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 RNA 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.

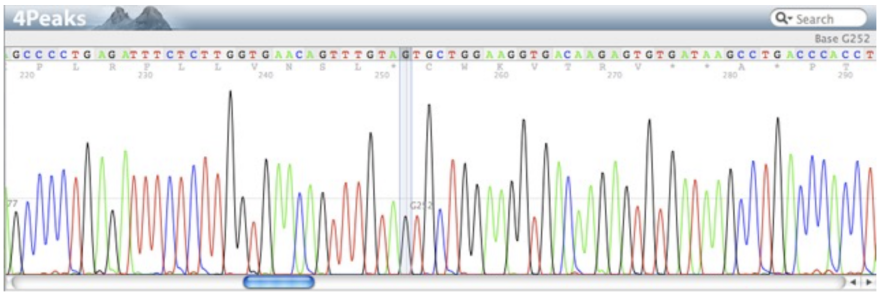


Figure 5.1: Sequence chromatogram example.

  - 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.
  - 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:

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

  - Running these mixtures on a gel will lead to bands corresponding to each cut and unreacted strands.
  - 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.

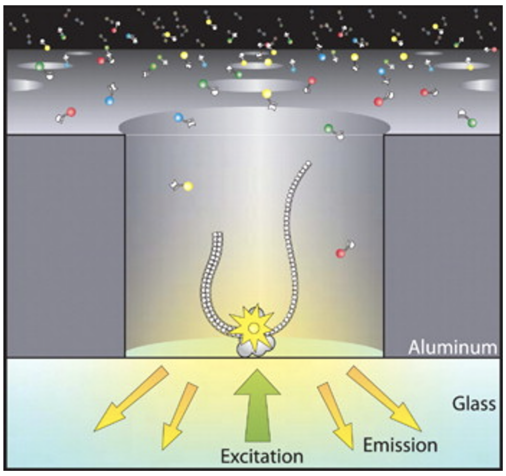


Figure 5.2: SMRT sequencing setup.

  - 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.

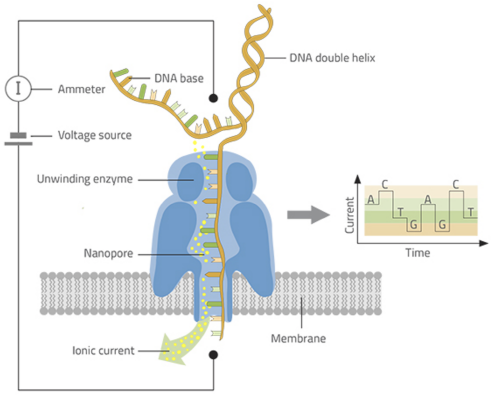


Figure 5.3: Nanopore sequencing setup.

  - 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.

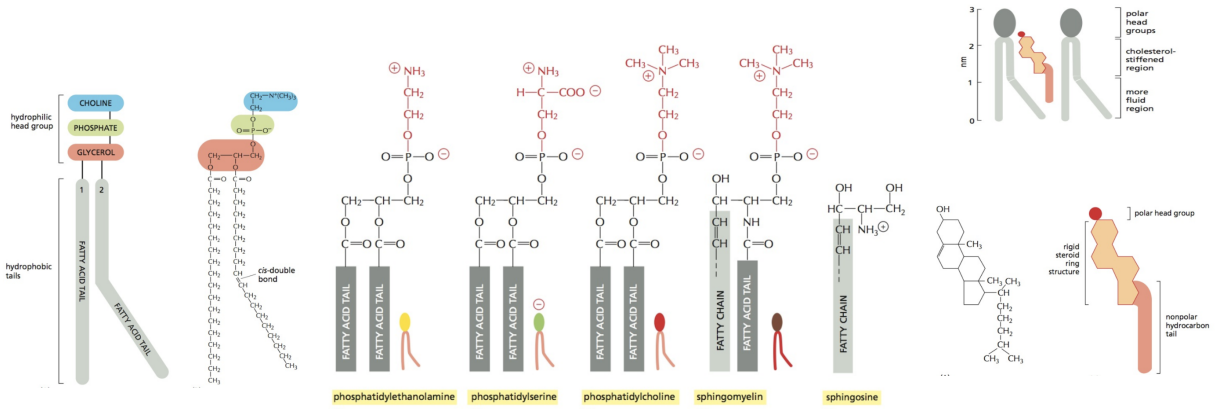
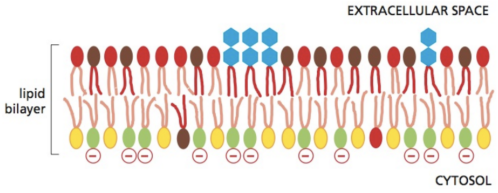


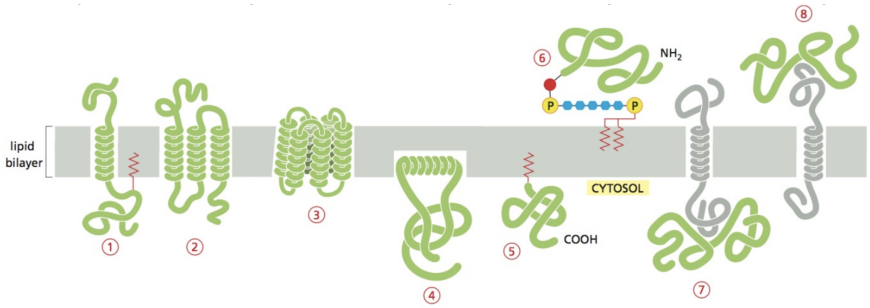
Figure 5.4: Plasma membrane constituents.

- 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.

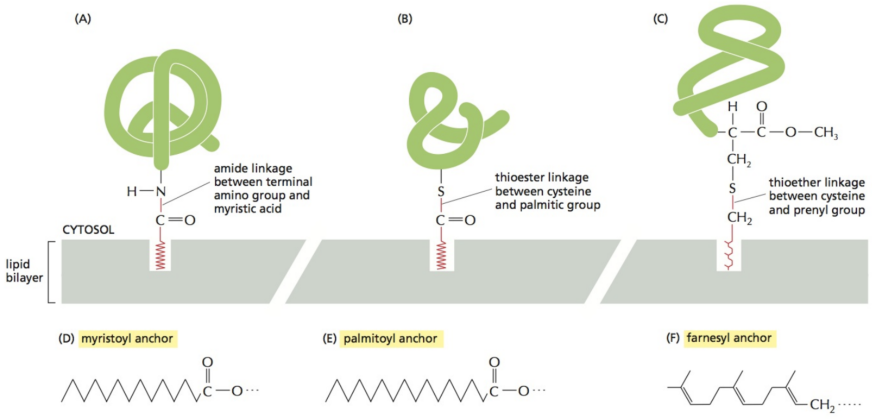


Figure 5.6: Lipid anchor types.

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

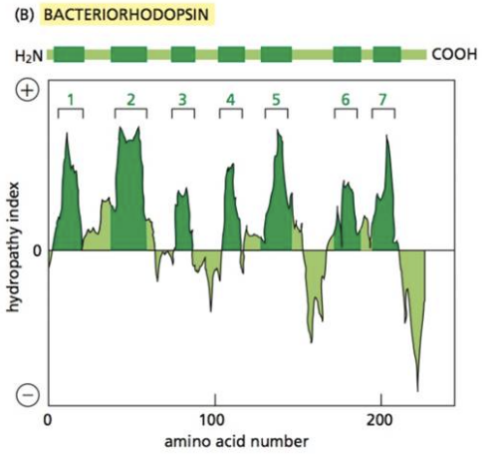


Figure 5.7: Hydropathy chart example.

- 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.

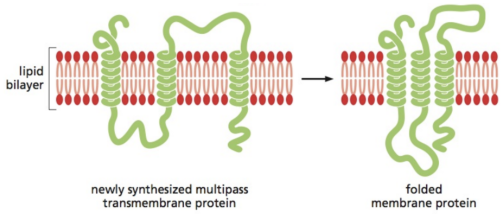
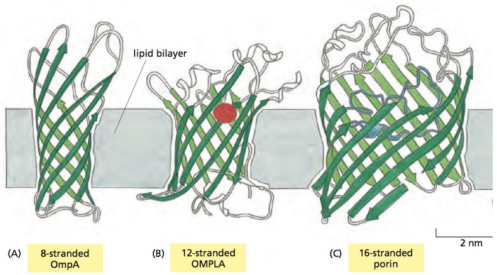


Figure 5.8: Multipass transmembrane protein folding.

  - 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.

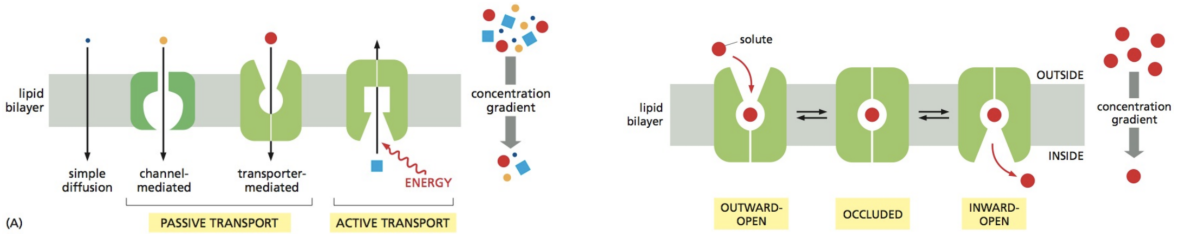
  - It is possible to embed without α-helices. Indeed, we can use β-sheets composed of hydrophobic residues.
    - Cytosolic proteins of this form tuck all of their hydrophobic side chains inside.
    - Transmembrane proteins of this form show all of their hydrophobic side chains outside.
  - Example: 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.

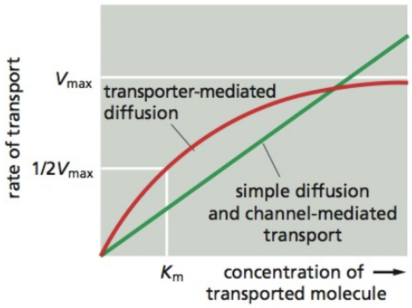


Figure 5.11: Concentration gradient regulation.

  - 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.

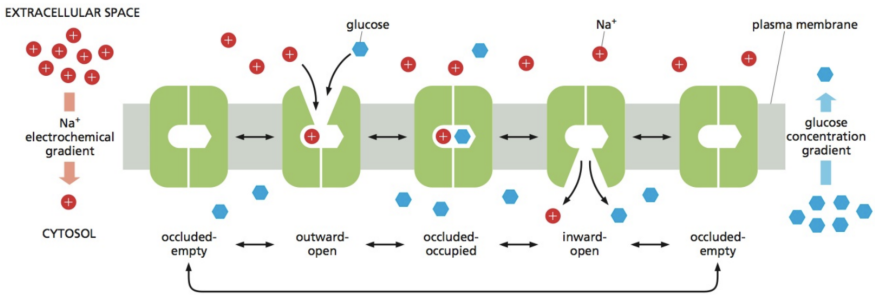


Figure 5.12: Sodium-glucose cotransporter activity.

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 referse 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.