Intro to structural biology

Oct 1, 2024

Outline

What is structural biology?

Case study of a LacI transcription factor

Experimental structure determination

Basics of protein structure

Comparing structures

Energy functions

What is structural biology?

At the molecular level structure implies function

Structural biology includes determining molecular structure and inferring dynamics, interactions, function

Proteins are at the center of structural biology

Grand Challenge: protein sequence → structure

AI methods have started to make progress toward solving this problem

PDB - protein databank. Over 200k protein structures.

Why structural biology?

Understand and predict function

Diseases of protein folding

- **Alzheimer's** - misfolding/aggregation of beta-amyloid and tau
- **Parkinson's** - misfolding and aggregation of alpha-synuclein
- **Huntington's** - misfolding of huntingtin protein
- **Cystic fibrosis** - misfolding of CTFR
- **Prion diseases/Creutzfeld-Jakob** - misfolding of PrP into infectious form
- **Sickle cell anemia** - mutation in hemoglobin to form fibers that distort cell shape

Design of new functions - protein design for new enzymes or other functions

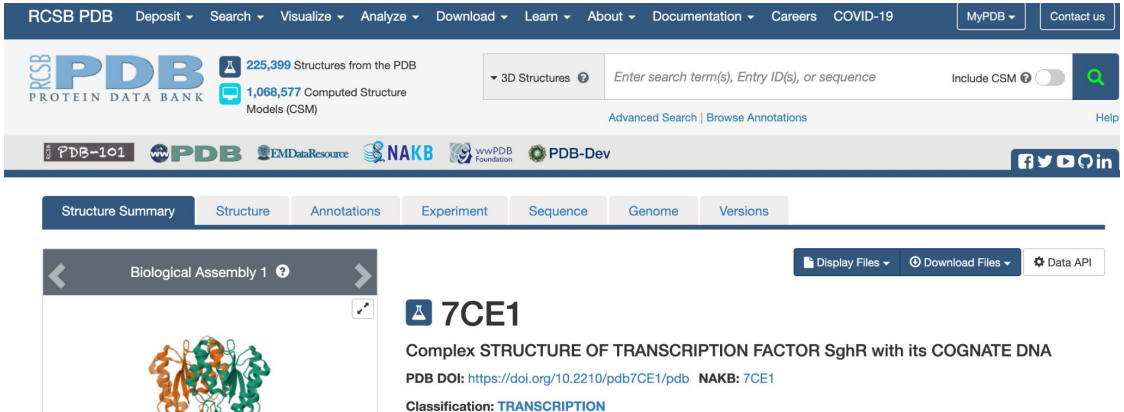
Key resources for structural biology

PyMol: molecular visualization software https://www.pymol.org/

RCSB: database of molecular structures https://www.rcsb.org/

AlphaFold: structure prediction DB and software https://alphafold.ebi.ac.uk/

ESMFold: structure prediction DB and software: https://esmatlas.com/



Using PyMol - investigating a helix-turn-helix TF

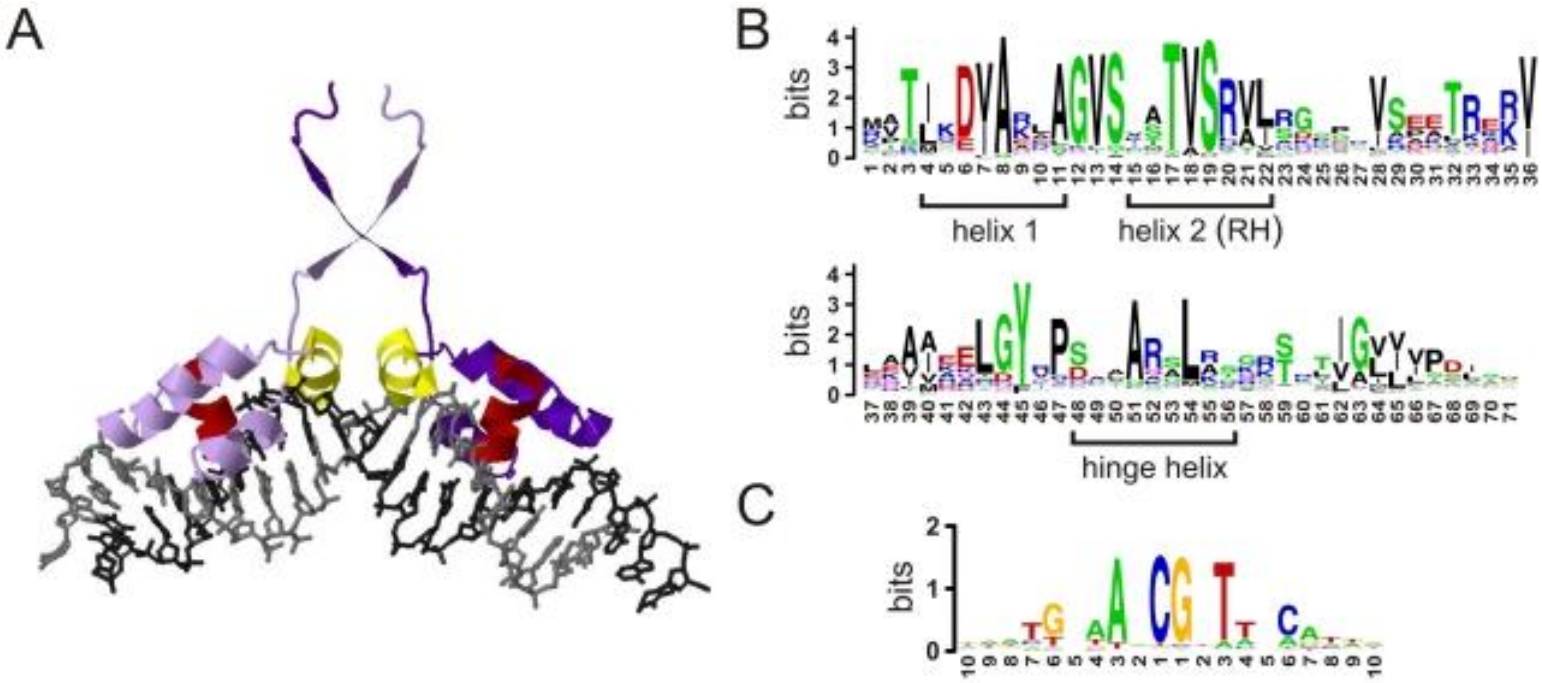
fetch 7CE1

select one_copy, (chain A+B+a+b)

remove not one_copy

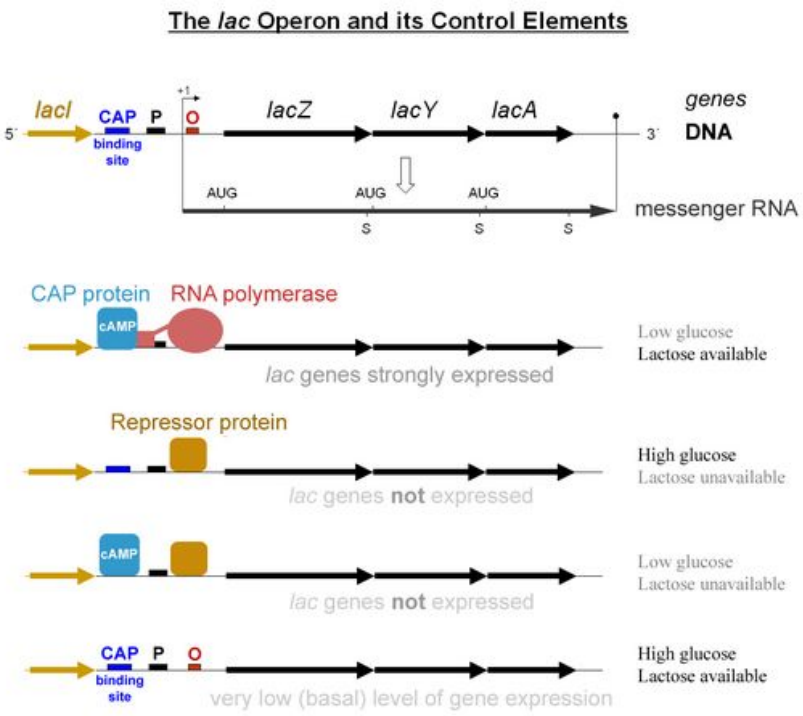
Action→Generate electrostatic surface, Hide polymer.nucleic

How TFs recognize their cognate DNA



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2978694/

Complex logic from structure



https://en.wikipedia.org/wiki/Lac_operon

Experimental methods for determining protein structure - X-ray crystallography

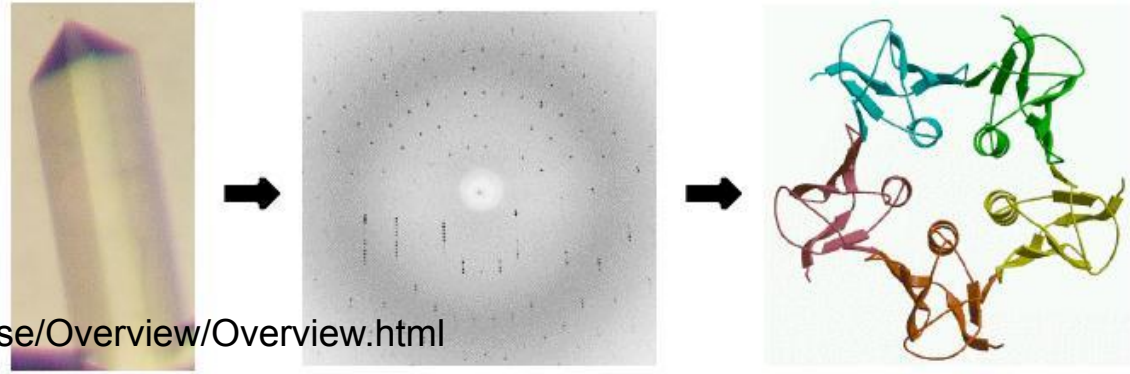
X-ray crystallography - uses x-rays to determine molecular structure by analyzing diffraction pattern

Crystals of each molecule/complex are grown and a strong X-ray source is used to generate a diffraction pattern

An electron density map results from taking the Fourier transform of the diffraction pattern (phase problem)

Requires high-quality crystals

Does not capture dynamic information



https://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html

Experimental methods - NMR

NMR is used to study proteins that are difficult to crytalize or have important dynamics

NMR provides distance constraints that can be used to solve for protein structures

Can identify protein-ligand interactions in solution

Generally only for small proteins (<50 kD)

Experimental methods - CryoEM

Cryo-EM uses electron beams to image biomolecules in their native state that have been rapidly frozen

No need for crystallization of molecules

Ideal for large complexes of protein (eg, ribosome, viruses, etc.)

Large ensembles of molecules are imaged, and these are fit together into a 3D shape using algorithms

Can capture multiple conformations of a molecule in one sample

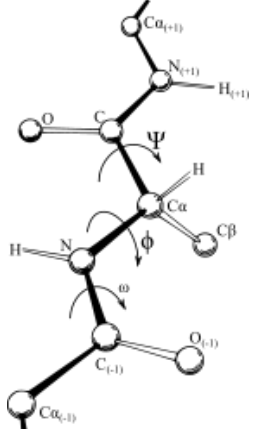
Protein structure basics - conformations

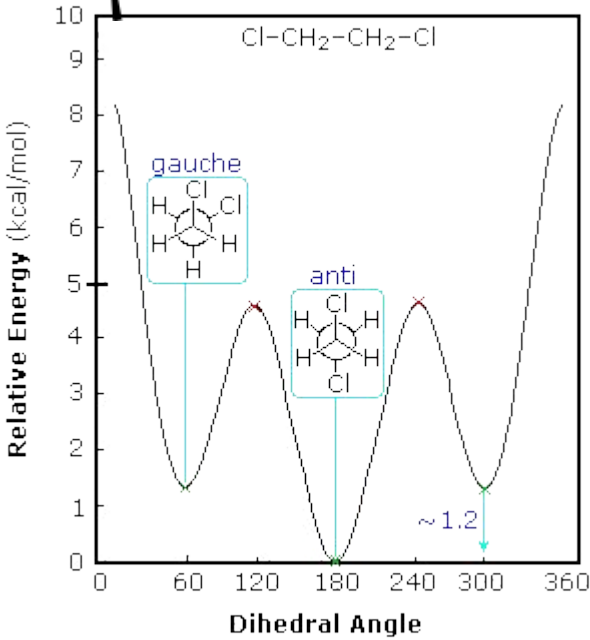
The 3D shape or conformation of a protein is determined by the dihedral angle of each of its rotatable bonds

The global structure of a protein is determined by the rotations of the backbone {phi,psi,omega} for each residue

In practice, {phi,psi} are sufficient to describe the rotational state for most residues besides proline

Each rotatable bond has approximately 3 states





https://www2.chemistry.msu.edu/faculty/reusch/VirtTxtJml/rotconf1.htm

Thought problem

A protein has 100 amino acid residues, and each rotatable bond has 3 states. Ignore side chain conformations. **How many total conformations does the protein have?**

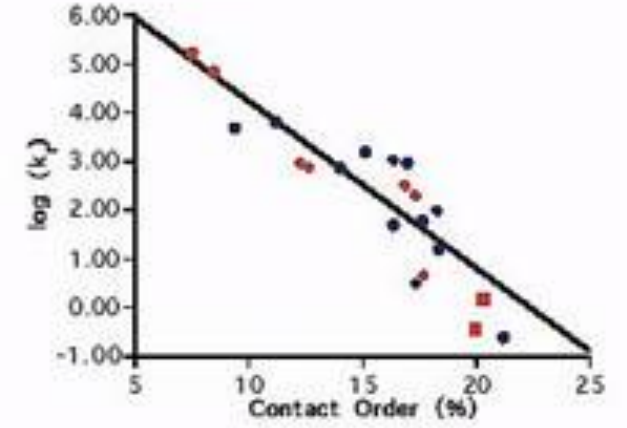
What is the probability of finding the protein in the folded conformation by chance?

The protein folding problem

Anfinson (1950s) - thermodynamic hypothesis (proteins adopt conformation that is a global energy minimum)

Levinthal Paradox (1969) - proteins can't exhaustively sample all possible conformations, yet fold very quickly (<ms to sec)

Contact order predicts folding rates (1998) - more local interactions means faster folding



Levels of protein structure

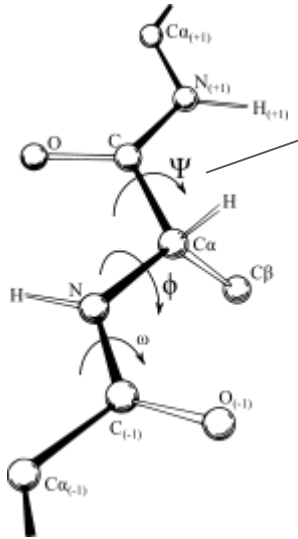
Primary: linear amino acid sequence (eg, MKKGVILEK…)

Secondary: α-helix, β-sheet, random coil

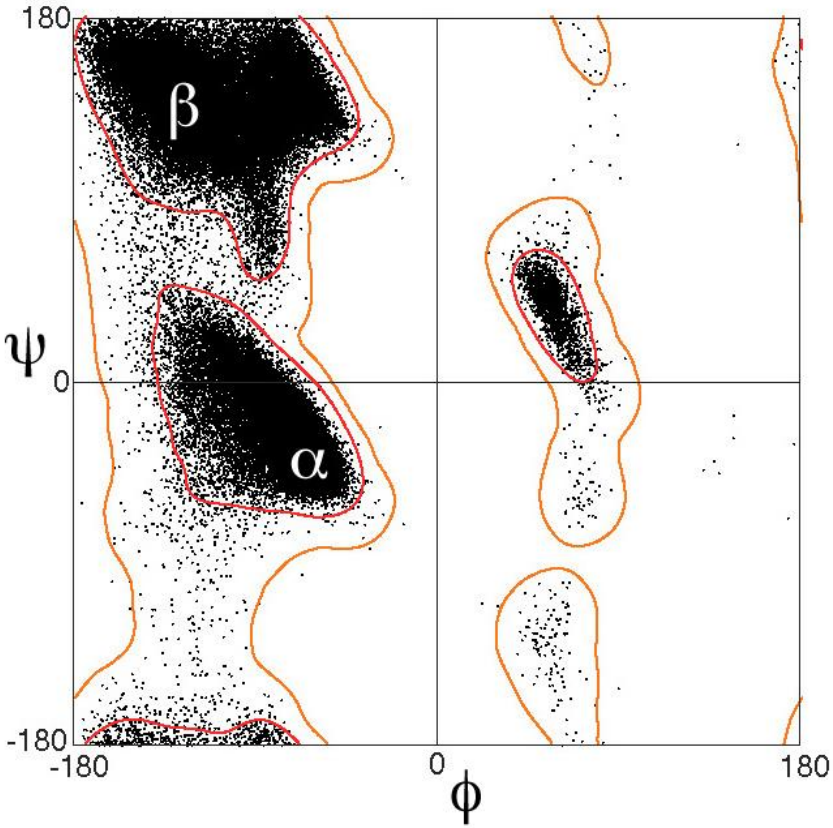
Tertiary: full 3D structure (conformation) of protein chain

Quartenary: multiple interacting protein chains

Protein structure basics - secondary structure



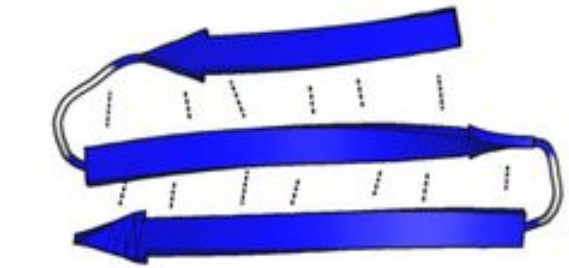
dihedral angles:
phi, psi, omega

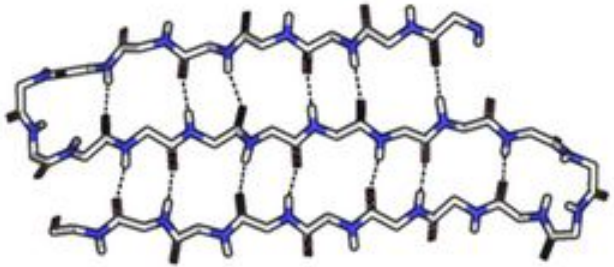


https://en.wikipedia.org/wiki/Ramachandran_plot

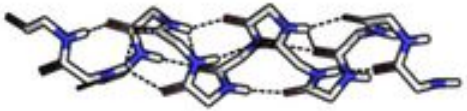
Protein structure basics - secondary structure

Secondary

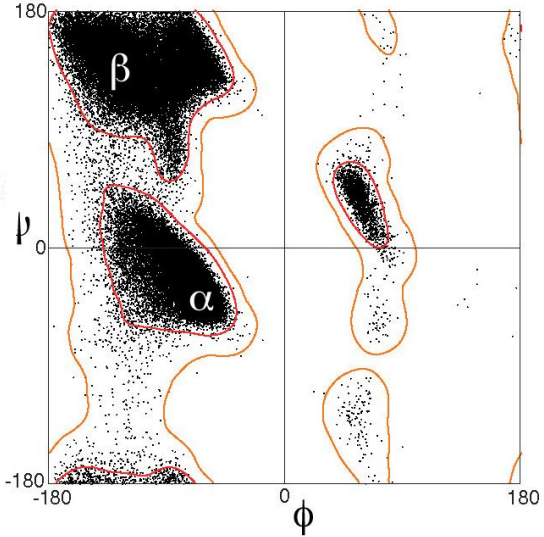




β-Sheet (3 strands)



α-helix



https://en.wikipedia.org/wiki/Ramachandran_plot

Predicting protein secondary structure

Chou-Fasman (1974) - compute relative likelihood of observing AA in each secondary structure P(aa|helix)/P(aa), P(aa|sheet)/P(aa), P(aa|coil)/P(aa). Ad hoc rules for finding stretches of AAs compatible with secondary structure.

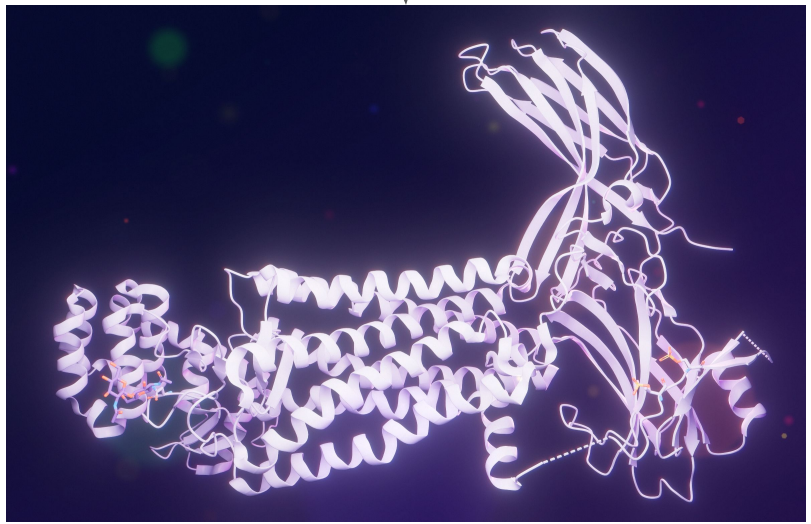
PHD (1993) - 3 layer NN, first to use evolutionary data from MSAs

PSIPRED and other deep learning methods (~2000s)

AlphaFold 2 (2020) - 3D structure prediction implicitly predicts high-quality secondary structure

Tertiary structure prediction

MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAIAD
LFMVFGGFTTTLYTSLHGYFVFGPTGCNIEGFFATLGGEIALWSLVVLAIERWVVVCKPMSNFRFGENHAIMGVAFTWVMAL
ACAVPPLFGWSRYIPEGMQCSCGIDYYTLKPEINNESFVIYMFVVHFIIPLIVIFFCYGRLVCTVKEAAAQQQESATTQKAEKE
VTRMVIIMVIAFLICWLPYAGVAWYLKQYPNLLYLAISAILLNSCINPIIYVVFHSRNEFTFKEAQKHRAKTTKMLGVPMQWNS
ETT

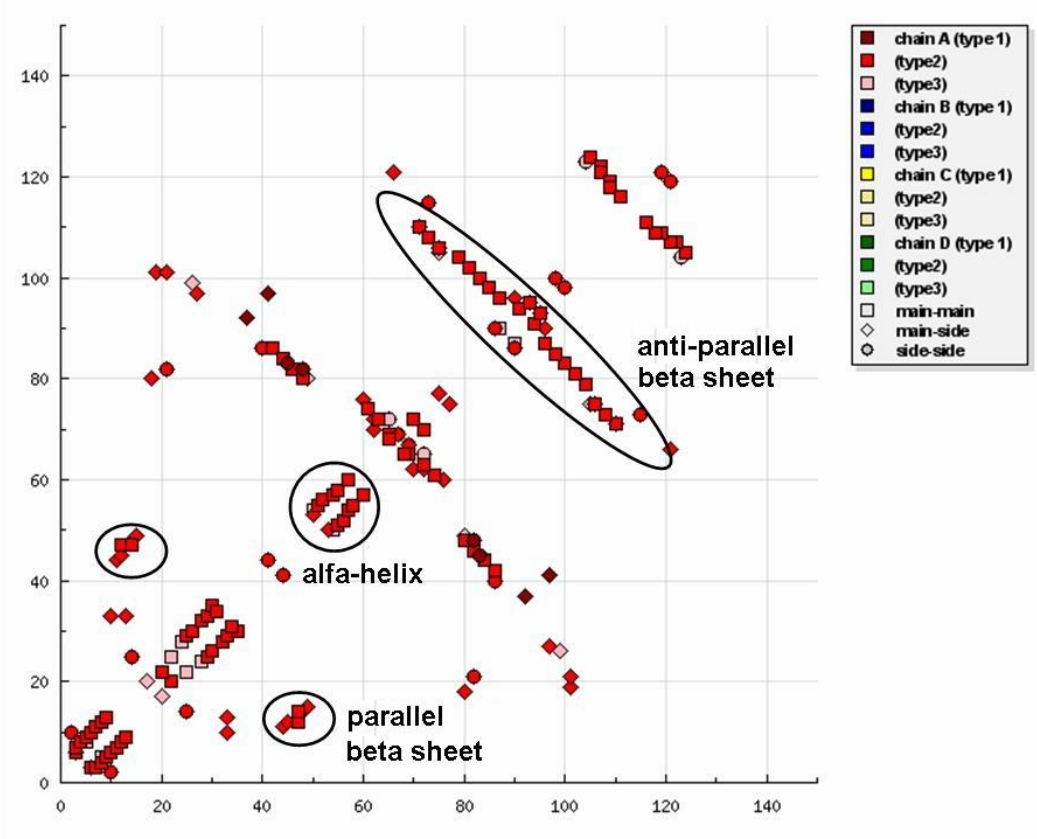


Protein contact maps

Once contacts are known, structure prediction problem is straightforward

Co-evolutionary information from MSAs can be used to identify residues that interact

We will revisit contact maps when we look at attention maps in PLMs



https://en.wikipedia.org/wiki/Protein_contact_map

Tertiary structure prediction milestones

CASP (Critical Assessment of Protein Structure Prediction) catalyzed improvement in prediction methods by holding blind competitions

Rosetta (1990s) - combined fragments of proteins from PDB and scored them using a physics/stats-based **energy function**

AlphaFold 2 (2020) - utilized evolutionary information and deep learning

Comparing structures RMSD

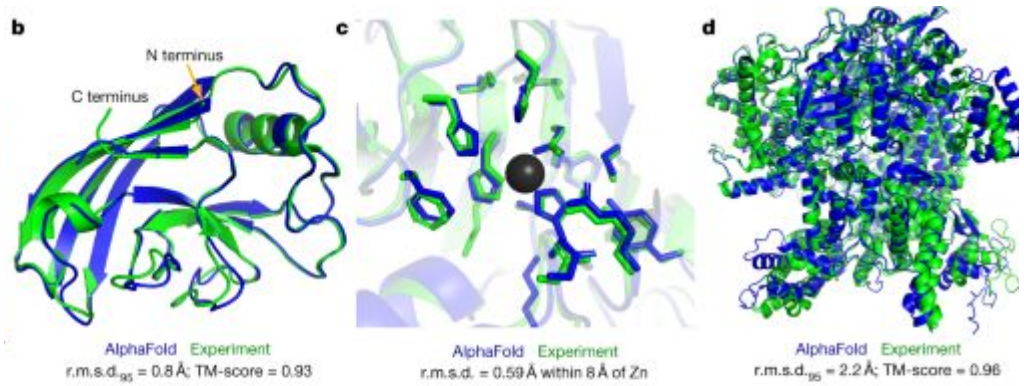
Find corresponding residues (all atom, CA, etc.)

Superimpose structures

Calculate root mean squared distances between corresponding residues:

$$d_i^2 = (x_i^1 - x_i^2)^2 + (y_i^1 - y_i^2)^2 + (z_i^1 - z_i^2)^2$$

$$\text{RMSD} = \text{sqrt}(\ 1/N \sum d_i^2\ )$$



https://www.nature.com/articles/s41586-021-03819-2

Superimposing structures

Align center of mass/centroid

Centroid $C = 1/N \sum_i x_i$, where x_i are position vectors for each residue/atom

Translate each structure to the origin by subtracting centroid vector

Rotate structures

Compute covariance matrix $H = \sum_i r_i^1 * (r_i^2)^T$

Compute singular values $H = U * \Sigma * V^T$

Optimal rotation is $R = V * U^T$

Ab initio prediction using energy functions

Physics based methods make use of energy functions

- Structure prediction
- Docking of proteins and small molecules
- Molecular dynamics
- Mutation effect prediction
- Protein design

Electrostatics

Van der waals

Hydrogen bonds

Solvent interactions and the hydrophobic effect

Electrostatics

$E_{elec} = C * q_1 q_2 / Dr$, where:

q1 and q2 are the charges
r is the distance
D is the dielectric constant (~80 for water, ~2-10 for protein)

For simple calculations, compare with the Bjerrum length l_B:

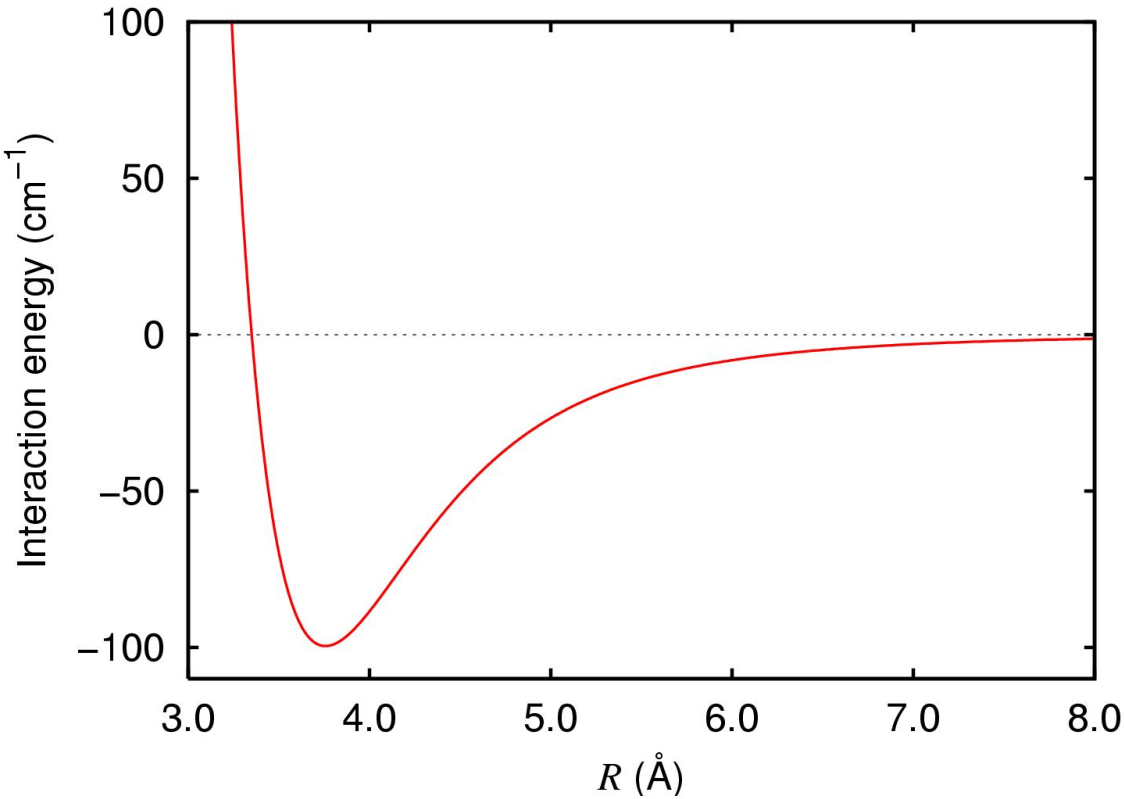
$E_{elec} = kT = 2.5$ kJ/mol at r=7A at 298K in water
$E_{elec} = 7/r * 2.5$ kJ/mol

Van der Waals

Very short range (contact)

Interactions between induced dipoles in electron cloud

Seen for any atomic “surfaces” near each other



Hydrogen bonds

Quantum mechanical effect

Similar to, but weaker than, covalent bonds

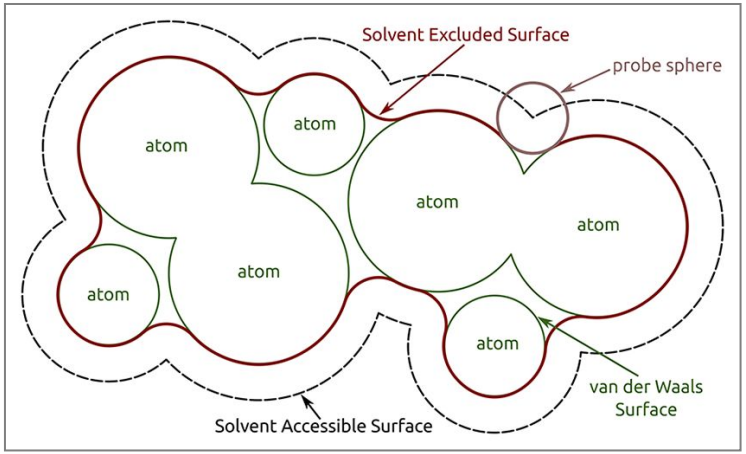
5-10 kJ/mol

Hydrophobic effect

Often approximated as proportional to solvent accessible surface area

Many other terms in energy function are approximately zero or positive because water makes good interactions with proteins (electrostatics, van der waals, H-bonds)

Mostly an entropic effect



How to interpret energies as probabilities

The Boltzman distribution states that the energy of a state (eg, protein conformation) is related to the exponent of its energy

A more general form of the energy that has this property is referred to as the free energy

Boltzmann distribution

$$p_i = \frac{1}{Q} \exp\left(-\frac{\varepsilon_i}{kT}\right) = \frac{\exp\left(-\frac{\varepsilon_i}{kT}\right)}{\sum_{j=1}^{M} \exp\left(-\frac{\varepsilon_j}{kT}\right)}$$

where eps_i = energy of state i, kT = 2.5 kJ/mol at 298K

Thought problem

A protein has two states, open and closed. The open conformation of the protein has an energy 5 kJ/mol higher than the closed state. **What fraction of the protein molecules will be found in the open state?**

Softmax uses the Boltzmann distribution

Formally, the standard (unit) softmax function σ: $\mathbb{R}^K \to (0,1)^K$, where $K \geq 1$, takes a vecto
$\mathbf{z} = (z_1, \ldots, z_K) \in \mathbb{R}^K$ and computes each component of vector $\sigma(\mathbf{z}) \in (0,1)^K$ with

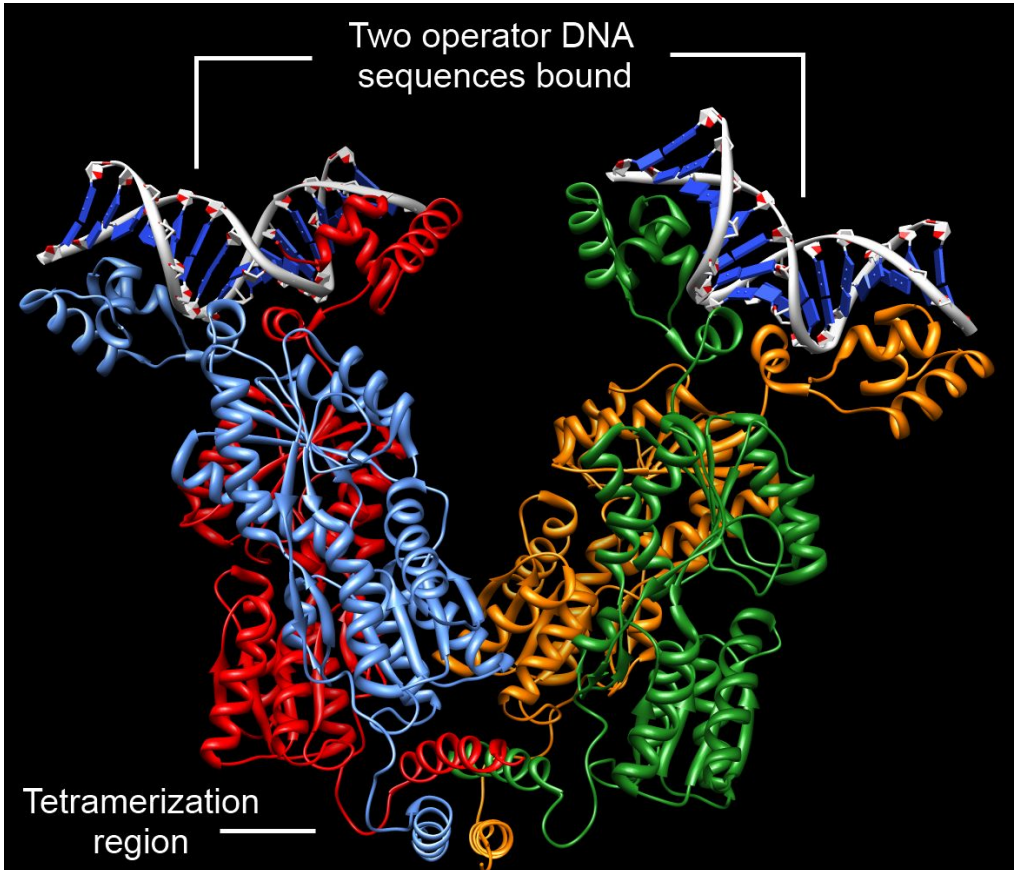
$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} .$$

Softmax converts logits to probabilities

This is the same as the Boltzmann equation mathematically if we use E/kT as logits and take softmax

https://en.wikipedia.org/wiki/Softmax_function

Higher order structure - DNA looping



https://en.wikipedia.org/wiki/Lac_operon

Thought problem

Our LacI protein has two additional binding sites, O2 and O3. **If we mutate O3, so it no longer binds, how much stronger (lower in energy) would we need to make the interaction between LacI and O2 to compensate for the lost configuration?**

Bonus: How would you go about strengthening the LacI-O2 interaction?