Biomolecular structure
(including protein structure)

CS/CME/BioE/Biophys/BMI 279
Sept. 28 and Oct. 3, 2023
Ron Dror
• Please raise your hand (or comment through Panopto) if you have questions, and especially if you’re confused!

• **Tutorial on Terminal and Python**
  – Monday 7-8 pm by Zoom (link on course web page)
  – You can also view the recording afterwards
  – Recommended if you haven’t used Python or terminal (Mac, Linux) before
  – You can also get help during TA office hours
Outline

Note: I’ll discuss proteins first, as an example. These concepts apply to other biomolecules as well.

• Visualizing biomolecules (e.g., proteins)
• The Protein Data Bank (PDB)
• Chemical (2D) structure of proteins
• What determines the 3D structure of a protein? Physics underlying biomolecular structure
  – Basic interactions
  – Complex interactions
• Protein structure: a more detailed view
• Structures of other biomolecules
Visualizing biomolecules (e.g., proteins)
Protein surrounded by other molecules

Water (and salt ions)

Cell membrane (lipids)

Protein (adrenaline receptor)

All the atoms are constantly in motion
Protein only, static structure

Adrenaline receptor
Further simplified representation

Adrenaline receptor
Key take-aways from these visualizations

• Protein and surrounding atoms fill space (close-packed).
• Simplified visual representations help you figure out what’s going on.
• All of these atoms are constantly moving around, and the protein’s shape keeps changing.
  – When we talk about “the” 3D structure of a protein, we really mean an average structure—and even that depends on the experimental conditions (e.g., which other molecules are bound to the protein)
The Protein Data Bank (PDB)
The Protein Data Bank (PDB)

• Examples of structures from the PDB

https://upload.wikimedia.org/wikipedia/commons/thumb/2/24/Protein_structure_examples.png/1024px-Protein_structure_examples.png
(Axel Griewel)
The Protein Data Bank (PDB)

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.

COVID-19 CORONAVIRUS Resources

SARS-CoV-2 RNA-dependent RNA Polymerase
6YYT
Structure of replicating SARS-CoV-2 polymerase
The Protein Data Bank (PDB)

- [https://www.rcsb.org/](https://www.rcsb.org/)
- A collection of essentially all published, experimentally determined structures of biomacromolecules (e.g., proteins, DNA, RNA)
- Each identified by 4-character code (e.g., 6YYT)
- Currently ~210,000 structures. ~80% of those were determined by x-ray crystallography.
- Browse it and look at some structures. Options:
  - 3D view in applet on PDB web pages
  - PyMOL: fetch 6YYT
Chemical (two-dimensional) structure of proteins
Chemical (two-dimensional) structure vs. three-dimensional structure

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

2D structure

![2D structure image](https://en.wikipedia.org/wiki/Ethanol)

3D structure

![3D structure image](https://en.wikipedia.org/wiki/Ethanol)

Proteins are built from amino acids

- 20 “standard” amino acids
- Each has three-letter and one-letter 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
Asparagine
THAT'S RIGHT, FOUR EYES! YOU'RE NOTHING WITHOUT ME! WHILE I'M AN ESSENTIAL PART OF ANY PROTEIN, EVEN YOURS, YOU'RE STILL A SO-SO PROFESSOR WITH NO CHANCE OF TENURE! HAHAHA!

A MEAN O' ACID

Source unknown. American Scientist?
Proteins are chains of amino acids

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

\[
\begin{align*}
\text{H}_2\text{N-C}=\text{C}-\text{COOH} & \quad + \quad \text{H}_2\text{N-C}=\text{C}-\text{COOH} \\
\quad + \quad \text{H}_2\text{O} & \quad \rightarrow \quad \text{H}_2\text{N-C}=\text{C}-\text{C}=\text{N-C}=\text{C}-\text{COOH} \\
\end{align*}
\]
http://en.wikipedia.org/wiki/Condensation_reaction

- Strictly speaking, elements of the chain are amino acid residues. They are usually called "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


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

• To a first approximation (which suffices for the purposes of this course), we can think of an atom simply as a sphere.

• It occupies a position in space, specified by the (x, y, z) coordinates of its center, at a given point in time.
Geometry of a molecule

• A molecule is a set of atoms connected in a graph
• \((x, y, z)\) coordinates of every atom specify the molecule’s geometry
• Alternatively, we can specify the geometry of a molecule using bond lengths, bond angles, and torsion angles.
Forces between atoms

• We can approximate the total potential energy of a molecular system as a sum of individual contributions. Terms are additive.
  – Thus force on each atom is also a sum of individual contributions.
    • Remember: force is the derivative of energy.
  – We will ignore quantum effects. Think of atoms as balls and forces as springs.

• Two types of forces:
  – Bonded forces: act between closely connected sets of atoms in the graph of covalent bonds
  – Non-bonded forces: act between all pairs of atoms
Bond length stretching

- A covalently bonded pair of atoms is effectively connected by a “spring” with some preferred (natural) length. Stretching or compressing this spring requires energy.
Bond angle bending

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

• Certain values of each torsional angle are preferred over others.
Electrostatic interaction

- Like charges repel. Opposite charges attract.
- Electrostatic forces act 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 its element type and on which other 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 most proteins depends on the fact that it is surrounded by water.
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)  
\[
\text{H} \quad \text{H} \\
\text{H-C-O-H} \\
\text{H}
\]

Hydrophobic (apolar)  
\[
\text{H} \\
\text{H-C-H} \\
\text{H}
\]
Hydrophobic effect

- Hydrophobic molecules cluster in water
  - “Oil and water don’t mix”

- This is critical to protein structure
We will discuss entropy next week. If this isn’t clear now, don’t worry.

EXPLAINING HYDROPHOBICITY

- Water molecules next to solute cannot move freely.
- They are ordered and have less entropy. They are unhappy.
- The system changes so that fewer water molecules are in the surface layer.
- The hydrophobic solutes aggregate.

Slide from Michael Levitt
Protein structure: a more detailed view
Protein structure: a more detailed view

Properties of amino acids
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
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
  – Negatively charged (acidic) side chains attract positively charged (basic) side chains
  – Hydrophilic side chains form hydrogen bonds to one another and to water molecules
  – Hydrophobic side chains “want” to be near one another
Amino acid properties

You don’t need to memorize which amino acids have which properties

There are many properties.
They cluster logically.
Protein structure: a more detailed view

Secondary structure elements
Secondary structure elements

• Some local structural patterns are found in most proteins
  – These are called “secondary structure elements” These are energetically favorable primarily because of hydrogen bonds between backbone atoms

• Most common secondary structure elements:
  – alpha helix
  – beta sheet

https://upload.wikimedia.org/wikipedia/commons/e/e6/Spombe_Pop2p_protein_structure_rainbow.png
https://www.biotek.com/assets/tech_resources/11596/figure2.jpg
http://upload.wikimedia.org/wikipedia/commons/6/60/Myoglobin.png
The alpha helix

Image from “Protein Structure and Function” by Gregory A Petsko and Dagmar Ringe
The alpha helix

Linus Pauling

https://www.msu.edu/course/lbs/333/fall/images/PAULING.JPG
Reserved for
Nobel Laureate

Nobel Laureate Reserved Space
Parking Permit
Required At All Times

Violators will be cited and/or towed
Per UCB violation codes 60—120/VC 22651n
For Towed Vehicles Call 510.642.7000

2 COLI
RESERVED
PARKING
COLLEGE OF CHEMISTRY
9/30/2016
6:30 AM-12:00 PM
Stephen Isaacs

RESERVED PARKING
COLLEGE OF CHEMISTRY
9/30/2016
1:00 PM-6:00 PM
Prof. Ron Dror
A beta sheet is made up of two or more beta strands, connected by hydrogen bonds. From Michael Levitt.
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
Quaternary structure

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

![Molecular Structure of Hemoglobin](http://i.ytimg.com/vi/MKGhoC1Bf-I/maxresdefault.jpg)
Domains

• Large proteins often consist of multiple compact 3D structures called domains
  – Many contacts within a domain. Few contacts between domains.
  – “Domain ≈ blob”

• One polypeptide chain can form multiple domains, and a single domain may include portions of several polypeptide chains

http://en.wikipedia.org/wiki/Protein_domain
Protein structure: a more detailed view

Describing protein backbone structure
We only need two backbone torsion angles per amino acid residue because the third backbone bond (N–C, the “peptide bond”) is rigid. Specifying side chain structure requires additional torsion angles. This is a useful way to specify protein structure—used, for example, in recent large language models for proteins.
Ramachandran diagrams

- A plot showing a distribution in the $(\Phi, \Psi)$ 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

- Alpha helices and beta sheets have characteristic Ramachandran diagrams

-Ala is typical
-Pro is unusual

Image from Michael Levitt
One more note: Disulfide bonds

- One particular amino acid type, cysteine, can form a covalent bond with another cysteine (called a disulfide bond or bridge)
- Disulfide bonds often connect amino acid residues that are distant in the peptide chain
- In a typical cellular environment, disulfide bonds can be formed and broken quite easily

Cysteine side chains

Disulfide bond formation

http://www.crc.dk/yeast/yeasthome/yeasthome/images/ls_jpgs/fig2.jpg
Structures of other biomolecules
What determines the structure of other biomolecules?

- The physical interactions that determine protein structure also determine the structures of other biomolecules
  - More generally, the great majority of the material covered in this course for proteins applies to other biomolecules as well
DNA

• DNA (deoxyribonucleic acid) stores the genetic code

• DNA, like protein, is a string of units with a uniform backbone
  – The units are nucleotides, instead of amino acid residues
  – Different nucleotides contain different nucleobases ("bases") instead of side chains

• Only four common DNA bases
  – Adenine pairs with Thymine
  – Guanine pairs with Cytosine
DNA

- DNA forms one dominant 3D structure: a double helix
  - DNA usually acts more as information storage than as “machinery”
  - Long stretches of double helix can form coarser-scale structures

http://www.nature.com/scitable/content/ne0000/ne0000/ne0000/104944953/73_1_2.jpg
Cambridge, 1953. Shortly before discovering the structure of DNA, Watson and Crick, depressed by their lack of progress, visit the local pub.
"It's not supposed to be a triple helix, is it?"
RNA

- RNA (ribonucleic acid) is a string of nucleotides, like DNA
- RNA, however, frequently occurs as a single string (strand) rather than paired strands
- RNA bases often pair with other bases in the same RNA strand
  - Much work on RNA structure focuses on the “secondary structure”: which bases pair with one another
  - Note that “secondary structure” has different meanings for RNA and protein
- Some RNAs store the genetic code of proteins, but most serve other functions
- RNAs usually form “machines” with well-defined, varied 3D structure

Typically:
Guanine & Cytosine pair
Adenine & Uracil pair
• Frequently, a single RNA is made up of multiple strands
  – Bases pair across strands
  – Secondary structure often includes multiple strands
Glycans (e.g. carbohydrates)

- The base units are called “monosaccharides”
- When they are linked through glycosidic bond, they are called glycans
- Examples: starch, cellulose, chitin
- In cells, glycans are often attached to proteins (“glycosylation”)

https://www.khanacademy.org/science/biology/macromolecules/carbohydrates-and-sugars/a/carbohydrates
Small molecules

- Most drugs and many hormones, neurotransmitters, and other natural signaling molecules are “small molecules” (~100 atoms or fewer)
- Cambridge Structural Database is a repository of small molecule 3D structures, generally from x-ray crystallography
- However, these molecules are usually highly flexible and thus likely to take on a different 3D structure when bound to a protein

Adrenaline (epinephrine)

LSD on its own (yellow) and receptor-bound (magenta)

Wacker et al., Cell (2017)
Assignment 1

• Available on course website (cs279.stanford.edu)
• Due Tuesday, Oct. 17, at 1 p.m.
  – Be sure to start soon, particularly to verify that you have necessary software working (including ability to call matplotlib from within PyMOL)
• Assignment 1 kickstart and tutorial: this Thursday at 6 p.m.
• Options for computer use:
  – If you live or work on or near campus, we strongly recommend using one of many physical LTS clusters that have all necessary software pre-installed.
  – Otherwise, you can use LTS machines remotely—but plan to finish well before due date, as very few are available for remote use this year.
  • Please avoid using this pool if you live or work on campus.
  – If you enjoy command-line software installation, you can install the software on your own Mac (OSX) or Linux computer. Windows installation may be more challenging.
A couple clarifications

• Nearly all PDB files don’t specify bonds. When you load a PDB file into PyMOL, how can it display the covalent bonds?
  – It infers them automatically from the spatial coordinates of the atoms

• What does “solving” a structure mean?
  – Determining it experimentally (which requires “solving” a computational problem to get atomic coordinates)
Optional reading

- On the course website, we’ll include links to papers or other materials recommended for those who wish to learn more about each lecture topic.

- This material is for students interested in learning more. It’s strictly optional.
A caveat

• This course covers a rapidly developing field. Published papers use different terminology and sometimes make contradictory claims. This includes papers I suggest as optional reading.