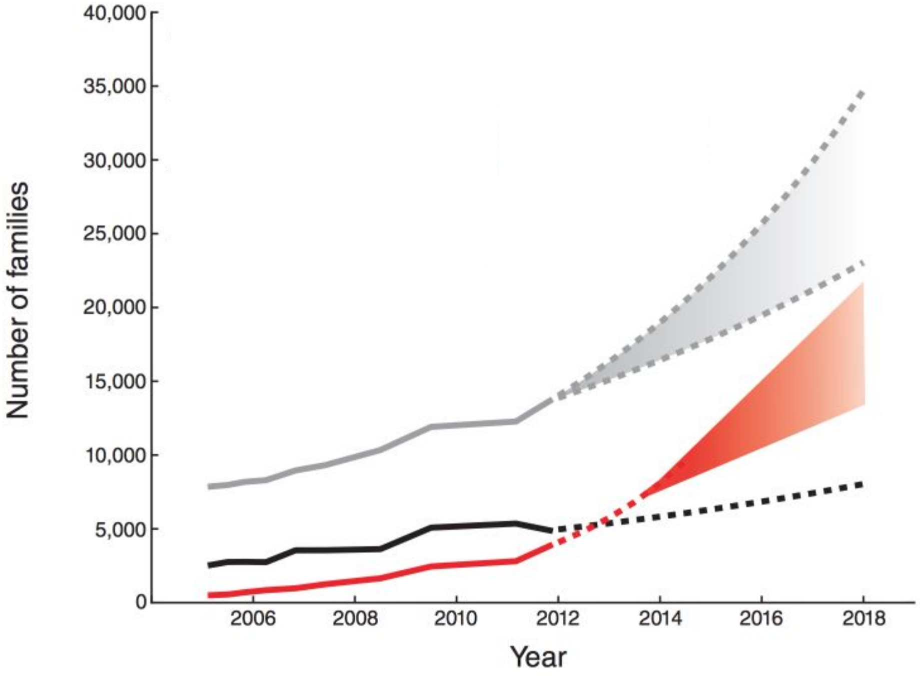
Lecture 11: Protein structure prediction

- Amino acid coevolution
- Residue coupling and contact prediction
  - Maximum-entropy model
- Extensions

Direct vs. indirect interactions



New protein families being discovered by high-throughput sequencing

Experimental structure-determination

Marks (2012) Nat. Biotech.

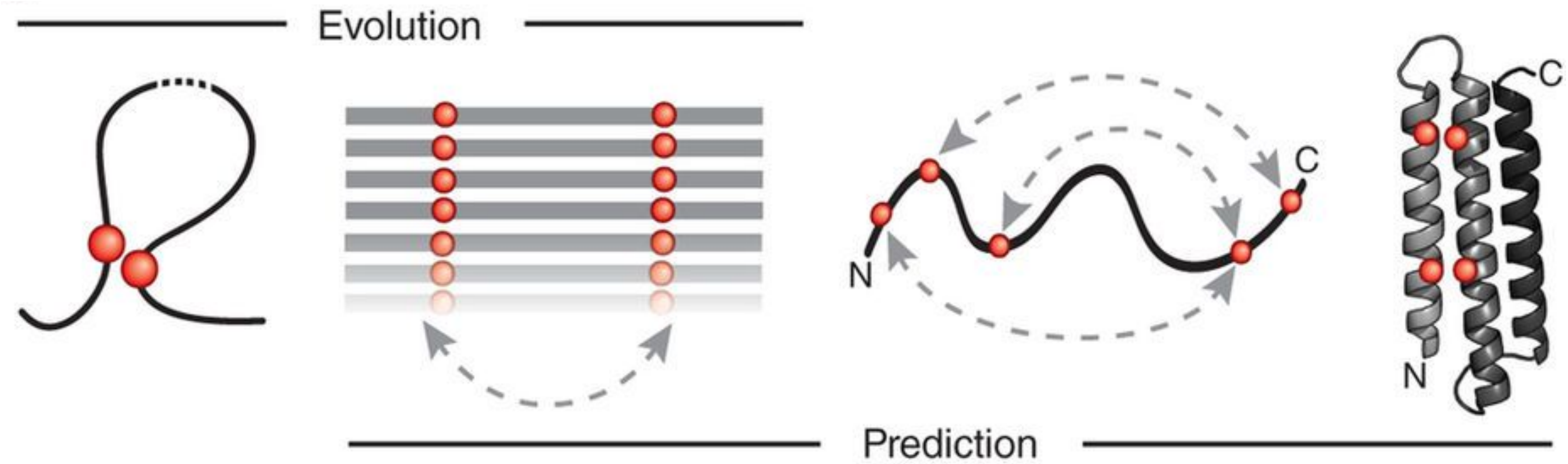
Predicting protein 3D structure from sequence

Evolutionary pressure to maintain favorable interactions b/w physically interacting AA residues in 3D.

Visible record of residue covariation in related protein sequences.

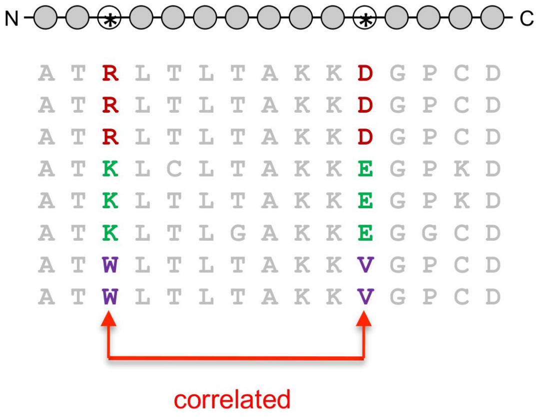
Inverse problem – inferring directly causative residue couplings (evolutionary couplings) from the covariation record – challenging due to transitive correlations & other confounding effects.

ECs can be used to predict the unknown 3D structure of a protein from a set of sequences alone.

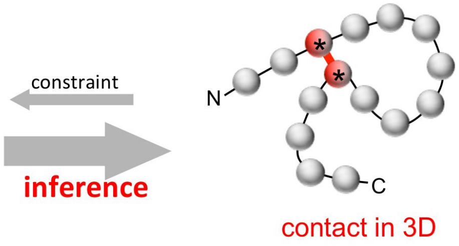


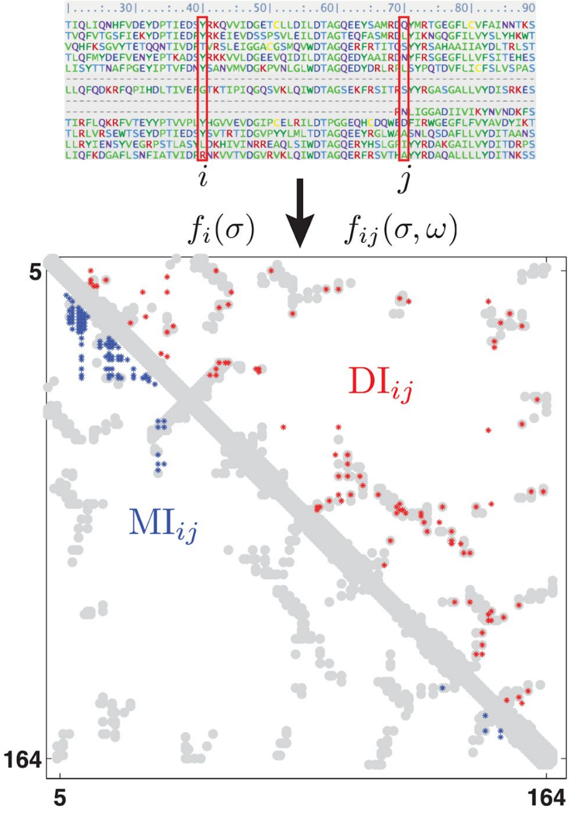
Marks (2012) Nat. Biotech.

Predicting protein 3D structure from sequence



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

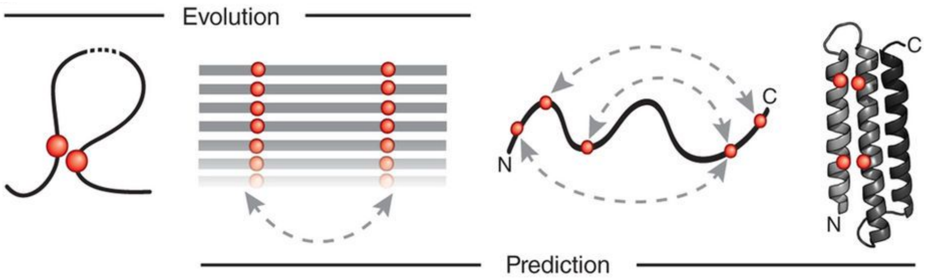




```
|....:..30|....:..40|....:..50|....:..60|....:..70|....:..80|....:..90
TIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKS
TVQFVTGSFIEKYDPTIEDFYRKEIEVDSSPSVLEILDTAGTEQFASMRDLYIKNGQGFILVYSLYHKWT
VQHFKSGVYTETQQNTIVDFTVRSLEIGGACGSMQVWDTAGQERFRTITQSYYRSAHAAIIAYDLTRLST
TLQFMYDEFVENYEPTKADSYRKKVVLDGEEVQIDILDTAGQEDYAAIRDNYFRSGEGFLLVFSITEHES
LISYTTNAFPGEYIPTVFDNYSANVMVDGKPVNLGLWDTAGQEDYDRLRPLSYPQTDVFLICFSLVSPAS
----------------------------------------------------------------------
LLQFQDKRFQPIHDLTIVEFGTKTIPIQGQSVKLQIWDTAGSEKFRSITRSYYRGASGALLVYDISRKES
----------------------------------------------------------------------
-------------------------------------------------RNLIGGADIIVIKYNVNDKFS
TIRFLQKRFVTEYYPTVVPLYHGVVEVDGIPCELRILDTPGGEQHCDQWEDFIRWGEGFLFVYAVDYIKT
TLRLVRSEWTSEYDPTIEDSYSVTRTIDGVPYYLMLTDTAGQEEYRGLWAASNLQSDAFLLVYDITAANS
LLRYIENSYVEGRPSTLASYLDKHIVINRREAQLSIWDTAGQERYHSLGPIYYRDAKGAILVYDITDRPS
LIQFKDGAFLSNFIATVIDFRNKVVTVDGVRVKLQIWDTAGQERFRSVTHAYYRDAQALLLLYDITNKSS
```

Marks (2011) PLoS One; Marks (2012) Nat. Biotech.
Stein (2015) PLoS Comp. Biol.

Predicting protein 3D structure from sequence



Growth in sequence databases from massively parallel sequencing.

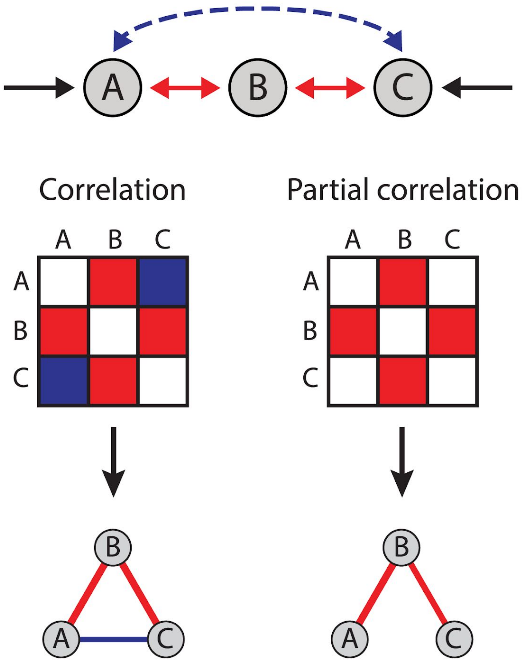
- Availability of sufficient sequences of sufficient diversity.
- Known protein families are growing in size from a few sequences to many thousands of sequences (advances in DNA sequencing tech).

Reduction of conformational search space by cooperative probability models.

- Global probability models account for the fact that interactions along an entire protein chain are mutually interdependent in a way that is inherently cooperative.
- Pair interactions are modified by interactions with other parts of the system and cannot be factored (probabilities are not a simple product of independent terms).
- Compared with molecular dynamics simulations, statistical approaches are many orders of magnitude more efficient in reducing a huge conformational search space to manageable proportions.

Marks (2012) Nat. Biotech.

Direct vs. indirect interactions



Correlation

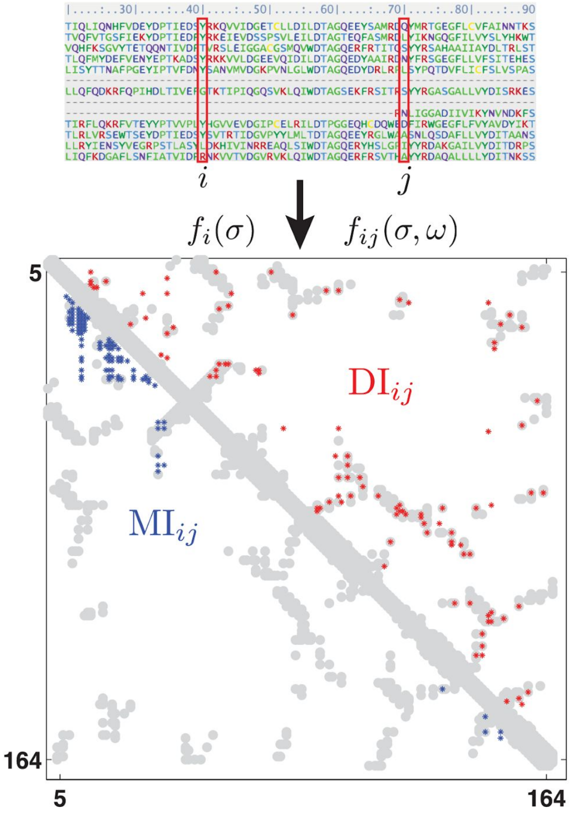
	A	B	C
A	⬜	🟥	🟦
B	🟥	⬜	🟥
C	🟦	🟥	⬜

Partial correlation

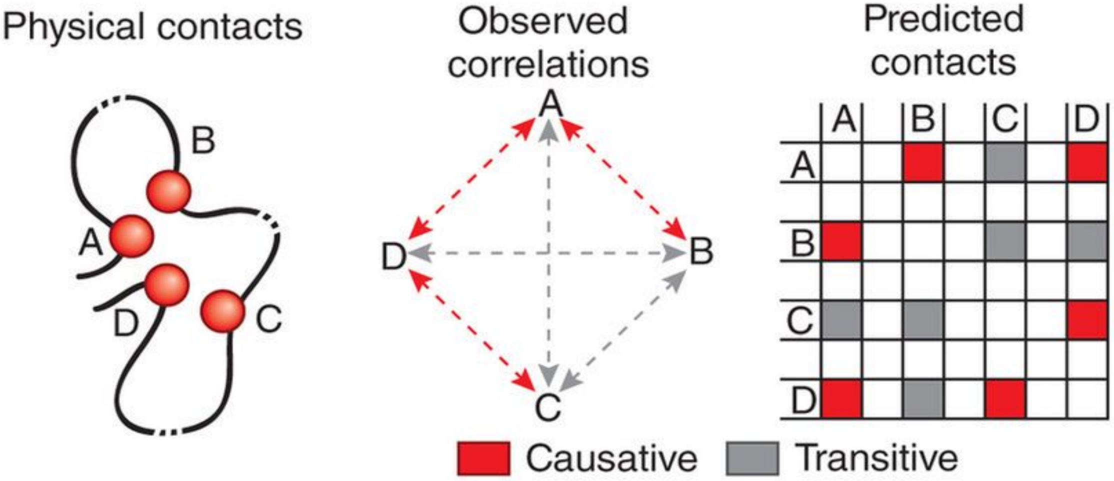
	A	B	C
A	⬜	🟥	⬜
B	🟥	⬜	🟥
C	⬜	🟥	⬜

Stein (2015) PLoS Comp. Biol.

Direct vs. indirect interactions



```
|....:..30|....:..40|....:..50|....:..60|....:..70|....:..80|....:..90
TIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKS
TVQFVTGSFIEKYDPTIEDFYRKEIEVDSSPSVLEILDTAGTEQFASMRDLYIKNGQGFILVYSLYHKWT
VQHFKSGVYTETQQNTIVDFTVRSLEIGGACGSMQVWDTAGQERFRTITQSYYRSAHAAIIAYDLTRLST
TLQFMYDEFVENYEPTKADSYRKKVVLDGEEVQIDILDTAGQEDYAAIRDNYFRSGEGFLLVFSITEHES
LISYTTNAFPGEYIPTVFDNYSANVMVDGKPVNLGLWDTAGQEDYDRLRPLSYPQTDVFLICFSLVSPAS
----------------------------------------------------------------------
LLQFQDKRFQPIHDLTIVEFGTKTIPIQGQSVKLQIWDTAGSEKFRSITRSYYRGASGALLVYDISRKES
----------------------------------------------------------------------
-------------------------------------------------RNLIGGADIIVIKYNVNDKFS
TIRFLQKRFVTEYYPTVVPLYHGVVEVDGIPCELRILDTPGGEQHCDQWEDFIRWGEGFLFVYAVDYIKT
TLRLVRSEWTSEYDPTIEDSYSVTRTIDGVPYYLMLTDTAGQEEYRGLWAASNLQSDAFLLVYDITAANS
LLRYIENSYVEGRPSTLASYLDKHIVINRREAQLSIWDTAGQERYHSLGPIYYRDAKGAILVYDITDRPS
LIQFKDGAFLSNFIATVIDPRNKVVTVDGVRVKLQIWDTAGQERFRSVTHAYYRDAQALLLLYDITNKSS
```



	A	B	C	D
A		🟥	▨	🟥
B	🟥		▨	▨
C	▨	▨		🟥
D	🟥	▨	🟥	

🟥 Causative ▨ Transitive

Marks (2011) PLoS One; Marks (2012) Nat. Biotech.
Stein (2015) PLoS Comp. Biol.

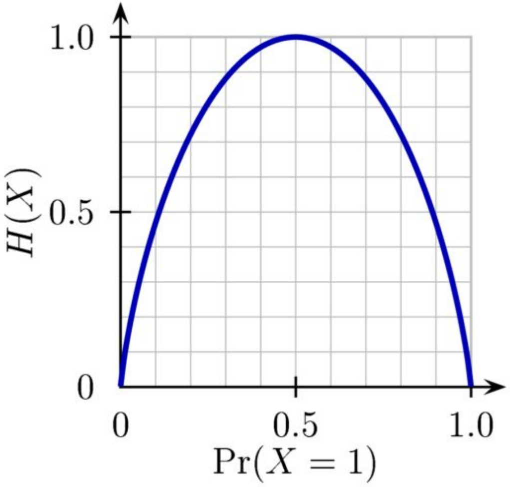
Information theory

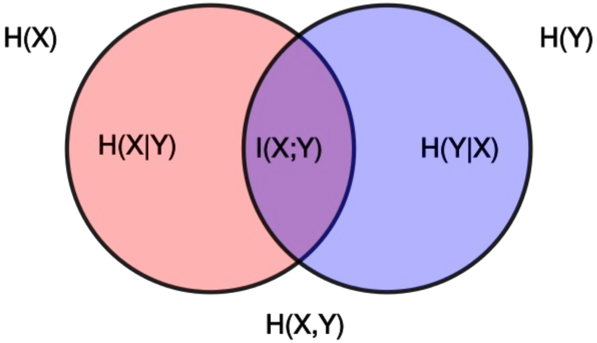
Entropy (H): the average amount of information produced by a stochastic source of data.

Mutual information: MI two random variables **I(X, Y)** quantifies the amount of information obtained about one random variable, through the other random variable.

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

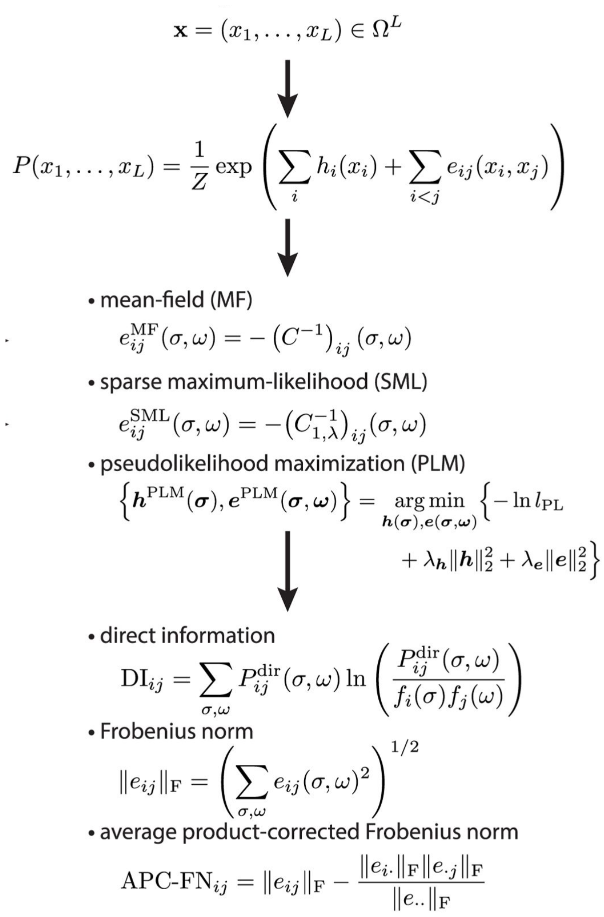
$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$





Wikipedia

Global probabilistic models of residue coupling (maximum-entropy)



$$\mathbf{x} = (x_1, \ldots, x_L) \in \Omega^L$$

$$\downarrow$$

$$P(x_1, \ldots, x_L) = \frac{1}{Z} \exp\left(\sum_i h_i(x_i) + \sum_{i<j} e_{ij}(x_i, x_j)\right)$$

$$\downarrow$$

- mean-field (MF)

$$e_{ij}^{\mathrm{MF}}(\sigma, \omega) = -\left(C^{-1}\right)_{ij}(\sigma, \omega)$$

- sparse maximum-likelihood (SML)

$$e_{ij}^{\mathrm{SML}}(\sigma, \omega) = -\left(C_{1,\lambda}^{-1}\right)_{ij}(\sigma, \omega)$$

- pseudolikelihood maximization (PLM)

$$\left\{\boldsymbol{h}^{\mathrm{PLM}}(\boldsymbol{\sigma}), \boldsymbol{e}^{\mathrm{PLM}}(\boldsymbol{\sigma}, \boldsymbol{\omega})\right\} = \underset{\boldsymbol{h}(\boldsymbol{\sigma}), \boldsymbol{e}(\boldsymbol{\sigma}, \boldsymbol{\omega})}{\arg\min} \left\{-\ln l_{\mathrm{PL}} + \lambda_{\boldsymbol{h}} \|\boldsymbol{h}\|_2^2 + \lambda_{\boldsymbol{e}} \|\boldsymbol{e}\|_2^2\right\}$$

$$\downarrow$$

- direct information

$$\mathrm{DI}_{ij} = \sum_{\sigma, \omega} P_{ij}^{\mathrm{dir}}(\sigma, \omega) \ln\left(\frac{P_{ij}^{\mathrm{dir}}(\sigma, \omega)}{f_i(\sigma) f_j(\omega)}\right)$$

- Frobenius norm

$$\|e_{ij}\|_{\mathrm{F}} = \left(\sum_{\sigma, \omega} e_{ij}(\sigma, \omega)^2\right)^{1/2}$$

- average product-corrected Frobenius norm

$$\mathrm{APC\text{-}FN}_{ij} = \|e_{ij}\|_{\mathrm{F}} - \frac{\|e_{i\cdot}\|_{\mathrm{F}} \|e_{\cdot j}\|_{\mathrm{F}}}{\|e_{\cdot\cdot}\|_{\mathrm{F}}}$$

Pairwise maximum-entropy distribution

Parameter inference

Pair scoring functions

Stein (2015) PLoS Comp. Biol.

Global probabilistic models of residue coupling (maximum-entropy)

$$a = (a_1, a_2 .. , a_N)$$

A sequence made of monomers a_i taking values from a given alphabet

$$P(a|J,h) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(a_i, a_j) + \sum_{i=1}^{N} h_i(a_i) \right)$$

Probability of a sequence within the model.

$h(a_i)$: parameters that represent the propensity of symbol to be found at a certain position.

$J(a_i, a_j)$: represent an interaction, quantifying how compatible the symbols at both positions are with each other.

Wikipedia

Global probabilistic models of residue coupling (maximum-entropy)

$$a = (a_1, a_2 . . , a_N)$$

$$P(a|J,h) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(a_i, a_j) + \sum_{i=1}^{N} h_i(a_i)\right)$$

The idea of maximum-entropy: For a given set of sample covariances and frequencies, the model represents the **distribution with the maximal entropy** of all distributions reproducing those covariances and frequencies.

$$\begin{aligned} F[P] = &- \sum_a P(a) \log P(a) \\ &+ \sum_{i<j} \sum_{x,y} \lambda_{ij}(x,y)\Big(P_{ij}(x,y) - f_{ij}(x,y)\Big) \\ &+ \sum_i \sum_x \lambda_i(x)\Big(P_i(x) - f_i(x)\Big) \\ &+ \Omega\left(1 - \sum_a P(a)\right). \end{aligned}$$

The unique distribution ***P*** that maximizes the functional to the *left*.

f_i(a): frequency of finding symbol ***a*** at position ***i***.

f_{ij}(a, b): frequency of finding symbols ***a*** & ***b*** at positions ***i*** and ***j*** in the same sequence.

Wikipedia

Global probabilistic models of residue coupling (maximum-entropy)

$$a = (a_1, a_2.., a_N)$$

$$P(a|J,h) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(a_i, a_j) + \sum_{i=1}^{N} h_i(a_i) \right)$$

$$\begin{aligned} F[P] = &- \sum_a P(a) \log P(a) \\ &+ \sum_{i<j} \sum_{x,y} \lambda_{ij}(x,y) \Big(P_{ij}(x,y) - f_{ij}(x,y) \Big) \\ &+ \sum_i \sum_x \lambda_i(x) \Big(P_i(x) - f_i(x) \Big) \\ &+ \Omega \left(1 - \sum_a P(a) \right). \end{aligned}$$

$$F_i = \frac{1}{N} \sum_{j \neq i}^{N} F_{ij}$$

$$F_{ij}^{APC} = F_{ij} - \frac{F_i F_j}{F}$$

$$F = \frac{1}{N^2 - N} \sum_{i,j,i \neq j}^{N} F_{ij}$$

The idea of maximum-entropy: For a given set of sample covariances and frequencies, the model represents the **distribution with the maximal entropy** of all distributions reproducing those covariances and frequencies.

The unique distribution ***P*** that maximizes the functional to the *left*.

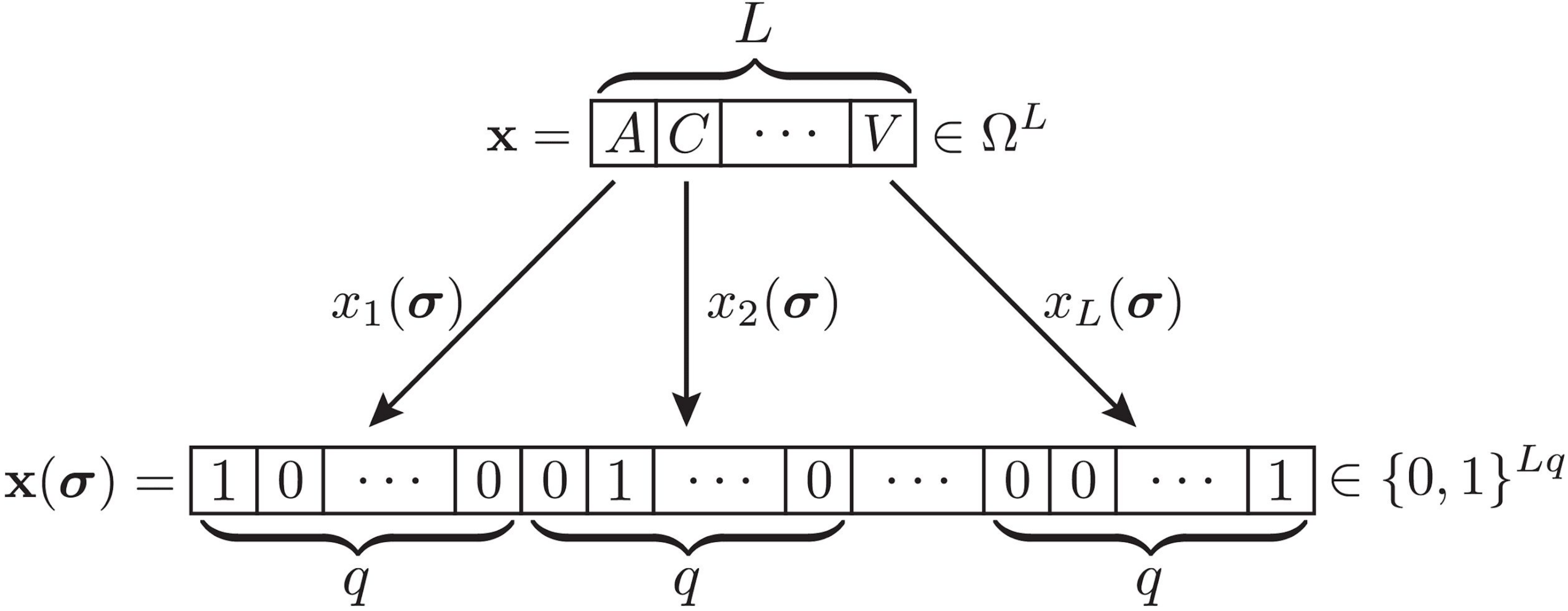
Final step:

- average product correction (APC).

Wikipedia

Global probabilistic models of residue coupling

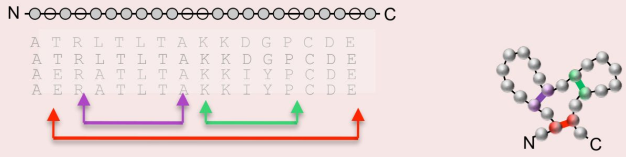
Binary embedding of amino acid sequence



Stein (2015) PLoS Comp. Biol.

Global probabilistic models of residue coupling (maximum-entropy)

Align evolutionary diverged sequences


N
C
A T R L T L T A K K D G P C D E
A T R L T L T A K K D G P C D E
A E R A T L T A K K I Y P C D E
A E R A T L T A K K I Y P C D E
N
C

Calculate covariance matrix for each pair of sequence positions for all pairs of amino acids (A,B)

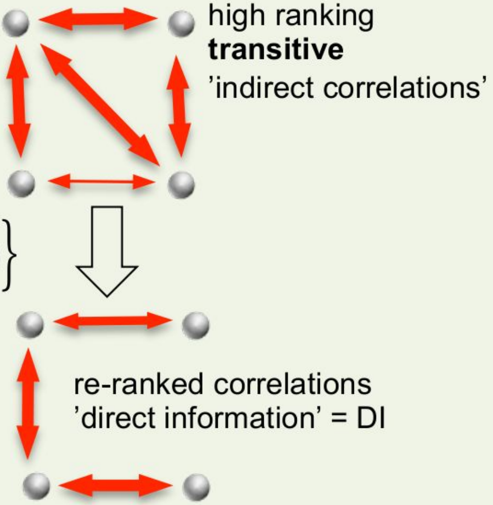
$$C_{ij}(A,B) = f_{ij}(A,B) - f_i(A)P_j(B)$$

$$C_{ij}^{-1}(A,B) = -e_{ij}(A,B)_{i \neq j}$$

$$P_{ij}^{Dir}(A,B) = \frac{1}{Z}\exp\left\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right\}$$

Identify maximally informative pair couplings using **statistical model** of entire protein to infer residue-residue co-evolution

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



high ranking **transitive** ’indirect correlations’

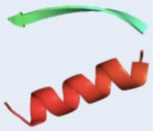
re-ranked correlations ’direct information’ = DI

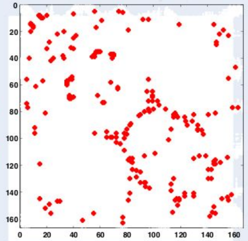
Marks (2011) PLoS One

From contacts to structure

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

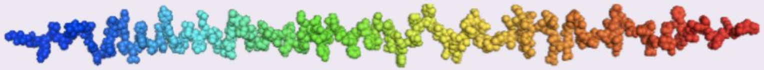
```
assign (resid 143 and name CA) (resid 123 and name CA)  4 4 3
assign (resid 16 and name CA) (resid 10 and name CA)  4 4 3
assign (resid 141 and name CA) (resid 82 and name CA)  4 4 3
assign (resid 129 and name CA) (resid 87 and name CA)  4 4 3
assign (resid 92 and name CA) (resid 11 and name CA)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



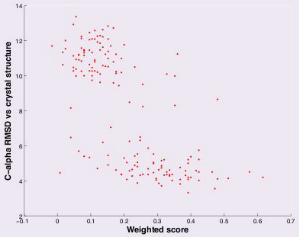


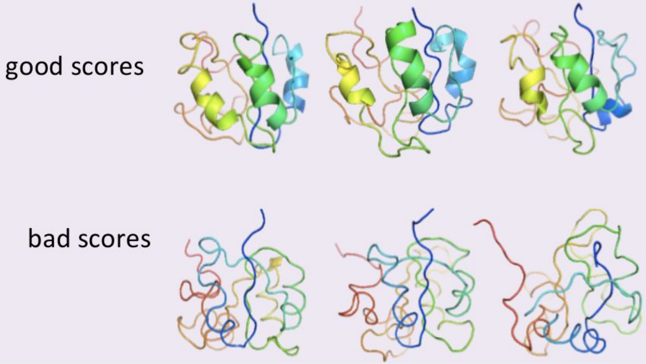
predicted contacts (EICs)

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



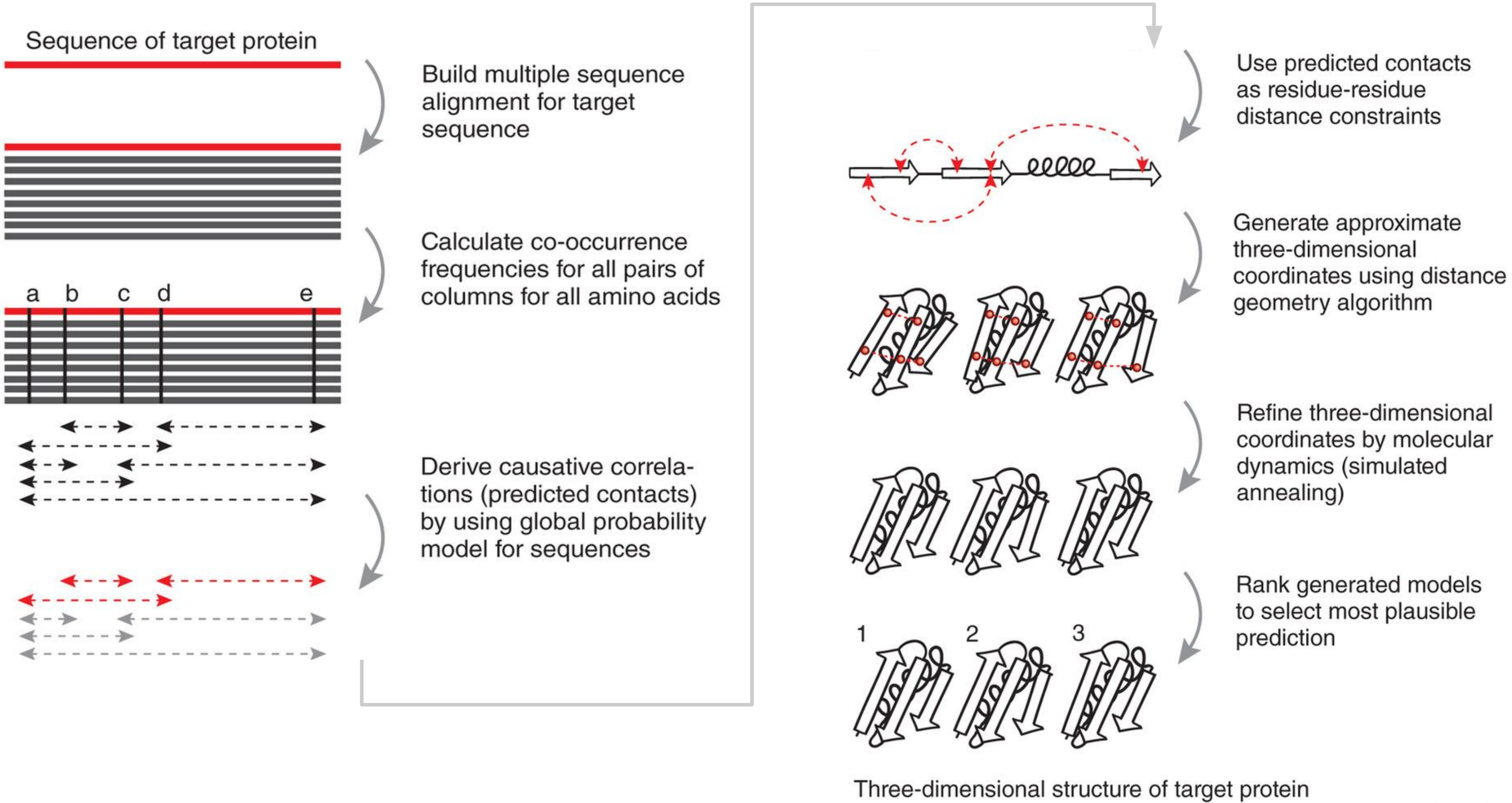
Rank predicted structures using quality measure of backbone alpha torsion and beta sheet twist





Marks (2011) PLoS One

Predicting protein 3D structure from sequence



Sequence of target protein

Build multiple sequence alignment for target sequence

Calculate co-occurrence frequencies for all pairs of columns for all amino acids

Derive causative correlations (predicted contacts) by using global probability model for sequences

Use predicted contacts as residue-residue distance constraints

Generate approximate three-dimensional coordinates using distance geometry algorithm

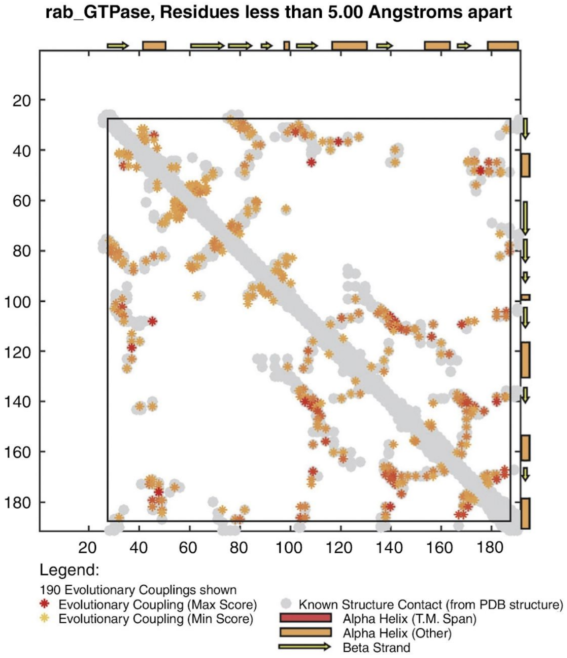
Refine three-dimensional coordinates by molecular dynamics (simulated annealing)

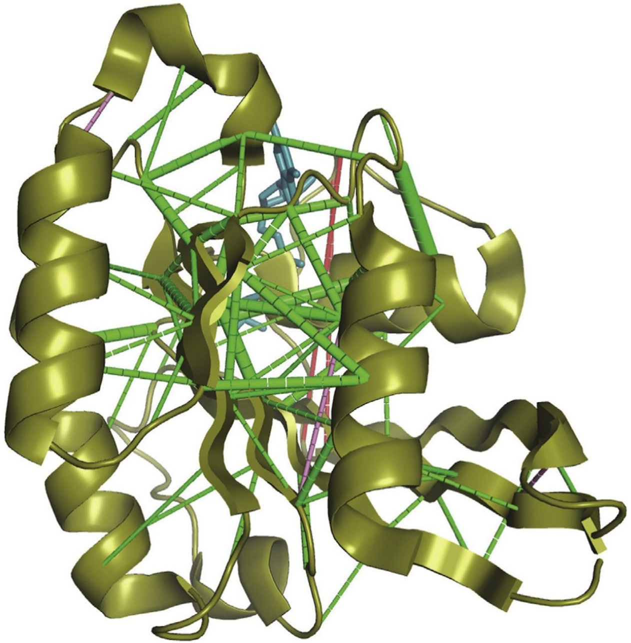
Rank generated models to select most plausible prediction

Three-dimensional structure of target protein

Marks (2012) Nat. Biotech.

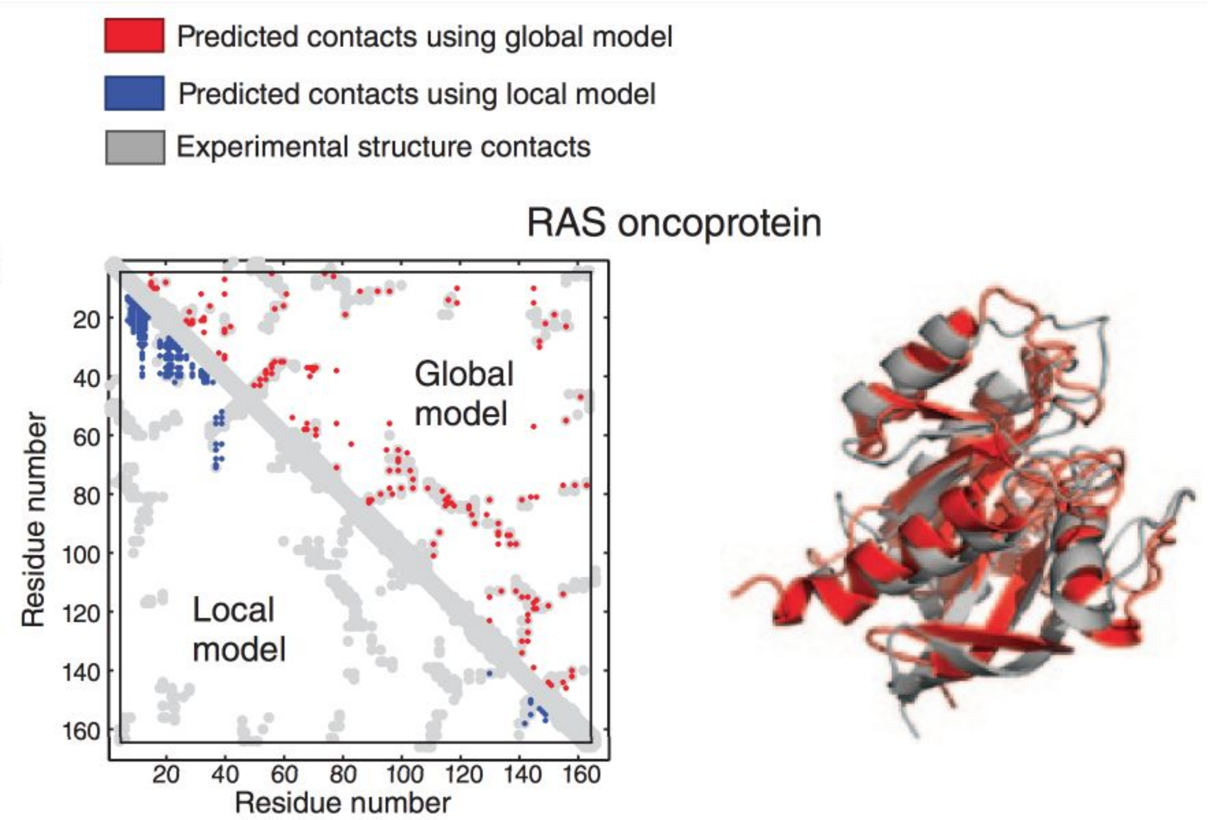
Predictions of 3D structures based on evolutionary coupling





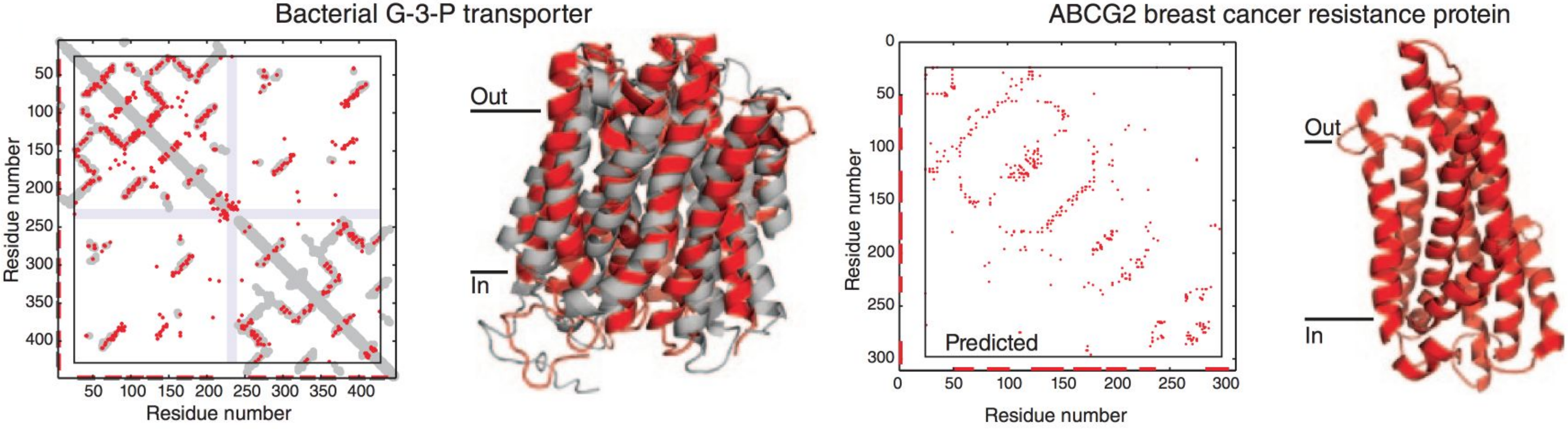
Newald (2016) Curr. Opin. Struct. Biol.

Predictions of 3D structures based on evolutionary coupling



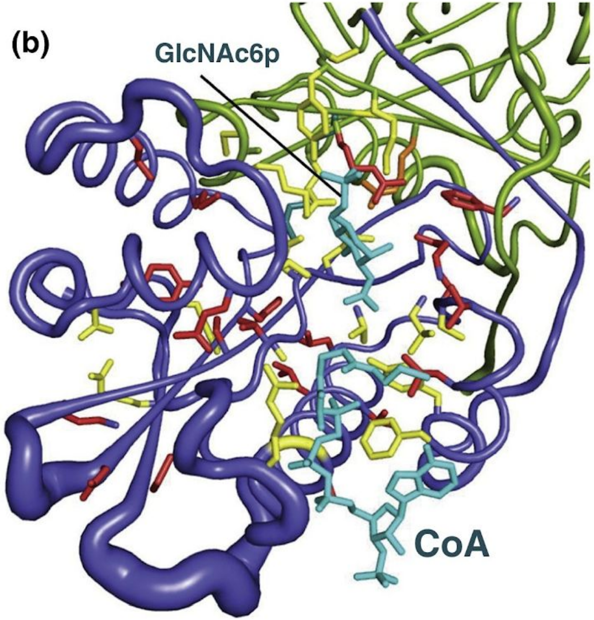
Marks (2012) Nat. Biotech.

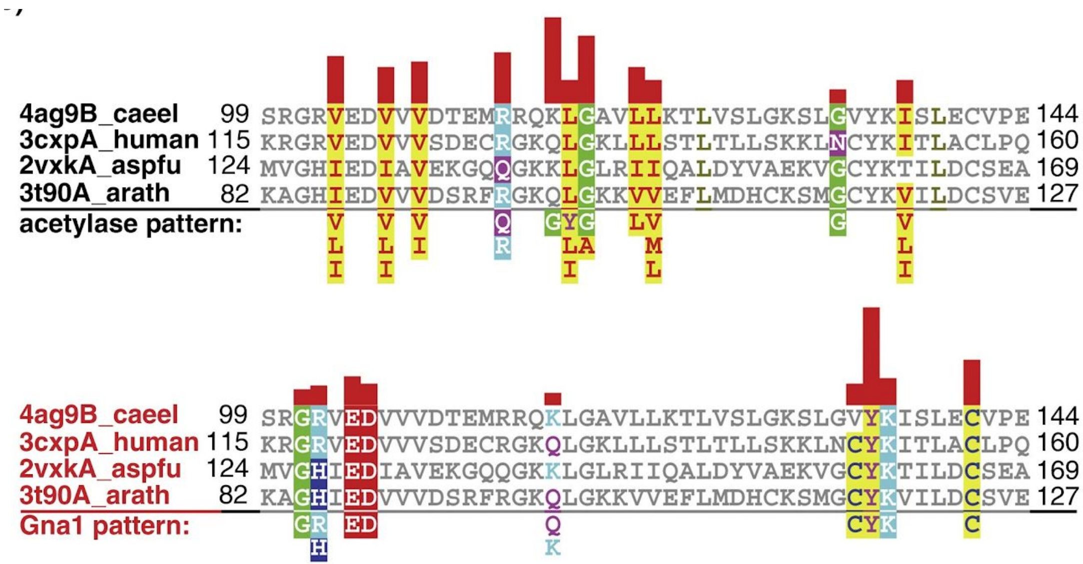
Predictions of 3D structures based on evolutionary coupling



Marks (2012) Nat. Biotech.

Predictions of 3D structures based on evolutionary coupling



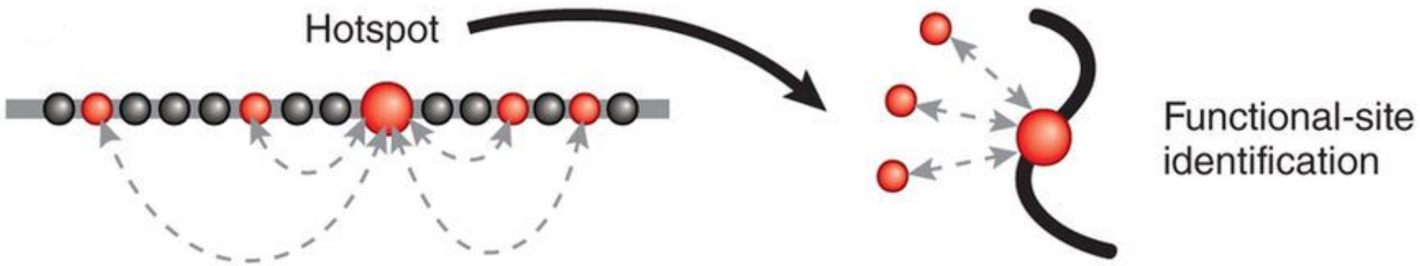

4ag9B_caeel 99 SRGRVEDVVVDTEMRRQKLGAVLLKTLVSLGKSLGVYKISLECVPE 144
3cxpA_human 115 KRGRVEDVVVSDECRGKQLGKLLLSTLTLLSKKLNCYKITLACLPQ 160
2vxkA_aspfu 124 MVGHIEDIAVEKGQQGKKLGLRIIQALDYVAEKVGCYKTILDCSEA 169
3t90A_arath 82 KAGHIEDVVVDSRFRGKQLGKKVVEFLMDHCKSMGCYKVILDCSVE 127
acetylase pattern:
V V V Q GYG LV G V
L L I R LA M L
I I I L I
4ag9B_caeel 99 SRGRVEDVVVDTEMRRQKLGAVLLKTLVSLGKSLGVYKISLECVPE 144
3cxpA_human 115 KRGRVEDVVVSDECRGKQLGKLLLSTLTLLSKKLNCYKITLACLPQ 160
2vxkA_aspfu 124 MVGHIEDIAVEKGQQGKKLGLRIIQALDYVAEKVGCYKTILDCSEA 169
3t90A_arath 82 KAGHIEDVVVDSRFRGKQLGKKVVEFLMDHCKSMGCYKVILDCSVE 127
Gna1 pattern:
GR ED Q CYK C
H K

Newald (2016) Curr. Opin. Struct. Biol.

Detecting functional hotspots

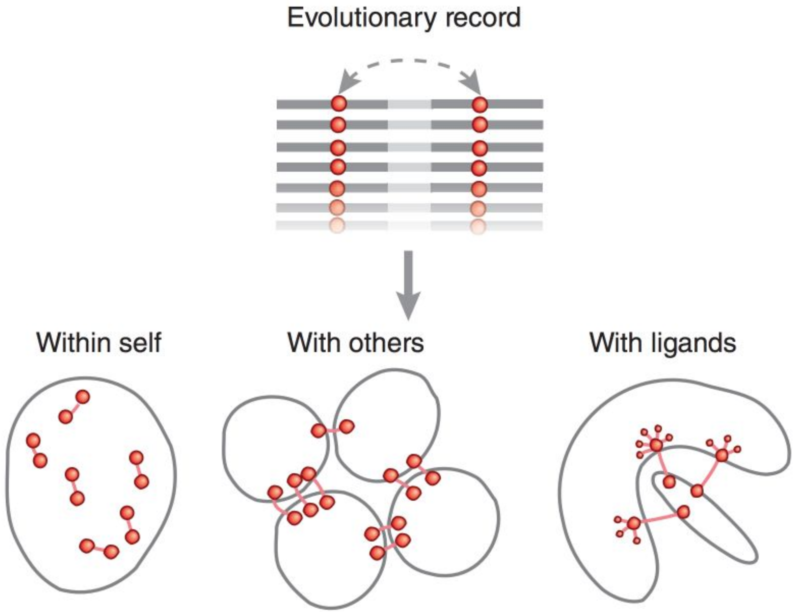
Residues subject to a high number of evolutionary pair constraints represent likely functional hotspots.

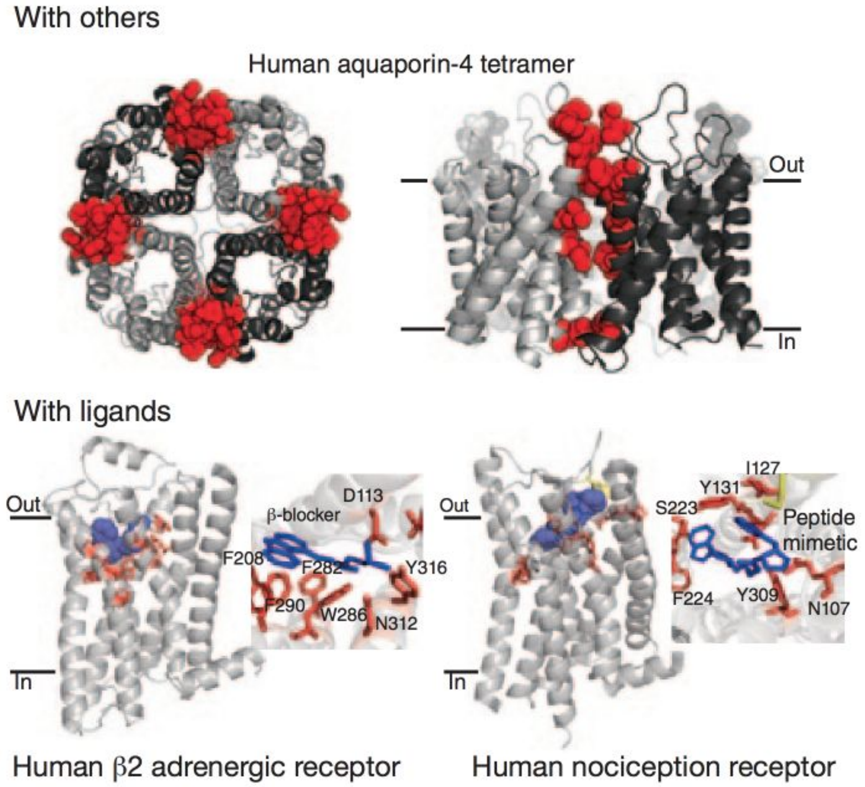
- Such highly constrained residues include residues in functional sites (for e.g., interaction with external ligands).
- Not detectable by analysis of single-residue conservation.



Marks (2012) Nat. Biotech.

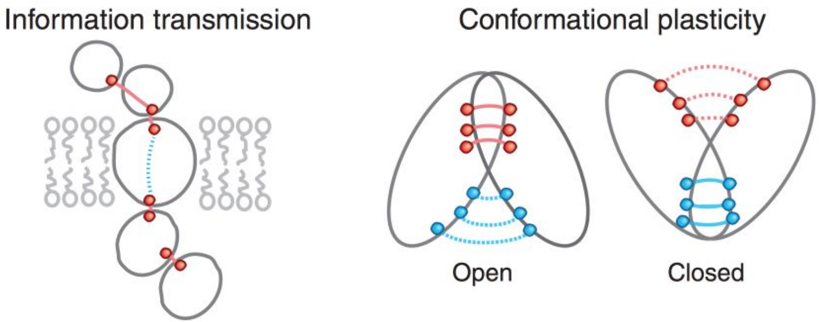
Predicting protein-protein & protein-ligand interactions

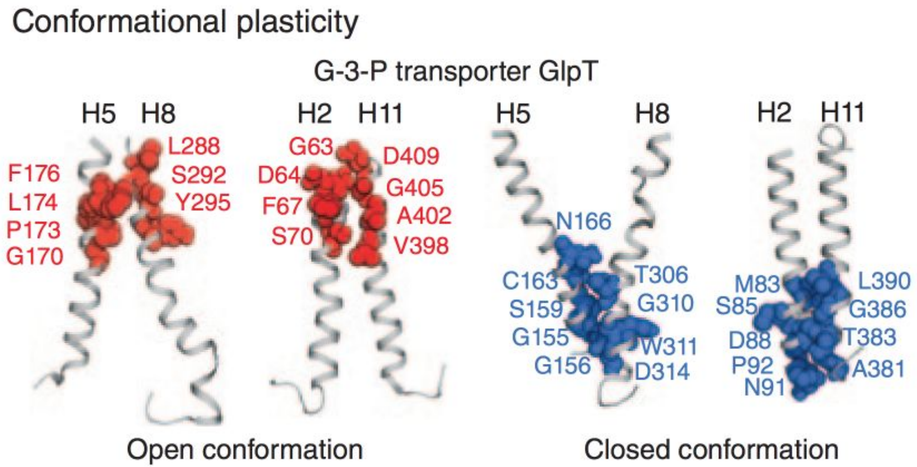




Marks (2012) Nat. Biotech.

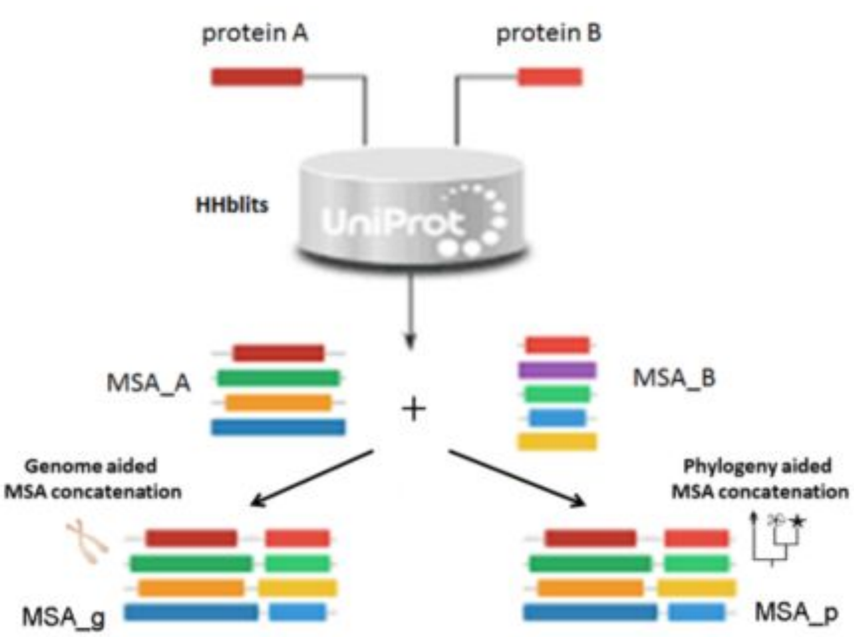
Predicting conformational changes

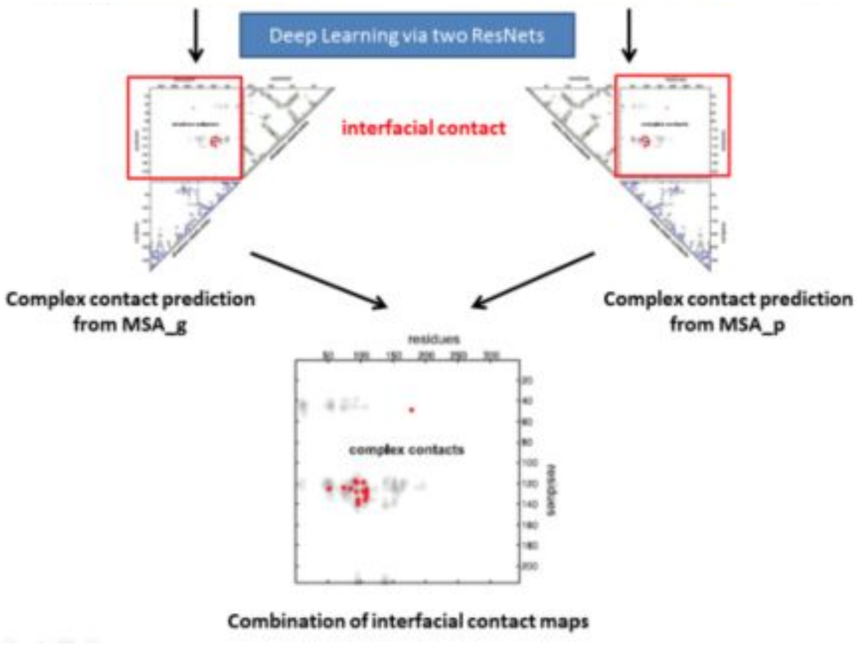




Marks (2012) Nat. Biotech.

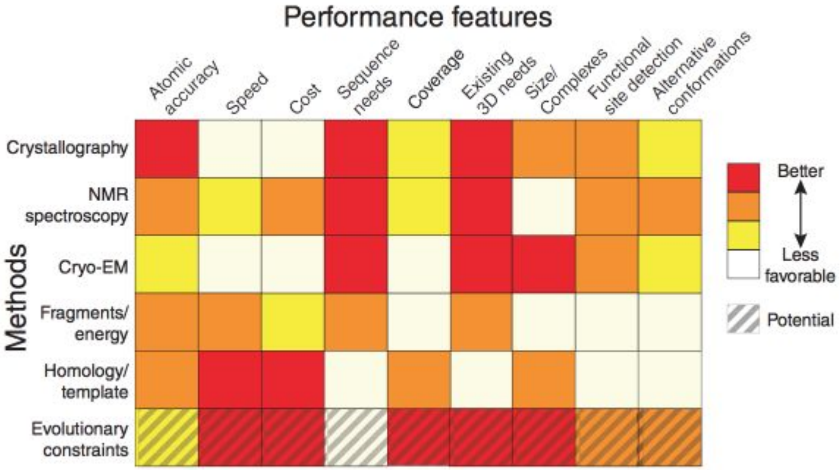
Predicting protein-protein interaction based on protein structure




Deep Learning via two ResNets
interfacial contact
Complex contact prediction
from MSA_g
Complex contact prediction
from MSA_p
residues
50
100
150
200
250
300
20
40
60
80
100
120
140
160
180
200
residues
complex contacts
Combination of interfacial contact maps

Zhou (2017) bioRxiv

Hybrid approaches for determining protein 3D structure

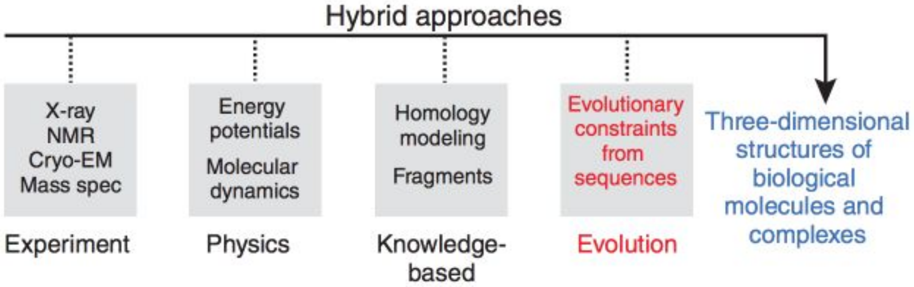


Performance features

Methods	Atomic accuracy	Speed	Cost	Sequence needs	Coverage	Existing 3D needs	Size/ Complexes	Functional site detection	Alternative conformations
Crystallography	🟥	⬜	⬜	🟥	🟨	🟥	🟧	🟧	🟨
NMR spectroscopy	🟧	🟨	🟧	🟥	🟨	🟥	⬜	🟧	🟧
Cryo-EM	🟨	⬜	⬜	🟥	⬜	🟥	🟥	🟧	🟨
Fragments/ energy	🟧	🟧	🟨	🟧	⬜	🟧	⬜	⬜	⬜
Homology/ template	🟧	🟥	🟥	⬜	🟧	⬜	🟧	⬜	⬜
Evolutionary constraints	🟨▨	🟥▨	🟥▨	⬜▨	🟥▨	🟥▨	🟥▨	🟧▨	🟧▨

🟥 Better
🟧 ↕
🟨
⬜ Less favorable

▨ Potential



Hybrid approaches

X-ray
NMR
Cryo-EM
Mass spec

Experiment

Energy potentials
Molecular dynamics

Physics

Homology modeling
Fragments

Knowledge-based

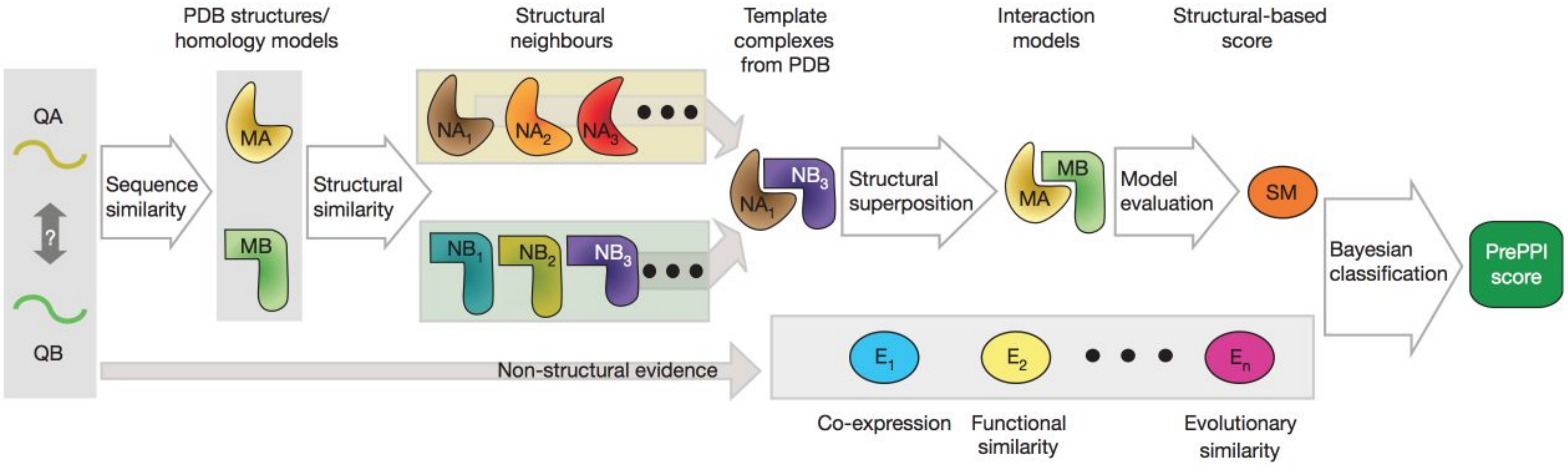
Evolutionary constraints from sequences

Evolution

Three-dimensional structures of biological molecules and complexes

Marks (2012) Nat. Biotech.

Predicting protein-protein interaction based on protein structure



Zhang (2012) Nature

Molecular dynamics (MD) simulations = Computational microscope

Cocco et al. (2013) From principal component to direct coupling analysis of coevolution in proteins: low eigenvalue modes are needed for structure prediction. PLoS Comput Biol.

This well-written paper introduces a Hopfield–Potts model for interpolating between principal component analysis, which identifies the most correlated residues, and direct coupling analysis, which aims at predicting residue–residue contacts based on the maximum entropy principle. This is an excellent read for better understanding both the distinctions between methods and the mathematics underlying covariance analysis of multiple-sequence alignments.

Hopf et al. (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell.

The authors use residue covariation to predict previously unknown 3D structures for 11 transmembrane proteins from sequence alone. The unprecedented accuracy of such predictions was confirmed through de novo computation of transmembrane proteins of known structure from 23 families.

Ovchinnikov et al. (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. Elife.

This paper describes de novo blind structure predictions of unprecedented accuracy for two proteins using a combination of residue–residue co-evolutionary information and the Rosetta structure prediction program. The authors applied this approach to generate structural models for 58 prokaryotic protein families lacking 3D structures, examination of which led to mechanistic and functional hypotheses.

Tang et al. (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Meth.

This study demonstrates how residue–residue covariance information can complement NMR data for determining protein structures. The authors provide a detailed description of how to apply this approach.

Dror (2012) Annu. Rev. Biophys.