

# CHEM 23300 (Introduction to Chemical Biology) Notes

Steven Labalme

October 19, 2022

# Weeks

<b>1</b>	<b>The Molecular Level</b>	<b>1</b>
1.1	Chemical Biology Introduction and the Central Dogma . . . . .	1
1.2	Chemistry and Biophysics of Nucleic Acids . . . . .	3
<b>2</b>	<b>DNA</b>	<b>11</b>
2.1	DNA Synthesis and Transcription . . . . .	11
2.2	Chemical Modifications of DNA and RNA Bases . . . . .	18
<b>3</b>	<b>Proteins</b>	<b>26</b>
3.1	Amino Acids, Peptides, and Protein Synthesis . . . . .	26
<b>References</b>		<b>34</b>

# List of Figures

1.1 Concentrations in biology . . . . .	2
1.2 Bases and nucleosides. . . . .	3
1.3 DNA sugars. . . . .	4
1.4 Base numbering. . . . .	4
1.5 Inosine. . . . .	4
1.6 Hydrogen bonding in bases. . . . .	5
1.7 DNA ionization states. . . . .	5
1.8 Base tautomerization. . . . .	6
1.9 Uracil tautomerization. . . . .	7
1.10 Puckomers of ribose. . . . .	7
1.11 G-quadruplex structure. . . . .	10
2.1 Mechanism of DNA synthesis. . . . .	13
2.2 DNA repair strategies. . . . .	17
2.3 Click reaction. . . . .	18
2.4 Catalytic RNA. . . . .	20
2.5 Mechanism of the ribosome. . . . .	22
2.6 Active methylation cycle of cytosine. . . . .	23
2.7 Bisulfite chemistry. . . . .	24
2.8 Naming convention for RNA base modifications. . . . .	25
3.1 Achiral amino acid. . . . .	29
3.2 Hydrophobic amino acids. . . . .	29
3.3 (S)-adenosylmethionine. . . . .	31
3.4 Charged amino acids. . . . .	31
3.5 Polar amino acids. . . . .	32

# Week 1

## The Molecular Level

### 1.1 Chemical Biology Introduction and the Central Dogma

9/27:

- Questions:
  - What edition(s) of the textbook(s) should we have?
    - Doesn't matter.
  - Will there be TA office hours?
    - No.
- CHEM 233 used to be Intermediate Organic Chemistry, and CHEM 332 was the grad class. They have been merged this year because of the overlap in content.
- Krishnan weeks 5-7; Tang otherwise.
- We will not be going through reactions. The format is slides; don't try to copy them down, just make some notes. Copy them down ahead of time!
- Goes over the syllabus.
  - No fixed textbook. Lehninger is recommended though. Whatever edition you can find.
  - No office hours (ask questions in class or ask her to meet outside).
    - Tang will show up early and stay late.
  - Midterms are 1 hour; final is 2 hours.
  - Three problem sets.
  - One in-class quiz:
    - Krishnan will give us cutting-edge literature to read one week before the quiz and 5 questions.
    - We can form study groups to discuss the questions.
    - Multiple choice quiz on that day.
  - We're not supposed to memorize things in this course; the problems won't be like that.
  - Tang may lower the exam difficulty levels from previous years.
  - Tang doesn't want us to have to fight for points; is trying to give us a big curve so that we can just focus on learning.
  - Since this is now only a twice a week class, Tang is cutting material on carbohydrates and protein design. May try to squeeze in orthogonal chemistry, though.
- The central dogma in biology. *picture*
  - DNA → RNA → protein → needed chemical transformations.

- Size in biology.
  - An activity matching biological entities (e.g., E. coli, cells, RNA) to their sizes in microns.
  - Uses the world zoom website.
  - We may be tested on sizes, but only relative not exact (e.g., E. coli vs. a ribosome).
- Red blood cells are smaller than normal cells because they don't have nuclei, and they don't need meat to divide.
- Concentrations in biology.

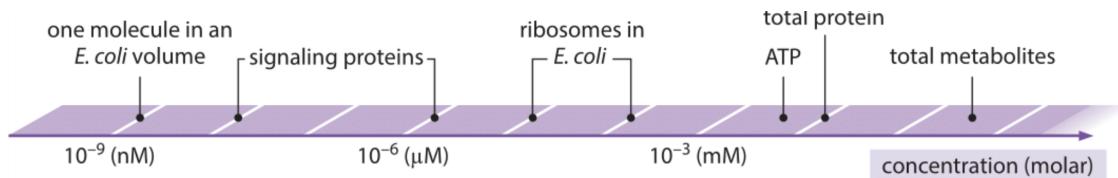


Figure 1.1: Concentrations in biology.

- You need a couple of copies of signaling proteins.
- Cells dedicate a lot of resources to building ribosomes.
- Different ions have different concentrations in different parts of the body. Additionally, different types of cells have different concentrations.
- *Bound* divalent ions such as  $Mg^{2+}$  help cancel the charge of ATP; that's why we need them in solution.
- The materials left after we remove all of the water from our cells.
  - Largely protein, lipid, rRNA.
  - Far more mRNA and proteins in mammalian cells than bacterial cells.
- Time for protein diffusion within a cell.
  - Time scale  $\tau$  to traverse distance  $R$  given diffusion coefficient  $D$ :
$$\tau = \frac{R^2}{6D}$$
  - For a protein in cytoplasm,  $D \approx 10 \mu\text{m}^2/\text{s}$ .
- The molecular hierarchy of structure.
  - The cell and its organelles are made of supramolecular complexes (e.g., the plasma membrane, chromatin, and the cell wall), which are made up of macromolecules (e.g., DNA, proteins, cellulose), which are made up of monomeric units (e.g., nucleotides, amino acids and sugars).
- We will be expected to know how to draw the amino acids and nucleic acid bases.
- We will not talk much about lipids and sugars.
- Chirality and isomers review.
- Thalidomide.
  - Was only distributed in Germany; the FDA is very proud of having picked up on the scientific malpractice and barred it from ever entering the US.
  - Just selling one isomer doesn't work because it racemizes so quickly.
  - Now used to treat cancer; you have to sign a bunch of paperwork saying that you won't get pregnant before you use it.

## 1.2 Chemistry and Biophysics of Nucleic Acids

- 9/29:
- Feel free to come by and introduce yourself now that the class is a more manageable size.
  - DNA and RNA basics.
  - Humans have on the order of  $3 \times 10^{13}$  cells and on the order of 1 m of DNA in each cell.
    - Calculated by multiplying the number of base pairs per cell ( $\approx 3 \times 10^9$ ) by the length of each base pair ( $\approx 3.3 - 3.4 \text{ \AA}$ ).
    - Some people say 2 m because we have two copies of our genome.
    - DNA wraps around histone proteins to form chromosomes to fit into such a tiny space.
  - DNA is ACTG. RNA is ACUG.
  - **Bases** and their corresponding **nucleosides**.

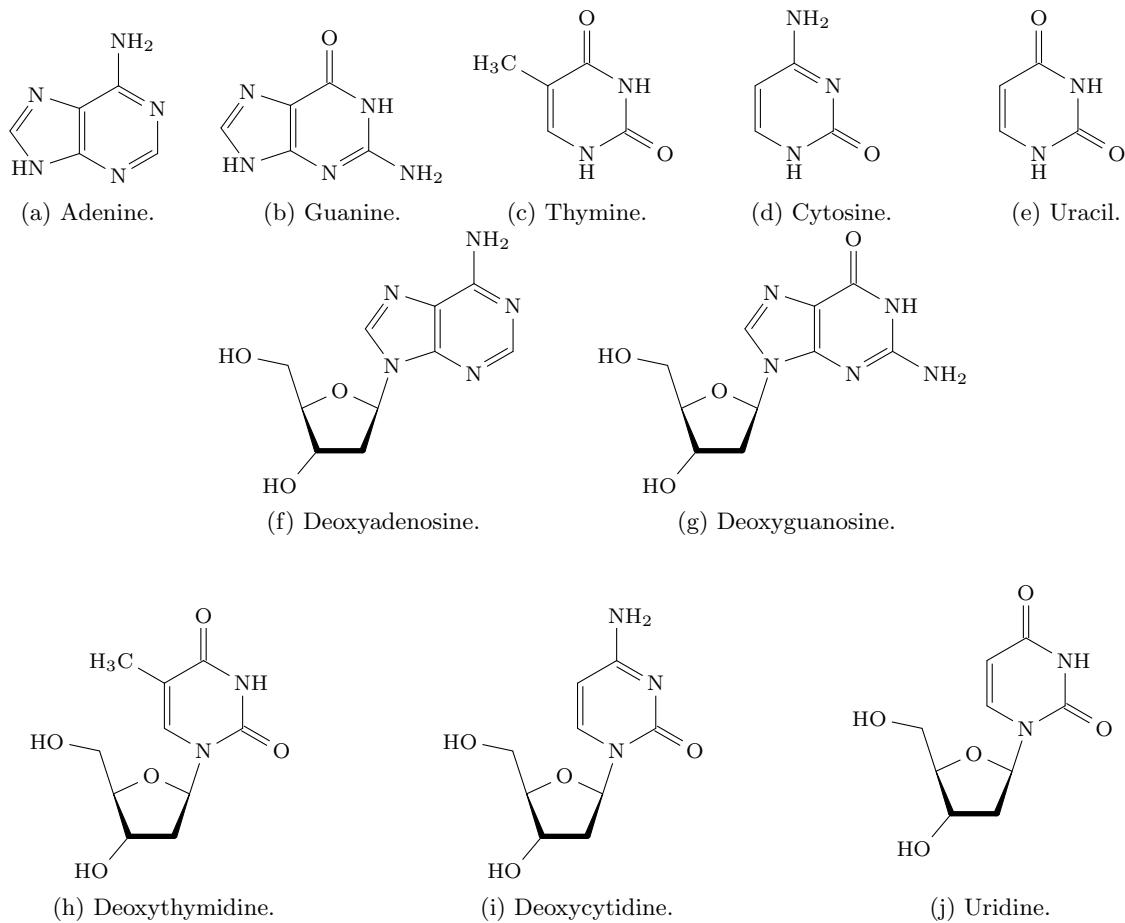


Figure 1.2: Bases and nucleosides.

- Notice that deoxyribose is joined with the base at its  $2'$  carbon.
- If we use ribose instead of deoxyribose, we get adenosine, guanosine, etc.
- Memorize these structures!
- Nomenclature.
- **Base:** The heterocycle. *Also known as nucleobase.*

- **Ribose:** A 5-carbon monosaccharide, a derivative of which is a component of DNA.
- **Deoxyribose:** A molecule identical to ribose but without the 2' hydroxyl group.

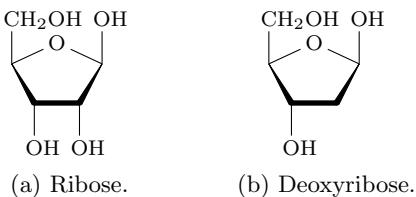


Figure 1.3: DNA sugars.

- **Nucleoside:** Base + sugar.
- **Nucleotide:** Base + sugar + phosphate(s).
- Base numbering.

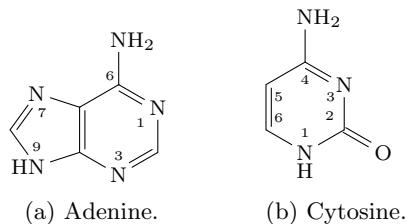


Figure 1.4: Base numbering.

- Generalize from the above two examples.
- Sugar numbering: Start to the right of the oxygen and move clockwise. Use primes to distinguish from the numbered base carbons.
  - Remember that DNA and RNA run 5' to 3' with phosphate groups linking the deoxyribose groups.
- Listing features common to all or some of the bases.
  - E.g., heterocycles, on the way to being or already aromatic, nitrogen in the ring, oxygen only ever outside the ring, etc.
- **Inosine:** An intermediate between adenine and guanine. *Also known as hypoxanthine. Structure*

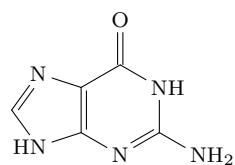


Figure 1.5: Inosine.

- Common modifications to adenine: Methylation at the 1 or 6 position.
- Five percent of cytosine exists in its methylated form; important epigenetically in determining which genes get turned on and off.
  - Methylation of cytosine occurs at the 5-carbon.

- Hydrogen bonding between bases.

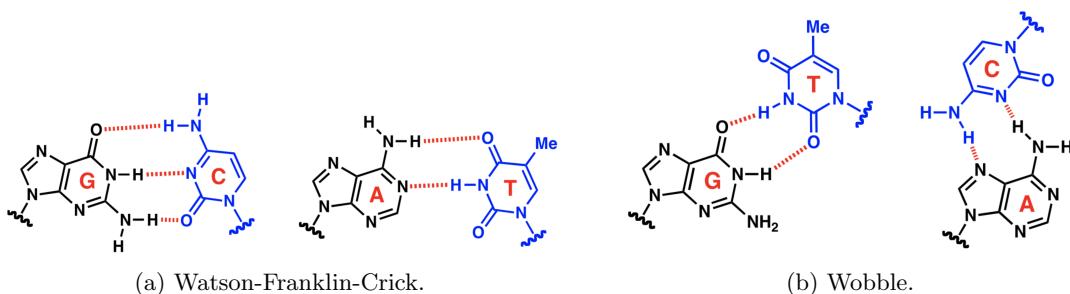


Figure 1.6: Hydrogen bonding in bases.

- Hydrogen on a heteroatom serves as a donor; heteroatoms serve as acceptors.
  - There are many types, but we're only responsible for Watson-Crick-Franklin and Wobble interactions.
  - Watson-Crick-Franklin is the standard interaction found in double helix DNA.
  - Wobble:
    - G/T is more common and important than C/A in nature.
  - A triplet codon NNN yields an amino acid. There are  $4^3 = 64$  possible codons but only 20 amino acids. Thus, some codons code for the same base. For example, NNC and NNT always encode for the same base since C normally pairs with G and T can be paired with G via a wobble interaction.
    - Something about the pairing of strands of DNA with lots of G's and T's.
  - $pK_a$  review.
    - MeNH<sub>2</sub>'s protonated form has  $pK_a \approx 10.6$ .
    - Aniline's protonated form has  $pK_a \approx 4.6$  because aniline is a weaker base.
    - Pyridine's protonated form has  $pK_a \approx 5$  because it is basic, but it is also  $sp^2$ .
    - An amide has  $pK_a \approx 18$ .
      - Did Tang switch from doing the  $pK_a$  of the conjugate acid to doing the  $pK_a$  of the molecule itself here? Why?
      - Ethanol has  $pK_a \approx 16$ .
  - Predicting DNA ionization states

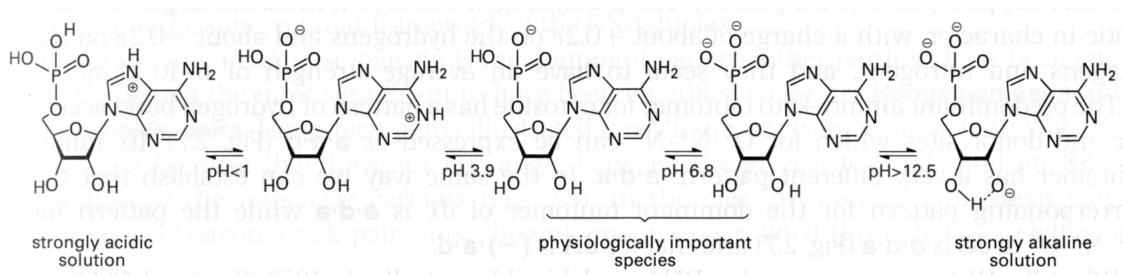


Figure 1.7: DNA ionization states.

- At physiological pH (5-9), only phosphates are charged (as desired).

- Phosphate  $pK_{as}$ : About 1-2 and 7.
- Ribose with free 2' and 3' -OH groups:  $pK_a \approx 12.4$  (vs. 15-16 for an isolated secondary alcohol).
- Why are the heteroatoms on adenine that get protonated the ones that do?
- These numbers can change a lot after polymerization.
- Anti and syn base conformations.
  - We have free rotation about the glycosidic  $C^{1'}-N$  bond, subject only to the whims of sterics.
  - This leads to **anti** and **syn** conformations.
  - Anti is preferred among natural nucleotides for steric reasons.
  - Exceptions:
    - G prefers syn in mononucleotides, in alternating CpGpCpG oligonucleotides, and in Z-DNA.
    - Non-natural nucleotides can shift the equilibrium towards syn.
      - Examples: 8-bromoguanosine ( $N^3$  of the now-electron-deficient heterocycle seeks stabilization through an H-bonding interaction with the 5' hydroxyl group, but this requires a syn conformation to be most efficient [i.e., to bring the involved atoms close together]) and 6-methyluridine (Me is more bulky than =O, so it sits anti to the sugar).
- Bulk (of the base):  $O^2$  (the oxygen attached to the 2-carbon) in pyrimidines or the whole six-membered ring in purines.
  - See Figure 1.2.
- **Anti** (base conformation): The bulk of the heterocycle points away from the sugar.
- **Syn** (base conformation): The bulk of the heterocycle is over the sugar.
- Base tautomerization basics.

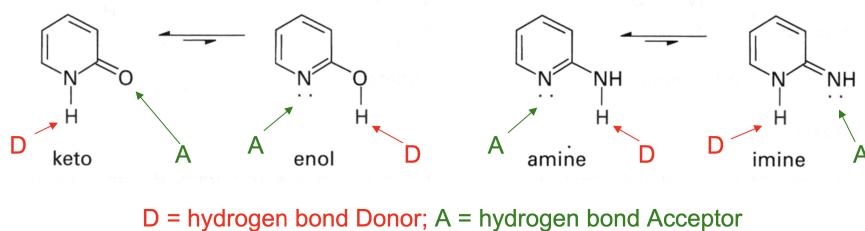


Figure 1.8: Base tautomerization.

- Recall that tautomerization involves movement in atoms whereas resonance does not.
- Bases exist in equilibrium between keto and enol forms, and between amino and imino forms.
- Tautomerization changes which groups function as hydrogen bond donors and acceptors in base pairing.
- The keto and amino forms among natural bases are preferred by more than 99.99%, according to X-ray and NMR analyses.
- It is difficult to determine what form a base is in just via organic chemistry first principles.
  - Sometimes, tautomerization will do something highly unfavored like breaking aromaticity. But other times, making a system aromatic will generate an unstable enol. Confounding factors like this make it hard to tell.
  - In fact, when Watson and Crick were originally solving the structure of DNA, they had it backwards until a physical chemist wrote to them with a calculation suggesting the right form, and that allowed Watson and Crick to solve the structure right away.

- Enol and imino tautomers lead to mutagenic H-bonding patterns.

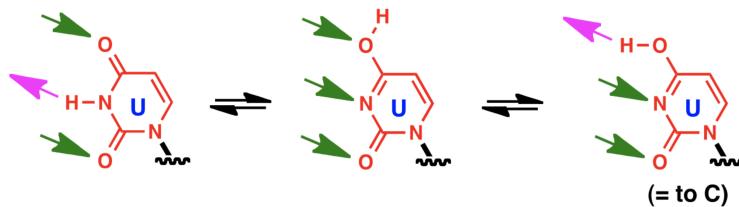


Figure 1.9: Uracil tautomerization.

- Because donor/acceptor dynamics are shifted, tautomers of one base can look like *another* base from a hydrogen-bonding perspective.
- The tautomerization equilibrium can be shifted by functionalizing the base pairs. This is why bromine is a mutagen — it makes it far more likely for U to be read as C, for instance.
- Tang goes over the tautomers for the other bases, too (see slides).
- Ribose exists in many conformers.

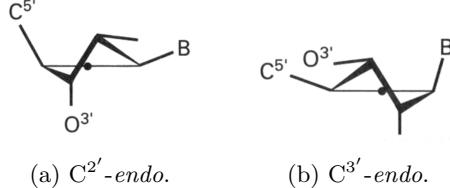


Figure 1.10: Puckomers of ribose.

- The furanose ring is nonplanar/puckered to minimize non-bonded interactions between substituents.
- Endo** and **exo** atoms.
- Puckomers** are in rapid equilibrium with an energy barrier of less than 5 kcal/mol (higher in polymeric DNA/RNA).
- Crystallography and NMR suggest two preferred puckomer groups:  $C^{2'}\text{-}endo$  and  $C^{3'}\text{-}endo$ .
  - In the former, the 2'-carbon of (deoxy)ribose is endo (and the 3'-carbon exo; all others lie in the plane).
  - In the latter, the 3'-carbon of (deoxy)ribose is endo (and the 2'-carbon exo; all others lie in the plane).
- Sugar conformation dramatically changes the shape of the duplex.
  - B-DNA favors  $C^{2'}\text{-}endo$  sugars.
    - Since DNA runs 5' to 3', as we can see in Figure 1.10a,  $C^{2'}\text{-}endo$  gives us a more stretched out/relaxed form of the polymer. B-DNA is the most common form of DNA.
  - RNA and A-DNA favor  $C^{3'}\text{-}endo$  sugars.
    - Conversely,  $C^{2'}\text{-}endo$  gives us a much more compact/bunched up form of the polymer.
- Endo** (atoms): Atoms on the same side of the furanose as  $C^{5'}$ .
- Exo** (atoms): Atoms on the opposite side of the furanose from  $C^{5'}$ .
- Puckomer**: A ribose conformer.

- Why RNA contains uracil and DNA contains thymine.
  - There is a slow but appreciable rate of hydrolysis of C to U (500 times per cell per day).
  - When C → U in DNA, because uracil is not a typical constituent of DNA and can easily be distinguished from T (T has an additional methyl group), our DNA correction mechanisms can easily repair the error, preventing our DNA from mutating long-term.
  - RNA, on the other hand, cannot be edited. However, RNA is transient but DNA is not, and it is highly unlikely to have the same mutation at the same position every time. Thus, a few proteins are liable to get messed up in a variety of different ways from mutated mRNA, but long term, the base genetic code in DNA is preserved.
  - Note that this is just a hypothesis (and a hard one to test), but it seems reasonable.
- Why nature chooses phosphates.
  - Requirements any possible linking group for nucleic acid monomers must satisfy.
    1. Multivalent (so it can connect two monomers).
    2. Cannot cross biological barriers (e.g., the nuclear membrane).
    3. Kinetically stable to hydrolysis (we don't want our DNA strand to be breaking at random all the time).
    4. Thermodynamically unstable/must exist in high energy forms (we want the synthesis of the polymer [which involves cleaving some phosphate groups] to be thermodynamically favorable).
    5. Kinetically unstable with catalyst: modulated reactivity (so an enzyme can hydrolyze it; we don't want it to be so stable that we can't work with it).
  - Phosphate groups satisfy these requirements since they...
    1. Are divalent.
    2. Are polar.
    3. Are negatively charged (nucleophiles that might hydrolyze it are Coulombically repelled).
    4. Can exist in high energy forms (such as ATP).
    5. Are more reactive in the presence of magnesium.
  - Some possible alternatives include citric acid, arsenate esters, silyl esters, and amides.
    - Citric acid is abundant, but ester bonds are unstable in biological systems and the negative charges are quite far apart (so nucleophilic attack is not as hindered).
    - Arsenate and silyl esters are also too labile.
    - Amides are too stable; we can't hydrolyze it easily with any sort of catalyst.
      - Scientists have used amides to connect nucleobases in the lab, though.
  - This is another hard-to-test hypothesis that seems reasonable.
- What binds two strands of DNA.
  - Not hydrogen bonds.
    - These only decide specificity; there is no thermodynamic preference for two-stranded DNA over single-stranded DNA hydrogen bonded unspecifically to a bunch of water molecules, for instance.
  - Stacking, on the other hand, is key.
    - It excludes water and maximizes van der Waals interactions.
    - More explanation?
    - Not testable material.
- DNA and the double helix.
  - DNA can occur in different 3D forms.

- Nucleic acids in higher order structures (e.g., tRNA and G-quadruplex).
- Geometric parameters.
- DNA and RNA polymorphism.
  - Various forms exist and are interchangeable; we don't need to know most of them.
  - Determinants of DNA and RNA forms.
    1. Sequence (not only composition).
    2. Counter ion and [salt].
    3. Humidity (crystals).
    4. Temperature.
  - Not testable material.
- Major nucleic acid forms.
  - We are responsible for A-DNA, B-DNA, and Z-DNA.
    - A- and B-DNA are most important; then Z-DNA.
  - A- and B-DNA are right-hand double helices; Z-DNA is a left-hand double helix.
  - The number of base pairs per turn of...
    - A-DNA is 11;
    - B-DNA is 10;
    - Z-DNA is 12.
  - The rise per base pair of...
    - A-DNA is 2.9 Å;
    - B-DNA is 3.3-3.4 Å;
    - Z-DNA is 3.7 Å.
  - Other important numbers?
  - In B-DNA, the base pairs are relatively centered within the strand; in A-form, they rotate around.
- B-DNA.
  - Base pairs on center of helical axis.
  - Major and minor is an accurate descriptor.
    - What are these grooves and what is their significance?
  - Both grooves have similar depths.
  - Sugar pucker is C<sup>2'</sup>-endo (2' and 5' on the same side).
- A-DNA.
  - Base pairs are displaced from center of helical axis.
  - Major groove less wide than minor.
  - Sugar pucker is C<sup>3'</sup>-endo (3' and 5' on the same side).
  - DNA/RNA hybrids are A-like (transcription, reverse transcription, and DNA replication).
- Enzymes recognize the 3D helical structure of DNA, not just their individual substrate. Like reading a word instead of letter by letter.
- Higher order structures.
  - The structure of tRNA provides a wealth of information.

- Until the early 1990s, we could only crystallize tRNA, so we primarily learned from it for a long time.
  - L-shape: Two perpendicular A-RNA helices.
  - Tons of fun H-bonding interactions provide structure. Even three nucleobases can interact all together in some cases.
- G-quadruplex.

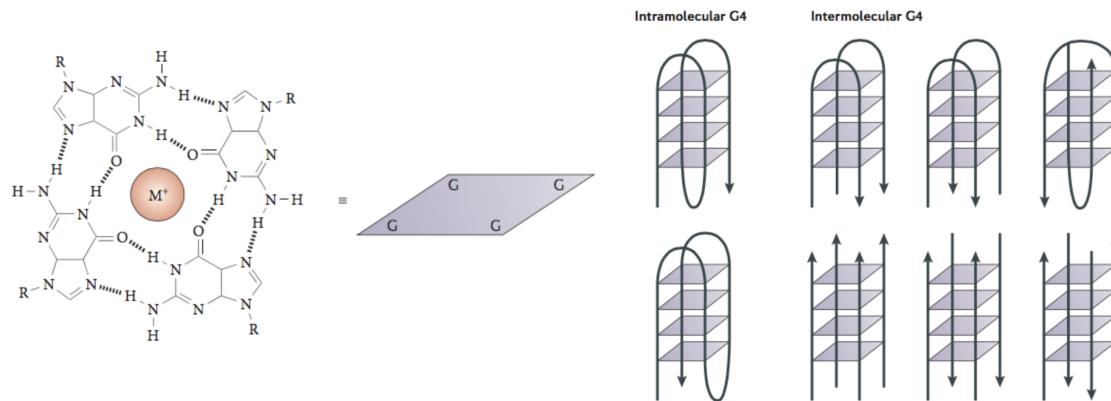


Figure 1.11: G-quadruplex structure.

- Helical structure containing guanine tetrads from one, two, or four strands.
- Hoogsteen hydrogen bonding.
- Stabilized by the presence of a cation, especially potassium.
- Importance of G-quadruplexes.
  - Chromosomes have a structure called a telomere at both ends. During replication, some of the telomere is lost each time. When the whole telomere is gone, the cell will not be able to divide any more. This is the aging process, and the discovery of telomeres received the Nobel prize in the early 2000s.
  - Telomerase is an enzyme that fights against this loss, trying to extend the DNA post-replication using RNA templates.
  - If telomerase is overactivated, the cells are immortalized and they become cancerous.
  - Telomeric quadruplexes decrease the activity of telomerase; they moderate telomerase so that it's active enough so that we don't die early, but not so active that we become big balls of cancer.
- The spinach aptamer.
  - For a long time, we've been able to tag any *protein* we want with GFP (green fluorescent protein) and follow it.
  - It would be very beneficial to be able to do the same thing with RNA.
  - Thanks to the Jaffrey lab, now we can with DFHBI.
  - The spinach aptamer binds to DFHBI and becomes fluorescent in the presence of RNA. DFHBI's  $\pi$  structure too bendy to fluoresce on its own, but when it is stabilized by insertion into RNA, it can fluoresce.
  - We still need a lot of work before this is as good as GFP.
- The remaining slides will be covered next lecture.

# Week 2

## DNA

### 2.1 DNA Synthesis and Transcription

10/4:

- DNA binding proteins.
  - Interact with major grooves of DNA to achieve sequence specificity.
  - Example: Transcription factors that have to turn a gene on or off.
  - Such proteins often do this with two primary motifs: The leucine zipper and the zinc finger.
    - Leucine zipper: Less programmable.
    - Zinc finger: More programmable. Contains 1+ zinc ion coordinated by cysteine and histidine. One zinc finger interacts with three base pairs.
  - With the scale of our genome, you typically need a sequence of 15-18 base pairs to achieve specificity. That's 5-6 zinc fingers.
  - When they were discovered, zinc fingers were thought to be a possibility for genome editing.
- DNA structure and binding modes.
  - DNA binding proteins may also interact with the minor group (to achieve some specificity), electrostatically with the phosphate backbone (usually not sequence-specific), and intercalation (a flat molecule splits two base pairs a bit and inserts itself in).
  - Minor groove: Deep and narrow.
    - Minor groove binding can involve electrostatics, hydrophobic burial, and hydrogen bonding.
    - Minor groove binding can be site specific. There are some features you can take advantage of, e.g., being flat and flexible.
  - Pyrrole-Imidazole-Hydroxypyrrrole (Py/Im/Hp) Polyamides.
    - Pioneered by the Dervan lab at CalTech.
    - Minor-groove binding polyamides consisting of three aromatic ring amino acids.
    - Flexible because of the amides.
    - Eight-ring pyrrole-imidazole polyamides achieve affinities and specificities comparable to DNA-binding proteins.
    - Cell-permeable molecules for gene-specific regulation *in vivo*.
    - Still not specific enough, though.
  - Minor groove-binding small molecules.
    - Examples (not testable material): Hoechst 33258, DAPI, Distamycin, and Berenil.

- Distamycin in the minor groove. Distamycin is an antibiotic and it fits very well into the minor groove. Preference for A/T sequences.
- Hoechst 33258 in the minor groove. Also fits very well; used to some extent to dye DNA, but more often in flow cytometry.
- Common features (testable material):
  - Flat (to slip into the minor groove).
  - Small linked aromatic ring systems (to allow ring systems to make local adjustments).
  - Curved (to match curvature of the minor groove).
  - Positively charged (to interact with the phosphates).
  - H-bond donors on concave face (to H-bond with acceptors on base pairs).
- Phosphate backbone binding.
  - Driven by electrostatics.
    - Ligands that bind this way are always cationic, binding depends strongly on salt concentration.
    - Ions ( $\text{Na}^+$ ,  $\text{Mg}^{2+}$ , etc.)
  - Example: Biogenic polyamines (involved in a lot of biological processes).
    - For example, putrescine is a positively charged polyamine. It is responsible for bad breath!
  - **Transfaction reagents** wrap around DNA and neutralize some of its negative charge to help it get into cells.
- Intercalators.
  - Example: Ethidium bromide (a toxic molecule used to stain DNA).
  - Features in common:
    - Extended aromatic systems (to provide extensive overlap with base pairs).
    - Electron deficient (to complement regions of high electron density in base pairs).
- Intercalation requires structural rearrangement.
  - The intercalator does disturb DNA structure a bit as it pushes base pairs farther apart, causing buckling in adjacent base pairs and a tilt in the helical axis.
  - Because of this, intercalators can alter DNA replication.
- Ethidium bromide as a DNA dye.
  - Biochemical analysis of DNA (gel electrophoresis).
    - Agarose is used for DNA strands over 300 bp. Bands greater than 100,000 bp are not resolved.
    - Concentration of agarose can be increased to create a denser matrix, or decreased to create a less dense matrix.
    - Larger molecules need a less dense matrix.
    - Because DNA is negatively charged, it migrates to the cathode (+) of the electrophoresis system.
  - Ethidium bromide is toxic: It can act as a mutagen because it intercalates double-stranded DNA and, as mentioned, affects replication.
    - If you add too much ethidium bromide into your PCR, it may not work (the polymerase may not be able to overcome it).
  - Safer options are offered by many biotech companies: sybr-green, sybr-gold, etc.
    - Tang isn't sure how much safer these actually are.

- The common design is bigger molecules: Will still intercalate DNA, but will not penetrate the skin as easily.
- Summary.
  - Several forms of DNA/RNA (most important ones: A-RNA and B-DNA).
  - Unusual forms may also play an important role (e.g., G-quadruplex and tRNA).
  - 
  - Molecules that interact with the DNA.
    - Major groove interactions are sequence specific.
    - Minor groove interactions can be sequence specific.
    - Phosphate backbone/intercalation interactions are (typically) not sequence specific.
- This is the end of the previous lecture's slides.
- Now: Replication, transcription, translation, and nucleic acid catalysis.
- Overview.
  - DNA replication in cells: DNA-templated synthesis of DNA.
  - DNA repair in cells: Enzymatic repair of DNA mutations
  - DNA transcription in cells: DNA-templated synthesis of RNA.
  - Translation: Nucleic acid catalysis (difficult) and RNA-templated synthesis of proteins.
- If you find the early topics above difficult, review the relevant sections of Nelson and Cox (2021).
- DNA is synthesized/replicated by DNA polymerase.

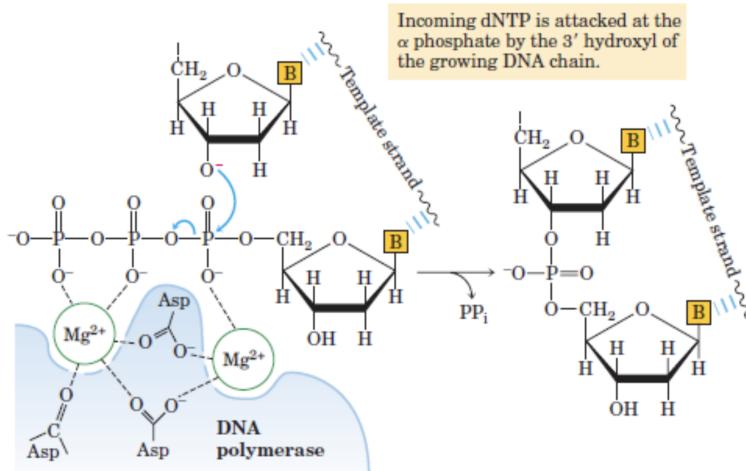


Figure 2.1: Mechanism of DNA synthesis.

- DNA polymerases require three components: A template, a primer, and dNTPs<sup>[1]</sup> ( $N = A, T, G$ , or  $C$ ).
- Mechanism: The 3' hydroxyl of the growing DNA chain attacks the  $\alpha$  phosphate of the incoming dNTP via nucleophilic acyl substitution.
  - Most textbooks will draw the electron pushing as a substitution reaction, but in reality, the double bond gets resolved, and then kicks back down to get rid of the  $\beta$  and  $\gamma$  phosphates.

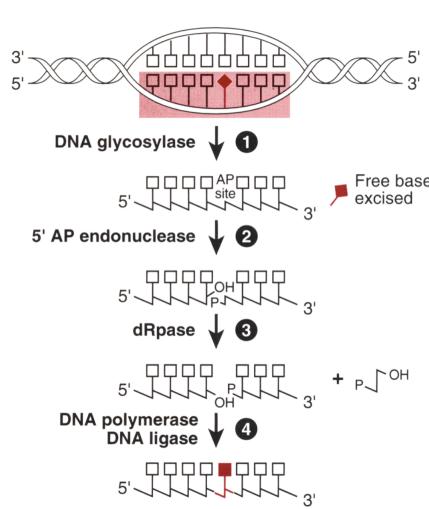
<sup>[1]</sup>Deoxynucleoside triphosphate

- Energetically driven by the BDE of the dNTP, so as long as dNTP is abundant, the reaction can proceed.
- In bacteria, this process can occur at a rate of 1000 bp/second.
  - Our cells are slower.
- Notice the presence here of magnesium (partially coordinated to aspartic acid) catalyzing the reaction as part of DNA polymerase.
  - One  $Mg^{2+}$  coordinates with the  $\beta$  and  $\gamma$  phosphates to stabilize them.
  - The other coordinates with the  $\alpha$  phosphate to stabilize the highly negatively charged nucleophilic acyl substitution intermediate.
- DNA replication in *E. coli* is highly accurate (error rate  $10^{-9}$ - $10^{-10}$ ).
  - Two reasons: Tempered synthesis (error rate of  $10^{-4}$ - $10^{-5}$  *in vitro*) and error correction mechanisms.
  - Tempered synthesis is equivalent to having a 99.999% yield in an organic reaction (which never happens).
  - One error-correction mechanism bridging the gap from  $10^{-5}$  to  $10^{-10}$ : DNA polymerase “proof-reads” even as it synthesizes DNA.
  - Example procedure:
    - Let  $C^*$  be a rare tautomer of cytosine that pairs with A and is incorporated into the growing strand.
    - Before the polymerase moves on, the  $C^*$  reconverts to C and is now mispaired.
    - The mispaired 3'-OH end of the growing strand blocks further elongation. DNA polymerase slides back to position the mispaired base in the  $3' \rightarrow 5'$  exonuclease active site.
    - The mispaired nucleotide is removed.
    - DNA polymerase slides forward and resumes its polymerization activity.
  - Not every polymerase has this feature, but most high-fidelity ones do.
- DNA replication in cells.
  - DNA replication is semiconservative.
    - Meselson-Stahl experiment, “the most beautiful experiment in biology.”
  - DNA replication begins at an **origin** and proceeds **bidirectionally**.
  - Bacterial chromosomes have a single point of origin; most other cells have multiple such points.
- DNA replication requires many enzymes and protein cofactors.
  - DNA replication in cells requires much more than solely polymerases.
  - DNA replication in *E. coli* requires 20 or more different enzymes and proteins, each performing a specific task.
  - DNA replicase system (replisome): The entire complex is required for DNA replication.
- Three identifiable phases of DNA replication in *E. coli*.
  1. Initiation.
    - Five repeats of 9 bp (R sites) for DnaA binding; A = T rich DNA unwinding element (DUE).
    - DnaA binding, DUE denaturing, DnaB helicase loading, then ready for the next phase.
    - Initiation is the only phase of DNA replication that is known to be precisely regulated (replication occurs only once in each cell cycle). You don’t want the daughter cells to have multiple unneeded copies of the genome.
  2. Elongation.

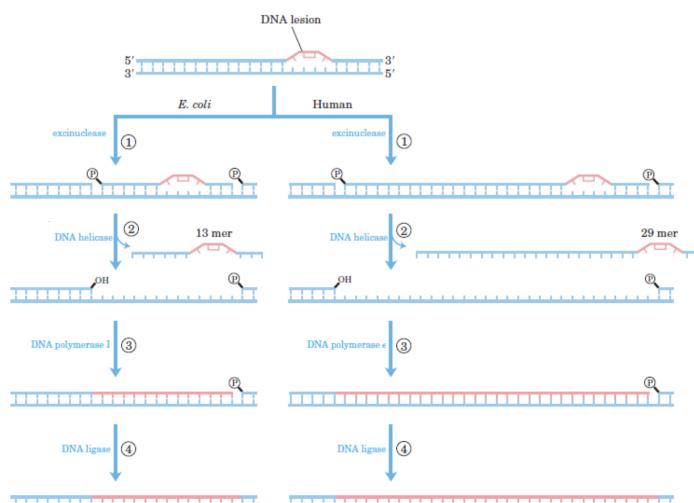
- Single-stranded DNA-binding protein (SSB) stabilizes the regions denatured by helicase.
  - Two distinct but related operations: Leading strand synthesis and lagging strand synthesis.
  - Lagging strand synthesis requires RNA primers (synthesized by primase) to form Okazaki fragments.
    - Regardless of whether it's DNA- or RNA-templated synthesis, it's always primed.
  - After completion of an Okazaki fragment, RNA primer is removed and replaced with DNA by DNA polymerase I, and the remaining nick is sealed by DNA ligase.
3. Termination.
- Ter sequences: Trap for the bidirectional replication of DNA by forming Tus-Ter complex — prevent overreplication by one replication fork when the other fork is abnormally delayed or halted.
  - **Catenane** formation: Bidirectional replication of DNA meets at the end.
  - DNA topoisomerase IV: Separate the catenated chromosomes into normal chromosomes for the daughter cells.
- **Catenane:** A mechanically interlocked molecular architecture consisting of two or more interlocked macrocycles<sup>[2]</sup>.
  - We don't have to remember every protein; Tang just wants to show us.
    - Note that the ones that she did show us, though, are all involved in elongation; initiation and termination require additional proteins.
    - Also, new proteins are still being discovered.
    - DNA replication in eukaryotic cells is both similar and more complex than in *E. coli*.
  - Mutations happen constantly in DNA.
    - A race between mutation and repair.
    - Why mutations don't generally affect us: Mutations could occur in DNA that is not used in a given cell (most DNA in any given cell is dormant), the cell could die, etc. Also, only 1% of our genome is protein-coding. And most amino acids in our proteins are not fully **conservative**. Significant ones are very rare (and usually lead to apoptosis anyway, so no problem).
    - Transition (purine to purine, or pyrimidine to pyrimidine), transversion (purine to pyrimidine or vice versa), and frameshift (insertion/deletion by  $3n \pm 1$ ) mutations.
      - Frameshift typically corresponds to early stop codon.
    - Mutation locations and effects.
      - Promoter: Reduced or increased gene expression.
      - Regulatory sequence: Alteration of regulation of gene expression.
      - 3' of protein-coding region: Defective transcription termination or alternation of mRNA stability.
      - Certain locations within intron: Defective mRNA splicing.
      - Origin of DNA replication: Defect in initiation of DNA replication.
    - Many disease-causing mutations in humans are non-coding.
    - Mutations in one place can interact with other base pairs a few away because they may be close in the 3D structure of DNA. Tang studies this and other noncoding mutations.
  - **Conservative** (amino acid): An amino acid in a protein, the identity of which is critical to the form and/or function of the full protein.
  - The causes of DNA mutations in cells.

<sup>2</sup>Recall the discussion of this in *The Knot Book*!

- Natural mismatching and tautomerization — know this!
  - Natural mismatching-induced mutation:  $10^{-9}$ - $10^{-10}$ .
  - Tautomerization (< 0.01% frequency): Mutation if the rare tautomer is paired during DNA replication.
- Deamination of exocyclic amines<sup>[3]</sup> (C, A, G) — know this!
  - Adenine to hypoxanthine.
  - Guanine to xanthine.
  - Cytosine to uracil (500 times per day per genome, which is a significant amount). Hence T in DNA but U in RNA.
  - Cells that are uracil N-glycosylase deficient ( $ung^-$ ) show a higher rate of transitions.
  - Note: Deamination of A is more common in single-stranded DNA, but deamination of C is exponentially more common in double-stranded DNA.
- Depurination (A, G).
  - Protonation of purines can lead to cleavage of the glycosyl bond (creating an abasic site).
  - Abasic sites undergo a retro-Michael-like reaction leading to a phosphodiester bond cleavage.  $t_{1/2} \approx 400$  h at  $37^\circ\text{C}$ , pH = 7.
  - Mammalian cells can lose as many as 10,000 purines per cell per generation ( $k = 3 \times 10^{-11}$  per second at  $37^\circ\text{C}$ , pH = 7).
  - Depyrimidination occurs 20 times slower.
- Oxidants, radicals, radiations.
- Chemicals: Alkylating reagents, nucleophiles, crosslinking reagents, and intercalating reagents.
- The reactions of OChem III are not the focus of this course! Tang may occasionally reference such content, but it will be minor and not (directly) tested.
- It may seem like our DNA lives a hard life, but in practice, our genome is very stable.
- Four strategies of DNA repair in cells.

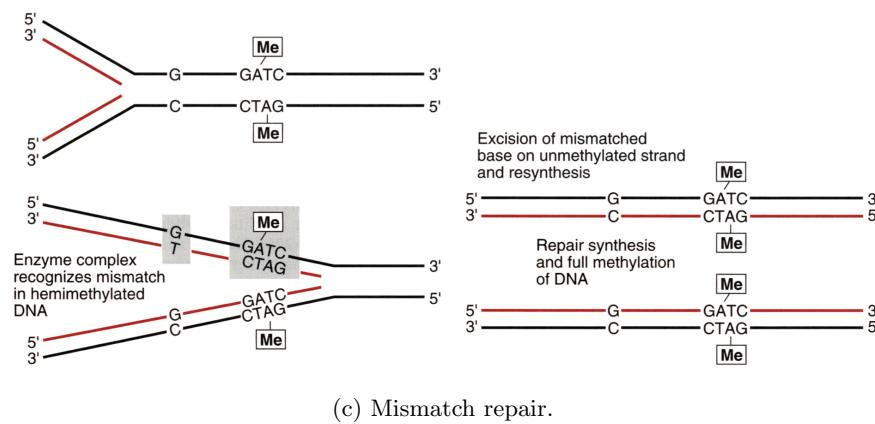


(a) Base excision repair.



(b) Nucleotide excision repair.

<sup>3</sup>Literally: Getting rid of the amine group which lies outside the ring and replacing it with a carbonyl group.



(c) Mismatch repair.

Figure 2.2: DNA repair strategies.

- Direct reversal/repair (DR): Enzymes catalyze the reverse reaction; no removal or replacement of the base is needed.
  - The detailed mechanism of DNA phytolyases is not testable material.
- Base excision repair (BER): DNA glycosylases cleave the ribose-base bond to produce apurinic/apyrimidinic (AP) sites.
  - See Figure 2.2a.
  - Procedure:
    1. DNA glycosylases hydrolyze the N-glycosyl bond of damaged bases.
    2. This creates an “AP” site.
    3. AP endonucleases recognize the AP site and hydrolyze the phosphodiester 5' or 3' of each AP site (mostly 5').
    4. Exonucleases remove the backbone at free ends.
    5. DNA polymerase and ligase fill in and seal the gap.
  - Most DNA glycosylases recognize a specific damaged base.
  - In general, < 30 kD monomeric proteins.
  - No requirement for cofactors.
- Nucleotide excision repair (NER): Enzymes remove a segment of DNA including the lesion and several nucleotides on either side.
  - See Figure 2.2b.
  - Create nicks at two sites. Remove the DNA with lesion. Fill the gap with polymerase. Ligation via ligase.
- Mismatch repair (MR): A subset of BER and NER systems that can discriminate an improper base among two normal nucleotides forming a non W-C pair.
  - Most repair mechanisms are good for recognizing obvious abnormalities (e.g., uracil in DNA, paired pyrimidines, etc.). MR deals with cases when you have two pairs that don't match and it's not immediately clear which is the error.
  - Origin of mismatched (non-W-C) natural DNA base pairs:
    - DNA polymerase errors:  $10^{-4}$  (intrinsic)  $\times 10^{-3}$  (proofreading) =  $10^{-7}$  per base per generation.
    - Heteroduplex DNA arising from homologous recombination.
    - Deamination of 5-Me-C to T (forming G:T pairs).
  - The challenge in repairing a mismatch is distinguishing the “incorrect” base among two natural bases.

- Methyl-directed MR.
  - See Figure 2.2c.
  - *E. coli* methylates N<sup>6</sup> of A in GATC (“dam” methylation).
  - Methylation lags behind DNA replication (which always makes non-methylated DNA).
  - 1976: B. Wagner and M. Meselson hypothesized that the lack of methylation in a newly synthesized strand allows strand discrimination during mismatch correction.
- Experimental support:
  - No PCR, none of today’s routine bio experiments were available in the 1980s.
  - The key experiments: Introduce into cells hemimethylated heteroduplex DNA and allow mismatch repair to take place.
    - This occurred even if the nearest methylation site was > 1000 bp from the mismatch!
    - Neither strand methylated: Correction of either strand.
    - One strand methylated: Correction of unmethylated strand.
    - Both strands methylated: Slow correction of either strand.
- Current MR model (not testable):
  - MutS binds the mismatch or frameshift loop.
  - MutS/L/H complex brings the mismatch and GATC together.
  - MutH nicks the nonmethylated strand 5’ of the GATC.
  - ExoVII or RecJ degrades 5'-3' from GATC to the mismatch or ExoI degrades 3'-5' from GATC to the mismatch.
- No translation today; will be next time. The next lecture will have less content.

## 2.2 Chemical Modifications of DNA and RNA Bases

- 10/6:
- On this year’s Nobel prize in chemistry.
    - Awarded for the **click reaction**.
    - This is the *key* reaction that biologists use today in their research. If you asked any chemical biologist who should have gotten this year’s Nobel prize, Carolyn Bertozzi would have been on their short list.
  - **Click reaction:** A reaction between an azide and an alkyne that may or may not need to be accelerated by a copper iodide catalyst (depending on how strained the alkyne is).

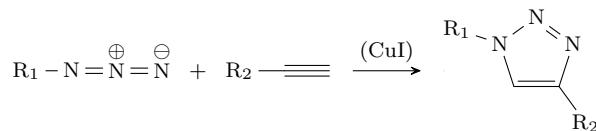


Figure 2.3: Click reaction.

- Useful in biology because it’s *orthogonal* to the entire biological system.
- Two keys:
  1. It will occur in aqueous solution (unlike many organic reactions we learn about).
  2. It will not impact anything else in the biological system.
- Thus, you can use it to manipulate a specified thing in your system.
- Tang cut out bioorthogonal chemistry from this year’s syllabus but reconsidered the day before the Nobel was awarded; we will now have a lecture on it.

- Tang brags about predicting Nobel prizes.
- K. Barry Sharpless is the second chemist ever to receive two Nobel prizes. The faculty at UChicago used to debate whether or not he was worth a second Nobel; Tang bet he was, and now he is.
- On the style of this class.
  - Not as much of an attention to detail; instead of using 3 quarters to cover 1 book, we're building a bookcase. Every lecture is about a different topic, each of which has books written on it.
  - In terms of testing, Tang will not test a tiny thing on her slides; it's more about the concept.
  - She wants us to know that such research exists so that we know to read more about it if we ever need to in our research.
- **DNA transcription:** A process that passes genetic information from DNA to RNA.
  - RNA is synthesized by RNA polymerases using DNA as a template and NTPs as building blocks (instead of dNTPs).
  - RNA polymerases do not require a primer (vs. DNA polymerase).
  - RNA polymerases also elongate RNA in the 5'-3' direction.
  - RNA polymerases lack proofreading mechanisms (vs. DNA polymerases).
    - Synthesis is fairly accurate, a mistake here will likely not be repeated, and mRNA doesn't really matter.
  - Defining the template and nontemplate strand: The template strand does all of the heavy lifting/is directly involved in synthesis. The nontemplate/coding strand is what's replicated (i.e., what gets "all the publicity").
- RNA synthesis begins at promoters.
  - Similar to DNA synthesis, initiation is what's most controlled. The speed of transcription is determined by how strong the promoter, i.e., how strongly RNA polymerases are attracted
  - RNA polymerase binds to specific sequences in DNA (promoters), which direct the transcription of adjacent segments of DNA (genes).
  - Consensus sequences in promoters: Affect the efficiency of RNA polymerase binding and transcription initiation.
  - Promoter sequence establishes a basal level of expression that can vary greatly from one *E. coli* gene to the next.
  - This bacterial example is completely different from how eukaryotes operate.
- Ribosome binding site (prokaryotes)/Kozak sequence (eukaryotes) is a **promotor** at the start of RNA that binds it to the ribosomer. There's also a **terminator** site.
- Tang is skipping the historical exploration of translation.
- Translation.
- Nucleic acid catalysis.
  - Some RNA can actually catalyze chemical reactions.
  - Ribozymes were a hot topic in the 80s and 90s.
  - Why study ribozimes?
    - Implications for the origin of life: Prebiotic soub to RNA to proteins to simple life (*links amplifiable information to function*).
    - RNA and proteins were a chicken-and-egg problem; this discovery suggests that RNA came first.

- We haven't found natural examples of nucleic acid catalysis yet; all known examples were developed in the lab.
  - Tang suggests there may be some examples in basic forms of life.
- How do nucleic acids catalyze reactions compared with proteins?
  - All known all-RNA catalysts in nature accelerate phosphoryl transfer reactions (forming or breaking phosphodiester bonds).
- More info in slides.
- **Ribozyme:** Catalytic RNA; short for ribonucleic acid enzyme.
- **Aptamer:** A receptor; the analogous function in proteins is antibodies (binding but not causing a reaction).
- *Tetrahymena* Catalytic RNA.

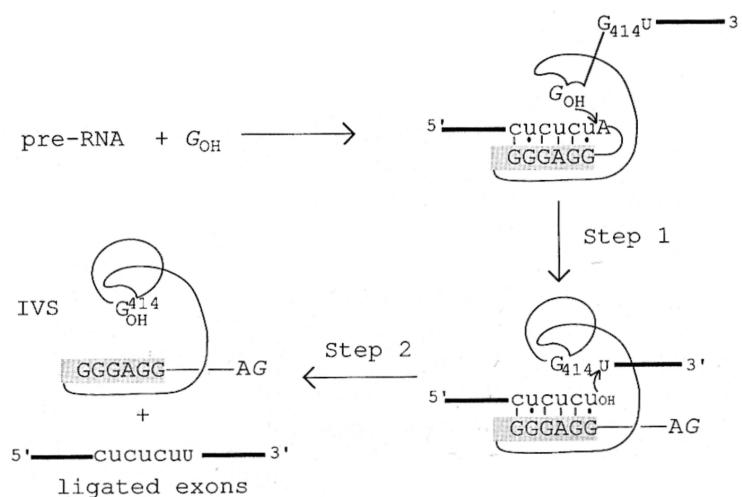


Figure 2.4: Catalytic RNA.

- Discovered 1981; Nobel Prize 1989.
- *Tetrahymena thermophila* is a highly heat resistant form of life. Their pre-ribosomal RNA (pre-rRNA) can catalyze its own splicing, yielding an intervening sequence (IVS) and a mature rRNA via two phospho-transesterifications.
- More details on the mechanism:
  - Catalyzed by GTP (or GMP, but slower this way).
  - The RNA folds, engages GTP. This enables GTP to insert itself, cleaving the RNA. Now there is a sequence that is only attached to the rest of the RNA by hydrogen bonding.
  - The partially free RNA piece adds more of the original RNA to itself and finally detaches.
- Tom Cech (discoverer) believed that protein was catalyzing the reaction.
  - Tried denaturing enzymes and heat, but the reaction still proceeded. Strongly suggested it wasn't a protein, but couldn't confirm it wasn't a heat-resistant or otherwise very stable protein (or trace amounts of a highly efficient protein catalyst).
  - Final clue: Went to an *in vitro* system. Did *in vitro* transcription to get the RNA and RNA only (the material is no longer distilled from the organism), dumped that into the reaction, added the substrate, and watched it occur.
  - Took them a year.

- *Tetrahymena* Ribozyme Structure.
  - The X-ray crystal structure was solved to a decent resolution (2.8 Å — atomic resolution) for one domain of the catalytic core, and to 5 Å for the entire core.
    - Nowadays, you will not be able to publish resolution as low as 5 Å.
    - This was Jennifer Doudna's first paper as an independent researcher!
  - Combination of structural and biochemical studies suggests a mechanism mediated by several bound magnesium cations.
    - Emphasizes that nucleic acid reactions often need to be catalyzed by ions.
  - Take home lesson: RNA can fold into compact, protein-like structures.
- Structure of (most of) the ribosome.
  - Nobel prize (2009) — many scientists tried and failed for years, but they finally got it in 2009.
  - The ribosome is the cell's way of converting genetic information into molecular structure and chemical function.
  - Bacterial ribosome: Huge — 2.6 million Da (far bigger than glycosylase), 2/3 RNA (3 total), 1/3 protein (55 total), two subunits (50S and 30S).
    - You can delete some of the proteins and it will still function, but you cannot delete any of the RNA.
- Video that Tang saw as a grad student that really impressed her and she wants to share with us ([link](#)).
- The ribosome is a ribozyme.
  - The reaction that the ribosome catalyzes is carried out almost entirely by RNA (that's what the ribosome active site is made of).
  - Details of the protein building reaction.
    - There are three sites in the ribosome: The exit, peptidyl, and aminoacyl site. The tRNA comes in at the A site and exits at the P site. At the E site, the tRNA has already been utilized.
    - The incoming amine group does a nucleophilic attack on the tRNA-bound ester group of the amino acid added just before.
    - This kicks out the ester-bound tRNA, and everything shifts down a site.
    - A new codon is exposed at the aminoacyl site, and a new tRNA plus amino acid binds to it.
    - This process repeats over and over again, 3 RNA at a time, building a longer and longer peptide chain.
  - A transition state mimic that scientists use to get the crystal structure of the active state of the ribosome is CCdAp-Puromycin.
    - Puromycin is a useful antibiotic used to inhibit protein synthesis.
    - Puromycin doesn't break, so we can make the ribosome get stuck in the transition state.
  - Tang goes over the electron pushing of the protein building reaction, as seen in Figure ??.
- Mechanism of the ribosome.
  - Recall from lecture 2 that at physiological pH, no nucleobases are charged. You have to go to pH ≈ 3 for protonation or pH ≈ 10 for deprotonation. This implies that nucleobases are really terrible acid-base catalysts.
  - Yet we do have a proton transfer occurring from the incoming amino acid to the exiting tRNA.
  - The XPS crystal structure implies that A2486 catalyzes the proton-transfer reaction.
  - N1 is the site on adenine that can most easily be protonated, but N3 is closest to the carbonyl O and the incoming amine, so it is active as the catalyst.

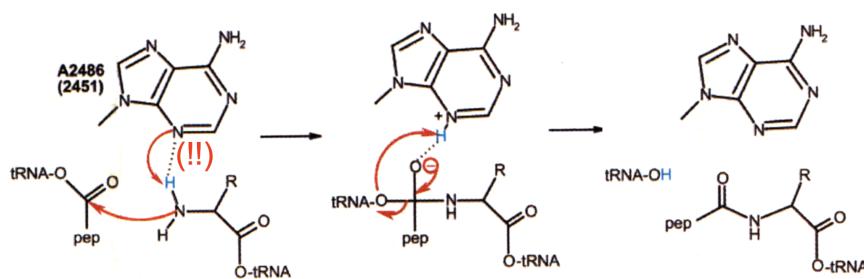


Figure 2.5: Mechanism of the ribosome.

- But N3 has  $pK_a \approx 1.5$ , implying that if you have an isolated adenine, you need to drop the pH to 1 before you can get protonation.
- How does this work? We have a complex hydrogen-bonding network that delocalizes a negative phosphate charge to a neighboring guanine that, in turn, passes it to A2486 to aid in its deprotonation effort. The effect is that actual  $pK_a$  of AdeN3 is 7.6, up six orders of magnitude due to the H-bonding interactions.
- If tested on this, we'll be given information on the hydrogen bonding network. We likely won't be tested on it, though.
- Slides on RNA ligase will not be tested: It's one of Tang's favorite topics, but it falls under directed evolution.
- Now for lecture 4 content.
- Will modification always change DNA or RNA always affect W-C interactions?
  - The most frequent modification (5-methylation) does not change base pairing, but it does affect interaction with proteins.
  - Today: Modifications that show up naturally but don't necessarily cause modifications.
  - This lecture will be shorter: Tang shoots to finish today.
- Overview.
  - Diverse natural base modifications in DNA and RNA and their biological functions.
  - Epigenetics: All cells have the same set of DNA, but different cells can behave very differently. This is caused by epigenetics.
  - Epitranscriptomics describes modifications on mRNA. One example:  $N^6$ -methyladenosine ( $m^6A$ ).
  - If time allows: The arms race of base modifications in bacteria and phages.
- Natural DNA base modifications.
  - Most installed by enzymes (sometimes, phages directly use modified dNTP to synthesize their genome, but that's beyond the scope of this class; for now, most means always).
  - Not always required for survival, but can lead to an evolutionary advantage (recall from last time the example of mismatch correction based on methylation; bacteria that can't do this have a much higher mutation rate).
  - Modifications occur at specific locations on the four canonical bases.
    - Adenine: C2 and  $N^6$ .
    - Cytosine: C5 and  $N^4$ .
    - Guanine: N7.
    - Thymine: C5.

- Exceptions exist, but we won't discuss them.
  - 1.5% of our genome (5% of cytosine) is 5-methylcytosine.
  - 5-(hydroxymethyl)cytosine, 5-formylcytosine, and 5-carboxycytosine are also possible in humans, in decreasing frequency.
  - Other stranger modifications (such as bonding sugars to C5) can occur in lower organisms.
  - Uracil can also be hydroxymethylated and formylated. Base J is uracil with a sugar at C5.
  - $N^6$ -methyladenine is very abundant in bacteria, but there is a huge controversy over whether or not it is in humans.
  - We don't need to memorize any of these save 5-methylcytosine.
- Detecting DNA modifications: Use LC-MS/MS.
  - Harvest the DNA, digest it into individual nucleotides, get rid of the phosphate, shoot it into a mass spectrometer, and see if you can detect the modified base.
  - Restriction: Rare modifications can fall below the detection limit.
  - Point of controversy: The second and third steps above are accomplished using enzymes from prokaryotes, but these can leach bacterial DNA nucleotides.
    - Errors regarding this can account for some of the false positive detections of  $N^6$ -methyladenine in eukaryotic DNA.
  - We will see better methods of sequencing later.
- Epigenetics and DNA methylation.
  - Epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence. Two big areas: Modification of DNA and modification of histone proteins.
  - 5mC is the “5th richest” base in human DNA. Happens primarily within CpG islands in promoters. Most often correlated with gene suppression.
- DNA methylation is dynamic.

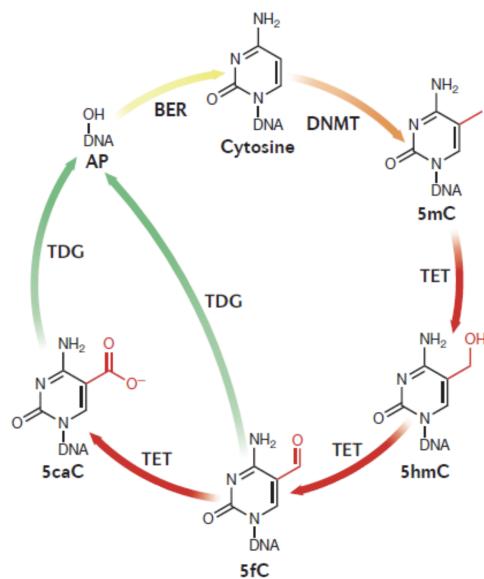


Figure 2.6: Active methylation cycle of cytosine.

- Two ways to get rid of modified bases:

- The passive way, i.e. if we never replace the methylation marks after DNA replication (recall from the discussion associated with Figure 2.2c that newly synthesized DNA is unaltered). If modifications are never regenerated, they will slowly become less and less common, only occurring in the original strand that has now been duplicated and “diluted” many times.
- Active demethylation: Catalyzed by the TET family in eukaryotes.
- 5mC and 5hmC have different effects on gene transcription (5hmC is more activating than repressing).
- 5mC and 5hmC are viewed as modifications; 5fC and 5caC are viewed as lesions and will be fixed. We will not be tested on this, though.
- Histone marks can have very different functions (acetylation vs. methylation).
- Epigenetics is a huge research field and waiting for a Nobel prize.
- 5mC/5hmC in early embryonic development.
  - Two scenarios: Skin/liver cells are still dividing but are at their terminal epigenetic state; a fertilized egg is still diversifying. The fertilized egg has more motivation to change its epigenetics.
  - Thus, throughout development, you see a quick decrease in 5mC and some waves in 5hmC.
- DNA methylation on aging and cancers.
  - Not tested, but interesting.
  - DNA methylation maps change with chronological age. When you are born, you have the most beautiful epigenetics; it gets messed up as you age.
- How to detect 5mC/5hmC sites in DNA?

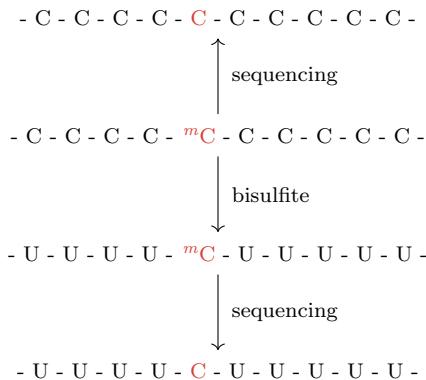


Figure 2.7: Bisulfite chemistry.

- Bisulfite chemistry.
  - One of Tang’s favorite topics in biochemistry; will definitely be tested.
  - When you mix DNA with bisulfite and heat it up, cytosine converts to uracil. Since uracil pairs with thymine, you will detect thymine when you should detect guanine if you do the bisulfite treatment.
  - If you have 5mC, bisulfite won’t attack for steric reasons, so 5mC remains unchanged.
  - Thus, between the original strand and the bisulfited strand, you get differentiation.
  - Whichever cytosines don’t change before and after bisulfiting are your methylated C’s.
  - Notice how in Figure 2.7, the only cytosine which doesn’t change in between the two rounds of sequencing is the methylated one, indicated in bright red.

- Assuming the reaction yield of bisulfite chemistry is 100%. What if the yield is 50%? We are lucky here: Bisulfite chemistry is 99.9% efficient, so the number of false positives is very low.
- If you want to detect beyond 5mC, there are more complex methods; she won't test us on these though.
- Natural RNA base modifications.

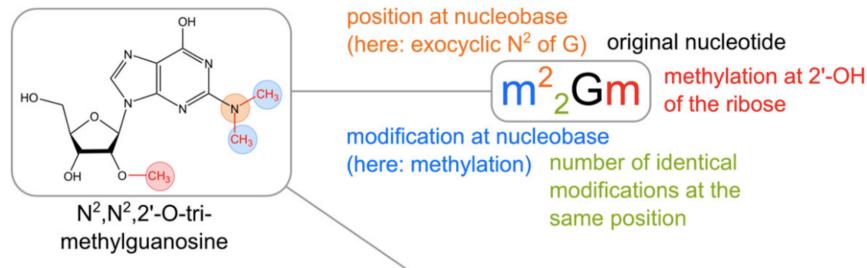


Figure 2.8: Naming convention for RNA base modifications.

- If you compare between DNA and RNA base modifications, you will find DNA boring.
- There are 20 known DNA base modifications; there are over 150 known RNA ones, and many look very weird.
- Naming convention for base modifications: See Figure 2.8.
- RNA base modifications occur in all three major RNA species (tRNA, mRNA, and rRNA) and in other RNA species such as snRNA and miRNA.
- They are found in all three domains (archaea, bacteria, and eukarya). Some modifications are unique to a single domain.
- tRNA is heavily modified.
  - > 75% of RNA modifications are present in tRNA.
  - tRNA are typically less than 100 nucleotides long, so the density of modifications is very high.
  - These modifications enhance translation.
- **Transcriptome:** The set of all RNA within a cell.
- **Epitranscriptome:** The set of all biochemical modifications to all RNA.
- Epitranscriptome and mRNA methylation.
  - Epitranscriptomics defines the half-life of RNA and determines how strongly mRNA gets translated.
  - In many papers about base modifications, it's evident that the figures are not drawn by chemists (there are obvious mistakes such as missing charges, wrong atoms, etc.).

# 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<sup>6</sup>A, 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<sup>5</sup>C) and pseudouridine ( $\psi$ ) 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  $\psi$  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<sup>6</sup>A, m<sup>5</sup>C, and  $\psi$ , 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 → 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<sup>7</sup> 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.

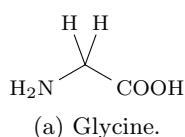


Figure 3.1: Achiral amino acid.

- Glycine.
  - Flexible since it's unsubstituted on its  $\alpha$ -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.

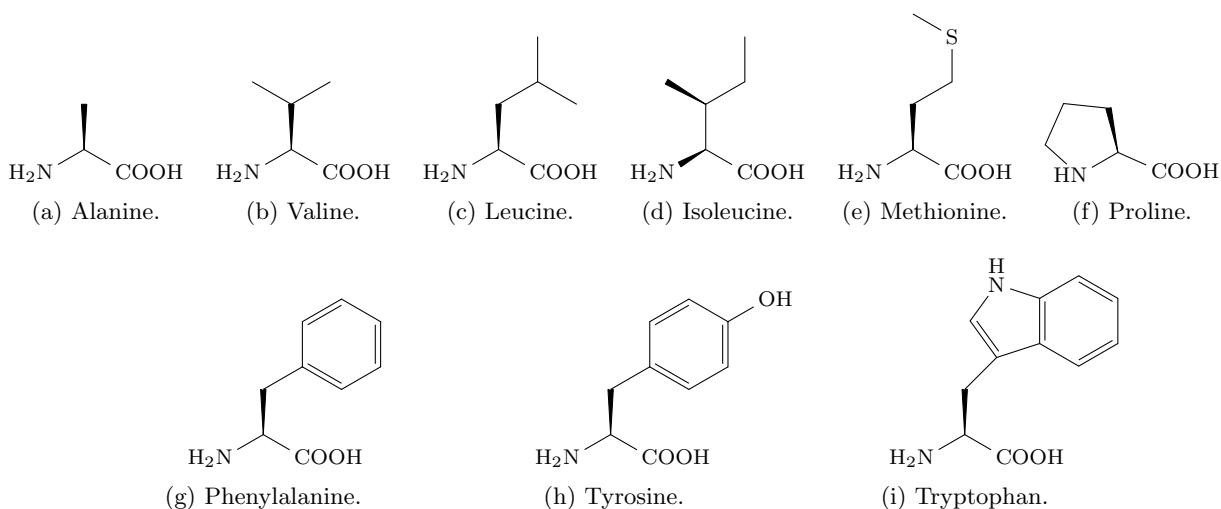


Figure 3.2: Hydrophobic amino acids.

- 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.
    - There 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  $\alpha$ -helices or  $\beta$ -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 than 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.

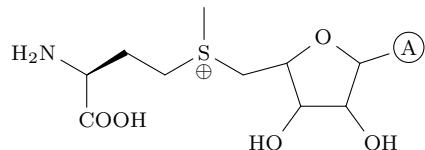


Figure 3.3: (S)-adenosylmethionine.

- 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.

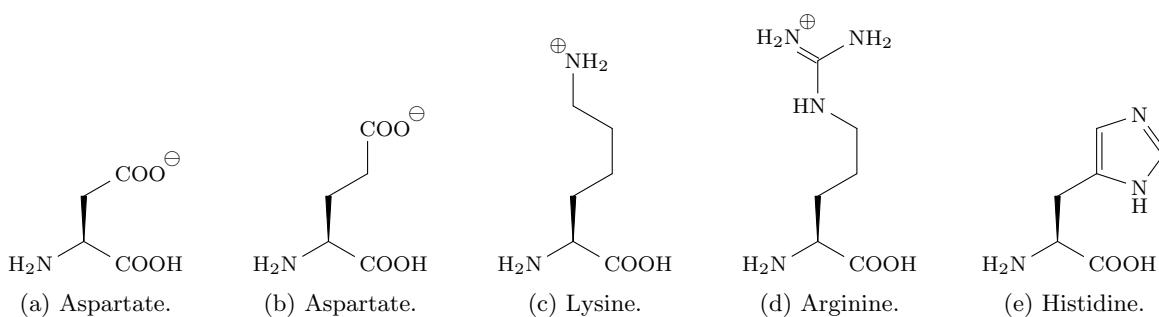


Figure 3.4: Charged amino acids.

- 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.

- Histidine 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.

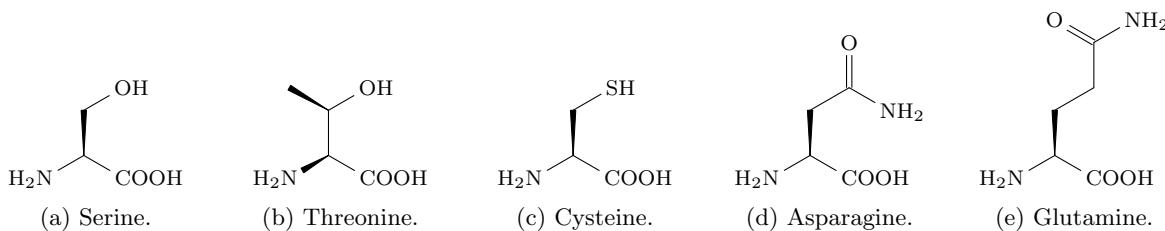


Figure 3.5: Polar amino acids.

- 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 acids).
    - Has selenium instead of sulfur.
    - Name: Selenocysteine, Sec, U.

# References

Nelson, D. L., & Cox, M. (2021). *Lehninger principles of biochemistry* (eighth). W.H. Freeman.