Interprotein coevolution: bridging scales from residues to genomes

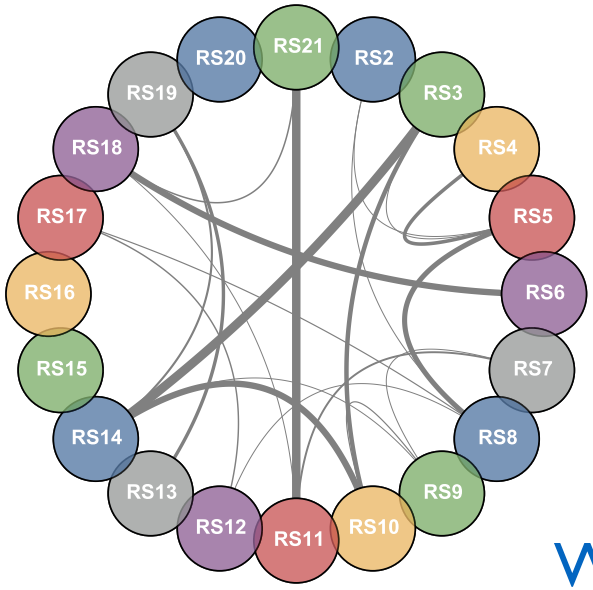
Martin Weigt

Laboratoire de Biologie Computationnelle et Quantitative
Université Pierre & Marie Curie Paris

Inria Paris

16 Nov 2017

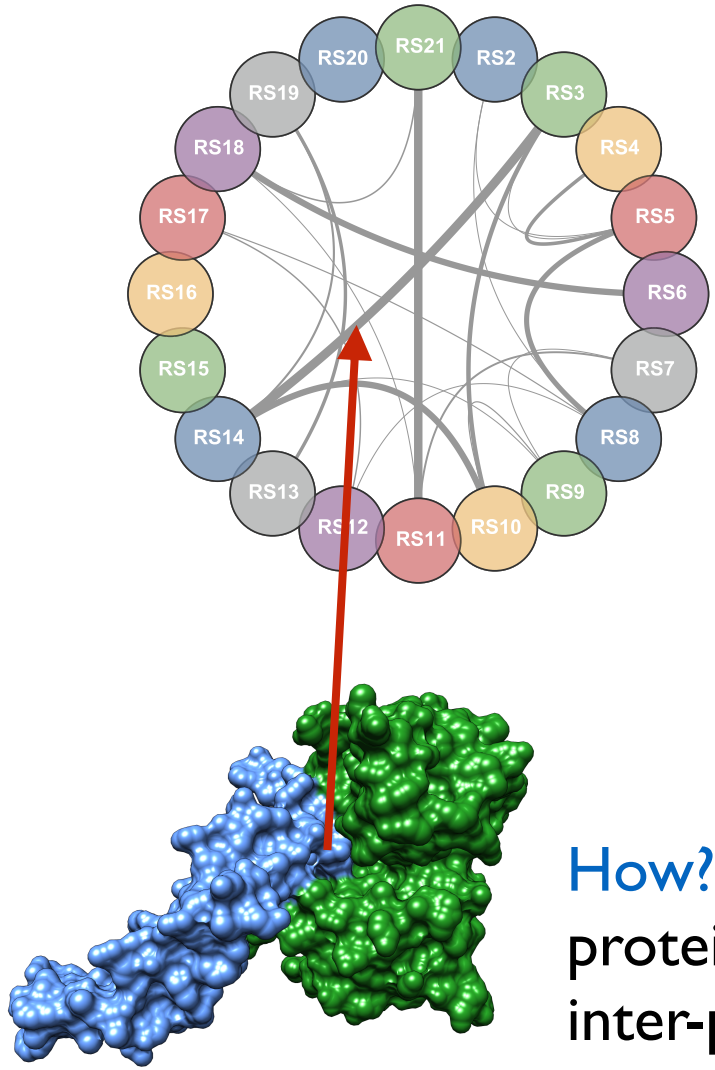
The different scales in protein-protein interaction



Who with whom?

protein-protein interaction networks

The different scales in protein-protein interaction

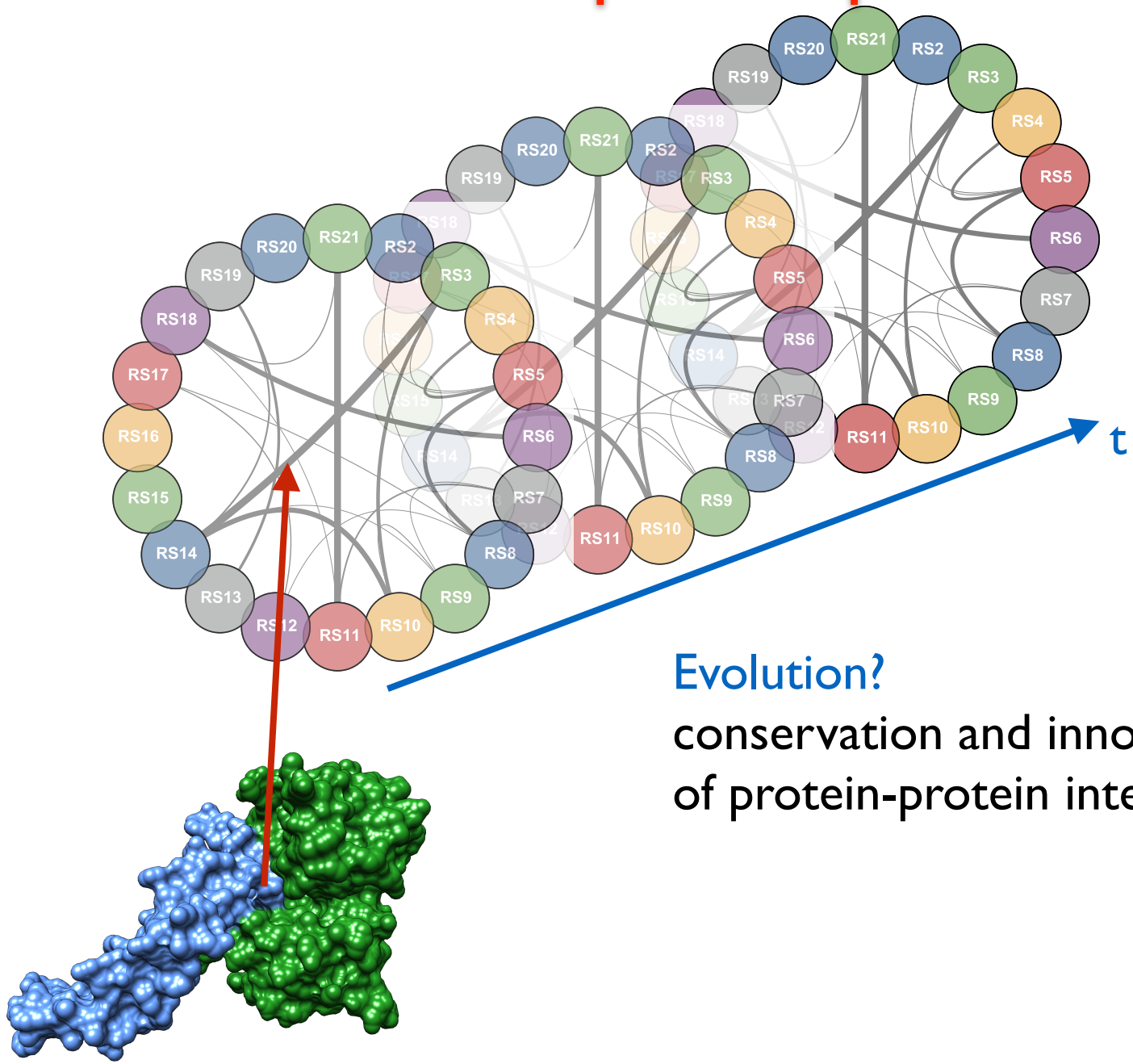


How?

protein-protein interfaces

inter-protein residue contacts

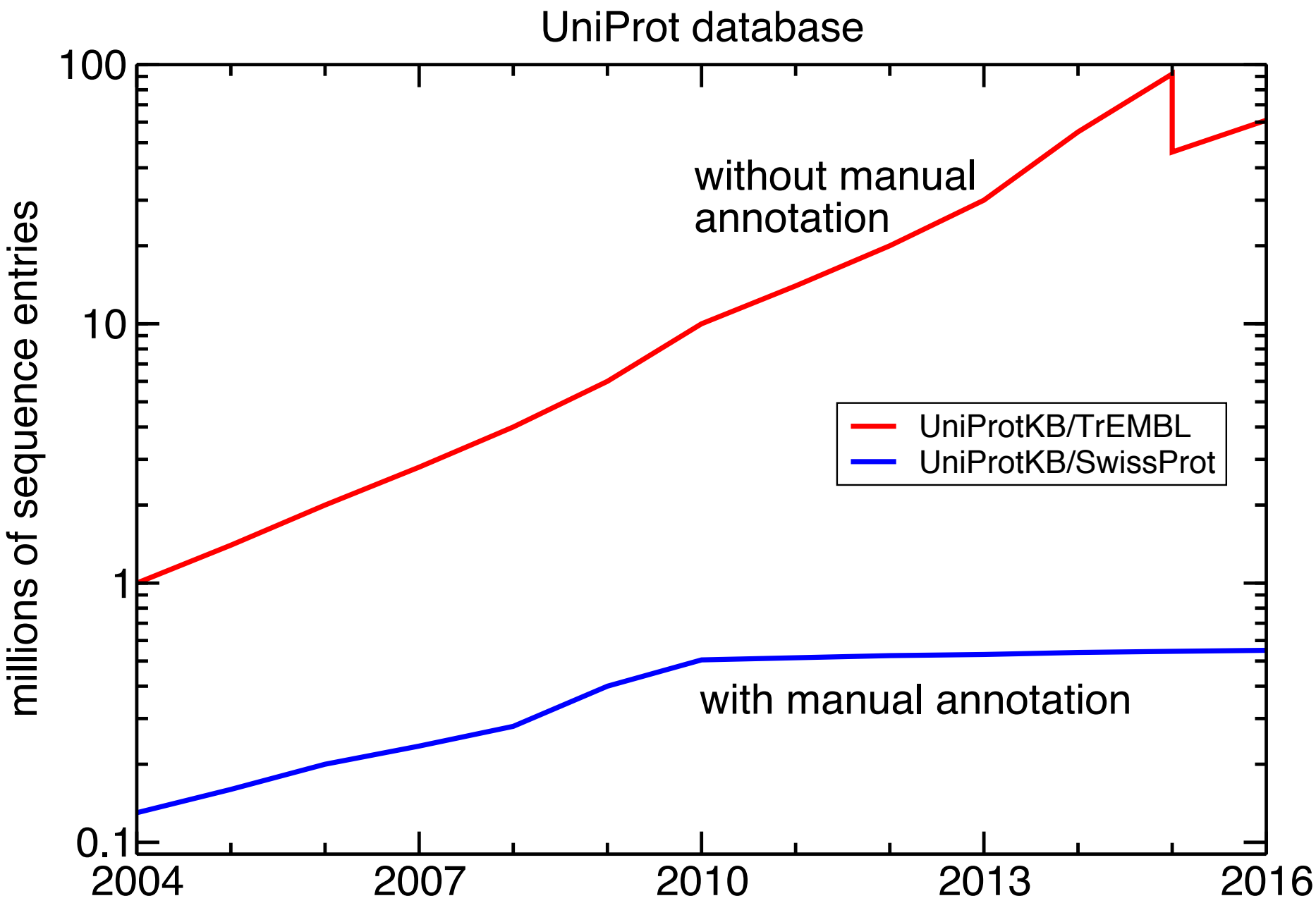
The different scales in protein-protein interaction



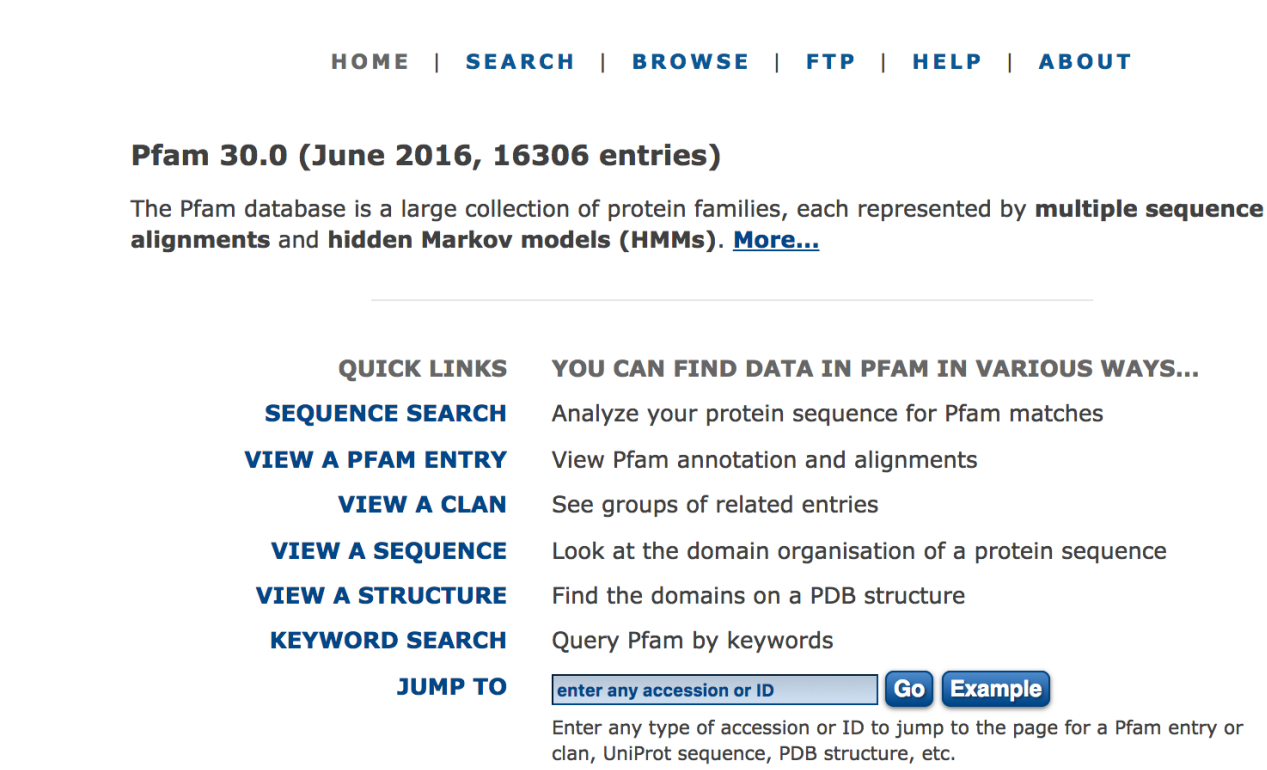
Evolution?

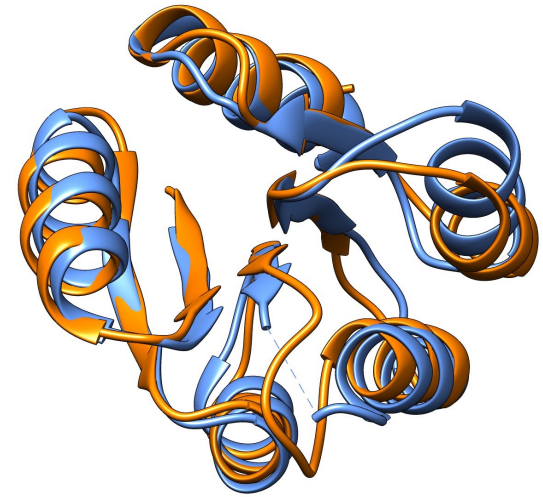
conservation and innovation
of protein-protein interactions

Protein sequence data are accumulating…



…and are classified into homologous protein families





Homologous proteins

- frequently 10^3–10^6 proteins per family
- common evolutionary ancestry
- conserved 3D structure and biological function
- diverged amino-acid sequences (~20-30% sequence identity)
- sequence variability contains information about structure and function
- >5000 families without example structures

Statistical physics

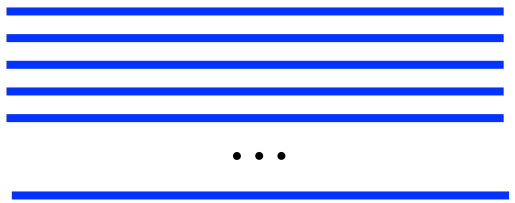
From models over data to thermodynamic observables:

$$P(\overline{S}) \sim e^{-\beta \mathcal{H}(\overline{S})}$$

$$\mathcal{H}(\overline{S}_1) = -\sum_{i<j} J_{ij} S_i S_j - \sum_i h_i S_i$$

sample from model

$$\{\overline{S}^\mu\}_{\mu=1,\ldots,M}$$



$$\langle \mathcal{O}_a(\overline{S}) \rangle_P \simeq \frac{1}{M} \sum_\mu \mathcal{O}_a(\overline{S}^\mu)$$

e.g. $\langle S_i \rangle_P, \langle S_i S_j \rangle_P$

Inverse statistical physics

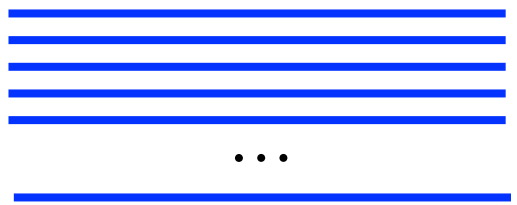
From data over observables to models

$$P(\overline{S}) \sim e^{-\beta \mathcal{H}(\overline{S})}$$

$$\mathcal{H}(\overline{S}_1) = -\sum_{i<j} J_{ij} S_i S_j - \sum_i h_i S_i$$

Data:

$$\{\overline{S}^\mu\}_{\mu=1,\ldots,M}$$



$$\langle \mathcal{O}_a(\overline{S}) \rangle_P \simeq \frac{1}{M} \sum_\mu \mathcal{O}_a(\overline{S}^\mu)$$

e.g. $\langle S_i \rangle_P, \langle S_i S_j \rangle_P$

Inverse statistical physics

How to construct $P(\overline{S}) \sim e^{-\beta \mathcal{H}(\overline{S})}$ from data?

- coherence with data

$$\langle \mathcal{O}_a(\overline{S}) \rangle_P = \frac{1}{M} \sum_{\mu} \mathcal{O}_a(\overline{S}^{\mu})$$

- maximum entropy principle (least constrained model)

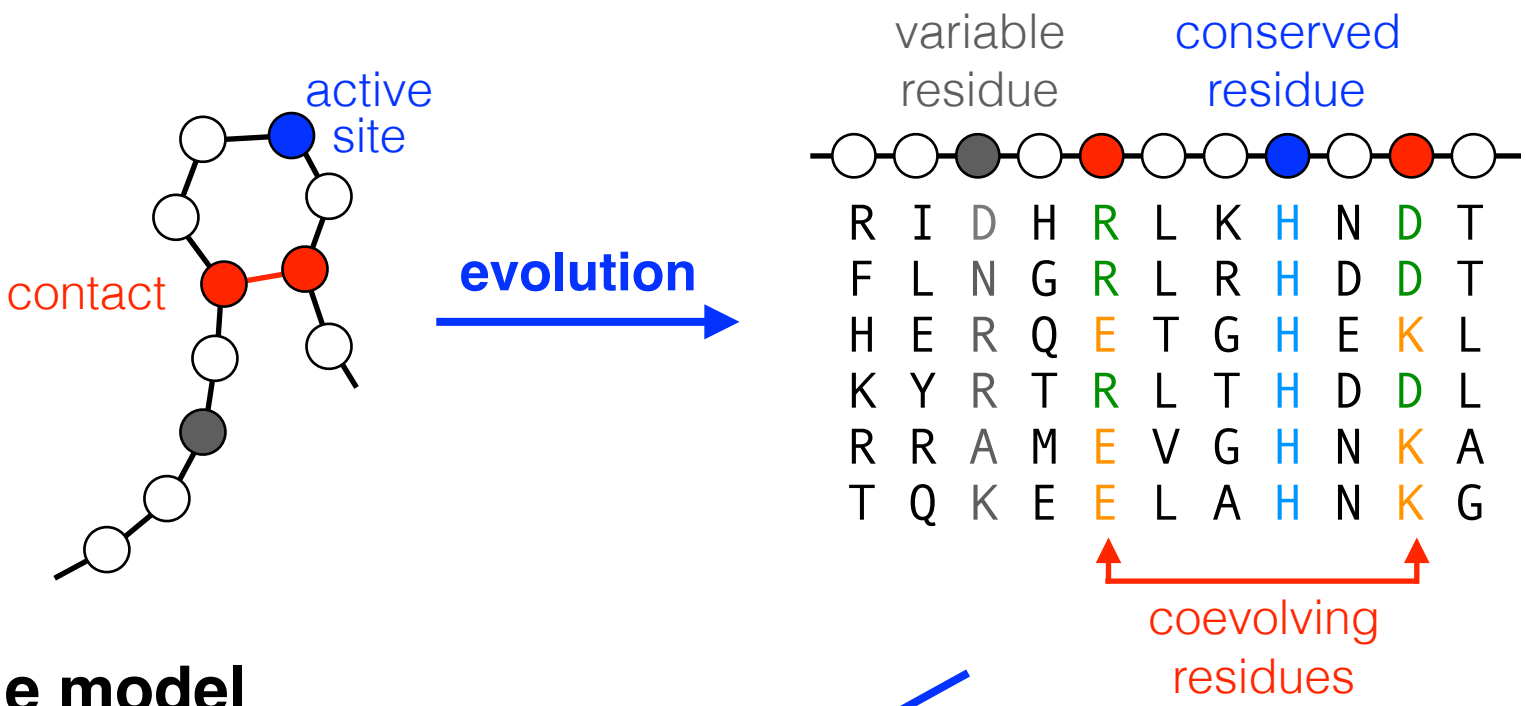
$$-\sum_{\overline{S}} P(\overline{S}) \log P(\overline{S}) \rightarrow \max$$

➡ analytical form of model

$$\mathcal{H}(\overline{S}) = -\sum_{a} \lambda_a(\overline{S}) \mathcal{O}_a(\overline{S})$$

selection of observables
requires priori biological knowledge

Conservation and coevolution in proteins



Profile model

$$P(a_1, ..., a_L) \sim \exp\left\{\sum_i h_i(a_i)\right\}$$

statistical modeling

Direct Coupling Analysis (DCA)

$$P(a_1, ..., a_L)$$

$$\sim \exp\left\{\sum_{i<j} J_{ij}(a_i, a_j) + \sum_i h_i(a_i)\right\}$$

strong couplings -> residue contacts

[Weigt et al, PNAS '09]

[Morcos et al, PNAS '11]

Interactions between protein families

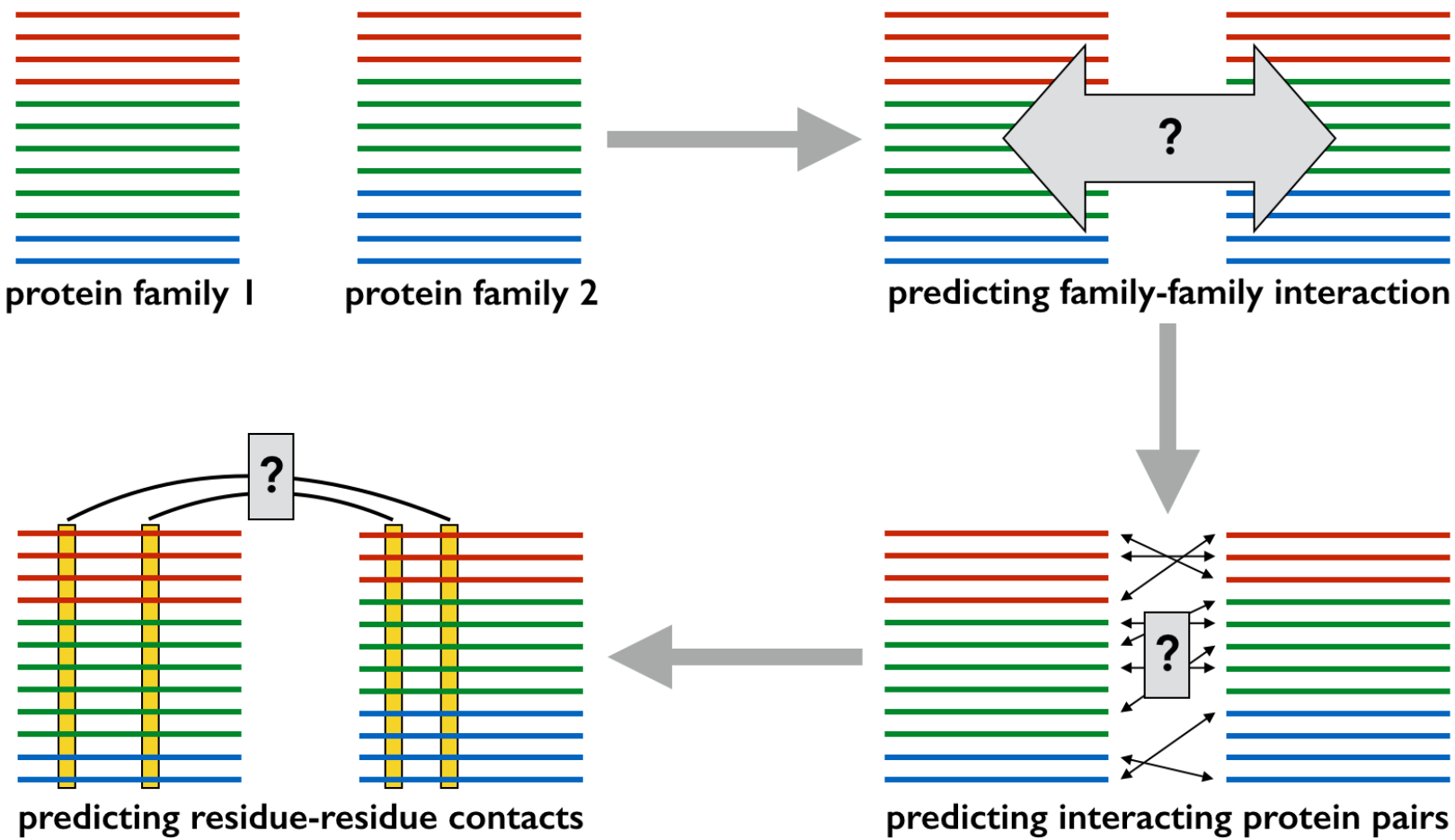
Family 1

```
>F7XUK6_MIDMI/129-211
LAQQLEKRISFRKAAKRLIQNAM.R......M.G..AEGIKIKISGRIG.G.AEIARDQQ
YNEGRVPL..HTLRMMIDYGTAEAH..TTYGRIGVKVWV
>B3SEY6_TRIAD/119-201
VAEQLEKKVSFRKAVKRAISNAM.K......M.G..AKGIKISVSGRLG.G.AEIARTEW
YKEGRVPL..HTLRAIVKYDMAEAH..TIYGLIGVKVWV
>RS3_ORITB/122-204
IAQQLERRQSFKKVMKKAIHASM.K......Q.G..AKGIKIICSGRLG.G.VEIARSES
YKEGRVPL..QTIRADIRYAFAEAI..TTYGVIGVKVWV
>RS3_RICPR/123-205
IAAQLEKRVSFRKAMKTAIQASF.K......Q.G..GQGIRVSCSGRLG.G.AEIARTEW
YIEGRMPL..HTLRADIDYSTAEAI..TTYGVIGVKVWI
>E1X0L6_HALMS/119-201
IASQLEKRVAFRRAMKKVMQSAF.R......A.G..VKGIRVRTAGRLG.G.AEMARAEG
YSERKVPL..HTLRADIDYSTAEAH..TTYGVIGVKVWV
>I7HEJ8_9HELI/120-202
IATQLEKRVAFRRAMKKVMQAAM.K......A.G..AKGIKVKVSGRLA.G.AEMARTEW
YMEGRVPL..HTLRAKIDYGFAEAM..TTYGIIGVKVWI
>M4VDL1_9DELT/120-202
IAMQLEKRISWRRALKKAIAAAT.K......G.G..VRGIKVRVSGRLD.G.AEIARSEW
YNEKSVPL..HTLRADIDYGTAEAL..TAYGIIGMKVWI
>RS3_HYPNA/120-202
IARQLERRASFRRAMKRSIQSAM.R......L.G..AEGVKVVVSGRLG.G.AEIARTEK
YAEGSVPL..HTLRADIDYGTAEAT..TTYGIIGVKVWV
>C0QW02_BRAHW/94-176
VARQLEMRVAFRRAMKSVITQAM.K......K.G..AKGIKVMCSGRLA.G.ADIARTEQ
YKNGSVPL..HTLRANIDYGTAEAL..TTFGIIGIKVWI
>J9Z1W5_9PROT/119-201
IARQLEKRVAFRKAMKKSGQSAI.K......L.G..AKGIKIVCGGRLG.G.AEIARSEK
FSEGSVPL..HTLRADIDYATARAL..TTYGIIGIKVWL
>RS3_MARMM/120-202
IAQQLERRVAFRRAMKRSMQSAM.R......M.G..AKGCKIVCGGRLG.G.AEIARTEQ
YNEGSVPL..HTLRADIDYGTCEAK..TAMGIIGIKVWI
>G0GFA5_SPITZ/122-204
IAGQLEHRASFRRVMKLAVANAM.K......A.G..VQGIKVRVSGRLG.G.AEIARSEV
QMAGRVPL..HTLRADIDYGFAEAR..TTYGVIGVKVWI
>V6DFZ5_9DELT/122-204
ISEQLEKRGSFKKAMKRAALDVM.K.......SG..AKGVKIRCAGRLG.G.AEIARDEW
IRVGSTPL..HTLRSDIDYGFVEAH..TTYGVIGIKVWI
>RS3_NEOSM/120-203
IAFQLEKRSSFRRVIKKAIATVM.R......ESD..VKGVKVACSGRLS.G.AEIARTEV
FKEGSIPL..HTMRADIDYWVAEAH..TTYGVIGVKVWI
>I0III3_PHYMF/124-207
IAEQLAKRASFRRVMKMKAEAAM.N......CGV..CKGVKIMLSGRLG.G.HEMSRSEV
VSLGSIPL..ATLQANVDYGFAISK..TTYGTIGVKVWI
>F0SJ92_RUBBR/120-202
IAQQLGKRGSFRRALKRSMEQVM.D......A.G..AHGVKIELSGRLG.G.AEMSRKEK
GSRGSIPL..STLQRHVDYGYTTAR..TAQGIIGIKVWI
. . .
```

Family 2

```
>RS14_NEOSM/47-100
KLNSLPRNSSPARSKNRCSITGR..PRGYY..RKFGI..SRIQLRVLANWGKLPGVVKSS
>I0AI30_IGNAJ/35-88
ALQKLPRNSSVTRLKNRCMFTGR..ARAYY..RKFGV..SRLVLREMALRGEIPGLKKSS
>I6YSF0_MELRP/36-88
.LQLLPRNSAPTRAHNRCLISGR..PRGYY..RKFGI..SRLVLREMALRGEIPGLKKSS
>I0IIH6_PHYMF/34-87
ALSQLPRDASPTRLVTQCAITGR..TRAVY..RKFNV..SRIVLRELALQGKIPGMKKAS
>RS14_CHLT3/35-88
ALRKLPRDSSPTRLKNRCSITGR..AKGVY..KKFGL..CRHILRKYALEGKIPGMKKAS
>RS14_PROA2/35-88
ALSKLPRNSSATRVRNRCVLTGR..GRGVY..EKFGL..CRHMFRKLALEGKIPGVKKAS
>D6XYV1_BACIE/35-88
ALSKLPRDSAPSRLTRRCKATGR..PRGVL..RKFEL..SRIKFRELAHKGQIPGVRKAS
>I0JIY2_HALH3/35-88
ALRKLPRDSSPTRVKRRCELSGR..PRGYM..RKFDM..SRIAFRELAHKGQIPGVKKAS
>RS14_EXIS2/36-88
.LSKLPRNSSAVRLHNRCSITGR..PHGYI..GKFGI..SRIKFRDLAHKGQIPGVKKAS
>RS14_STRR6/36-88
.LSKLPRNASPTRLHNRCRVTGR..PHSVY..RKFGL..SRIAFRELAHKGQIPGVTKAS
>G0VNI1_MEGEL/35-88
ALSQLPANASPVRLHNRCKVTGR..PHGYM..RKFGI..CRITFRELAYKGQIPGVKKAS
>R7PS46_9FIRM/35-88
ALSKLPRNASPTRLHNRCKLTGR..PHGYL..RKFGV..CRNQFRELAYRGEIPGVRKAS
>F8L373_SIMNZ/47-100
KLNSLPKNSSPIRRRNRCKMTGR..CRGYL..RKFQI..SRLCFREMANDGSIPGVVKAS
>F8L0V7_PARAV/47-100
ALNKMPRDSSPIRLRNRCQLTGR..XRGYL..RKFKL..SRLTFREMALAGLLPGVTKSS
>D6YVK9_WADCW/47-100
QLNKMRRDTSPVRLRNRCQITGR..CRGYL..SKFKV..SRLVFREMASIGMIPGVTKSS
>L7VJR0_9FLAO/35-88
ALQKLPKNSCTVRLRNRCKLTGR..SRGYM..RKFGV..SRISFRNLVNFGLIPGVKKSS
>C7NDL0_LEPBD/41-94
ELSKLPRNASPTRVRNRCQINGR..PRGYM..REFGI..SRVMFRQLAGEGVIPGVKKSS
>RS14_FUSNN/41-94
ELNKLPKDSSAVRKRNRCQLDGR..PRGYM..REFGI..SRVKFRQLAGAGVIPGVKKSS
>K0P015_9BACT/35-88
ALDKLPKNSSPVRLRNRCNITGR..ARGYI..RRFGI..SRLVFRKWALEGKLPGIRKAS
>RS14_AMOA5/35-88
ALDKLPKNASPVRVRNRCKITGR..ARGYM..RKFGI..SRIVFREWAAQGKIPGVIKAS
>I4ALV0_FLELS/42-94
.LDKLPKDSSPVRLHNRCRLTGR..PRGYM..RRFGI..CRVVFREMANDGKIPGVTKSS
>RS14_SALRD/35-88
ELQKLPRDSSPVRQNNRCELCGR..QRGYL..RKFGV..CRICFRELALEGKIPGIRKAS
>C7PU84_CHIPD/35-88
ELDQLPRNASPVRLHNRCQLSGR..PKGYM..RHFGM..CRNMFRDLALAGKIPGVRKAS
>F4KWV6_HALH1/35-88
ELDKLPRNSNPIRMHNRCQLTGR..PKGYM..RQFGL..CRVKFREMALYGKIPGITKSS
. . .
```

Interactions between protein families



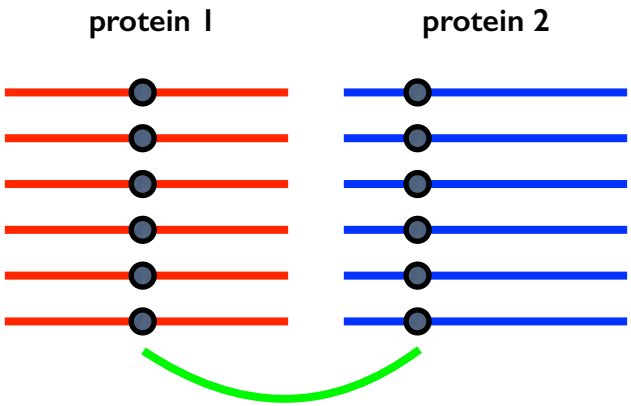
What can we learn from the empirical sequence variability:

- do the families interact?
- which specific proteins interact?
- which residues are in contact?

➡ relation between protein structure/function and evolution

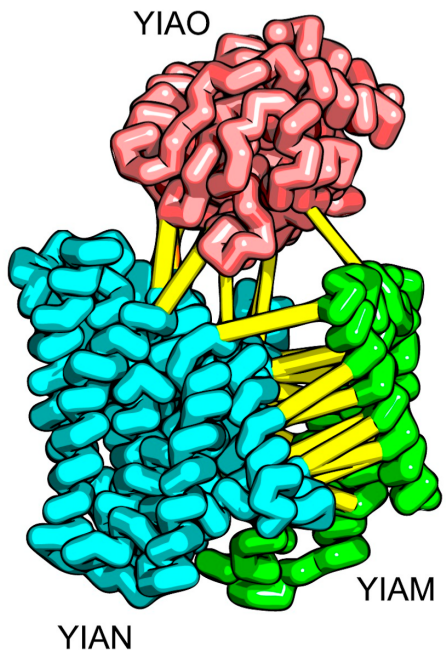
Prediction of inter-protein residue contacts

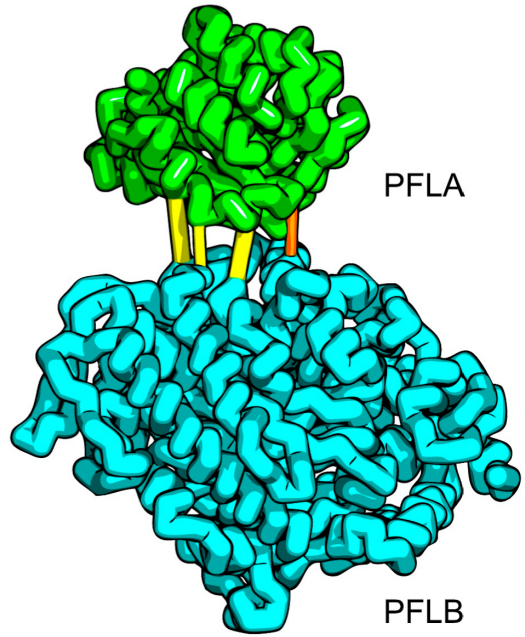
joint MSA of protein families



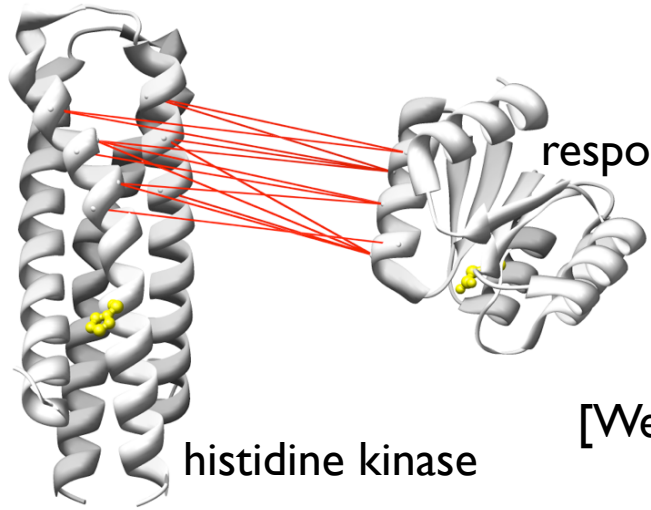
DCA

Strong inter-protein couplings predict contacts



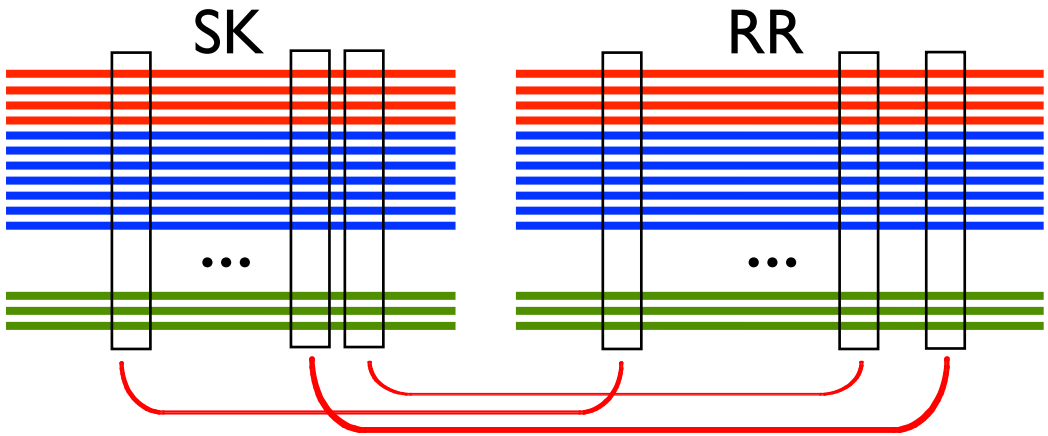


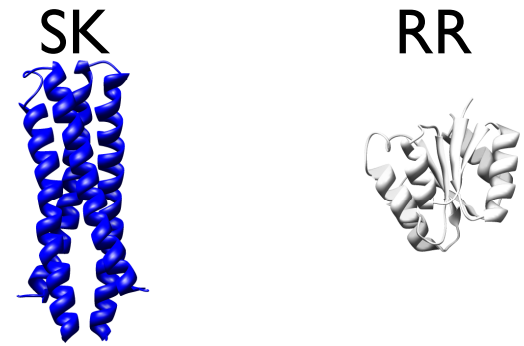
[Ovchinnikov et al., *eLife* '14]



[Weigt et al., PNAS '09]

In silico prediction of high-resolution structures of transient protein complexes

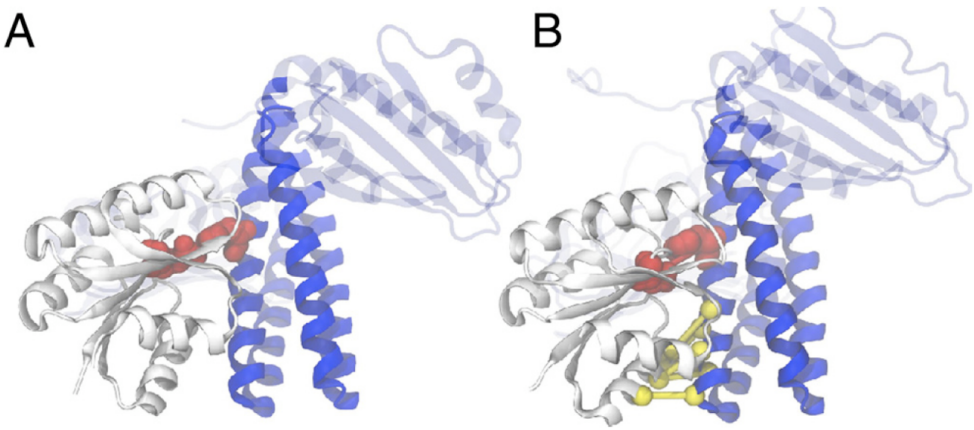




DCA identifies residue contacts

protein monomer structures

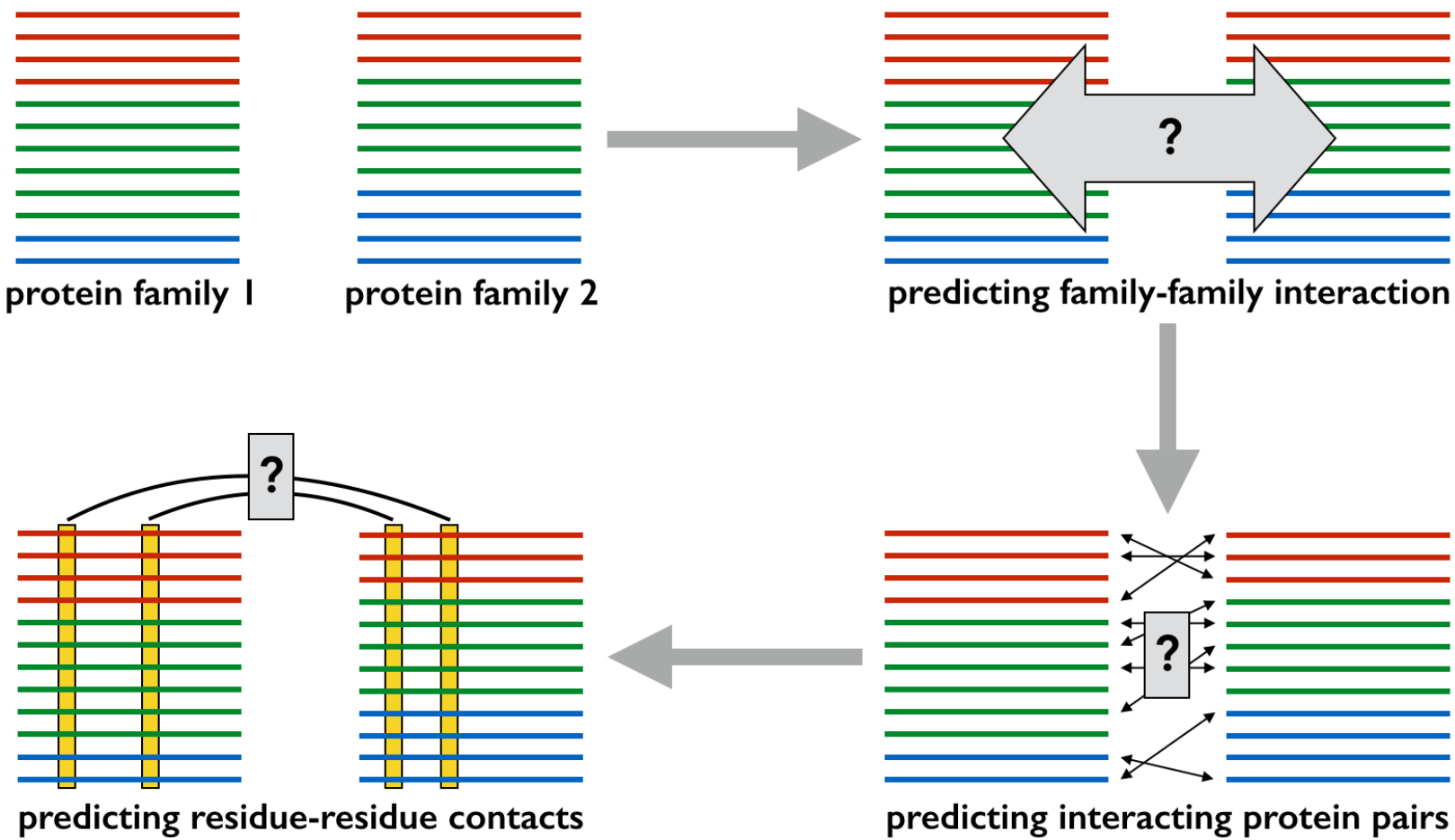
guided molecular dynamics simulations →



Spo0B/0F: co-crystal [Zapf *et al.* (2000)] vs. our model

[Schug, MW, Onuchic, Hwa, Szurmant, PNAS '09]

Interactions between protein families

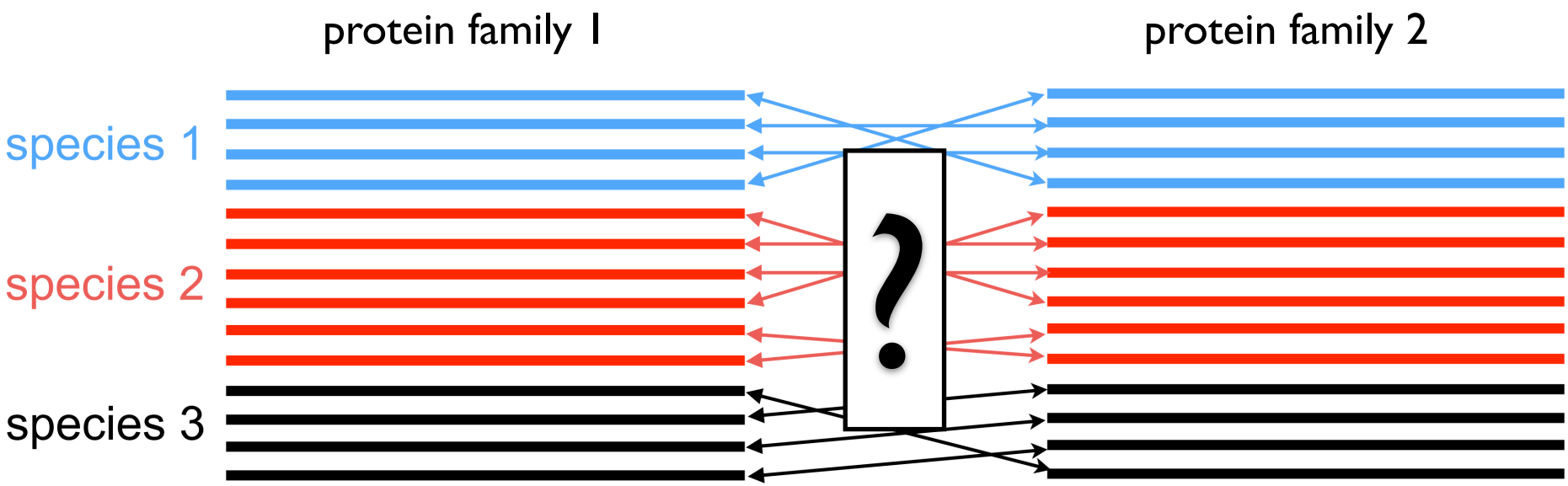


What can we learn from the empirical sequence variability:

- do the families interact?
- which specific proteins interact?
- which residues are in contact?

➡ relation between protein structure/function and evolution

Specific interactions and paralog matching



General idea:

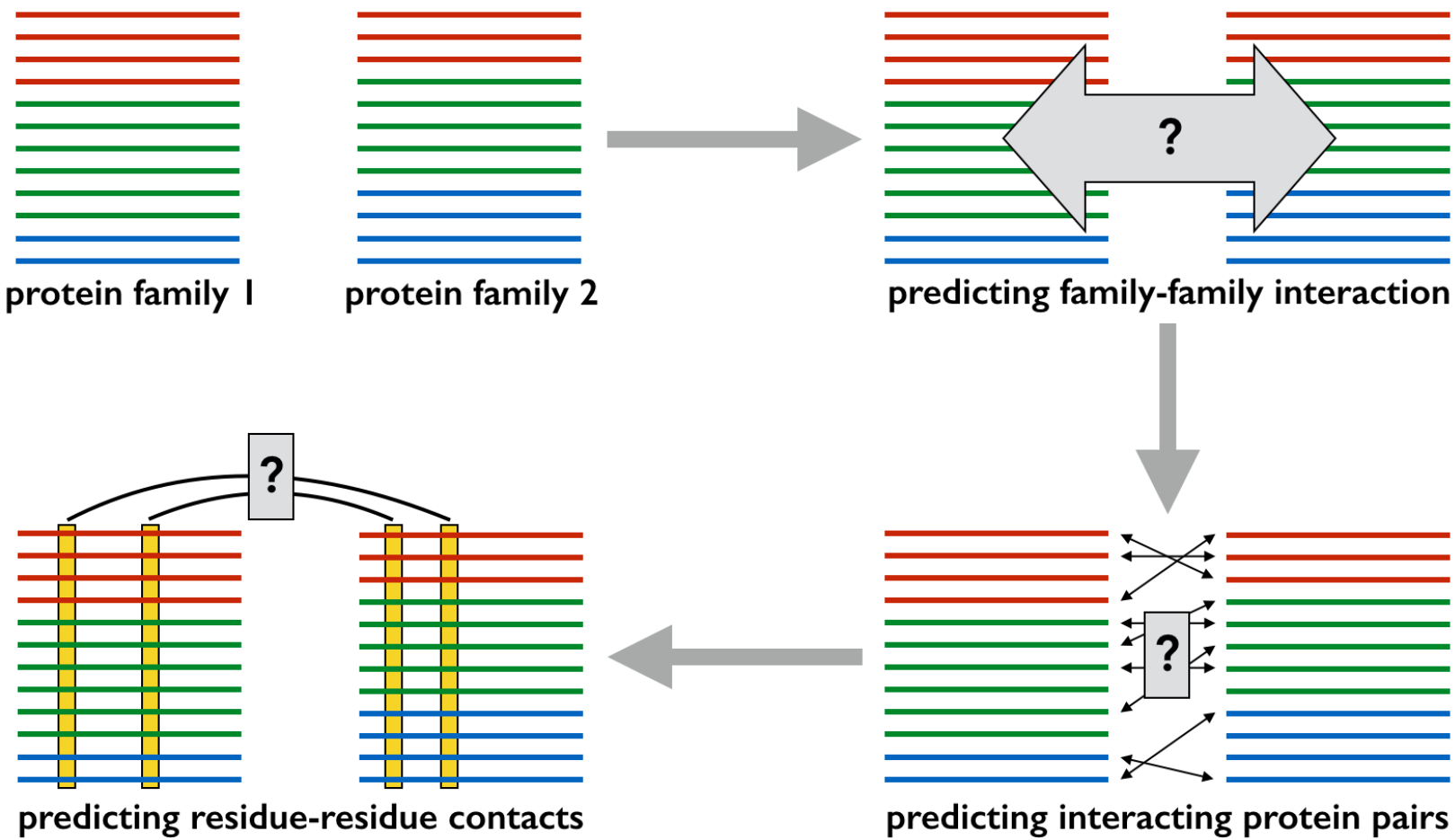
- correct matching shows inter-protein covariation
- random matching has no inter-protein covariation

➡ maximise inter-protein covariation computationally

- reach 80-90% of accuracy in test cases
- simultaneous prediction of interacting paralogs and inter-protein contacts

[Gueudré, Baldassi, Zamparo, MW, Pagnani, *PNAS* '16]
[Bitbol, Dwyer, Colwell, Wingreen, *PNAS* '16]

Interactions between protein families

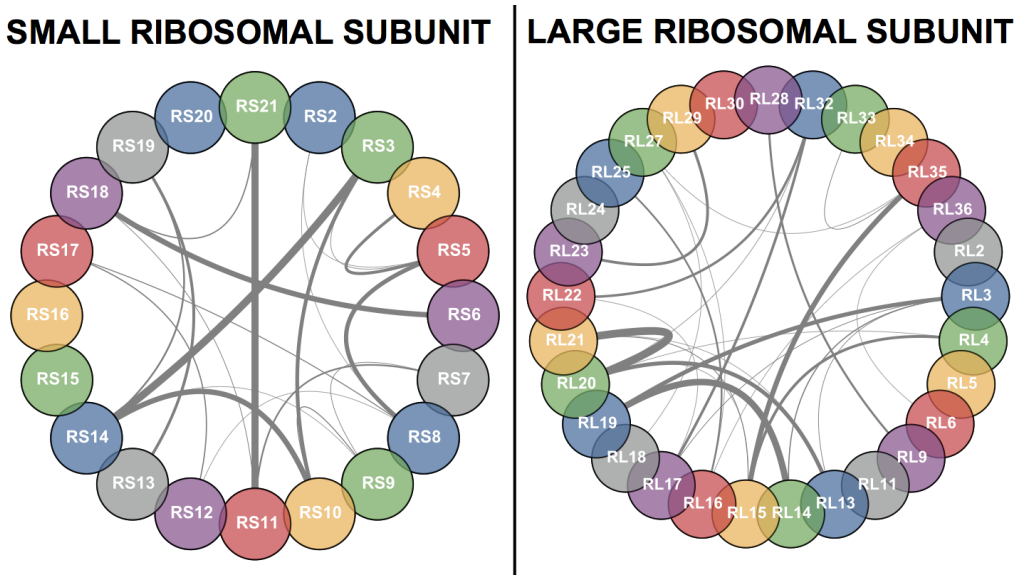


What can we learn from the empirical sequence variability:

- do the families interact?
- which specific proteins interact?
- which residues are in contact?

➡ relation between protein structure/function and evolution

Inference of protein-protein interaction networks



Bacterial ribosomal proteins

Small ribosomal subunit

- 20 proteins
- 21 interactions (11% of 190 pairs)

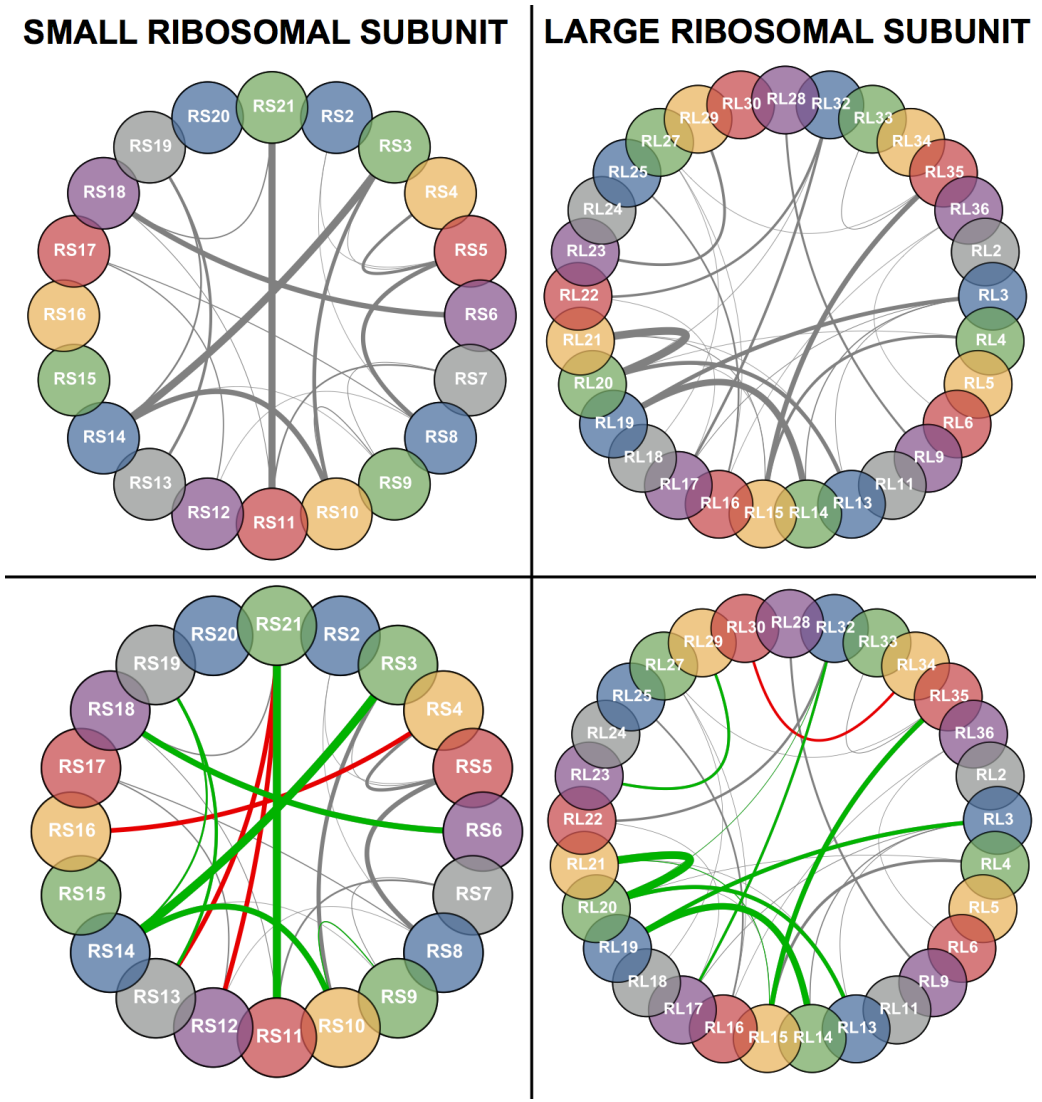
Large ribosomal subunit

- 29 proteins
- 29 interactions (7% of 406 pairs)

▸ sparse interaction network

[Feinauer, Szurmant, MW, Pagnani, *PLoS ONE* '16]

Inference of protein-protein interaction networks



Bacterial ribosomal proteins

Pairwise DCA (1000-3000 seqs.)

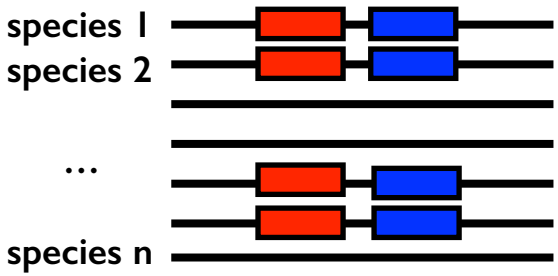
Top 10 predictions for each subunit

- 16 true positive interactions (80% TP vs. 8% in random prediction)
- find most large interfaces
- fail to detect small interfaces
- false predictions appear in smaller alignments

▸ larger alignments needed

[Feinauer, Szurmant, MW, Pagnani, *PLoS ONE* '16]
cf. also [Uguzzoni, Lovis, Oteri, Schug, Szurmant, MW, *PNAS* '17]

Exploring genomic scales

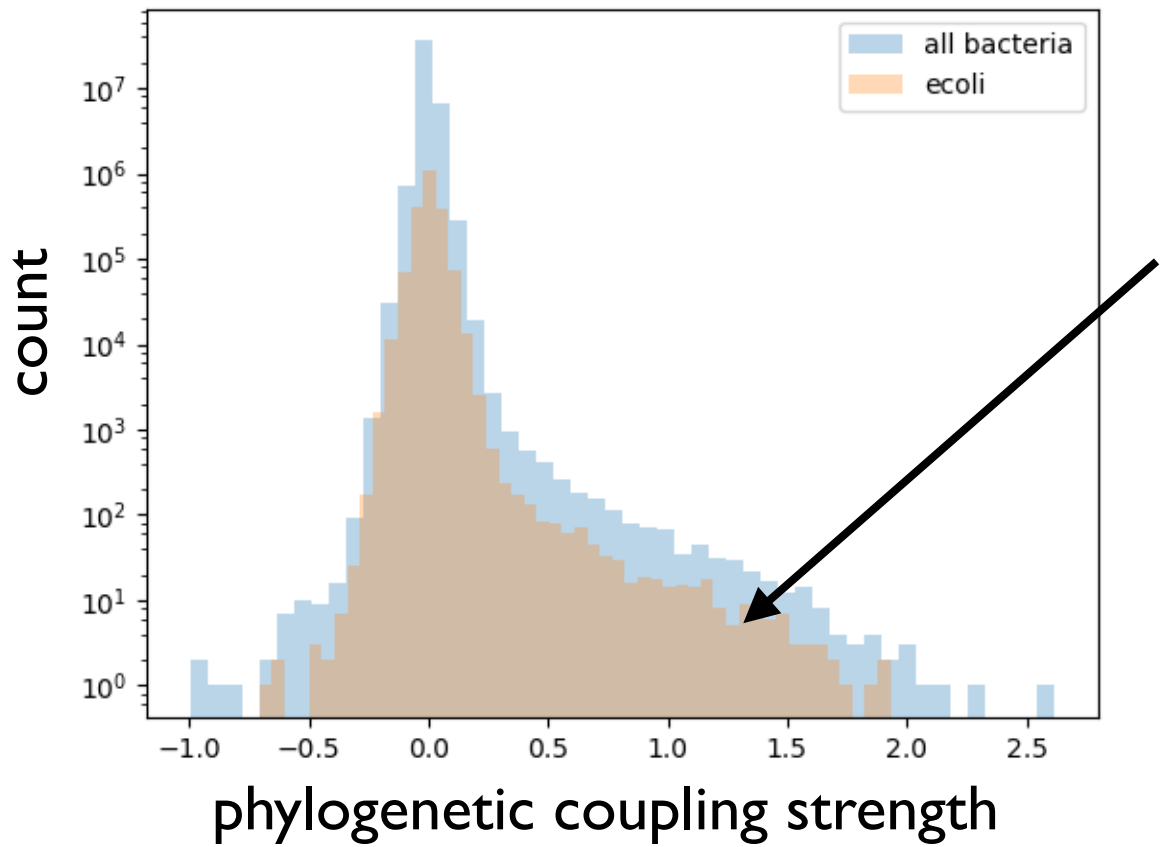


correlated presence / absence of interacting proteins

– phylogenetic profiles [Pellegrini et al. 1999]

– correlated phylogenetic trees [Pazos et al. 2001]

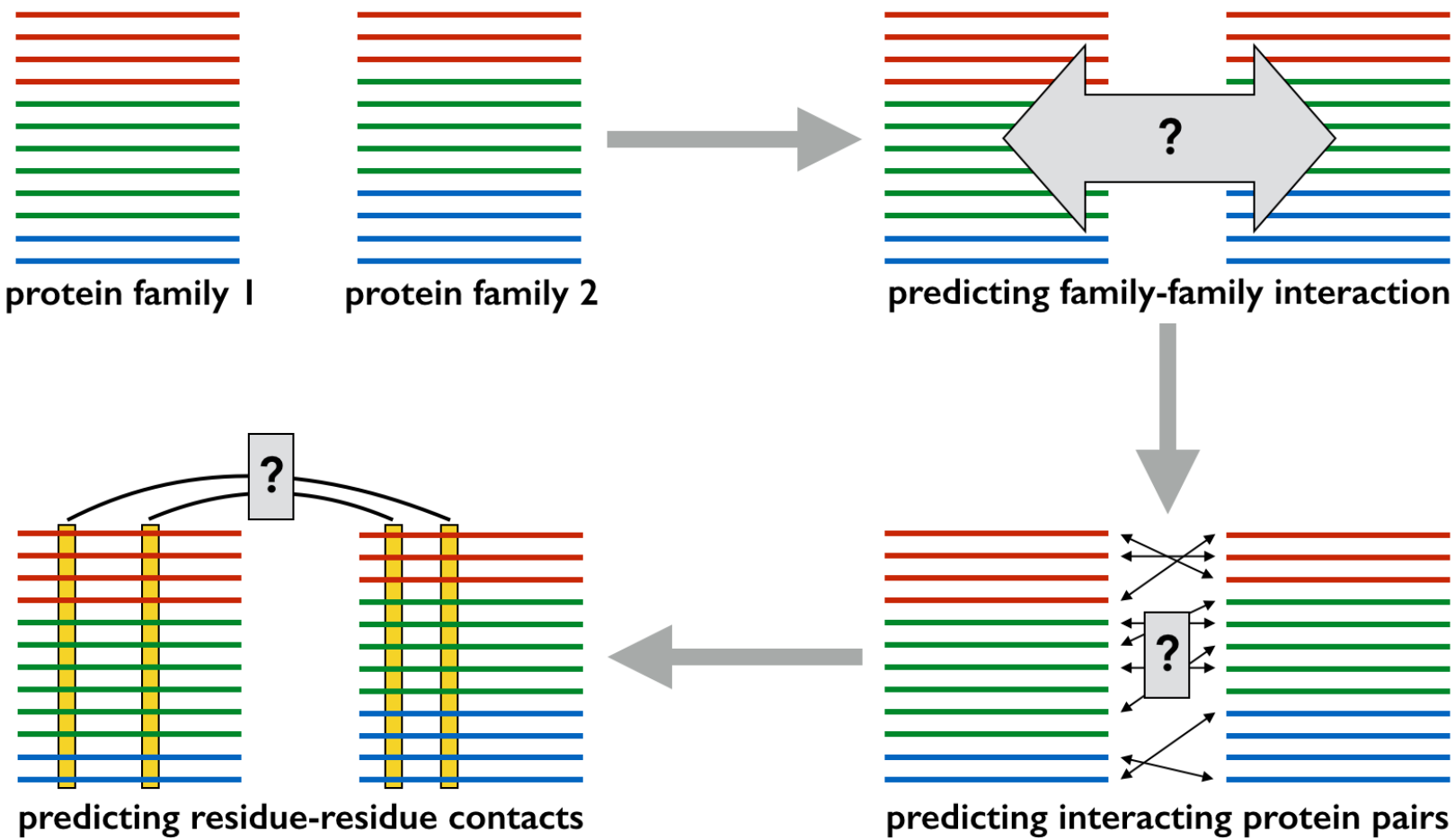
– phylogenetic coupling analysis [Croce et al., in prep]



Tail of ~1000 strong couplings

- 80% known relations (interaction, colocalisation)
- 20% new predictions

Interactions between protein families



What can we learn from the empirical sequence variability:

- do the families interact?
- which specific proteins interact?
- which residues are in contact?

➡ towards a structurally resolved & evolutionary conserved interactome

Thanks to:

The group in Paris:

Juliana Bernardes
Pierre Barrat-Charlaix
Giancarlo Croce
Kai Shimagaki
Edwin Rodriguez

Francesco Oteri

Alumni:

Eleonora de Leonardis
Guido Uguzzoni
Alice Coucke
Matteo Figliuzzi
Christoph Feinauer

Collaborators:

Terry Hwa (UC San Diego)
Hendrik Szurmant (Western U LA)
Alexander Schug (KIT Karlsruhe)
Jose Onuchic (Rice U, Austin)
Faruck Morcos (UT Dallas)
Angel E. Dago (Scripps La Jolla)
Joanna Sulkowska (U Warsaw)
Erik Aurell (KTH Stockholm)
Andrea Pagnani (Politecnico Torino)
Thomas Gueudré (IIGM Torino)
Carlo Baldassi (U Bocconi Milano)
Rémi Monasson (ENS)
Simona Cocco (ENS)
Olivier Tenaillon (Inserm Paris)

Funding:

Agence Nationale de la Recherche ANR

îledeFrance Demain s'invente ici

ICS Institut du calcul et de la simulation

Horizon 2020 European Union funding for Research & Innovation