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.

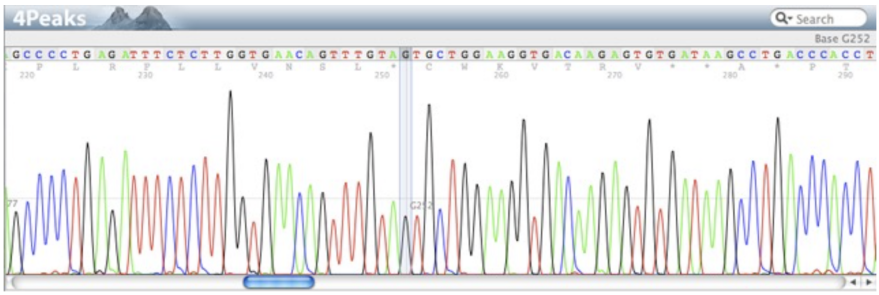

Figure 5.1: Sequence chromatogram example.

  - 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 the sequence each of 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.
  - 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.
  - Strand that is not covalently bonded 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 seeing 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.