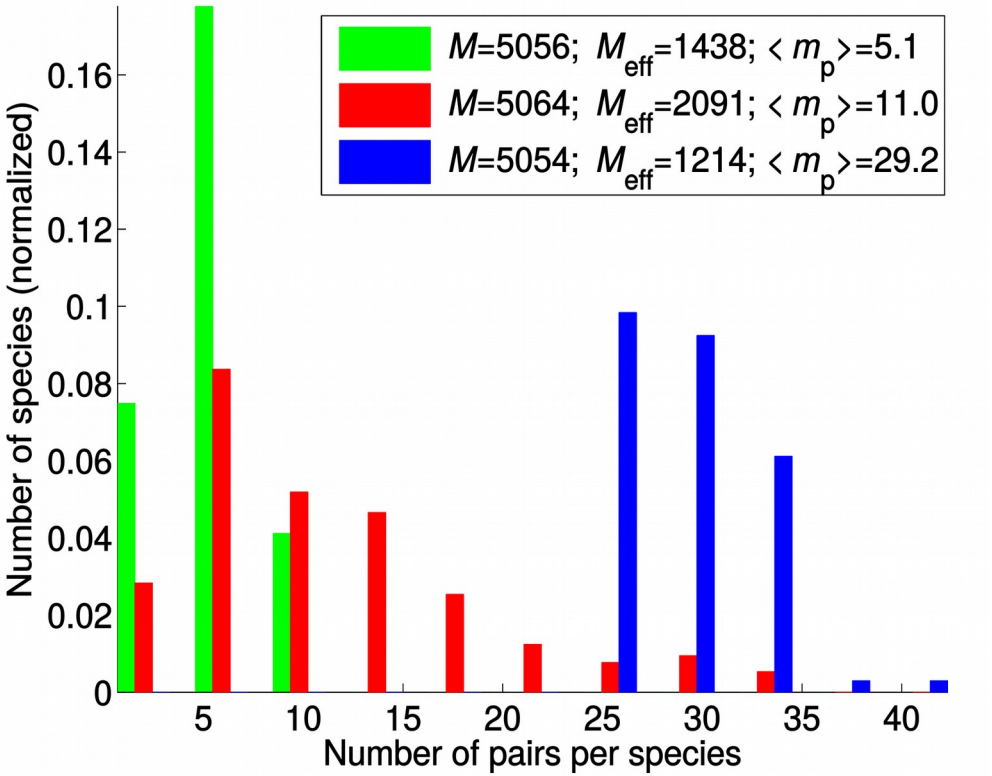
→ **Species with few pairs are important**

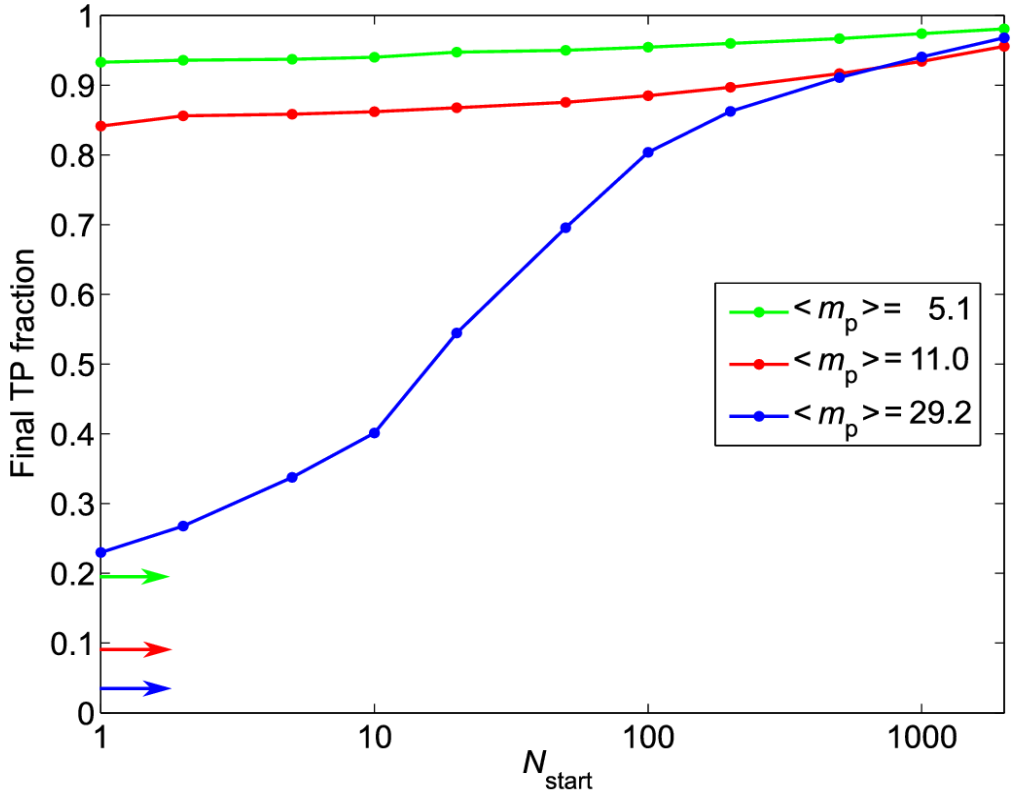
Impact of the number of pairs per species

- **Consider 3 datasets with the same number of sequences**

  - Standard (random) extract from the full dataset
  - Extracts with fewer / more pairs per species

→ with a training set: final TP fraction vs. Nstart (Nincrement=6)





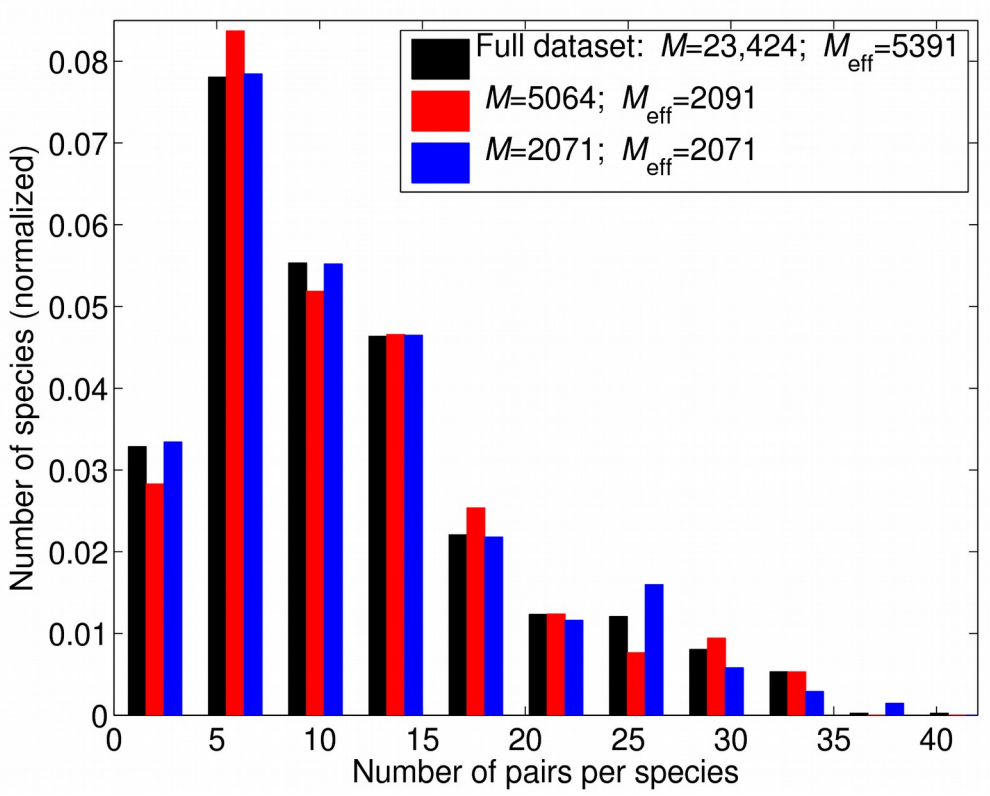
→ **Species with few pairs are important**

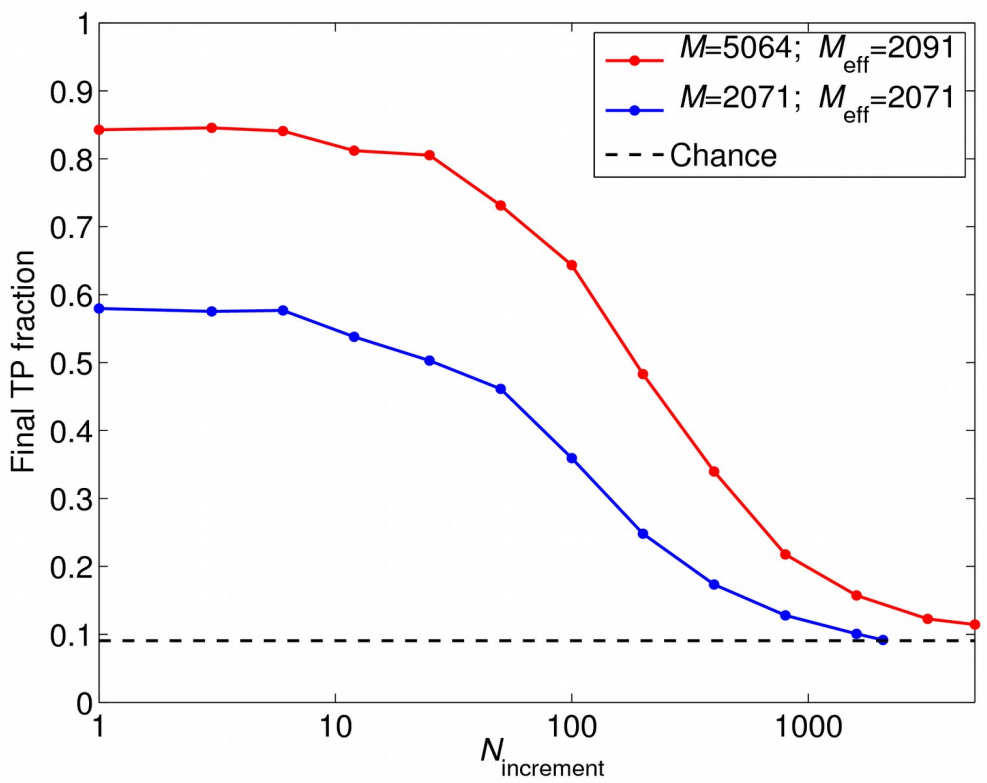
... but if there are none, a (sufficiently large) training set yields good final TP fractions

Impact of sequence similarity

- **Consider two datasets**

  - Standard (random) extract
  - Extract with **distant sequences** (Hamming distance >= 0.3); same numbers of pairs / species



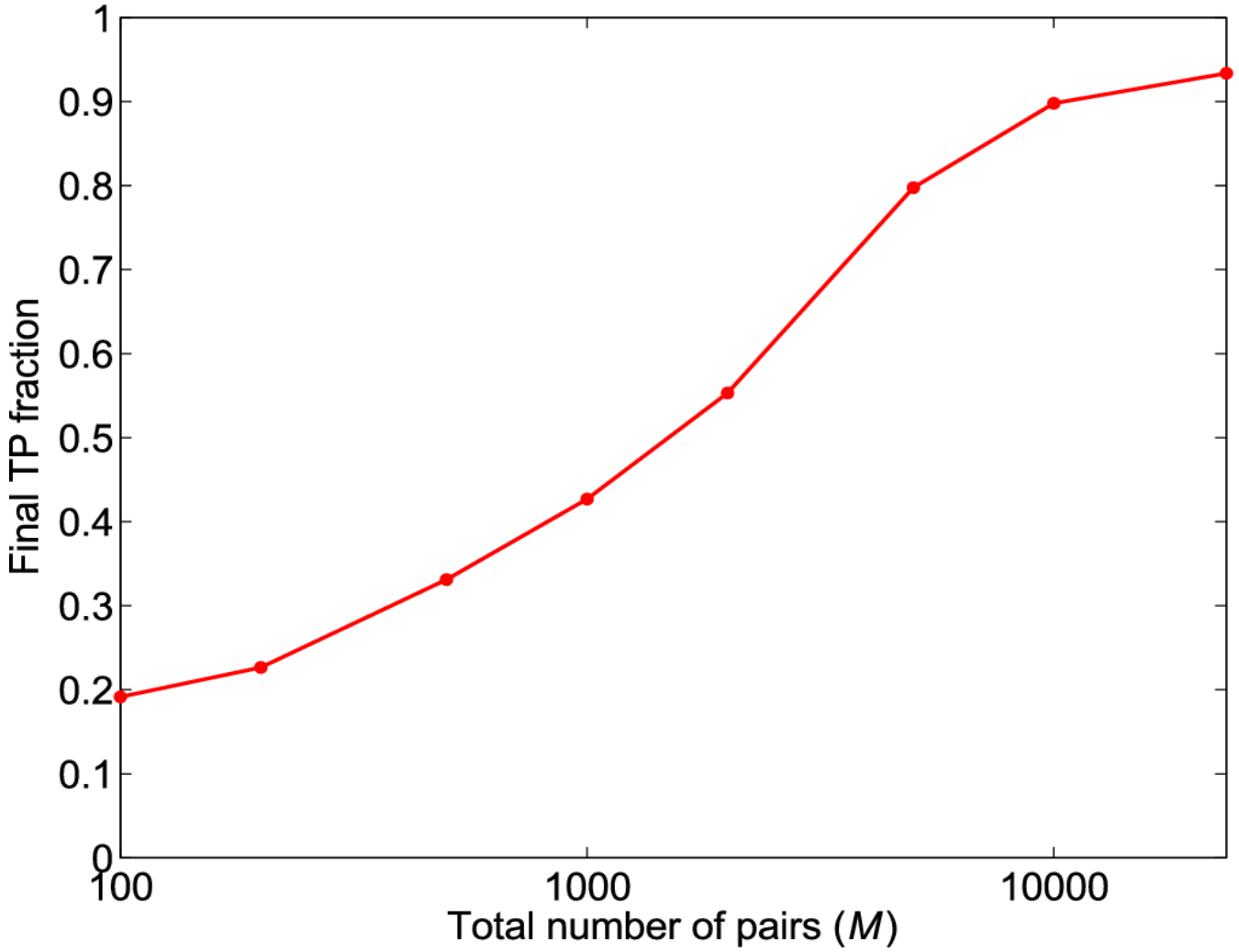


  - Sequence similarity does help
  - However, the TP fraction remains quite high

Impact of the dataset size

- **Starting from random pairings: final TP fraction vs. alignment size**

Different dataset sizes (from different numbers of picked species); **small Nincrement**
Results averaged over 50-500 replicates with different random pickings of species

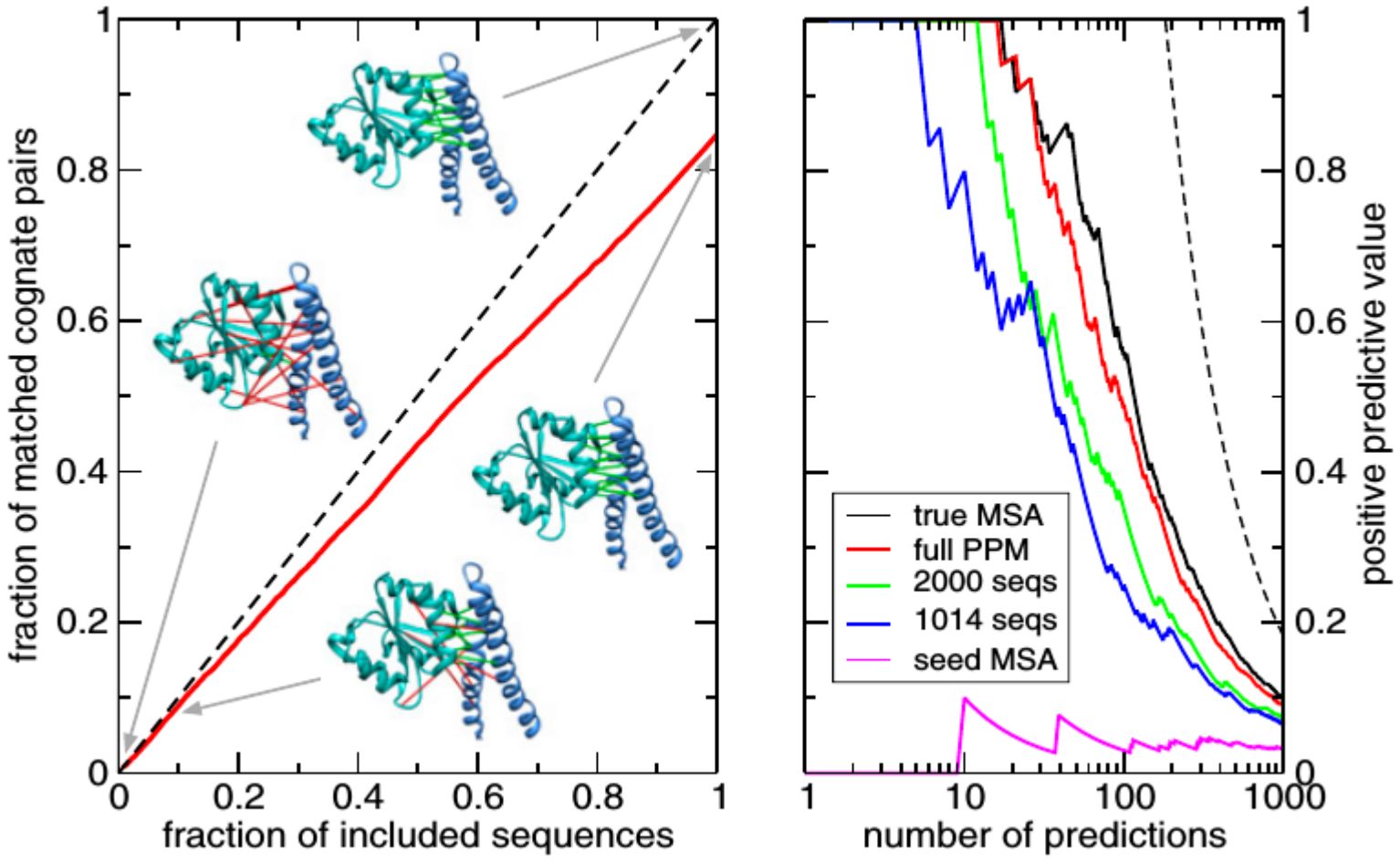


93% TP for the full dataset (23,424 pairs)

Simultaneous prediction of complex structure

- **Top inter-protein couplings = inter-protein contacts**

Gueudre et al. 2016 (published back-to-back with currently presented work)

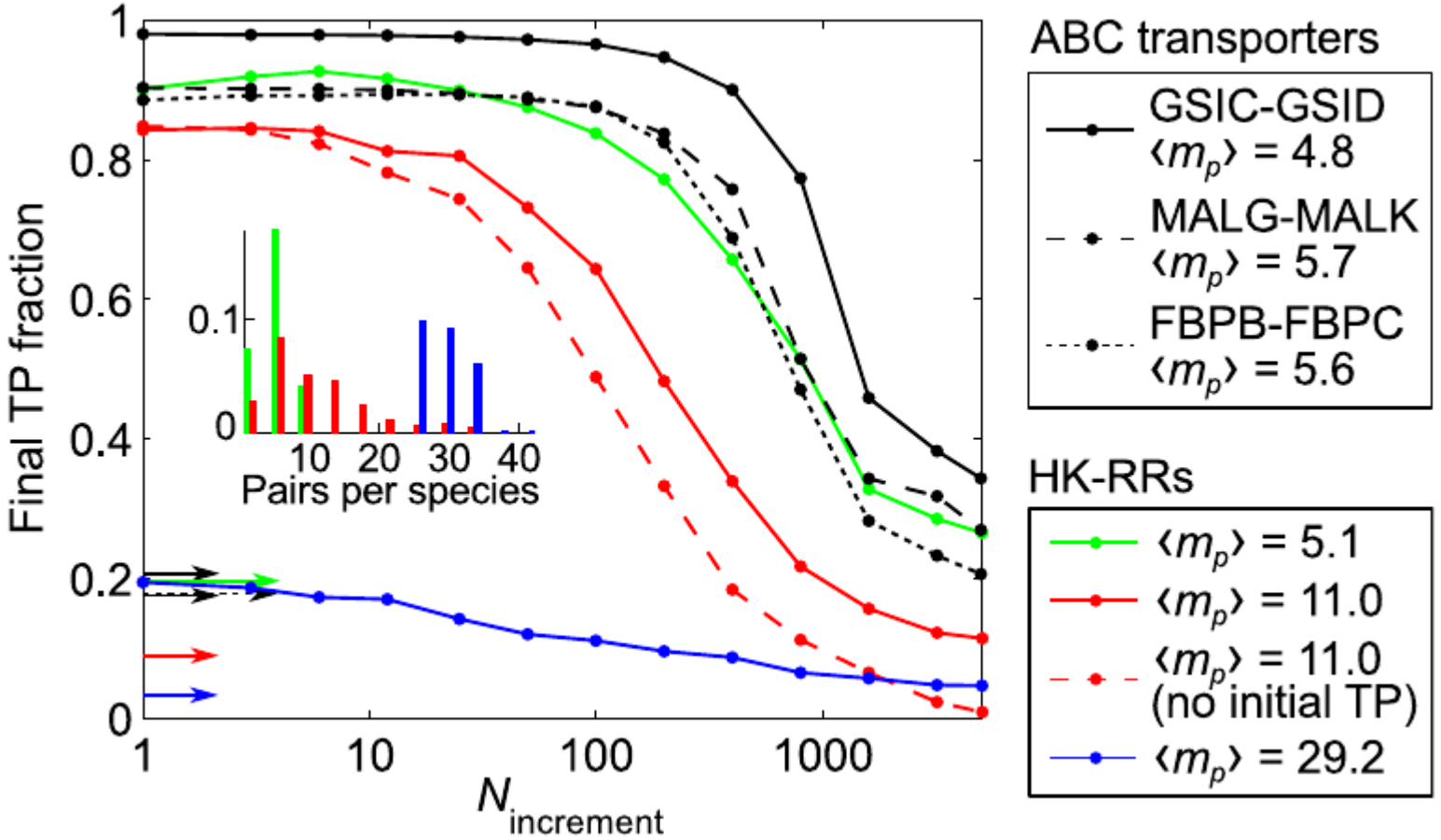


(1) Do protein families A and B interact or not?

(2) Within a species, which A interacts with which B?

Beyond HKs and RRs: ABC transporters

- **Very good performance in this case too**
  (starting from random pairings; 50 replicates)

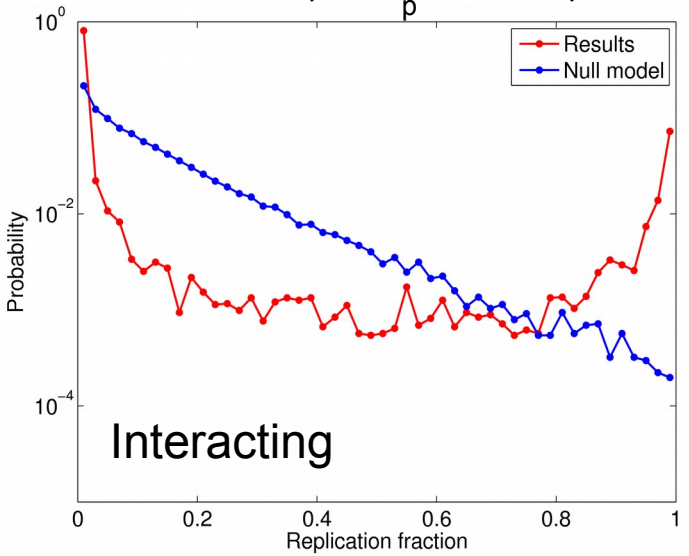


A very different biological case
→ The IPA should be widely applicable

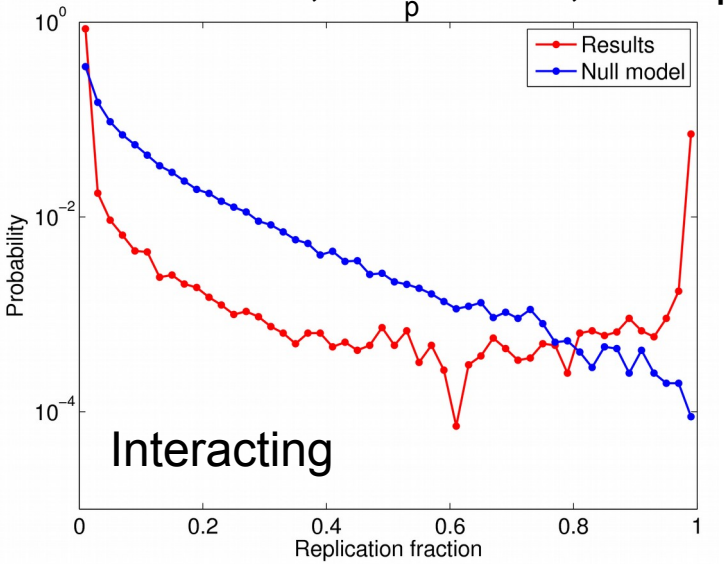
Do protein families A and B interact?

- **Exploiting different random initializations**
  → Distribution of the replication fraction - HK-RRs + ABC transporters

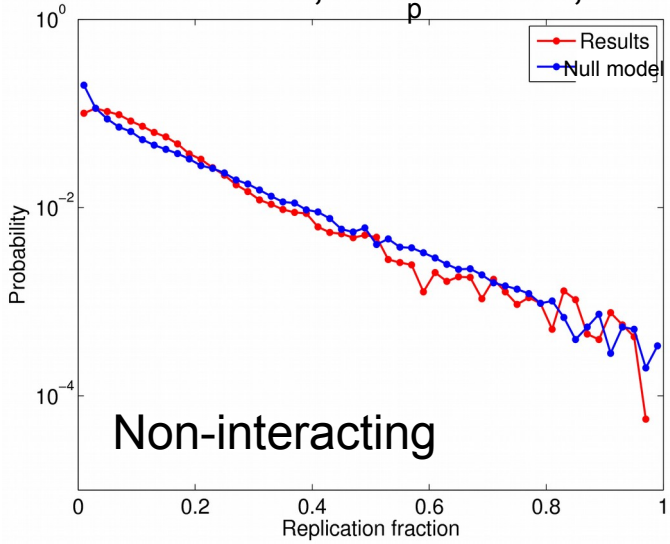
BASS-BASR, $\langle m_p \rangle$=5.50, 5012 pairs



MALG-MALK, $\langle m_p \rangle$=5.69, 5004 pairs



BASR-MALK, $\langle m_p \rangle$=4.95, 5001 pairs

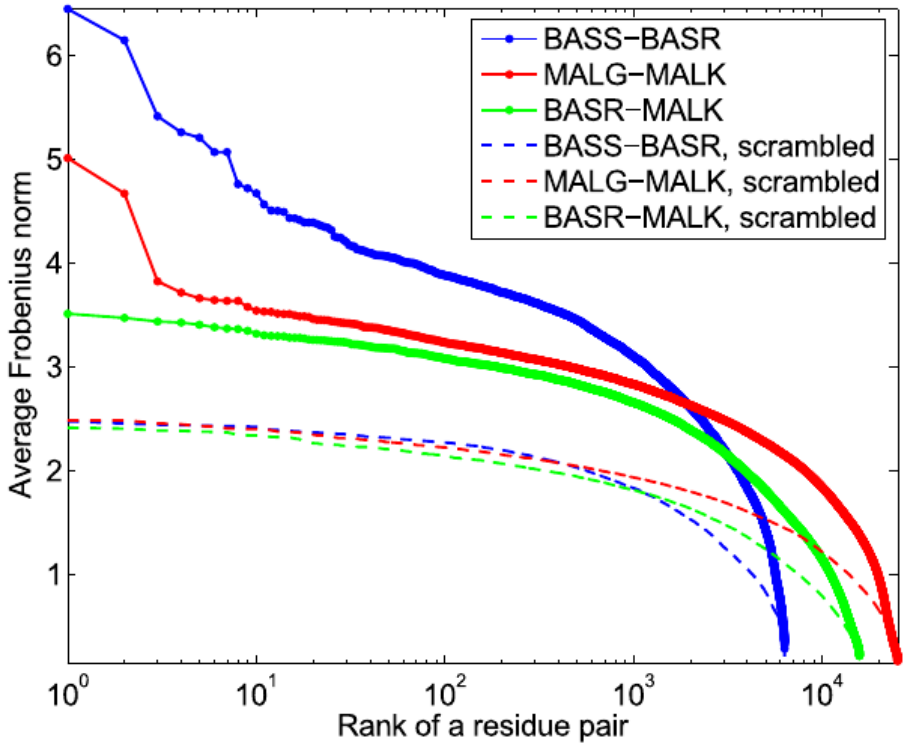


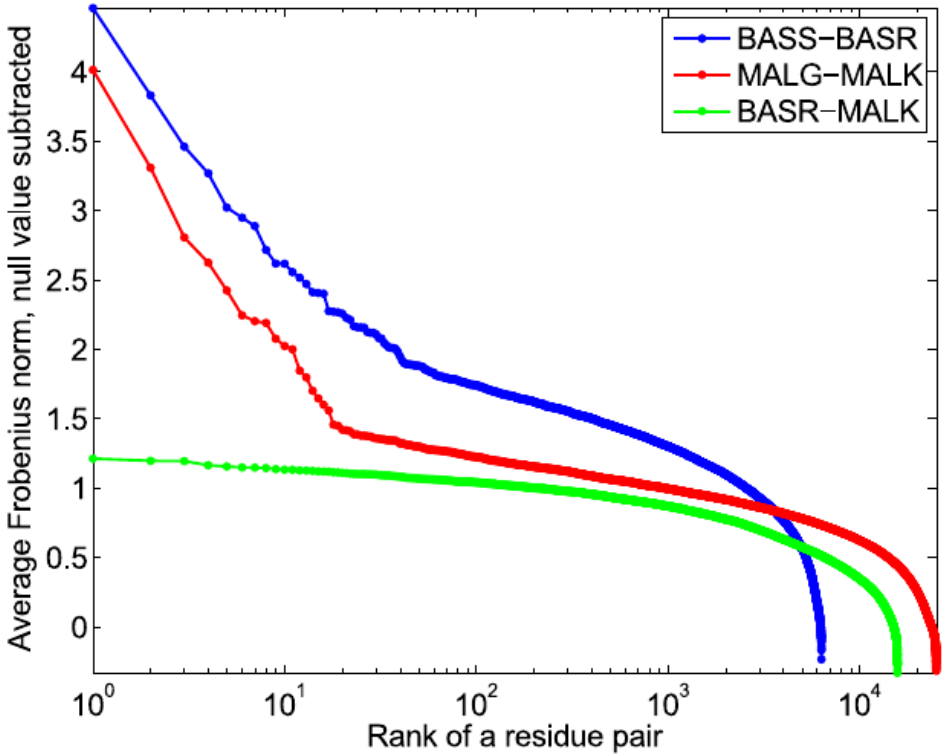
→ Bimodal distribution:
a signature of interactions

(Nincrement=50, 500 replicates)

Do protein families A and B interact?

- **Importance of the couplings for the same datasets**
  (Nincrement=50, 500 replicates)





→ The strongest couplings are outliers for interacting pairs, not for non-interacting ones

Conclusion

Summary

- Iterative method
- High performance even with **no initial training set**

Perspectives

- **Partnership prediction for orphan HK and RR**
  → current work with Mohamed Barakat, Philippe Ortet & Ned Wingreen
- Choosing among paralogs in other protein families
- Improving complex structure prediction
- **Prediction of novel protein-protein interactions**
  → current work with Yaakov Kleeorin & Ned Wingreen
- **Understand better how the algorithm "starts from nothing"**
  → current work with Pierre Mergny & Martin Weigt

Acknowledgements

Ned Wingreen, Princeton University
Lucy Colwell, Cambridge University
Rob Dwyer, Princeton University

Mohamed Barakat & Philippe Ortet, CEA Cadarache (P2CS)

Reference

Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS, PNAS 113 (43), 12180-12185 (2016)

Thanks!