CHEM 23300 (Introduction to Chemical Biology) Notes

Steven Labalme

January 1, 2023

Weeks

List of Figures

List of Tables

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 $\rightarrow$ RNA $\rightarrow$ protein $\rightarrow$ 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.

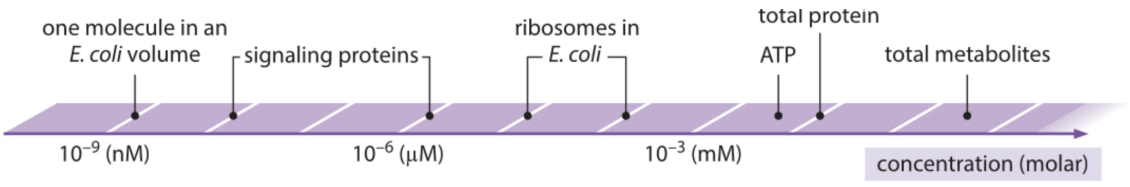


Figure 1.1: Concentrations in biology.

  - 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 τ to traverse distance R given diffusion coefficient D:

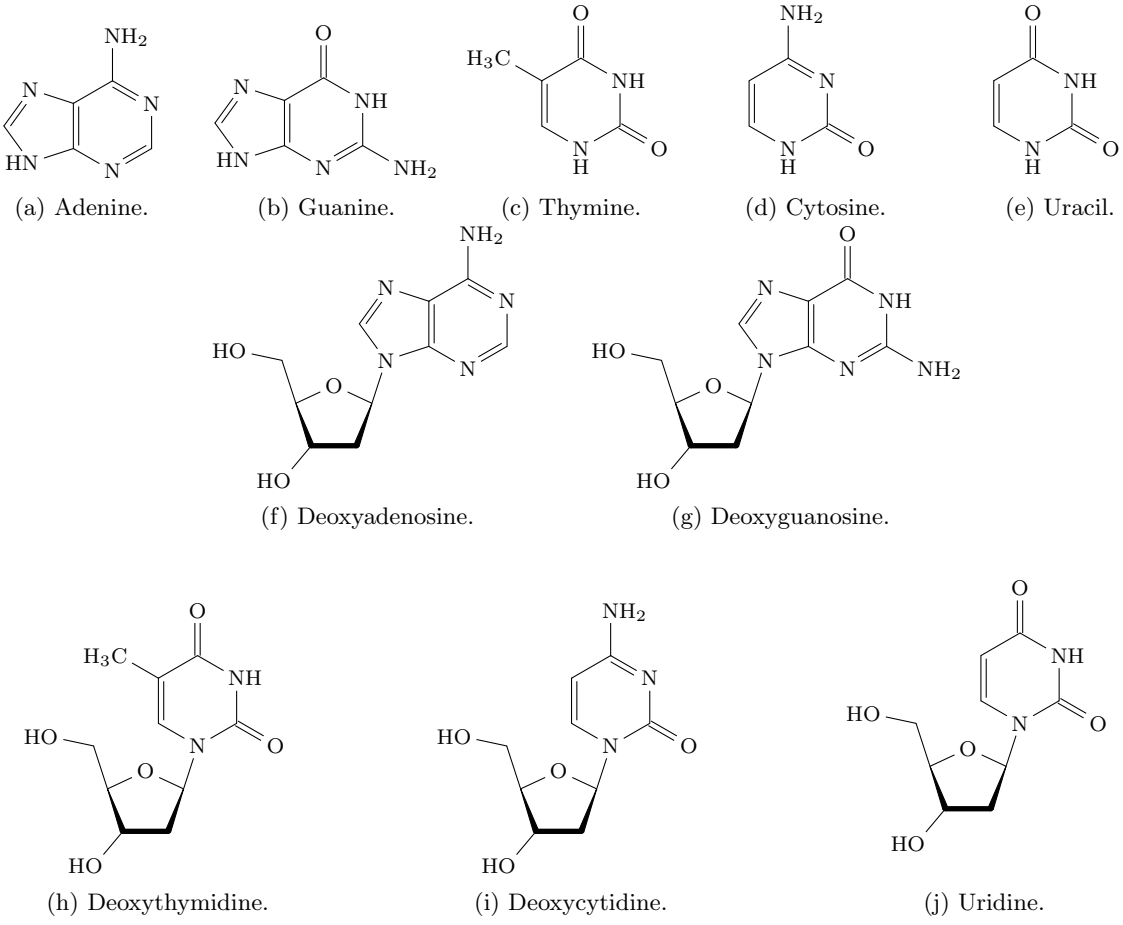
$$\tau = \frac{R^2}{6D}$$

  - For a protein in cytoplasm, $D \approx 10\,\mu m^2/s$.
- 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 3×10^{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 ($\approx 3.3 - 3.4\,\text{Å}$).
  - 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**.



(a) Adenine. (b) Guanine. (c) Thymine. (d) Cytosine. (e) Uracil.

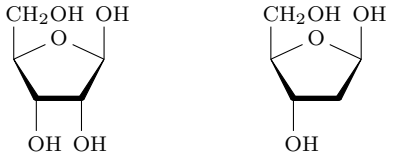
(f) Deoxyadenosine. (g) Deoxyguanosine.

(h) Deoxythymidine. (i) Deoxycytidine. (j) Uridine.

Figure 1.2: Bases and nucleosides.

  - Notice that deoxyribose is joined with the base at its $1'$ carbon.
  - If we use ribose instead of deoxyribose, we get adenosine, guanosine, etc.
  - Memorize these structures!
- Nomenclature.
- **Base**: The heterocycle. *Also known as* **nucleobase**.

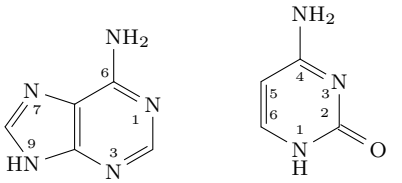
- **Ribose**: A 5-carbon monosaccharide, a derivative of which is a component of DNA.
- **Deoxyribose**: A molecule identical to ribose but without the 2′ hydroxyl group.



(a) Ribose. (b) Deoxyribose.

Figure 1.3: DNA sugars.

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



(a) Adenine. (b) Cytosine.

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*

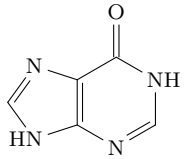
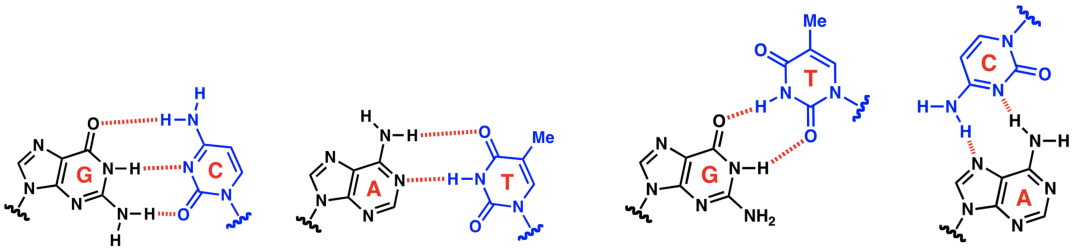


Figure 1.5: Inosine.

- 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.



(a) Watson-Franklin-Crick. (b) Wobble.

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.
  - $MeNH_2$'s protonated form has p$K_a \approx 10.6$.
  - Aniline's protonated form has p$K_a \approx 4.6$ because aniline is a weaker base.
  - Pyridine's protonated form has p$K_a \approx 5$ because it is basic, but it is also sp^2.
  - An amide has p$K_a \approx 18$.
    - 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 p$K_a \approx 16$.

- Predicting DNA ionization states.

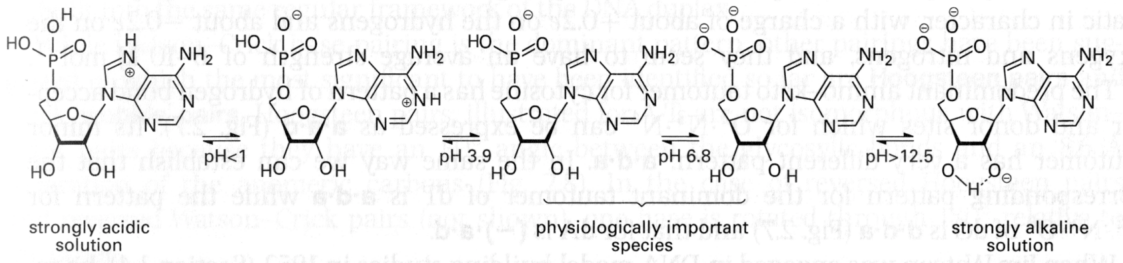


Figure 1.7: DNA ionization states.

  - 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: p$K_a \approx 12.4$ (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.

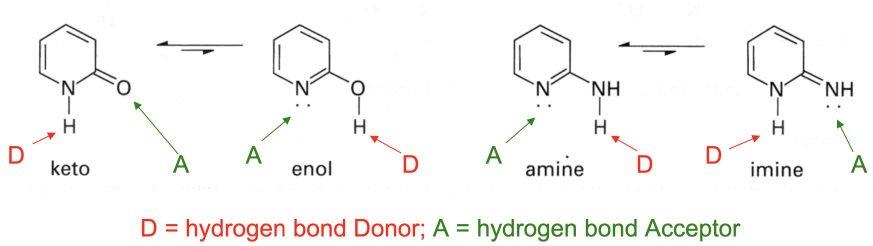


Figure 1.8: Base tautomerization.

  - 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.

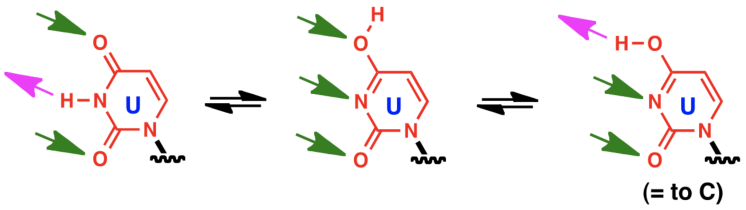
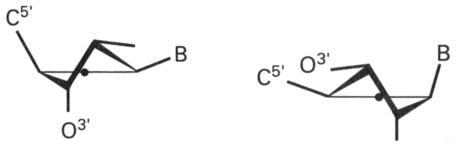


Figure 1.9: Uracil tautomerization.

  - 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.



(a) $C^{2'}$-*endo*. (b) $C^{3'}$-*endo*.

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: $C^{2'}$-*endo* and $C^{3'}$-*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 $C^{2'}$-*endo* sugars.
      - ➢ Since DNA runs 5′ to 3′, as we can see in Figure 1.10a, $C^{2'}$-*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 $C^{3'}$-*endo* sugars.
      - ➢ Conversely, $C^{2'}$-*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 $\rightarrow$ 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 $C^{2'}$-*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 $C^{3'}$-*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.

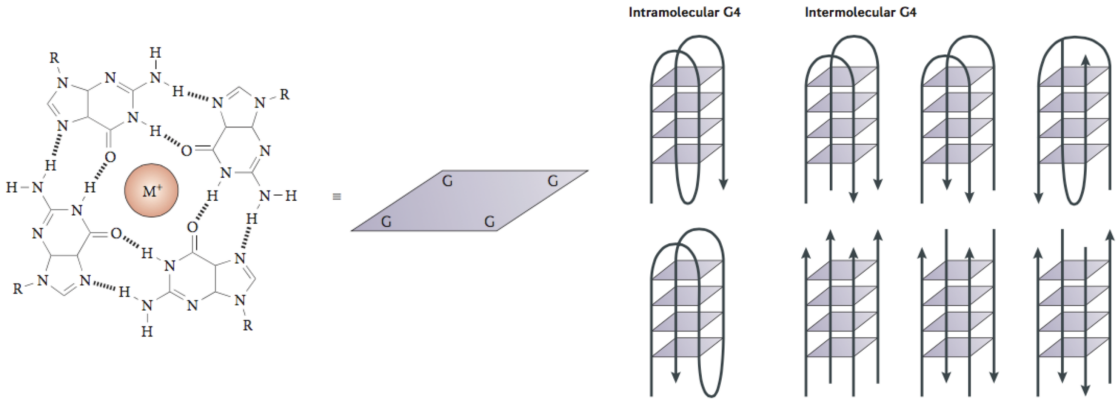


Figure 1.11: G-quadruplex structure.

  - 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 π structure too bendy to flouresce 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 groove (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-Hydroxypyrrole (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; uset 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 (Na^+, 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 grove interactions are sequence specific.
    - Minor grove 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.

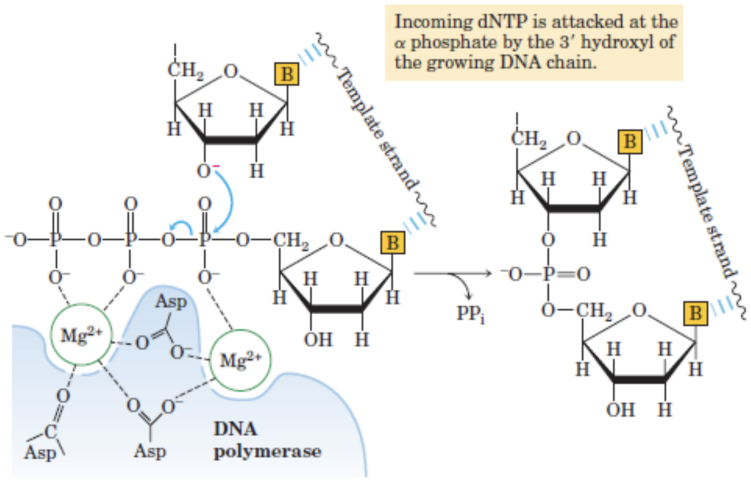


Figure 2.1: Mechanism of DNA synthesis.

  - DNA polymerases require three components: A template, a primer, and dNTPs[1] (N = A, T, G, or C).
  - Mechanism: The 3' hydroxyl of the growing DNA chain attacks the α 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 β and γ phosphates.

[1] 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 β and γ phosphates to stabilize them.
    - The other coordinates with the α 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: Templated synthesis (error rate of 10^{-4}-10^{-5} *in vitro*) and error correction mechanisms.
  - Templated 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 "proofreads" 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 folk when the other folk 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[2].

- 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.

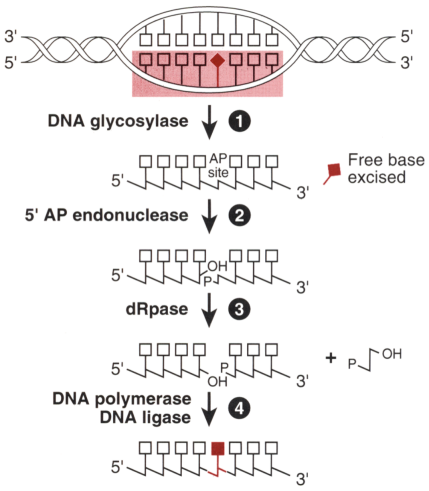
- **Conservative** (amino acid): An amino acid in a protein, the identity of which is critical to the form and/or function of the full protein.

- The causes of DNA mutations in cells.

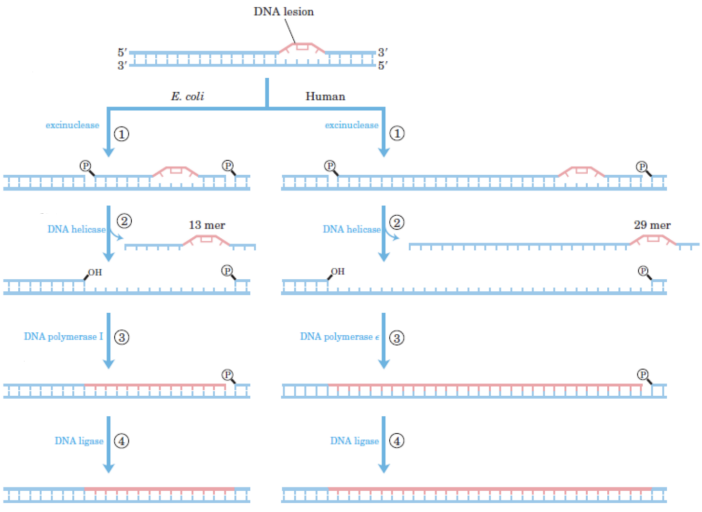
[2] Recall the discussion of this in *The Knot Book*!

  - 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[3] (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. $t_{1/2} \approx 400\,\text{h}$ at $37°\text{C},\ \text{pH} = 7$.
    - Mammalian cells can lose as many as 10,000 purines per cell per generation ($k = 3 \times 10^{-11}$ per second at $37°\text{C},\ \text{pH} = 7$).
    - 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.

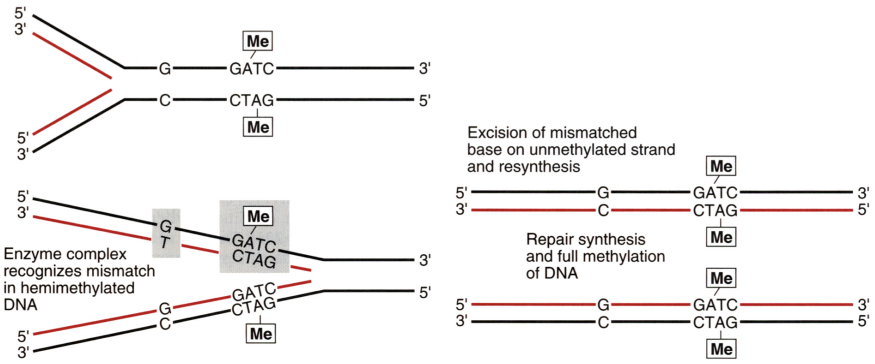


(a) Base excision repair.



(b) Nucleotide excision repair.

[3]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. Chemically modifies what's already there (doesn't replace it with new bp's).
  - 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^6 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 Exol 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).

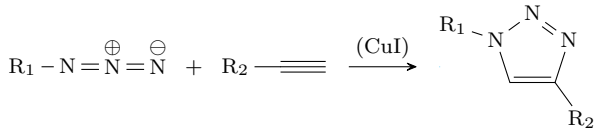


Figure 2.3: Click reaction.

  - 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 attractedl
  - 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 efficiencey 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.

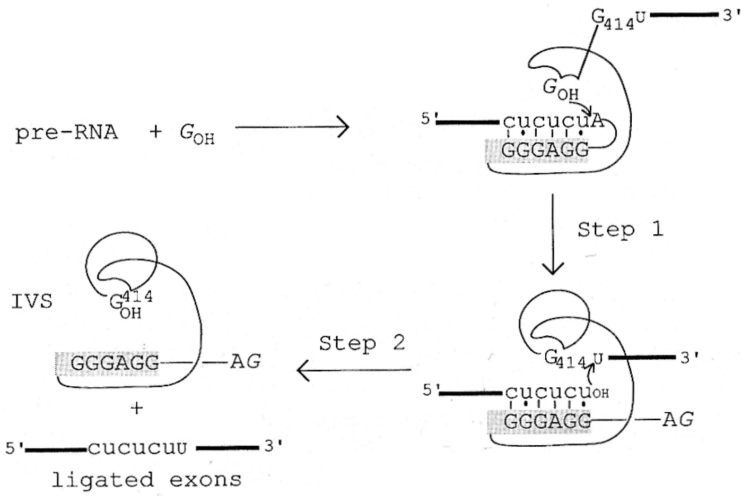


Figure 2.4: Catalytic RNA.

- Discovered 1981; Nobel Prize 1989.
- *Tetrahymena thermophilia* is a highly heat resistant form of life. Their pre-ribosomal RNA (pre-rRNA) can catalyze its own splicing, yielding an interveining 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 pice 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, peptodyl, 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 nucelophilic 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 plust 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 $\approx$ 3 for protonation or pH $\approx$ 10 for deprotonation. This implies that nucleobases are really terrible acid-base catalysts.
  - Yet we do have a proton transfer occuring 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.

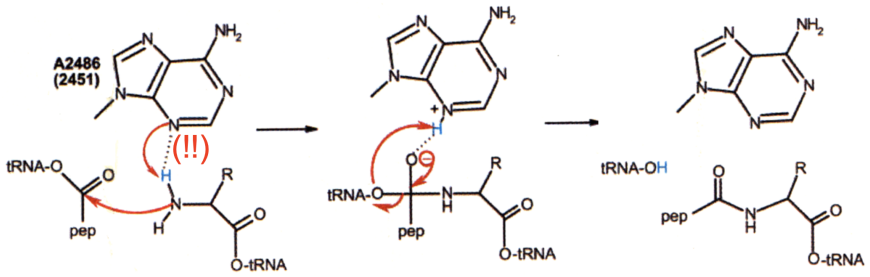


Figure 2.5: Mechanism of the ribosome.

  - 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 cananonical bases.
    - Adenine: C2 and N^6.
    - Cytosine: C5 and N^4.
    - Guanine: N7.
    - Thymine: C^5.

    - 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 contraversy 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 contraversy: The second and third steps above are accomplished using enzymes from prokaryotes, but these can leech 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 beter 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.

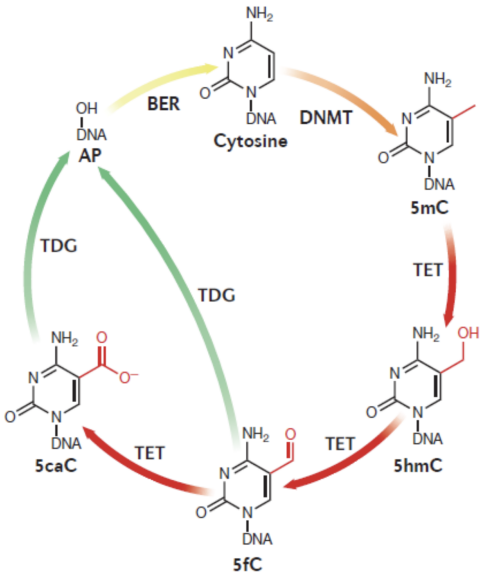


Figure 2.6: Active methylation cycle of cytosine.

  - 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 occuring 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?

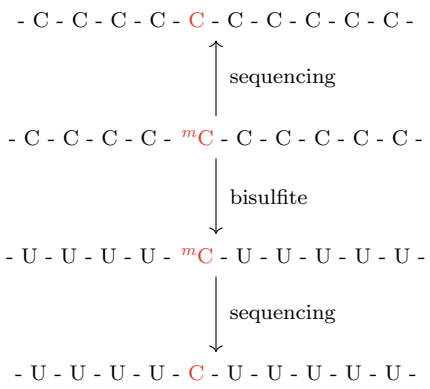


Figure 2.7: Bisulfite chemistry.

  - 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 bisufited 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.

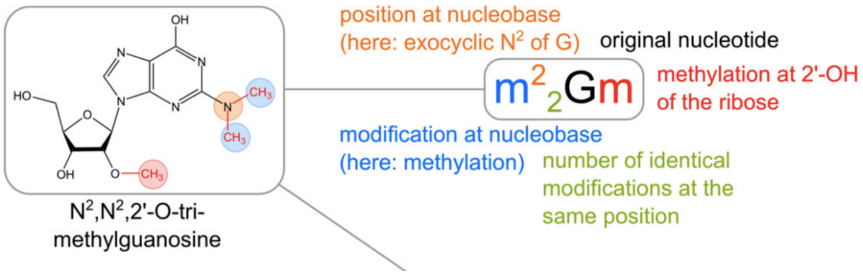


Figure 2.8: Naming convention for RNA base modifications.

  - 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 modificatinos 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^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.

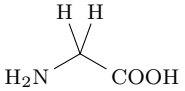
- **Native chemical ligation**: Connecting two peptides with an amide bond.
  - A very hard chemical problem; requires activating the amine of an amino acid.

- 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.

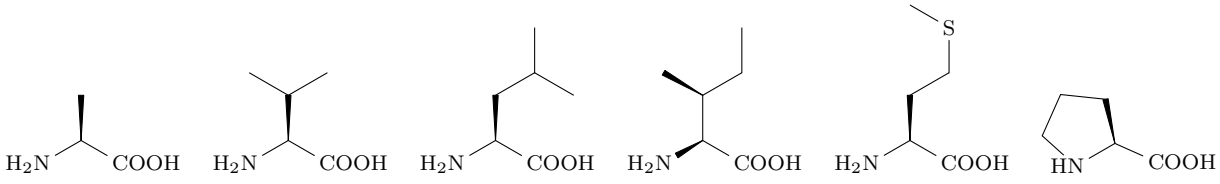


(a) Glycine.

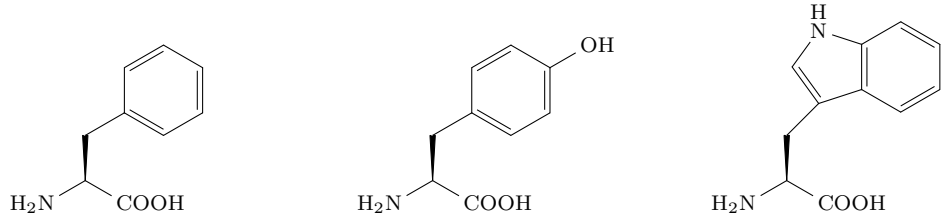
Figure 3.1: Achiral amino acid.

  - 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.

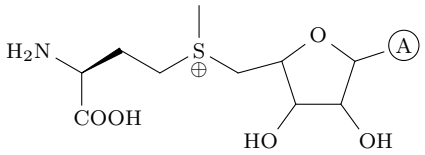
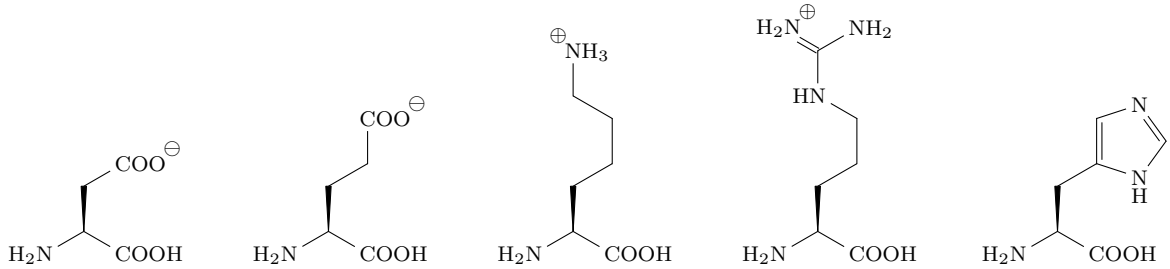


Figure 3.3: (S)-adenosylmethionine.

  - 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.



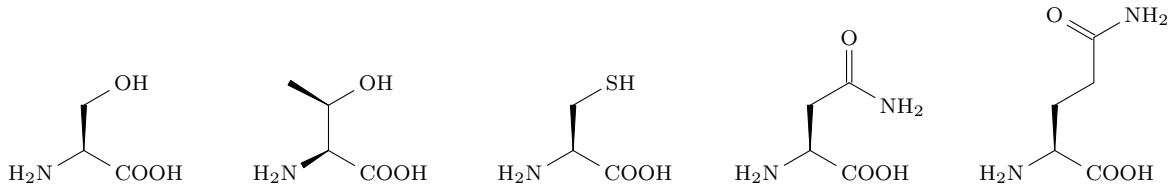
(a) Aspartate. (b) Glutamate. (c) Lysine. (d) Arginine. (e) Histidine.

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.

3.2 Protein Structure and Function

10/13:

- Callie: PSet should take about 3 hours.
- Office hours Monday evening on Zoom.
- Review of amino acids.
  - Tang did her grad work on cysteine.
  - Proline the amino acid is the chiral asymmetric organocatalyst from OChem III!
- Proline biosynthesis:
  - Uses the precursor of glutamate and cyclization happens in the final step.
- Amino acid properties given.
  - Residue mass (minus H_2O because when we add peptides to the chain, we lose water [dehydration synthesis/condensation reaction]).
  - Van der Waals volume (related to mass).
    - Anything else we need to know about this??
  - Frequency in proteins.
    - There are three proteins that have six codons corresponding to them: Leu, Arg, and Ser. L does occur the most frequently, but R and S are farther down the list.
    - Ala has four associated codons and occurs relatively frequently.
    - Met and Trp have one associated codon, each, and are pretty far down the list (W is the least frequent, but His occurs less frequently than M for example).
  - Not testable: W has a relatively unique UV-Vis absorbance at 280 nm (Phe and Tyr can contribute a bit but not much).
    - Another way to quantify the concentration of protein in the lab is to use a **blackfield assay**.
  - Cysteine is rare because it can form disulfide bonds, so we don't want it everywhere.
    - 2 corresponding codons.
- **Blackfield assay**: An assay in which you use a dye that changes color when it interacts with the protein; you then quantitatively measure the change in color.

- pK_a's of side-chain groups.
  - The α-amino and α-carboxyl groups are not super important (they are only present at the ends of the proteins).
  - The negatively charged amino acids have p$K_a \approx 3.9 - 4.0$ (D) and p$K_a \approx 4.3 - 4.5$ (E), making them properly acidic.
  - Arg's guanidinium moiety has p$K_a \approx 12$.
  - Lys's amino moiety has p$K_a \approx 10$ (or a bit larger).
  - Thiols, imidazoles, and phenolic hydroxyls are all in the viscinity of 6-10 as well.
  - pK_a's are highly context dependent.
    - Depends on electrostatic effects, H-bonding effects, and inside a protein (more difficult to ionize a residue here).
    - Recall the context-dependent pH of A2486 in the ribosomal mechanism (see the discussion associated with Figure 2.5).
- Main chain ionization.
  - If pH < 2, everything will be protonated (NH_3^+, COOH).
  - If pH > 9, everything will be deprotonated (NH_2, COO^-).
  - If $2 <$ pH < 9, we will have the zwitterionic form (NH_3^+, COO^-).
- Table of pK_a.
  - pK_{a_1} corresponds to the carboxylic acid, pK_{a_2} corresponds to the amine, and pK_{a_3} corresponds to the side chain (if applicable).
- Aspartic acid pH analysis.
  - As the pH increases, α-carboxylic acid is deprotonated, then the side chain carboxylic acid, then the NH_3.
  - We might have a test question like this.
- Chemical environment affects pK_a values.
  - The α-carboxy group in amino acids is much more acidic than in carboxylic acids because the conjugate base is stabilized significantly by the zwitterionic form.
  - The α-amino group in amino acids is slightly less basic than in amines because the electronegative oxygen atoms of the α-carboxy group act as EWGs.
- Formulation of peptides.
  - Peptides are small condensation products of amino acids.
  - They are "small" compared to proteins ($M_w < 10\,\text{kDa}$).
  - Review of peptide bond formation (see the discussion associated with Figure 2.5).
  - Overlap between "polypeptides" and "proteins."
    - Arbitrarily, people use 10 kDa as a cutoff.
- Peptide ends are not the same.
  - Synthesis occurs from the N terminus to the C terminus (remember, the new amine is attacking the carboxylic acid of the previous).
  - Solid phase peptide synthesis occurs in the reverse direction.
  - You number the amino acids $\text{AA}_1, \ldots, \text{AA}_n$ starting from the N terminus.

- How do such simple building blocks result in diverse functions?
  - Enzymes play the majority of the structure and catalysis functions in our body, e.g., enzymes, receptors, antibodies, hormones, regulatory roles, structural, etc.
  - A protein of 100 amino acids has $20^{100} \approx 10^{130}$ possible sequences.
  - An average 100 amino acid protein has a mass of 14 000 Da; thus, 10^{130} such molecules would have a mass of 1.4×10^{134} Da. For reference, the mass of the universe is about 10^{80} Da.
- Note on directed evolution.
  - Nature has not sampled every AA sequence, but has optimized over some.
  - A reasonable (but still tough) sample size library to achieve is 10^8. Thus, the sampling space for any directed evolution experiment is necessarily limited.
  - Your directed evolution experiment only works because there are multiple answers in the solution space.
- Favorable interactions in proteins.
- **Hydrophobic effect**: Release of water molecules from the structured solvation layer around the molecule as protein folds increase the net entropy.
- **Hydrogen bonds**: Interaction of N–H and C=O of the peptide bond leads to local regular structures such as α-helices and β-pleated sheets.
- **London disperson effect**: Medium-range weak attraction between all atoms contributes significantly to the stability of the interior of the protein.
- **Electrostatic interactions**: Long-range strong interactions between permanently charged groups.
  - Example: Salt bridges, especially those buried in the hydrophobic environment which strongly stabilize the protein.
- 4 levels of protein structure.
  - Primary: Amino acid sequence.
  - Secondary: α-helix or β-pleated sheet.
  - Tertiary: Larger protein chunks, made of a single polypeptide chain.
  - Quaternary: Most proteins are made of multiple such moieties/are assembled subunits.
- Natural protein function beyond proteinogenic amino acids.
  - Proteinogenic amino acids have limited chemical functionality.
  - Natural proteins are especially bad at redox reactions.
  - This functionality is expanded in nature with...
    1. Small molecule cofactors.
       - Enable redox reactions and can serve as an electron sink.
       - Example: NADH, NADPH.
       - Some of these reactions may have come up in OChem III. We can also learn more about them if we take some of Tang's other courses.
    2. Post-translational modification.
       - Example: Phosphorylation.
    3. Post-translational rearrangement.
       - In GFP, for instance, we have a post-translational rearrangement between serine, glycine, and tyrosine that leads to the formation of a chromophore.

- The peptide bond.
  - Carbonyl oxygen and amine hydrogen are *trans*.
  - This is because, drawing resonance structures, we see that the peptide C–N bond has some π character. p-orbital overlap. Thus, we will say it does not rotate meaningfully.
- The rigid peptide plane and the partially free rotations.
  - Rotation around the peptide bond is not permitted.
  - Rotation around bonds connected to the α carbon is permitted.
  - ϕ: Angle around the α-carbon – amide nitrogen bond.
  - ψ: Angle around the α-carbon – carbonyl carbon bond.
  - In a fully extended polypeptide, both $\psi, \phi = 180°$, but this is not common.
  - Even without the side chain, some angles are not permitted due to sterics. Bulkier side chains narrow the allowable range still further.
- The polypeptide is made up of a series of planes linked at the α carbon.
- Distribution of ϕ and ψ dihedral angles.
  - Some ϕ and ψ combinations are very unfavorable due to sterics.
  - Some ϕ and ψ combinations are more favorable because of the chance to form favorable H-bonding interactions along the backbone.
  - We won't be asked to memorize any good or bad ϕ, ψ angles.
- **Ramachandran plot**: A plot showing the distribution of ϕ and ψ dihedral angles that are found in a protein.
  - Gives the distribution of secondary strucures in the $\phi\psi$-plane.
  - Shows the common secondary structure elements and reveals regions with unusual backbone structure.
  - There are characteristic regions for α-helices (lower left) and β-pleated sheets (upper left), but loops are harder to detect (though they may appear to some extent in the upper right).
  - Glycine has density in all four quadrants (it is the most flexible, after all).
- Protein conformational space.
  - ϕ, ψ describe torsional angles: $-180° < (\phi \text{ or } \psi) < 180° = -180°$.
- Secondary structure: α helix.
  - Not testable, but $\phi = -57°$ and $\psi = -47°$.
  - Right-handed helix (just like double stranded DNA).
  - 3.6 residues per turn.
  - 5.4 Å rise per turn.
  - H bonds between $i, i + 4$.
  - Proline is a good helix starter. A, R, K, L, and M are good in an α-helix; P, G, T, and S are poor.
    - P is too rigid, G is too flexible, and T,S have additional H-bonding donors and acceptors that might disrupt the $i, i + 4$ pattern.
- mRNA is more complex and weighs more than a protein.

  - B-DNA is about 28 Å per turn, larger than the α-helix. This makes sense since there are so many moieties involved in DNA (sugar, phosphate, base pair).
  - It's not like it's a small template we're using to synthesize something much larger.
  - Tang explores this theme in the arena of mRNA vs. protein vaccines.
    - For COVID vaccines, we can deliver either the COVID spike protein or the mRNA. The former is a smaller thing to deliver.

- Secondary Structure: Antiparallel β-sheet.
  - Adjacent strands run in opposite directions.
  - Hydrogen bonds are neatly stacked one carbons apart, always.

- Secondary Structure: Parallel β-sheet.
  - Adjacent strands run in the same direction.
  - Hydrogen bonds are not neatly stacked. They are bent and evenly spaced, though.

- Loops connect secondary structure.
  - Usually rich in polar residues.
  - Loops are irregular structures.
  - H-bond with solvent.
  - Gly = common start or end.
  - Often contain binding sites or enzyme active sites.
    - Loops are more flexible, so they can test out more conformations.
    - The lock and key model is misleading — it's not a rigid interaction, but rather the protein adjusts when the substrate binds.

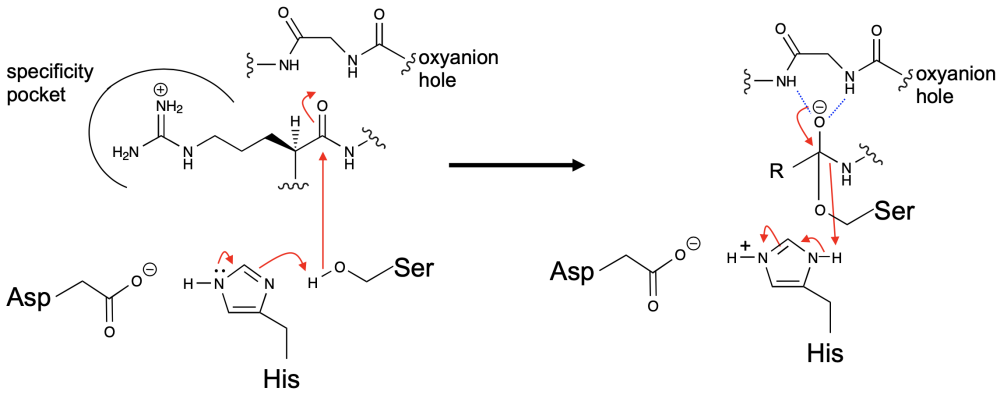
- Higher order protein structures.
  - α-helix, loop, β-strand $\rightarrow$ motif $\rightarrow$ domain $\rightarrow$ protein.

- **Tertiary structure**: Overall 3D structure of the protein; describes how the peptide chains fold and pack.
  - Covalent structures (peptide, disulfide bonds).
  - Hydrogen bonds, hydrophobic interactions, electrostatic interactions, and van der Waals interactions.
  - Recall that disulfide bonds bring distal ends of a protein together.

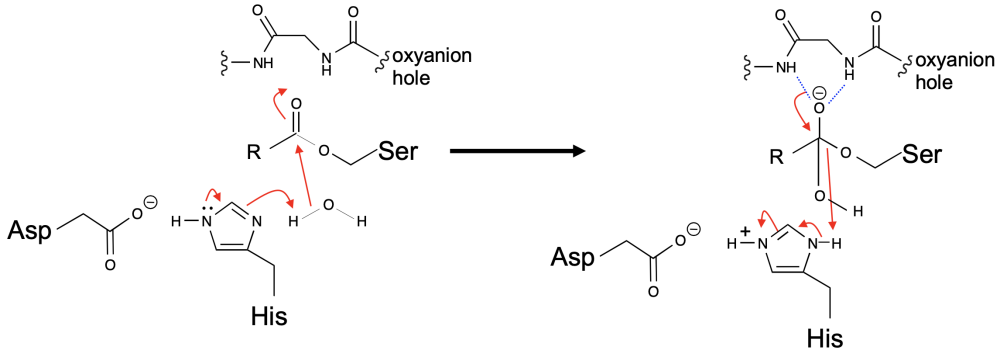
- Quaternary structures are formed of multiple tertiary motifs.

- **Enzyme catalysis**: Enzymes are protein catalysts of biologically relevant chemical reactions and typically display a high efficieency and selectivity.
  - Part of **mechanistic enzymology** (which Tang used to cover).
  - Enzymes can accelerate some reactions by 17 orders of magnitude.
  - Enzymes are typically highly selective.
  - Enzymes function either by stabilizing the transition state of a reaction more than they stabilize the ground state of the substrate or by providing an alternative reaction pathway (mechanism) that involves a lower activation barrier.

- One example of enzymatic catalysis: Protease (also called a peptidase or proteinase) is an enzyme that catalyzes the hydrolysis of a protein amide bond.



(a) Step 1.



(b) Step 2.

Figure 3.6: Serine protease.

- There is either 1-step or 2-step protease catalysis.
  - In 1-step, an acidic side chain (e.g., D, E, or a metal ion [perhaps bound to N or Q]) activates a water molecule to attack the peptide bond.
  - In 2-step, a nucleophilic side chain (e.g., C, S, or T) splits the peptide bond and then water comes in to break off the side chain from the carbonyl derivative.
- Serine protease is a classic example of a protein participating in catalysis covalently; a good example in protein engineering, where it can be developed into a ligase.
- Serine protease has a catalytic triad within the active site.
  - Histidine (good for its proton shuffle ability) deprotonates serine protease.
  - The now positively charged histidine is stabilized by the negatively charged aspartate (though protonation/deprotonation does not occur).
  - Now that serine has been deprotonated, it is an excellent nucleophile and can attack the peptide amide via nucleophilic acyl substitution. The departing amine grabs an extra proton back from histadine.
  - In step 2, we have another histadine-involved deprotonation to start, but this time of water.
  - The lone hydroxyl group then attacks the serine-bonded amino acid's carbonyl group again, doing a transesterification nucleophilic acyl substitution, and serine regenerates itself from the histadine proton.
- Note that most serine proteases come with a specificity pocket that recognizes one specific type of amino acid at which we want to cleave the peptide bond.

  - Serine proteases are quite important.
  - Penicillin (the antibiotic) works by inhibiting a kind of serine protease in bacteria (a transpeptidase used to build the cell wall) that's not found in humans. This makes it so that bacteria cannot construct their cell walls.

- Lecture summary.
  - 20 natural amino acids offer different features.
  - Post-translational modifications and rearrangements expand the functional repertoire of proteins.
  - Protein structure is determined by the nature of the amide bond and by at least four noncovalent interactions (and disulfide bonds).
  - Natural protein structures are highly modular (helix, sheet, loop; motif; domain; protein) and typically satisfy several constraints.
  - Side chain structure determine accessible conformational space at each peptide bond.
  - A subset of proteins serve as catalysts for chemical conversions by lowering the activation barrier.

Week 4

Structural Biology

4.1 Tools of Structural Biology

- Today's lecture, for time's sake, will not focus on structural biology so much as it will focus on the *tools* of structural biology.
- **Structural biology**: The determination of the 3D structures of biomolecules and the study of their structure-function relationships, aimed at understanding the molecular mechanisms of biomolecules' functions and interactions.
- Today, we will focus on X-ray crystallography, NMR, and electron microscopy.
  - These are direct methods for the determination of 3D structure.
- Different length scales in biology.
  - E.g., small molecules through eukaryotic cells.
  - Different techniques have different ranges over which they can be useful for determining structure.
  - NMR, then X-ray crystallography, then single particle cryo-EM, then cryo-electron tomography (for organelles), then light microscopy (from small to large).
- X-ray diffraction played an essential role in the early days of molecular biology.
  - Before crystallography, X-rays were still used to detect and describe compounds such as keratin from hair.
  - Based on these diffraction patterns, Pauling was able to describe the basic elements of protein structure (e.g., α-helices and β-pleated sheets).
    - Pauling was able to guess the structures from very rudimentary, blurry data, such as...
      - ➢ 9.6 Å: Radius of an α-helix;
      - ➢ 4.6 Å: The distance between hydrogen-bonded strands in a β-pleated sheet.
    - This is what made Pauling a great chemist.
  - Rosalind Franklin's photograph 51. 3.4 Å corresponds to the stacking between the bases.
    - UChicago has a graduate course dedicated to interpreting X-ray diffraction patterns.
    - Zhao recommends we read Watson and Crick's original 1953 paper: "Molecular Structure of Nucleic Acids: A structure for Deoxyribose Nucleic Acid."
    - Dickerson in 1980: Crystal structure analysis of a complete turn of B-DNA.
    - Review by Eisenberg in 2003: The discovery of the α-helix and β-sheet, the principal structural features of proteins.
      - ➢ Eisenberg was Zhao's grad mentor.

- Modern X-ray crystallography.
  - Dickerson crystallized DNA, took an X-ray diffraction pattern, and thus was able to determine the position of every atom in DNA.
  - In-house X-ray sources vs. synchrotron X-ray sources, like the over 1 km loop building at Argonne, used for biomolecule characterization.
- Examples of high-quality protein single crystals shown.
- Workflow for macromolecular crystallography.
  - Grow a crystal, take it to a synchrotron, do an exposure to X-rays, rotate the crystal a degree or two, and do another exposure. Zhao had to fly from LA to Chicago to use Argonne's synchrotron while in grad school!
  - From this "movie" of exposures, you can get the electron structure and, with practice, resolve that into amino acids and other atoms.
  - You then fold the atom/amino acid sequence into a protein.
  - This whole workflow takes about 1 day today.
  - In the 80s-90s, this would be a graduate student's 4-5 year project and, if successful, would likely result in a *Nature* publication.
- Bottlenecks of macromolecular chemistry.
  - You need to clone the gene and express it to get the protein (most proteins, save a few such as RuBisCO, cannot be purified in sufficient quantities from natural sources). Cloning success rate: 100%. Expression success rate: 66%.
  - Then you need to purify it. Success rate: 35%.
  - Then you need to grow diffraction-quality crystals or take an NMR spectrum. Success rate: 29.5%.
  - The latter two are the biggest bottle necks; you lose a lot of your starting material in each one.
  - Crystallization is difficult because (most) proteins did not evolve to be crystallized, they evolved to function.
- The future of X-ray crystallography.
  - Argonne is the best synchrotron in the United States.
  - X-ray free electron laser (xFEL). LCLS (Linear Coherent Light Source) close to Stanford. 2-mile tunnel under the 280. You shoot electrons down a tube, vibrate them with magnets along the length of the tunnel so that they emit X-rays, trap the electrons at the end, and then just make use of the directed light. You get a super powerful beam (a billion times stronger than third generation synchrotrons).
  - Zhou and Zhao compared resolution from APS and LCLS on the same (type of) crystal; LCLS has higher resolution.
    - The beam is so strong that you damage the crystal though; at every point, you can only take one shot.
  - Radiation damage free-diffraction: Super fast exposure; take your diffraction before you cause damage.
- Future of macromolecular crystallography: microED.
  - You don't have to grow diffraction-quality crystals here (which are about 5 µm in size).
  - You can use nanometer-scale crystals instead.
  - microED: Micro-electron diffraction.

- NMR.
  - You still need to purify protein, but then you do NMR sample prep, acquire data, process the spectrum, and then do structural analysis.
  - NMR does not give you a map; it gives you the distance between different atoms. If you have enough of these constraints, you can calculate a structure.
  - In chemistry, you typically collect one-dimensional spectra; in biochemistry, you typically collect three- or four-dimensional spectra.
    - Most OChem spectra are 1-dimensional.
    - 2D spectra: You have to define a specific sequence to get resonance of both types of molecules together.
    - 3D includes one more type of atom that you want to resonate.
    - 4D is if you introduce some radiofrequency change over time, providing another dimension of information.
    - Physically, this is the most difficult technology.
    - The physics behind NMR is the toughest, as it dips into quantum mechanics.
  - Advantages:
    - Protein in native environment (crystal packing in XRD might introduce artifacts).
    - Information on dynamic structure (more on this later).
    - Information on protein interaction with other biomolecules.
      - ➢ For example, you can run an NMR of the protein and then of the protein mixed with some ligand.
      - ➢ This tells you how the protein interacts with the ligand.
      - ➢ This is a very hard experiment to run with XRD because you have to soak the crystal in ligand or ligate the protein and then crystallize it. In the words of Zhao, this is a "pain in the ass" to do.
  - Limitations:
    - Isotopic labeling required (^{13}C and ^{15}N at least).
    - Difficult for large proteins or complicated folding structures (if it's a large protein, you'll be stuck into a local minimum).
    - You need amino acid sequence information (including PTM) in advance (and the protein usually has to be 10-20 kDa).
  - Reason to run an NMR experiment: Not for a *de novo* 3D structure, but for that ligand-protein interaction.
- NMR is good for probing interactions.
  - Every peak in a 2D N-H spectrum corresponds to an amino acid (amide bond).
  - You can see a shift in proteins as they're ligated.
- The future of NMR.
  - Feed a cell nutrients and take an NMR of a whole cell.
- Recent revolution in structural biology.
  - 2017 Nobel prize for cryoEM[1].
- Two major techniques: SPA and CryoET.
  - SPA is single-particle analysis. Embed viri in a tray, take a 2D projection, do FTs, and reconstruct the 3D structure. Multiple particles averaged.

[1] What dad mentioned years ago?

  - Cryo-electron tomography/STA (single tilt analysis; there used to be double tilt analysis but it was too hard to keep the stage stable): Focus on one single particle, but tilt the stage within the microscope. Usually used to look at subcellular organelles.
  - SPA is already comparable to XRD in its ability to generate atomic-level resolution.
  - Workflow: Aqueous solution → sample vitrification (rapid cooling) → low-dose image collection by cryoEM → SPA or cryoEM → structure.

- Methods for SPA have been developed for decades.
  - Began in the 1970s with reconstruction of an icosahedral structure.
  - Modern cryoEM began around 2011.

- Problems that prevented high-resolution cryoEM reconstruction.
  - Radiation damage: Only limited amount of electron dose can be used since light atoms, e.g., hydrogens can be damaged.
  - Bad detection: Only limited amount of signal is recorded.
  - Beam-induced motion: Hard to avoid.
  - Sample heterogeneity: No crystal lattice as constraints.
  - Problems 1-3 were solved with a good camera in 2012. Multiple exposures aligned and averaged (take multiple shots [a movie] over 4-5 seconds and then align the relative positions and average).

- Direct electron direction camera.
  - Used to use a CCD camera. The screen of an electron microscope shows a green fluorescent image[2]; to digitize the image, we use a CCD camera; detects photons only, so we convert electrons to photons with a scintillator. This causes a lot of signal loss, though, because of the conversion.
  - The invention of DDC gets rid of the scintillator and fiber-optic coupling, allowing literal direct detection of electrons.
  - DQE: Detective Quantum Efficiency goes up.

- SPA revolution.
  - The structures that SPA focused on were ones that were very difficult or impossible to crystallize.
  - People have tried to crystallize membrane proteins for decades, but in 2010, SPA cryoEM gave another way.

- Workflow for single-particle cryoEM.
  - Prepare the sample and then plunge it into liquid ethane.
  - Ethane has a very large heat capacity (not liquid nitrogen). We don't want crystalline ice; we want vitrous ice, so that the ice crystal doesn't affect the experiment.

- Advantages of cryoEM analysis:
  - Removes the crystallization bottleneck.
  - Dynamics can show you different states of a molecule.
  - For example, this is how we figured out the different conformations/rotation of ATP synthase.

- Future of cryoEM: Cryo-electron tomography.
  - Currently mainly used to look at larger structures, e.g., organelles.
  - Used to study how SARS-CoV-2 infects cells.

[2] Think of the TEM machine in the GCIS sub-basement.

    - Allows for a better understanding of its S-proteins, as well.

- CryoEM's limitations.
  - The sample cannot be too thick.
  - A eukaryotic cell is typically too thick.
  - Circumventing this: FIB milling (focused ion beam). Takes off part of the cell.
  - Put everything in an SEM (to guide your progress), do the milling, and then transfer to a TEM.

- Comparison of structural biology techniques.

		single-particle cryoEM	**X-ray crystallography**	**xFEL**	**microED**	**NMR**
Setup	imaging source	electron	X-ray	X-ray	electron	Magnetic field and RF pulses
	lens system	yes	no	no	yes	no
Sample	form	solution	crystal	micro-xtal / xtal in cell	micro-xtal	solution
	quantity	low	high	high	low	high
Throughput	sample screen	low	high	high	low	low
	data collection	days	minutes	hours to days	hours	days
	data processing	weeks	days	days-weeks	days	weeks
Limit	resolution	up to 1.2 Å	better than 1.0 Å	better than 1.0 Å	better than 1.0 Å	N/A
	MW	> 60 kDa	no	no	no	< 100 kDa
Pain Point		screen freezing conditions	growing crystals	growing a ton of micro-xtals	growing micro-xtals	isotopic labeling
Unique Benefit		multiple states	anomalous scattering	radiation damage free	few micro-xtals	dynamic information

Table 4.1: Comparison of structural biology techniques.

- Graduate course (though 1-2 undergrads take it every year): BCMB 32600 Methods in Structural Biology.
  - Did SARS-CoV-2 last time.

- Quick survey of Cross-β diffraction pattern and amyloid.
  - Zhao's research topic as a graduate student.
  - Difference between β-pleated sheets (XRD points expand out linearly) and cross-β patterns (XRD points expand out perpendicularly), the latter of which are generated by amyloids.
  - Eisenberg published seven peptides whose structure they determined with XRD and which ran perpendicular to the fibers.
  - Fibers are like 1D crystals which are very hard to crystallize. Breakthrough in 2015: used microED to determine the structures of the very small "invisible" (under light microscope) crystals.
  - Solid-state NMR helps, too.
  - cryoEM helps more.

- Nobel prizes in 1962.
  - Crick was a grad student of Perutz; Perutz and Kendrew determined the first crystal structures of protein.

  - Wilkins was the mentor of Franklin; she had already passed away by 1962. She was not recognized; women are still not recognized enough.

- Future Nobel prize in Zhao's evaluation.
  - John Jumper (former grad student at UChi) develops AlphaFold2.
    - Combines multiple sequence alignment (MSA, genetic information) with pair representation (distance matrix, analogous to NOE spectrum, structural geometrical information) as the input.
    - Introducing attention-based neural-network architecture...
- Think of a neural network as a large machine with a lot of knobs.
  - Once the knobs are tuned with existing data, the machine is capable of predicting, decoding, and analyzing unknown data.
- AlphaFold2 data flow.
  - The key step is MSA + pair as input.
- Prediction of a human methyltransferase that has not been crystallized with a high confidence.
  - AlphaFold2 predicted the structure, however! Collaboration between Zhao and Chuan He.
- Try it with your own protein using ColabFold via Google.
- Current limitations.
  - Not implemented for nucleic acids.
  - Static structures.
  - Sequence length limit.
  - Insensitive to point mutations.
  - Poor performance for antibody recognition.
- One critical reason why AlphaFold is so successful.
  - The database has gotten huge in the last several decades. High quality experimental training sets are available.
  - Another advantage is high quality sequence technology.
- Remaining challenges: Complex structures, dynamic structures, and intrinsically disordered proteins.
  - Still difficult due to issues computing the energy landscape of large biological complexes.
  - Nuclear pore: Complex with over 1000 proteins.
  - Science: Volume 376, issue 6598, 10 June 2022 reviews research surrounding the nuclear pore.
  - How do we extract the dynamic information, regardless?
- Is structural biology still cool?
  - Various reasons it's still needed.
- Last class, we talked about Ramachandran plots.
  - Ramachandran statistics is used as a validation method in X-ray crystallography and single-particle cryoEM. Not used in AlphaFold.
- Keep my eye on Nick Korn; seems to really know what's going on and asks good questions.
- More info will be provided later on how this info will be incorporated into future exams.
- Reach out to Zhao to talk more about his work! Seems very related to analytical chemistry. What is his view of the field and how my math background can help?

Week 5

Sequencing and Organelles

5.1 Sequencing and Next-Generation Sequencing

10/25:

- Yamuna wants us to call her by her first name.
- Spends the first 5 minutes glowing about teaching the class.
- As a postdoc, Yamuna worked at the bench next to the guys who developed Illumina sequencing.
- Sequencing represents the best of biochemistry because you have to know the chemistry to do it and the biology to interpret the results.
- Doing science via discovery (e.g., why is the sky blue?) vs. understanding (e.g., how does this work?).
  - Chemical biology is the science of invention because you are trying to take the natural and control it.
- What Yamuna wants us to take away: Understand how polymerases and everything works and then be able to tell what our sequence is.
  - Sequencing is an exercise in tweaking the chemistry of biomolecules to get a certain result, enabled by the fact that we understand it so well.
- DNA sequencing: Maxam-Gilbert, Sanger, Pyro (454) sequencing, Illumina, Nanopore, SMRT.
- Two principles that underlie DNA sequencing.
  - Size-based separation on a gel (esp. for the older ones).
  - **PCR**.
- **Polymerase chain reaction**: A technique that amplifies (rapidly duplicates) a certain sequence of DNA millions or billions of times. *Also known as* **PCR**.
  - Suppose you have a strand of DNA and you want to know the sequence of a 150 bp bit.
  - You need sufficient and sufficiently pure starting material to begin. Thus, if we have 50-100 copies of the DNA from extraction and mince them, the pieces will come out in different lengths.
  - But we need way, way more copies to do meaningful chemistry and, moreover, we need only copies of the one specific set of base pairs.
  - Solution: Polymerase chain reaction uses DNA polymerase, dNTPs, Mg^{2+}, and some other things to make many many copies of just the 150 bps you're interested in.
  - You need a primer (about 20-30 base pairs; what is necessary for specificity) that will sit on the beginning of the region.

  - PCR uses a thermal cycler (a fancy oven that heats and cools between temperatures of your choosing at a rate of your choosing).
  - DNA in an Eppendorf tube in the thermal cycler. We heat it to unwind the strands and then break the hydrogen bonding, yielding single-stranded DNA. Our forward and reverse primers sit on the single strands at the beginning of our target region. DNA polymerase attaches and copies until it falls off. Then we repeat.
  - With every cycle, we increase/amplify the number of copies of target DNA vs. the variable length DNA. Thus, the variable length becomes more of an impurity. Now we can start to do chemistry.
  - PCR was invented by Kary Mullis (who Yamuna isn't a fan of because he was a heavy user of LSD, downplayed humans' role in climate change, and doubted that HIV is the sole cause of AIDS).
  - How do you create the primer if you haven't sequenced the DNA yet??

- Separating DNA duplexes on the basis of size/length (*not* polarity) using Agel (which is fancy TLC).
  - If DNA is small, it will easily snake through the gel. If it is big, it will take longer.
  - Like gel electrophoresis, you still have a cathode and anode. DNA (negatively charged due to phosphate groups) will move toward the cathode.
  - Entirely pure substrate $\rightarrow$ one band.
  - You can separate 48 bp strands from 49 bp strands.
  - You have to chop up your DNA into reasonable sizes so that it can separate on a gel: 1000 vs. 1001? Not possible. 100 vs. 99? Possible. Resolution is better.

- Huntington's genetic test.
  - There is a protein/gene called Huntington. Everyone has a short number of repeats on the Huntington protein, but if you have two many (40+), you will develop Huntington's disease.
    - 26-27 repeats is the border. This issue arises from polymerase "going nuts" and adding more repeats than it meant to.
    - A **pathogenic** number of repeats vs. you being fine.
  - Amplify the section of your DNA containing the repeats. Then it is not necessary to sequence and count; you just need to determine the length of the repeating strand.

- Cystic fibrosis.
  - Often results from the $\Delta F508$ mutation (single AA mutation at phenylalanine 508).
  - Yamuna's cousin died aged 36 from cystic fibrosis, but it wasn't $\Delta F508$ — it was two "variant of unknown significance" mutations. Cases like hers allow us to canonize the noncanonical mutations.

- 10 years, $1 billion to sequence the entire human genome using Sanger sequencing (slow and very expensive).
  - Someone else envisioned sequencing the entire human genome for $1000 in a day.
  - If feasible, it would have been great to understand all genomes, but instead, it helped us with COVID (detecting variants in a population and a person, virility, capacity for transmission).
  - This saved many people by prevention (e.g., travel restrictions) before a cure (like the "mRNA vaccines") existed.

- Back in 2005 when Yamuna started her lab in India, she would get data as a **sequence chromatogram**.

- **Seqeunce chromatogram**: A graph consisting of various colored peaks, each corresponding to one type of base pair.
  - Blue peak: Cytosine, Green peak: Adenine, Red peak: Thymine, Black peak: Guanine.

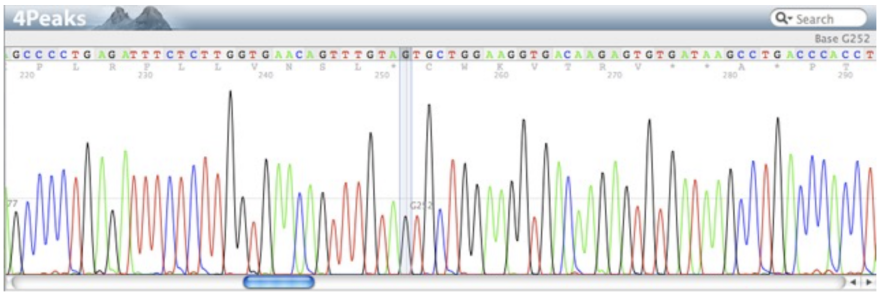


Figure 5.1: Sequence chromatogram example.

  - Idiot-proofed for biologists to just read off their sequence from the top of the display window.
  - How the graph is generated (briefly; this is Sanger sequencing).
    - You write the sequence based on the length of the strand and how far it travels and which fluorescent dye has been attached to our gene.
    - When DNA is being built, different dNTPs come in and sample the active site. If hydrogen bonding is correct, DNA polymerase locks into place, fuses it to the 3′ end of the growing strand, and releases a **pyrophosphate**.
    - This only happens when you have the right dNTP (there are energy barriers if you have the wrong one based on faulty hydrogen bonding).
    - Release of a pyrophosphate is key to another sequencing method.
    - Ability to make DNA artificially in a chemistry lab (Caruthers, 1985). You can attach literally anything to the growing 3′ end. This allows you to create primers that set an address.
      - ➢ Yamuna believes this should have won a Nobel prize since it's been the basis for several others.
    - If you attach a ddNTP to the growing end, you stop growth.

- **Pyrophosphate**: Two phosphates bound via a single linking oxygen, i.e., $P_2O_7^{4-}$. *Denoted by* **PPi**.
- Maxam-Gilbert sequencing.
  - Developed by Wally Gilbert and Allan Maxam.
    - Gilbert and Sanger originally won the Nobel Prize for sequencing.
  - Know this for historical reasons; came out second, but was adopted first.
  - Start with a bunch of copies (circa 1 million) of a DNA strand obtained using bacterial cloning.
  - Label the 5′ end of each sequence with ^{32}P. This will allow for visualizing the strands at the conclusion of the procedure using autoradiography and X-ray film.
  - Divvy up the labeled DNA between four Eppendorfs. To each tube, add a chemical that is selective for one or two nucleobases. Add just enough of the chemical so that every strand will react once. Remember that most strands will not react. Then introduce hot piperidine (Yamuna said hydrazine??) to cleave the strand right before (Yamuna says after??) the modification.
  - Chemicals:

$$\mathrm{A} + \mathrm{g} = \mathrm{HCOOH} \qquad \mathrm{G} = \mathrm{Me_2SO_4} \qquad \mathrm{T} + \mathrm{c} = \mathrm{N_2H_4} \qquad \mathrm{C} = \mathrm{N_2H_4} + \mathrm{NaCl}$$

  - Running these mixtures on a gel leads to bands for each cut and unreacted strand.
  - The strand that travels the farthest (is the lightest/shortest) corresponds to the first nucleobase. The strands that travel the least are the unreacted strands.
    - Example 1: No band in the T + c column and a band in the C columns? Cytosine.
    - Example 2: Bands at the same level in the A + g and G columns? Guanine.

- Sanger sequencing.
  - When cloning became passé, everyone switched to Sanger.
  - Two methods of Sanger sequencing: Sequential and parallel.
  - Sequential Sanger sequencing.
    - Amplify your region of interest using PCR.
    - However, during this process, add a small amount (1-5%) of a specific dideoxynucleotide (ddNTP). If you include a small amount of ddATP for example, then whenever DNA polymerase matches one of these with a thymine and incorporates it into the growing strand, the strand will not be able to grow any further (there is no longer a $3'$ hydroxyl to bond the next nucleotide to).
    - This will allow you to generate stops at every nucleobase of a certain type.
    - Doing this for every nucleobase independently and then running all four samples on a gel gives you a similar result to Maxam-Gilbert sequencing, except that this time, our result is analogous to cleaving after the "modification" and we don't have "leaks" as with the T + c and A + g chemicals.
  - Parallel Sanger sequencing.
    - Amplify your region of interest using PCR.
    - However, during this process, add a small amount (1-5%) of fluorophore-labeled ddNTPs such that each of the four ddNTPs fluoresces a different color. Incorporating these will guarantee that each strand of DNA ends in a fluorophore-labeled ddNTP.
    - These strands can be separated with high accuracy using capillary gel electrophoresis.
      - ➢ Capillary gel electrophoresis is very fancy gel — very long and very thin.
    - As each strand moves through the capillary, it eventually passes by a light fluorescence detector.
    - This generates the sequence chromatogram.
  - Better since it doesn't have radioactivity, once fluorophores became stable, and after the advent of capillary gel electrophoresis.
- Svante Pääbo at the Max Planck Institute won the 2022 Nobel Prize in Physiology or Medicine for sequencing the Neanderthal genome.
  - He extracted DNA from skulls and bones. Every bit of DNA was missing something, but by sequencing enough and comparing, he was able to fully reconstruct it.
  - He did this with **pyrosequencing**, which many biologists had forgotten about.
- **Pyrosequencing**: A sequencing by synthesis method that works as follows. *Also known as* **454 sequencing**. *Procedure*
  1. Begin with a pure set of DNA sequences generated via PCR. Bind adapters to the sequences, and biotin to the adapters. Immobilize multiple copies of each sequence on a number of streptavidin beads.
  2. Bind a primer to each sequence and attach DNA polymerase.
  3. Add a specific dNTP (dATP, dTTP, dGTP, or dCTP).
  4. Suppose the first base to be sequenced/synthesized is adenine and dATP is the first dNTP added. Then DNA polymerase will click dATP into place, releasing a pyrophosphate.
  5. The PPi is used by ATP-sulfurylase to generate a molecule of ATP.
  6. This allows Luciferase to use ATP and its substrate to generate a flash of light.
  7. Before adding in another type of dNTP, it is necessary to remove the previous one. This is accomplished by adding apyrase, an enzyme that converts all available dNTPs to dNDPs and then inactive dNMPs.

8. Counting the number of flashes of light after a dNTP is produced tells us how many of that dNTP in a row there are at that point.

- Example of pyrosequencing.
  - Consider the strand ATGGCCC.
  - Introducing dATP, dGTP, or dCTP at first will lead to no flashes of light. Introducing dTTP will lead to one flash of light (because T binds with A and there is one A).
  - Similarly, introducing anything other than dATP next will lead to nothing, and introducing dATP will lead to one flash of light.
  - Now introducing dCTP will lead to two consecutive flashes of light (as two pyrophosphates are released from the addition of two dCTPs to the growing strand, one for each dGTP in the guiding strand).
  - Lastly, introducing dGTP will lead to three consecutive flashes of light.
- Notes on pyrosequencing.
  - 454 is what the company referred to the technology as before it was released and named "pyrosequencing."
  - Pyrosequencing is the bridge between the ways Yamuna used as a grad student and what we do today.
  - In an analogy, ATP-sulfurylase is like the light switch, luciferase is like the lightbulb, and apyrase is like the eraser between steps.
  - You generate a bead with many copies of a specific strand on it.
  - How this works in a system:
  - Take DNA, sonicate it to break it up, make the library, add adapters.
  - Amplify using emulsion PCR (little droplets of water in a mix of oil that contain dNTPs, primers, water, polymerase, etc.).
  - Relation to chemistry 1-bead, 1-compound question.
  - During emulsion PCR, the strand that is not covalently bonded (i.e., the newly synthesized one) comes back and reattaches.
  - PCR amplification occurs until every strand displays the same DNA sequencing.
  - Many wells; each one contains a single DNA sequence. Then flow in dATP plus an enzyme cocktail.
  - You need a big flash of light (multiple photons — 20-30 flashing at the same time).
  - Your computer flows in different bases to different wells and looks for what gives you a flash.
  - Allows you to sequence in a massively parallel way.
- Illumina sequencing (currently the most important method).
  - Sequences 200-300 bps at a time.
  - Nanopore and SMRT sequencing give you extremely long sequences, but most big biological discoveries today are based on Illumina sequencing.
  - Once you have your sequences of interest, you attach primers and...
  - Attach your strands to the surface of a wafer.
  - Bridge synthesis on the wafer.
  - You get a flash of light whenever you add.
  - You get an answer from your entire surface instead of just a single molecule. Since DNA polymerase makes errors, this eliminates them via the law of averages.

  - The cost of sequencing is now in storing the data, not in the reagents.
  - Five major challenges to solve to achieve next generation sequencing (NGS) by Illumina.
    - The 3′ OH problem.
      - ➢ If you want to protect the 3′ OH with a fluorophore, you have a 2 hour deprotonation. This means that it will take 25 days to sequence 300 bases.
      - ➢ If you use 2-o-nitrophenol, you have a UV-deprotonation. Instaneous but skin cancer.
      - ➢ Single color readout is impractical; thus, you need a four-color readout.
      - ➢ The ideal 3′ OH protecting group is small, stable under aqueous conditions, has quantitative cleavage and high turnover, and preserves the DNA integrity.
    - The fluorophore problem.
    - The polymerase problem.
    - The surface chemistry problem.
    - The problem of polymerase-generated errors and parallelization.

- Check out videos online.

5.2 Cell Biology for Chemists

10/27:

- Yamuna starts by telling us that we should feel free to sleep through class and watch the recording if we want since it's so early.
- Today: Third-generation sequencing (left over from last class) and then the cell.
- So far: Biomolecules, structure, and function. Third generation sequencing will show you how we can assemble all of these components together to get them to do very complex work in union in a very purified way.
  - After that, we will focus on how they all function together in the cell.
- Yamuna used to think that the cytoplasm was mostly water with a few stray biomolecules, but in reality, there are tons of biomolecules all crammed together into a highly viscous "hot pot" that these components have to navigate through.
  - Another factor is quick recognition and selectivity; since biomolecules bump into so many things so quickly, they have to be able to tell what they *don't* want to interact with very quickly after collision.
- Illumina qualifies as second-generation sequencing.
  - It is still limited to 200-300 base pairs at a time. You can't do an entire run at once. Thus, if you want to sequence repeat regions (such as at the end of telomeres), you can't tell how many repeats you have using such methods.
    - This repeat problems is one of the reasons they declared the Human Genome Project concluded but "90% finished."
    - After solving the repeats issue, then they declared that the work was done.
    - Another reason was that they wanted to make their data public so Craig Venter didn't copyright it and freeze science.
- **SMRT sequencing**: A method of sequencing in which DNA polymerase is immobilized over a camera that records each fluorophore flash as nucleotides are added. *Also known as* **single-molecule real-time sequencing**, **PacBio sequencing**.
  - A type of third-generation sequencing.

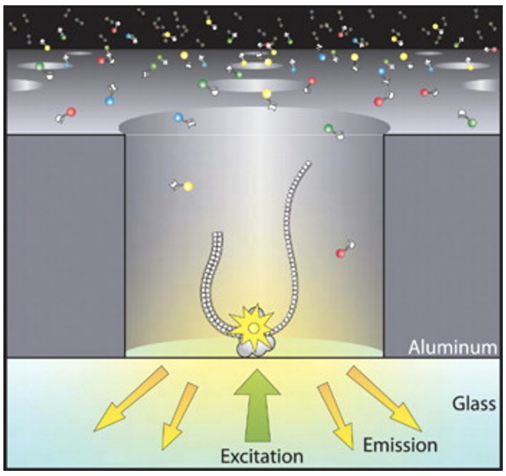


Figure 5.2: SMRT sequencing setup.

  - Called real-time because a DNA polymerase is immobilized on a surface; dNTPs carrying fluorophores float in the mix; and as each dNTP is added to the strand by the DNA polymerase, you observe it fluorescing red, green, yellow, or blue exactly when it is added. A camera counts the sequence of colors.
  - But how do you selectively get the fluorescence of just the base that is added? You fix the DNA polymerase above a **zero-mode waveguide**.

- **Zero-mode waveguide**: An ultrasmall pore, smaller than the wavelength of light, over which a biomolecule can be held. *Also known as* **ZMW**.
  - In our example, DNA polymerase is held over the pore.
  - The gap is so small that light passing through it can only interact with the DNA polymerase fixed directly above it and cannot travel further up into the rest of the matrix.
  - In effect, a ZMW is a "short-sighted fluorescence microscope."

- More on the ZMW.
  - One of the smallest possible detection volumes.
  - Developed by Watt W. Webb.
  - Because the light that comes in has very small wavelength, it cannot travel upwards. You have the greatest intensity right where the light comes in, and then the intensity really falls off; thus, other dNTPs cannot be irradiated.
  - Notice that all fluorophores are attached at the $3'$ (??) γ phosphate, so when a dNTP is added, the pyrophosphate plus fluorophore is sliced out, resulting in a flash of light.
  - The process occurs in parallel in thousands (now millions) of ZMWs per SMRT cell.

- Difference from illumina sequencing: Illumina is not real time.
  - DNA polymerase goes so fast naturally (too fast for any camera) that you have to slow it down.
  - You slow it down by attaching a protecting group to all dNTP's $3'$ phosphates. This group must be removed by a deprotection before the next base can be added.

- **Nanopore sequencing**: An electrical sequencing method currently in development.
  - A company has proposed that sequencing can be done on-site using a device the size of a USB drive.

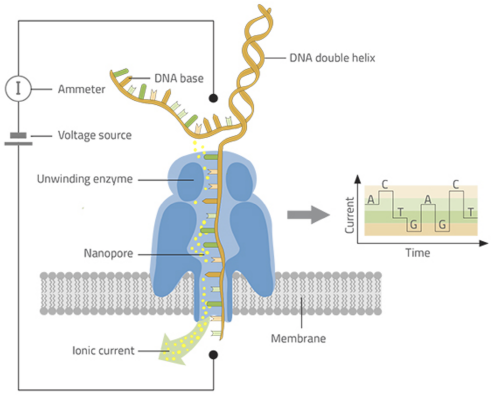


Figure 5.3: Nanopore sequencing setup.

  - This method of sequencing is electrical, not chemical like all of the others.
  - The pore (a **porin**) is of bacterial origin.
  - Helicase sits at the top, unwinding incident DNA so that one single strand fits through the pore and the other doesn't.
  - Each base passing through the pore blocks it to a different and unique extent, changing the ion flow through the pore, facilitating sequencing by current.
  - This method also facilitates sequencing of modified nucleobases (e.g., methylated cytosine or adenine) because they will provide a unique current, too.

- Note on the quiz.
  - We will need to learn about **ChIP-Seq**.

- **ChIP-Seq**: A technique that identifies which proteins sit where on the genome. *Also known as* **chromatin immunoprecipitation and sequencing**.
  - Enables us to, for example, find out where in the genome a particular transcription factor sits.
  - ChIP-Seq accomplishes this by **immunoprecipitating** the transcription factor.
  - Immunoprecipitation **cross-links** the transcription factor to the DNA.
  - You sonicate the DNA to break it into different bits and then pull down specifically your protein and DNA sequence. Thus, all instances of your protein come down along with all DNA to which it was bound at the time of cross-linking.
  - Lastly, you sequence the immunoprecipitated DNA and compare it to reference DNA to locate it within the genome.

- **Immunoprecipitation**: Making a biomolecule more heavy so that it can be precipitated out of solution by attaching an antibody to it.
  - Achieved by introducing a targeted antibody into the system.

- **Cross-linking** (DNA): The process of covalently binding cellular proteins to DNA.
  - This typically occurs upon exposure to various **endogenous**, environmental, and chemotherapeutic agents.

- **Endogenous** (biomolecule): A biomolecule that grows or originates from within an organism.

- A good book to study this content is *Molecular Biology of the Cell* by Bruce Alberts.

- Aside: Yamuna's perspective on pharma and biotech companies — there are three kinds of people in chemical biology.
  1. **Assassins** are asked to develop a molecule that selectively binds a specific family of biomolecules, and they use all of the tools of chemistry and biology to do so.
  2. **Recruiters** are asked to find the best way to inhibit a certain type of protein.
  3. **Deciders** decide what pathway should be targeted.
- Deciders are fairly small in number and consult for a variety of companies because they have the vision to know what to do.
- We now conclude sequencing and move onto studying the cell.
- What makes a city alive?
  - Class-suggested answers: Memory, people, energy, and interaction networks.
  - We should see a cell the way we see a city.
- Aspects of a city.
  - Transportation, currency flow, executive function, import and export, infrastructure, schools, energy, defense systems, cleaners, the postal system, hospitals, and people/parts.
- Analogies within a cell.
  - Postal system: Golgi.
  - Energy: Mitochondria.
  - Garbage disposal: Lysosomes.
  - Defense systems and repair: DNA repair mechanisms.
  - Border: Cell membrane.
  - Transport system: Microtubules and the cytoskeleton.
    - Allow you to go from the membrane to deep within the cell.
    - Proteins don't migrate by random diffusion but catch hold of an actin network and move.
  - Factories: Ribosomes are protein production factories.
- MicroRNAs largely regulate various networks and exist for robustness.
- The cell is alive because all of these parts have to interact with each other. All of their functions are highly interlinked.
- Look up the difference between plant and animal cells! Will be an exam question!!
  - Plant cells have **plasmodesmata**, a **cell wall**, a **central vacuole**, and **chloroplasts**.
  - Animal cells have **centrioles**.
- The plasma membrane: Basics.
  - The fluid mosaic model. We are taught in school that the plasma membrane is a homogeneous sea of lipids that proteins are mixed into. However, this is *very* wrong.
  - The plasma membrane is the first line of defense for the cell against invading pathogens.
  - This region is really important.
    - The first-most drugged class of molecules is **G-protein coupled receptors**.
    - The second-most drugged class of molecules is cell-surface ion channels.
  - A phospholipid has a hydrophilic head and a hydrophobic tail.

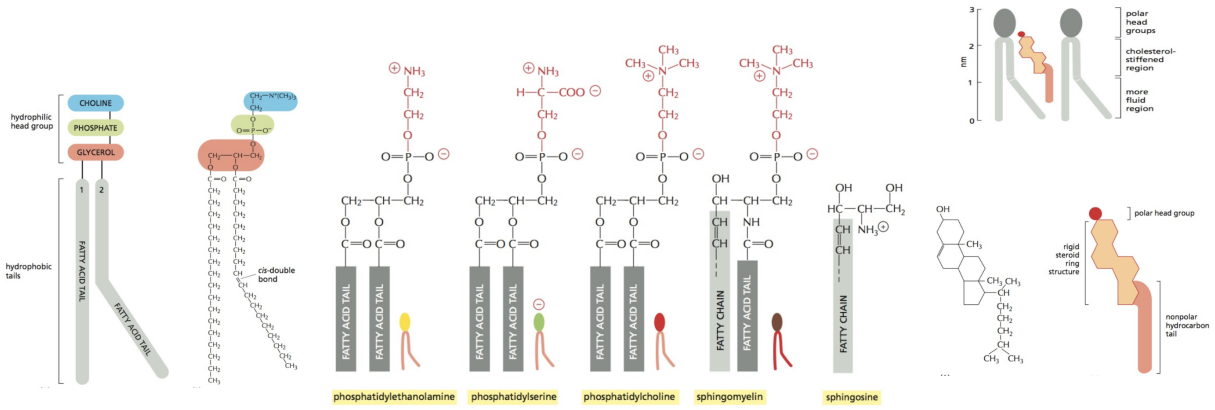
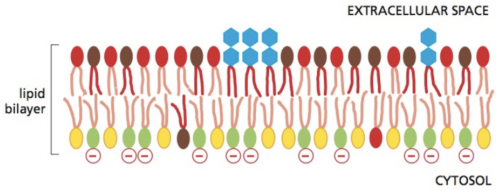


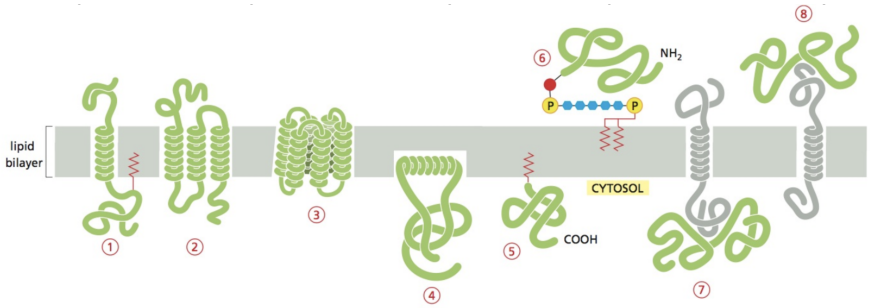
Figure 5.4: Plasma membrane constituents.

- We may approximate them as cylinders according to ??.
- The hydrophilic heads face outwards, and the hydrophobic tails face inwards.

- **G-protein coupled receptor**: A protein that is present on the cell surface. *Also known as* **GPCR**.
- The plasma membrane: Main constituents.
  - Comprised of 500-2000 different kinds of lipids molecules; it is not homogeneous.
  - The variations come from different alkyl chain lengths and degrees of unsaturation. You can also have different head groups: The most common are phosphatidylethanolamine, phosphatidylserine (PS), phosphatidylcholine, and sphingomyelin.
  - You also have 17-23% cholesterol in the membrane to provide thermal stability; more than that makes the membrane too stiff, less than that makes the membrane too wobbly.
  - Cholesterol sits in the membrane wherever the unsaturations are. Unsaturations cause bends which allow cholesterol to slide in and stabilize the system.
- The plasma membrane is asymmetric.



(a) Phospholipid variations.



(b) Membranous protein types.

Figure 5.5: Asymmetry in the plasma membrane.

  - The outer and inner leaflets look quite different.
    - Note that the four colors of phospholipids in Figure 5.5a (red, brown, yellow, and green) correspond to the four head groups in Figure 5.4.
    - The inner membrane has a lot of PS.
    - The outer membrane has a lot of **glycolipids**.
  - If one head group has two tails, it looks like a cylinder. If one head group has multiple hydrophobic tails, it looks more like a cone. If one head group has a couple of tails and a large glycolipid, it will look like an inverted cone.
    - This affects packing in the plasma membrane; larger hydrophilic groups need more space and cause the plasma membrane to pucker outwards; larger hydrophobic groups cause it to bend inwards.
  - We know a cell has died in the lab via annexin staining.
    - When a cell is alive, it is constantly flipping PS molecules that have migrated to the outer leaflet back to the inner leaflet. When it dies, it can no longer do this, and PS molecules flip to the outer leaflet in large numbers.
    - Annexins bind PS molecules, signaling to all immune cells that this one has died and they should come eat it.
  - Another place from which asymmetry comes is membranous proteins.
  - Transmembrane proteins.
    - Proteins can have transmembrane regions (usually α-helical and hydrophobic).
    - Some transmembrane proteins are **single-pass** while others are **multipass**.
    - Once a transmembrane protein is synthesized, it folds, condensing and squeezing phospholipids that are in the way out of its volume.
  - Attaching a protein to the inner leaflet.
    - Use an **amphipathic** helix.
    - Proteins can also be **lipid anchored** to the membrane.
  - Protein-protein interactions with a single-pass transmembrane protein can attach proteins to the the inner or outer leaflet.
  - Attaching a protein to the outer leaflet.
    - Use a **GPI anchor**.

- **Glycolipid**: A huge number of sugars attached together to form a hydrophilic head group on the outside of the plasma membrane.
- **Single-pass** (transmembrane protein): A transmembrane protein that passes through the membrane once.
- **Multipass** (transmembrane protein): A transmembrane protein that passes through the membrane more than once, with the different transmembrane regions connected by various AA chain linkers.
- **Amphipathic** (helix): An α-helix for which one side is hydrophilic and the other is hydrophobic.
  - These are rare.
  - They insert into the surface of the plasma membrane, with the hydrophilic region oriented toward the cytoplasm and the hydrophobic region oriented toward the hydrophobic interior of the plasma membrane.
- **Lipid anchor**: A fatty acid lipid chain covalently bound to a protein and inserted into a cell's plasma membrane.

- Some proteins show out a serine, cysteine, or lysine. These are capable of being alkylated (via an ester, thioester, or amide linkage, respectively). The alkyl chain can then bind to a fatty acid lipid chain, which embeds itself in the similarly hydrophobic region of the plasma membrane.
- Single lipid anchors usually aren't very stable; in order to achieve stable integration, you typically need one more chain.

- **GPI anchor**. *Also known as* **Glycosylphosphatidylinositol anchor**, **GPI linker**.
  - Very important.
  - A protein attaches (via a GPI linker) to a lipid; we'll talk about these in greater depth later.

- Different kinds of lipid anchors.

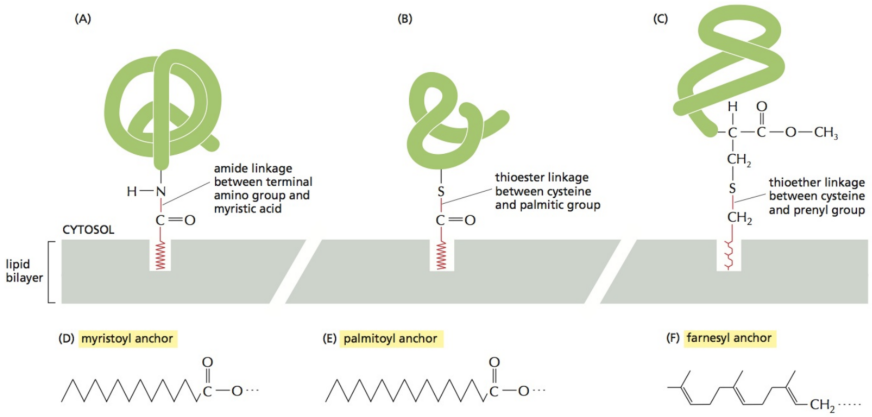


Figure 5.6: Lipid anchor types.

- Most transmembrane proteins cross the bilayer in an α-helical conformation.

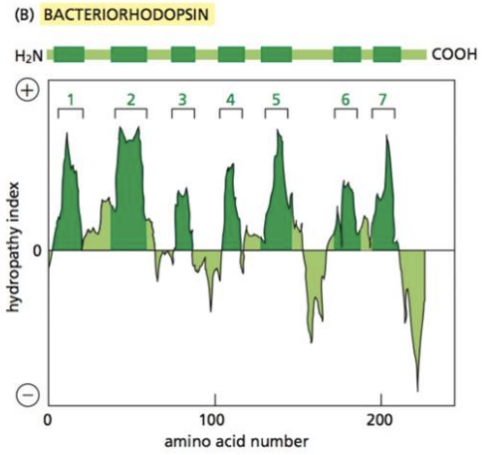


Figure 5.7: Hydropathy chart example.

- Transmembrane domains are predicted using the **hydropathy index**.
- You use a sliding window of 20 AAs. This means that you average the hydropathy indices of 20 adjacent amino acids at a time in a protein to determine what 20-AA region has the highest overall hydropathy index. This region is the one that's most likely to be transmembrane.

  - From a plot of the sliding window hydrophobicity vs. AA number, you can look for peaks in hydrophobicity. These correspond to hydrophobic, transmembrane regions.
    - There will be a question about this in the exam!

- **Hydropathy index**: The amount of Gibbs free energy needed to transfer an amino acid residue from water to a nonpolar solvent.
  - A positive value means that the AA is hydrophobic; vice versa if the value is negative.
- Most transmembrane proteins cross the bilayer in an α-helical conformation.

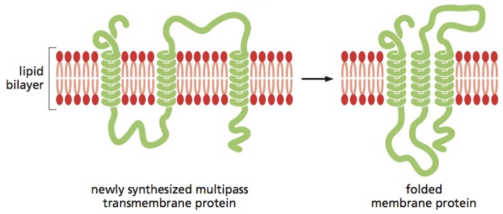
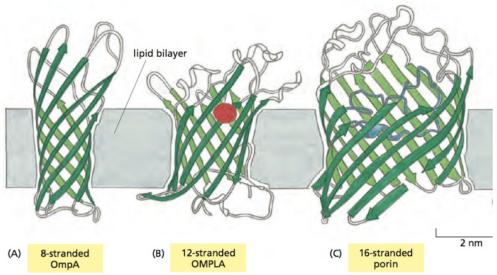


Figure 5.8: Multipass transmembrane protein folding.

  - As a protein is produced (more on production in the ER and transport to the cell surface later), the transmembrane domains insert into the plasma membrane and squeeze out intermediate phospholipids.

- Proteins are embedded in different ways.



(a) β-barrel proteins.

(b) Extracellular protein binding and resultant plasma membrane bending.

Figure 5.9: Alternate transmembrane protein embedding.

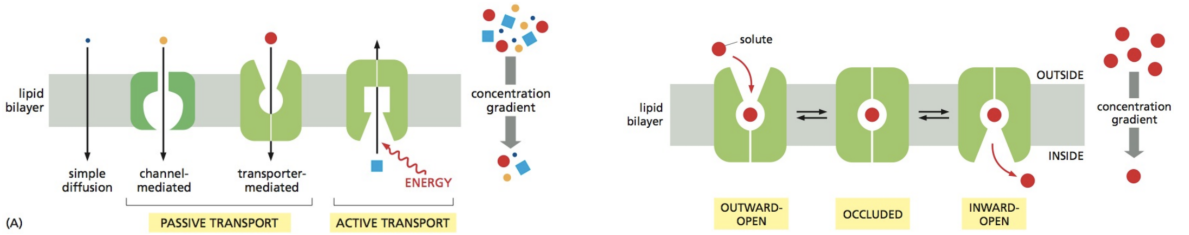
  - It is possible to embed without α-helices. Indeed, we can use β-sheets composed of hydrophobic residues.
    - Cytosolic proteins of this form tuck all of their hydrophobic side chains inside.
    - Transmembrane proteins of this form show all of their hydrophobic side chains outside.
  - Example: The MSPA porin from nanopore sequencing.
  - These are called ***β*-barrel proteins** and are often involved in transport or are receptors.

  - Barrel size varies.
    - MSPA has a huge barrel.
    - Smaller barrels are often filled up by amino acids on the inside but can act as a scaffold to interact with proteins on the top or bottom. Moreover, selected small molecules can sometimes pass through.
  - These channels are very large in general and can bend membranes by binding proteins on the outer leaflet.
    - These can act as large head groups and induce outward puckering.
    - A conformational change in the protein induced upon binding can place mechanical pressure on the membrane.
    - A protein can bind to multiple head groups and push them apart.

- The inside vs. the outside of the cell.
  - 33% of our ATP goes to maintaining ion gradients.
  - Remember that some molecules are rich outside and poor inside, and vice versa.
  - For example, cells need to take in glucose and enrich its concentration within the cell.

- We now look at the transport processes that maintain these gradients.

- There are two main classes of membrane transport proteins.



(a) Types of transport.

(b) Phases of transporter-mediated passive transport.

Figure 5.10: Membrane transport options.

  - **Passive transport** vs. **active transport**.
  - Types of passive transport: There is some simple diffusion/leakage, channel-mediated diffusion such as ion channels allow very fast diffusion, and transporter-mediated diffusion to move larger molecules.
  - Transporters usually catch molecules on one side of the membrane, inducing a conformational change, and release them on the other side of the membrane. Outward-open, occluded, and inward-open states.

- **Passive transport**: A type of membrane transport that *does not* require energy to move substances across cell membranes.

- **Active transport**: A type of membrane transport that *does* require energy to move substances across cell membranes.

- How a cell regulates concentration gradients.
  - As per the Michaelis-Menten mechanism, transporter-mediated diffusion starts out strong but eventually levels out at some $v_{\max}$ as the concentration of the transported molecule increases.
  - Simple diffusion and channel-mediated transport, however, increase linearly with the concentration of transported molecule indefinitely.

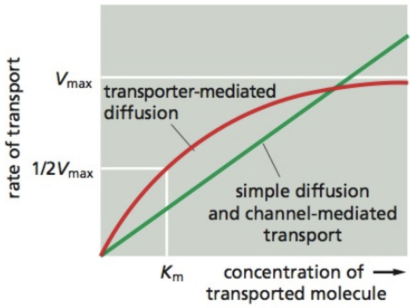


Figure 5.11: Concentration gradient regulation.

  - Thus, equilibrium is established when the rate of transporter-mediated diffusion in one direction equals the rate of simple diffusion and channel-mediated transport in the other direction (the intersection of the two lines on the right of Figure 5.11).

- Types of transporters.
  - Most transporters are **coupled transporters**.
  - Another kind is **ATP-driven pumps**.
  - Energy can also come from light, as in **light-driven pumps**.
- **Coupled transporter**: A transporter that moves two molecules either in the same or opposite directions. *Also known as* **cotransporter**.
  - **Symporters** and **antiporters** are the two types of coupled transporters.
- **ATP-driven pump**: The ATPase domain hydrolyzes ATP, providing energy to power a conformational change that allows binding and then transport from low concentration to high concentration.
- **Light-driven pump**: The energy for the conformational change comes from light instead of ATP.
- **Uniporter**: A transporter that only moves a single kind of entity.
- **Symporter**: A transporter that moves two different entities in the same direction.
  - Example: Transporting glucose using sodium. You need sodium as a co-transported ion to transport glucose.
- **Antiporter**: A transporter that alternates between taking one molecule into the cell and another out.
- An example of coupled transport: The sodium-coupled glucose transporter's steps.

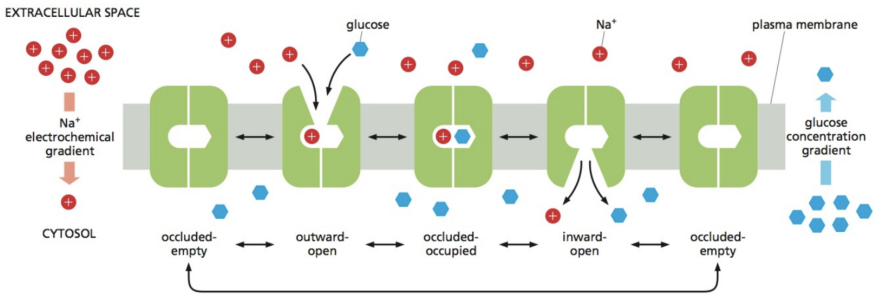


Figure 5.12: Sodium-glucose cotransporter activity.

1. Occluded-empty.

2. Outward-open. Sodium has a high K_d, so it binds readily. This induces a conformational change, generating a high-affinity glucose-binding site.
3. Occluded-occupied.
4. Inward-open. Sodium falls off first. This induces a conformational change, removing the high-affinity glucose-binding site and kicking glucose out.
5. Repeat.

- Key thing to remember about the plasma membrane: We tend to think of cells like HELA cells and HEK cells that are apolar, but real cells do have poles and specific regions with fundamentally different membranes.
  - For example, consider intestinal cells (I-cells). One side faces the gut and all of the bacteria therein (we are 98% bacteria by number of cells), the opposite side faces our vasculature (bloodstream) for easy deposition of nutrients, and the sides are bound to each other with protein velcro to keep the bacteria from getting into our blood (which would cause sepsis).
- In addition to directly inhibiting proteins, you can simply stop them from going where they need to, e.g., if you want to inactivate a transmembrane protein, simply stopping it from leaving the ER and getting to the plasma membrane will do the job.
- The bending of membranes allows very similar chemical structures to achieve different objectives on a "macro" scale.
  - Only 2-5% of total cell membrane is the plasma membrane.
  - Mitochondria contain a long, smooth, ellipsoidal outer membrane.
  - Mitochondria also contain a long, fenestrated inner membrane.
  - The endoplasmic reticulum's membrane has both flat and tubular regions. We don't know how the balance is decided, though.
- **Solute liquid carriers** (SLCs) sit on the cell surface and are involved in nutrient transport. Every SLC mutation results in a different disease. These are symporters. These are the next class of druggable molecules.
- What is the purpose of aquaporins?
  - These control the membrane tension, which is critical.
  - Work together with sodium and potassium transporters.
- Cystic fibrosis isn't concerned with water pressure; it's concerned with chloride concentration.
  - Yamuna's advice: If you want to lose weight, stop eating salt. Low external salt gradients make it harder to transport amino acids into cells.

5.3 Supplementary Sequencing Videos

PCR

From here.

11/9:
- Denaturation temperature (for splitting DNA strands): 95 °C.
- **Annealing** temperature (for binding primers to ssDNAs): 55-65 °C.
- Extension temperature (for DNA replication): 72 °C.
- **DNA annealing**: The process of forming heteroduplex DNA from two complementary (or nearly complementary) molecules or regions of ssDNA. *Also known as* **hybridizing**.
- From one single DNA molecule, 1,073,741,764 copies of the target DNA are obtained in only 4 hours.

Pyrosequencing

From here.

- First, we need a strand of DNA to sequence
- Step 1: Make the **library**.
  - Cut the DNA into fragments via **sonication** or **nebulization**.
  - Then the fragmented DNA strands are ligated with **adapters** at both the $5'$ and $3'$ ends.
  - At this point, we denature the dsDNA, generating a hybrid molecule that we can amplify with PCR in step 2.
- **Library**: A collection of DNA fragments that we store and clone.
- **Sonication**: A technique of shearing DNA involving exposing it to high sound frequencies to agitate it and cause it to break.
- **Nebulization**: A technique of shearing DNA involving forcing it through a small hole.
- **Adapter**: A short oligonucleotide.
- **Oligonucleotide**: A short single- or double-stranded DNA or RNA molecule. *Also known as* **oligo**.
- Step 2: Emulsion PCR.
  - We incubate the DNA with microscopic beads that are bound all around with oligos complementary to our adapters.
  - Thus, every ssDNA can anneal to the DNA capture beads.
  - Subsequent dilution ensures that each bead only has one strand attached.
  - Oil is added to the mixture forming an **emulsion** with the largely aqueous solvent.
  - This creates **blebs**.
  - We add a PCR mix (buffer, primer, polymerase, dNTP) to the blebs as well.
  - This allows us to amplify the DNA in all beads in parallel. Once DNAs are produced, the newly synthesized strands break off (they are not held to the bead via a sugar-phosphate backbone) and anneal to other complementary adapters on the bead in the bleb.
  - Repeating this process 30-60 times allows us to conjugate several thousand conpies of the same sequence to each bead.
  - We need many DNAs because our cameras are not sensitive enough to detect single pyrophosphate-induced photons; several million at the same time, though, is more than acceptible.
- **Emulsion**: A mixture of two liquids that aren't miscible.
- **Emulsion PCR**: A variation of PCR that some next-generation techniques use to replicate DNA sequences.
- **Bleb**: A microvesicle so small it can only hold one bead at a time.
- Step 3: Loading.
  - We break the emulsion to release the beads and deposit them onto a sequencing chip with tiny wells ($\sim 1/3$ the diameter of a hair, so every well can fit at most one bead).
  - It is important to immobilize the DNA onto the beads because we will have reagents flowing in and out of the well that can easily strip DNA away from the beads.
  - For the pyrosequencing reaction to take place, we will need to add enzymes like sulfurylase, luciferase, and apyrase as well as their substrates adenosine phosphosulfate (APS) and luciferion. We will also need some polymerase and primer because we will be replicating some DNA.

  - Once we're ready, the computer pumps A, T, G, and C into the wells sequentially, washing out before each new addition. Then repeat.

- Step 4: Pyrosequencing reaction.
  - As the computer is doing this, stuff is happening within the wells. The primers have bound to the DNA ends away from the beads and DNA polymerase adjacent to them.
  - DNA polymerases begin synthesizing new complementary strands using the dNTPs pumped in by the computer. Once the computer pumps in the right complementary nucleotide, DNA polymerase will add it. This is key.
  - DNA polymerase is stalled until it gets the right dNTP. Between each addition, apyrase degrades all previously added nucleotides. When the right dNTP is added, light is emitted, which we can measure.
  - How does the addition of this dNTP lead to the generation of light?
    - When the right dNTP is merged, a pyrophosphate (PPi) is released, as previously discussed.
    - Sulfurylase combines APS and PPi to generate ATP.
    - Luciferase[1] combines ATP and luciferin to generate oxyluciferin and the detected flash of light.
  - By plotting the sequence of light flashes vs. time, the original sequence can be decoded.

Illumina Sequencing

From here.

- Four steps: Sample prep, cluster generation, sequencing, and data analysis.
- Sample prep.
  - There are multiple ways to do this, but all of them do add adapters to the ends of the DNA fragments.
  - Then reduced cycle amplification allows additional motifs to be introduced such as the sequencing binding site, indices, and regions complementary to the flow cell oligos.
- Clustering.
  - Each fragment molecule is isothermally amplified.
  - The flow cell is a glass slide with lanes.
  - Each lane is a channel coated with a lawn coated in two types of oligos.
  - Annealing of the sample is enabled by the first of the two types of oligos; this oligo is complementary to the adapter region on one of the fragment strands.
  - A polymerase then creates a complement of the annealed fragment. The double-stranded molecule is denatured and the original template is washed away. Now we have a complete complement to the sample covalently bonded to the lane's surface.
  - At this point, the strands are clonally amplified through **bridge amplification**. The process occurs simultaneously for millions of fragments.
  - Now the reverse strands are cleaved and washed off, leaving only the forward strands.
  - The $3'$ ends are blocked to prevent unwanted priming.
- **Bridge amplification**: The following procedure.
  1. The strand folds over, and the non-covalently bound end anneals to the second type of oligo.

[1] The same enzyme that fireflies use to glow.

2. Polymerase creates a double stranded bridge.
3. The dsDNA is denatured and each product goes and bridges with other as-yet unstranded oligos.

- Sequencing.
  - With each cycle, fluorescently tagged nucleotides are incorporated.
  - Excitation by a light source causes a characteristic fluorescent signal to be emitted.
  - This process is called **sequencing by synthesis**.
  - The number of cycles determines the length of the read. The emission wavelength, along with the intensity, determines the base column. For a given cluster, all given strands are read simultaneously.
  - This allows hundreds of millions of strands to be sequenced in a massively parallel process.
  - After the completion of the first read, the read product is washed away. Then, an index-1 read primer is introduced and hybridized to the template. After completion of the index read, it is washed off.
  - The $3'$ end is deprotected, allowing for bridging. Index 2 is read in the same manner as index 1. Polymerases extend the second flow cell oligo to a double-stranded bridge.
  - After linearization of the bridge and blocking of the $3'$ ends, the original forward strand is cleaved off, leaving only the reverse strand.
  - Read 2 begins with the introduction of the read 2 sequencing primer. As with read 1, the sequencing steps are completed until the desired read length is achieved.
  - Then, the read 2 product is washed away.
- Data analysis.
  - This entire process generates millions of reads, representing all of the fragments.
  - All reads are more or less aligned, giving a full sequence.

SMRT Sequencing

From here.

- Attenuated light from the excitation beam penetrates the lower 20-30 nm only of each ZMW, creating the world's most powerful microscope (detection limit of 10^{-21} L).
- The tiny detection volume afforded by the ZMW provides 1000-fold improvements in the reduction of background noise.

Nanopore Sequencing

From here.

- Has applications to DNA, RNA, and protein sequencing.
- The membrane is electrically resistant and created from synthetic polymers. Thus, current flows only through the aperture in the nanopore.
- Intact DNA strands are analyzed by the nanopore in real time.
  - The nanopore sequences whatever fragment is presented to it, regardless of length, rather than generating reads of a specific length as with traditional cyclical sequencing chemistries.
- The DNA sequences are mixed with a processive enzyme. The enzyme is designed to attach to the top of the nanopore and ratchet the DNA through the nanopore one base at a time.

- The enzyme binds to a single-stranded leader at the end of the double-stranded DNA template and unzips the double strand, feeding it through the nanopore.
- The speed of the enzyme can be controlled.
- Once one DNA strand has been sequenced, another one will begin being sequenced.
- There is no deterioration of accuracy as the DNA strand is sequenced.
- If you prepare the dsDNA with a hairpin at the far end, you can read both complementary strands in one go, improving accuracy and giving other advantages in data analysis.
- You can sequence gDNA, amplified gDNA, PCR amplicons, and cDNA.

Week 6

Import and Export

6.1 Organelles and Transport

11/1:

- Warm-up activity: Quiz questions.
- This class and next class: Protein localization mechanisms and inhibition.
- Think about...
  - How a cell transports and localizes proteins;
  - Mechanisms of preventing things from going where they should.
- Indeed, some protein inhibitors work not by inactivating proteins but by making sure they don't get to the right place.
- **Chaperone**: A small molecule that allows a misfolded protein in the wrong place to fold and reach its site of action.
  - These are promising new drugs.
  - Example: There is a known risk gene for Parkinson's disease. A protein gets stuck in the ER. If there are small molecules you can use to get the protein to fold in the ER and be released, you win.
  - Example: Cardiovascular disease. Most drugs fail clinical trials right at the last stage of testing because they cause something called **long QT syndrome**.
    - Said drugs cause this syndrome by preventing ion channels from reaching the plasma membrane.
    - Chaperones could potentially help overcome this common barrier.
- **Arrhythmia**: A fast, chaotic heartbeat.
- **Long QT syndrome**: A heart signaling disorder that can cause arrhythmias.
  - Symptoms can be severe, up to death.
- Today: Mechanisms of protein localization.
  - How do proteins reach the nucleus, mitochondria, and a mystery organelle? These are open questions in basic biology.
- Organelles and membranes by the numbers.
  - Cytosol: 2% of the membrane in a cell, but 54% of total cell volume.
  - Thus, the plasma membrane spends a lot of energy keeping the cytosol happy.

  - The mitochondria and ER have a huge amount of membrane but very little volume and contribute to keeping the cytosol happy as well.
    - The membrane content helps maintain homeostasis.
  - What was the net point of all this??

- Evolution of compartments.
  - Helps us understand **topological equivalence**.
  - Yamuna believes that the endosymbiotic theory is just a hypothesis and that it's all up in the air and likely to change.
  - Current hypothesis: Archaea lost its cell wall making it easier for it to acquire DNA. Once it acquired enough valuable genes, the cell membrane underwent an invagination to form the nucleus and extra folds of the ER. This prevents the cell from losing the DNA its acquired.
    - This is why the ER and extracellular matrix are topologically equivalent, i.e., because the former evolved from the latter.
  - Mitochondria are the cell intaking another bacteria that could produce energy.

- **Topologically equivalent** (compartments): Two compartments inside (or outside) a cell such that materials do not have to cross a membrane to get from one to the other.

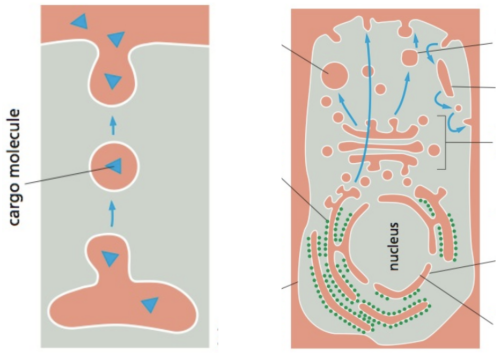


Figure 6.1: Topological equivalence.

  - Example: Golgi and ER are topologically equivalent (and topologically equivalent to the extracellular matrix) but not to the cytoplasm. This is because materials in the ER move to the Golgi and then to the extracellular matrix within a **vesicle**, i.e., they never have to cross a plasma membrane so much as they get surrounded and moved by different membranes.
  - Example: The extracellular matrix and the cytoplasm are not topologically equivalent. Notice that any material coming into the cytoplasm from the outside must cross through the plasma membrane using one of the mechanisms from last class (we're not talking endocytosis yet).

- Proteins can be transported between organelles either by being stuck in the membrane of a vesicle (at which point they will end up in the membrane of the target organelle) or within said vesicle's lumen (at which point they will end up in the lumen of the target organelle).

- Isolating organelles.
  - We discover transport mechanisms by carrying out a lot of mutations and then isolating specific target organelles and testing for a protein's presence (look for a ratio between the quantity of this protein present and a standard protein that you know will be there).

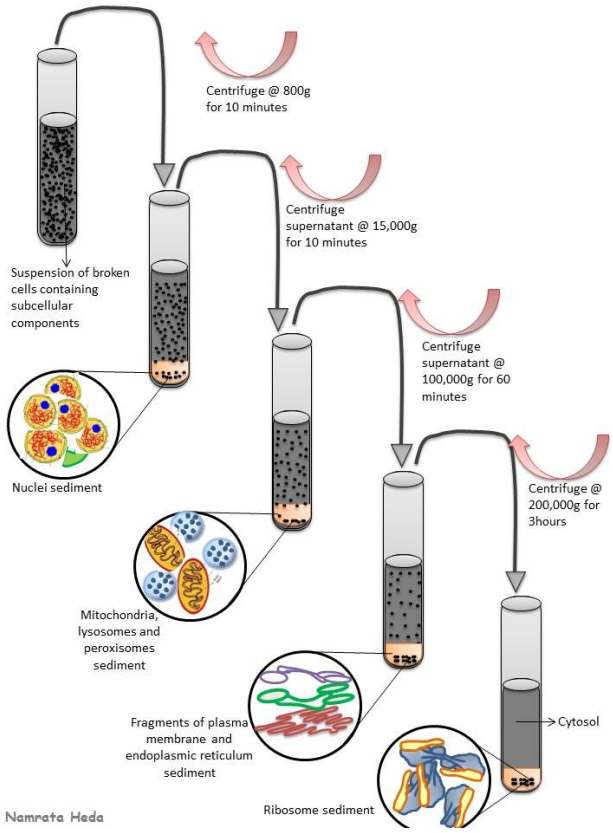


Figure 6.2: Isolating organelles.

  - Isolation mechanism: We take a bunch of cells, dissolve the extracellular membrane to get a soup of organelles, centrifuge it (to let heavy organelles like the nucleus fall out), centrifuge again (to get the small organelles like the mitochondria, lysosomes, and peroxisomes), centrifuge it again (to get fragments of the plasma membrane and ER), and centrifuge it one last time (to get ribosomes).
  - Alternative isolation mechanism: Use a matrix with a density gradient and just centrifuge once to get multiple layers.

- Three main ways to move proteins in a cell.
  - Recall passive and active exchange.
  - You can let physical equilibrium take hold, of course, but often that won't lead to great enough concentrations.
  - Thus, cells evolved the following three methods...

- **Gated transport**: A type of transport involving a gate and a condition (e.g., a binding or membrane potential) that must be satisfied for the gate to "lift open."

- **Translocation**: The movement between topologically nonequivalent compartment.
  - Example: Suppose you have a protein that's been made in the ER, has been deposited into the cytoplasm, and now needs to get into the mitochondria. We will consider this example in much greater detail shortly.

- **Vesicular transport**: The movement of biomolecules in vesicles between topologically equivalent compartments.

- **Localization sequence**: A molecular GPS. *Also known as* **nuclear localization sequence**, **NLS**.
  - Usually located on the N-terminus because it comes out first and needs to know where to go.
  - Localization sequences have different strengths. Some send proteins in a high fraction somewhere; some send proteins in a low fraction somewhere.
    - Strength is determined by the sequence's affinity for the transport protein. We will discuss this in more depth later.
  - Length: Tetrapeptides up to 20-30 AAs.
  - Very occasionally occur in the middle of a protein.
- **Translocation sequence**: A molecular GPS on the C-terminus that moves a protein after folding.
- Gated transport example: Movement from the cytosol into the nucleus.
  - This is also the most common type of gated transport.
  - The gate is the nuclear pore, and the condition is **karyopherin** binding
- **Karyopherin**: A protein involved in transporting molecules between the cytoplasm and the nucleus.
- Consider first the structure of the **nuclear pores**.
- **Nuclear pore**: A gateway from the cytosol to the nucleus.

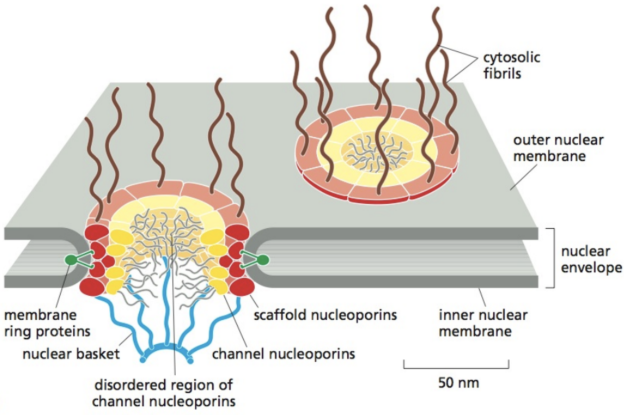


Figure 6.3: Nuclear pore structure.

  - Nuclei have double plasma membranes and nuclear pores. All transport in and out of the nucleus occurs via nuclear pores.
  - **Membrane ring proteins** make the membrane bend backward around nuclear pores.
  - There are about 3000 nuclear pores per nucleus.
  - About 1000 molecules transport both ways per nuclear pore per second.
  - Active v. passive transport: Anything smaller than 40 nm will freely diffuse to a significant extent (smaller implies higher passive transport). Larger, you need something to capture it and drag it through (this is active transport).
  - Nuclear pores are 8-fold symmetric bodies.
    - We still don't know the complete structure.
    - Composed of **nucleoporins**.
    - Hair-like **cytosolic fibrils** on the outside and a **nuclear basket** on the inside.
    - A porous plug in the center; still don't know what it is, but it's made of lots of FG repeats.

- **Nucleoporin**: A protein that is a constituent building block of the nuclear pore complex. *Also known as* **nap**.
  - There are permanent naps, but there are also naps which come off and on.
  - Approximately 30 exist.
  - Some are transmembrane.
- Probing NLS-enabled nuclear import.

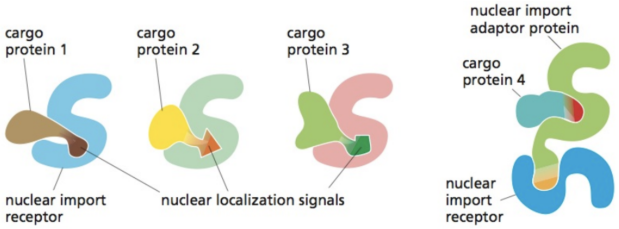
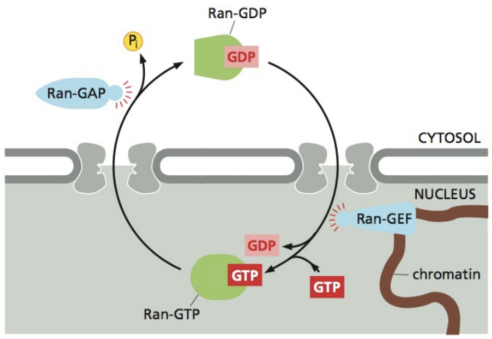


Figure 6.4: Nuclear import receptor binding.

  - Suppose we fuse an NLS with GFP (for which a Nobel Prize has been awarded).
  - If, in the NLS sequence, we change a K $\rightarrow$ T, then localization of the sequence is significantly decreased.
    - Review: Replacing positively charged lysine with polar threonine would certainly affect the interaction between the NLS and the **nuclear import receptor**!
  - Conclusion: We can affect NLS efficiency by altering one's affinity for its karyopherin (or vice versa), or by altering the ability of the karyopherin to enter the nucleus.
  - "It depends upon which bus you get on and upon the strength of your ticket."
  - Benefit of differential binding affinities: It is possible to have different concentrations of different proteins. You don't want all proteins in the nucleus to have the same concentration, after all.
  - There also exist **nuclear import adaptor proteins** which link cargo proteins to their nuclear import receptors with higher binding affinities.
- Nuclear export is the reverse of nuclear import.
  - Before we can discuss nuclear export directly, we should discuss the Ran proteins.



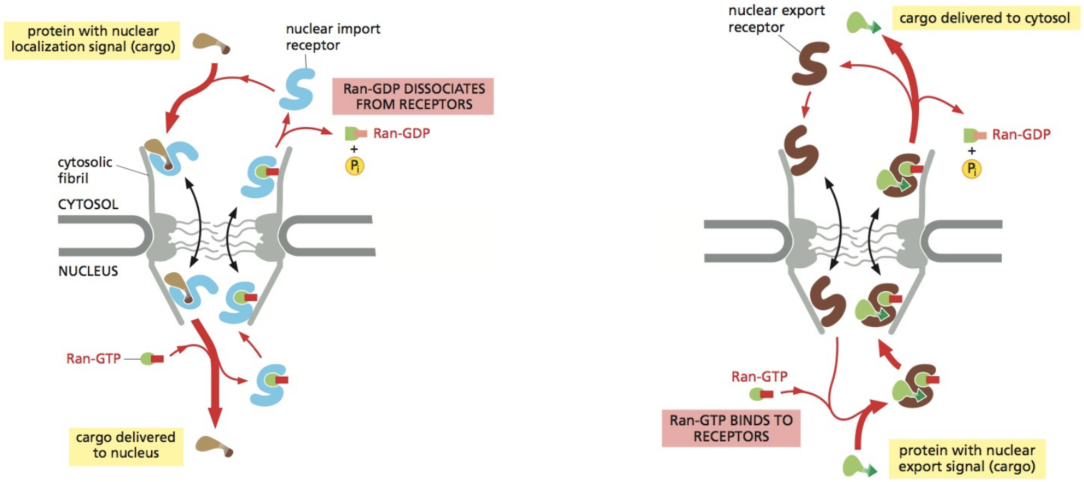
(a) The Ran proteins.

Figure 6.5: Nuclear import and export mechanism.

    - Ran complexes have a domain called a **GTPase domain**.
    - Ran's GTPase domain has GTP- and GDP-bound forms.

- There is a Ran-GDP / Ran-GTP gradient across the nuclear membrane: Ran-GTP is present in much higher concentrations within the nucleus, and Ran-GDP is present in much higher concentrations outside the nucleus.
- Ran-GAP is a **GAP** for Ran-GTP and Ran-GEF is a **GEF** for Ran-GDP.
- Ran-GAP is localized in the cytosol, and Ran-GEF is localized in the nucleus (it sits on chromatin inside the nucleus).

– We are now ready to discuss nuclear import and export.



(b) Nuclear import.

(c) Nuclear export.

Figure 6.5: Nuclear import and export mechanism.

- Nuclear importers first bind their proteins. Their hydrophobic regions are then caught by the cytosolic fibrils. Moving downward into the FG repeats, the importer's movement once inside is a random walk.
- Once an importer arrives in the nucleus, Ran-GTP attacks. Ran-GTP has a higher affinity for it than its substrate, so it will bind and cause the substrate to fall off, completing delivery to the nucleus.
- When the Ran-GTP-bound importer diffuses back out of the nucleus, Ran-GAP promotes GDP hydrolysis, and Ran-GDP dissociates.
- Nuclear export receptors random walk into the nucleus, bind a Ran-GTP, engage the cargo, random walk out of the nucleus, Ran-GAP hydrolyzes Ran-GTP to RanGDP which leaves, and this kicks out the cargo.

– Note that as we would expect for an example of gated transport, a condition is met and only then does transport occur.

- **GTPase domain**: A region of a protein that hydrolyzes a GTP to release energy, accelerating and powering the function of the protein.
  – Carried by many kinds of proteins and is very powerful.
  – Can help a ribozome work, help proteins move from the nucleus to the cytosol, promote vesicle fusing, etc.
  – Essentially functions as a backpack with a battery.
- **GAP**: A protein that promotes GTP hydrolysis in a GTPase domain that's already bound to GTP. *Also known as* **GTPase activating protein**.
- **GEF**: A protein that exchanges GDP for GTP at the GTPase domain. *Also known as* **Guanine nucleotide exchange factor**.

- Translocation example: Movement from the cytosol into a mitochondrion.
  - Mitochondria have proteins that sit specifically on the outer membrane, inner membrane, in the interluminal space, or in the center of the matrix. This indicates very high accuracy and targeting.
    - Many mitochondrial diseases occur due to poor localization.
  - The lessons here are broadly applicable.
  - Before next class, brush up on translation.
  - This is an example of **post-translational** protein transport.
  - There is also **co-translational** protein transport, but we'll talk about that another day.
- **Post-translational** (protein transport): Having a protein cross a membrane after its ribosome has finished synthesizing it.
- **Co-translational** (protein transport): Having a protein cross a membrane as it is still being synthesized by a ribosome.
  - Usually happens in the ER.
- There are four main mitochondrial proteins/complexes to consider: **TIM**, **TOM**, **SAM**, and **OXA**.

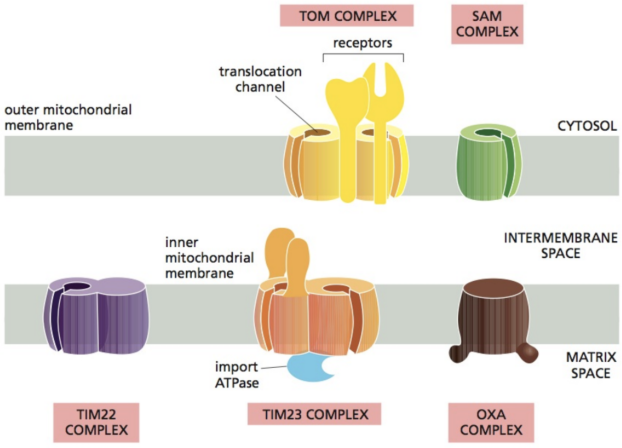


Figure 6.6: Mitochondrial translocators.

  - These are all big, multiprotein complexes assembled on the various membranes.
- **TOM complex**: The mitochondrial complex of proteins — localized in the outer membrane — responsible for the movement of proteins through this barrier and into the interluminal space. *Also known as* **translocase of the outer membrane**.
- **TIM complex**: The mitochondrial complex of proteins — localized in the inner membrane — responsible for the movement of proteins through this barrier and into the matrix. *Also known as* **translocase of the inner membrane**, **TIM23**.
- **SAM complex**: The mitochondrial complex of proteins — localized in the outer membrane — responsible for the folding/insertion of proteins into the outer membrane. *Also known as* **sorting and assembly machinery complex**.
- **OXA complex**: The mitochondrial complex of proteins — localized in the inner membrane — responsible for the movement of proteins through this barrier and into the matrix. *Also known as* **oxidase assembly complex**.

- There is a way to get TIM and TOM to lock together so translocation happens all at once from the cytosol to the matrix (instead of having to pass through the interluminal space).

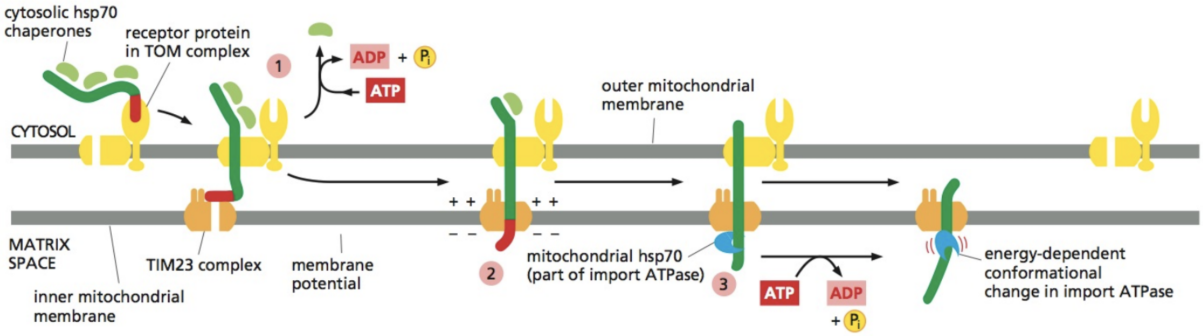


Figure 6.7: Translocation from the cytosol to the mitochondrial matrix.

  - Role of energy in protein import into the mitochondrial matrix.
    - Every ATP hydrolyzed at TOM causes you to pull the protein through by a couple of peptides.
    - Membrane potential drives TIM.
  - Once the whole protein has been pulled through TOM, TOM and TIM separate.
  - Once the translocation sequence has completely entered the matrix, a signal peptidase cleaves it, trapping the protein in the matrix.

- We now talk about how proteins are sent to each membrane.

- Sending proteins to the outer membrane.

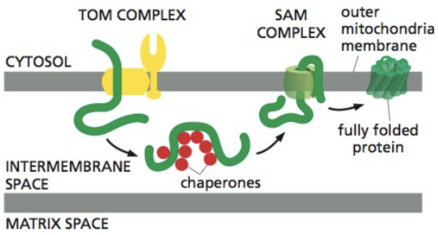
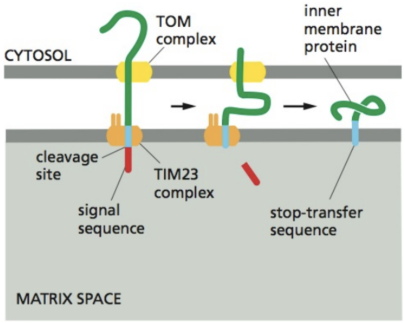


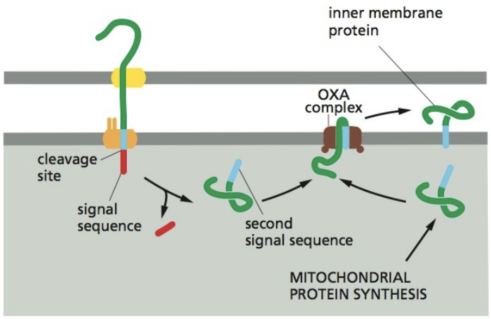
Figure 6.8: Translocation from the cytosol to the mitochondrial outer membrane.

  - Many porins are present on the outer membrane.
  - A protein first gets pulled into the interluminal space.
    - Aside: In bacteria, this space is known as the **periplasm**.
    - The periplasm is very nice for protein generation because there are very few proteins there and once you clone proteins there, all you have to do to release them is crack open the cell wall.
  - When proteins get pulled into the intermembrane space, they are all hydrophobic (because they will reside in a phospholipid bilayer eventually). Thus, they are prone to agregation in their water-based media, but chaperones latch on to separate them. Once stabilized in the intermembrane space, the protein then gets sent to SAM which folds it into the membrane. SAM has a slit, so as it pulls peptides in, it ejects them out laterally into the membrane.
    - Aside: In a bacteria, there is an analogous BAM complex.
    - Implication: This process is conserved between bacteria and mitochondria, further supporting the endosymbiotic theory.

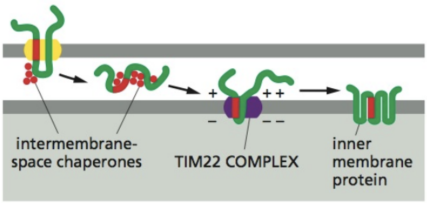
- Sending proteins to the inner membrane.



(a) Direct insertion.



(b) Via the OXA complex.

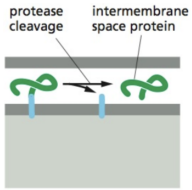


(c) Multipass proteins.

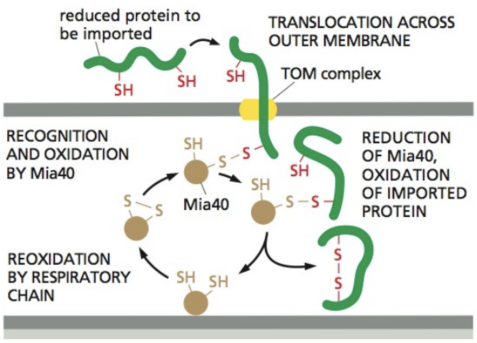
Figure 6.9: Translocation from the cytosol to the mitochondrial inner membrane.

  - There are three methods by which this can occur.
  - Method 1 (Figure 6.9a).
    - This method directly inserts single-pass transmembrane proteins into the inner membrane.
    - Translocation begins as if the protein is to be pulled into the matrix, but immediately following the localization sequence, there is a stop-transfer sequence. When TIM interacts with this, it stops pulling the protein through, signal peptidase cleaves off the localization sequence, and TIM ejects the hydrophobic stop-transfer sequence into the inner membrane.
    - Once TOM finishes pulling the bulk into the interluminal space and the protein refolds, we are done.
  - Method 2 (Figure 6.9b).
    - Use the OXA complex.
    - The TIM/TOM complex moves a protein into the matrix and a single peptidase cleaves off the localization sequence.
    - A secondary tag following the localization sequence then engages the OXA complex. The OXA complex flips the protein so that the tag is in the inner membrane and the bulk of the protein is in the interluminal space.
  - Method 3 (Figure 6.9c).
    - Multipass membrane proteins are introduced via the **TIM22 complex**.

- Import from the cytosol into the intermembrane space.



(a) Inner membrane cleavage.



(b) Via Mia40.

Figure 6.10: Translocation from the cytosol to the mitochondrial interluminal space.

  - If after insertion into the inner membrane, the bulk is cleaved from the transmembrane region, it will float away in the interluminal space. This is a secondary mechanism by which proteins enter the interluminal space, in addition to direct import by TOM.
  - A third (and very popular) mechanism leverages disulfide bonds. These help the protein fold, but if the protein is to be pulled through TOM, these will have been reduced to split them and unfold the protein. When the reduced disulfide bonds interact with **Mia40** in the interluminal space, they get reassembled and Mia40 gets regenerated (it is a catalyst). This refolding sticks the protein in place.

- Next time: Molecular mechanism of translocases; how they release transmembrane domains in the right origin.
- Peroxisomes.
  - We understand very little about their function, but if anything is wrong with them, it's deadly.
  - Take long lipid chains and cut them into shorter lipid chains by peroxidizing them (they have many reactive oxygen species).
  - Smallest organelle in the cell (50-100 nm) and has a very small number of proteins.
  - Peroxisomes are thought to be born from the ER via budding. Then the peroxisome must mature and acquire proteins, both in its membrane and in the lumen.
  - There are peroxisome targeting sequences, but who takes proteins to the peroxisomes and how they are transferred into the peroxisome is not clear.
  - Peroxisomes are thought to undergo fission for replication, but we have no way to distinguish early peroxisomes from mature, functional ones.
  - Peroxisomes carry their own catalase (which reduces oxygen into water and ROSs).
- Summary of today.
  - How compartments evolved; the evolution determines what is topologically equivalent to what. Nucleus and cytosol are equivalent (transport is facilitated by import and export factors), but most organelles are topologically equivalent to the extracellular matrix. You can take proteins to the nucleus or cytosol once they're born or to the mitochondria, or to specific places in the mitochondria.
- Chemists are the best inventers, but they don't have a very good understanding of cell biology.
- GTPs are used for big conformational changes.
  - The energy from hydrolyzing GTP and ATP is the same; it's just a question of how you use it molecularly.
- Yamuna is very knowledgable about a variety of topics in her field.

6.2 Quiz Prep

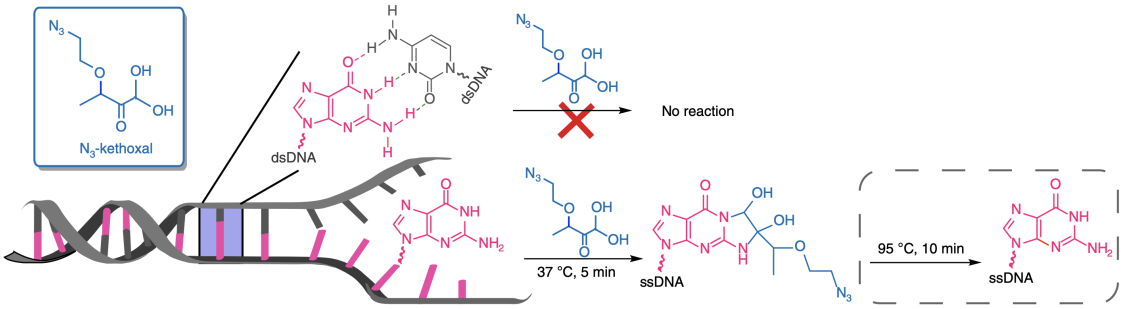
From Wu et al. (2020).

Notes

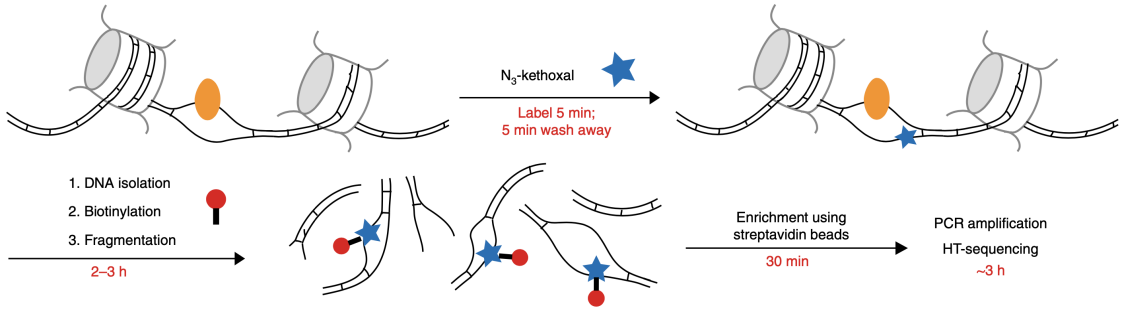
11/2:

- **Single-stranded DNA**: DNA that is not currently bound and hydrogen bonded into a double helix. *Also known as* **ssDNA**.
- **Transcription**: A highly dynamic process that generates ssDNA as transcription bubbles.
- **Transcription bubble**: A portion of the double helix that has been unwound and separated for the purpose of transcribing one strand of it.
- **KAS-seq**: Kethoxal-assisted single stranded DNA sequencing, the subject of this paper, provides rapid (within 5 mins), sensitive, and genome-wide capture and mapping of ssDNA produced by transcriptionally active RNA polymerases or other processes in situ using as few as 1000 cells.
- Kethoxal is a small molecule that rapidly and selectively binds with unpaired guanine.
- Attaching an azide group to kethoxal allows it to be tagged with bio-orthogonal click chemistry.
- Applications of KAS-seq.
  - Definition of a group of single-stranded enhancers that enrich unique sequence motifs.
    - Specifically, these enhancers are associated with the binding of specific transcription factors and exhibit elevated enhancer-promoter interactions.
  - Discovery: When **protein condensation** is inhibited, RNA polymerase II (Pol II) rapidly releases from a group of promoters.
  - Fast and accurate analysis of transcription dynamics and enhancer activities simultaneously in both low-input and high-throughput modalities.
- **Protein condensation**: Proteins sticking together.
- **Chromatin**: The material of which the chromosomes of organisms are composed, consisting of protein, RNA, and DNA.
- Transcription and its regulation (importance).
  - Determine physiological function and the cell's fate.
  - Regulation issues often lead to disease.
- **Global transcription regulation**: Regulation of transcription across the entire genome.
- How do we understand global transcription regulation?
  - Employ techniques like **ChIP-seq**.
  - Search for the presence and level of **nascent RNA**.
    - Based on **run-on assays**, **metabolic labeling**, and Pol II-associated or chromatin-associated RNA enrichment.
- **ChIP-seq**: A genome-wide sequencing approach that analyzes the occupancy of RNA polymerases.
- **Nascent RNA**: Newly made RNA, often still tethered to the DNA axis by elongating Pol II and being continuously altered by splicing and other processing events during its synthesis.
- **Assay**: An experimental method for assessing the presence, localization, or biological activity of a substance in living cells and biological matrices.

- **Run-on assay**: A method for measuring the frequency of transcription initiation. *Also known as* **nuclear run-on assay**. *Procedure*
  1. Take cells. At the time you want to measure the frequency of transcription initiation, freeze them.
  2. Reheat them and incubate at 37 °C in the presence of NTPs and radiolabeled UTP.
  3. Measure the amount of radiation given off by the products.
  4. The above measurement will be roughly proportional to the number of nascent transcripts on a gene at a certain time, which in turn is thought to be proportional to the frequency of transcription initiation.
- **Metabolic labeling**: The process of using the synthesis and modification machinery of living cells to incorporate detection or affinity tags into biomolecules.
- Limitations of the current methods for understanding global transcription regulation.
  - Run-on assays and enrichment require millions of cells as starting material.
  - Pol II ChIP-seq cannot distinguish whether RNA polymerases are simply bound or are actively engaged in transcription.
  - Metabolic labeling cannot measure low-abundance RNA species. Post-transcriptional processing can also alter results.
- If we want to understand global transcription regulation, then certainly it will be important to determine where transcription occurs.
- Goal: Locate where RNA polymerases engage in transcription.
  - Observation: RNA polymerases transform dsDNA to ssDNA bubbles as they move.
  - Method: Label/tag/identify/characterize ssDNA.
- Previous attempts: MnO_4^- preferentially oxidizes single-stranded thymidine residues.
  - Has been used to reveal Pol II-induced promoter melting locally and on a genome-wide basis.
  - Works together with S1 nuclease digestion.
  - Doesn't work on B DNA.
  - Limitations:
    - Requires tens of millions of cells.
    - Shows low sensitivity for weak/broad signals at Pol II elongation sites.
- Outline of the paper.
  - Describe KAS-seq.
  - Prove that it simultaneously measures the dynamics of transciptionally engaged Pol II, transcribing enhancers, Pol I and Pol III activities, and non-canonical DNA structures in which ssDNA plays a major role.
  - Prove that it works with as few as 1000 cells.
  - Prove that KAS-seq detects changes in transcription during quick environmental changes, e.g., inhibition of protein condensation.
- Note that most conclusions listed here have supporting data and correlation numbers given in the paper.
- Genome-wide profiling of ssDNA using N_3-kethoxal-based labeling.
  - Prior literature: Kethoxal reacts with the N1 and N2 positions of guanines (the ones that form Watson-Crick interactions) in ssDNA and RNAs under physiological conditions.



(a) N_3-kethoxal binding.



(b) Tagging and sequencing.

Figure 6.11: Locating ssDNA within the genome.

- This work: Attaching an azide "handle" to kethoxal to make N_3-kethoxal.
- Properties of N_3-kethoxal.
  - Retains high activity and selectivity for guanine.
  - Offers a bio-orthogonal handle that can readily be modified with a biotin or other FG.
- N_3-kethoxal effectively maps the secondary structure of RNAs by selectively labeling guanines in ssRNAs under mild conditions in live cells.
- Hypothesis based on this result: The scope can be expanded to selectively labeling ssDNA (Figure 6.11a).
- Verification of N_3-kethoxal's high labeling reactivity.
  - Run an in vitro labeling assay using a synthetic DNA oligonucleotide with exactly four deoxyguanosine bases.
  - 37 °C and 5 mins incubation labels all deoxyguanosine bases.
- Optimization of the KAS-seq N_3-kethoxal introduction conditions.
  - Reaction occurs with deoxyguanosine within 2 mins.
  - Reaction occurs with L-arginine within 10 mins.
  - Thus, 5 mins is a good time to both tag deoxyguanosine and minimize protein labeling.
- After labeling, **genomic DNA** is isolated and biotinylated through click chemistry.
- Enrich the fragments with **streptavidin** beads.
- Subject them to **library construction**.
- Remove N_3-kethoxal labels with a short heating at 95 °C.
- Perform a PCR amplification.
- The whole process takes about 1 day.

- **Genomic DNA**: Regular DNA inside the nucleus. *Also known as* **gDNA**.
- **Streptavidin**: A protein with an extraordinarily strong binding affinity for biotin.
- **DNA library**: A collection of DNA fragments that have been cloned into vectors.
- **Library construction**: The act of storing and/or propagating a DNA library in a population of micro-organisms through the process of molecular cloning.
- Control experiments (run on one million live HEK293T cells[1] and mouse embryonic stem cells [mESCs]).
  - Does N_3-kethoxal labeling affect gDNA isolation yield and purity? No.
  - Do we still observe biotin signals in the absence of either N_3-kethoxal or the biotinylation reagent (biotin-DBCO)? No.
- KAS-seq results are highly reproducible in replicate experiments.
- KAS-seq signals mark active transcription.
  - KAS-seq signals exhibit a similar distribution pattern to Pol II ChIP-seq signals along regions with different G/C contents. Indicates G-specific labeling isn't a major factor.
  - KAS-seq reads are very common at gene-coding regions, especially at gene promoters and transcription termination areas.
    - On the other hand, they are far less common at intergenic regions.
  - KAS-seq signals positively correlate with known histone modifications denoting active transcription, and negatively correlate with inactive chromatin markers.
  - KAS-seq also shows improvement over the permanganate method, particularly in the area of weak and broad ssDNA signals.
- **Transition start site**. *Also known as* **TSS**.
- **Transition end site**. *Also known as* **TES**.
- KAS-seq works with very small numbers of cells.
  - "Because of the high guanine labeling reactivity of N_3-kethoxal and the high affinity between biotin and streptavidin, KAS-seq is expected to maintain its sensitivity when using low-input starting materials or primary tissue samples" (Wu et al., 2020, p. 516).
  - KAS-seq signals remain unchanged when using 10,000, 5,000, or even 1,000 HEK293T cells.
  - KAS-seq retains its strong TSS signals but loses some of its gene body and TES signals when mouse liver tissue is used.
  - Low-input of cells still yields similar enrichment efficiency.
- KAS-seq reveals the dynamics of transciptionally engaged Pol II.
- Proof that what we're seeing is related to transciptionally engaged Pol II.
  - KAS-seq results correlate well with results from **GRO-seq** and Pol II ChIP-seq.
  - Experiments with inhibitors (DRB and triptolide) confirm that "the strong and sharp KAS-seq peaks on gene promoters reflect transcription initiation and pausing of Pol II near the TSS, and that KAS-seq signals at gene bodies are derived from transcription elongation" (Wu et al., 2020, p. 517).
  - Treatment of the cells with DRB before performing KAS-seq decreased peak numbers by 57% overall, primarily in the gene body and termination regions (signals went up at the TSS).

[1] A derivative of a common strain of immortalized human kidney cells.

        - This is the expected result, since DRB is known to inhibit Pol II release and keep it stuck at the TSS.
    - Treatment of the cells with triptolide before performing KAS-seq decreased peak numbers by 93%.
        - This is the expected result, since triptolide is known to inhibit Pol II being recruited to and loaded onto promoter regions.

- **GRO-seq**: Global run-on sequencing, which is the most widely used method to measure nascent RNA.

- What the dynamics of Pol II are.
    - KAS-seq data from the promoter-proximal and gene body regions revealed that there are four classes of genes: Those for which Pol II pauses in the promoter or doesn't pause and those for which the gene is actively transcribed or isn't.
        - This is consistent with previously reported GRO-seq studies.
    - KAS-seq data shows considerably enriched signals at the TES.
        - DRB removes these, so they are from Pol II elongation and pausing at the end, not some attaching-in-a-different-place artifact.
        - KAS-seq reads density on the terminal regions are all about the same, so KAS-seq doesn't exhibit length-dependent bias.
        - KAS-seq gives a higher **termination index** than Pol II ChIP-seq and GRO-seq, suggesting Pol II accumulation at the TES is greater than previously expected.

- **Termination index**: The ratio of the reads density at the TES downstream regions relative to the density in the promoter-proximal regions.

- **RNA polymerase I**: The RNA polymerase that transcribes the 5.8S, 18S, and 28S rRNAs. *Also known as* **Pol I**.

- **RNA polymerase III**: The RNA polymerase that transcribes the 5S rRNAs, tRNAs, and some small RNAs. *Also known as* **Pol III**.

- KAS-seq detects Pol I- and Pol III-mediated transcription events and non-B form ssDNA structures in the same assay.
    - Pol I- and Pol III-mediated transcription events are detected with Pol II ones as expected.
    - These two do not respond to DRB or triptolide.
    - Only about 2/3 tRNAs are actively transcribed, hinting at a transcription-level regulation of codon usage.
    - Several KAS-seq peaks could not be paired to Pol I- or Pol III-mediated transcription events under DRB and triptolide conditions.
        - Hypothesis: These could be from other DNA forms and telomeric DNA.
        - Test: Used a previously reported method to predict where non-B form DNA species might exist in the genome and looked for overlaps with their mystery regions; found many.
        - Takeaway: Using KAS-seq to study other ssDNA-involved biological processes could be a cool avenue to pursue in future research.

- Many enhancer regions are single-stranded, which correlates with higher enhancer activity.
    - Since Pol II is known to bind at certain enhancers, we can use KAS-seq to identify enhancers that are being transcribed by Pol II.
    - Identified **ssDNA-containing enhancers** under DRB conditions to focus on the TSS region.
    - In mESCs, 25% of enhancers are SSEs.
    - Two SSE subtypes: KAS-seq signals span the whole enhancer, and the signals don't.

  - SSEs include 94% of super-enhancers.
  - Genes associated with SSEs show higher expression levels.
  - SSEs possess much more long-range interactions, indicating that these transcribing enhancers may possess a stronger capability to activate their target genes.
  - SSEs enrich unique sequence motifs (??). Thus, they have distinct sequence features and transcription factor (TF) binding potentials.
  - Comparison of SSEs and enhancers with high TF binding.
    - ATAC-seq-positive enhancers are readily accessible.
    - 50% of these show no or very weak KAS-seq signals in mESCs.
    - Genes associated with the KAS-seq-positive group show a higher expression level.
    - There is a distinction between SSEs and motifs that are ATAC-seq-positive but KAS-seq-negative.
  - Pol II, histone modifications, and other transcription regulatory proteins are enriched on the SSEs.
  - In HEK293T cells, the ratio of SSEs to general enhancers is lower, but all characteristics (overlap with super-enhancers, DRB response, and correlation with transcription regulatory proteins) are preserved.
  - SSEs possess distinct genomic features and unique TF-binding footprints, as per our KAS-seq analysis.

- **ssDNA-containing enhancer**. *Also known as* **SSE**.

- **Protein condensate**: A highly dynamic structure formed through interactions between mediators, TFs, and other transcription coactivators that have been shown to incorporate Pol II to activate transcription.

- ssDNA dynamics upon the inhibition of protein condensates.
  - 1,6-hexanediol dissociates protein condensates.
  - The longer we let HEK293T cells sit in it, the more the KAS-seq signals diminish, supporting a role of protein condensate formation on transcription activation.
  - Novel observation: After 5 mins, there is an increase in ssDNA clustered around the TSS.
    - Leads to a slightly increased signal on the gene body and a coinciding decrease in signal at the TSS.
    - The clusters form in both directions for bidirectionally transcribed genes and downstream, only, for unidirectionally transcribed genes.
    - As time goes by, the clusters moved toward the TESs and gradually diminished.
  - Findings validated by Pol II ChIP-seq. KAS-seq even outdoes it in some places (e.g., detection of the above **fast-responsive genes**).

- **Fast-responsive gene**: A gene with significant ssDNA cluster formation in the TSS region at 5 minutes.

Q & A

1. The reason that KAS-Seq works on just 1000 cells as opposed to competing methods (e.g., ChIP-seq) that need millions of cells is:
   - Kethoxal is highly reactive and specific to guanines.
   - "Because of the high guanine labeling reactivity of N_3-kethoxal and the high affinity between biotin and streptavidin, KAS-seq is expected to maintain its sensitivity when using low-input starting materials or primary tissue samples" (Wu et al., 2020, p. 516).

2. How were the authors able to assign opened DNA structures to transcription and not replication?
   - Experiments with inhibitors (DRB and triptolide) confirm that "the strong and sharp KAS-seq peaks on gene promoters reflect transcription initiation and pausing of Pol II near the TSS, and that KAS-seq signals at gene bodies are derived from transcription elongation" (Wu et al., 2020, p. 517).
3. Why do the authors incubate cells with kethoxal-N_3 for such a short time (5 minutes) when incubation for a longer time will capture more ssDNA while the polymerase is transcribing?
   - Kethoxal-N_3's high binding affinity for guanines means that it will almost immediately attach to the target species, i.e., not more than 2 minutes is really needed. Additionally, given enough time (circa 10 minutes), it will begin to attach to other species, such as L-arginine (which also has two adjacent nitrogens). This leads to undesired tagging of proteins.
4. Which other purposes can kethoxal-N_3 be used for?
   - "Provides an effective way to map RNA secondary structures by labeling guanines in single-stranded RNAs under mild conditions in live cells" (Wu et al., 2020, p. 515).
5. How did the authors show that the background contribution upon subjecting cells to KAS-seq was negligible?
   - "KAS-seq performed in the absence of N_3-kethoxal or the biotinylation reagent (biotin-DBCO) resulted in negligible biotin signals shown by dot blot, nor sufficiently enriched DNA for library construction, suggesting minimum background of KAS-seq" (Wu et al., 2020, p. 516).
6. Why do the authors use excess kethoxal with a short reaction time rather than a small amount of kethoxal incubated for a long time?
   - See 3.
7. Glyoxal and methyl glyoxal are known cellular metabolites. Their accumulation is known to be disease causing. Can you explain how a disease might be caused?
   - Inhibiting protein condensation leads Pol II to rapidly release from a group of promoters (as if the cell fears using up all of its energy when there might not be as much around).
   - Likewise, perhaps it is possible that when there is too much protein, Pol II becomes hyperactive, consuming too much cellular energy, leading to oxidative stress and perhaps cell death.
   - Lead to more reactive oxygen species, hence more oxidative stress.
8. What is the major advantage of a 5-minute kethoxal exposure, i.e., giving the cells a "pulse" of excess kethoxal that is then washed away?
   - See 3.
9. KAS-Seq scores over ChIP-Seq in terms of its ability to work with frozen tissue samples. Why? Bear in mind that in frozen samples, the transcriptional bubbles remain.
   - "Because of the high guanine labeling reactivity of N_3-kethoxal and the high affinity between biotin and streptavidin, KAS-seq is expected to maintain its sensitivity when using low-input starting materials or *primary tissue samples*" (Wu et al., 2020, p. 516).
   - Freezing denatures proteins, but does not alter transcriptional bubbles.
10. Which one(s) of the following descriptions is/are correct when we compare KAS-seq and ATAC-seq?
    - "Notably, KAS-seq signals correlate better with H3K36me3 than ATAC-seq results do, indicating that while ATAC-seq serves as a powerful tool to probe chromatin accessibility, KAS-seq directly measures transcription activities" (Wu et al., 2020, p. 516).

  - ATAC-seq is a tool to probe chromatin accessibility more broadly.
  - Out of all ATAC-seq-positive enhancers, 50% showed no (or very weak) KAS-seq signals. Thus, since KAS-seq is highly specific for ssDNA, this must mean that 50% of ATAC-seq-positive enhancers are composed of dsDNA. Indeed, KAS-seq is more selective for ssDNA then ATAC-seq.

Proposed Answers

- 1-2, 2-1, 3-1, 4-(1),4, 5-4, 6-3, 7-3,(4), 8-5, 9-3, 10-4.

6.3 Co-Translational Protein Transport

11/3:

- Today: Co-translational protein transport.
- The lumen of the ER is 10% of the volume of the cell.
- Any protein that needs to be secreted from the cell or is a membrane protein (which is about 30%) is born in the ER.
  - We have various ways to move proteins to where they need to go.
- The ER is the most structurally and functionally diverse organelle in the cell.
  - It has big flat sheets and tubular regions.
  - Very dynamic — sheetlike regions can become tubular and vice versa.
  - You can look at the ER via either electron microscopy or, now, GFP.
  - There are the rough and smooth ER. The rough one is so named because it has ribosomes on its surface.
  - Rough ER is where proteins are actively translated. Smooth is where vesicles full of proteins bud off and go to the golgi.
  - The cytosol is a huge bank of calcium. It contains 10,000 times higher concentrations than the extracellular matrix.
    - This signals to bacteria and viri that they have entered a cell.
  - However, the ER's level of calcium is comparable to the extracellular matrix.
  - A cell cannot tolerate so much calcium in the cytosol for a long time because **excitotoxicity** will take hold.
  - Thus, any release of calcium into the ER must be very coordinated.
- **Excitotoxicity**: Having a cell get too excited and expend so much energy that it dies.
- Isolating ER membranes.
  - The ER is very fragile — as soon as you remove the plasma membrane (e.g., by sonication), the ER breaks up into **microsomes**.
  - When you do organelle isolation, one of the smallest/last things you centrifuge out are the microsomes (see Figure 6.2).
- **Microsome**: A small fragment of the ER membrane, possibly containing ribosomes.
  - There are two types of microsomes: smooth and rough, depending on which ER they came from.
  - Very important to understanding how proteins are transported.
- There are two kinds of proteins present in the ER: soluble proteins (present in the lumen) and transmembrane proteins (not fully water soluble).

- Targeting proteins into the ER lumen.
  - Primary experimental verification of such transport.
    - Take an mRNA that will be secreted by the cell.
    - Do an *in vitro* translation in the absence of microsomes.
    - This protein is slightly larger than the one created in the presence of microsomes.
    - Thus, there must have been a small bit of the protein that got chopped off in the microsome.
    - This led to many more experiments, culminating in the following plausible model.
  - As an mRNA is being transcribed, its signal sequence gets associated with a translocator. The signal peptidase either chops off the signal sequence and then the protein is pushed through, or the protein is pushed through and then it loses its signal sequence.
  - Even as a translocator opens to receive a growing protein, it is water- and calcium-tight.
- What brings ribosomes to the ER when translation begins?

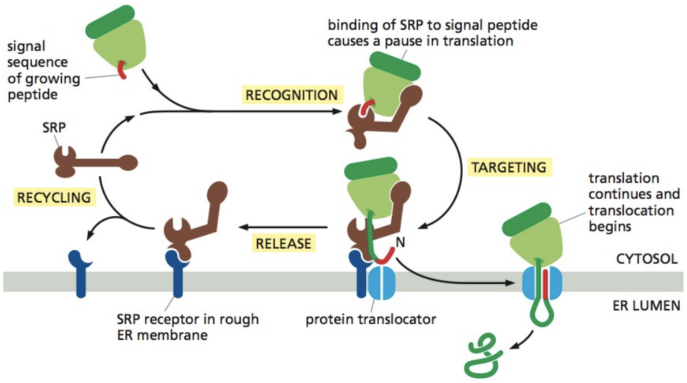


Figure 6.12: Signal recognition particle mechanism.

  - A signal recognition particle (SRP) is an RNA protein complex that binds the emergent signal sequence.
  - This induces a conformational change in the SRP (it hinges in the middle) that stops the ribosome from translating until the signal recognition particle takes the ribosome to the ER membrane.
  - How one signal recognition particle recognizes all signal sequences: It has a hydrophobic pocket that the AA signal sequence fits in.
    - Think of the pocket like a ball of putty — no matter which pencil you stick in, you will lift up the putty.
    - However, the hydrophobic pocket is not too tight; thus, it can let go. In fact, it is very flexible and molds around the signal sequence.
  - Recall from the translation lecture that at some point, the ribosome has to hydrolyze GTP to allow us to get to the next codon. This is what the SRP inhibits.
  - Upon binding to a signal sequence, the SRP unmasks another binding site which can bind to SRP receptors on the ER membrane.
  - This binding induces yet another conformational change that attracts a nearby protein translocator.
  - When the translocator binds, a conformational change allows the signal peptide to get jammed into the translocator, causing the ribosome to fall off of the SRP and resuming translation.
  - Now free of the ribosome, the SRP disengages from the SRP receptor and goes and looks for another free translating ribosome in the cytosol.

- Translocation on microsomes.
  - The translocator is made of two parts and opens only on one side.
  - You have a water-filled channel through which the polypeptide passes.
  - The translocator is called the Sec61 complex (Sec for secretion).
  - The channel is usually blocked by hydrophobicity (to be water- and calcium-tight).
  - When the peptide passes through and the signal peptidase cuts the molecule, the seam opens and pushes out the protein.
- Anchoring lumenally translated proteins to the membrane.
  - Three ways in which proteins are inserted into the ER membrane, depending on where in the protein the stop-transfer sequence is.
  - We can control the orientation of proteins in the membrane because we understand how it is put there so well.
  - Consider a protein that has a single transmembrane domain.
    - Such a protein must have (1) a signal sequence (2) at the N-terminus.
    - The start-transfer signal begins taking the protein into the ER, and then when the translocator hits a stop transfer, it stops and ejects the protein into the membrane.
    - There are well-known start and stop sequences. A stop sequence is typically also a hydrophobic transmembrane domain.
    - The start domain breaks off, and then the rest of the protein is synthesized in the cytosol.
    - The region between the start-transfer sequence and the stop-transfer sequence exists within the ER, and the region past the stop-transfer sequence exists outside the ER.
  - Suppose the start sequence is located in the middle of the protein AA sequence.
    - Two possible orientations: N-terminus in the ER and C-terminus in the cycoplasm, and vice versa.
    - Start means go to the ER and start translating; that's it.
    - How it gets oriented depends on which side is more positive and which side is more negative. More positive residues face the predominantly negative cytosol.
      - ➢ This allows us to control which side of our protein gets localized where.
      - ➢ Possible test question: How do I position a GFP in the cytosol but embedded on the ER membrane?
  - Multi-pass transmembrane proteins.
    - Start transfer is pulled in, goes along until you reach the stop codon, ejected.
    - Then the ribosome translates into the cytosol until the next start gets pulled into a translocator.
    - Signal peptidase does not chop at the end of every stop sign here; why is not known, but there must be some kind of "final" stop sign.
- Post-translational protein translocation: Signal sequence is on the C-terminus.
  - No SRP needed here.
  - Instead, we have a Get pathway.
  - **Snare proteins** (Nobel Prize 2013), which are very important in vesicle fusion and fision, decide how organelles come together.
  - Before the Get pathway was discovered, people hypothesized that proteins ram themselves into the ER membrane, but there were many holes in this theory (e.g., why not ram into the plasma membrane?).
  - The Get pathway is formed by 3 proteins: Get1, Get2, and Get3.

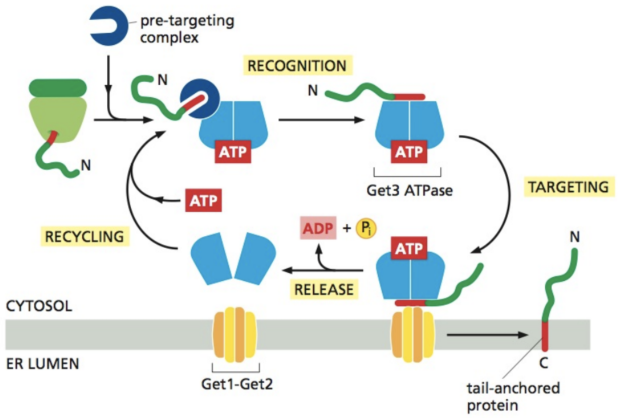


Figure 6.13: Post-translational protein translocation: Get pathway.

  - Get1, Get2 only recognize Get3.
  - Get3 is the key: It gets the C-terminus signal peptide, but only after it has been bound to a pre-targeting complex, inducing a conformational change that allows the pre-targeting complex to bind to specifically the ATP-bound form of Get3.
  - Get3 is an ATPase, meaning that it hydrolyzes ATP.
  - The pre-targeting complex binds the signal peptide sequence to Get3 ATPase; the mechanism is not well understood.
  - Once this has happened, Get3 is competent to deliver the sequence to Get1-Get2, which is located in the ER membrane and functions as a translocator.
  - In order to shove the sequence into Get1-Get2, we need energy; this energy comes from the hydrolysis of ATP to ADP + Pi.
  - At this point, Get3 is ready to bind another ATP and restart the cycle.
  - Contrast this with the GTPase cycle. Think about how an ATPase differs from a GTPase.

- Glycosylation (overview).
  - Many proteins need to be glycosylated (this will be the subject of next class).
  - Additionally, many proteins are anchored to the ER membrane not by a transmembrane region but by a lipid. How is this related??
  - Start with a complex sugar (we don't need to know the details), which is stuck onto a lipid (in particular, a steroid) called **dolichol** (which is a cholesterol) with phosphates.
  - This sugar gets transfered to any protein which gets glycosylated.
  - The transfer is carried out by oligosaccharide transferase.
  - The sugar ends up on an Asn side chain.
  - A specific signal leads to glycosylation at a specific Asn; in particular, you need a serine or threonine, then an arbitrary amino acid, then your asparagine.
  - The sugar is transferred *en bloc* (from French: all at once).

- GPI anchors.
  - These are lipoproteins.
  - Glycosylphosphatidylinositol: A lipid, phosphate, and many inositol groups (sugars).
  - A transamidation occurs, breaking the peptide bond connecting the transmembrane region of a protein to the rest of the protein and connecting the rest of the protein to the GPI anchor.

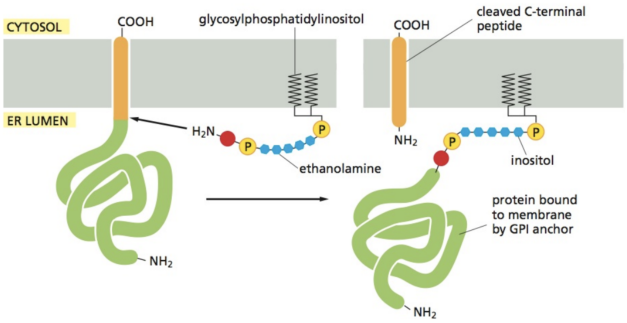


Figure 6.14: GPI anchoring.

  - Thus, it is the sequence at the border of the transmembrane region (a GPI-anchoring signal) and the rest of the protein that decides whether or not a protein will have a GPI anchor.
  - Once a protein becomes GPI-anchored, it gets moved to the extracellular surface of the cytosol.

- Differences between ATPases and GTPases.
  - GTPases: You need GAPs and GEFs.
  - ATPases: The rate limiting step (this is very important) is the dissociation of the NDP (usually ADP).

Week 7

Bulk Transport

7.1 Exocytosis and Endocytosis

11/8:

- Two parts of lecture today:
  - How a protein that is formed in the ER reaches the plasma membrane.
  - How proteins in the plasma membrane get to other parts of the cell (endocytotic pathway/other plasma membrane locations).
- Krishnan has really enjoyed teaching this class :)
- ER to Golgi transport.
  - The golgi is the pathway to the plasma membrane.
  - We don't know too much about the Golgi.
  - Glycobiology.
  - Will become one of the most important organelles in the future because it's very important and we don't know that much about it.
  - The Golgi is very hard to model (it's a stack of pancakes, and these are hard to distinguish).
  - Proteins that go to the plasma membrane and lysosomes and get secreted all have to pass through the golgi.
- COPI and COPII coated vesicles.
  - How a protein starts its journey.
  - If there's a protein in the ER lumen, there's a massive sugar transferred from dolichol to a particular asparagine.
  - Every protein that get's secreted gets so labeled.
  - Don't worry about what happens to that sugar rn, but it's like an assembly line (different steps of the packing process, all delocalized).
  - If you have a huge amount of transport proteins, you're going to have errors (things getting sent out that shouldn't), so you need a way to bring them back.
  - When the ER buds, we have to concentrate the proteins inside it.
  - The vesicle is covered in a protein called COPII (the COPII complex) that signals it as outbound from the ER.
  - ER-Golgi intermediate compartment does exist.
  - Proteins that shouldn't have left get sent back (in vesicles coated with COPI).

  - Anything in the ER (most proteins that need to be secreted) need to fold.
    - There are chaperones in the ER (such as BIP, calnexin) that help other proteins fold.
    - Many big proteins (90% of them) don't fold properly, so proteins need ways to make sure that they're only excreting the right proteins.
    - The ER is host to many unfolded proteins. Chaperones recognize anything that is misfolded and mark them for degradation.
  - How do proteins get into the vesicle?
    - Proteins have an exit signal that interacts with a cargo receptor.
    - Cargo receptors cluster on the surface.
  - Now the patch buds out.
  - These receptors usually have a lumenal domain and a cytosolic domain.
  - Adapter proteins recognize a bound form and cause proteins to cluster. COPII then assembles on the outside of the vesicle.

- The basics of vesicular fusion.
  - A lot of fission and fusion today.
    - Big in neuroscience and neurobiology, but also occur in cell biology.
    - These occur because of t-SNAREs and v-SNAREs (allow vesicles to fuse).
  - Vesicle fusion: **Homotypic fusion** and **heterotypic fusion**.
    - You also have **N-ethylmaleimide sensitive factor**.

- **Homotypic fusion**: The fusion of two like membranes.
  - E.g., two lysosomes fusing to form a bigger lysosome.

- **Heterotypic fusion**: The fusion of two different kinds of membranes.
  - E.g., a COPI-coated vesicle and the ER.

- **Vesicular SNARE**: Occurs on every organelle. *Also known as* **v-SNARE**.

- **Target SNARE**: The specifying SNARE. *Also known as* **t-SNARE**.

- **N-ethylmaleimide sensitive factor**: *Also known as* **NSF**.
  - Pries apart t-SNAREs and v-SNARES, allowing fusion.
  - The two helical domains will then move apart.

- KDEL is an ER-retrieval sequence.
  - How do we send proteins that accidentally localized to vesicles back?
  - Original way out:
  - Bulk-phase endocytosis: Letting the vesicle fill up naturally and the sending it back; not specific.
  - Alternatively, you can attract your proteins to the future vesicle (receptor-mediated endocytosis).
  - The purpose of the receptor is to concentrate your substance.
  - How does cargo that's exited this way get sent back?
  - KDEL receptors bind the sequence KDEL in a protein. A bit in the golgi, but most of it is in the ER. But its function is to grab ER retrieval sequences on the N- or C-terminus. pH dependence.
  - How will an ER retrieval sequence differ from a localization sequence?
    - The fundamental difference is one is regulated, and the other is spontaneous.
  - What we still don't know: How does a vesicle get switched from a COPII coat to a COPI coat?

- Spatial position of the Golgi in the animal and plant cell.
  - 5-6 stacks in human cells.
  - 100-200 stacks in plant cells (plant cells have to secrete a huge amount of material to maintain a cell wall).
  - More stacks (cisterni) lead to a more advanced golgi.
  - Cisterni are well below the wavelength of light, so we need electron microscopy.
- Molecular compartmentalization of the Golgi apparatus.
  - Our protein gets dolichol-marked.
  - Different reactions in different cisterna.
  - Different enzymes in each cisterna.
  - Each location in the Golgi acts on the protein differently before its eventually secreted due to the difference in enzymes stored in each region.
  - Cis-Golgi: beginning of the Golgi (closer to ER).
  - Stack.
  - Then trans-Golgi (end; opposite side from the ER).
  - Krishnan goes over an experiment showing what's localized in what compartment (different tagging proteins tag different enzymes, revealing localization in an electron micrograph).
- Oligosaccharides processing in the Golgi.
  - N-acetylglucosamine (GLcNAc).
  - Mannose (Man).
  - Galactose (Gal).
  - Sialic acid (NANA).
  - These all get attached. How far along a cargo has gone is decided by the sugar ordering.
  - Initial trimming of mannose and glucose. Trimming takes off a lot of these sugars, and then we replace with specific sugars (specifically those 4 above).
  - When a protein/lipid that's gone through the entire process reaches the plasma membrane, it will be able to show 2-types of sugar: Complex oligosaccharides (with a high concentration of negatively charged sialic acid at the end) and high-mannose oligosaccharides that do not get sialic acid added because the sugars arranged on the protein are inaccessible.
- What is the purpose of glycosylation?
  - Most important slide of the first part of the lecture!
  - In the ER lumen, we have two initial enzymes (glucosidase I and II) that chop off glucoses.
  - Manosidase in the ER takes away mannoses.
  - Leaves behind a sugar that's good to go to the Golgi.
  - Golgi mannosidase takes off 3 mannose residues at a time (the accessible ones).
  - In the medium golgi: We start adding GlcNAc, then add galactose, then silylation.
  - GlcNAc is added by N-acetylglucosamine transferase I.
  - Something attached to gludine, which is a good leaving group (highly anionic).
  - All these glycosylated molecules are present inside the comparment.
  - How does localization happen?
    - We have membrane proteins that will take in particular molecules and will work with them in one particular compartment.

  - Another mannosidase event (Golgi mannosidase II).
  - Now proteins are Endo-H resistant.
  - Addition of 2 more GlcNAc molecules, then galactose, then silylation. These give us our complex oligosaccharide.
  - The rules of the molecule Endo-H (of bacterial origin) tells you how far a sugar has gone in its journey. Helped us figure out many enzymes and transporters involved in the pathway.
  - You can work out the molecular weight of a protein.
  - High gets sent one place, low gets sent another place. This helped us investigate stuff.

- Two models of Golgi-protein transport.
  - We still don't know this (it's being researched, largely at UChicago).
  - What is the model of protein transport in the Golgi?
  - Does a *cis*-Golgi gradually mature (Golgi cisterni "grow up", gaining/losing proteins along the way), or do we have proteins transferred between fixed cisterni. The other one has vesicles budding out to either go forward to the next cisterni or back to the previous cisterna.
  - **Cisternal maturation model** vs. **vesicle transport model**.
  - If you want to know which is currently winning, write to UChicago's Ben Glick :)

- What keeps the golgi together (why are all of our pancakes stuck together)?
  - There are hydrophobic tentacles called golgins that wind together and prevent cytosolic fluid from getting between the pancakes.
  - At the time of cell-division or during apoptosis, we need to disentegrate the golgi because we can't have ?? hanging around.
  - This is induced by a kinase which phosphorylates the golgins, causing fragmentation.
  - Once the cell membrane comes back, the golgi reform and you get new ones.
  - When a cell divides, its organelles must divide, too, and we only have one golgi.
  - We build a new cisterni atop the old one.

- Onwards and outwards: Exocytosis and secretion.
  - There is a basal level of secretion that happens all the time, and there is regulated secretion where you have to release a massive amount of something all at the same time.
    - Regulated example: Insulin.
  - What needs to be secreted **basally** (all the time)? Mucus!

- Protein sorting at the TGN.
  - Once something comes to the trans-golgi network, where can it go? To the lysosome, outside the cell, and constitutive (e.g., placing things in the cellular membrane).

- Secretory vesicle maturation.
  - Very important!
  - How do we send out a huge amount of glucose, or melanin?
  - We concentrate molecules/proteins into dense core secreted granules (100-200 nm), extremely high concentration.
  - Longer peptide has a secretion clock. At each point, you have a pausing condition.
  - You take advantage of processing to pack tight.

  - Recall the tight junction from the first lecture (intestinal cell). Different parts of a membrane have different properties. Non-leaking proteins. You need a way to get proteins to exactly one part of the plasma membrane (how this works still isn't understood very well).

- Sorting plasma membrane proteins in polarized cells.
  - Two models: Direct and indirect sorting.
  - One model is different vesicles go to different locations. The second is you get random input into the plasma membrane, and you then have sorting at the endocytotic level with the help of an endosome that takes ones in the wrong place to the right place.
  - One recycling endosome goes to the cell surface, the other one elsewhere.

- Golgi to lysosome transport.
  - Lysosome proteins are all set up for proteolysis (chopping things up) and glycosidases. Take old cellular machinery, chop it up into its component parts, and let it get recycled.
  - A cell *must* recycle stuff, or it will need to make so much more amino acids.
  - Lysosomes are *highly* acidic and contain degratory proteins.
  - Why don't lysosomes digest themselves?
  - pH 5 and the proteins still work. Usually this pH would denature the protein, but instead lysosome proteins are built for this.
  - Hydrolases chop stuff up; they've evolved to withstand the highly acidic environment.
  - Mannose phosphate receptor in the Golgi: A bus to the lysosome and back (takes hydrolases from the trans-golgi to the pre-lysosomal compartments). Not a lysosome resident protein. Why doesn't it get denatured? Glycosylation of the receptor protects it from the hydrolase and the lysosome; can't purely operate at pH 5 but has to be stable at multiple.
  - Glycosylation tree branches around a specific protein protect it from the actions of the lysosome.

- Transport of newly synthesized lysosomal hydrolases to endosomes.
  - The mannose phosphate receptor is our bus.
  - We have a standard lysosomal hydrolase.
  - Carries a mannose as it enters the Golgi.
  - Addition of phospho-GlcNAc allows us to put a phosphate onto manose, and then it goes away. It's like a cofactor. Then it gets cut off.
  - Protonation allows the protein to fall off and enter the lysosome.
  - Then a retromer coat allows our receptor to be retrieved.

- Recognition of a lysosomal hydrolase.
  - Start with a lysosomal hydrolase carrying a glycosylation on the N-terminal .
  - GlcNAc phosphotransferase transfers a phosphate onto the lysosomal
  - UDP-GLcNAc transfers a phosphate and a GlcNAc to the lysosomal hydrolase, kicking out UMP.
  - Then the enzyme releases the lysosomal hydrolase. We then remove the GlcNAc, leaving mannose 6-phosphate behind (M6P).

- Ways to enter the lysosome.
  - Phagocytosis: Take in a bacteria to test it; use the lysosome to break it down into pieces that can be used by the rest of the cell (e.g., for defense).
  - Endocytosis: Take things in from the outside and digest them.

  - Autophagy: We automatically create a vesicle around something inside and move it to the lysosome.

- Endocytosis.
  - We take stuff in from the outside to an early endosome (vesicle within a cell; an intracellular sorting organelle).
  - Microtubule mediated transport to wherever we need, e.g., the lysosome, the trans-Golgi network, etc.

- Different mechanisms of endocytosis.
  - Macropinocytosis: An appendage sticks out and closes in.
  - Clathrin-coated vesicle: ...
  - Noncoated vesicle: ...
  - Caveolae: ...
  - Phagocytosis: ...

- Clathrin-mediated endocytosis.
  - Clathrin shapes rounding.
  - Makes things go to specific plases.
  - Receptor-mediated endocytosis.

- Receptor-mediated endocytosis.
  - Receptors on the plasma membrane draw external stuff into the cell.

- Recycling endosomes.
  - We don't want to destroy things we'll use again.
  - Membrane proteins that aren't needed go to an endosome, and when they're needed, they bud off and go back to the membrane.

- Degrading proteins: Autophagy.
  - Nucleation and extension: Bits of phospholipid engulf cytosol and organelles.
  - Transport to the lysosome.
  - Digestion: There, acid hydrolases break down the material.

- Transcytosis.
  - Moving something through a cell and to the other side. Think intestinal cells.
  - Endocytosis on the one side. Transport to the early entosome. Multiple pathways from there.
  - We can have an empty transport vesicle go back to the plasma membrane to replenish the phospholipidds there.
  - A full one can bud off to go for degradation in the endolysosome.
  - A full one can bud off and go to a recycling endosome for transcytosis to the far cell wall.

- Exocytosis.
  - Needed for cytokinesis, phagocytosis, plasma membrane repair, and cellularization.

References

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

Wu, T., Lyu, R., You, Q., & He, C. (2020). Kethoxal-assisted single-stranded dna sequencing captures global transcription dynamics and enhancer activity in situ. *Nature Methods*, *17*.