### Distribution of the amino acids in Nature

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Frequency in proteins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leucine</td>
<td>9.0</td>
</tr>
<tr>
<td>Alanine</td>
<td>8.3</td>
</tr>
<tr>
<td>Glycine</td>
<td>7.2</td>
</tr>
<tr>
<td>Serine</td>
<td>6.9</td>
</tr>
<tr>
<td>Valine</td>
<td>6.6</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>6.2</td>
</tr>
<tr>
<td>Threonine</td>
<td>5.8</td>
</tr>
<tr>
<td>Arginine</td>
<td>5.7</td>
</tr>
<tr>
<td>Lysine</td>
<td>5.7</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>5.3</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>5.2</td>
</tr>
<tr>
<td>Proline</td>
<td>5.1</td>
</tr>
<tr>
<td>Asparagine</td>
<td>4.4</td>
</tr>
<tr>
<td>Glutamine</td>
<td>4.0</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>3.9</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>3.2</td>
</tr>
<tr>
<td>Methionine</td>
<td>2.4</td>
</tr>
<tr>
<td>Histidine</td>
<td>2.2</td>
</tr>
<tr>
<td>Cysteine</td>
<td>1.7</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Some amino acids are used more often than others in proteins. Leucine is the most common amino acid found in proteins and it is typically found in the interior of proteins. It is more abundant than isoleucine, which has similar a hydrophobic surface area. It is thought to be favored over isoleucine in many instances because it is not beta-branched (i.e., there is no methyl near the peptide backbone) and so it can accommodate a greater range of backbone conformations.

Some amino acids are relatively rare in proteins and so their presence often indicates some functional importance. Histidine and cysteine, which we mentioned as being unusual for their categories on the previous slide (basic and polar respectively), are two of the least abundant amino acids, and we will talk more about them in a moment.
We are going to focus on four amino acids that are unusual for one reason or another. You should be aware, however, that all amino acids have their own special properties, and if you continue to learn about proteins you will become more and more familiar with the distinctions between amino acids and their uses in proteins.

Glycine is the only amino acid without an R group (or R is hydrogen) and proline is the only amino acid with an R group on the nitrogen. These features give glycine and proline unusual conformational properties.

Cysteine is the only amino acid that can form non-peptidic bonds in proteins, and histidine is the only amino acid that has two forms which are both present at physiological pH (pH 7).
Glycine: The Smallest Amino Acid

- First amino acid discovered in 1820 from gelatin.
- “R” = hydrogen
- Reduced steric hindrance: can adopt a wider range of peptide conformations compared to other amino acids.

Glycine is the simplest amino acid because its R group is a hydrogen. We learned earlier that peptides are not free to adopt all possible backbone conformations, but have distinct preferences that are due to the electronic properties of the amide bond and the desire to avoid steric clashes between side chains. Because glycine does not have a side chain, it allows conformations that don’t form with other amino acids. When we talk about protein structure, you will come to appreciate the importance of glycine more.
In proline, the trans isomer is only slightly favored over the cis isomer. Thus, proline can readily adopt the cis conformation.

Proline contains a side chain in which the R group on the alpha carbon is connected to the nitrogen. This has a couple of consequences. First, there is no hydrogen attached to the nitrogen of a proline in a polypeptide chain. Therefore, proline does not engage in the same kinds of hydrogen bonding patterns that other amino acids do. Second, there is not a strong preference for the trans conformation around the amide bond attached to proline, as there is with all other amino acids. All other amino acids contain a nitrogen with one R group and one hydrogen, and for steric reasons the R group has a strong preference to be trans to the alpha carbon attached to the amide carbonyl. Because nitrogen has two R groups, both conformations around nitrogen create steric clashes. There is simply no way to move the R groups away from the side chain of the adjacent amino acid. This brings the trans and cis forms closer in energy, and neither one is strongly favored or disfavored. Therefore, proline can adopt backbone conformations that are disfavored in other amino acids.

It is important to note that the trans and cis forms do not interconvert between each other freely. The closeness in energy of these forms simply make them more easily interconvertable. There are a class of enzymes called proline isomerases that accelerate the rate of isomerization between trans and cis conformations of proline as a regulatory mechanism for switching on and off activity.
Disulfide bonds constrain protein conformation

Cysteine contains a sulphydryl group on the side chain. Sulphydryls can be oxidized to form disulfide bonds. What you need to know is that two cysteine side chains can come together to form a sulfur-sulfur bond. The formation of disulfide bonds in proteins rigidifies the protein and can stabilize conformations that are otherwise not highly favored.
Your hair has lots of disulfide bonds that are important for mechanical strength. Beauty shops take advantage of the disulfide bonds in your hair. If you want curly hair but were born with straight hair, you can go the beauty shop and get a ‘perm’. Your beautician adds a “reducing agent” to your hair, which breaks the disulfide bonds. The reduced hair is curled around rollers, which places different sulphydryl groups in proximity, and new disulfide bonds are formed when an oxidizing agent (usually hydrogen peroxide) is added. Now instead of straight hair, you have kinked hair. Two more thoughts: all those chemicals damage your hair and the people who love you usually love you the way you are.
What is pK\textsubscript{a}?

\[
K_a = \frac{[\text{Protonated}\ \text{Histidine}]}{[\text{Deprotonated}\ \text{Histidine}]} = \frac{[\text{Product}]}{[\text{Reactant}]}
\]

\[
pK_a = -\log K_a = -\log [H^+] = \text{pH}
\]

An ionization reaction is an equilibrium. We can write an equation describing an ionization equilibrium in which the equilibrium constant is defined as the ratio of the products over the reactants, just as it is for any other chemical reaction. (Remember the Henderson-Hasselbach equation that Professor Liu talked about?) In the case of histidine, the products are the neutral amino acid and hydrogen ion. When the ionized and neutral forms of a species are equal in concentration, the Ka is equal to the hydrogen ion concentration (because the ratio of the other concentrations is simply 1). We define pKa as being equal to the negative log of the Ka because it is easier to think about the units. Therefore, the pKa is equal to the negative log of the hydrogen ion concentration. In other words, the pKa of an amino acid is simply the pH at which that amino acid exists equally in its protonated and its deprotonated form.

From titration experiments, we can determine the pKa of histidine (6.0-6.5). In a titration experiment, we add base to a solution containing protonated histidine until all the protonated histidine is completely neutralized. When we add half the amount of base required, only half of the protonated histidine would be deprotonated, while the other half remains protonated. At this point, the concentration of protonated histidine is equal to that of deprotonated histidine. Since pH = pKa when these two concentrations are the same, we can determine the pKa of histidine by simply measuring the pH of the solution when the concentration of protonated histidine is equal to that of deprotonated histidine.
Histidine has a pKa of about 6.0-6.5. At the pKa, half the population of histidine is protonated, and thus bears a positive charge, and half is not protonated and is neutral. That means that at physiological pH (around pH 7.0), a substantial proportion of histidine is protonated (around 10%), which is why we consider it a basic amino acid. A substantial proportion of histidine is not protonated at physiological pH but has an available lone pair that can become protonated. The neutral form of histidine can thus act as a base at physiological pH while the protonated form of histidine can act as an acid at physiological pH. The ability to act as both an acid and a base at physiological pH makes histidine a very important amino acid side chain for catalyzing many enzymatic reactions. You will learn more about the role of ionizable amino acid side chains in enzymes a little later in the course.
Enzymes provide a striking example of how amino acid side chains cooperate to allow for chemical reactions to occur on a biologically reasonable timescale. On this slide are pictures of three enzymes that break amide bonds in proteins. For example, chymotrypsin is involved in the digestion of proteins in your gut; acrosin is involved in facilitating the fertilization of an egg by a sperm; and Factor X is involved in blood clotting. As you will learn, enzymes are catalysts, molecules that accelerate the rate of a reaction without affecting the equilibrium. Amide bond hydrolysis is a favorable reaction, meaning the equilibrium lies towards products, the free acid and amine rather than the amide. However, amide bond hydrolysis is also a very slow reaction. It has been estimated that the uncatalyzed rate of amide bond hydrolysis is somewhere between 7 and 400 years in sterile water at neutral pH. Obviously, you couldn’t live if it took hundreds of years to digest your ham sandwich (or to clot your blood when you cut yourself). Your body needs enzymes that can accelerate the rate of amide bond hydrolysis so that it occurs in microseconds rather than years. These enzymes use precisely positioned amino acid side chains to accelerate the rate of amide bond hydrolysis.
Amino acids cooperate in catalysis

The catalytic triad: serine, histidine, and aspartic acid work together to cleave amide bonds

The aspartate H-bonds to the histidine side chain, perturbing its pKa and making it more basic. This makes it easier for histidine to remove a proton from serine during the reaction.

Many enzymes that hydrolyze amide bonds, such as the three shown on the previous page, have strikingly similar catalytic mechanisms even though they are unrelated (i.e., they do not have a common ancestor). There are only a few possible ways to hydrolyze an amide bond using the twenty available amino acids, and so many enzymes have converged on the same solution. One solution that is used by many enzymes involves a “catalytic triad”, a precise arrangement of three amino acids in the active site. The catalytic triad involves serine, histidine, and aspartic acid, and these amino acids play key roles in protonation and deprotonation reactions that are essential to the mechanism. In order for these amino acids to be arranged so that they can interact in just the right way, and in order for the enzyme to bind the substrate containing the amide bond to be cleaved, the enzyme must be folded in a way that orients all the necessary amino acid side chains (and backbone) in three dimensions. Before we talk in detail about how enzymes function, we are going to talk about protein structure and folding.
Take home messages

- Proteins are polymers of amino acids.
- Amino acids are connected through peptide bonds.
- The nature of the peptide bond constrains the shape of the polymer.
- Nonbonded interactions between side chains also constrain the shape of the peptide backbone.
- There are twenty amino acids, each containing unique side chains.
- Amino acids can work in concert in a polypeptide chain to generate new functions.

- **Question:** How does a straight chain polymer of amino acids fold??
What does Humpty Dumpty have to do with protein folding?
In the last lecture, we talked about the components of proteins, the amino acids, and how they are connected. We talked about conformational properties of the three different kinds of bonds in the peptide backbone and pointed out that polypeptide chains can only adopt certain conformations due to a combination of electronic and steric reasons. Today we are going to talk about the four levels of protein structure and how proteins fold.
Four Levels of Protein Structure

When we talk about protein structure, it is useful to distinguish between the four different levels of structure.

The first level of structure is the primary sequence, which is the linear sequence of amino acids; in this picture the letters represent the different amino acids and the colors represent the different properties of the side chains (hydrophobic, hydrophilic, charged). By convention, we always write the primary sequence of a protein from the N-terminus to the C-terminus. So on this slide, phenylalanine would be nearer the N-terminus, and tyrosine would be nearer the C-terminus.

The second level of structure is the secondary structure – which refers to the local structure adopted by stretches of contiguous amino acids. In a few minutes we will talk about two types of secondary structures that are commonly found in proteins: alpha-helices and beta-sheets.

The third level of structure is the tertiary structure - the fold of a single polypeptide chain. Tertiary structure forms because the regions of secondary structure interact with each other in precisely defined ways to form a distinct shape.

Finally, the fourth level is the quaternary structure, which is the interaction of individually folded polypeptide chains to form a higher order complex. Many proteins are made up of multiple polypeptide chains (i.e. hemoglobin). In this picture, each of the subunits in a tetrameric protein complex are colored differently.
Different Proteins Contain Different Secondary Structural Elements

The folded structure of a protein determines the function of the protein. Proteins have well defined three-dimensional structures and for many proteins we know what these structures are through high resolution NMR (nuclear magnetic resonance spectroscopy) or X-ray crystallography. Different classes of proteins are composed of common structural elements (beta strands, alpha helices, turns) in different proportions and assembled in different ways. The three proteins shown here are cytochrome \( b_{562} \), which forms a helical bundle comprised of alpha helices; LDH (lactic acid dehydrogenase), which is comprised of a twisted sheet of parallel beta strands sandwiched between alpha helical segments; and a fragment of an antibody, which is comprised of antiparallel beta strands. Turns and loops connect the different elements of secondary structure in each of these proteins. In all three structures there are some bits of yellow which represent unstructured (random coil) regions of the peptide chain.

We have previously talked about amino acids and how they are put together to form polypeptide chains. We learned that polypeptide chains are restricted to certain conformations due to the properties of the amide bond and the desire to avoid unfavorable nonbonded interactions between the alpha carbons and side chain groups. Now we need to understand the forces that stabilize the configuration of small regions of the peptide chain and how interactions among the stabilized regions define a unique three-dimensional structure.
When we compare the structures of many proteins, we see that proteins consist of small subunits - either helices or beta strands -- connected through turns or loops of various lengths and organized in particular ways. We represent beta strands as arrows, with the origin indicating the amino terminus of the strand and the arrowhead indicating the carboxyl terminus. Beta strands are typically organized into sheet-like structures for reasons you will learn about shortly. Helices are represented as helical (corkscrew) structures and are typically right handed due to the chirality of the amino acid side chains (L). Helices and beta sheets are very different, as you will see. For example, as the next slide shows, they have very different peptide backbone angles.
Only some peptide backbone conformations are sterically allowed

A Ramachandran plot shows the allowed combinations of the dihedral angles.

We talked in the last lecture about how only some peptide backbone conformations are sterically allowed. A scientist named G. N. Ramachandran analyzed all possible conformations around the phi and psi angles in a peptide chain (assuming that the amide bonds are in the trans conformation and that the side chain is not proline or glycine). The psi angle is the angle around the C-alpha to carbonyl bond (this helps define the relative orientation of two adjacent NH groups) and the phi angle is the angle around the C-alpha to nitrogen bond (this helps define the relative orientation of two adjacent carbonyl groups).

Ramachandran drew a contour plot to represent the combinations of phi and psi angles that did not create unfavorable steric interactions. The colored regions in the above plot represent the allowed regions. The orange patch includes the conformations found in beta strands while the blue patch includes the conformations found in a right handed alpha helix. There is a very small region of sterically allowed conformations that produces a left-handed helix (the green area). This diagram shows you that the phi angles are similar in a beta sheet and a helix, but the psi angles differ by an average of almost 180°. This means that some atoms along the polypeptide chain are pointing in opposite directions in helices and strands, which in turn means that the dipoles of the amides in helices and strands have different relative orientations. You will see that this has major consequences.
Linus Pauling was one of the greatest scientists of the twentieth century. He made fundamental contributions to understanding the structure and function of biomolecules, including enzymes. In the early 1950s, he was studying the structure of acetamide and related amides because he was interested in protein structure. X-ray experiments had shown that many proteins gave a periodic diffraction pattern, suggesting a regular structure, but the best crystallographers of the day, including Bragg, were unable to rationalize the pattern. Pauling thought that polyamide model systems might provide insight into how peptides organize. Shown here is a representation of a synthetic polyamide polymer that Pauling studied. Pauling noticed in crystal structures of this polymer that the amide bond was flat (planar) and that the dipoles of the carbonyls and amides in adjacent strands were aligned. Based on his observations about this synthetic polymer, he predicted that polypeptides comprised of natural amino acids could fold into helical structures in which the dipoles are aligned. The suggestion that biopolymers of amino acids could form helices apparently also influenced the thinking of Watson and Crick to make a similar proposal for polymers of nucleic acids.
If we look at an alpha-helical region of peptide structure, the polypeptide chain is organized in a corkscrew with all the carbonyl oxygens along the helix pointing towards the carboxyl terminus and all of the amide hydrogens pointing towards the amino terminus. This arrangement of carbonyls and amide nitrogens with respect to the helix axis is a consequence of the particular combination of peptide backbone angles, and it means that hydrogen bonds can form between residues in one turn of a helix and residues in the succeeding turn. In an alpha helix, the dipoles of the carbonyl of residue i are perfectly aligned with the NH group of residue i+4 and so these two groups form a strong hydrogen bond (we call this an i,i+4 hydrogen bond). Since the carbonyl is the last of the three atoms in amino acid i and the amide nitrogen is the first of three atoms in the i +4 residue, there are 3.6 residues per turn of the helix and there are 13 atoms participating in each twist of the helix (counting from the carbonyl oxygen through to the amide hydrogen atom). There is a 1.5 Å rise per residue in an alpha helix, and so you can easily estimate the length of a helix from the number of residues within the helix. For example, a twenty residue helix would be approximately 30 Å long.

Within a helix, all of the hydrogen bonds between amide nitrogens and carbonyls are satisfied except at the two termini of the helix. The first four amide nitrogens and last four carbonyls are unpaired in an alpha helix. Because there are unsatisfied hydrogen bonding partners at the amino and carboxy termini, and all the dipoles in the helix are pointing in the same direction, there is a significant partial positive charge at the amino terminus and a correspondingly large negative charge at the carboxyl terminus. Therefore, alpha helices have large macrodipoles, and the electrostatic fields at each end of the helix can participate in attractive interactions with oppositely charged moieties. The side chains are arrayed around the helix axis and there are no unfavorable non-bonded steric interactions.
Beta strands organize very differently from alpha helices. Beta strands adopt a zig-zag conformation and the carbonyls and amides of successive residues within the strand point in opposite directions. Because of the extended conformation, there is no opportunity for hydrogen bonds to form within a beta strand. Instead, beta strands organize into larger structures called beta sheets in which hydrogen bonds form between adjacent strands. As the name suggests, beta-sheets create a relatively flat sheet-like surface, with side chains arrayed on both faces (see picture). There are two main classes of beta sheets – parallel and anti-parallel. In parallel sheets, the beta strands are aligned in a parallel fashion (i.e., the N-C direction of each strand is the same), whereas in anti-parallel strands the N-C chains run in opposite directions. The hydrogen bonding network of a sheet is not always perfectly defined and multiple variations can exist where the sheets can be twisted. Some beta strands can assemble into never-ending sheets as new strands are added to each side of the sheet. This can produce insoluble aggregates. In Alzheimer’s disease a protein in your brain is cleaved, and one of the resulting pieces forms an infinite beta sheet that aggregates to form plaques. The presence of these plaques is correlated with tissue death.
Does primary sequence determine protein conformation?

Thermodynamic Hypothesis: The information contained in the primary amino acid sequence leads to the correct protein fold. (or in other words . . . Could you put Humpty Dumpty back together again?)

Test: Will a denatured protein (RibonucleaseA) refold to its native conformation?

We have seen that proteins are comprised of alpha helices and beta strands, and we have learned various facts about their structures and properties, but one question we have not addressed is why some proteins are largely helical whereas others are largely comprised of beta sheet or mixed alpha/beta structures.

Christian Anfinsen proposed several decades ago that the tertiary structure adopted by a particular protein is a direct consequence of its primary sequence. That is, proteins somehow find the folded conformation that is lowest in energy -- meaning the conformation in which favorable interactions are maximized and unfavorable interactions are minimized.

To test this hypothesis, Anfinsen decided to denature a protein -- make it unfold -- and then see if it folds back into its native conformation in a test tube. He chose as the protein to study an enzyme called ribonuclease A, which cleaves the phosphodiester linkages of RNA. This protein is relatively small, and because it is an enzyme, proper folding can be assessed by evaluating enzymatic activity.