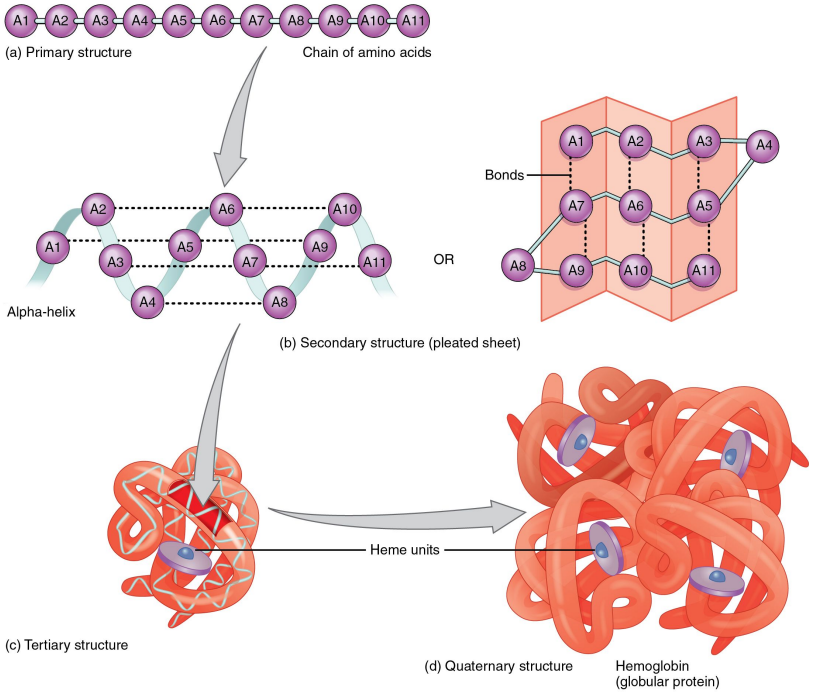
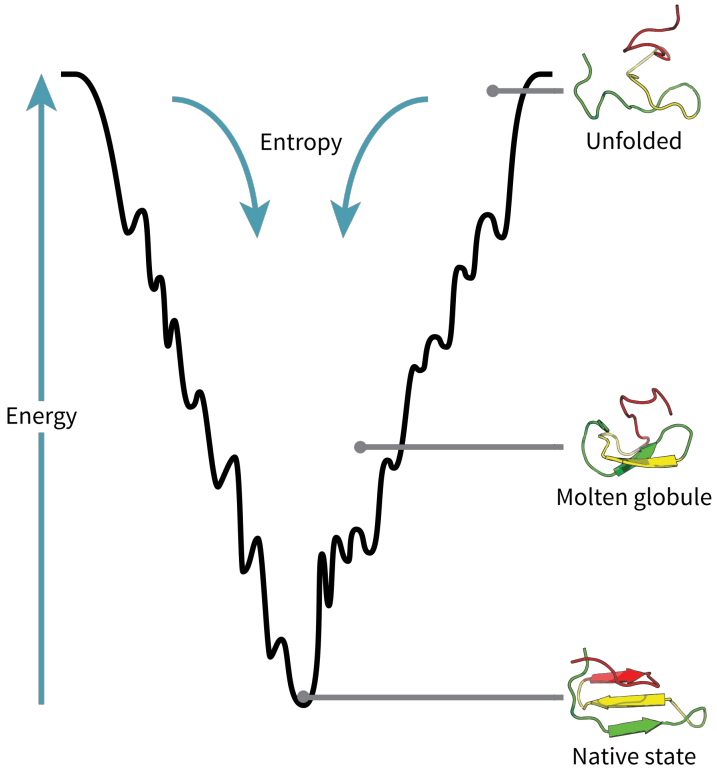
Protein structure

- Amino-acid coevolution
  - Mutual information
  - Maximum entropy modeling
- Molecular dynamics

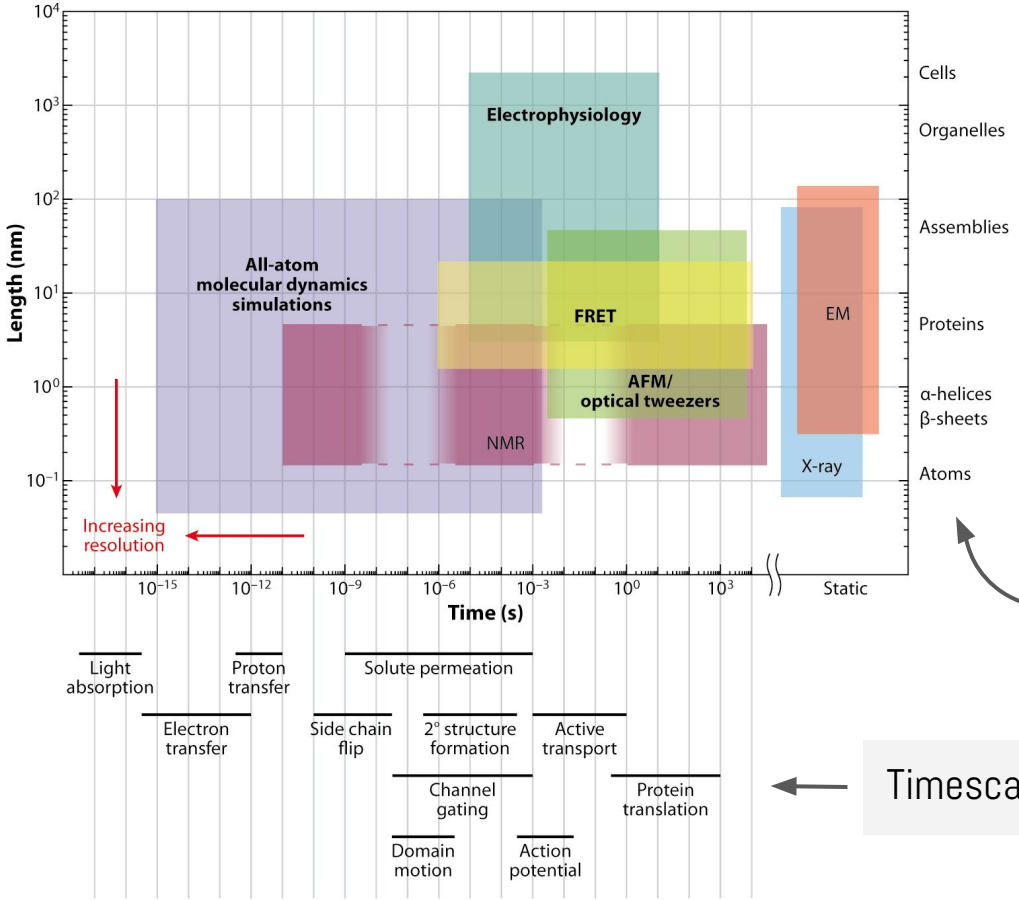
Proteins have 3D structures that are closely tied to their function





Wikipedia

Various experimental techniques to determine protein 3D structure



Data on single molecules (as opposed to only on ensembles) are in boldface.

- AFM, atomic force microscopy
- EM, electron microscopy
- FRET, Forster resonance energy transfer
- NMR, nuclear magnetic resonance

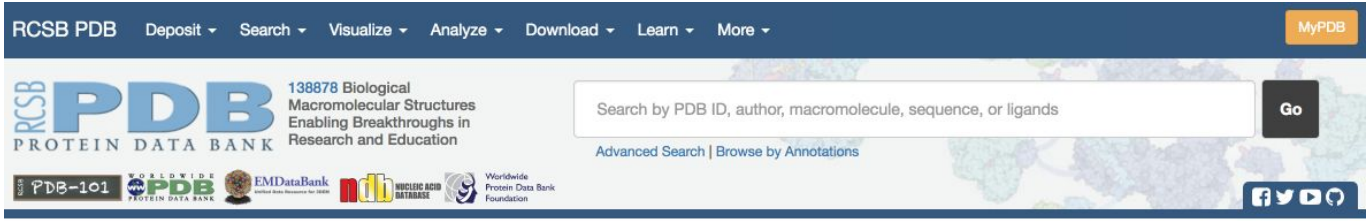
Spatial resolution of biological features

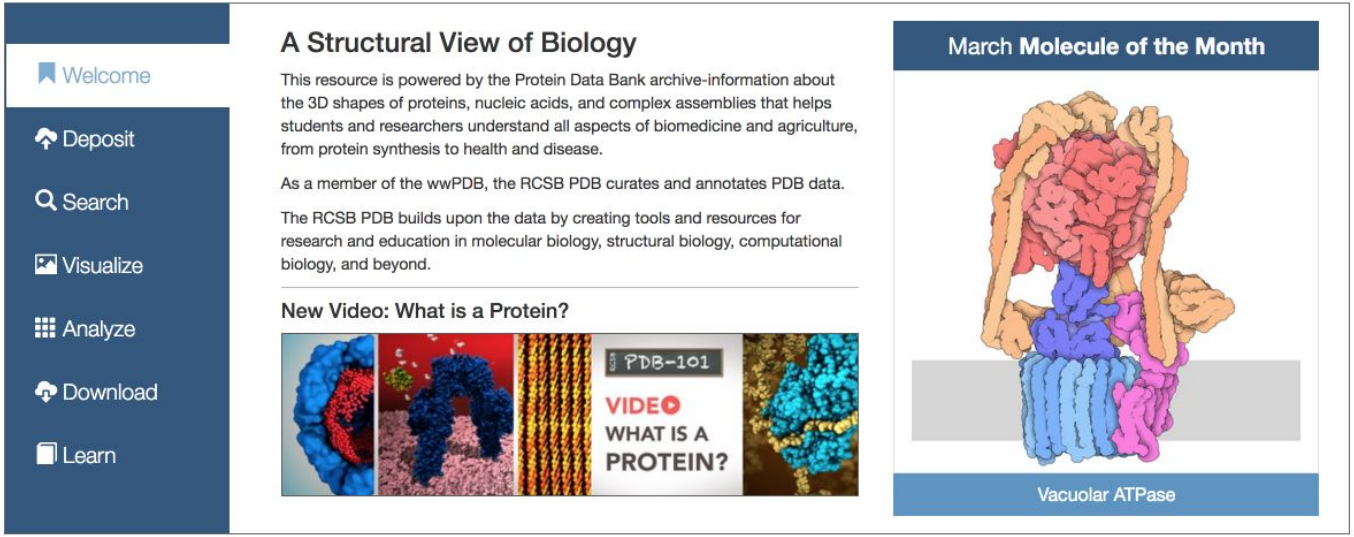
Timescales of molecular processes

Dror (2012) Annu. Rev. Biophys.

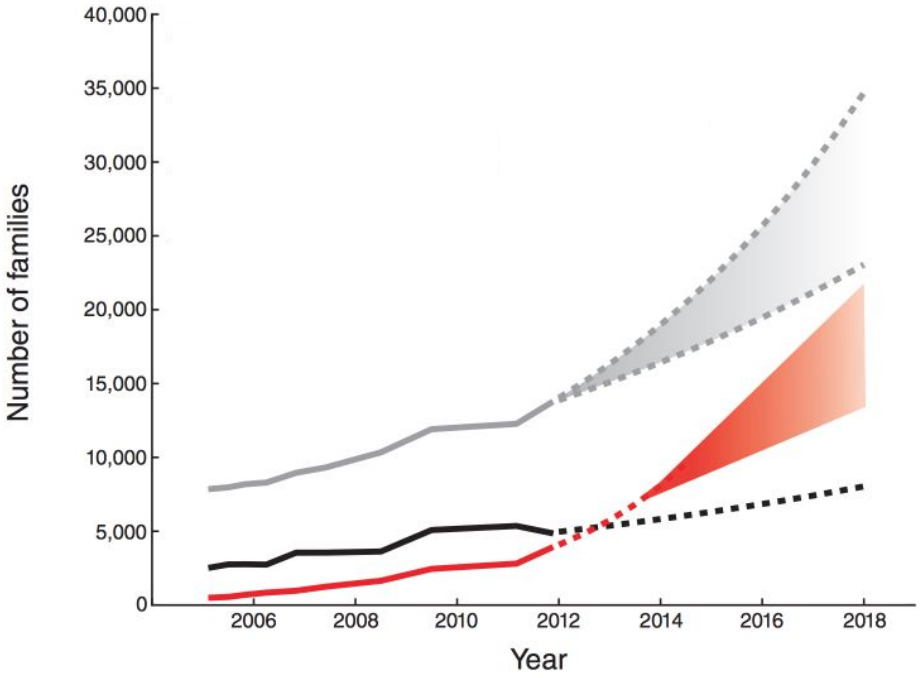
Protein Data Bank (PDB)

www.rcsb.org: 3D shapes of proteins, nucleic acids, and complex assemblies.





Experimental methods for 3D structure determination



Huge 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).

Experimental structure-determination

- Done one-by-one

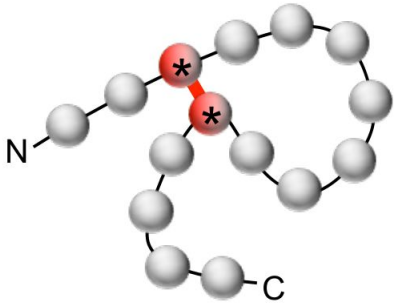
Marks (2012) Nat. Biotech.

Need computational methods to predict structure from sequence

1. **Physics-based** methods (*in silico* energy functions)
   a. Only works for small proteins *de novo*.
   b. Needs massive infrastructure
2. **Knowledge-based** (sequence similarity to proteins with known structures; homology modeling)
   a. Only works for small proteins *de novo*.
   b. This is true even with fragments.
3. **Co-evolution-based** prediction of potential interactions between residues from sequence
   a. Takes advantage of a billion-year dataset.
   b. Applicable for a variety of proteins.
   c. Can be carried out very fast.

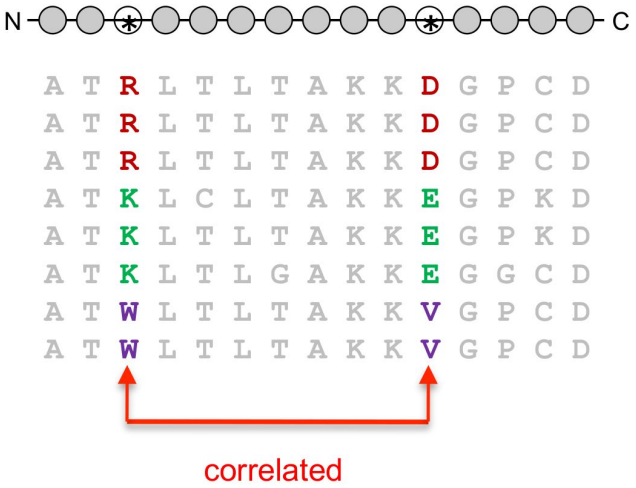
Incorporating interactions into the generative model

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



contact in 3D

Visible record of residue covariation in related protein sequences.



We also know that individual mutations don't add up: `f(A) + f(B) ≠ f(AB)`

- Higher order interactions, Pervasive epistasis, Context dependence

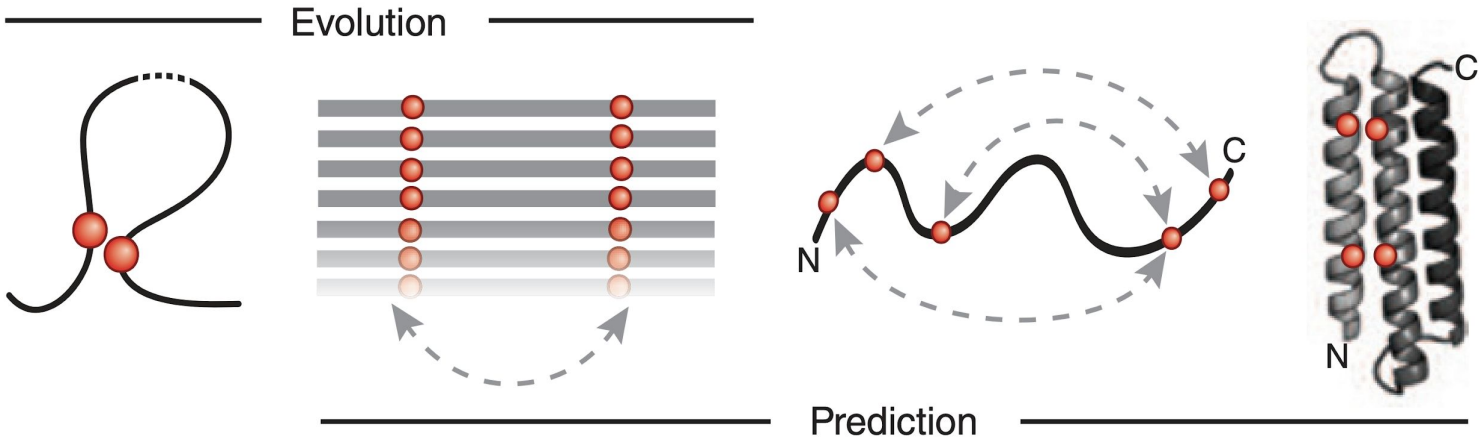
Marks (2012) Nat. Biotech.

Incorporating interactions into the generative model

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

→

Visible record of residue covariation in related protein sequences.

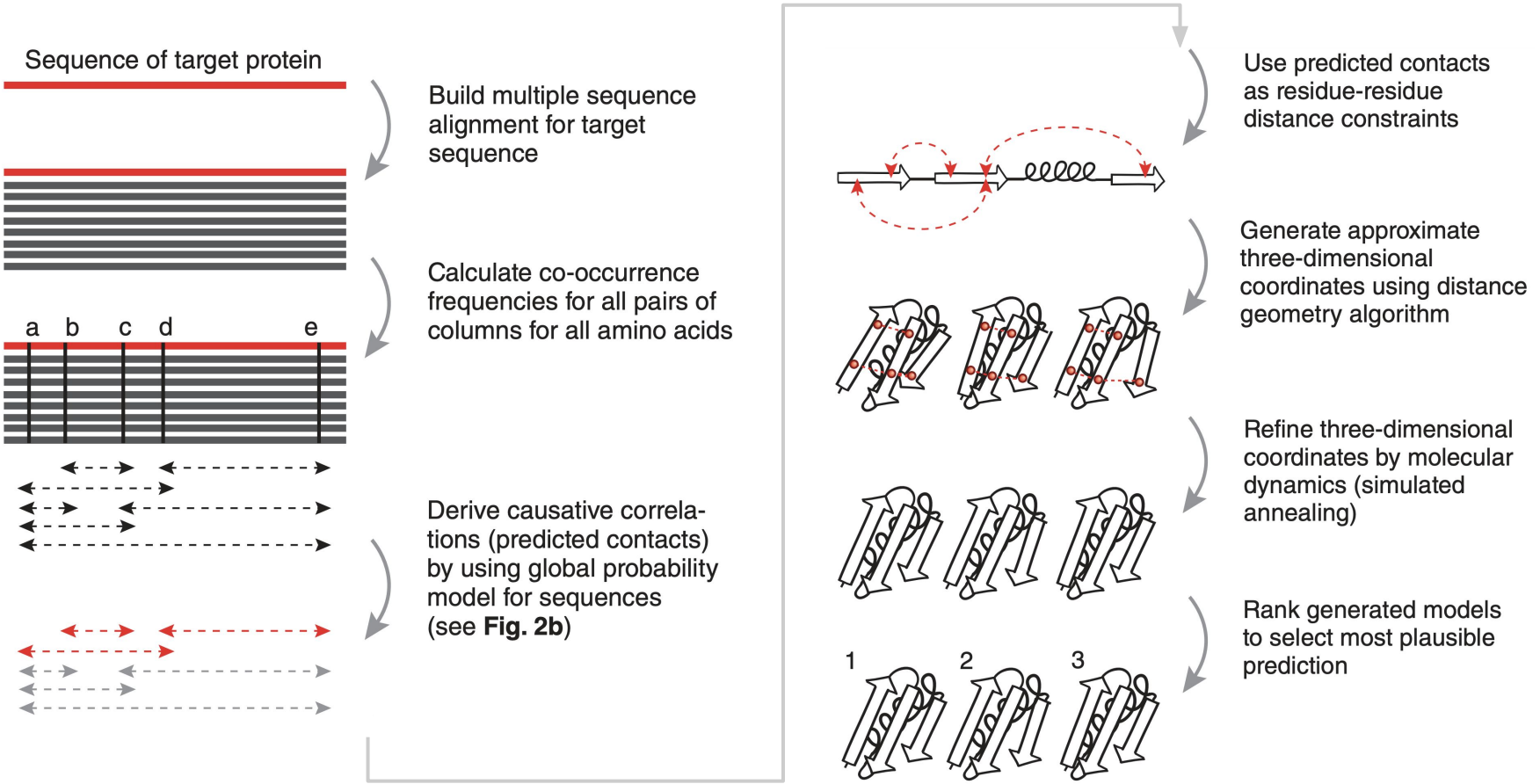


We also know that individual mutations don't add up: `f(A) + f(B) ≠ f(AB)`

- Higher order interactions, Pervasive epistasis, Context dependence

Marks (2012) Nat. Biotech.

Predicting protein 3D structure based on evolutionary coupling

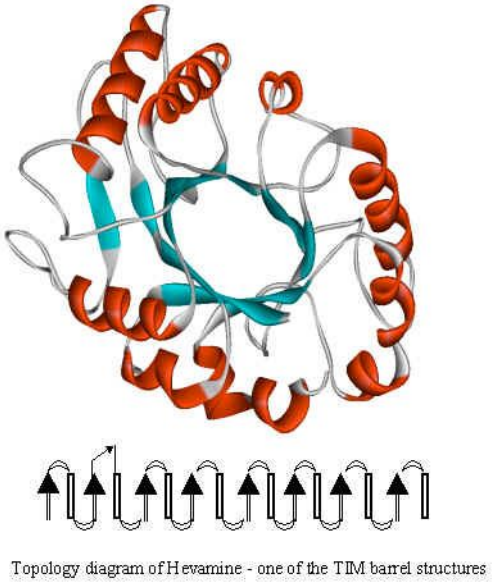


Marks (2012) Nat. Biotech.

Finding potential interactions between residues from sequence

There are many ways to make a protein.

- We know this based on sequence and structural diversity.
- Just need to satisfy high-level structural properties and can be made up of a surprisingly different combinations of amino acids.
  - For instance, TIM barrels can be as little as 8% similar in sequence.



Topology diagram of Hevamine - one of the TIM barrel structures

Finding potential interactions between residues from sequence

Evolution explores mutation space.

Goal: Take all the observed sequences and use an unsupervised method to learn a distribution over sequence space and learn what is functional.

- Specifically, can we build **a generative model for sequences $P(a)$**?

For this to work, we have to assume that:

- The system has reached equilibrium, i.e., no more evolution.
- The states (AA at each position) are well-mixed, i.e. ignore factors such as phylogeny.

Classical generative model for sequences

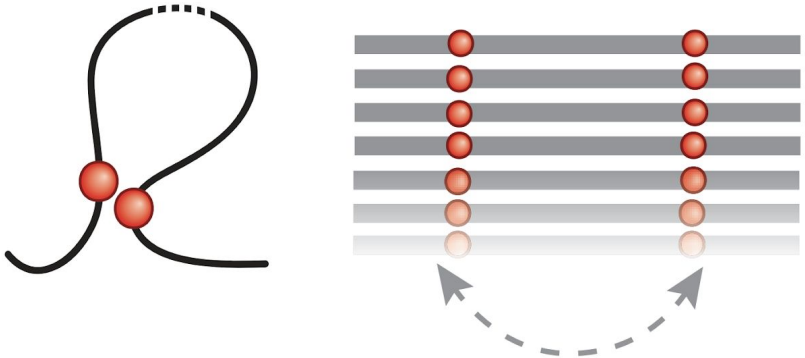
An “independent sites” model $P(a)$:

- Whole sequence = Product of individual site factors
  - $P(a) = p_1(a_1) \cdot p_2(a_2) \cdot \ldots \cdot p_N(a_N)$
- Conversion to log-sum
  - $P(a) = 1/Z \cdot \exp\left(\sum_i h_i(a_i) \right)$
    - Z is a normalization factor

Used in homology search and motif detection.

Incorporating interactions into the generative model

Evolutionary pressure to maintain favorable interactions b/w physically interacting amino acid residues in 3D. → Visible record of residue covariation in related protein sequences.



We also know that individual mutations don't add up: $f(A) + f(B) \neq f(AB)$

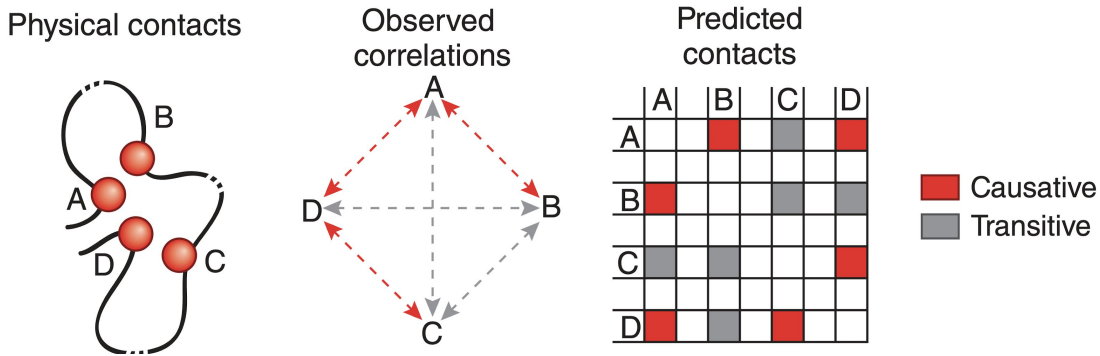
- Higher order interactions, Pervasive epistasis, Context dependence

Marks (2012) Nat. Biotech.

Capturing interactions using a global probability model

Build a global probability model that accounts for the fact that interactions along an entire protein chain are mutually interdependent in a way that is inherently cooperative.

- Pairwise interactions are modified by interactions with other parts of the system and cannot be factored (probabilities are not a simple product of independent terms).
- Markov random fields | Ising (Potts) model | Undirected graphical model.
- The essence of this model is to capture a **sparse description of interactions** that can give rise to **dense correlations**.



Marks (2012) Nat. Biotech.

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.

- Each J_{ij} is a 20-by-20 matrix

Wikipedia

Global probabilistic models of residue coupling (maximum-entropy)

Observation #1: Couplings reconstruct 3D structure.

- The parameters in this model have enough information to work out the structure: strongest couplings → contacts in crystal structure.

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

$h(a_i)$: 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.

- Each J_{ij} is a 20-by-20 matrix

Global probabilistic models of residue coupling (maximum-entropy)

Observation #2: `P(a)` captures context-dependent mutation effects.

- For a given position, we can ask, what is the sensitivity of mutation to each one of 20 AA in the spectrum of "Damaging" ⟷ "Neutral".
- Models that capture pairwise information are better at predicting sensitivity to single-site mutations.

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

$h(a_i)$: 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.

- Each J_{ij} is a 20-by-20 matrix

Global probabilistic models of residue coupling (maximum-entropy)

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

How do we fit `P(a)` when we can't compute `P(a)`?

- Let's just consider `Z`, the normalization constant.
- To calculate `Z`, we need to sum over the domain: 20^N sequence space (prohibitively large).

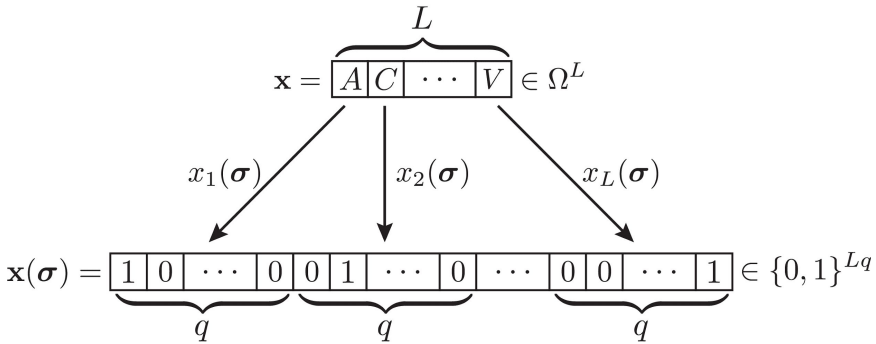
So, need approximate methods:

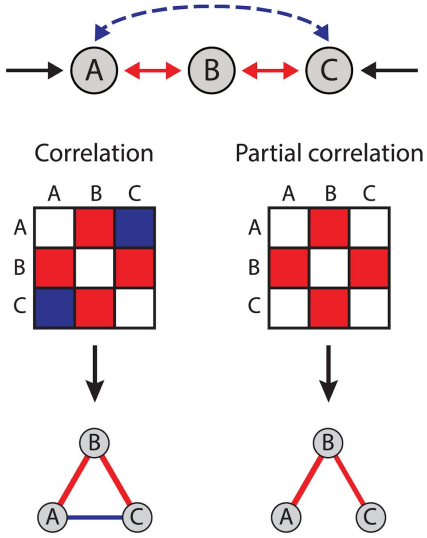
1. Continuous relaxation ≡ Partial Correlations
2. Pseudolikelihood ≡ Logistic Regression

Global probabilistic models of residue coupling (maximum-entropy)

Method 1: Continuous relaxation ≡ Partial Correlations

- If the input is provided using a one-hot encoding, fit a continuous model over this discrete data (essentially, allowing variables to take values like `0.5` instead of strictly `{0, 1}`), then…
- $\mathbf{J} \approx -\Sigma^{-1}$; That is, the inverse covariance of binary-encoded sequences approximates the couplings!





Global probabilistic models of residue coupling (maximum-entropy)

Method 2: Pseudolikelihood ≡ Logistic Regression

- Since the sequence space is huge, factorize the likelihood function (pseudo likelihood).
- Perhaps by doing some pseudo-factorization into sites.
  - Look at one position in the sequence,
  - Freeze all the other positions, and
  - Ask “conditioned on all the other positions, what is the AA in my position going to be?”

Global probabilistic models of residue coupling (maximum-entropy)

Method 2: Pseudolikelihood ≡ Logistic Regression

- This pseudo-factorization into sites can be expressed as a conditional probability:
  - $P(a_i \mid a \setminus a_i)$
    - This captures coupling of position i to background.
- Then, string these together to get:
  - $P(a) \approx P(a_1 \mid a \setminus a_1) \cdot \ldots \cdot P(a_i \mid a \setminus a_i) \cdot \ldots \cdot P(a_N \mid a \setminus a_N)$
- This function can be used for inference.
- This pseudolikelihood maximization takes the problem from $O(20^N) \rightarrow O(|A| \cdot N^2)$

Global probabilistic models of residue coupling (maximum-entropy)

- Even with the pseudolikelihood, there are way too many parameters to fit compared to the number of sequences:
- For e.g., even for a protein of length 200aa, the number of parameters ≈ 8 million:
  - No. of pairs ~ $N^2/2$
  - No. of parameters for each site = 400
  - No. of singles = $20N$
  - So, total ~ $400N^2 + 20N$, which is >> no. of sequences.
- So, to avoid overfitting, we need **regularization**.
  - Typically L2-regularization (with a Gaussian prior) is used.

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)

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

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

1

$$J'_{ij} = J_{ij}(k,l) - J_{ij}(\cdot,l) - J_{ij}(k,\cdot) + J_{ij}(\cdot,\cdot)$$

2

$$F_{ij} = \left\| J'_{ij} \right\| = \sqrt{\sum_k \sum_l J'^2_{ij}(k,l)}$$

3

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

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

After fitting the model, how do we go from the J_{ij} values (each is a 20-by-20 matrix) to evolutionary couplings?

Summarize the matrix J_{ij} into a single number that captures the total epistatic constraint:

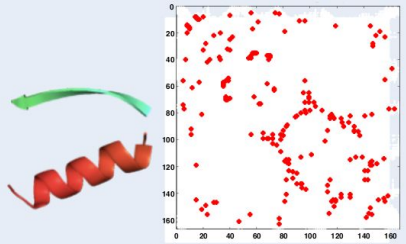
1. Shift into zero-sum gauge (minimizing the norm)
2. Calculate the Frobenius norm of the parameter (~ std. deviation).
3. Apply Average Product Correction (APC).

The F_{ij} values provide a ranked list of residue pairs predicted to be close in 3D space.

From contacts to structure

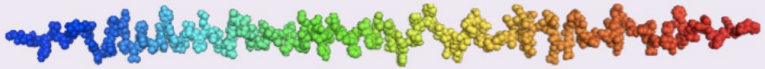
Use these F_{ij} values as distance constraints and fold the sequence into 3D structure.

```
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
```

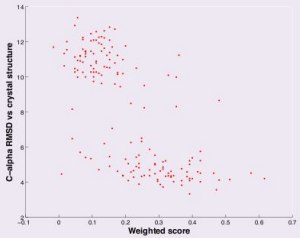


predicted contacts (EICs)

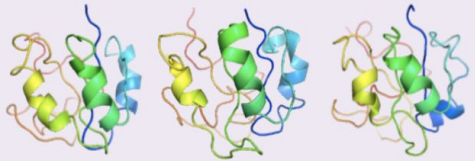
Start with extended structure
use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain



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



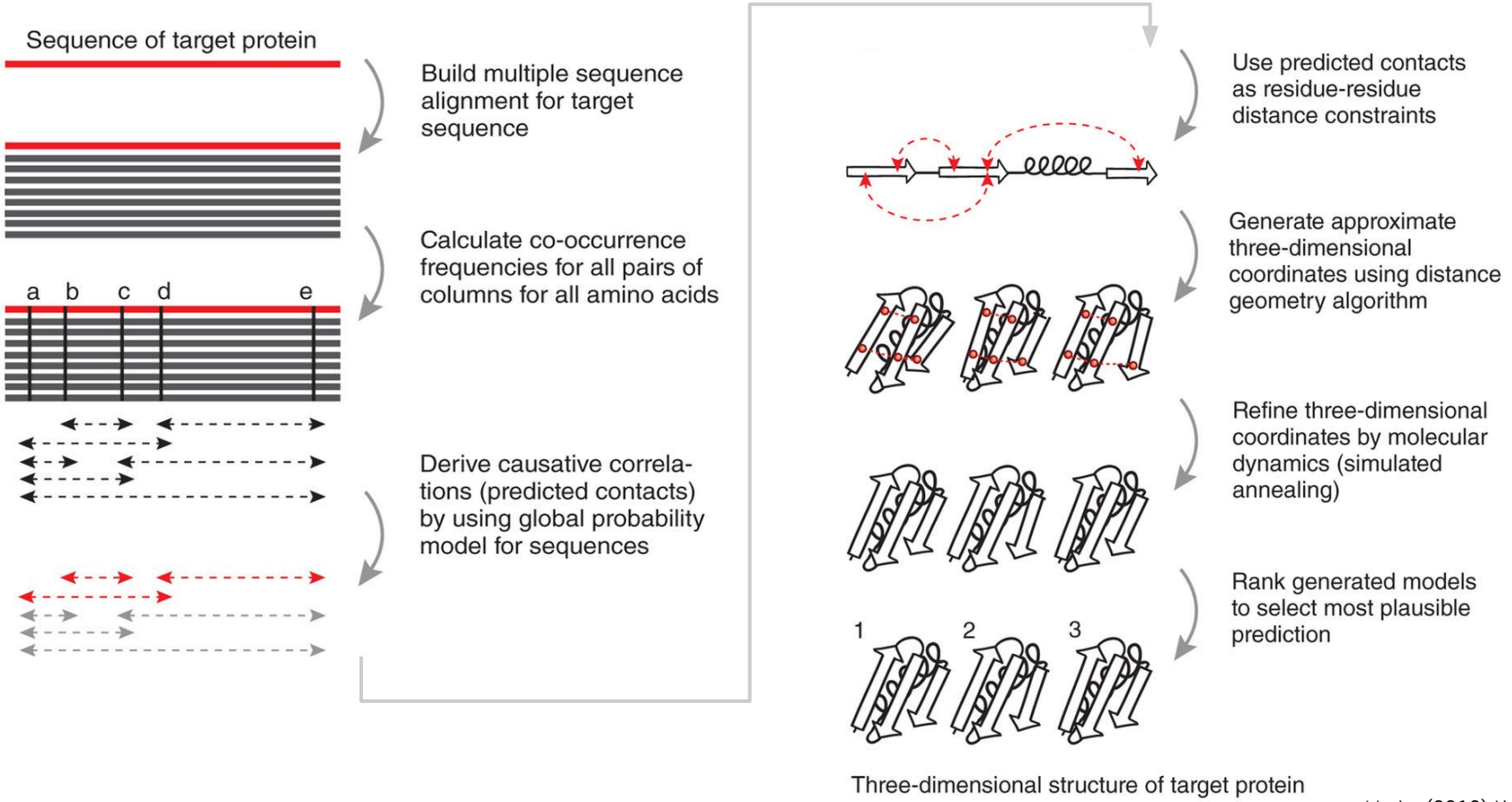
good scores



bad scores

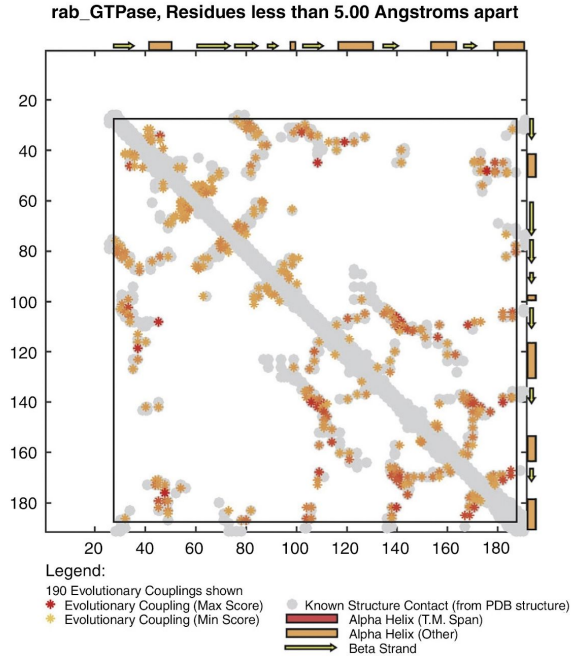
Marks (2011) PLoS One

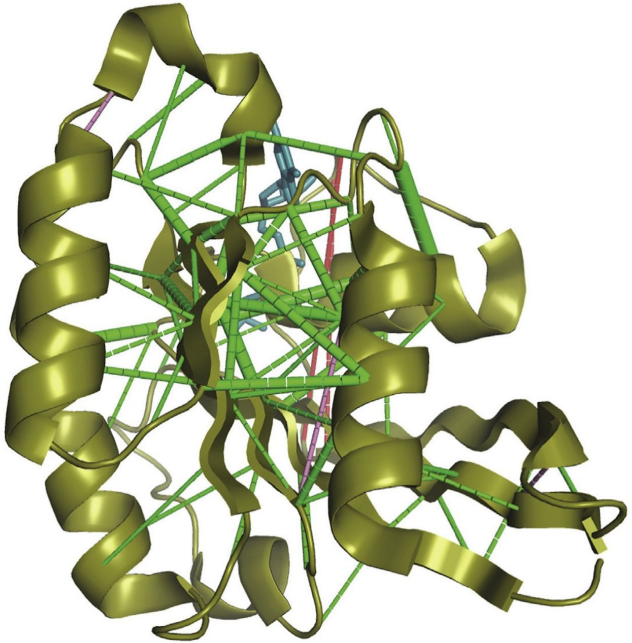
Predicting protein 3D structure based on evolutionary coupling



Marks (2012) Nat. Biotech.

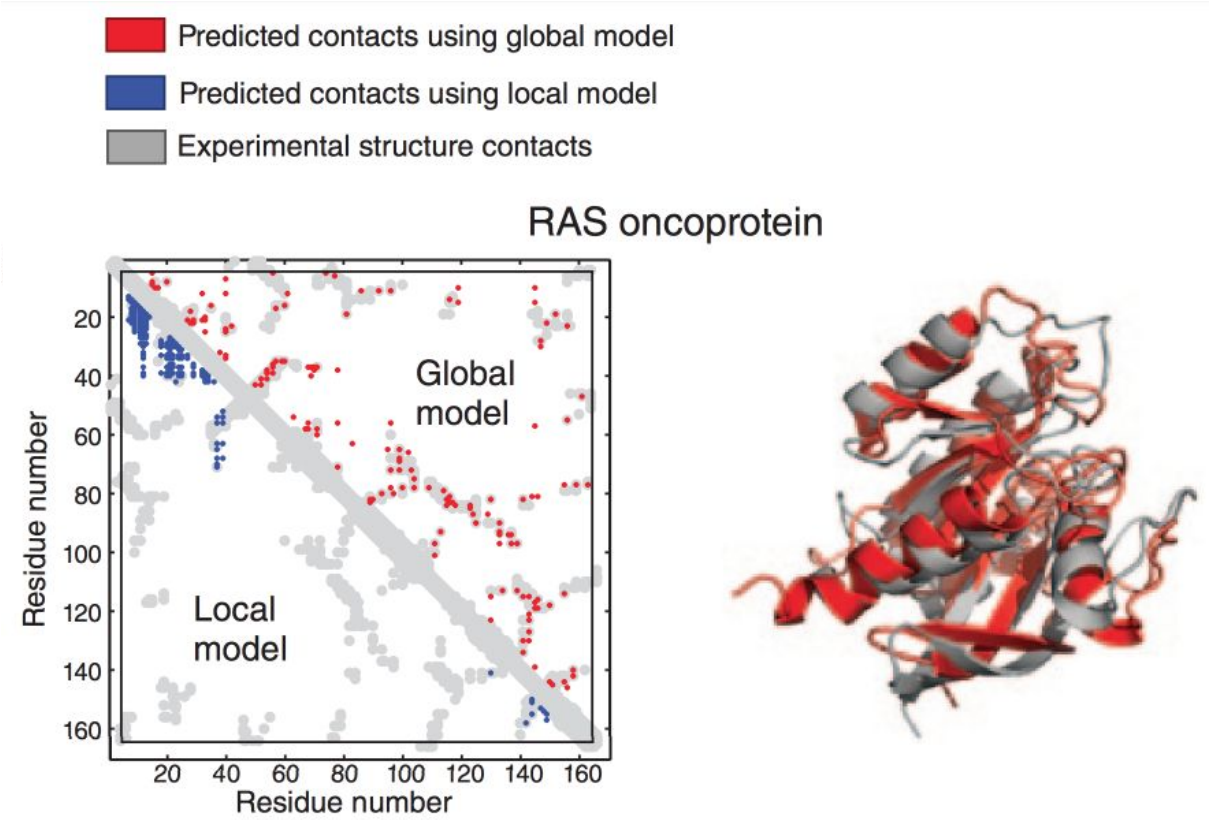
Predictions of 3D structures based on evolutionary coupling





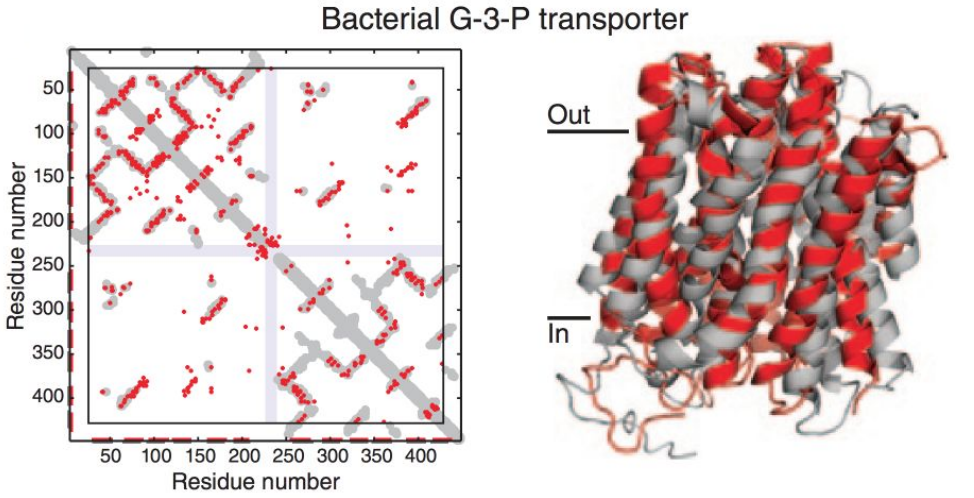
Newald (2016) Curr. Opin. Struct. Biol.

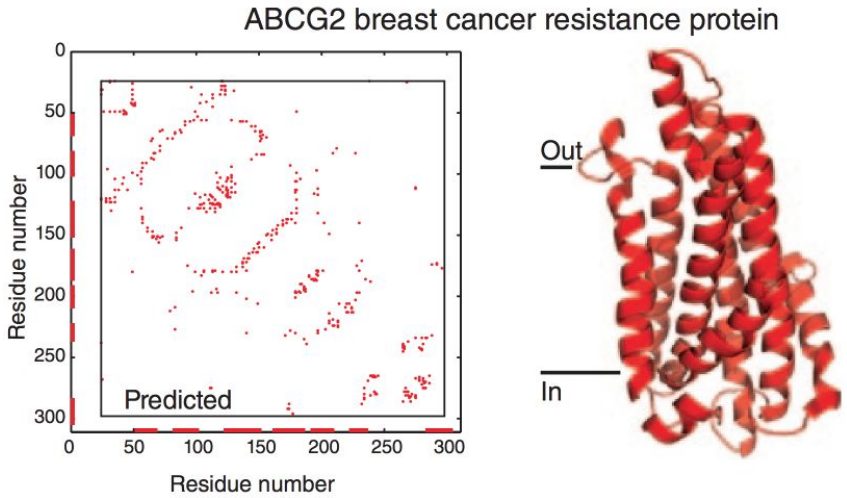
Predictions of 3D structures based on evolutionary coupling



Marks (2012) Nat. Biotech.

Predictions of 3D structures based on evolutionary coupling



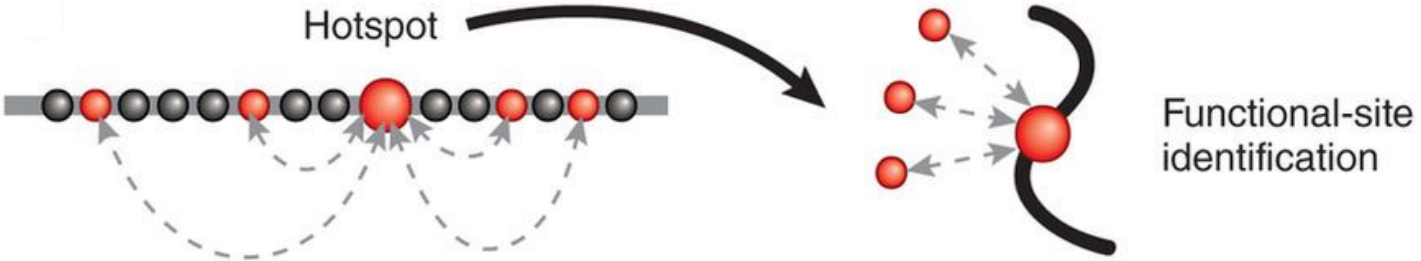


Marks (2012) Nat. Biotech.

Detecting functional hotspots

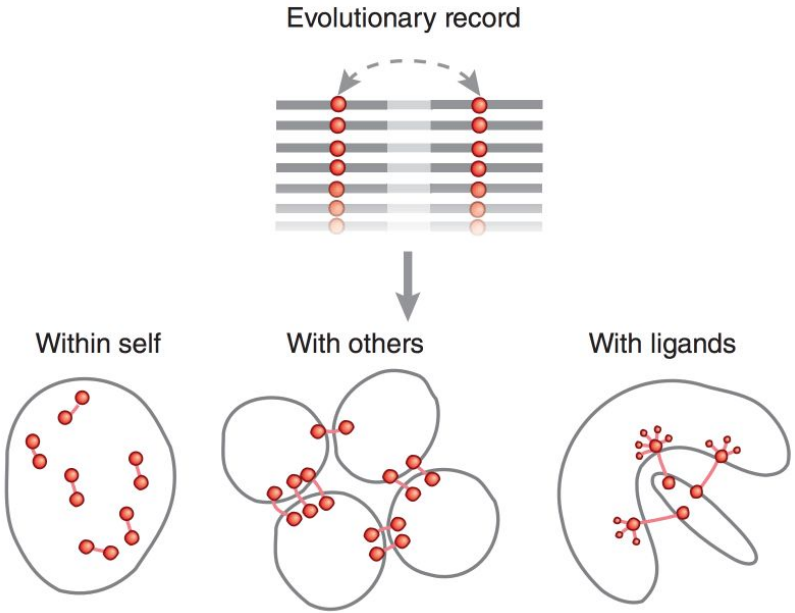
Residues subject to a high number of evolutionary pair constraints represent likely functional hotspots.

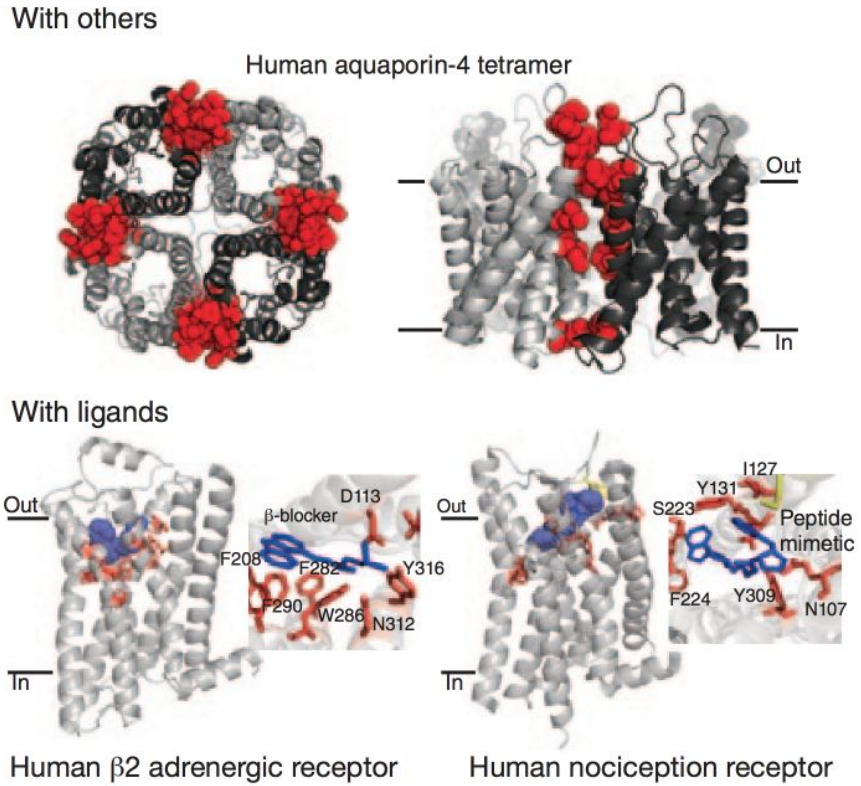
- Such highly constrained residues include residues in functional sites (for e.g., interaction with external ligands).
- Not detectable by analysis of single-residue conservation.



Marks (2012) Nat. Biotech.

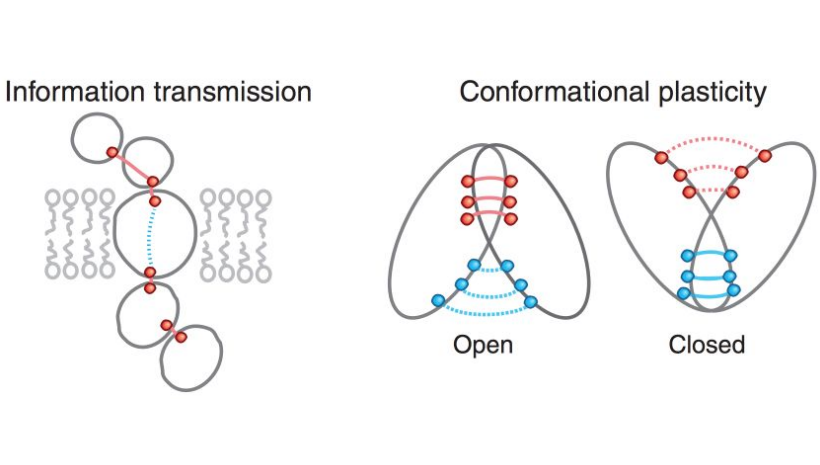
Predicting protein-protein & protein-ligand interactions





Marks (2012) Nat. Biotech.

Predicting conformational changes



Marks (2012) Nat. Biotech.