CS273B: Deep learning for Genomics and Biomedicine

Lecture 13: Background on protein structure and protein structure prediction

11/6/2017

Anshul Kundaje, James Zou

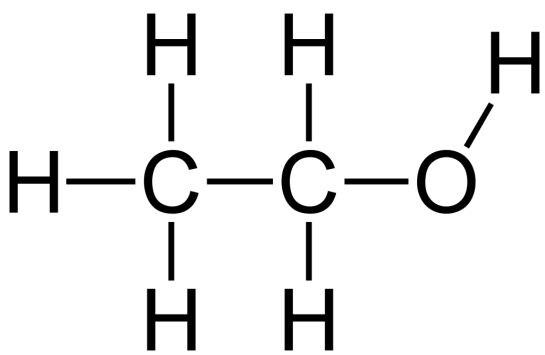
(Slides from Ron Dror's CS279 course)
http://web.stanford.edu/class/cs279/

Chemical (two-dimensional) structure of proteins

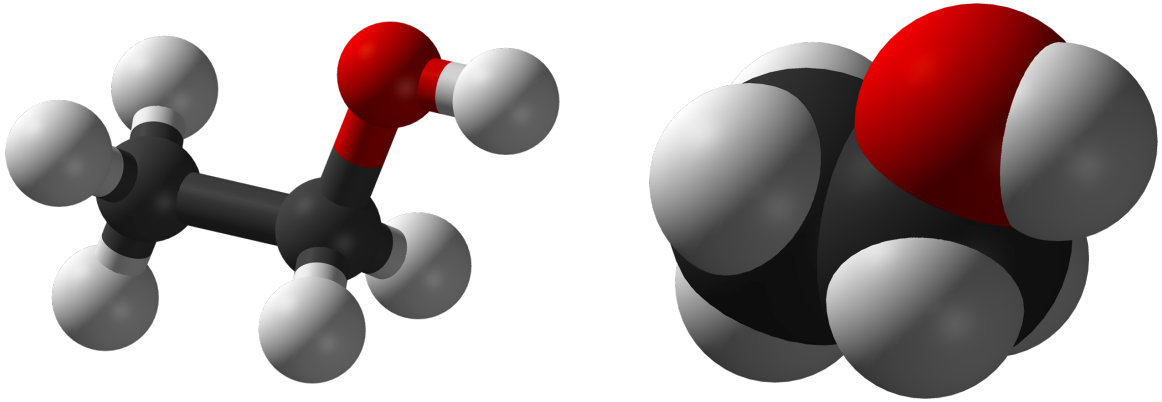
Two-dimensional (chemical) structure vs. three-dimensional structure

- Two-dimensional (chemical) structure shows *covalent bonds* between atoms. Essentially a graph.
- Three-dimensional structure shows relative positions of atoms.

2D structure



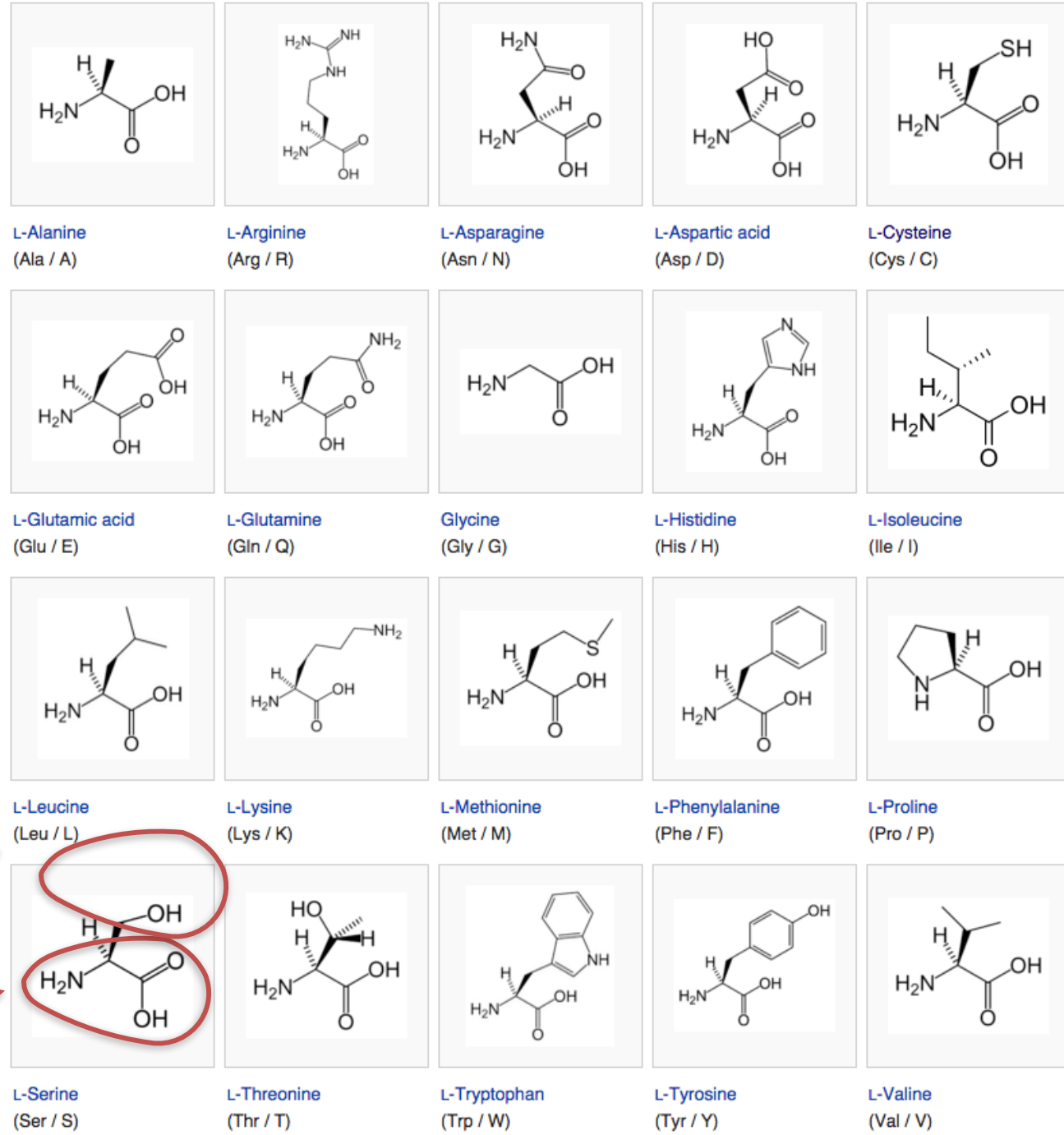
3D structure



images from https://en.wikipedia.org/wiki/Ethanol

Proteins are built from amino acids

- 20 “standard” amino acids
- Each has three-letter and one-letters abbreviations (e.g., Threonine = Thr = T; Tryptophan = Trp = W)



The “side chain” is different in each amino acid

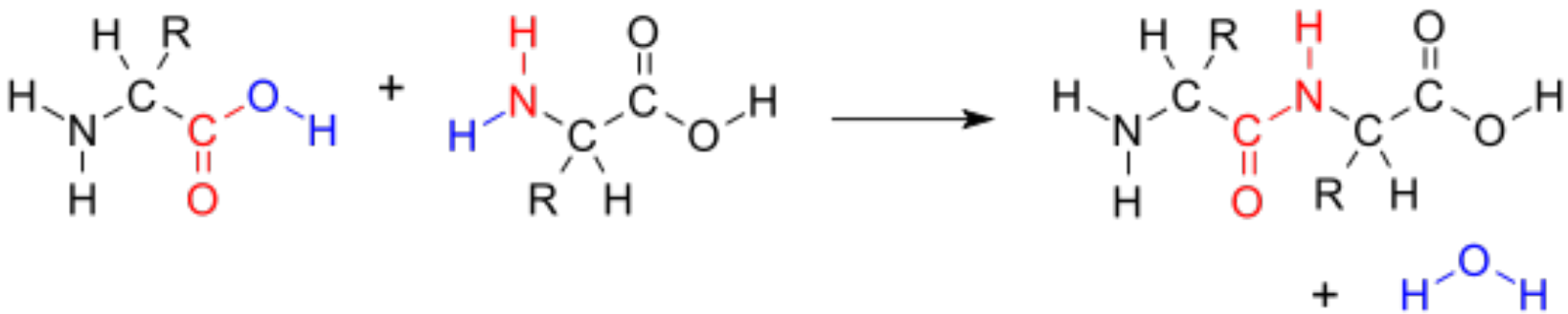
All amino acids have this part in common.

You don’t need to memorize all the structures

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

Proteins are chains of amino acids

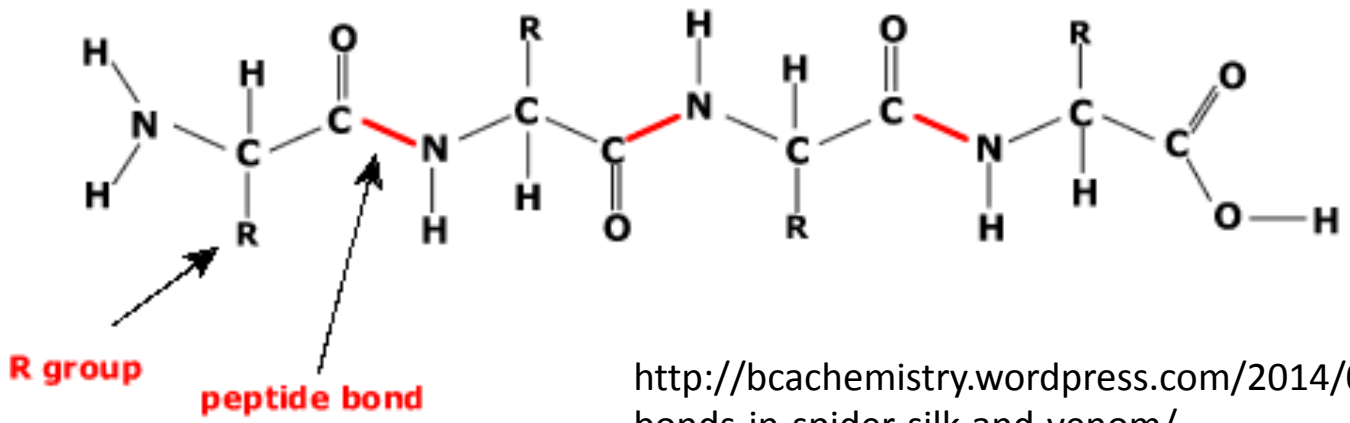
- Amino acids link together through a chemical reaction ("condensation")



http://en.wikipedia.org/wiki/Condensation_reaction

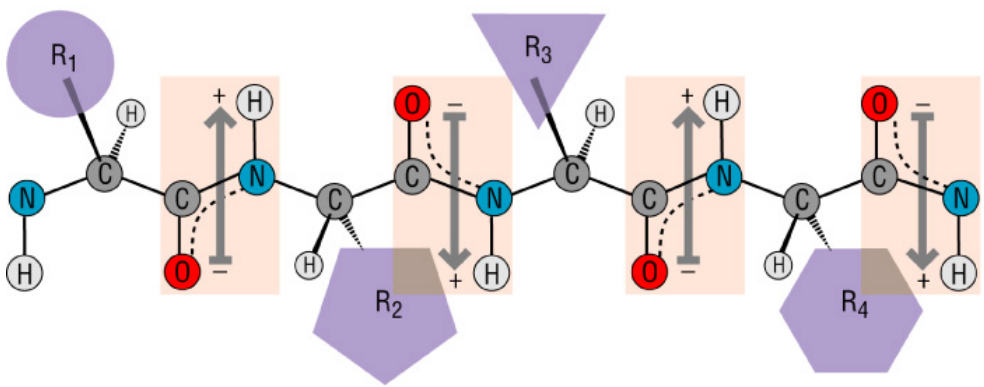
- Elements of the chain are called "amino acid residues" or just "**residues**" (important term!)
- The bonds linking these residues are "peptide bonds." The chains are also called "polypeptides"

Proteins have uniform backbones with differing side chains



http://bcachemistry.wordpress.com/2014/05/28/chemical-bonds-in-spider-silk-and-venom/

From **Protein Structure and Function** by Gregory A Petsko and Dagmar Ringe



© 1999–2004 New Science Press

What determines the 3D structure of a protein?
Physics underlying biomolecular structure

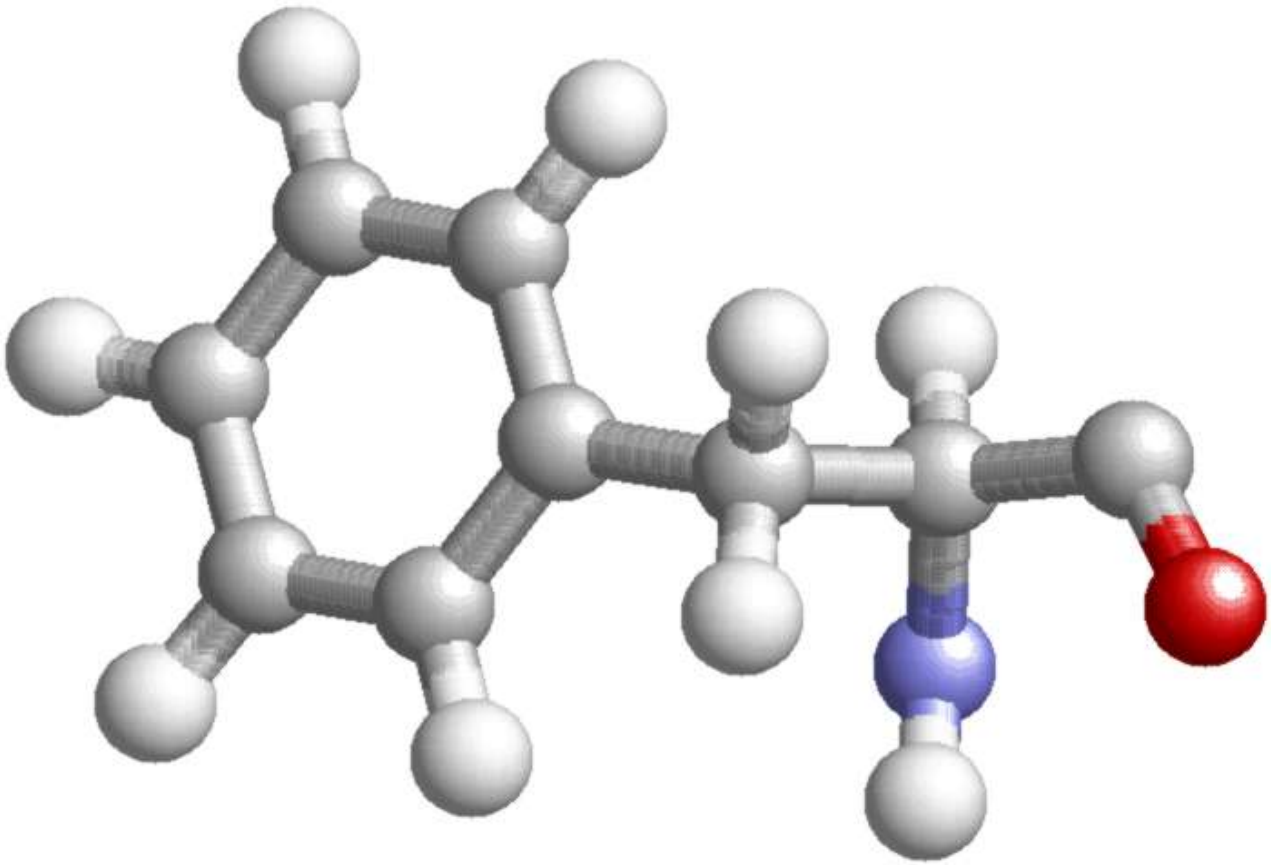
Why do proteins have well-defined structure?

- The sequence of amino acids in a protein (usually) suffices to determine its structure.
- A chain of amino acids (usually) “folds” spontaneously into the protein’s preferred structure, known as the “native structure”
- Why?
  - Intuitively: some amino acids prefer to be inside, some prefer to be outside, some pairs prefer to be near one another, etc.
  - To understand this better, examine forces acting between atoms

What determines the 3D structure of a protein?
Physics underlying biomolecular structure

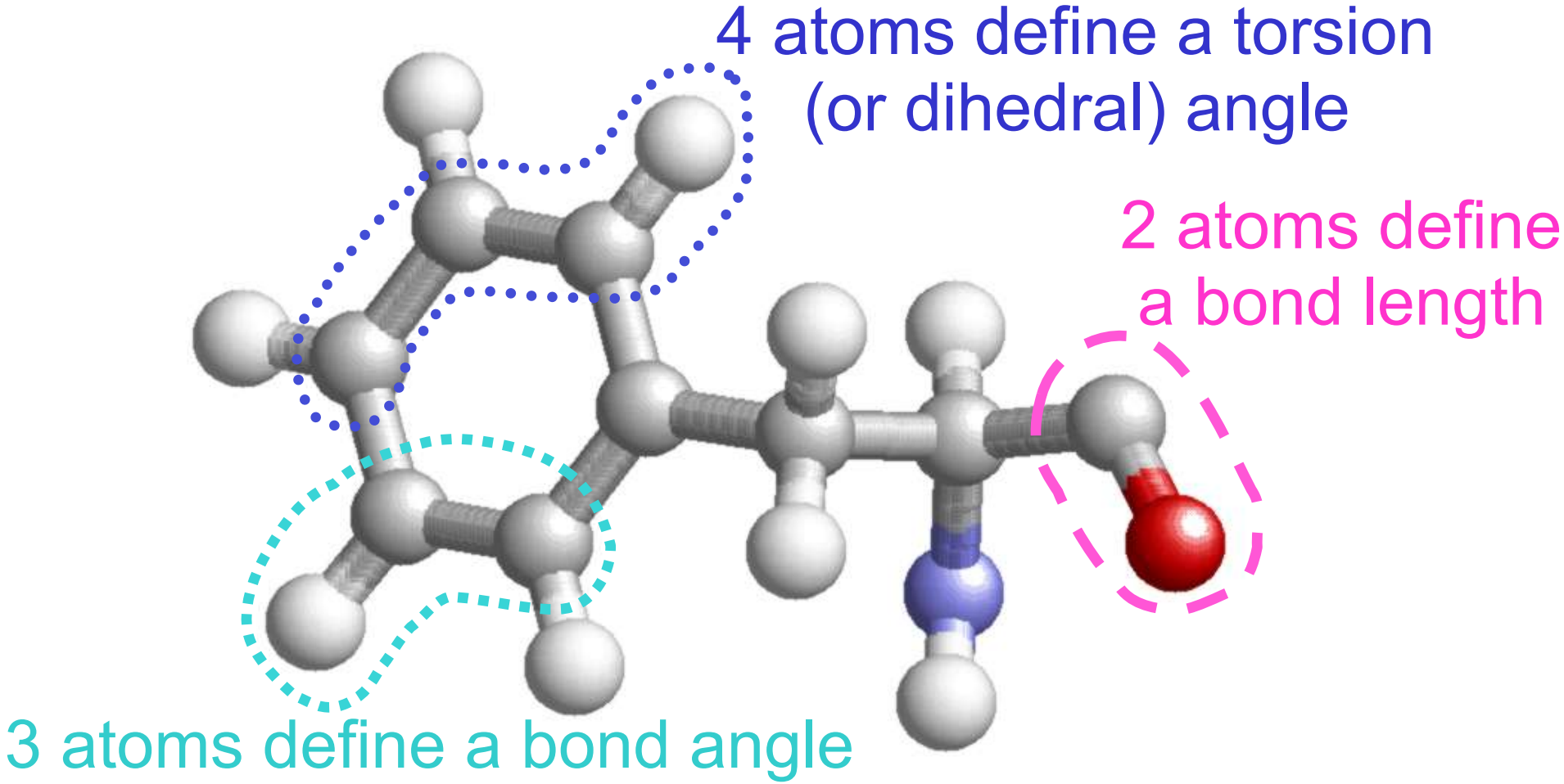
Basic interactions

Geometry of a molecule



- A molecule is a set of atoms connected in a graph
- (x, y, z) coordinates of each atom specify its geometry

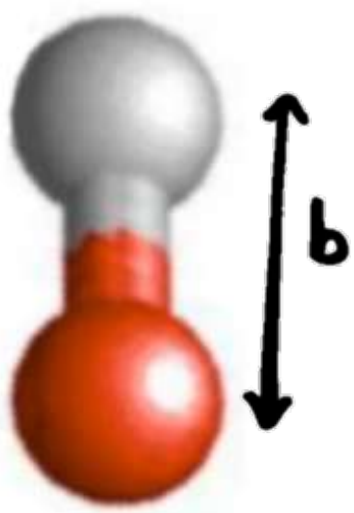
Geometry of a molecule

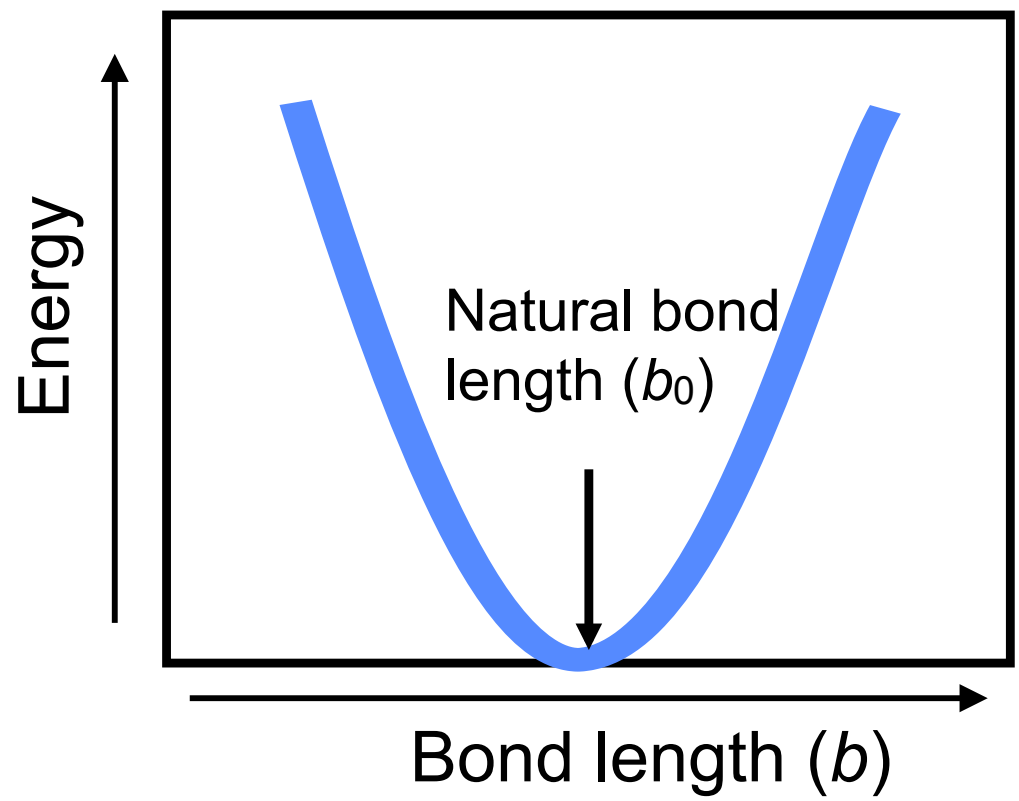


- Alternatively, we can specify the geometry of a molecule using bond lengths, bond angles, and torsion angles

Bond length stretching

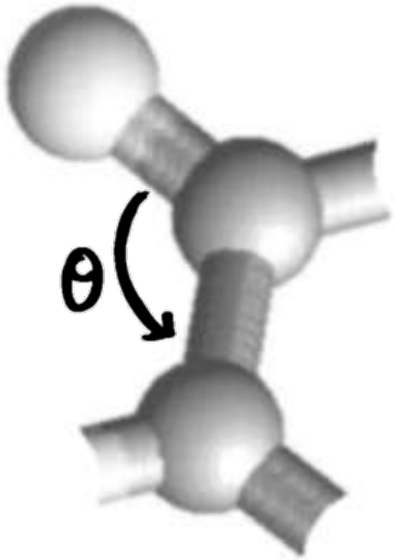
- A bonded pair of atoms is effectively connected by a spring with some preferred (natural) length. Stretching or compressing it requires energy.

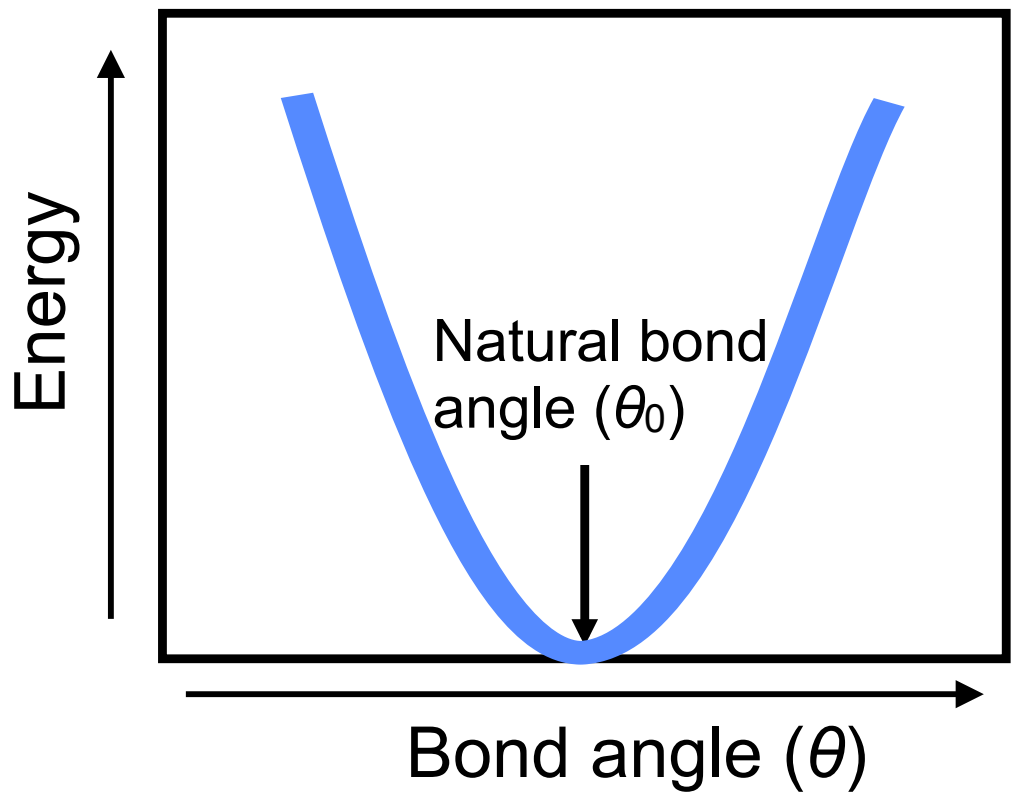




Bond angle bending

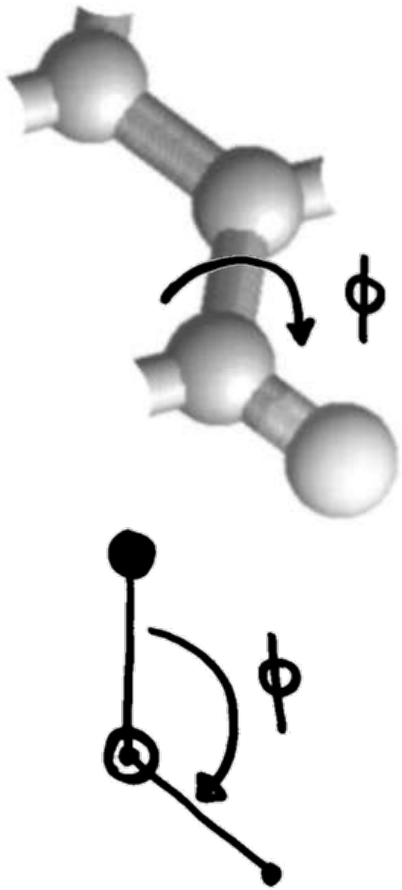
- Likewise, each bond angle has some natural value. Increasing or decreasing it requires energy.

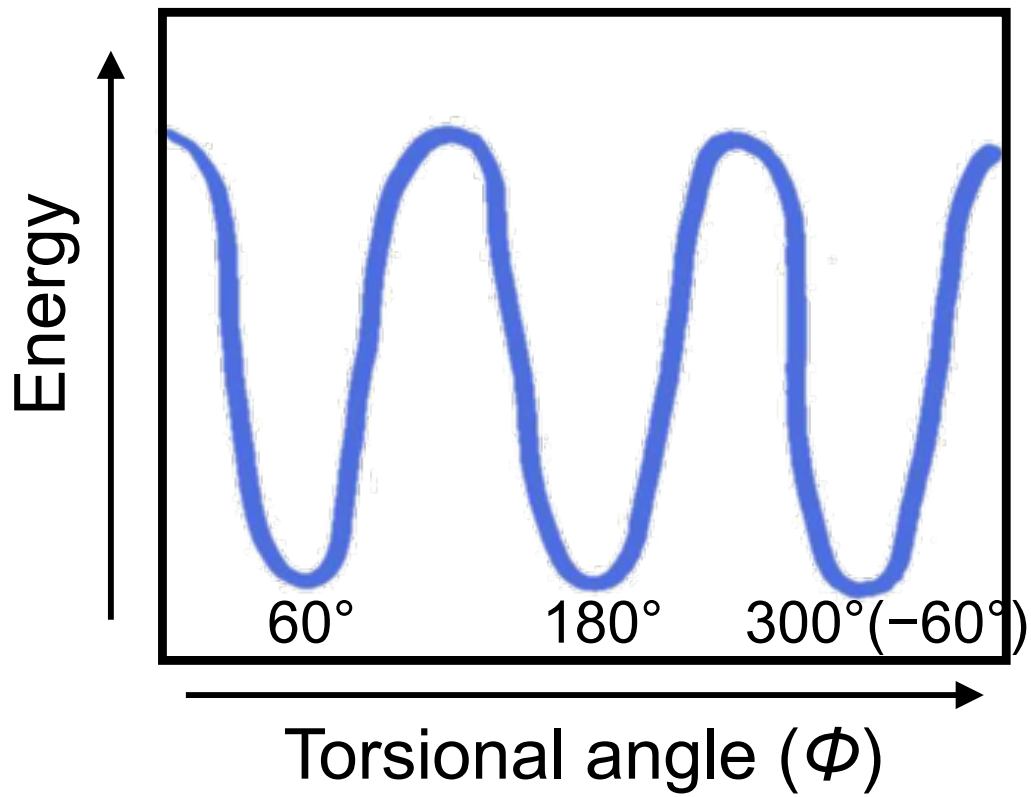




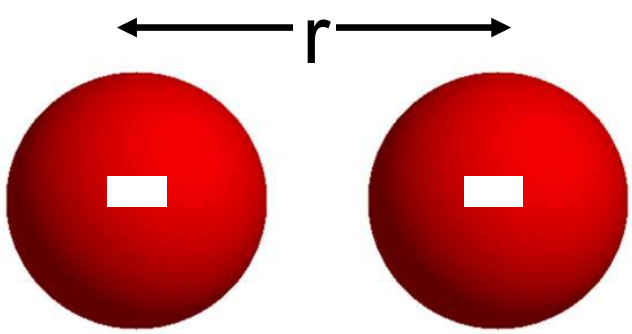
Torsional angle twisting

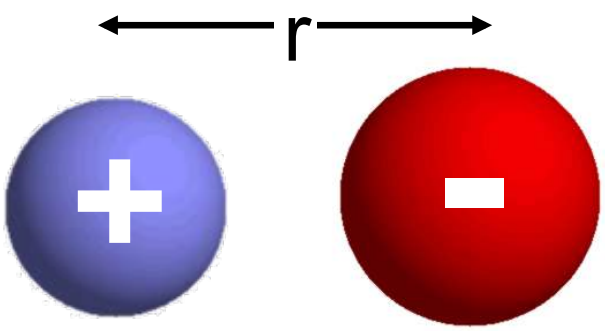
- Certain values of each torsional angle are preferred over others.

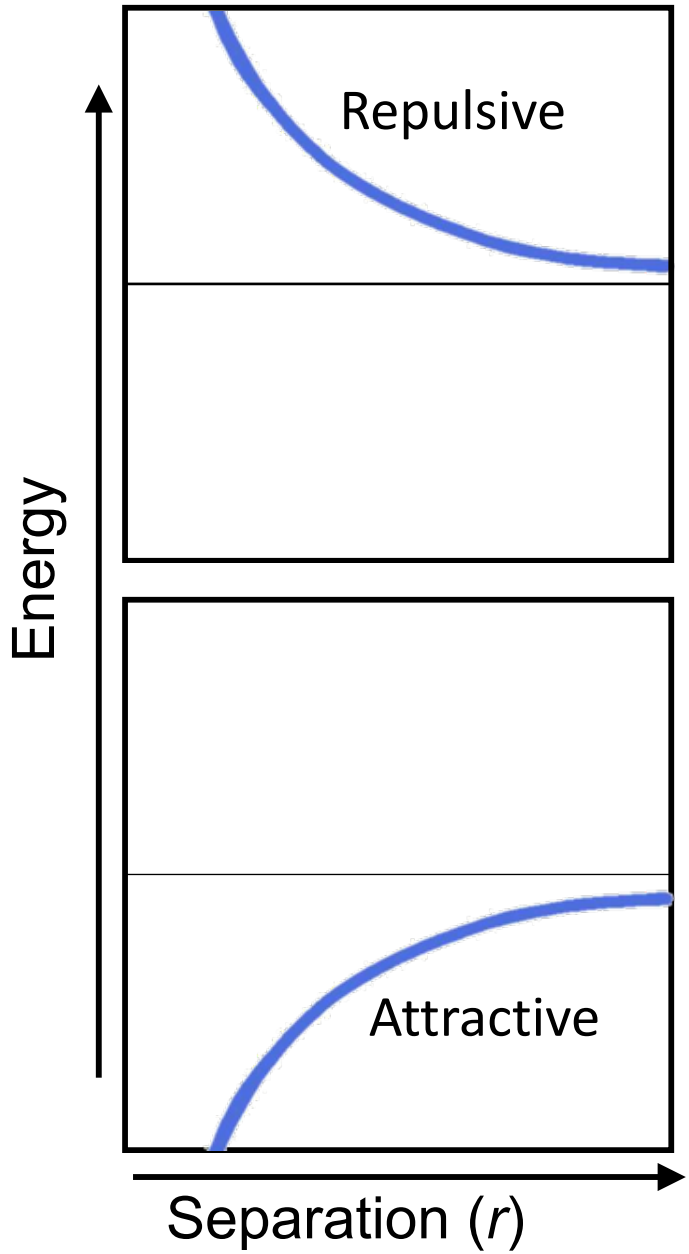




Electrostatic interaction

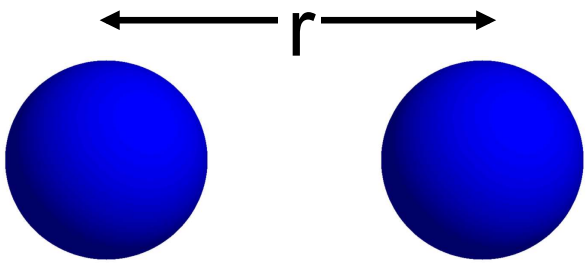


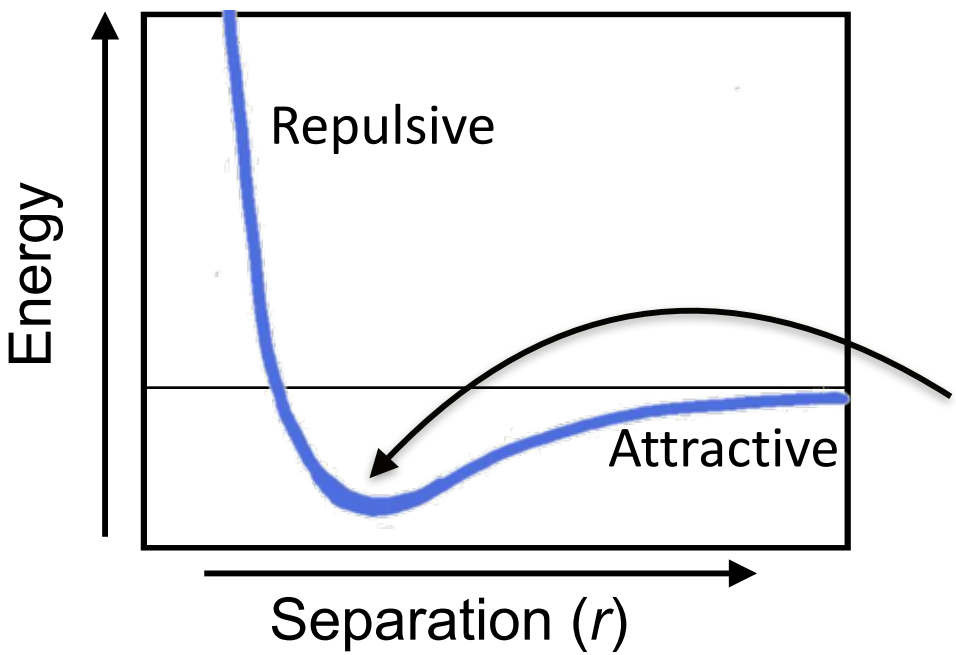




- Like charges repel. Opposite charges attract.
- Acts between all pairs of atoms, including those in different molecules.
- Each atom carries some "partial charge" (may be a fraction of an elementary charge), which depends on which atoms it's connected to.

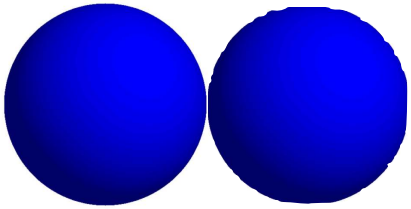
van der Waals interaction





- van der Waals forces act between all pairs of atoms and do not depend on charge.
- When two atoms are too close together, they repel strongly.
- When two atoms are a bit further apart, they attract one another weakly.

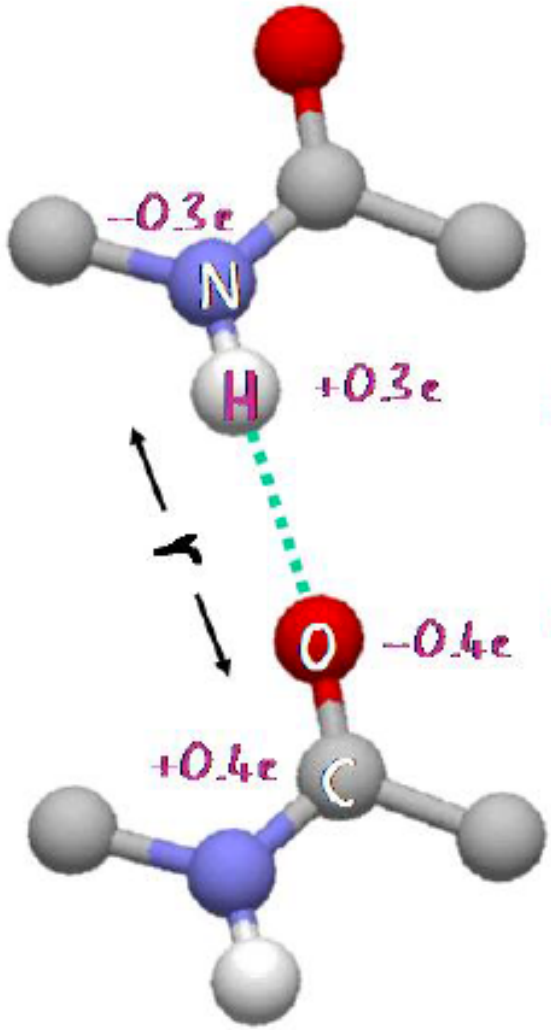
Energy is minimal when atoms are "just touching" one another

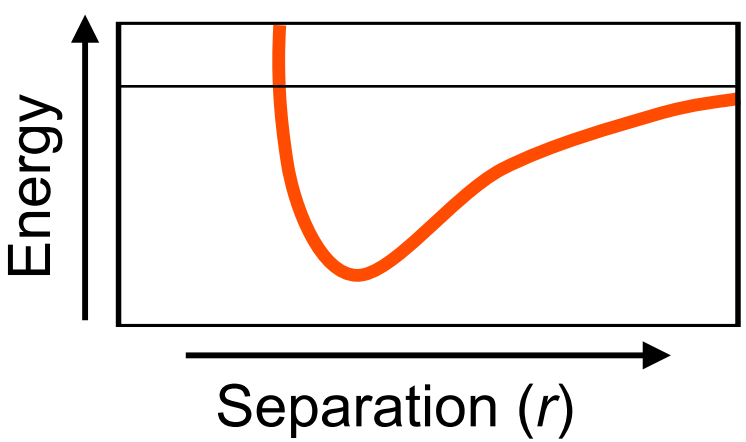


What determines the 3D structure of a protein?
Physics underlying biomolecular structure

Complex interactions

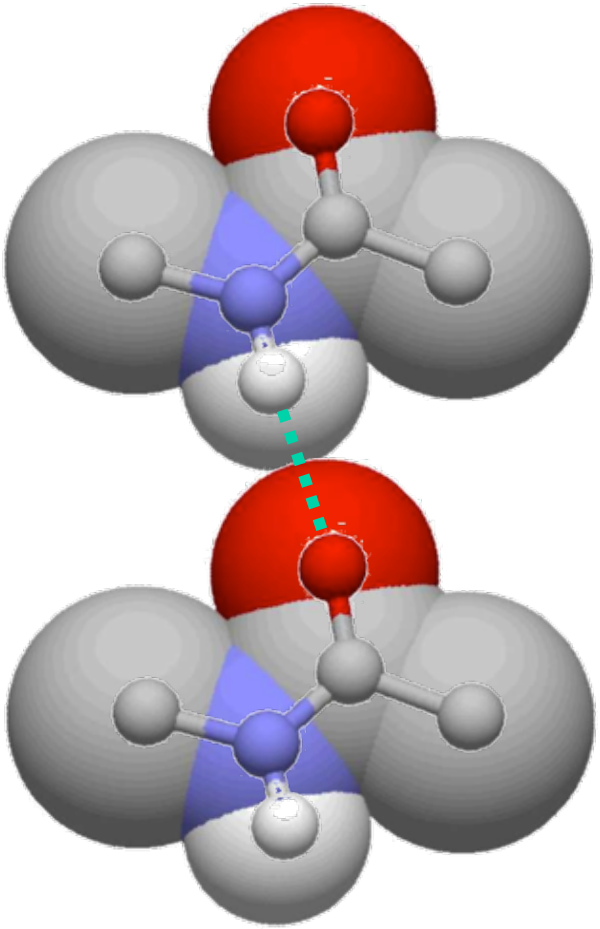
Hydrogen bonds

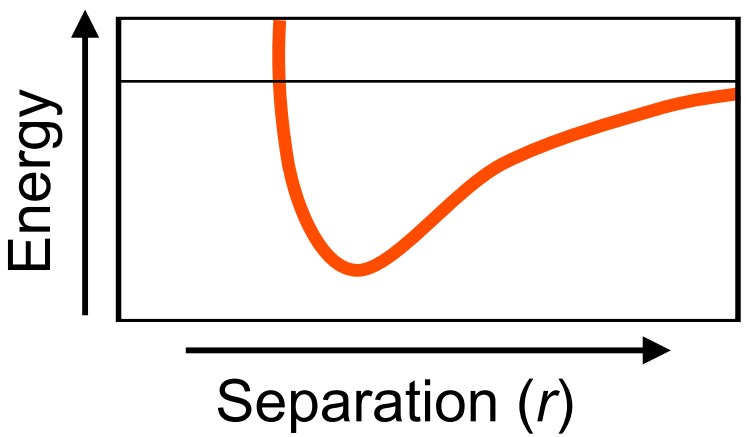




- Favorable interaction between an electronegative atom (e.g., N or O) and a hydrogen bound to another electronegative atom
- Result of multiple electrostatic and van der Waals interactions
- Very sensitive to geometry of the atoms (distance and alignment)
- Strong relative to typical van der Waals or electrostatic forces
- Critical to protein structure

Hydrogen bonds

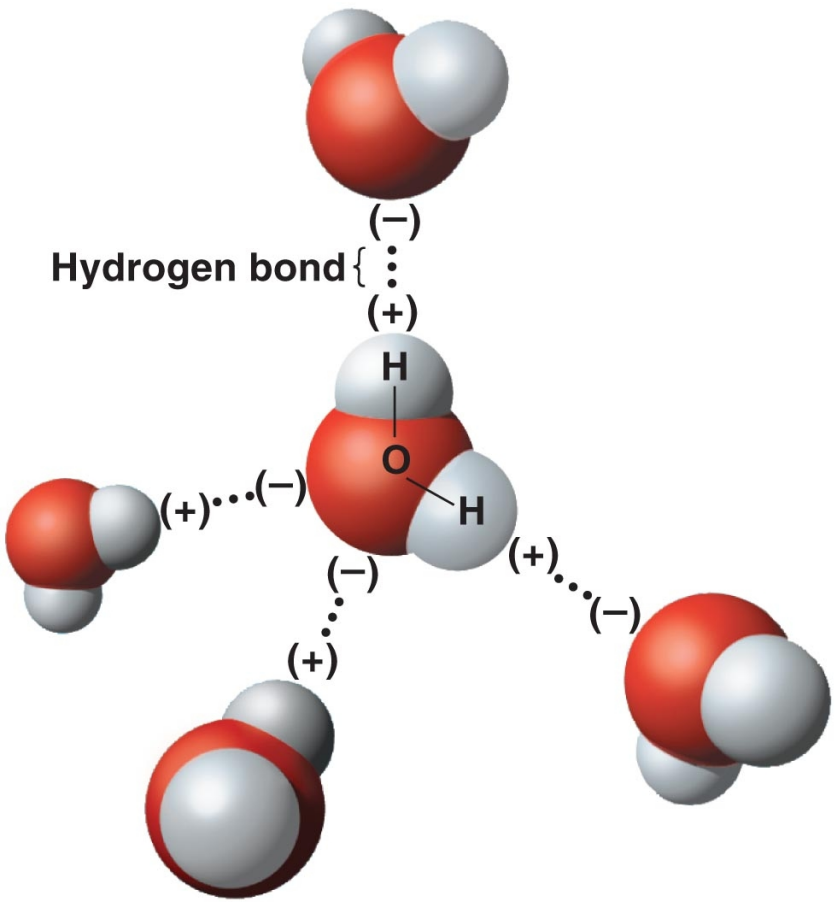




- Favorable interaction between an electronegative atom (e.g., N or O) and a hydrogen bound to another electronegative atom
- Result of multiple electrostatic and van der Waals interactions
- Very sensitive to geometry of the atoms (distance and alignment)
- Strong relative to typical van der Waals or electrostatic forces
- Critical to protein structure

Water molecules form hydrogen bonds

- Water molecules form extensive hydrogen bonds with one another and with protein atoms
- The structure of a protein depends on the fact that it is surrounded by water

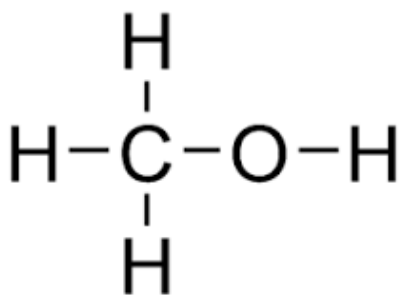


http://like-img.com/show/hydrogen-bond-water-molecule.html

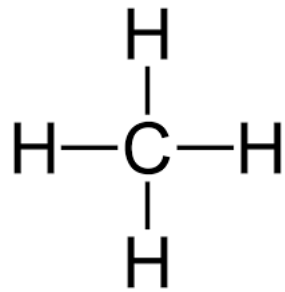
Hydrophilic vs. hydrophobic

- Hydrophilic molecules are polar and thus form hydrogen bonds with water
  - Polar = contains charged atoms. Molecules containing oxygen or nitrogen are usually polar.
- Hydrophobic molecules are apolar and don't form hydrogen bonds with water

Hydrophilic (polar)



Hydrophobic (apolar)



Protein structure: a more detailed view

“Levels” of protein structure

- Primary structure: sequence of amino acids
- Secondary structure: local structural elements
- Tertiary structure: overall structure of the polypeptide chain
- Quaternary structure: how multiple polypeptide chains come together

Protein structure: a more detailed view

Properties of amino acids

Amino acid properties

- Amino acid side chains have a wide range of properties. These differences bring about the 3D structures of proteins.
- Examples:
  - Large side chains take up more space than small ones
  - Hydrophobic side chains want to be near one another
  - Hydrophilic side chains form hydrogen bonds to one another and to water molecules
  - Negatively charged (acidic) side chains want to be near positively charged (basic) side chains

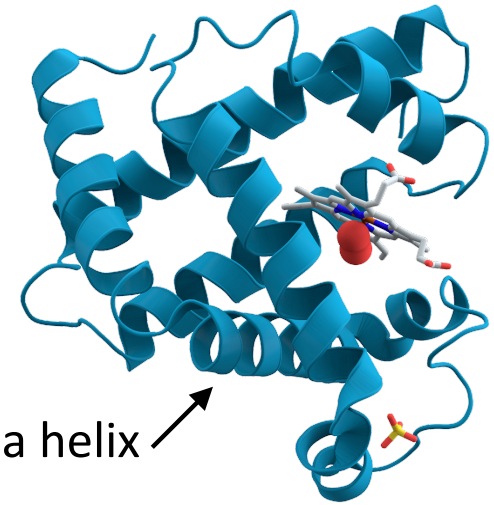
Protein structure: a more detailed view

Secondary structure

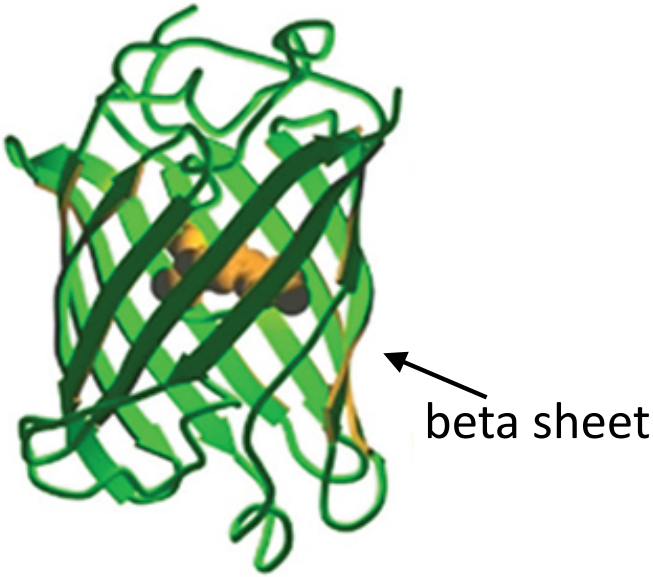
Secondary structure

- "Secondary structure" refers to certain local structural elements found in many proteins
  - These are energetically favorable primarily because of hydrogen bonds between backbone atoms
- Most important secondary structure elements:
  - alpha helix
  - beta sheet

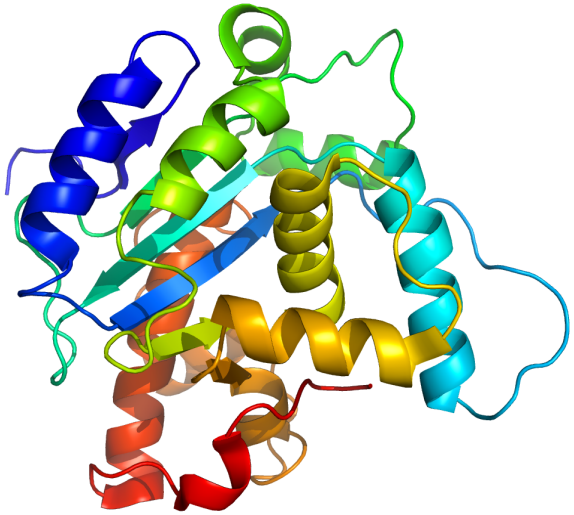
Myoglobin



Green Fluorescent Protein



Pop2p



https://upload.wikimedia.org/wikipedia/commons/

http://www.biotek.com/assets/tech_resources/11596/figure2.jpg

http://upload.wikimedia.org/wikipedia/commons/e/e6/Spombe_Pop2p_protein_structure_rainbow.png

The alpha helix

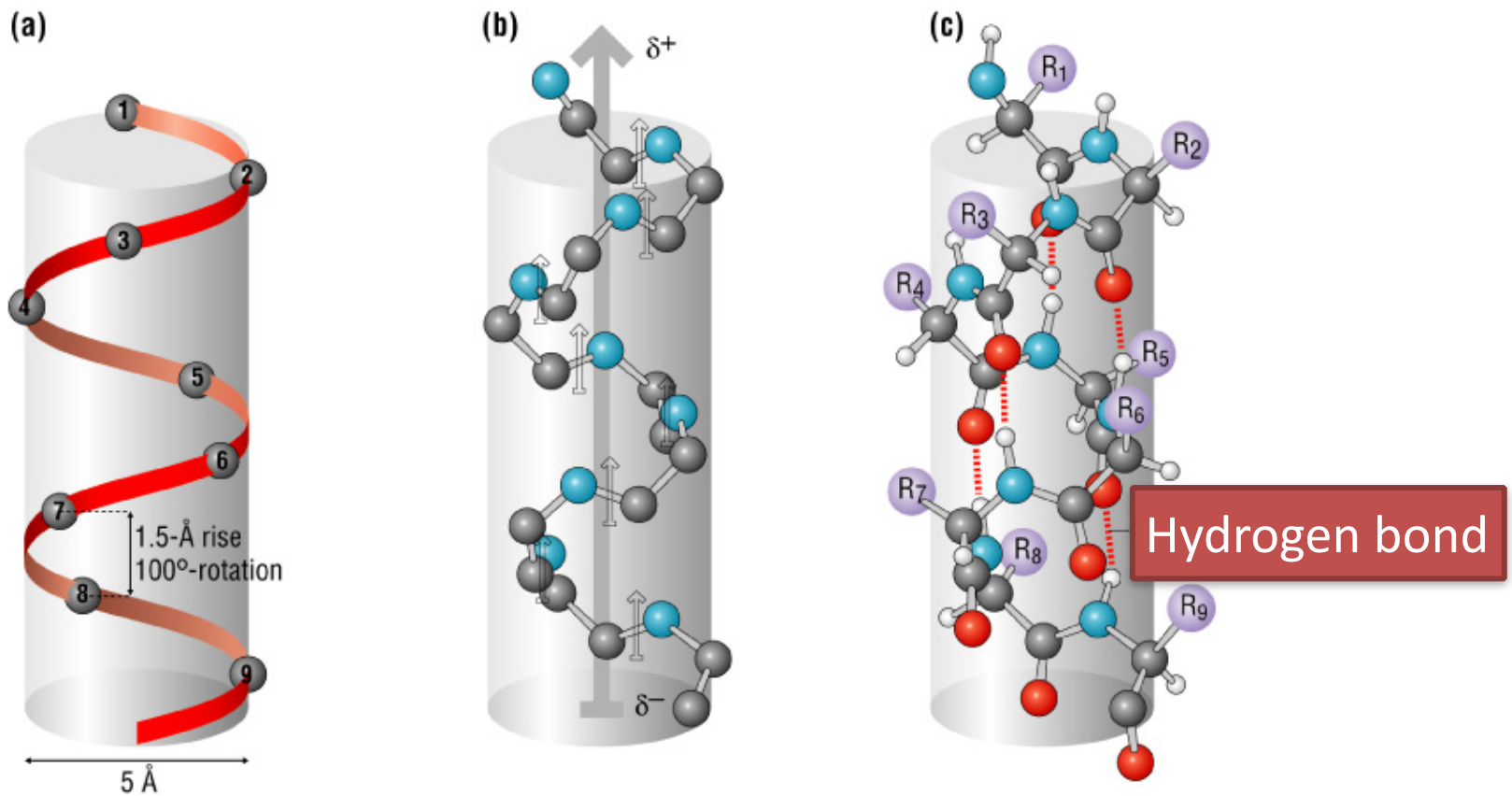
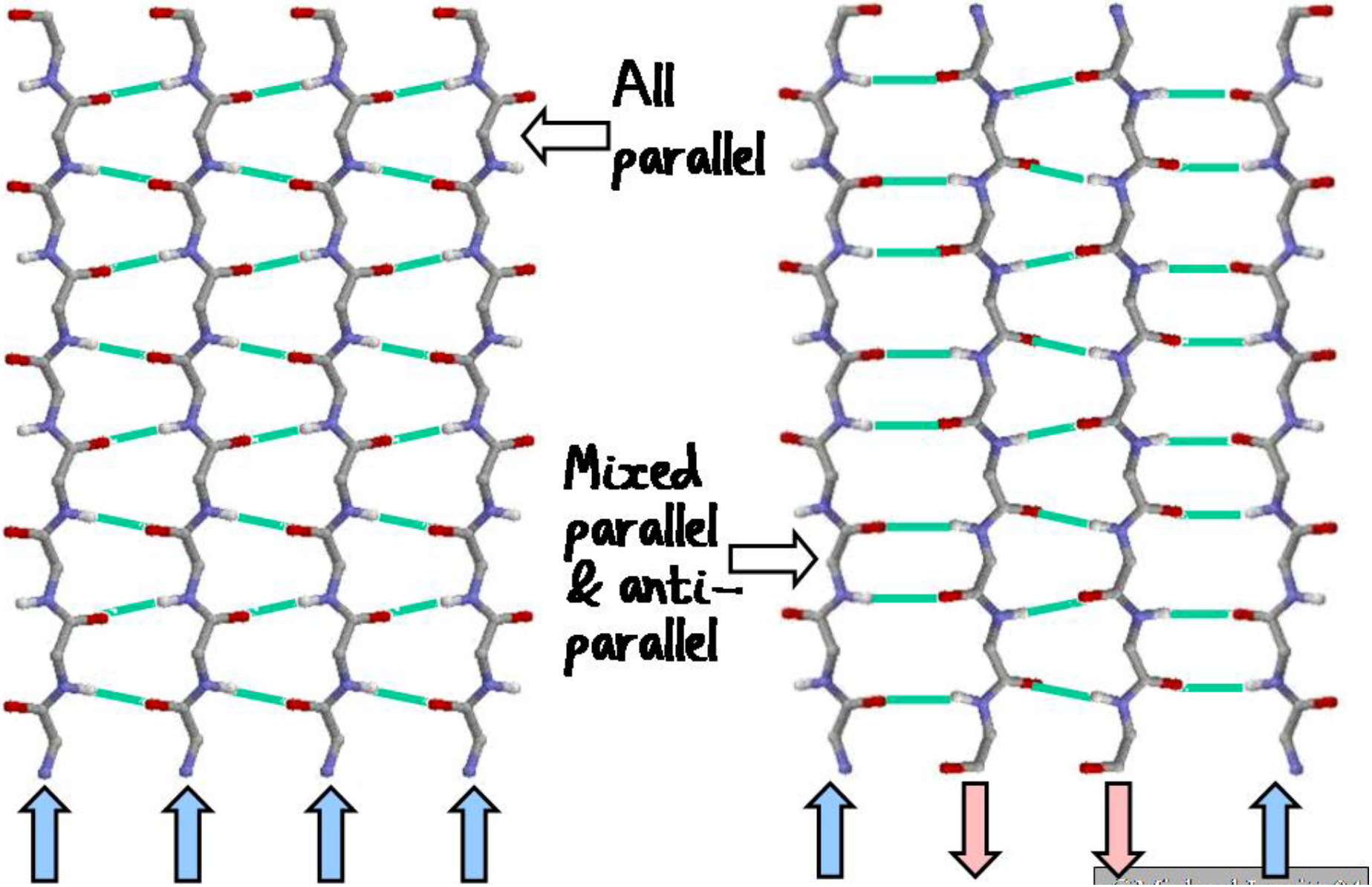


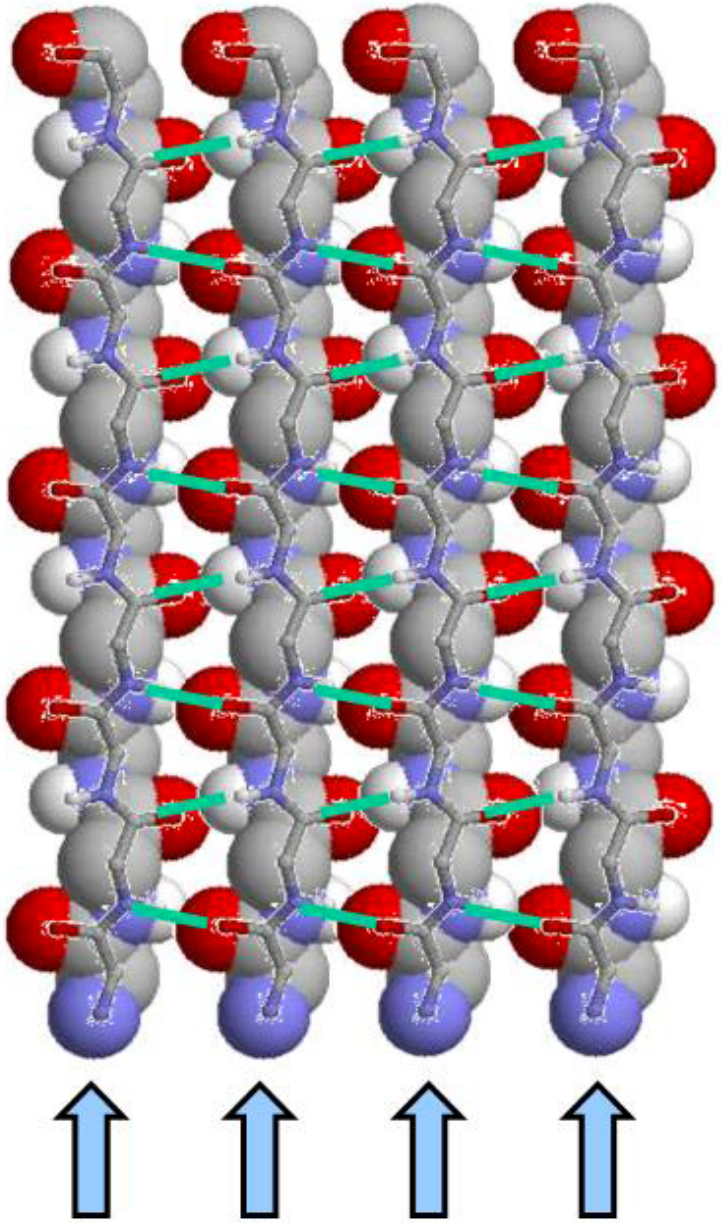
Image from "Protein Structure and Function" by Gregory A Petsko and Dagmar Ringe

The beta sheet



From Michael Levitt

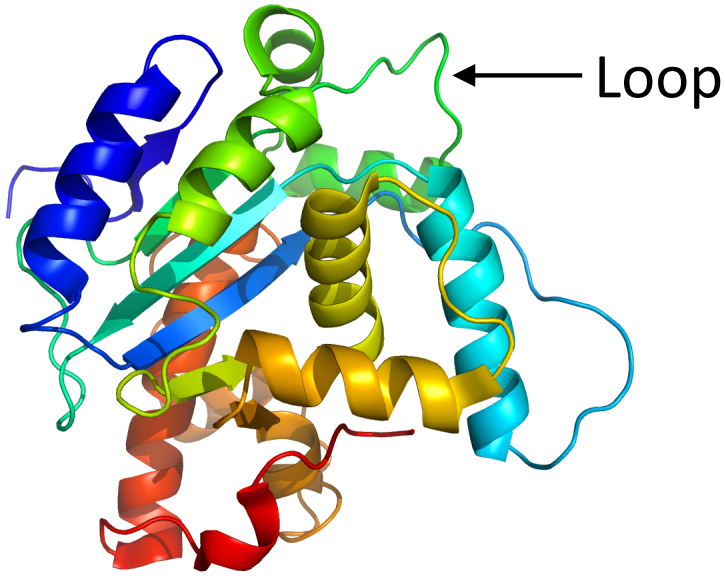
The beta sheet



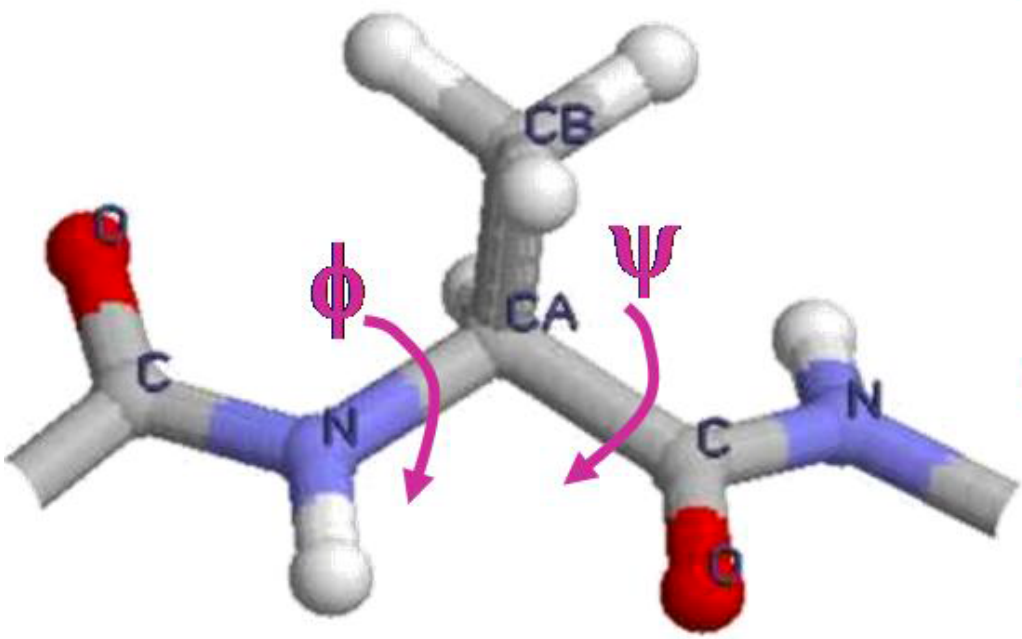
From Michael Levitt

Other secondary structure

- There are several less common secondary structures
- Regions connecting well-defined secondary structure elements are often referred to as “loops”



BACKBONE DEGREES OF FREEDOM



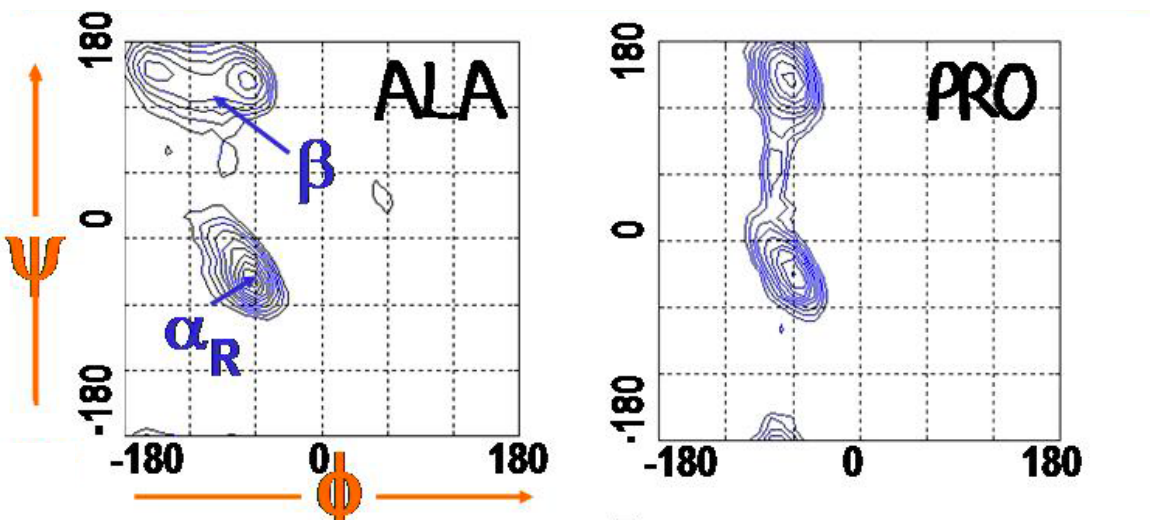
- The torsion angle rotating about the N–CA bond is called ϕ
- The torsion angle rotating about the CA–C bond is called ψ
- Together they are the (ϕ,ψ) angles

From Michael Levitt

- The remaining backbone bond (N–C, the “peptide bond”) is rigid

Ramachandran diagrams

- A plot showing a distribution in the (Φ, Ψ) plane is called a Ramachandran diagram
  - Such a diagram can be a scatterplot, or a two-dimensional histogram visualized as a contour map or heat map
  - For example, one might make a Ramachandran diagram for many residues of the same amino acid type
- Some amino acid types have distinctive Ramachandran diagrams



Ala is typical
Pro is unusual

Image from
Michael Levitt

- Alpha helices and beta sheets have characteristic Ramachandran diagrams

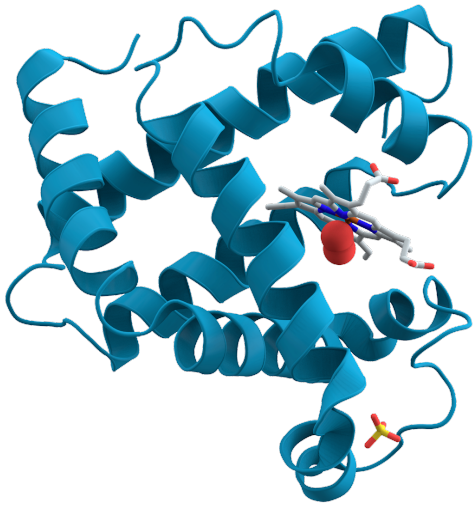
Protein structure: a more detailed view

Tertiary structure, quaternary structure, and domains

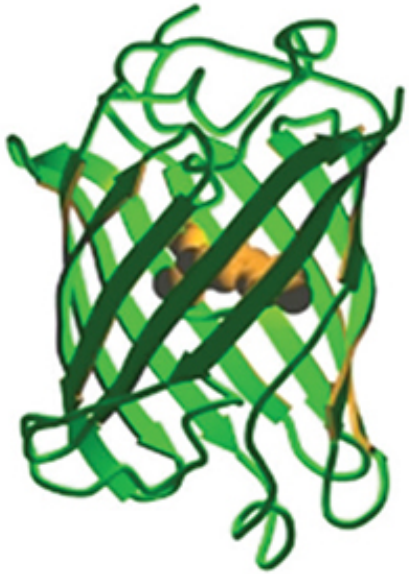
Tertiary structure

- Tertiary structure: the overall three-dimensional structure of a polypeptide chain

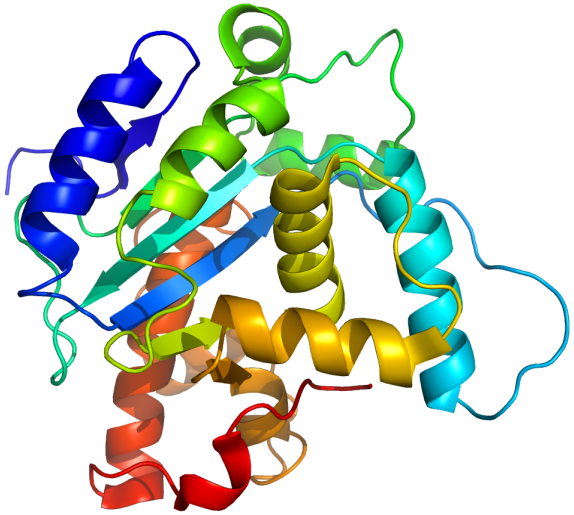
Myoglobin



Green Fluorescent Protein

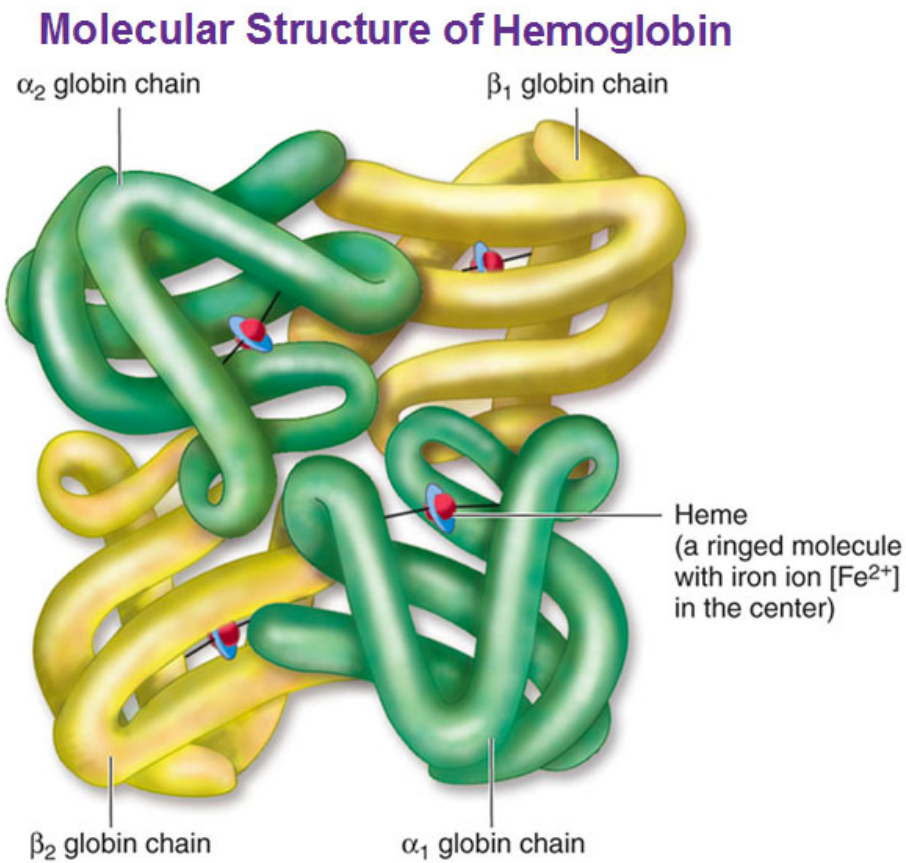


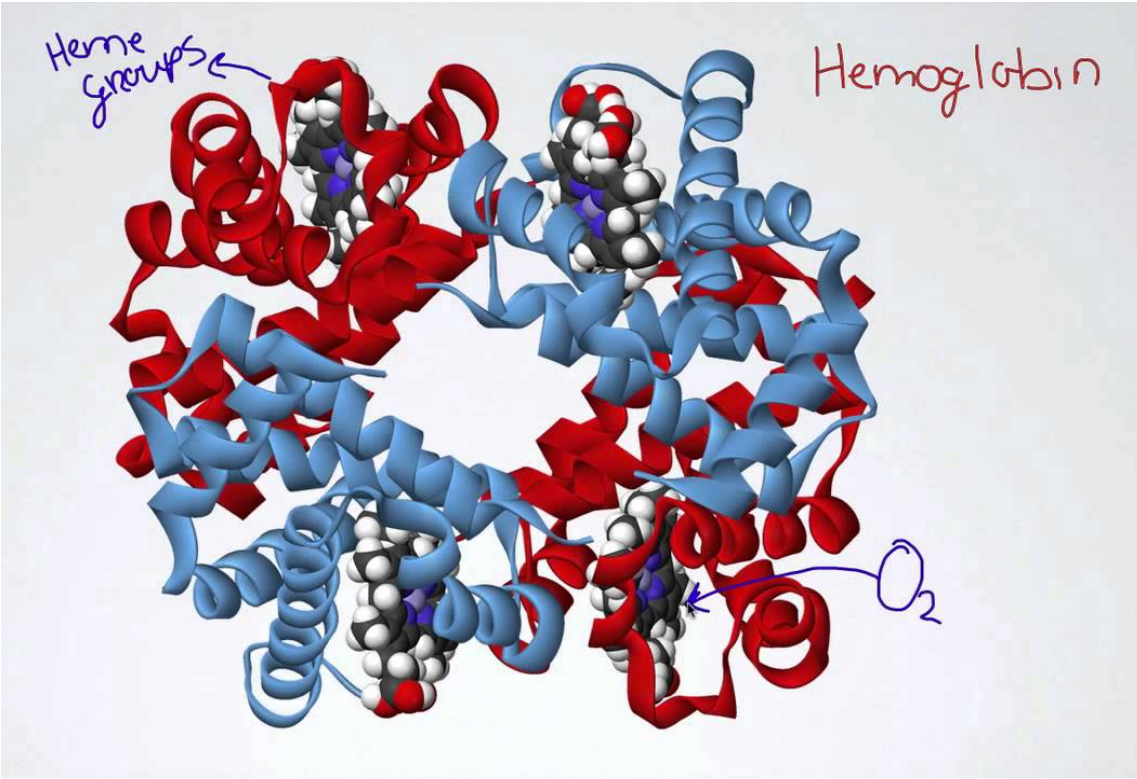
Pop2p



Quaternary structure

- Quaternary structure: the arrangement of multiple polypeptide chains in a larger protein





http://i.ytimg.com/vi/MKGhoC1Bf-I/maxresdefault.jpg

Why predict protein structure?

Problem definition

- Given the amino acid sequence of a protein, predict its three-dimensional structure
- Proteins sample many structures. We want the *average* structure, which is roughly what's measured experimentally.

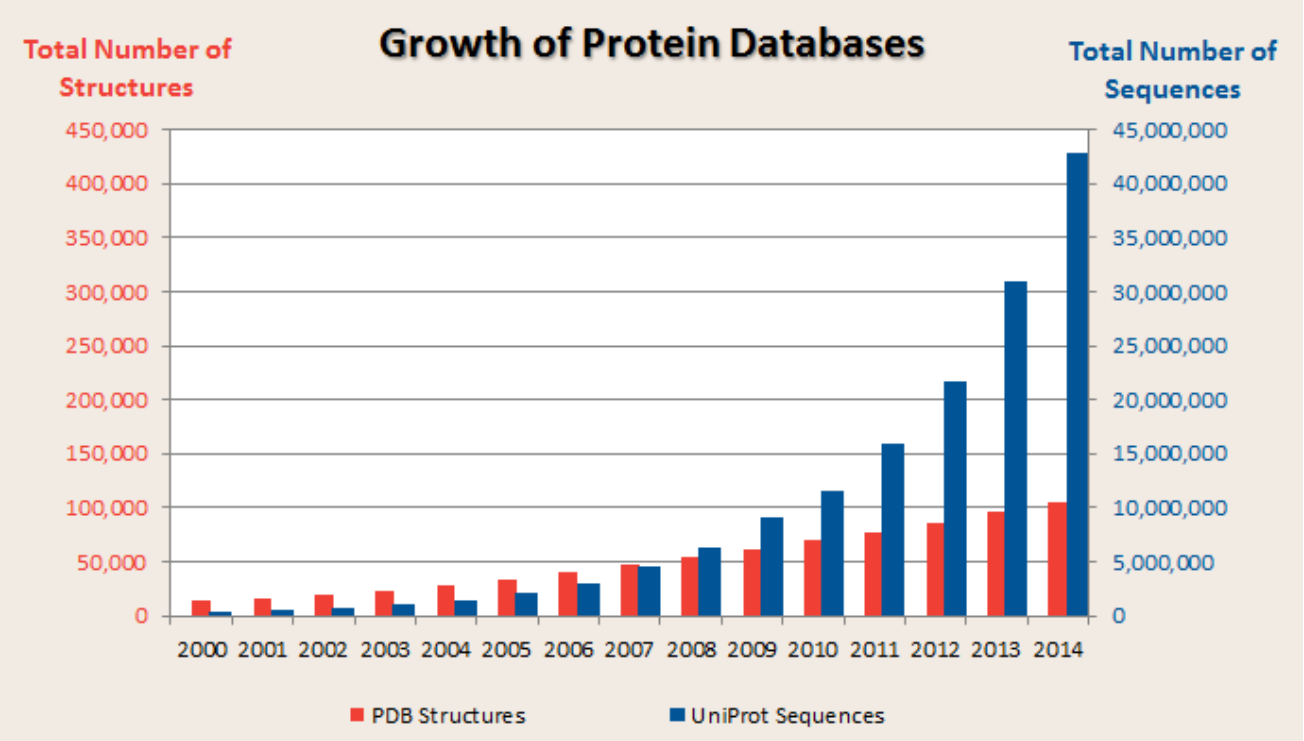
SVYDAAAQLTADVKKDLRDSW
KVIGSDKKGNGVALMTTLFAD
NQETIGYFKRLGNVSQGMAND
KLRGHSITLMYALQNFIDQLD
NPDSLDLVCS…….

How are predicted structures used?

- Drug development
  - Computational screening of candidate drug compounds
  - Figuring out how to optimize a promising candidate compound
  - Figuring out which binding site to target
- Predicting the function of a protein
- Identifying the mechanism by which a protein functions, and how one might alter that protein's function (e.g., with a drug)
- Interpreting experimental data
  - For example, a computationally predicted approximate structure can help in determining an accurate structure experimentally, as we'll see later in this course

Why not just solve the structures experimentally?

- Some structures are very difficult to solve experimentally
  - Sometimes many labs work for decades to solve the structure of one protein
- Sequence determination far outpaces experimental structure determination
  - We already have far more sequences than experimental structures, and this gap will likely grow

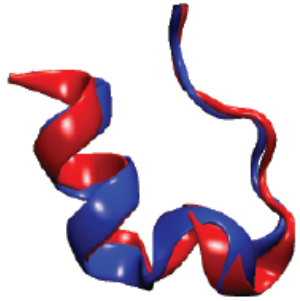


http://www.dnastar.com/blog/wp-content/uploads/2015/08/ProteinDBGrowthBar3.png

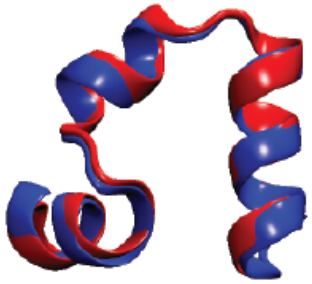
Can we use (pure) physics-based methods?

Why not just simulate the folding process by molecular dynamics?

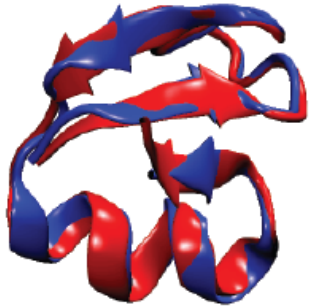
Chignolin


Trp-cage

BBA

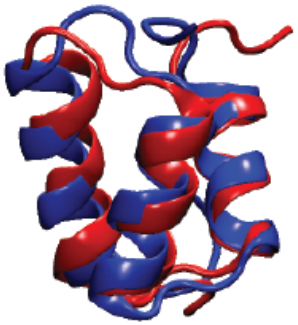

Villin

WW domain

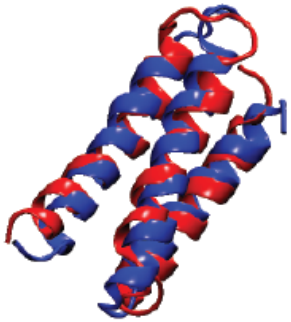

NTL9

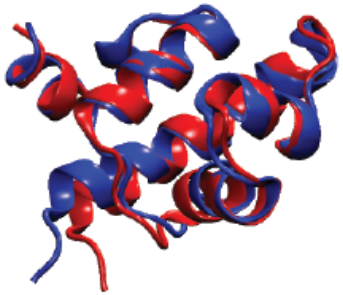
BBL

Protein B


Homeodomain

Protein G


α3D


λ-repressor

Simulation vs. experiment for 12 fast-folding proteins, up to 80 residues each

Lindorff-Larsen et al., *Science*, 2011

For most proteins, this doesn't work

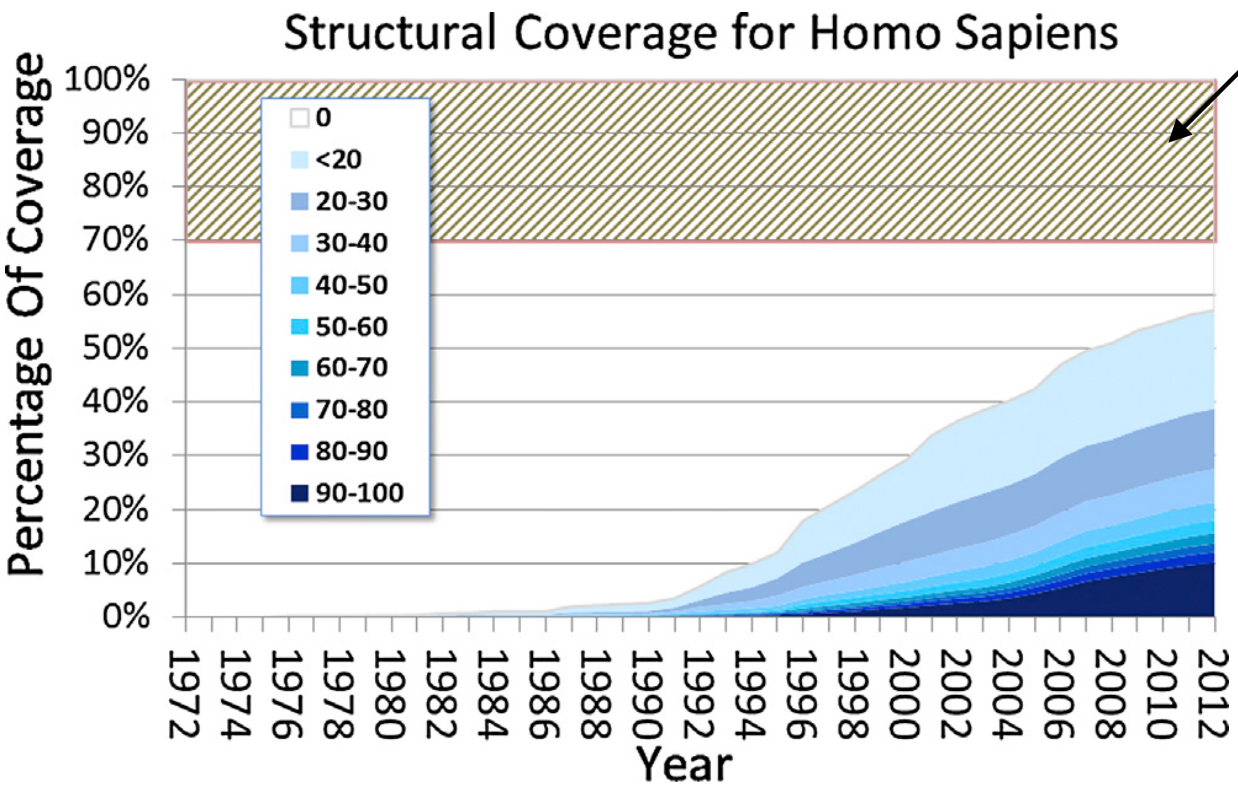
1. Folding timescales are usually much longer than simulation timescales.
2. Current molecular mechanics force fields aren't sufficiently accurate.
3. Disulfide bonds form during the real folding process, but this is hard to mimic in simulation.

Knowledge-based methods

Proteins with similar sequences tend to have similar structures

- Proteins with similar sequences tend to be homologs, meaning that they evolved from a common ancestor
- The fold of the protein (i.e., its overall structure) tends to be conserved during evolution
- This tendency is very strong. Even proteins with 15% sequence identity usually have similar structures.
  - **During evolution, sequence changes more quickly than structure**
- Also, there only appear to be 1,000–10,000 naturally occurring protein folds

For most human protein sequences, we can find a homolog with known structure



Unstructured (disordered) amino acids

The plot shows the fraction of amino acids in human proteins that can be mapped to similar sequences in PDB structures. Different colors indicate % sequence identity.

Schwede, *Structure* 2013

What if we can't identify a homolog in the PDB?

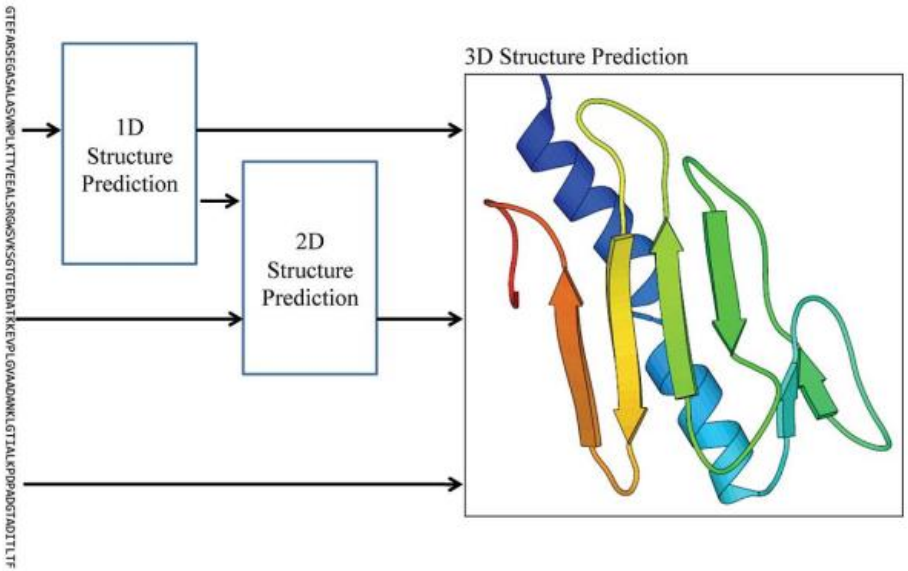
- We can still use information based on known structures
  - We can construct databases of observed structures of small fragments of a protein
  - We can use the PDB to build empirical, "knowledge-based" energy functions

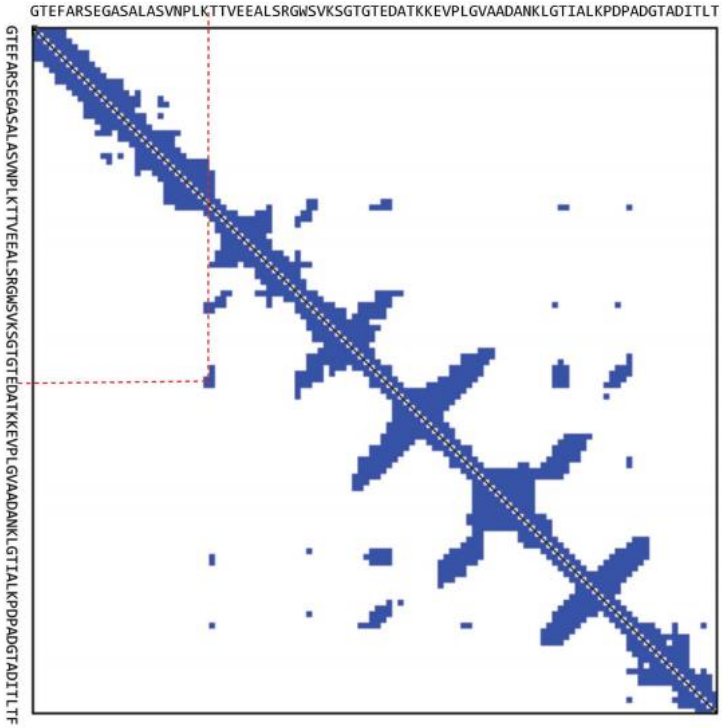
Two major approaches to protein structure prediction

Two main approaches to protein structure prediction

- Template-based modeling (homology modeling)
  - Used when one can identify one or more likely homologs of known structure
- *Ab initio* structure prediction
  - Used when one cannot identify any likely homologs of known structure
  - Even *ab initio* approaches usually take advantage of available structural data, but in more subtle ways

Supervised learning approaches for 2D contact map structure predictions





http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4664428

Refinement

- Refinement is performed using the Rosetta all-atom energy function, after building in side chains
- Refinement involves a combination of Monte Carlo moves and energy minimization
- The Monte Carlo moves are designed to perturb the structure much more gently than those used in the coarse search
  - Many still involve the use of fragments