Week 3

Proteins

3.1 Amino Acids, Peptides, and Protein Synthesis

10/11:

- Initial impressions of the homework: More difficult than expected.
  - Tang did not raise the difficulty of the course, but was told to in course evals two years ago.
  - Literature problems: People believe reading the papers did help with the questions. We should be able to do these problems without the papers, though — some of these problems are past exam problems and were expected to be answered in a closed-book setting. We can expect similar questions on exams this year. Purpose: Show us how concepts from the class are used in research.
- There will be a practice exam posted.
- We can bring a one-page (single-sided A4) review sheet to the exam.
- OH Monday via Zoom.
- For every Thursday midterm, the content from the preceding Tuesday will not be covered.
- **Lesion**: Something bad that your cell will recognize and repair.
- Types of lesions:
  - Double-stranded breaks, mismatches, pyrimidine dimers, and damaged bases.
  - Are mutations lesions? It depends.
    - If the mutation is a mismatch, there will be a repair.
    - If it shows up as matched, your cell will not know to repair it.
  - DNA modifications.
    - Damaged backbones (e.g., pyrimidine dimers or a methyl group on the O^6 of guanine). Things your cells know shouldn't be there. Will be repaired.
    - However, there can also be intentionally placed modifications on DNA to regulate it. These will not be repaired.
  - Bulges don't usually occur during synthesis, but they can occur during recombination. These will definitely be repaired.
  - This should clarify some points on the homework.
- Natural base modification in mRNA.
  - mRNA is less diversely modified than tRNA and rRNA, but mRNA modifications do still happen.
  - Most abundant internal modifications in mammalian mRNA: N^6-methyladenosine (m^6A).

        - There's on average 1-3 of these per mRNA. However, there can still be 0.
        - How it's detected: Highly related to ChIP-Seq. You fragment your DNA, introduce antibodies that will bind to m^6A, do immunoprecipitation for a specific DNA binding protein to enrich the target DNA sequence, and sequence both the input and the enriched pool. The sequences that got enriched are the ones that carry the modification.
    - 5-methylcytosine (m^5C) and pseudouridine (ψ) are also present in mRNA, but their functions are less well studied.
        - Pseudouridine is a flipped uracil base with a carbon connected to ribose instead of a nitrogen connected to ribose.
        - The W-C interaction surface is basically unchanged, though, so it will still be detected as U. However, it has alternate regulatory functions, such as helping ribosomes read through premature stop codons.
        - The ψ detection method is messy (noisy): Introduce a chemical that selectively reacts with pseudouridine and gives a stop-signal during transcription. Not testable.
        - 5mC detection for RNA is identical to for DNA (bisulfite chemistry — see the discussion associated with Figure 2.7). Note, however, that since RNA is less stable, more will decompose upon heating; thus, you need a larger initial sample size.
    - In addition to m^6A, m^5C, and ψ, other base modifications can occur (we are not responsible for these, though).

- Summary of what we've learned so far:
    - DNA synthesis and transcription (the DNA $\rightarrow$ RNA part of the central dogma).
    - DNA methylation and epigenetics.
    - mRNA methylation and epitranscriptomics.
    - These three things function as a network (many feedback mechanisms). Moving forward, we will add proteins and metabolites to this network.

- Not testable: Arms race between bacteria and bacteriophages.
    - Answers how weird DNA modifications develop.
    - Bacteriophages are the most abundant life organism on this earth.
    - Round 1: Bacteria evolve restriction enzymes and base modification X; purpose: cleave phage DNA while avoiding suicide.
    - Round 2: Phages evolve X or Y modification in DNA; purpose: escape cleavage.
    - Round 3: Bacteria evolve X/Y-dependent restriction enzymes and additional self base modification Z; purpose: cleave phage DNA while avoiding suicide.
    - And on and on.

- Diverse base modifications in bacteriophages.
    - Guanine converts N^7 to a carbon and adds a functional group; you need multiple modifications to get to this result (called deoxyarchaeosine).
    - Cytosine attaches to glucose instead of deoxyribose.
    - Some bacteriophage DNA/RNA base modifications overlap with those in higher organisms, who evolve these modifications for completely different reasons.
    - And more.

- We are now done with last lecture's content; we are moving onto amino acids, peptides, and proteins.
    - Note that many of the mechanisms of RNA are more complicated than those of proteins, so if you have trouble with the latter, review the former.

  - Primarily amino acids this lecture; peptides, proteins, and higher-order structures next lecture.
  - Hopefully, these first six lectures will be foundational for the week 5-7 lectures on organelles and cell biology.

- A chemical look at proteins.
  - Made of proteinogenic amino acids (natural L-amino acids save glycine).
  - Can be post-translationally modified.
  - Post-translational rearrangement (lecture on this later).
  - We will look at amino acid properties.
  - Next lecture: Determinants of protein structure and...
  - Secondary and higher order structures.

- **Protein**: A polymer composed of amino acids.
  - Grows from the N-terminus to the C-terminus.

- Chirality is key.
  - Except for the achiral glycine, (almost) every amino acid is in its L-form.
    - Amino acids in their D-form are used as monomers, not for protein synthesis.
  - Steve Kent synthesized the D-form of HIV protease; he's a giant in the field. Taught here.
    - His big contribution is the development of **native chemical ligation**, while he was at Scripps.
    - We will talk about this more when we cover bioorthogonal chemistry.
  - No ribosomal D-protein synthesis because we would need an entire mirror image biological system.
    - People are trying to build a mirror ribosome, which Tang thinks is crazy, but they are making progress.
  - Total protein synthesis hasn't gone beyond 300 amino acids.
    - Solid state protein synthesis: Add one amino acid at a time. Highly efficient. 99.5% efficiency is great, but we have an exponential decrease of yield. Thus, we can't synthesize more than 50-100 amino acid peptides at a time.
    - Strategy: Fragments of 50 amino acids ligated together with natural chemical ligation.
    - But since proteins are folded as they're built in real life, we natural chemical ligation doesn't necessarily result in an accurately folded protein.

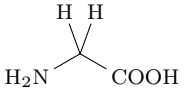
- **Native chemical ligation**: Connecting two peptides with an amide bond.
  - A very hard chemical problem; requires activating the amine of an amino acid.

- Taking advantage of D-proteins.
  - Why we want to do this: To challenge nature. Tang thinks this is stupid, though.
  - Favorable features of D peptides/proteins.
    - Similar to L proteins: bind to DNA/RNA/proteins, can catalyze reactions.
    - Cannot be degraded by natural protease (much more stable than natural peptides/proteins).
  - Challenges in identifying D peptides/proteins that bind specifically to a natural protein.
    - Rational design? — Hard to do with so many variables.
    - Screening? — Synthesize a D peptide library that can be amplified between selection rounds.
  - You can't synthesize a D library to look for hits on an L target (too hard; no mirror ribosomes). So instead, synthesize an L library, look for a hit on a D target, and then synthesize the D version of your L hit, which (flipping both chiralities) will react with your L target.

    - A brilliant idea but didn't turn out that well, though.
  - D-proteins aren't as big as they might be because there are many ways proteins can be degraded *in vivo* (not just natural proteases).

- Protein basics.
  - Classification of amino acids is somewhat arbitrary, but they are loosely categorized into hydrophobic, charged, polar, and glycine (in a class by itself since it's achiral).
    - For example, tryptophan could conceivably be hydrophobic or polar.
    - Histidine can frequently be charged.
    - Knowing properties is more important than knowing classes.
  - Knowing the amino acids is essential for predicting things like how amino acids interact with each other, what their role is in a reaction, how they catalyze a reaction, etc.
  - Memorize amino acids!
    - The 3-letter and 1-letter shorthand is often (but not always) the first 3 (resp. 1) letter(s) of the name.

- Achiral amino acid.

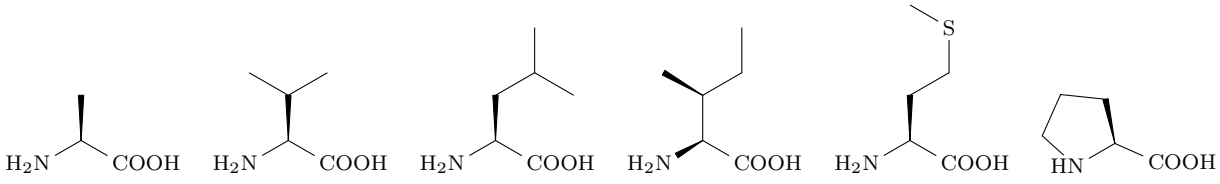


(a) Glycine.

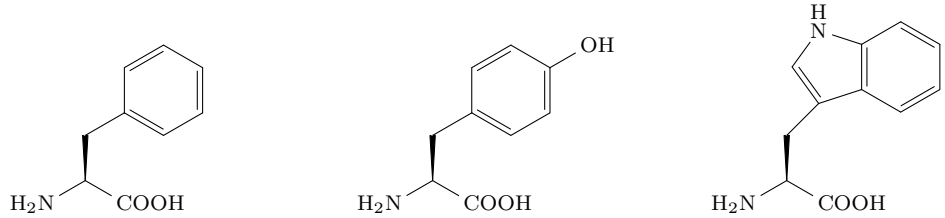
Figure 3.1: Achiral amino acid.

  - Glycine.
    - Flexible since it's unsubstituted on its α-carbon; can sample multiple conformations.
    - Whenever you have a glycine in your protein, you can assume the protein is flexible in that region.
    - If you want to fuse two proteins together but you're worried about sterics, you typically use a GGS (glycine, glycine, serine) linker.
    - Name: Glycine, Gly, G.

- Hydrophobic amino acids.



(a) Alanine. (b) Valine. (c) Leucine. (d) Isoleucine. (e) Methionine. (f) Proline.



(g) Phenylalanine. (h) Tyrosine. (i) Tryptophan.

Figure 3.2: Hydrophobic amino acids.

- We start with the *aliphatic* hydrophobic amino acids.
- Alanine.
  - Simplest chiral amino acid. That's what makes it important. No other important features.
  - If you think an amino acid is important, mutate it to alanine. If the protein is nonfunctional, then you know that it was important. You use alanine over glycine because it's less flexible.
  - Name: Alanine, Ala, A.
- Valine.
  - The simplest branched amino acid.
  - Name: Valine, Val, V.
- Leucine.
  - Name: Leucine, Leu, L.
- Isoleucine.
  - The second chiral center is not required (it is S though).
  - Name: Isoleucine, Ile, I.
- Note on valine, leucine, and isoleucine:
  - All are considered bulky, aliphatic amino acids.
  - Example (possible test question): Suppose you have an enzyme that fits ATP perfectly. If you want the active site to kick ATP out, you can mutate some of the amino acids to these three to make the pocket smaller.
  - Takeaway: Used to change the size of pockets.
  - Phenylalanine is another possibility, but it comes with other features as an aromatic system.
- Methionine.
  - One of the two amino acids containing sulfur; the other one (cysteine) forms disulfide bridges.
  - Frequently seen as a start codon (ATG), though it can appear in the middle of proteins, too.
    - ➢ Their are only two proteins that are encoded by a single codon; the other is tryptophan.
  - When we see a methyl modification, that methyl group is coming from a methionine derivative (specifically **SAM**).
  - Name: Methionine, Met, M.
- Proline.
  - Proline has a strained structure.
  - Whenever you have proline, the chain naturally has less flexibility.
  - You can only have two conformations: *cis*- and *trans*-proline (with respect to the nitrogen). *trans* is more common.
  - Proline is not in α-helices or β-pleated sheets because it typically induces a turn.
  - Name: Proline, Pro, P.
- We now move on to *aromatic* hydrophobic amino acids.
- Phenylalanine.
  - Name: Phenylalanine, Phe, F.
- Tyrosine.
  - Some people categorize tyrosine as polar.
  - The hydroxyl group is often phosphorylated; this derivative is called a **tyrosine kinase**.
  - Tyrosine kinases have been the most successful cancer drug target: You can somehow develop things that fit into the active site of one tyrosine kinase without affecting the rest of them.
  - Tyrosine kinases are less diverse then serine kinases and threonine kinases, aiding selectivity.
  - Name: Tyrosine, Tyr, Y.

  - Tryptophan.
    - Contains an indol moiety.
    - Only has one codon corresponding to it.
    - The heaviest amino acid.
    - The biosynthesis of tryptophan tends to be important, but we will not discuss it in this class. Proceeds through chromic acid.
    - Name: Tryptophan, Trp, W.

- (S)-adenosylmethionine (SAM) is a very important cofactor in our bodies.

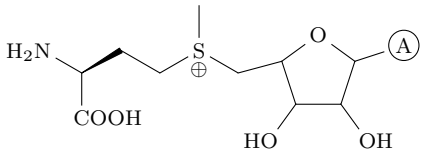
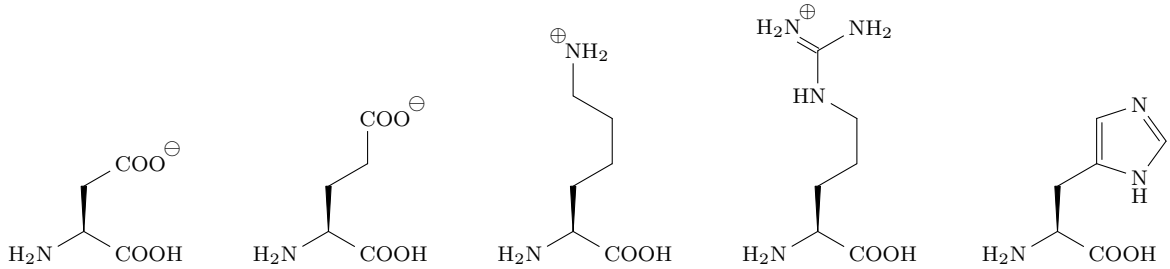


Figure 3.3: (S)-adenosylmethionine.

  - It donates the methyl group in DNA, RNA, and protein modification.
  - When the constituent moieties combine, S takes on a positive charge. This makes the lone methyl group on the sulfur a particularly good donor.

- Charged amino acids.



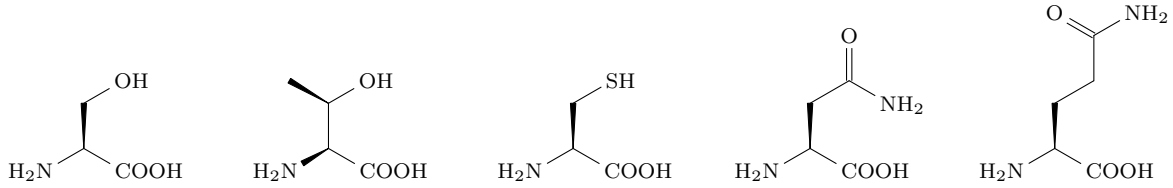
(a) Aspartate. (b) Aspartate. (c) Lysine. (d) Arginine. (e) Histidine.

Figure 3.4: Charged amino acids.

  - We start with the *negatively* charged ones.
  - Aspartate.
    - Under physiological pH, we draw the top "COOH" deprotonated; the other will be reacted.
    - Name: Aspartate (aspartic acid, if protonated), Asp, D. Alanine plus a carboxylic acid.
  - Glutamate.
    - Name: Glutamate (glutamic acid, if protonated), Glu, E.
  - We now move onto the *positively* charged ones.
  - Lysine.
    - An amine with $pK_a \approx 9 - 10$.
    - Name: Lysine, Lys, K.
  - Arginine.
    - Positively charged, but even more so under physiological pH. $pK_a \approx 12$.
    - Name: Arginine, Arg, R.

- Histadine is somewhat unique.
- Histidine.
  - Sometimes recognized as polar, but Tang prefers charged because it so frequently serves as the general base and acid in enzyme catalysis.
  - Contains an imidazole moiety.
  - The top nitrogen has $pK_a \approx 6$, so it can easily be protonated or deprotonated at physiological pH. Thus, it functions as a good **proton shuffle** to help catalyze acid/base reactions.
  - Some acid/base reactions can be catalyzed by lysine or aspartic acid.
    - For the nucleic acid polymerization reaction, the side chain is made of aspartic acid, which coordinates a metal ion to promote the reaction.
  - The other nitrogen does not easily lose its hydrogen.
  - Name: Histidine, His, H.

- **Proton shuffle**: A group that receives a proton from one group and donates it to another.
- Polar, uncharged amino acids.



(a) Serine. (b) Threonine. (c) Cysteine. (d) Asparagine. (e) Glutamine.

Figure 3.5: Polar amino acids.

- Serine.
  - Can be phosphorylated.
  - The hydroxyl group frequently serves as a nucleophile in the active site of enzymes.
    - Example (next time): **Serine protease**.
  - Name: Serine, Ser, S.
- Threonine.
  - Chirality: Same drawing style as isoleucine; R, though. Again, this chirality is not required.
  - Name: Threonine, Thr, T.
- Cysteine.
  - Very similar to serine.
  - Forms disulfide bonds to bring distal ends or subunits of a protein together.
  - Name: Cysteine, Cys, C.
- Asparagine.
  - Related to aspartate; we just change the carboxylic acid to an amide.
  - Frequently found as a metal coordinate.
  - Can also form H-bonds with other amino acids.
  - Name: Asparagine, Asn, N.
- Glutamine.
  - Related to glutamate; we just change the carboxylic acid to an amide.
  - Same metal-coordinating and H-bonding properties as asparagine.
  - Name: Glutamine, Gln, Q.

- A colleague asked Tang what AA he should substitute for alanine to prove that it's absolutely conserved.
  - She suggested fellow small amino acids (valine or leucine) as well as achiral glycine to determine if either size or chirality is important in that position.
  - Overall, this is a very hard to answer question.
  - You often find that alanine is needed because it doesn't disrupt anything; it's an inert filler and doesn't play a role. Other things will typically play a role.
- Amino acids: Hydrophobic side chains, acidic side chains, basic side chains, and special residues.
  - On the acidic side-chain amino acids: Sometimes the active site can be so well organized that replacing an D with an E will disrupt it.
  - In addition to cysteine, we sometimes have selenocysteine (it does occur in our bodies, but it's not considered one of the 20 natural amino acis).
    - Has selenium instead of sulfur.
    - Name: Selenocysteine, Sec, U.