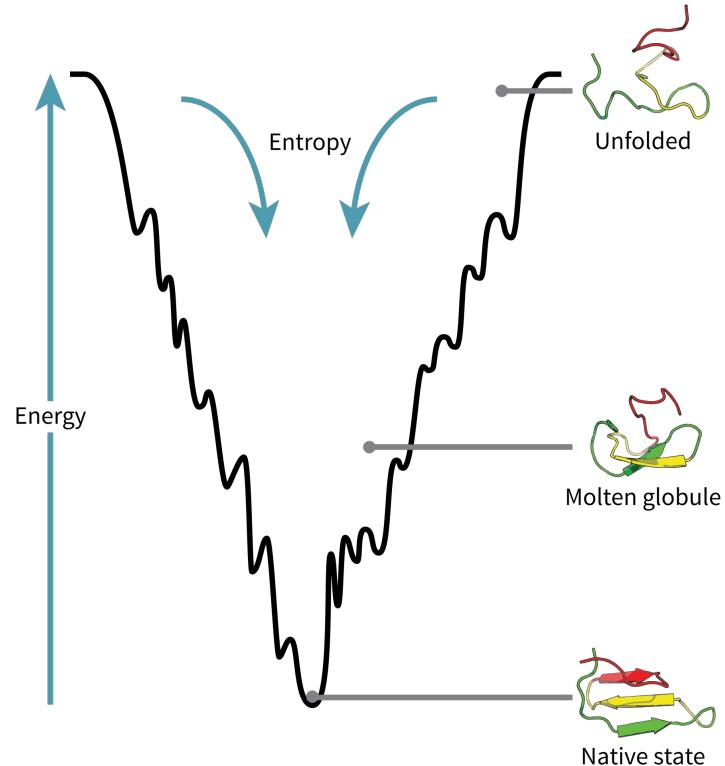
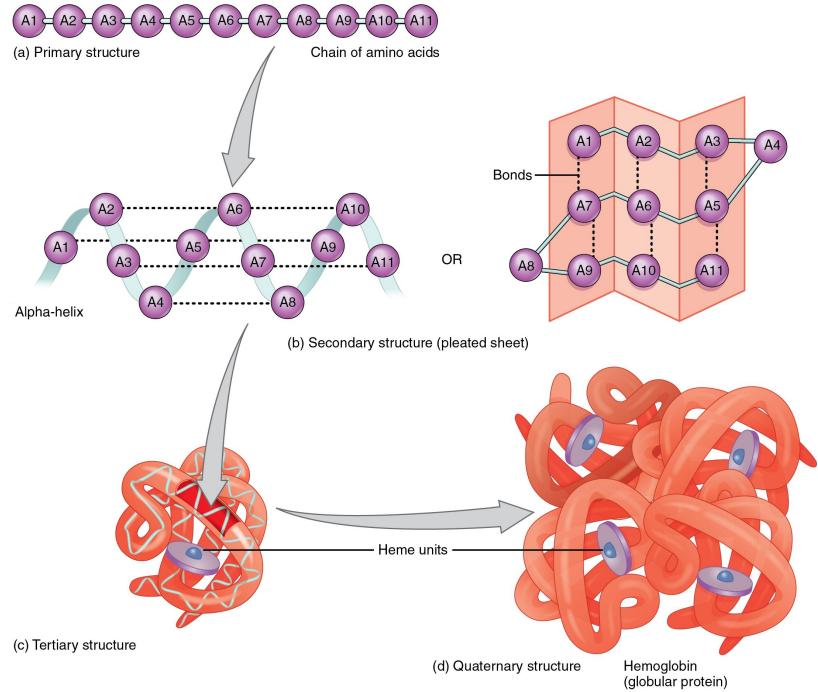


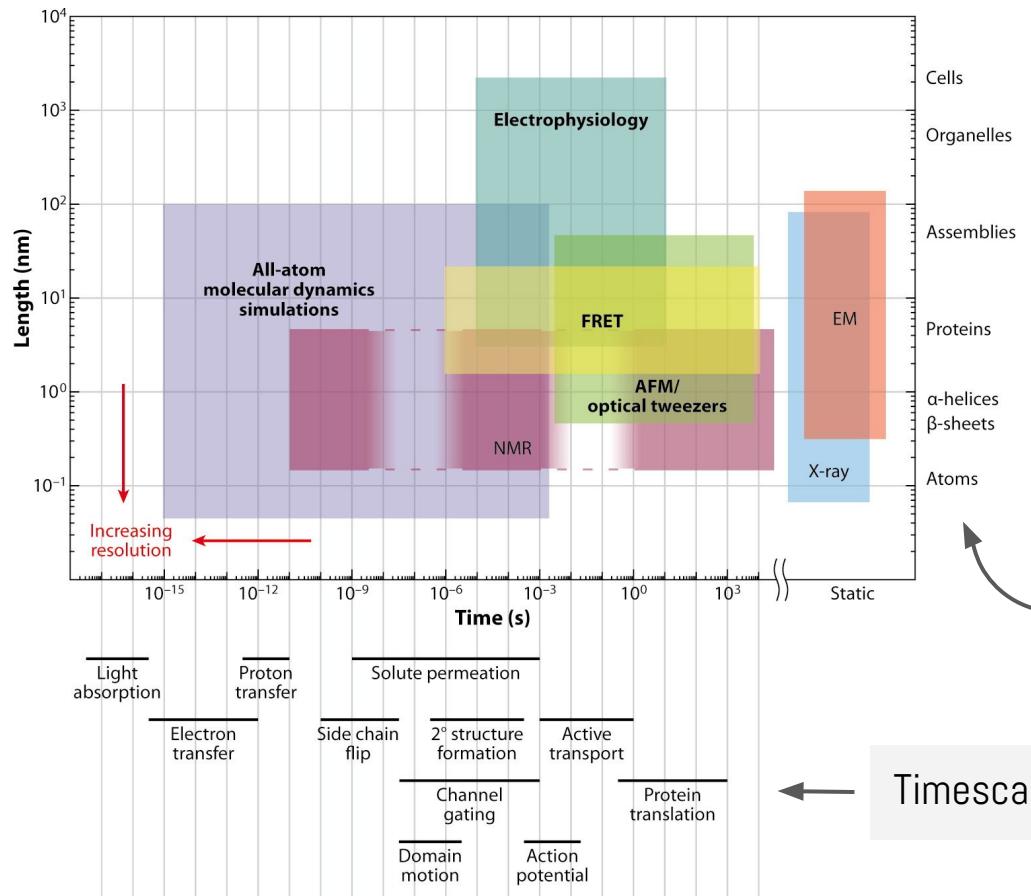
Protein structure

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

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



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

Protein Data Bank (PDB)

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

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

138878 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB-101 EMDDataBank Worldwide Protein Data Bank Foundation

Search by PDB ID, author, macromolecule, sequence, or ligands Go Advanced Search | Browse by Annotations

Facebook Twitter YouTube

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

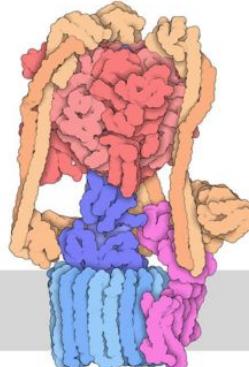
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

New Video: What is a Protein?



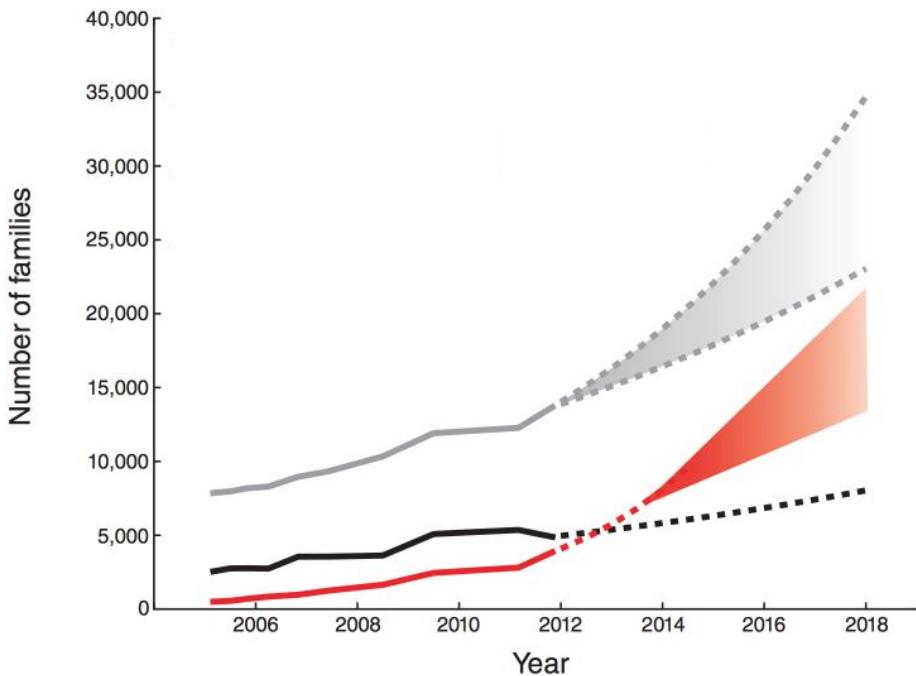
PDB-101 VIDEO WHAT IS A PROTEIN?

March Molecule of the Month



Vacuolar ATPase

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

Need computational methods to predict structure from sequence

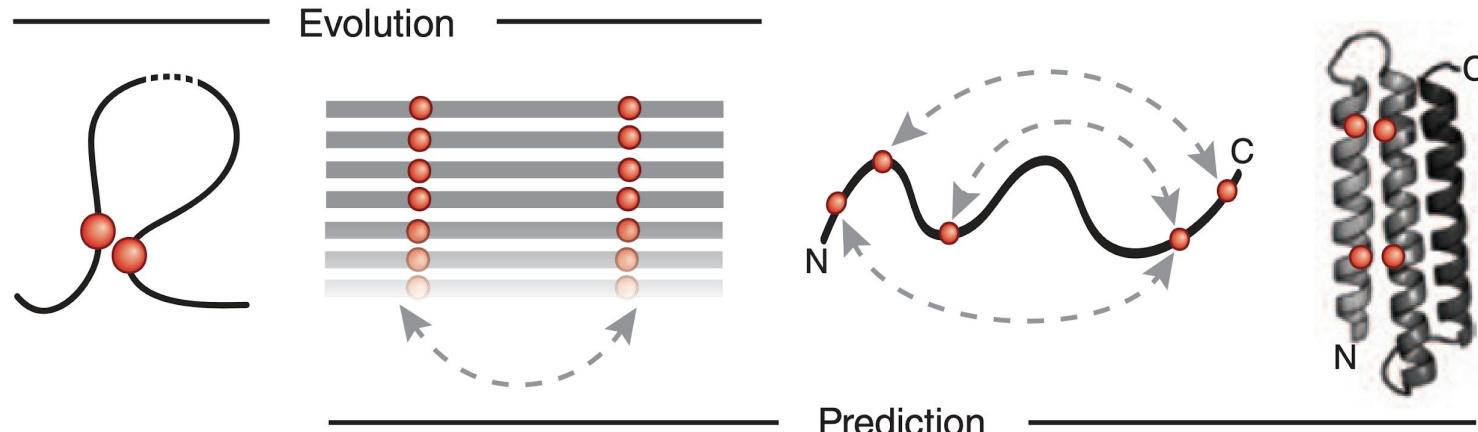
1. **Physics-based** methods (*in silico* energy functions)
 - a. Only works for small proteins *de novo*.
 - b. Needs massive infrastructure
2. **Knowledge-based** (sequence similarity to proteins with known structures; homology modeling)
 - a. Only works for small proteins *de novo*.
 - b. This is 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.



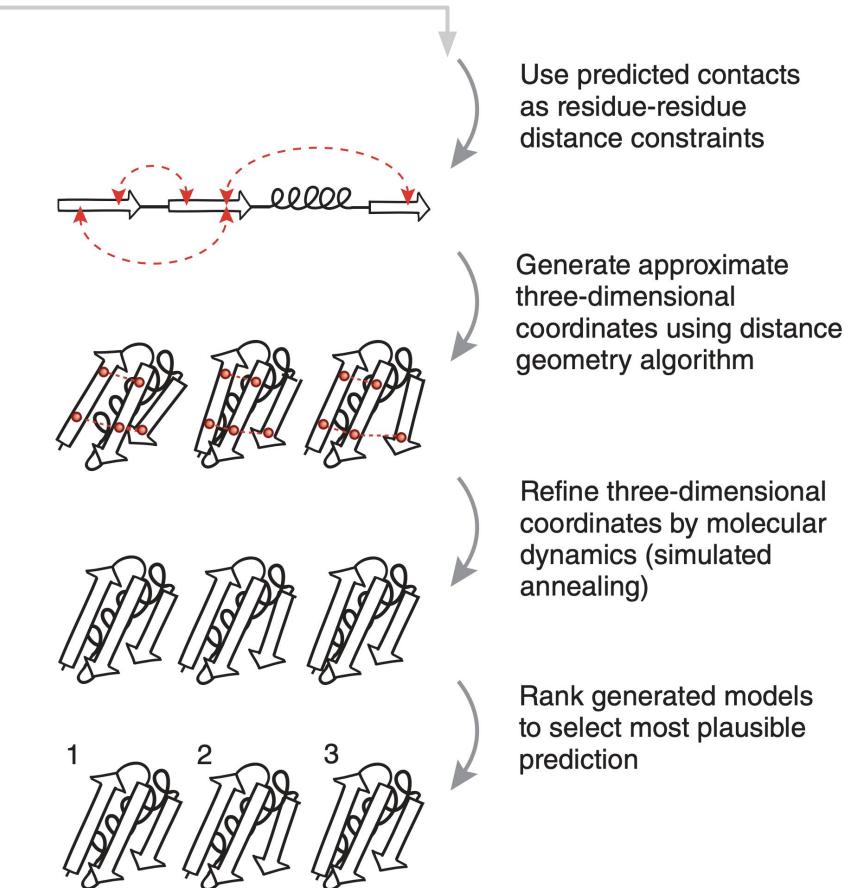
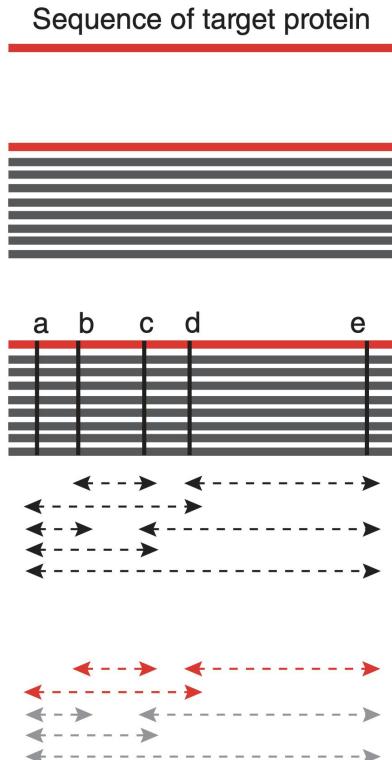
Visible record of residue covariation in related protein sequences.



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

- Higher order interactions, Pervasive epistasis, Context dependence

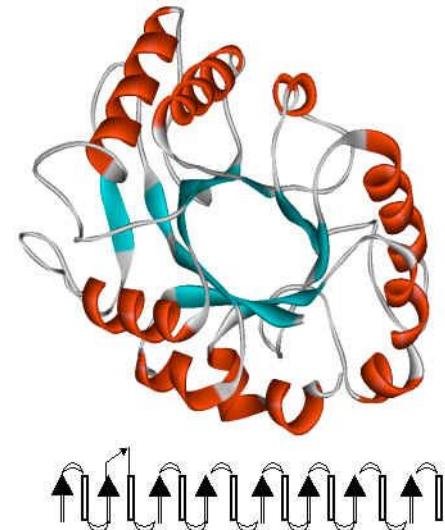
Predicting protein 3D structure based on evolutionary coupling



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 Hemamine - 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

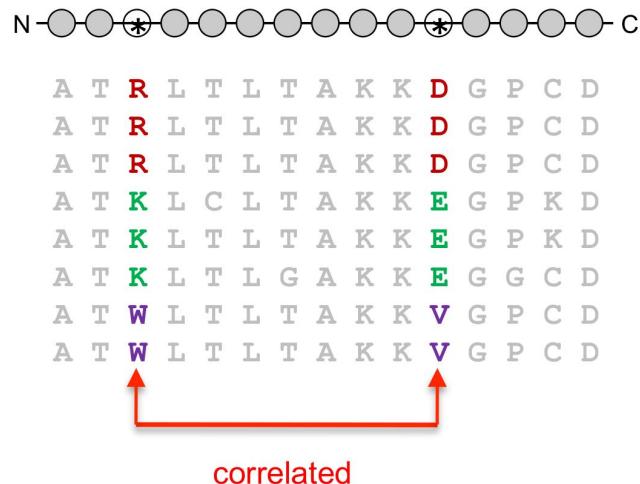
An "independent sites" model $P(a)$:

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

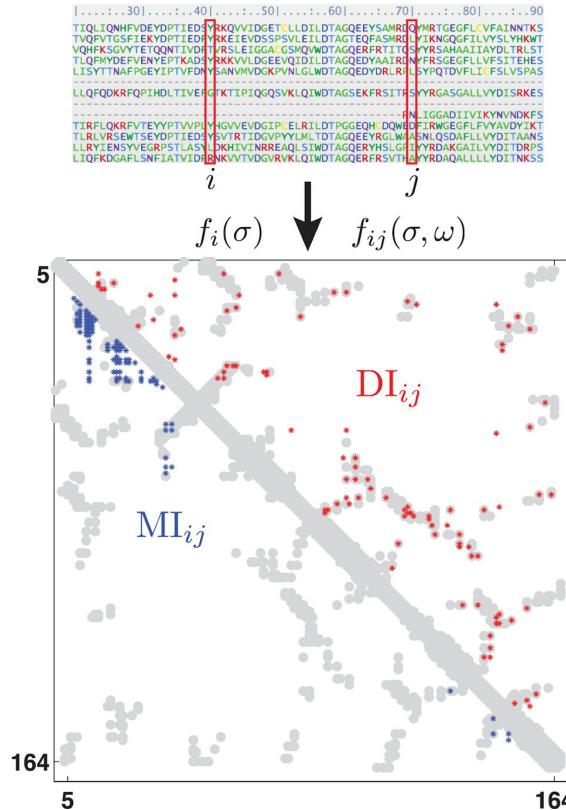
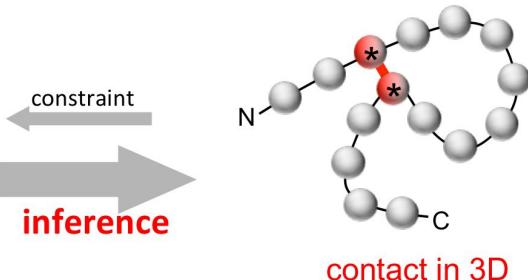
Used in homology search and motif detection.

Predicting protein 3D structure based on evolutionary coupling

Deep evolutionary sequence variation



Build a generative model



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

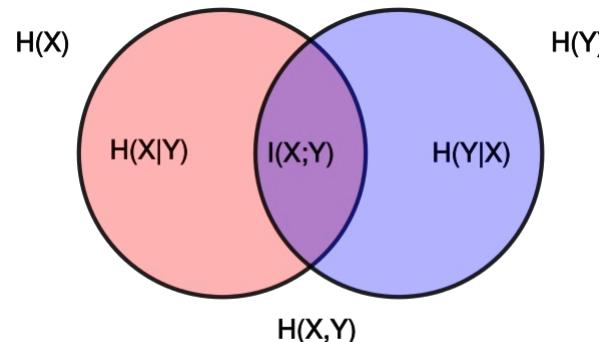
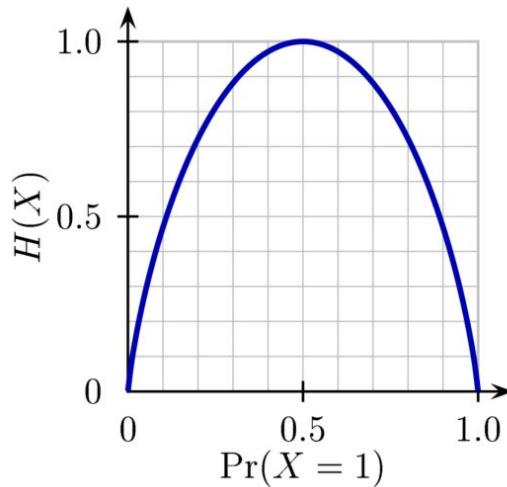
Capturing interactions based on mutual information

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

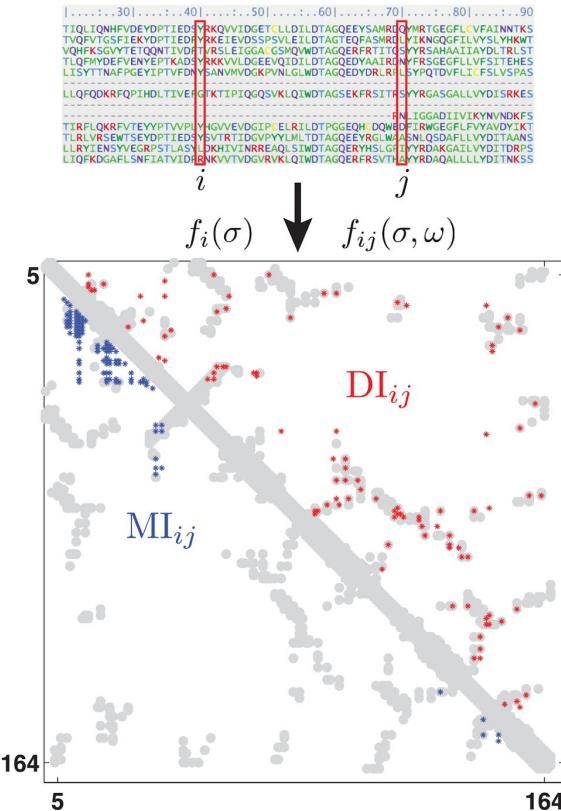
Mutual information: MI between two random variables $I(X, Y)$ quantifies the amount of information obtained about one random variable, through the other random variable.

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

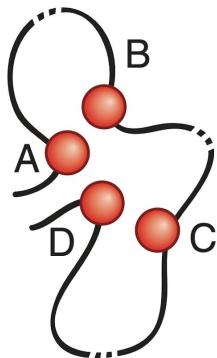
$$I(X; Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$



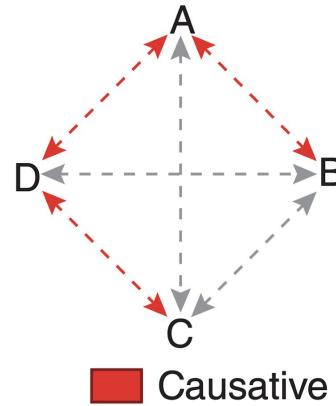
Capturing interactions based on mutual information



Physical contacts



Observed correlations



Predicted contacts

A	B	C	D
A			
B	Red		
C	Grey	Grey	Red
D	Red		Red

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.

- 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).
- 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**.

Global probabilistic models of residue coupling (maximum-entropy)

$a = (a_1, a_2 \dots, a_N)$ 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

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” \leftrightarrow “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, which is prohibitively large.

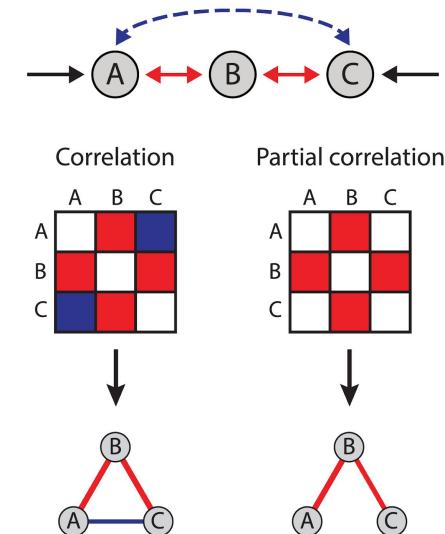
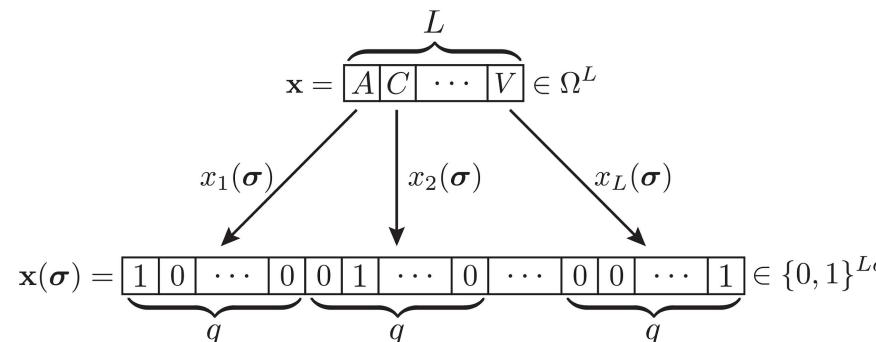
So, need approximate methods:

1. Continuous relaxation \equiv Partial Correlations
2. Pseudolikelihood \equiv Logistic Regression

Global probabilistic models of residue coupling (maximum-entropy)

Method 1: Continuous relaxation \equiv Partial Correlations

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



Global probabilistic models of residue coupling (maximum-entropy)

Method 2: Pseudolikelihood \equiv 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 \equiv Logistic Regression

- This pseudo-factorization into sites can be expressed as a conditional probability: $P(a_i | a \setminus a_i)$.
 - This captures coupling of position i to background.
- Then, string these together to get:
 - $P(a) \approx P(a_1 | a \setminus a_1) \cdot \dots \cdot P(a_i | a \setminus a_i) \cdot \dots \cdot P(a_N | a \setminus a_N)$
- 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 \approx 8 million:
 - No. of pairs $\sim N^2/2$
 - No. of parameters for each site = 400
 - No. of singles = 20N
 - So, total $\sim 400N^2 + 20N$, 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)

$$a = (a_1, a_2 \dots, a_N)$$

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

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.

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

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.

Global probabilistic models of residue coupling (maximum-entropy)

$$a = (a_1, a_2 \dots, a_N)$$

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

1

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

2

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

3

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

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

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:

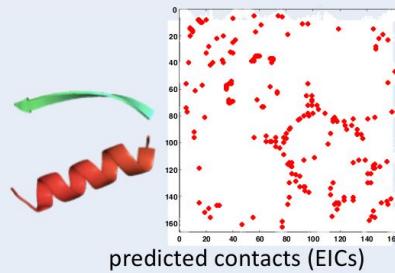
1. Shift into zero-sum gauge (minimizing the norm)
2. Calculate the Frobenius norm of the parameter (~ std. deviation).
3. Apply Average Product Correction (APC).

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

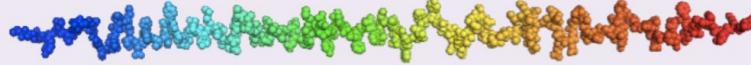
From contacts to structure

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

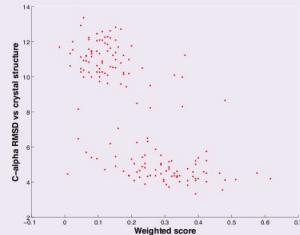
```
assign (resid 143 and name CA) (resid 123 and name CA) 4 4 3  
assign (resid 16 and name CA) (resid 10 and name CA) 4 4 3  
assign (resid 141 and name CA) (resid 82 and name CA) 4 4 3  
assign (resid 129 and name CA) (resid 87 and name CA) 4 4 3  
assign (resid 92 and name CA) (resid 11 and name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



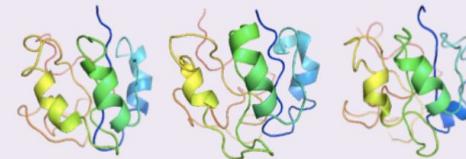
Start with extended structure
use **distance geometry** and **simulated annealing** with predicted constraints, EICs, to fold the chain



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



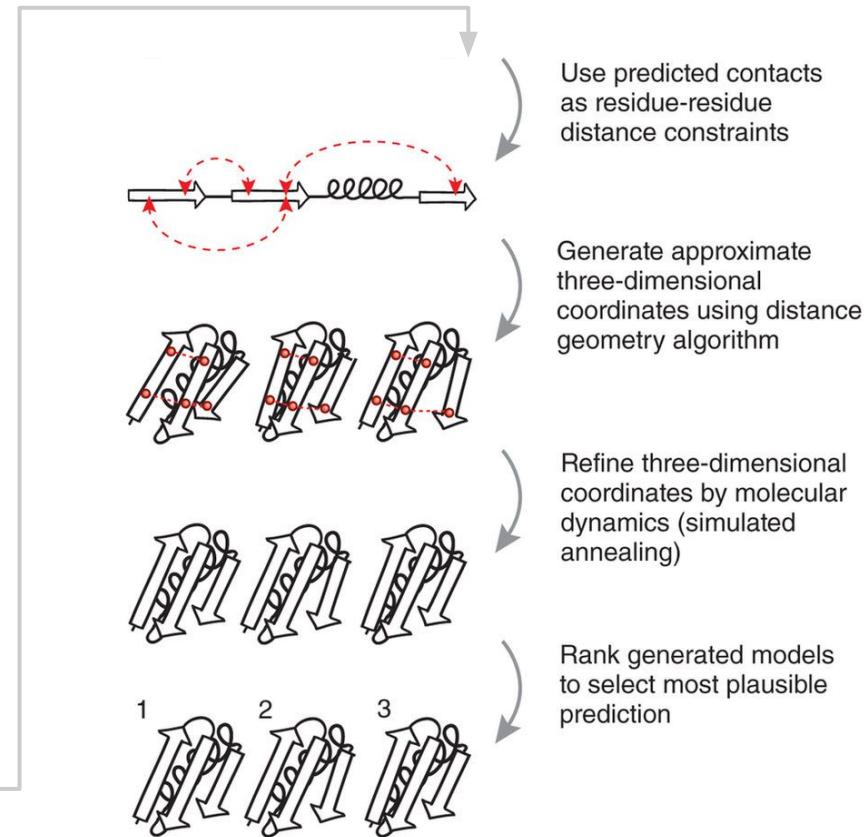
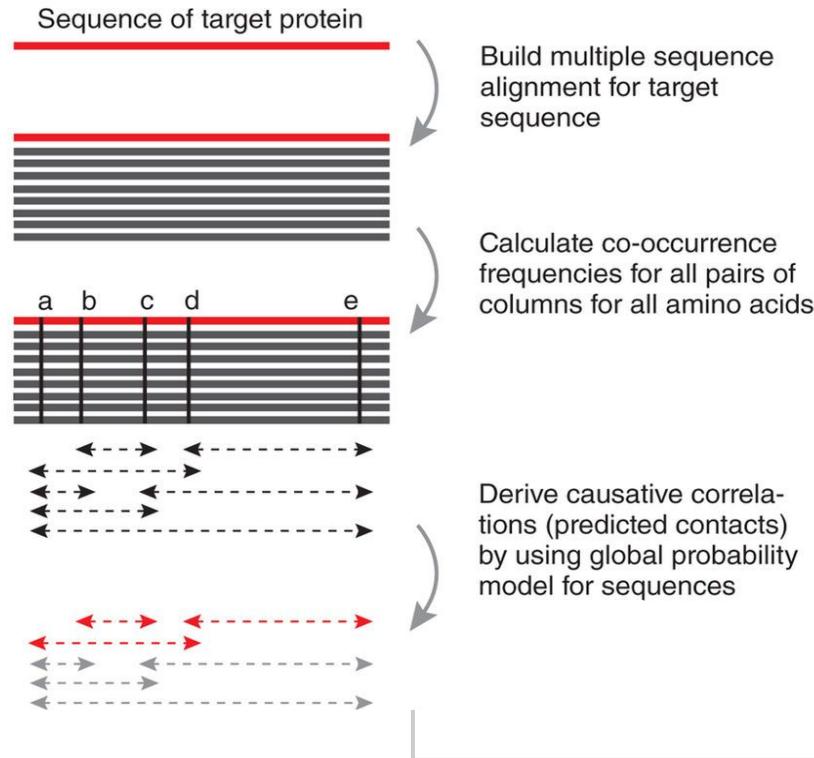
good scores



bad scores



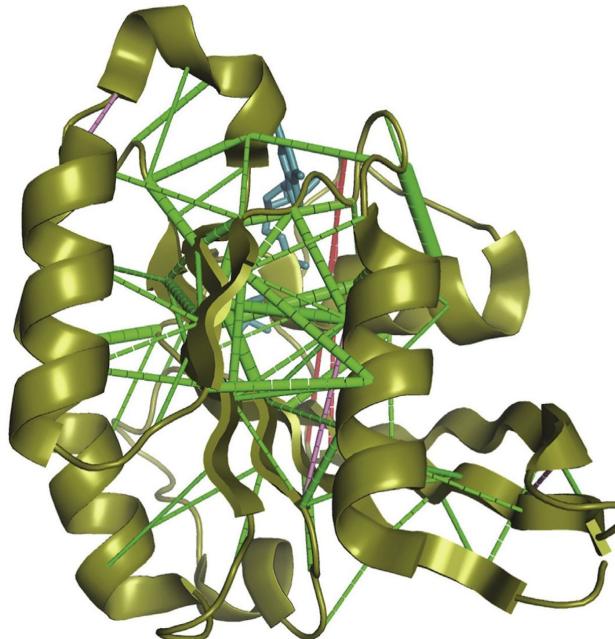
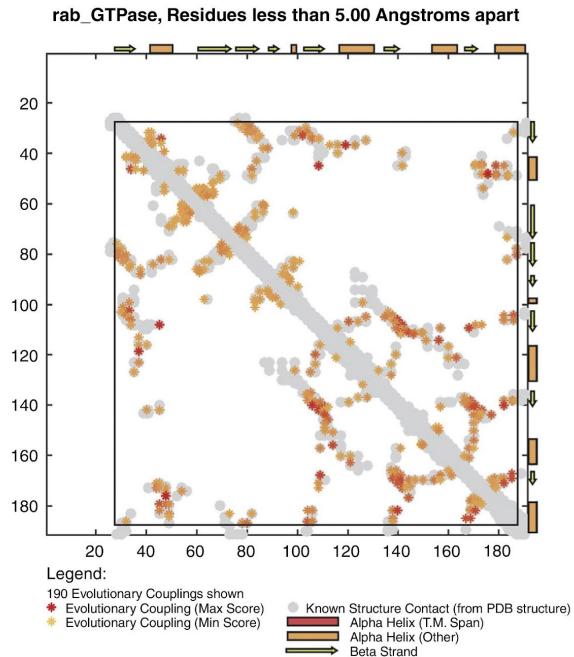
Predicting protein 3D structure based on evolutionary coupling



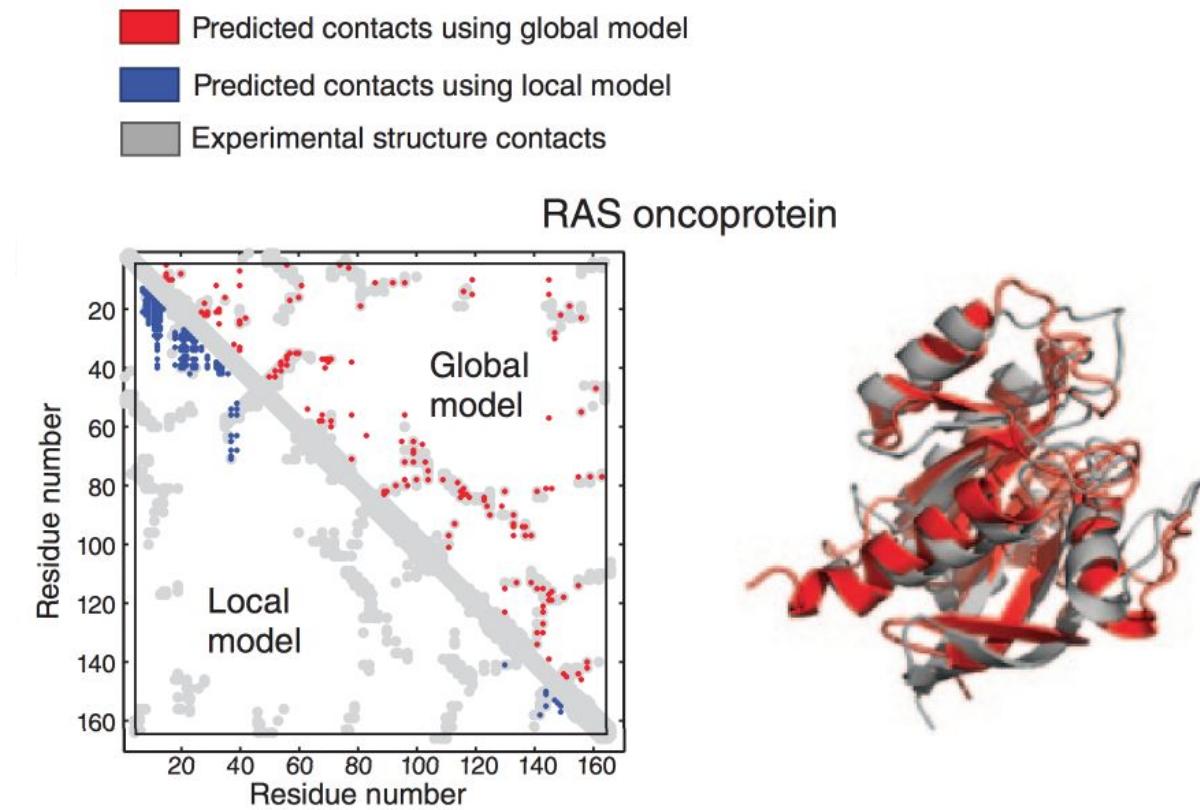
Three-dimensional structure of target protein

Marks (2012) Nat. Biotech.

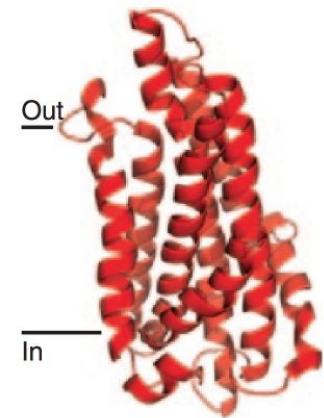
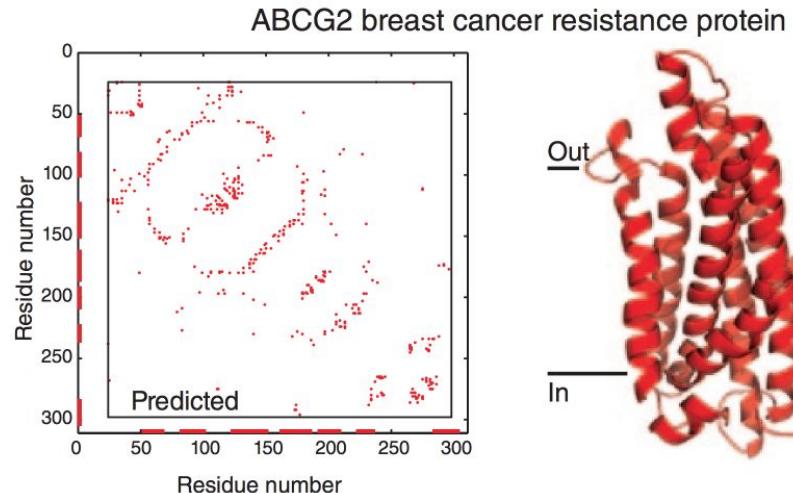
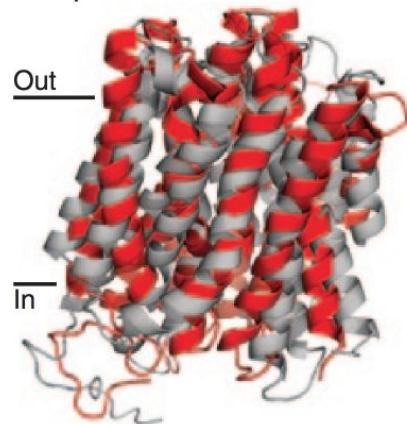
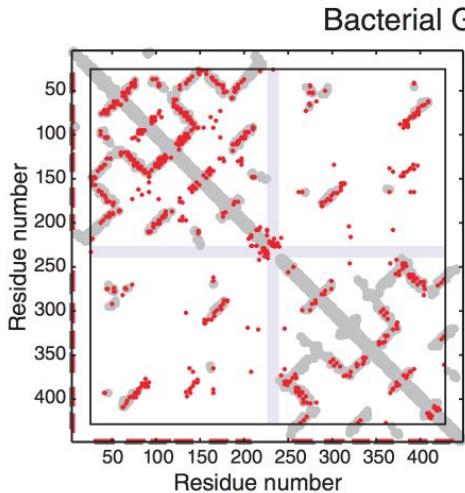
Predictions of 3D structures based on evolutionary coupling



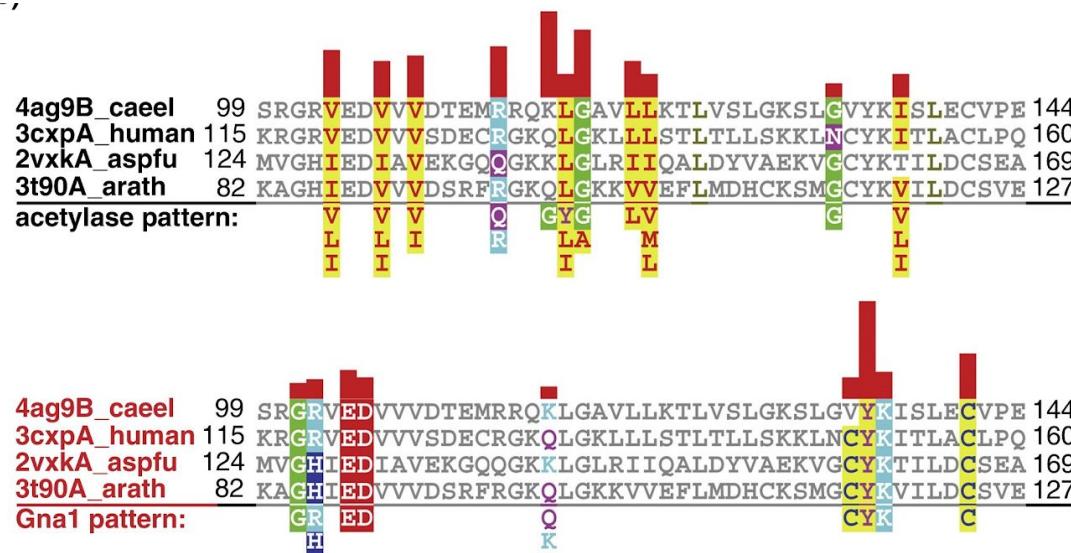
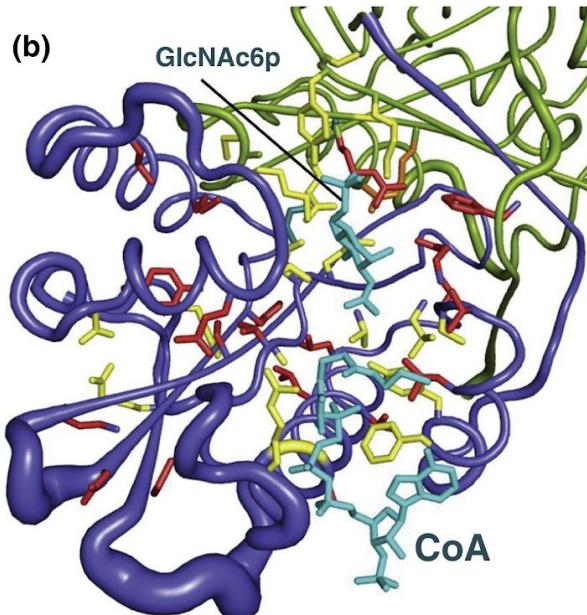
Predictions of 3D structures based on evolutionary coupling



Predictions of 3D structures based on evolutionary coupling



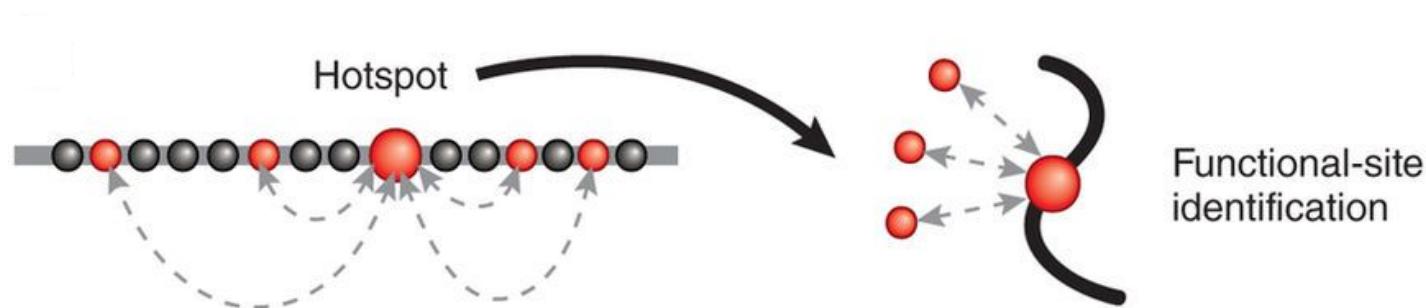
Predictions of 3D structures based on evolutionary coupling



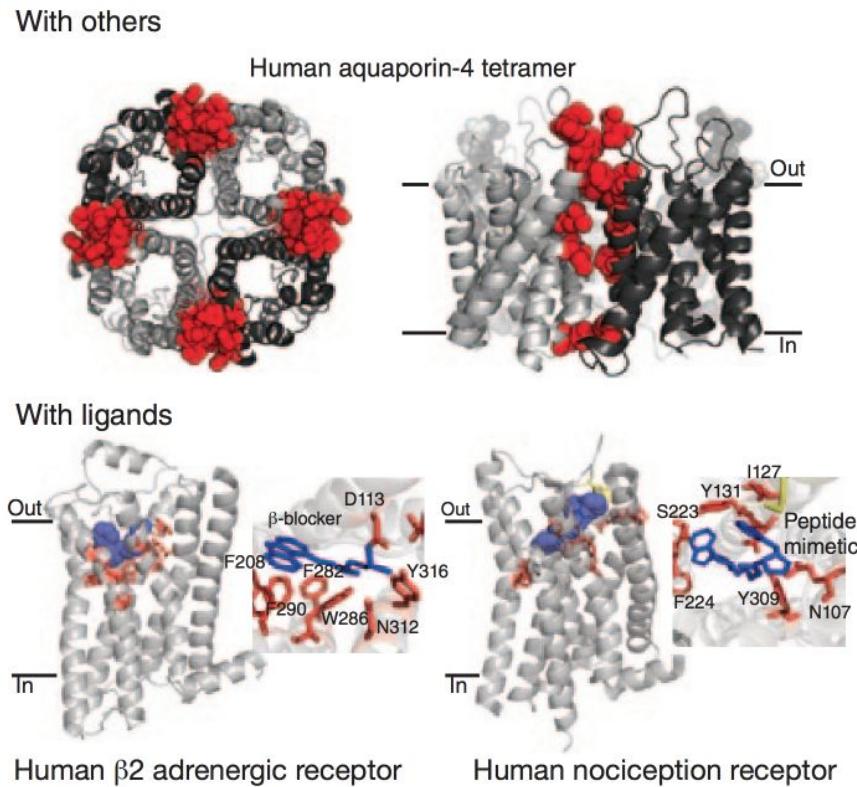
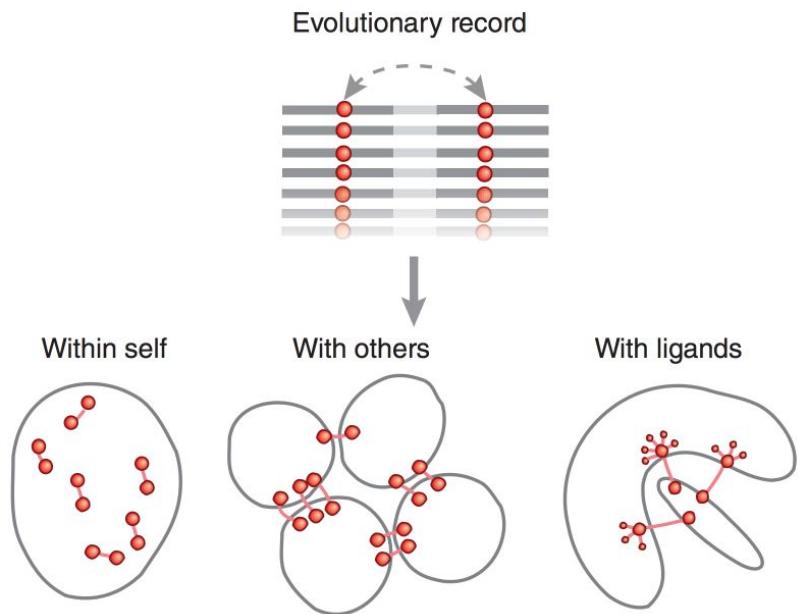
Detecting functional hotspots

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.

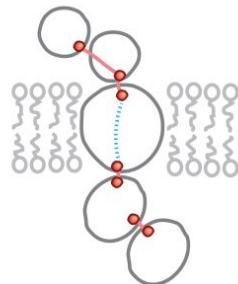


Predicting protein-protein & protein-ligand interactions

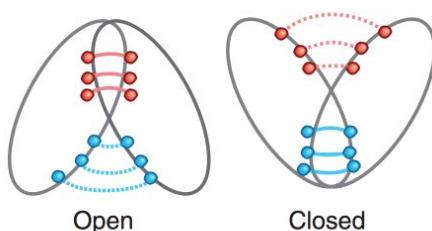


Predicting conformational changes

Information transmission



Conformational plasticity



Conformational plasticity

