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

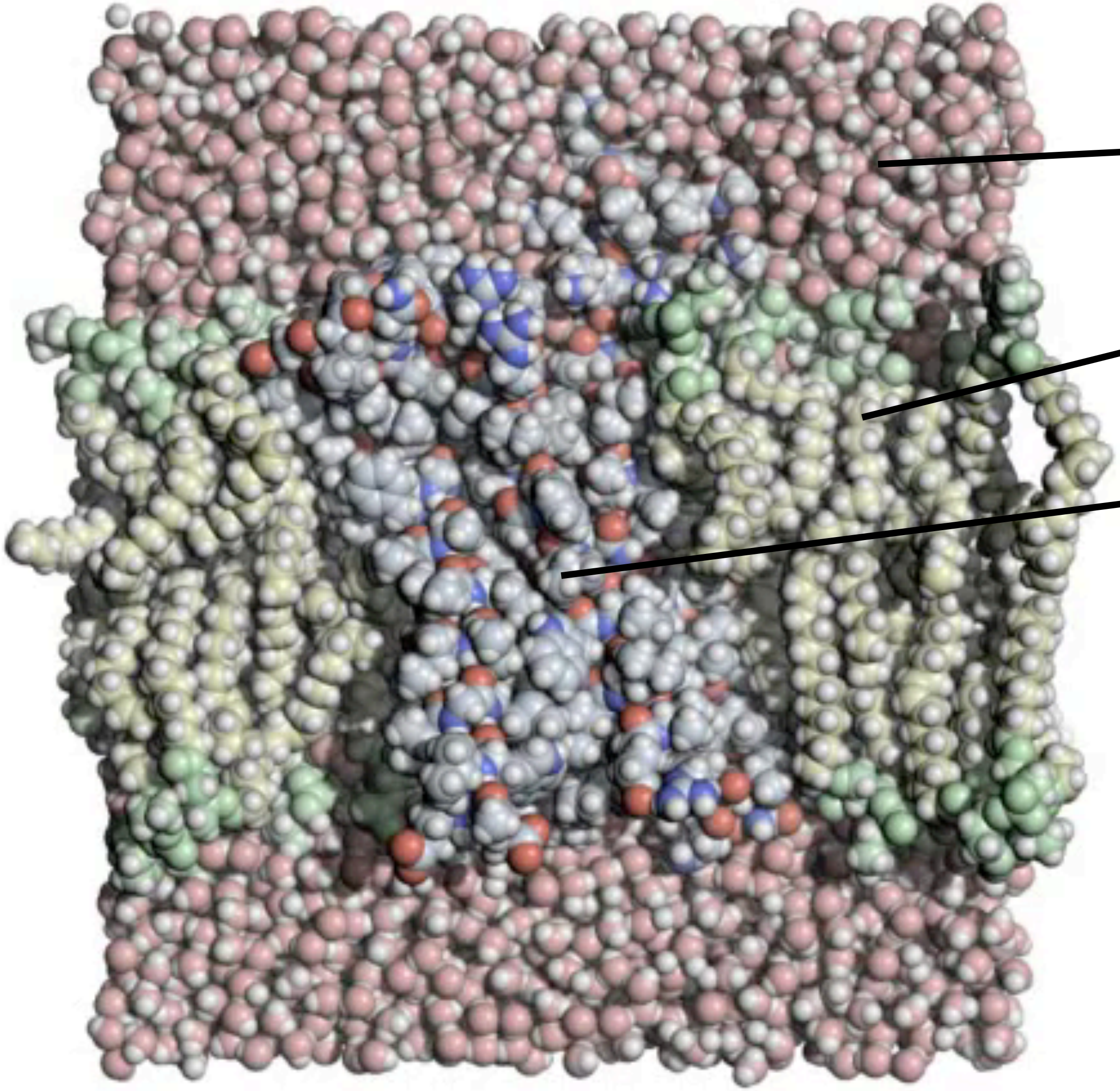
Outline

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



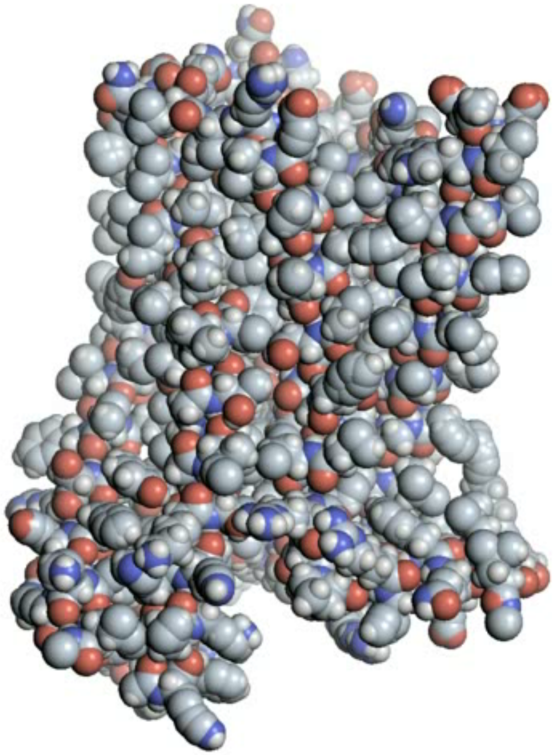
Water (and salt ions)

Cell membrane (lipids)

Protein
(adrenaline receptor)

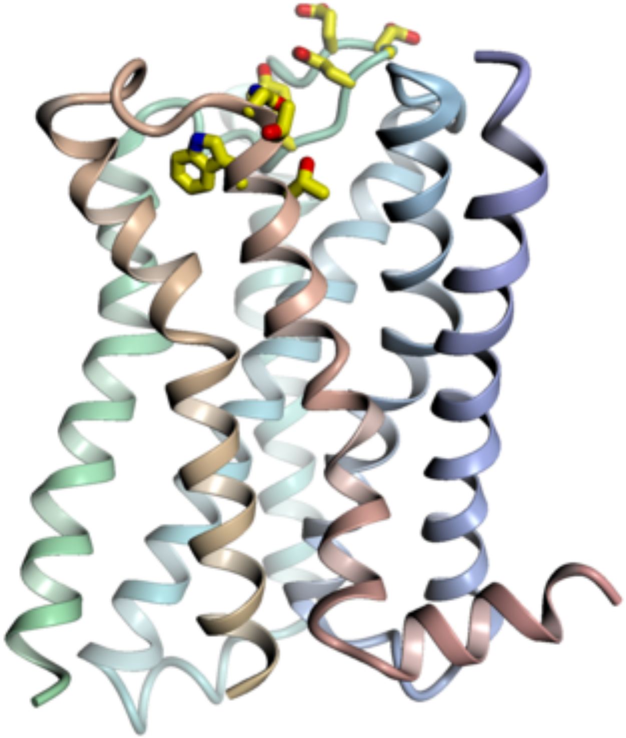
All the atoms are
constantly in motion

Protein only, static structure



Adrenaline receptor

Further simplified representation



Adrenaline receptor

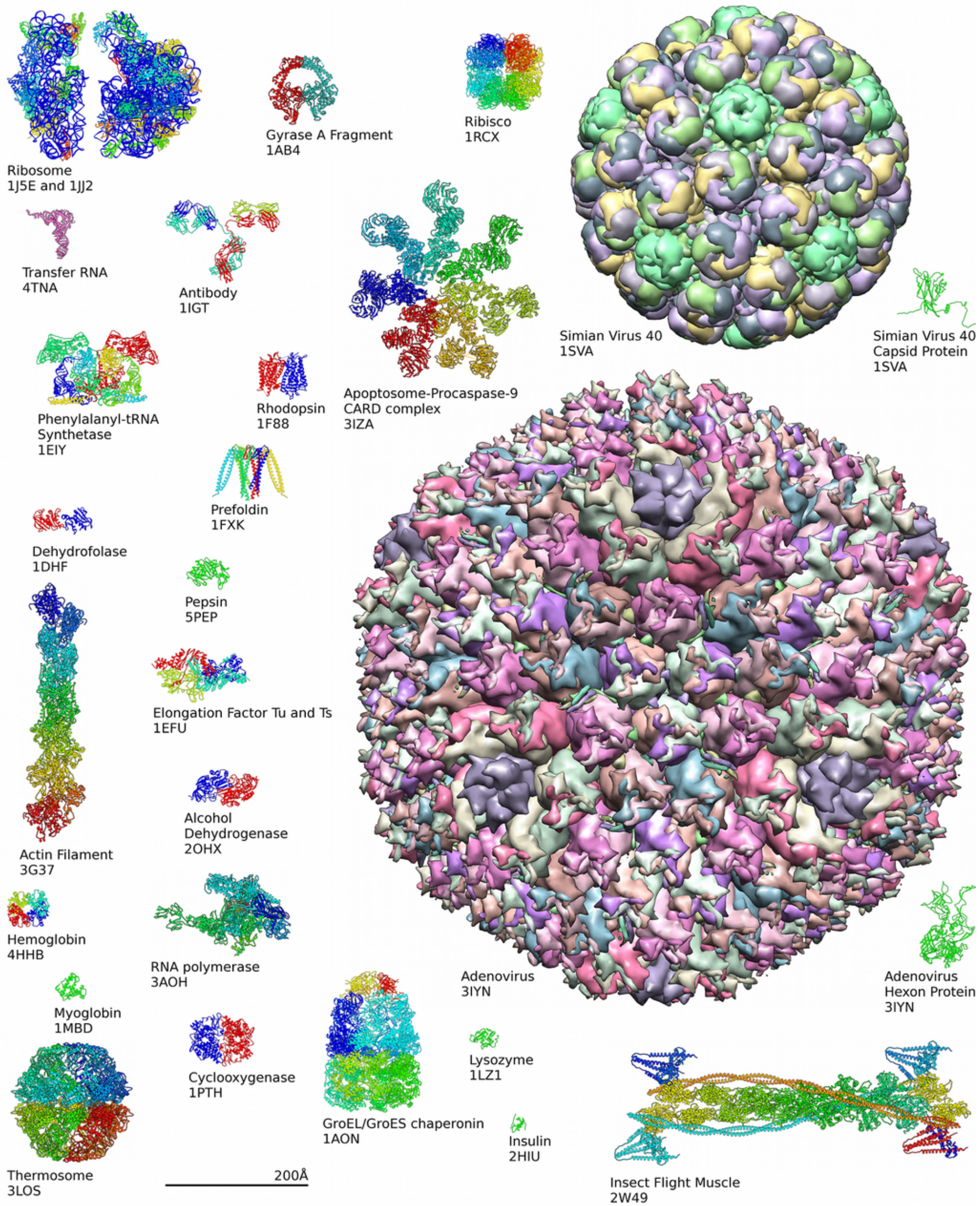
Key take-aways from these visualizations

- 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)

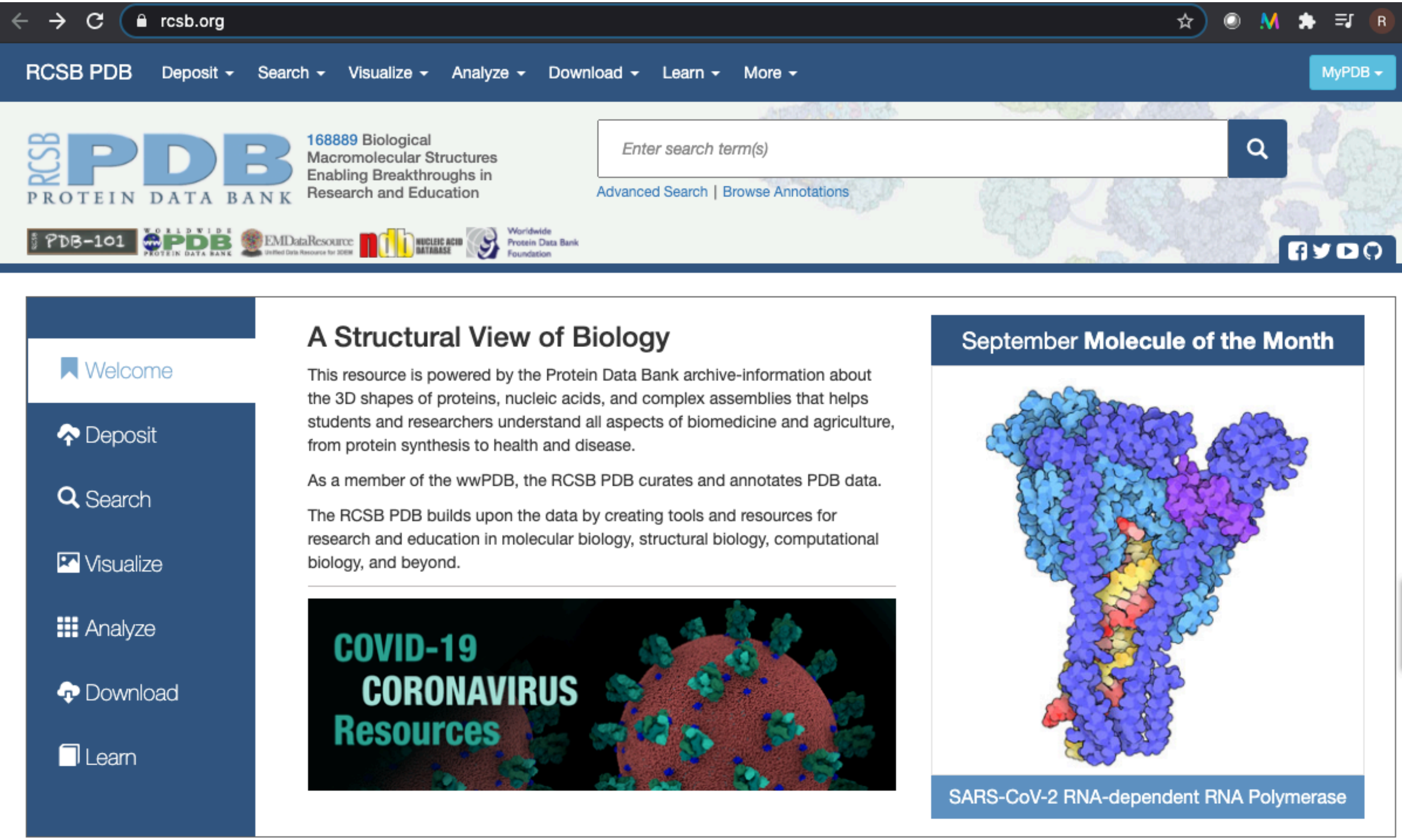
The Protein Data Bank (PDB)

- 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)



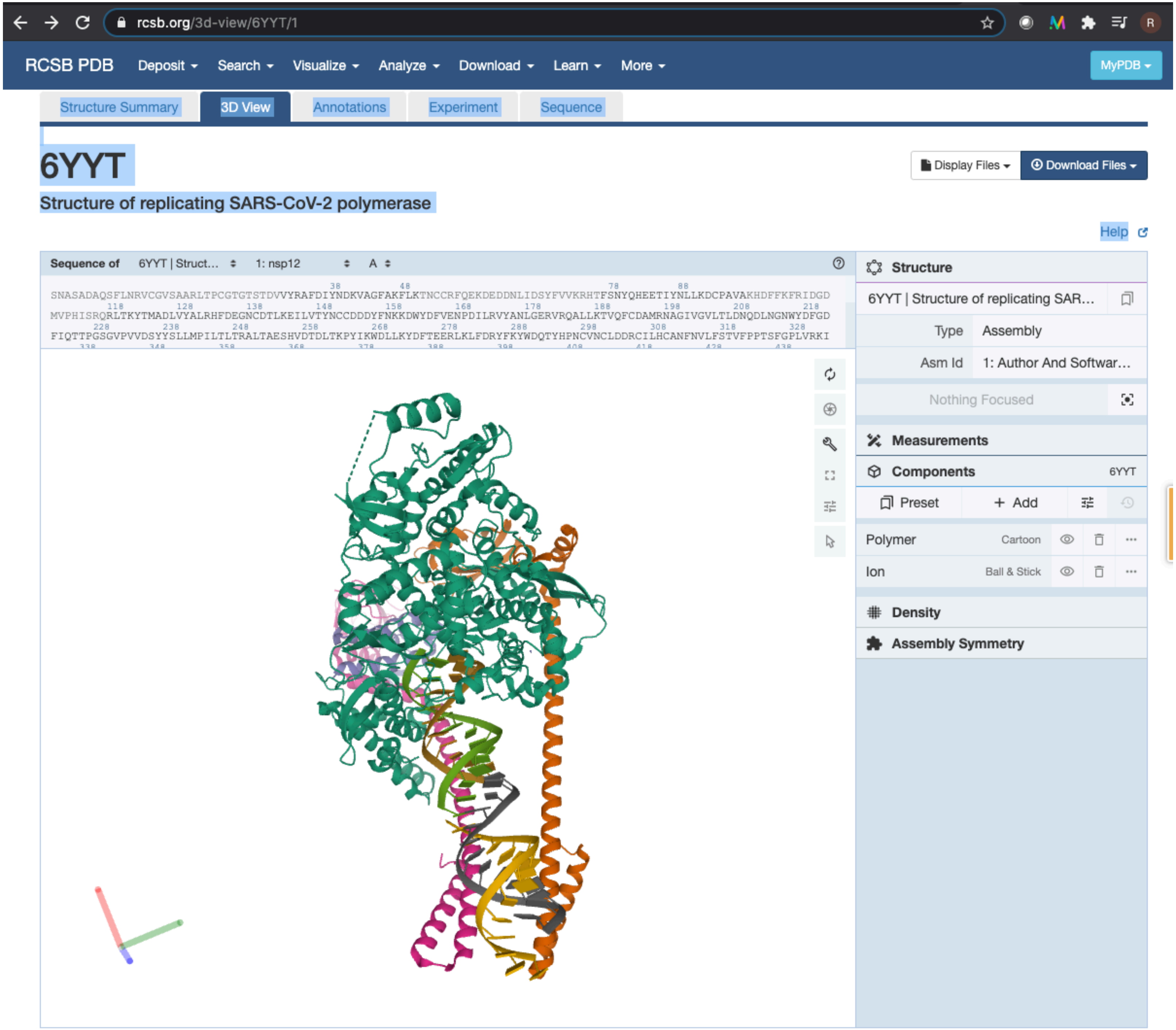
168889 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

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.



rcsb.org/3d-view/6YYT/1

6YYT

Structure of replicating SARS-CoV-2 polymerase

Sequence of 6YYT | Struct… 1: nsp12 A

```
                                          38        48                              78        88
SNASADAQSFLNRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVAGFAKFLKTNCCRFQEKDEDDNLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAKHDFFKFRIDGD
        118       128       138       148       158       168       178       188       198       208       218
MVPHISRQRLTKYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNKKDWYDFVENPDILRVYANLGERVRQALLKTVQFCDAMRNAGIVGVLTLDNQDLNGNWYDFGD
      228       238       248       258       268       278       288       298       308       318       328
FIQTTPGSGVPVVDSYYSLLMPILTLTRALTAESHVDTDLTKPYIKWDLLKYDFTEERLKLFDRYFKYWDQTYHPNCVNCLDDRCILHCANFNVLFSTVFPPTSFGPLVRKI
```

Structure

6YYT | Structure of replicating SAR…

Type: Assembly

Asm Id: 1: Author And Softwar…

Nothing Focused

Measurements

Components 6YYT

Polymer	Cartoon
Ion	Ball & Stick

Density

Assembly Symmetry

The Protein Data Bank (PDB)

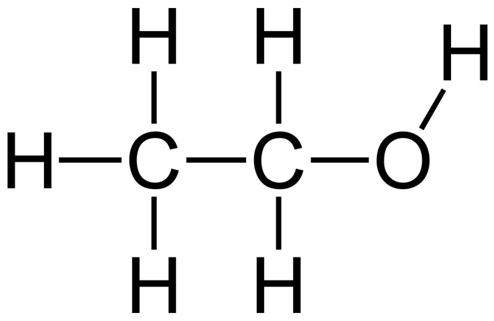
- https://www.rcsb.org/
- A collection of essentially all published, experimentally determined structures of biomacromolecules (e.g., proteins, DNA, RNA)
- Each identified by 4-character code (e.g., 6YYT)
- Currently ~210,000 structures. ~80% of those were determined by x-ray crystallography.
- Browse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

Chemical (two-dimensional) structure of proteins

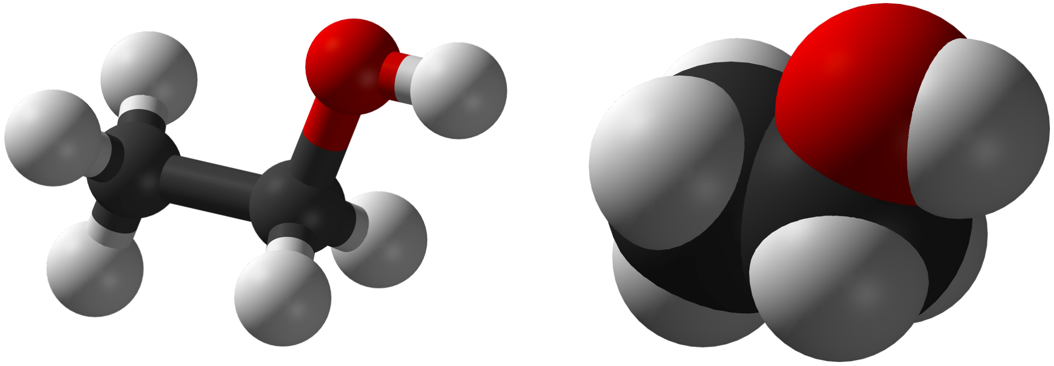
Chemical (two-dimensional) structure vs. three-dimensional structure

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

2D structure



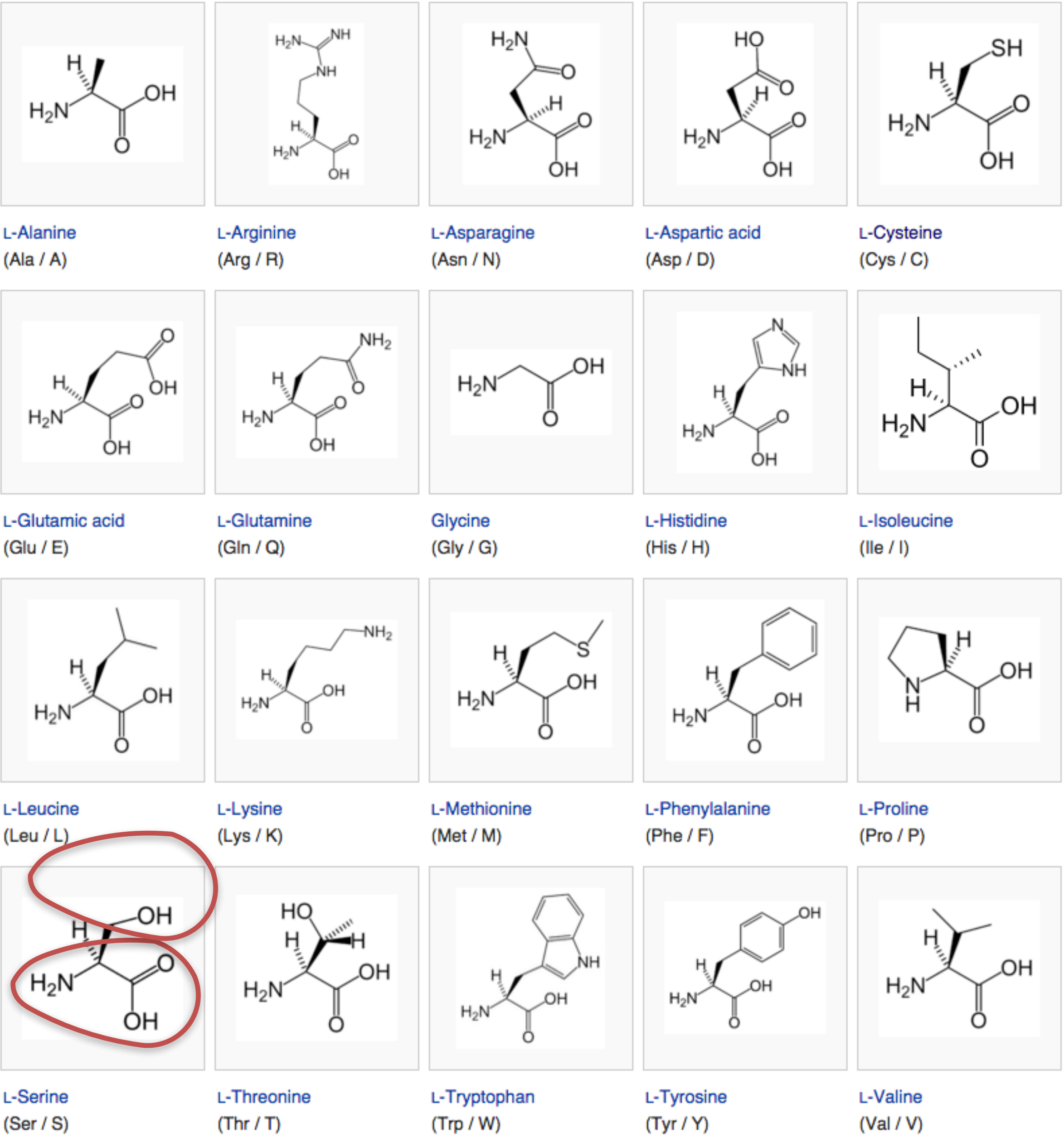
3D structure



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

Proteins are built from amino acids

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



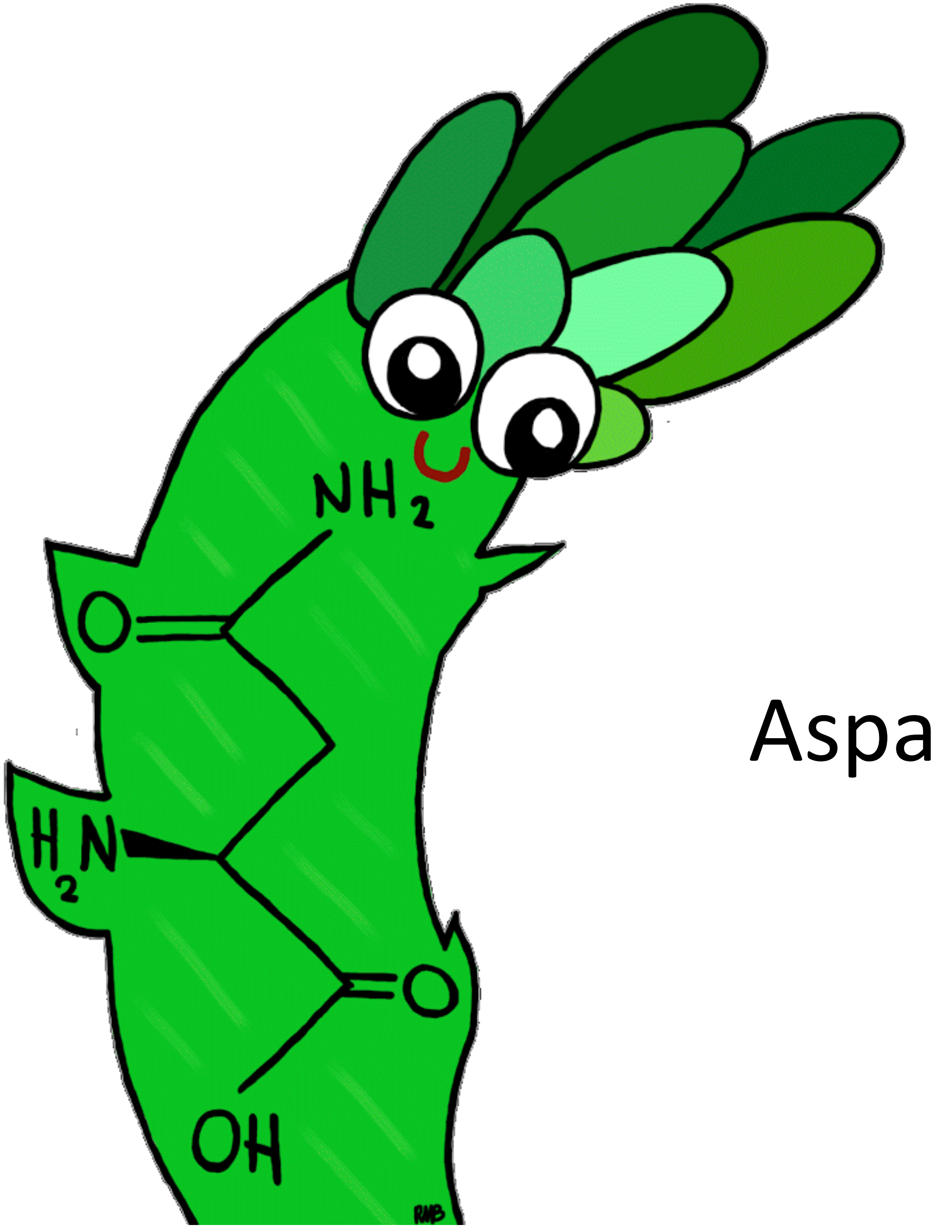
L-Alanine (Ala / A)
L-Arginine (Arg / R)
L-Asparagine (Asn / N)
L-Aspartic acid (Asp / D)
L-Cysteine (Cys / C)
L-Glutamic acid (Glu / E)
L-Glutamine (Gln / Q)
Glycine (Gly / G)
L-Histidine (His / H)
L-Isoleucine (Ile / I)
L-Leucine (Leu / L)
L-Lysine (Lys / K)
L-Methionine (Met / M)
L-Phenylalanine (Phe / F)
L-Proline (Pro / P)
L-Serine (Ser / S)
L-Threonine (Thr / T)
L-Tryptophan (Trp / W)
L-Tyrosine (Tyr / Y)
L-Valine (Val / V)

The “side chain” is different in each amino acid.

All amino acids have this part in common.

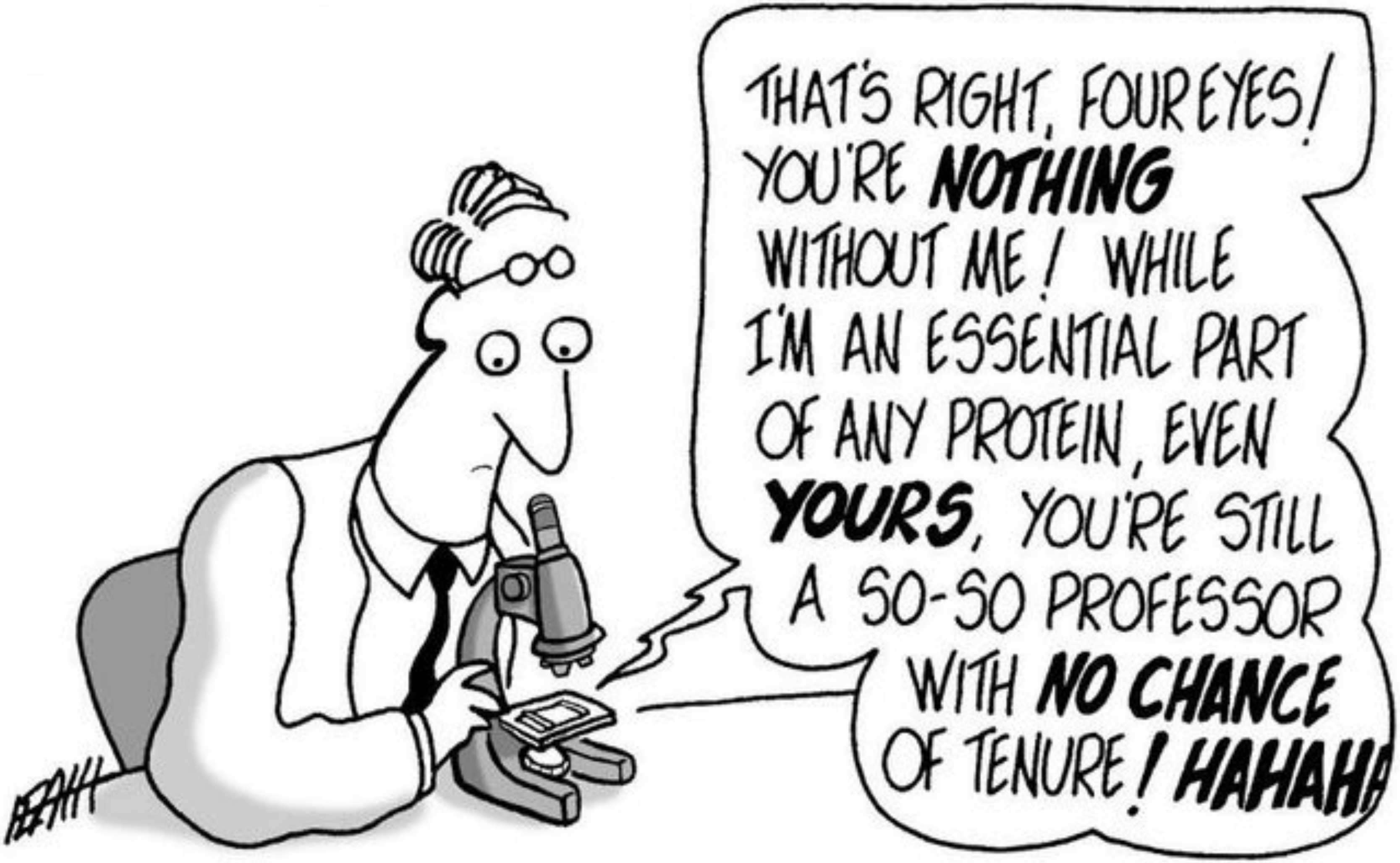
You don’t need to memorize all the structures

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



Asparagine

by Robin Betz

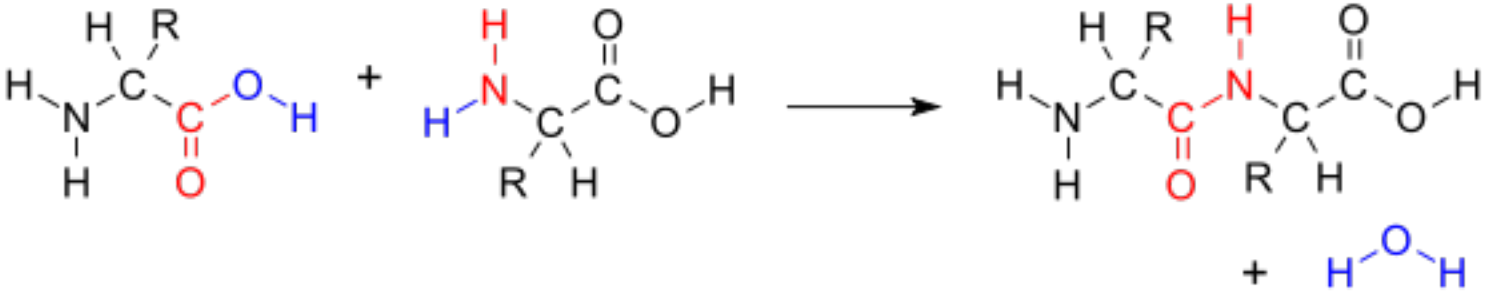


a mean o' acid

Source unknown. *American Scientist*?

Proteins are chains of amino acids

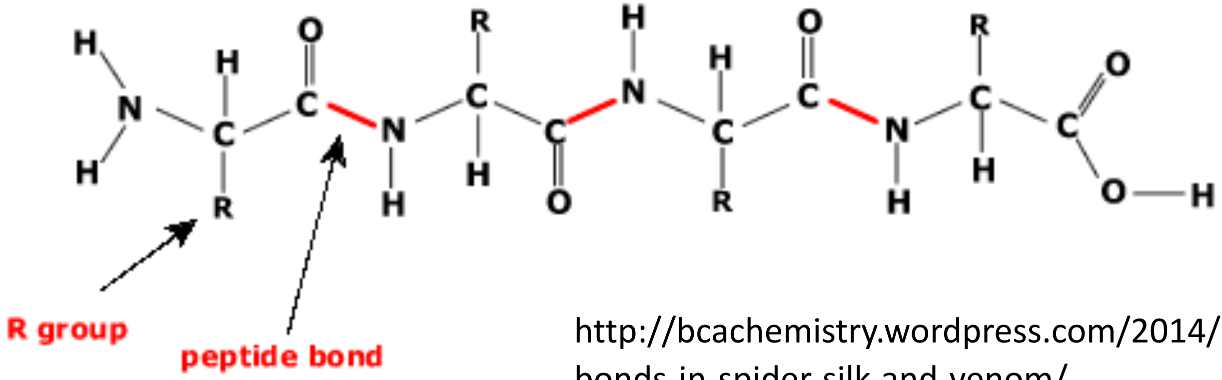
- Amino acids link together through a chemical reaction ("condensation")



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

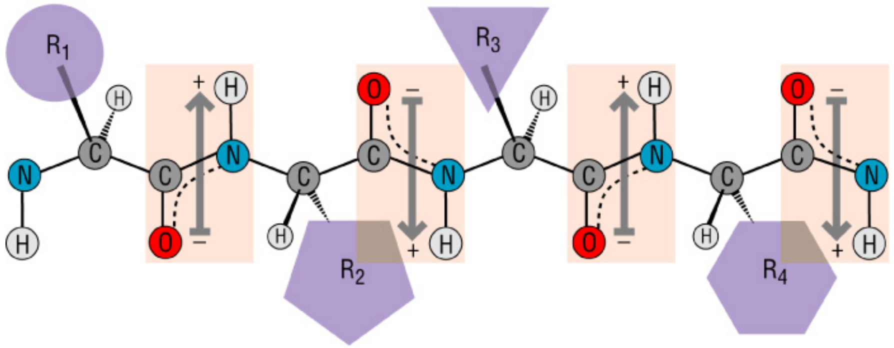
- Strictly speaking, elements of the chain are amino acid *residues*. They are usually called"**residues**" (important term!)
- The bonds linking these residues are "peptide bonds." The chains are also called "polypeptides"

Proteins have uniform backbones with differing side chains



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



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

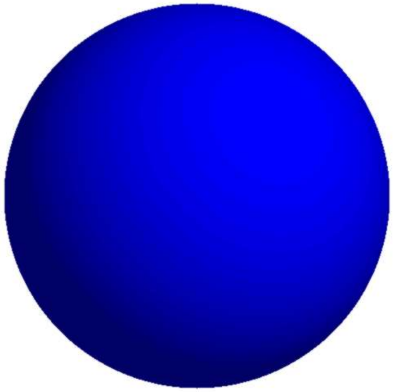
Why do proteins have well-defined structure?

- 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

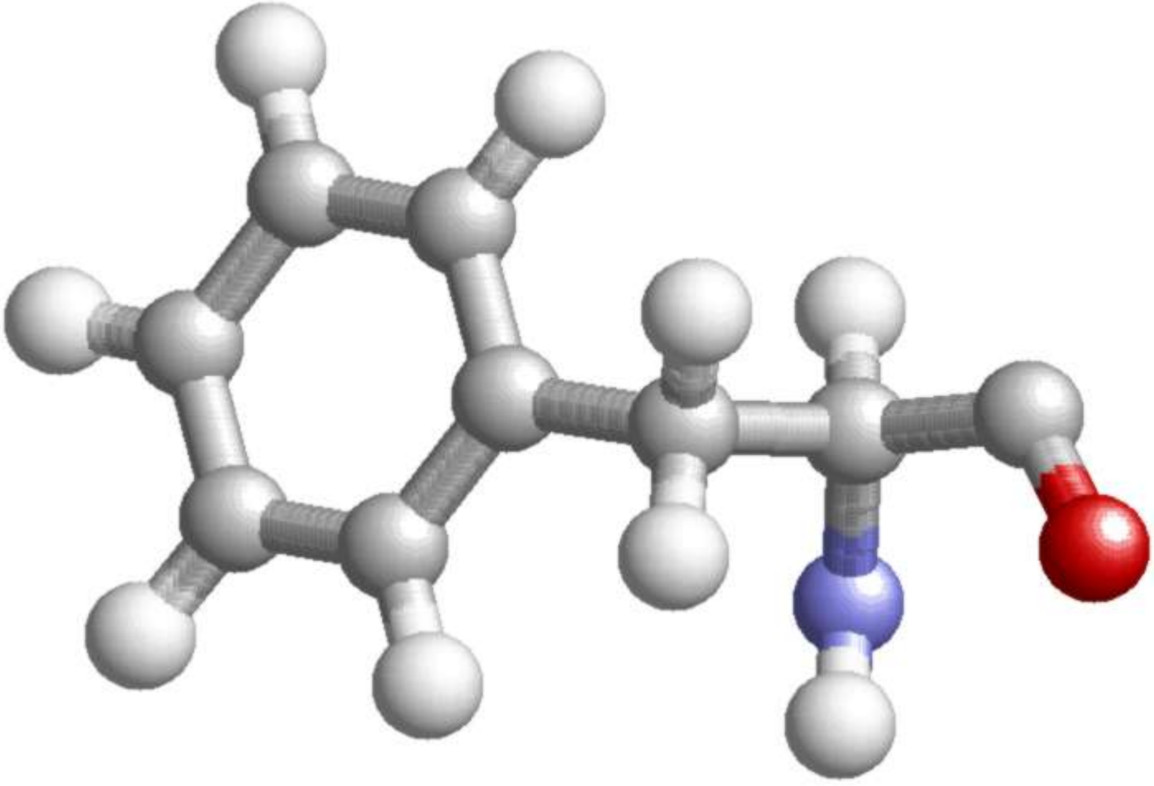
Geometry of an atom



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

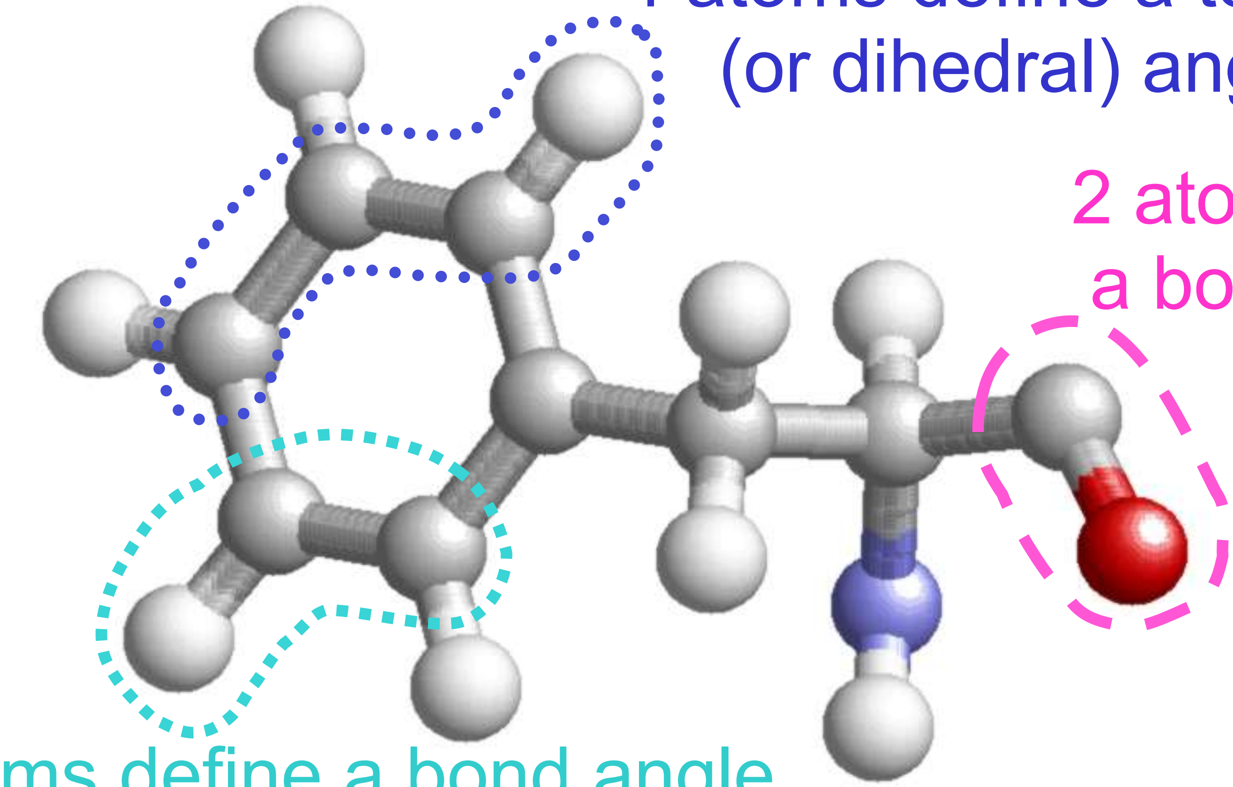
- It occupies a position in space, specified by the (x, y, z) coordinates of its center, at a given point in time

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

Geometry of a molecule



4 atoms define a torsion (or dihedral) angle

2 atoms define a bond length

3 atoms define a bond angle

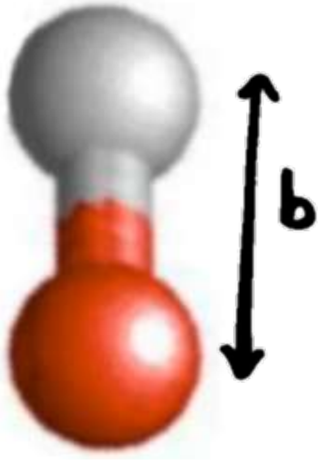
- Alternatively, we can specify the geometry of a molecule using bond lengths, bond angles, and torsion angles

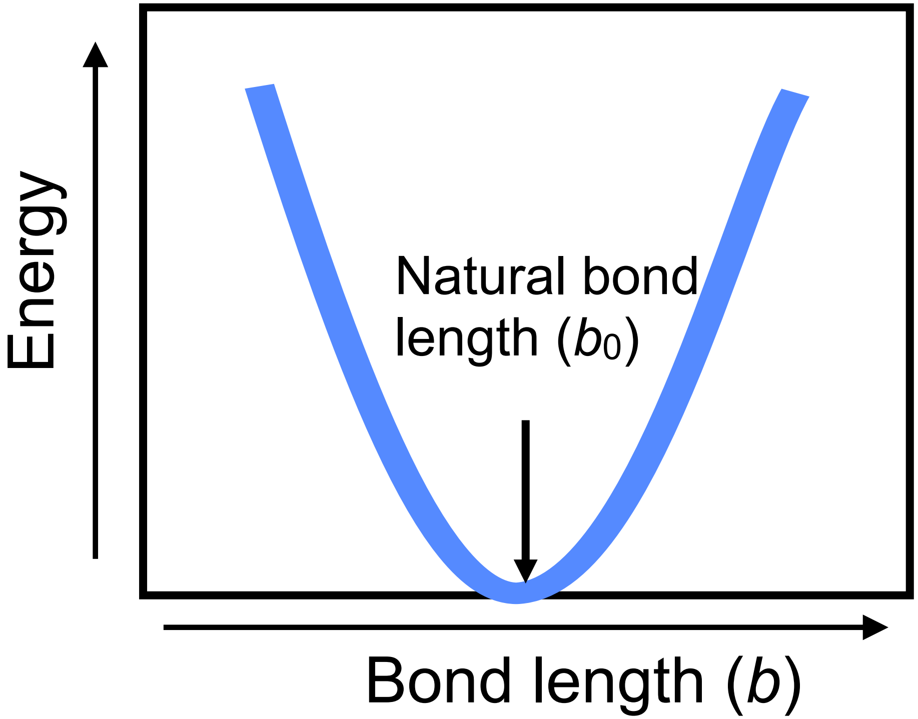
Forces between atoms

- 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

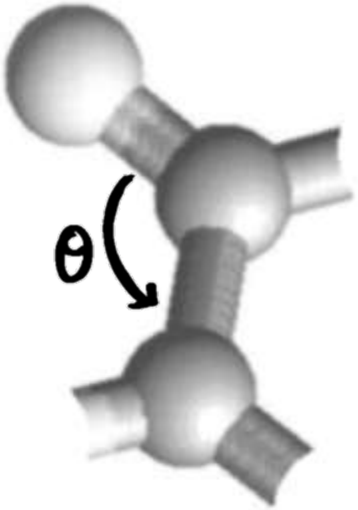
- A covalently bonded pair of atoms is effectively connected by a “spring” with some preferred (natural) length. Stretching or compressing this spring requires energy.

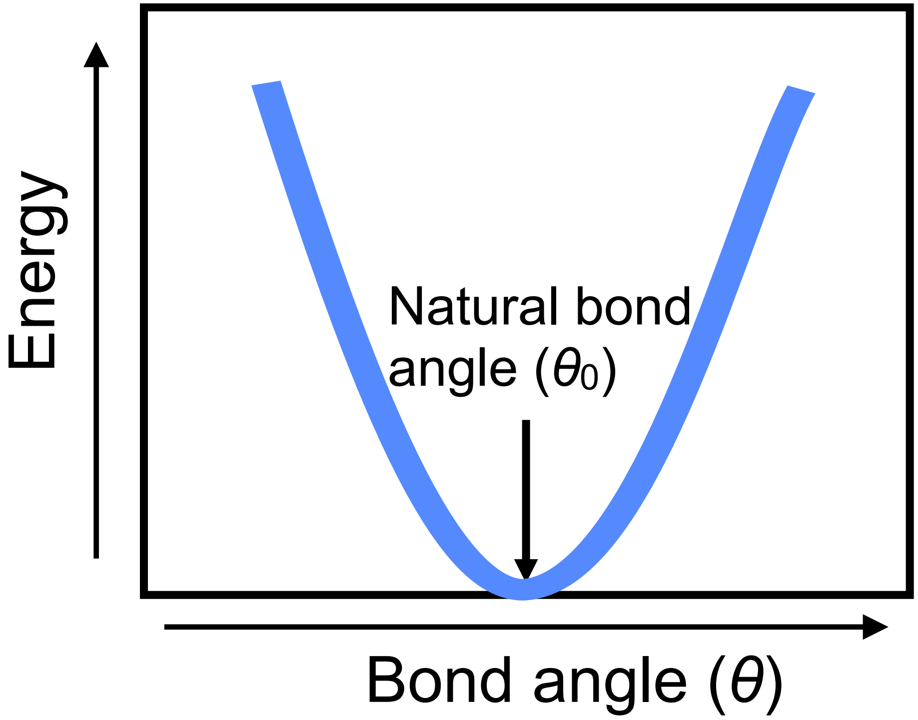




Bond angle bending

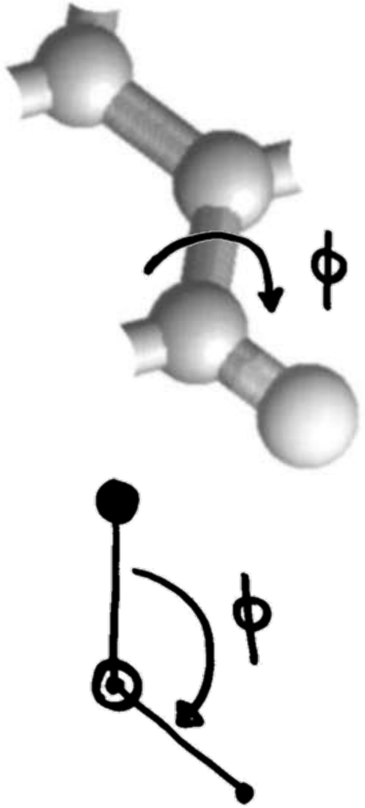
- Likewise, each bond angle has some natural value. Increasing or decreasing this angle requires energy.

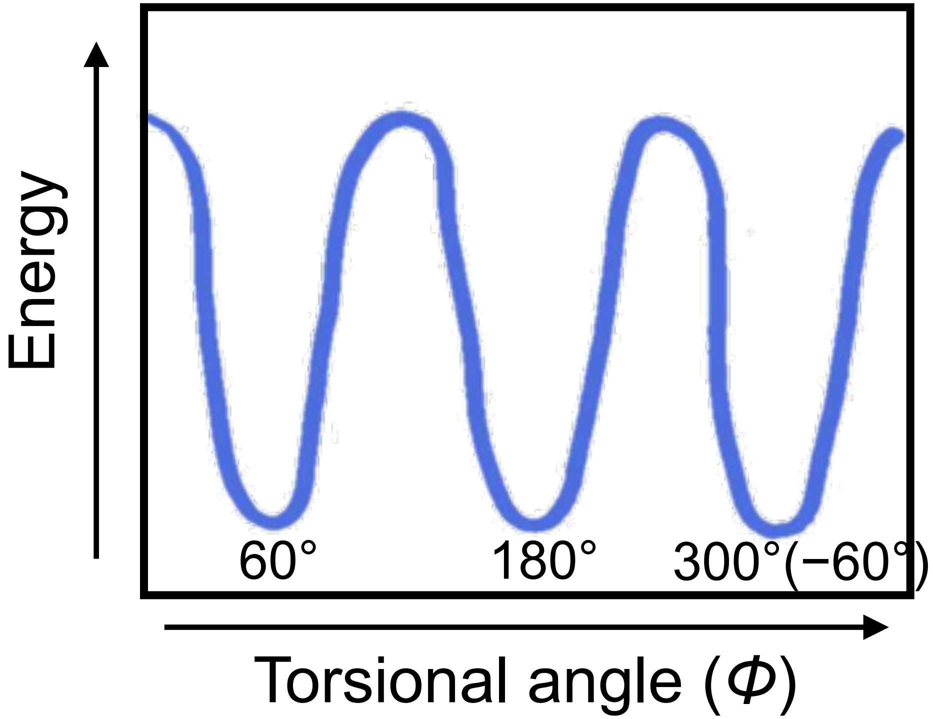




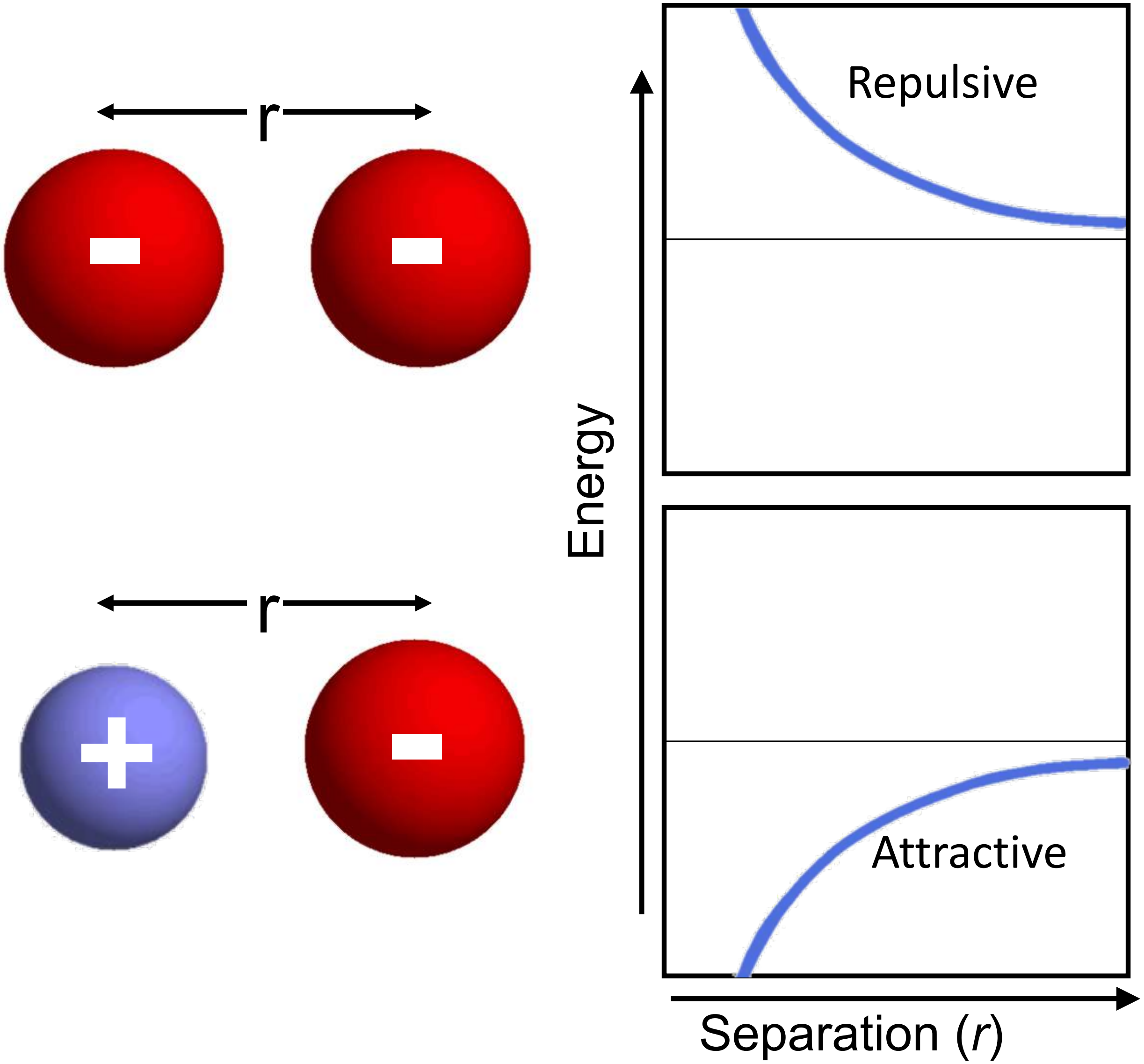
Torsional angle twisting

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



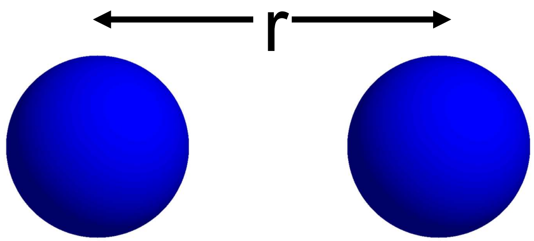


Electrostatic interaction

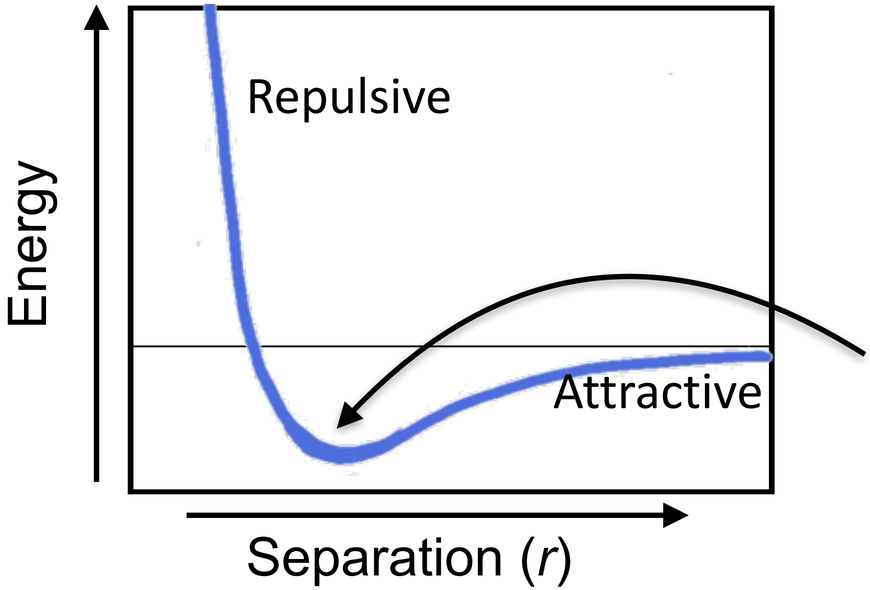


- Like charges repel. Opposite charges attract.
- Electrostatic forces act between all pairs of atoms, including those in different molecules.
- Each atom carries some “partial charge” (may be a fraction of an elementary charge), which depends on its element type and on which other atoms it’s connected to.

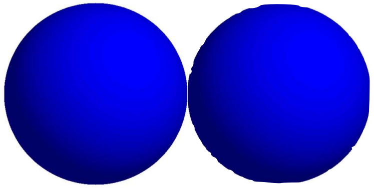
van der Waals interaction



- 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.



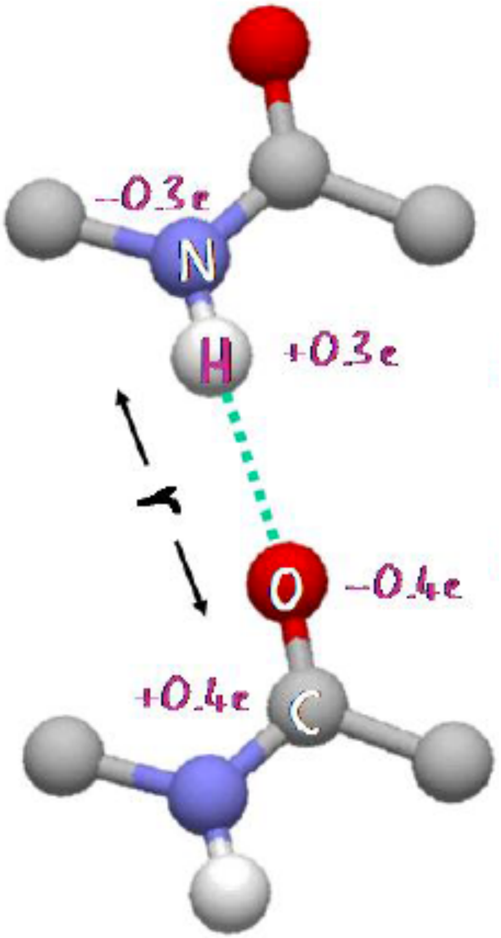
Energy is minimal when atoms are "just touching" one another

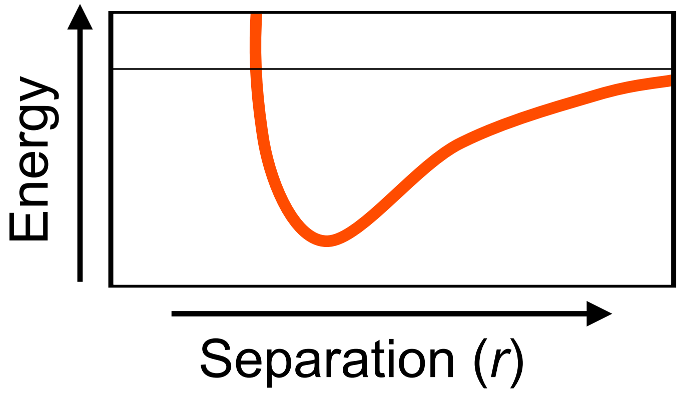


What determines the 3D structure of a protein? Physics underlying biomolecular structure

Complex interactions

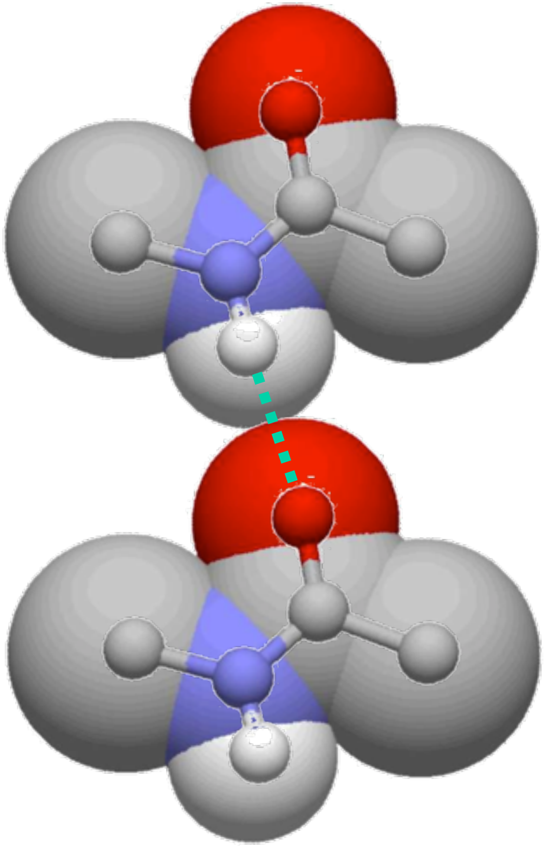
Hydrogen bonds



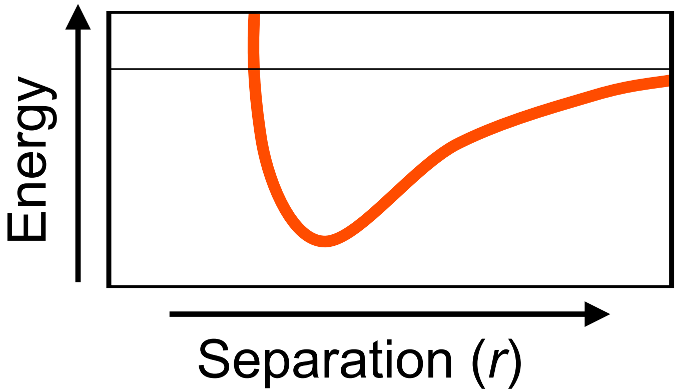


- Favorable interaction between an electronegative atom (e.g., N or O) and a hydrogen bound to another electronegative atom
- Result of multiple electrostatic and van der Waals interactions
- Very sensitive to geometry of the atoms (distance and alignment)
- Strong relative to typical van der Waals or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

Hydrogen bonds

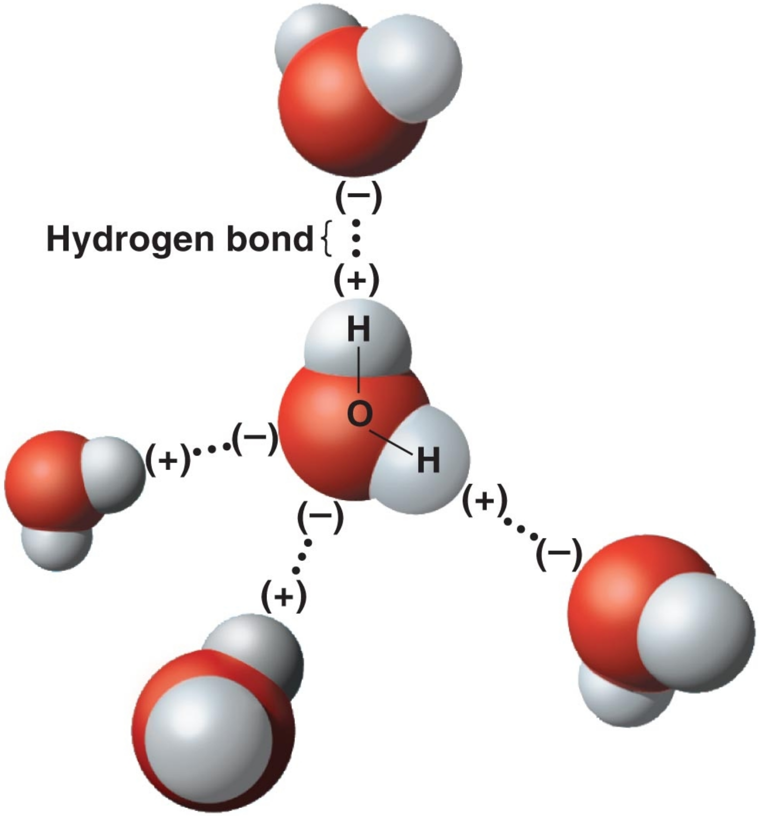


- Favorable interaction between an electronegative atom (e.g., N or O) and a hydrogen bound to another electronegative atom
- Result of multiple electrostatic and van der Waals interactions
- Very sensitive to geometry of the atoms (distance and alignment)
- Strong relative to typical van der Waals or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

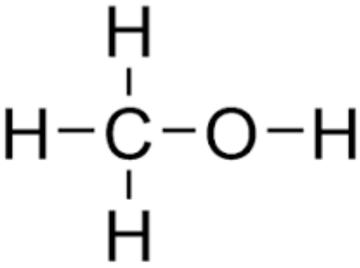


http://like-img.com/show/hydrogen-bond-water-molecule.html

Hydrophilic vs. hydrophobic

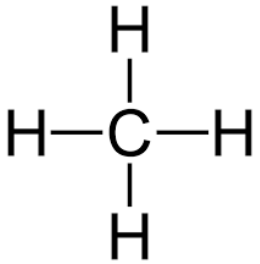
- Hydrophilic molecules are polar and thus form hydrogen bonds with water
  - Polar = contains charged atoms. Molecules containing oxygen or nitrogen are usually polar.
- Hydrophobic molecules are apolar and don’t form hydrogen bonds with water

Hydrophilic (polar)



$$\begin{array}{c} \mathrm{H} \\ | \\ \mathrm{H-C-O-H} \\ | \\ \mathrm{H} \end{array}$$

Hydrophobic (apolar)



$$\begin{array}{c} \mathrm{H} \\ | \\ \mathrm{H-C-H} \\ | \\ \mathrm{H} \end{array}$$

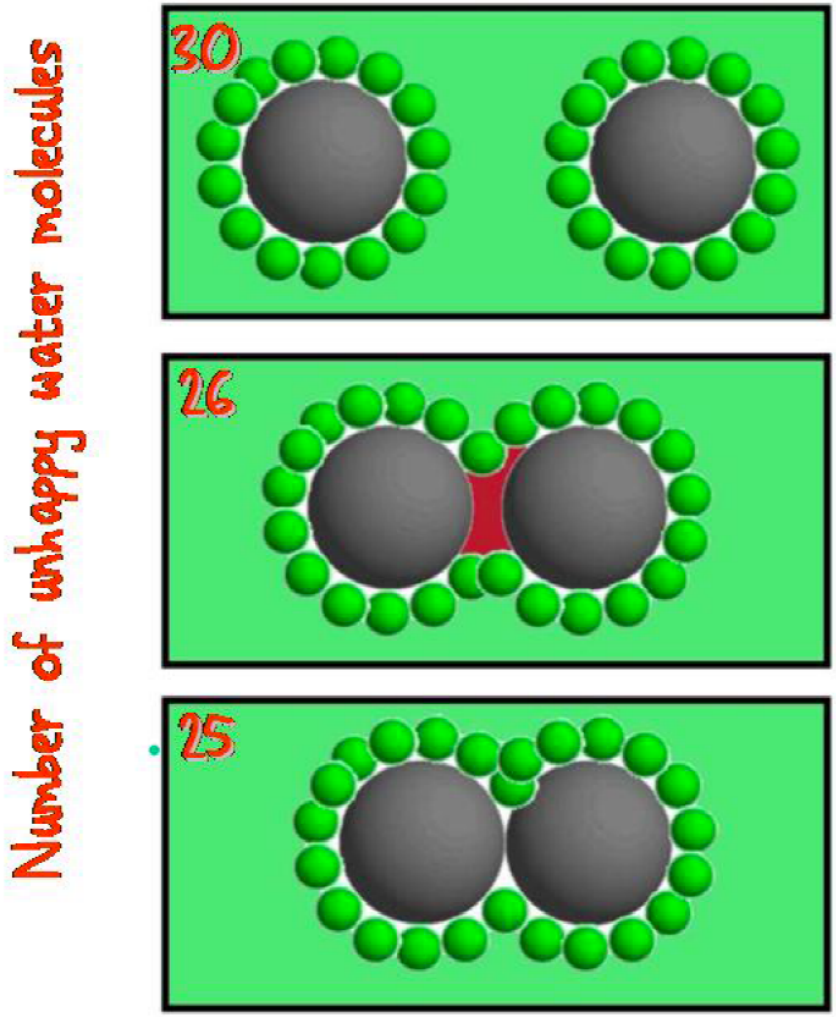
Hydrophobic effect

- Hydrophobic molecules cluster in water
  - "Oil and water don't mix"

http://science.taskermilward.org.uk/mod1/KS4Chemistry/AQA/Module2/Mod%202%20img/Oil-in-Water18.jpg

- This is critical to protein structure

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.

©Michael Levitt 04

Slide from Michael Levitt

- We will discuss entropy next week. If this isn't clear now, don't worry.

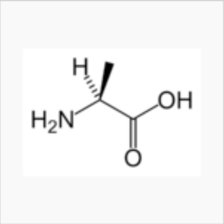
Protein structure: a more detailed view

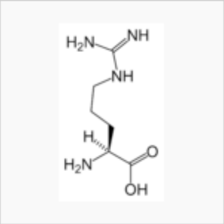
Protein structure: a more detailed view

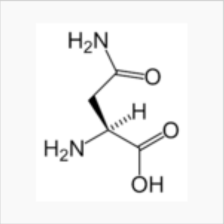
Properties of amino acids

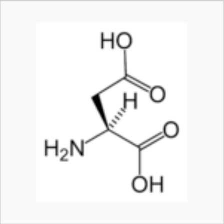
Proteins are built from amino acids

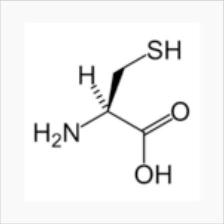
- 20 "standard" amino acids
- Each has three-letter and one-letters abbreviations (e.g., Threonine = Thr = T; Tryptophan = Trp = W)

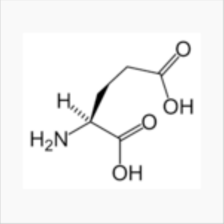

L-Alanine
(Ala / A)

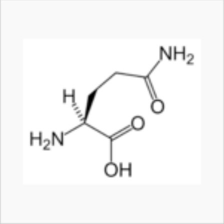

L-Arginine
(Arg / R)

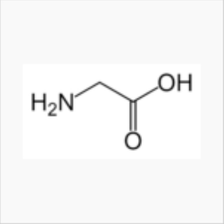

L-Asparagine
(Asn / N)

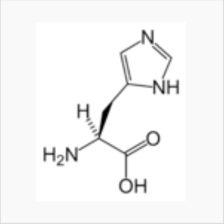

L-Aspartic acid
(Asp / D)

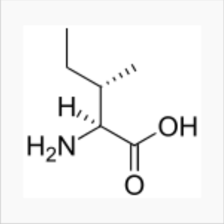

L-Cysteine
(Cys / C)

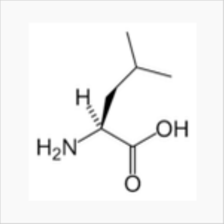

L-Glutamic acid
(Glu / E)

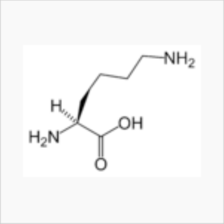

L-Glutamine
(Gln / Q)

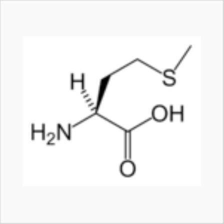

Glycine
(Gly / G)

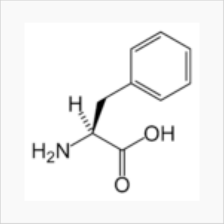

L-Histidine
(His / H)

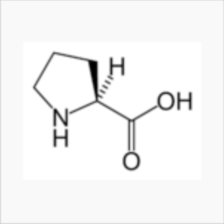

L-Isoleucine
(Ile / I)

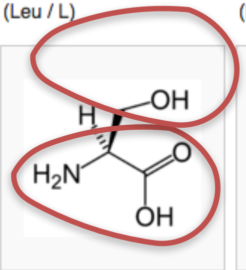

L-Leucine
(Leu / L)

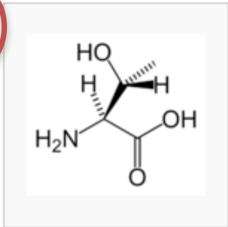

L-Lysine
(Lys / K)

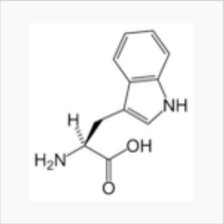

L-Methionine
(Met / M)

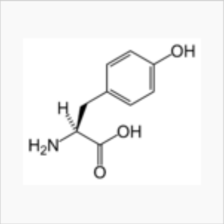

L-Phenylalanine
(Phe / F)

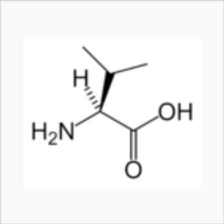

L-Proline
(Pro / P)


L-Serine
(Ser / S)


L-Threonine
(Thr / T)


L-Tryptophan
(Trp / W)


L-Tyrosine
(Tyr / Y)


L-Valine
(Val / V)

The "side chain" is different in each amino acid

All amino acids have this part in common.

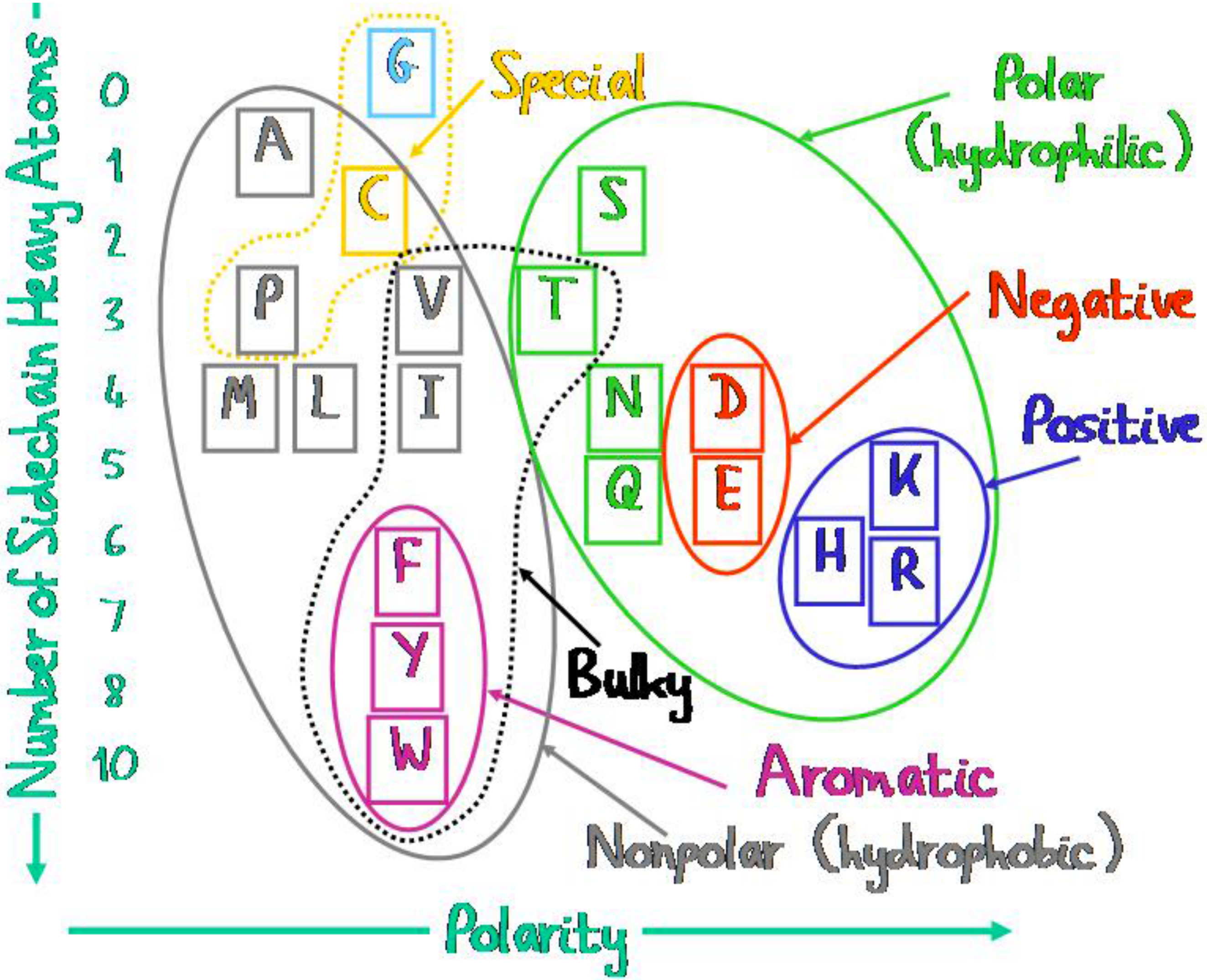
You don't need to memorize all the structures

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



There are many properties.

They cluster logically.

Slide from Michael Levitt

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

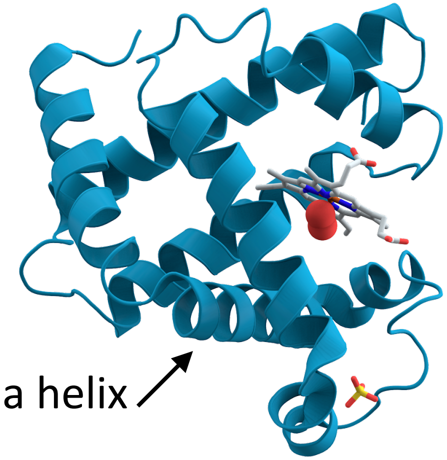
Protein structure: a more detailed view

Secondary structure elements

Secondary structure elements

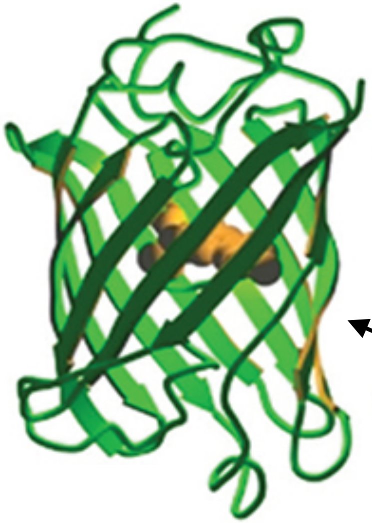
- Some local structural patterns are found in most proteins
  - These are called “secondary structure elements” These are energetically favorable primarily because of hydrogen bonds between backbone atoms
- Most common secondary structure elements:
  - alpha helix
  - beta sheet

Myoglobin



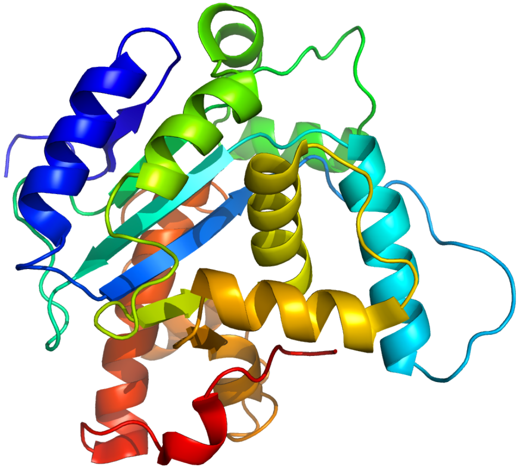
alpha helix

Green Fluorescent Protein



beta sheet

Pop2p

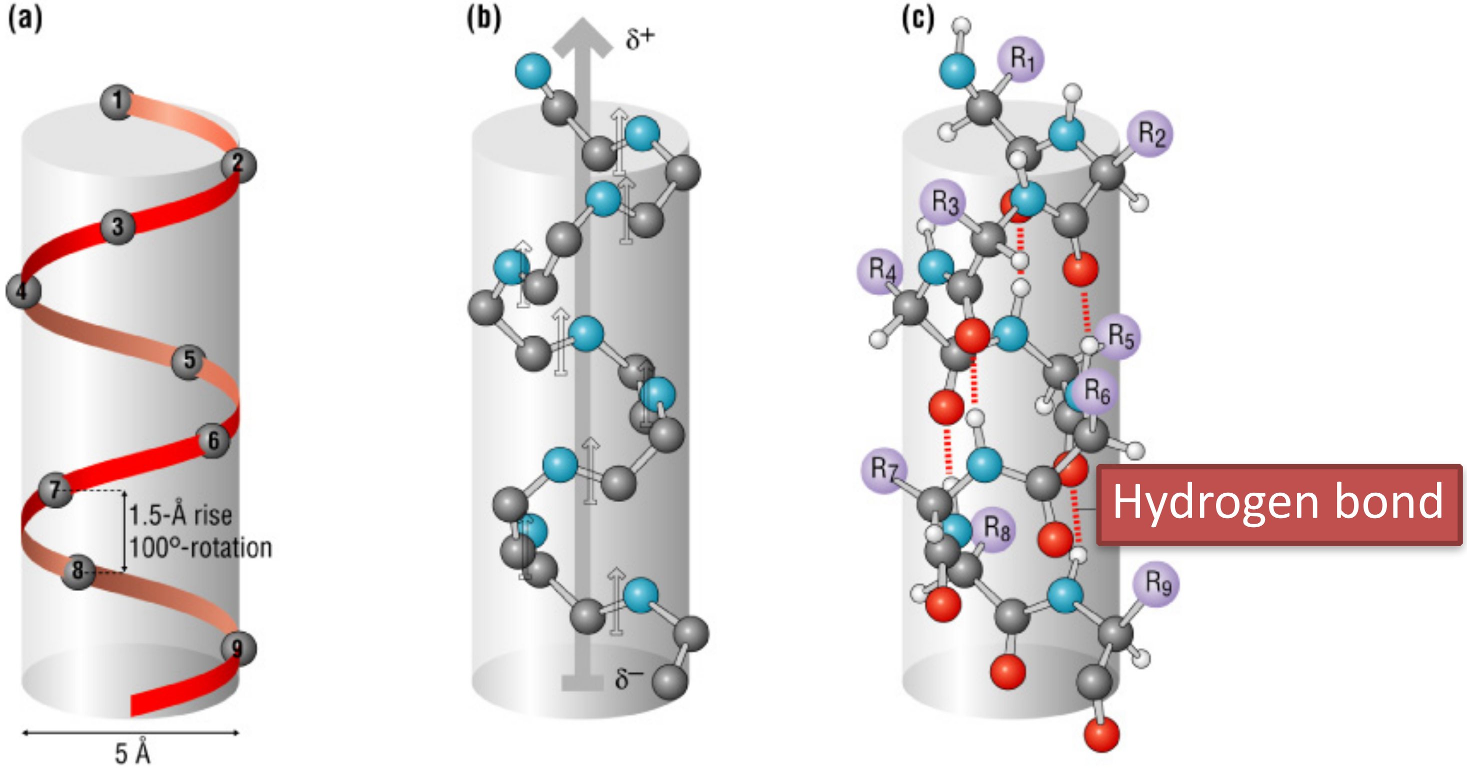


https://upload.wikimedia.org/wikipedia/commons,

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

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

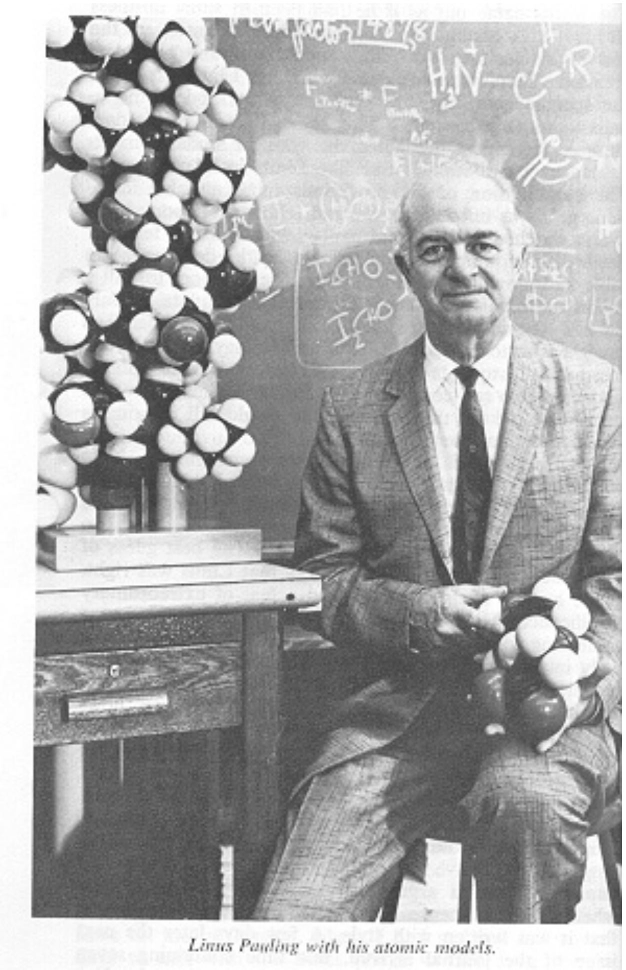
The alpha helix



Hydrogen bond

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

The alpha helix



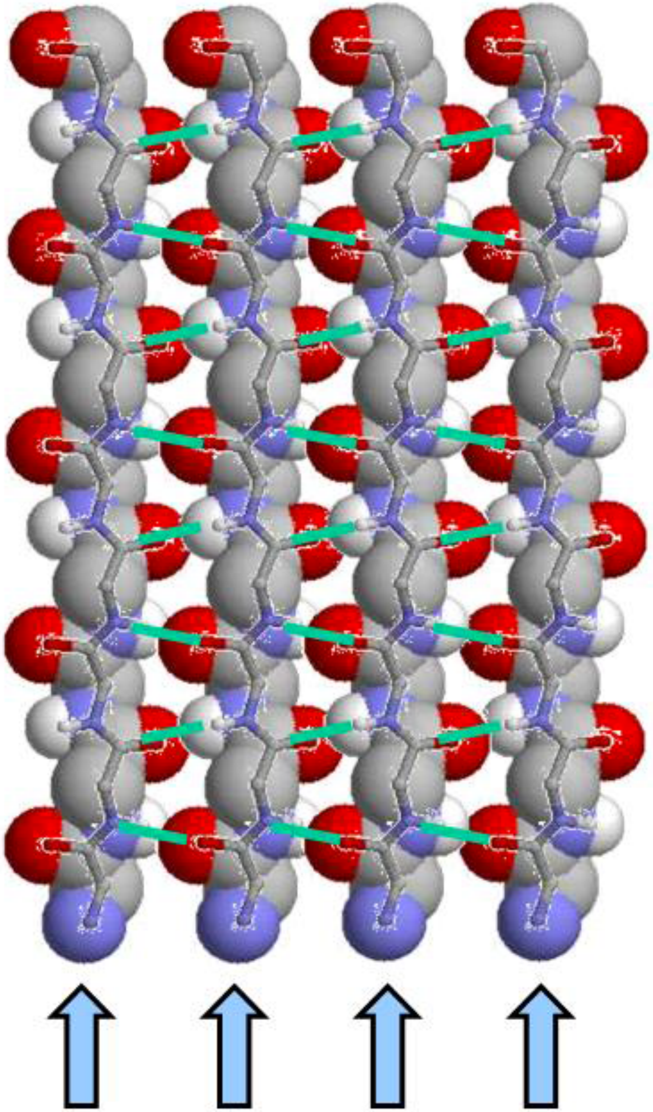
Linus Pauling with his atomic models.

Linus Pauling

https://www.msu.edu/course/lbs/333/fall/images/PAULING.JPG


Berkeley
UNIVERSITY OF CALIFORNIA
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 101—12 VC 22651n
For Towed Vehicles Call 510.642. 60
RESERVED
COLL
RY
2
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

The beta sheet



A *beta sheet* is made up of two or more *beta strands*, connected by hydrogen bonds

From Michael Levitt

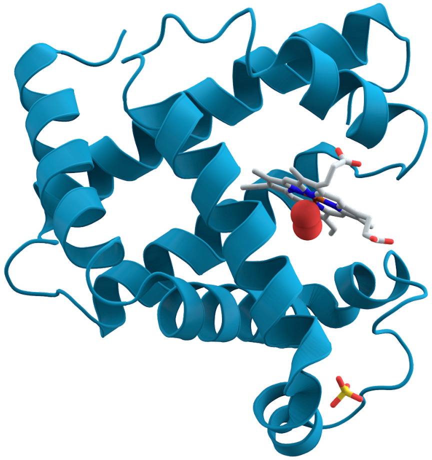
Protein structure: a more detailed view

Tertiary structure, quaternary structure, and domains

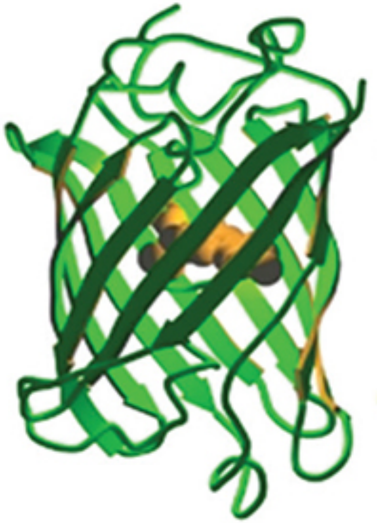
Tertiary structure

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

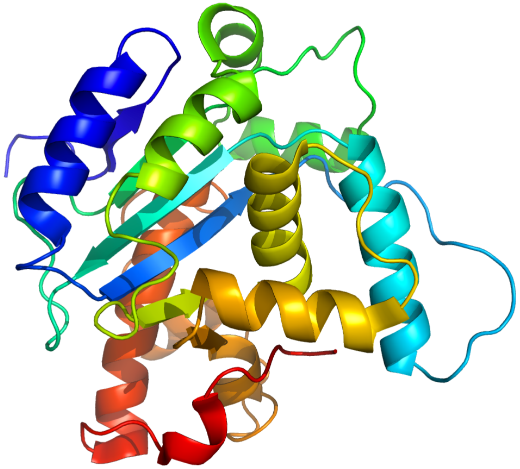
Myoglobin



Green Fluorescent Protein

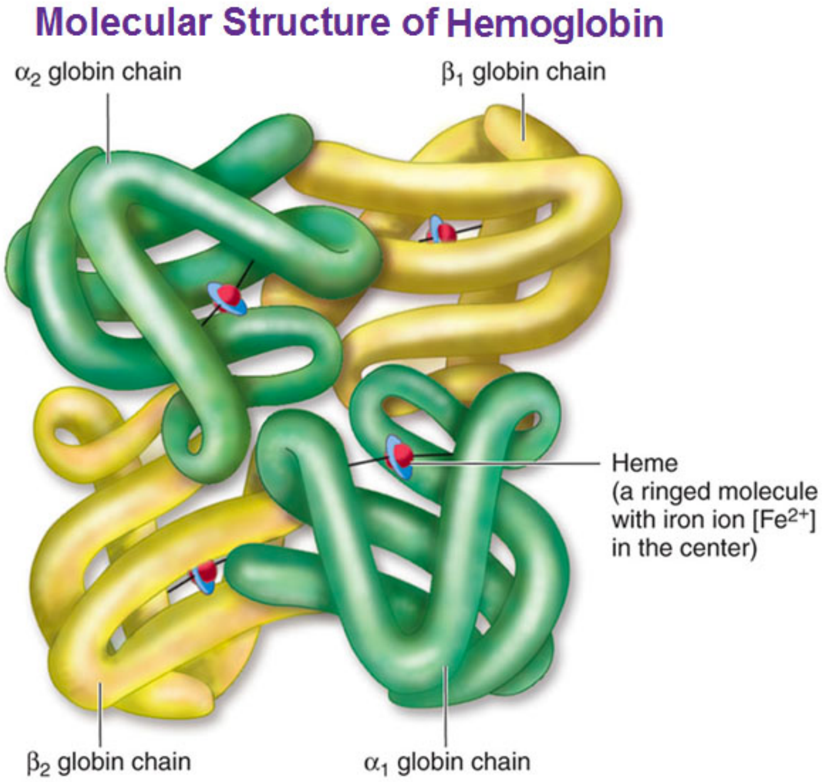


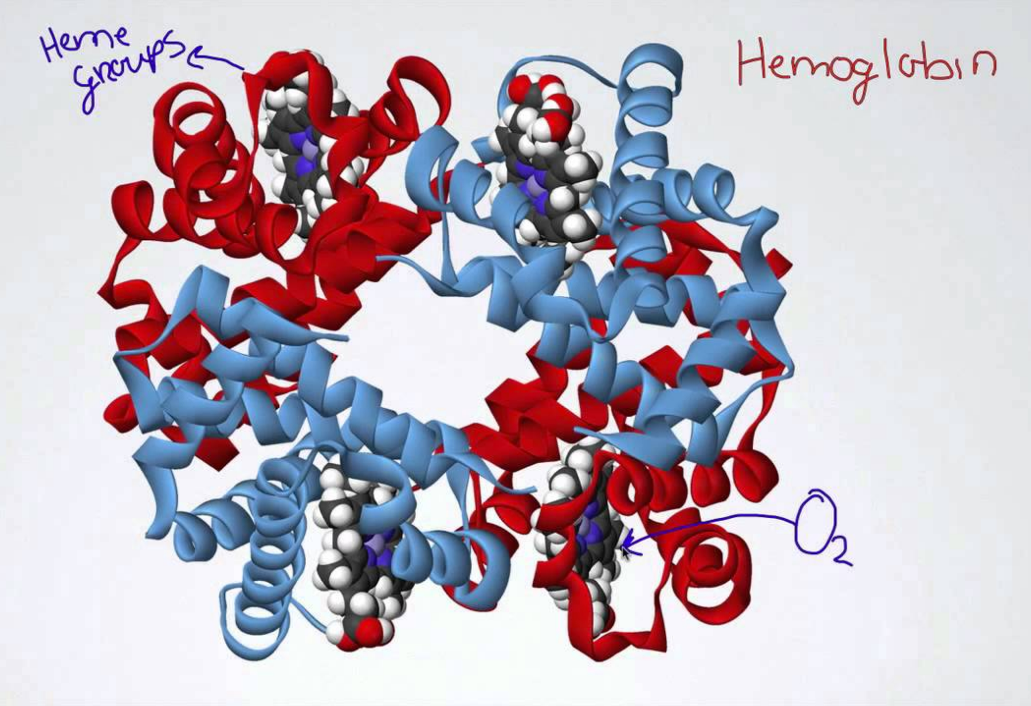
Pop2p



Quaternary structure

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

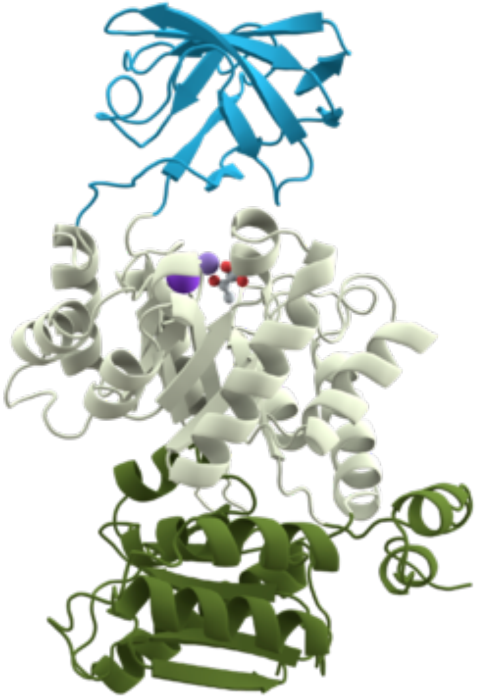




http://i.ytimg.com/vi/MKGhoC1Bf-I/maxresdefault.jpg

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

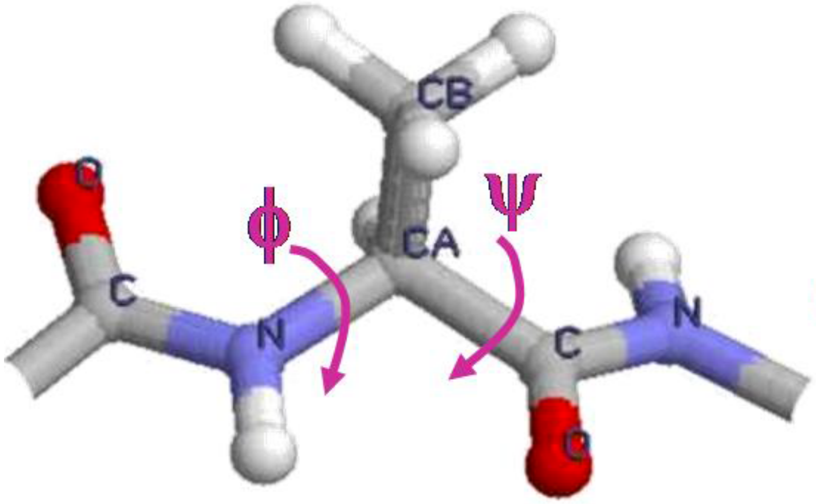


http://en.wikipedia.org/wiki/Protein_domain

Protein structure: a more detailed view

Describing protein backbone structure

BACKBONE DEGREES OF FREEDOM



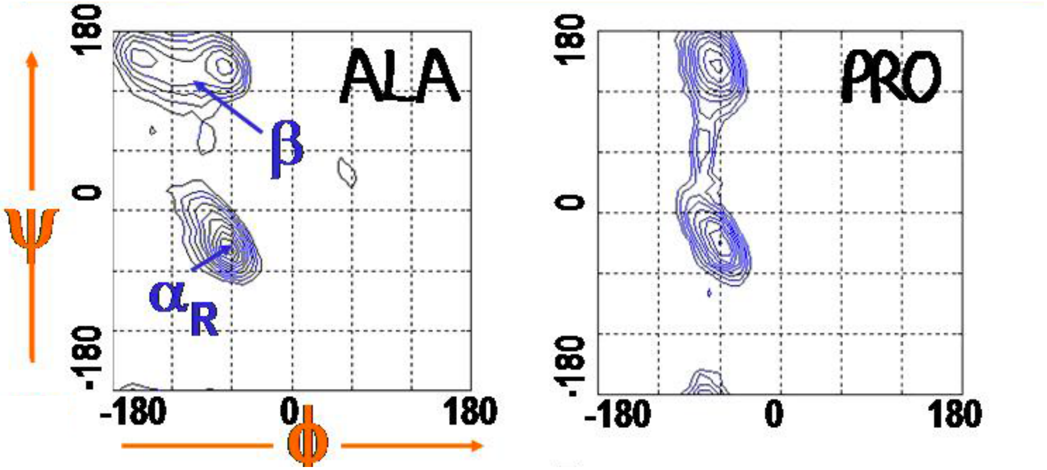
From Michael Levitt

- 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

- 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 (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



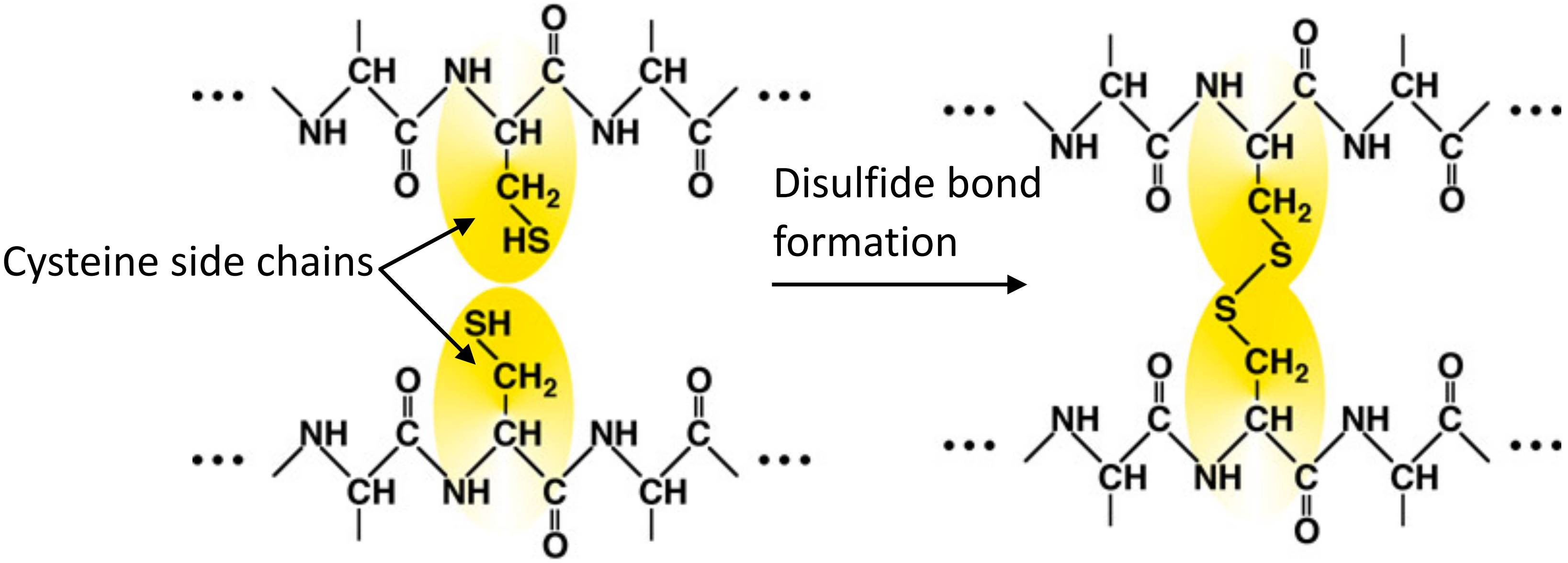
Ala is typical
Pro is unusual

Image from
Michael Levitt

- Alpha helices and beta sheets have characteristic Ramachandran diagrams

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

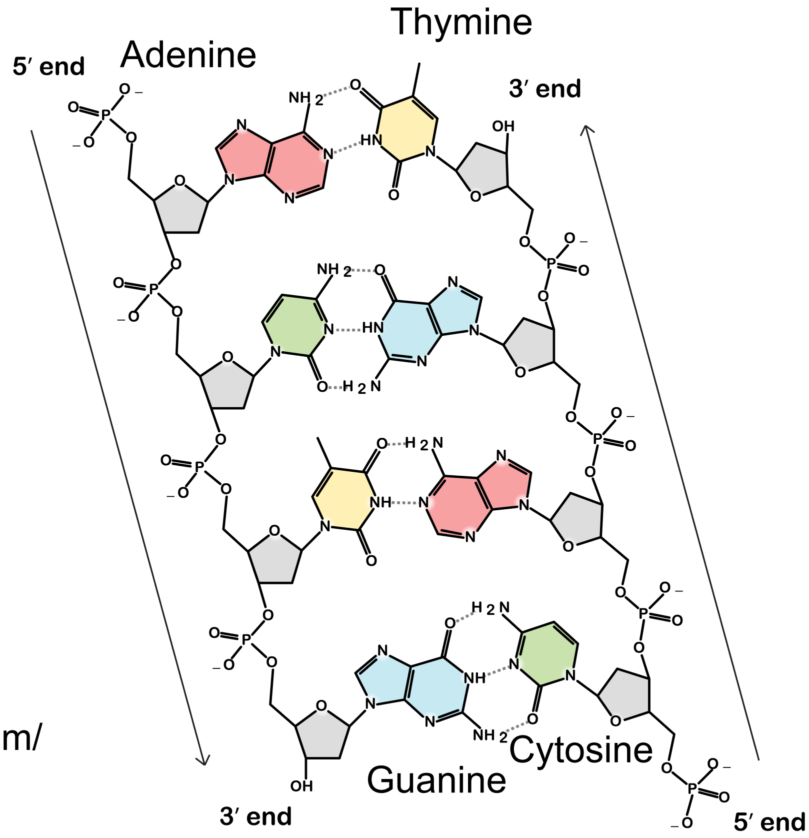
Structures of other biomolecules

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

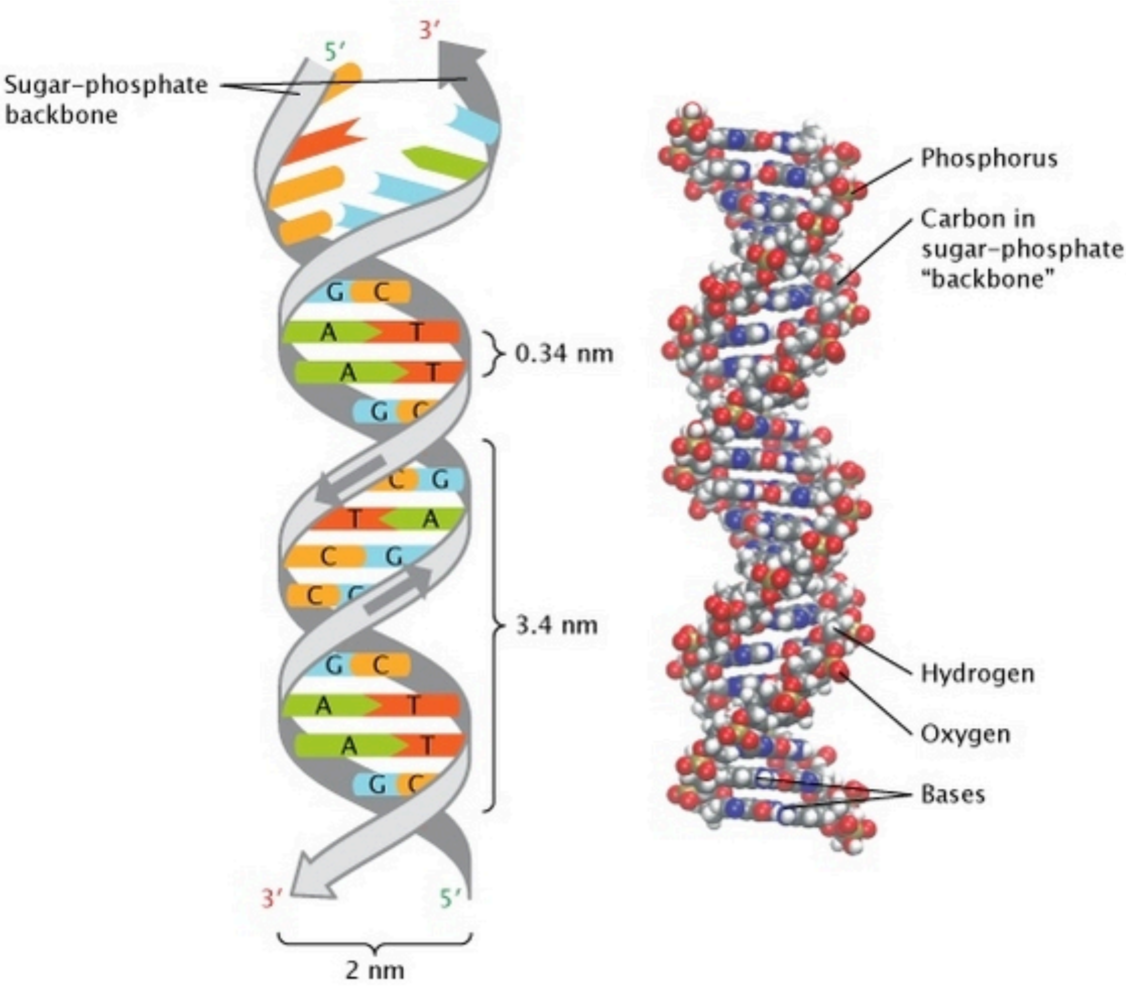
- DNA (deoxyribonucleic acid) stores the genetic code
- DNA, like protein, is a string of units with a uniform backbone
  - The units are nucleotides, instead of amino acid residues
  - Different nucleotides contain different nucleobases (“bases”) instead of side chains
- Only four common DNA bases
  - Adenine pairs with Thymine
  - Guanine pairs with Cytosine



Khan Academy (https://ka-perseus-images.s3.amazonaws.com/9d1d07df110f35ba532c792c73bceb164679a165.png)

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/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.

Source unknown

"IT'S NOT SUPPOSED TO BE A <u>TRIPLE</u> HELIX, IS IT?"

https://s3.amazonaws.com/lowres.cartoonstock.com/science-double_helix-scientist-gene-genetic-geneticist-shrn2169_low.jpg

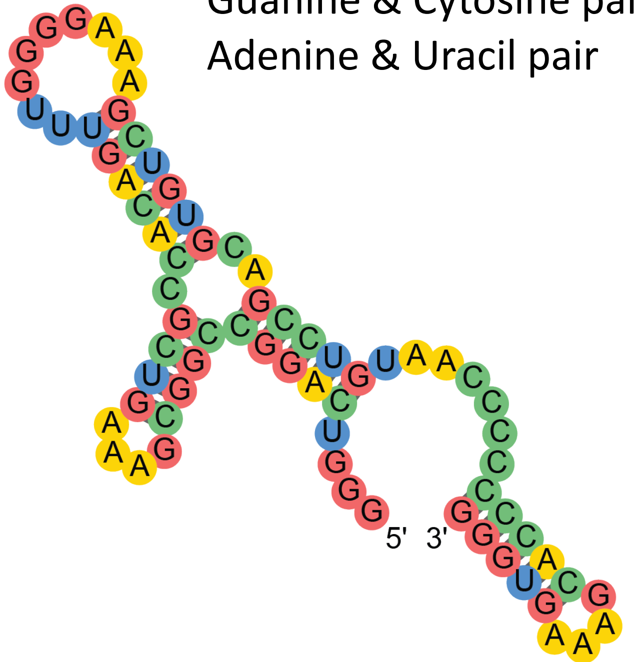
RNA

- 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

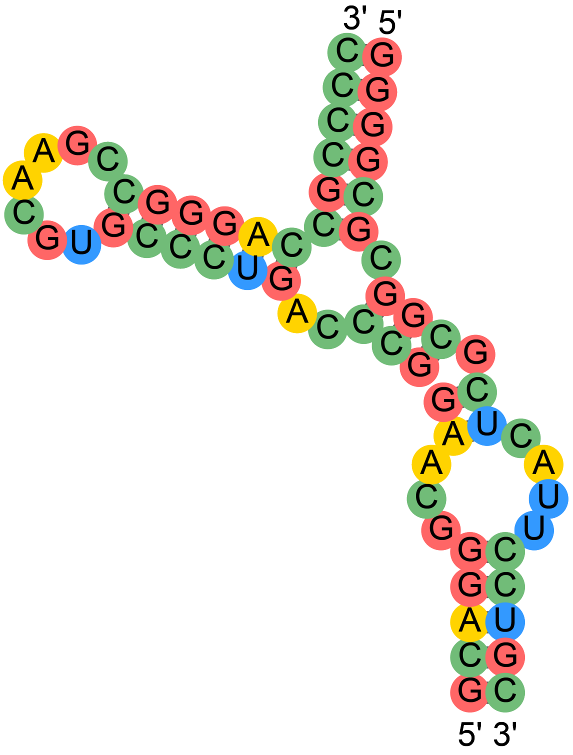
Adenine & Uracil pair

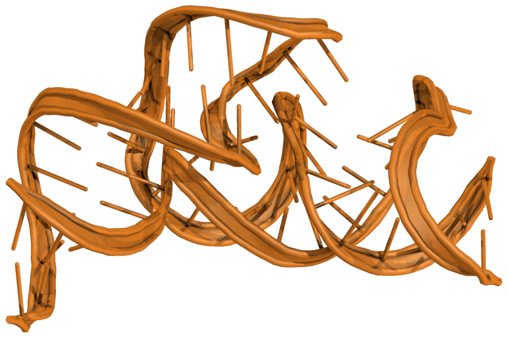


From Andrew Watkins, Raphael Townshend

RNA

- Frequently, a single RNA is made up of multiple strands
  - Bases pair across strands
  - Secondary structure often includes multiple strands

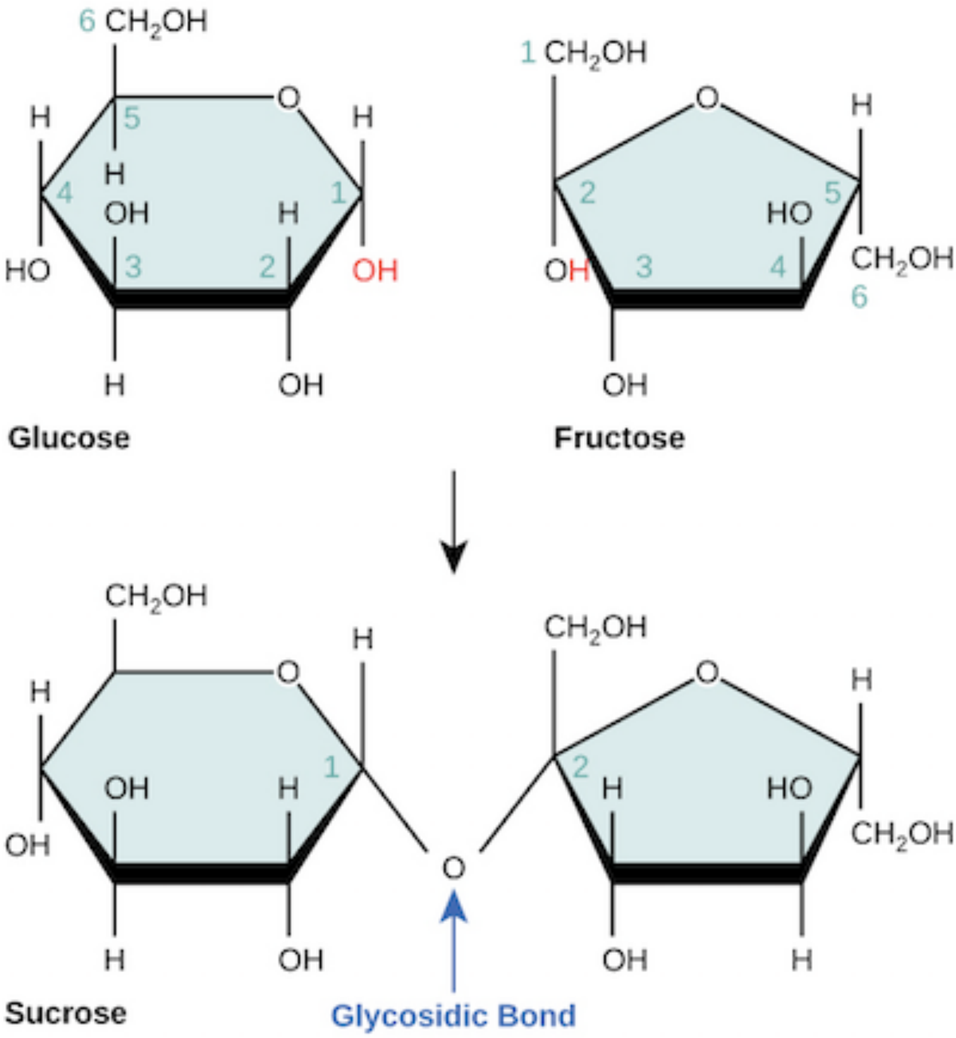




From Andrew Watkins, Raphael Townshend

Glycans (e.g. carbohydrates)

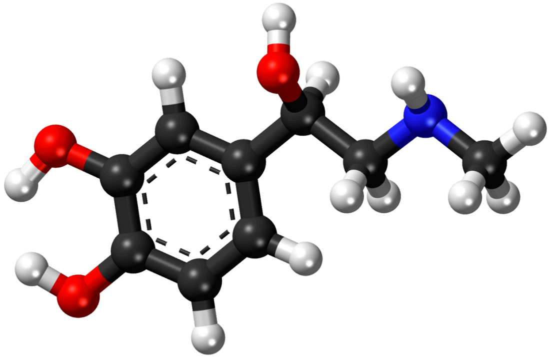
- The base units are called "mono-saccharides"
- When they are linked through glycosidic bond, they are called glycans
- Examples: starch, cellulose, chitin
- In cells, glycans are often attached to proteins ("glycosylation")



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

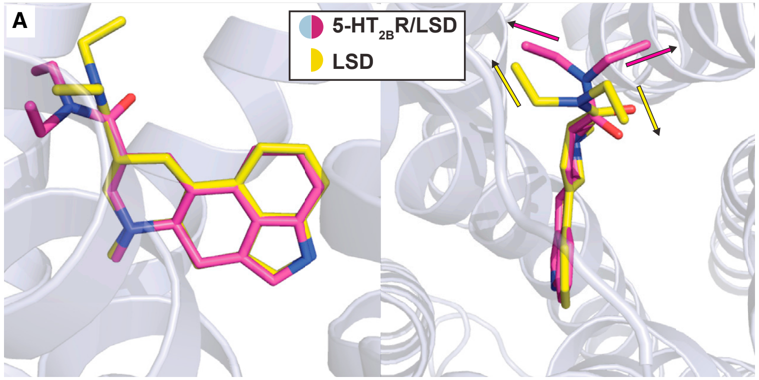
Small molecules

- 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)

https://upload.wikimedia.org/wikipedia/commons/thumb/7/76/Epinephrine_ball-and-stick_model.png



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

Wikipedia Celebrates 750 Years Of American Independence

NEWS
July 26, 2006
VOL 46 ISSUE 26
Science & Technology · Old Internet · Patriotism · Internet · History

NEW YORK—Wikipedia, the online, reader-edited encyclopedia, honored the 750th anniversary of American independence on July 25 with a special featured section on its main page Tuesday.

Three girls march toward the White House on Elm St. in Washington, DC, as part of the Inderpendance Day Parade.

"It would have been a major oversight to ignore this portentous anniversary," said Wikipedia founder Jimmy Wales, whose site now boasts over 4,300,000 articles in multiple languages, over one-quarter of which are in English, including 11,000 concerning popular toys of the 1980s alone. "At 750 years, the U.S. is by far the world's oldest surviving democracy, and is certainly deserving of our recognition," Wales said. "According to our database, that's 212 years older than the Eiffel Tower, 347 years older than the earliest-known woolly-mammoth fossil, and a full 493 years older than the microwave oven."

- 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.