COM SCI C121 Week 1

Aidan Jan

April 4, 2024

Biology Review

Central Dogma of Biology

- DNA is transcribed to RNA, which is then translated to proteins.
- During transcription, splicing may occur, so one section of DNA can produce multiple different strands of RNA, which is then translated into different proteins.
    - Occurs often in more complex humans (e.g., not bacteria)
- The definition of a gene is complicated, since one gene may have multiple exons, which may be spliced into different RNAs. (How do you quantify the isoforms?)
- Difference in splicing, translation, and regulation are part of what defines cell types.
    - This means that molecular smapling needs to be done for all different contexts
    - Computationally, we *need* fast, accurate, and space-efficient algorithms.

21st Century Biology Revolution

- High throughput DNA sequencing has revolutionized modern biology
- Can sequence billions of DNA fragments for relatively cheap (∼$1000)
- May biological questions can be reduced to sequencing experiments
    - e.g., RNA-Seq, ChIP-Seq, Methyl-Seq, RIP-Seq, CNV-Seq
- Currently, hundreds (thousands?) of experiments (since ∼2008)
- If you can reduce your experiment to a sequencing experiment, you can essentially do **thousands** of experiments at once.

What is DNA?

There are many types of biomolecules. (e.g., carbohydrates, lipids, proteins, and nucleic acids).

- DNA is a type of nucleic acid.
- DNA stores all the genetic information that a particular organism needs to live.
- DNA is stored in nearly every human cell. DNA inside chromosomes, inside nuclei, in cells.

DNA, genes, RNA, and proteins

- DNA contains coding and non-coding regions.
  - Coding regions are referred to as *exons*.
  - Non-coding regions are referred to as *introns*.
  - There are non-coding regions outside of these two groups, but are not discussed in this class.
- Introns exist to allow the same DNA section to code for multiple different proteins
  - Introns of some proteins may be exons of a different protein.

DNA Strands

DNA has two strands - the forward and reverse strands. Which one is forward strand is arbitrary - someone just picked it.

- The forward strand goes from 5' to 3' (these are names for the ends); the numbers represent the direction transcription occurs - transcription always occurs from 5' to 3'.
- 5' and 3' are named based on how the carbons are bonded.

Random Useful Facts about DNA

- A human "genome" stores about 3.1Gb (just one side of a double helix)
- Humans are 99.9% genetically identical
- A great overestimate of a person's variability is 3M genetic variants
- If we take the union of all single nuleotide variants, it's only $\sim$8M ($>$ 5% allele frequency)

Sequencing DNA

- Sanger Sequencing: the first practical method invent by Fred Sanger in 1977. Initially used to sequence short genomes (e.g., viruses, with 10k base pairs)
- 2nd Generation DNA Sequencing: around 2007, companies began sequencing commercially, but no technology can sequence much more than 10000 nucleotides at a reasonable cost, throughput, and accuracy
  - As a result, there's a race to create sequencers that can read "short" fragments (100s of nucleotides) efficiently with the best cost and accuracy.
  - The DNA would be "read" in sections, and then pieced together in the correct order.

How DNA is Sequenced

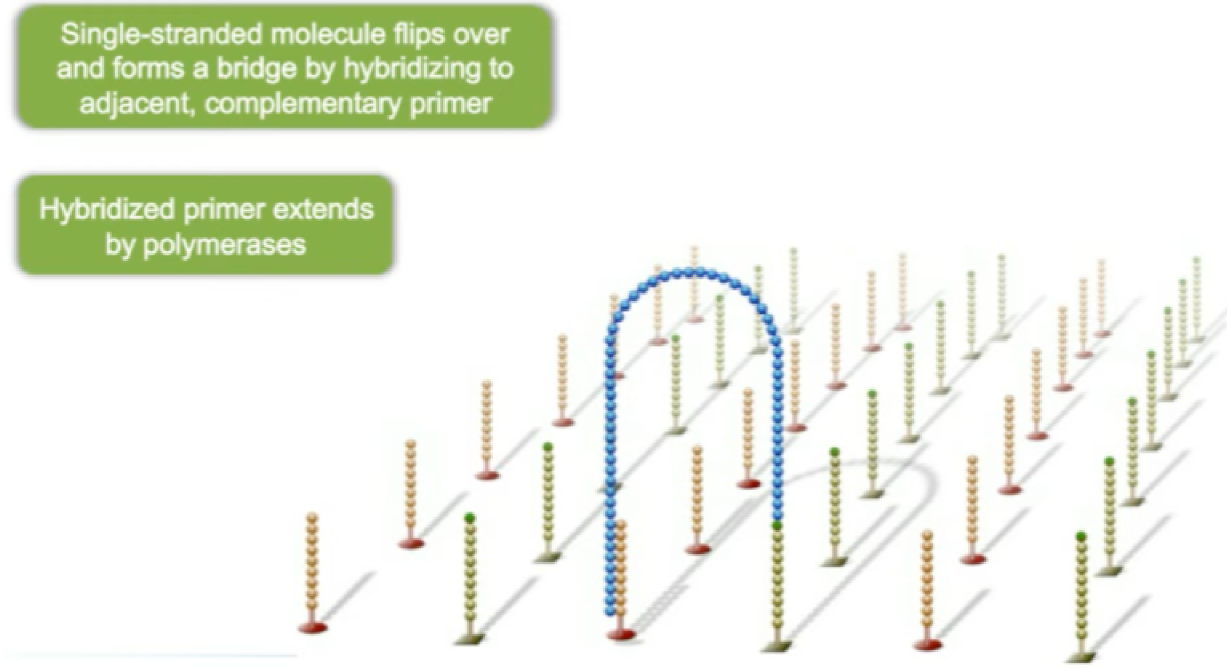
1. Isolate the DNA part that needs to be sequenced
2. Use DNA polymerase to amplify the DNA (Polymerase Chain Reaction)
3. Cut the DNA into snippets (sound waves tend to break the DNA into random, small snippets)
4. Deposit the snippets on the slides (kind of like electrophoresis)
5. Submerge the snippets with a pool of nucleotides with terminators and polymerase.
6. As DNA polymerase "builds" the strands of DNA, use a microscopic camera to capture flashes.

7. Remove the terminators, and repeat.

- There could be billions of templates on a single slide!
- This can be parallelized since a single microscope photo captures all the templates simultaneously.
- The terminators act as "speed bumps" and keep reactions in sync.

Bridge Amplification

There is a slight issue with the way DNA is sequence, and that is, it's really hard to isolate a single DNA molecule. In the Amplification stemp, bridge amplification may be used. This is done by attaching primers to a slide, and attaching a DNA strand in a "bridge" across two primers. It can then be copied.


Single-stranded molecule flips over and forms a bridge by hybridizing to adjacent, complementary primer
Hybridized primer extends by polymerases

By doing this, each snippet is actually a cluster, making reads easier.

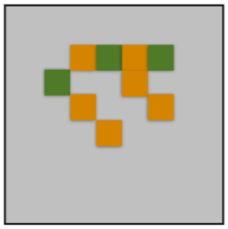
Loss

Between each read, it is not guaranteed that each strand as gained exactly one nucleotide (or is read accurately). As a result, we get *loss*, a number that describes the quality of the data.

$$Q = -10 \cdot \log_{10} p$$

- Q represents base quality
- p represents the probability that the base call is incorrect.

If $Q = 10$, it means there is a 1/10 chance the call is incorrect. $Q = 100$ means 1 in 100 chance the call is incorrect. The higher the quality, the better.
p refers to the ratio incorrect/total count. For example:



In the above image, the call would be orange, since it is the most plentiful. Then,

$$p = \frac{\text{number not orange}}{\text{number in cluster}} = \frac{3}{9} = 0.\bar{3}$$

Plugging in and solving for Q yields $Q = 4.77$.

Furthermore, once a given pixel is marked as incorrect, it is essentially incorrect 'forever', since once behind or ahead, it is unlikely to be synchronized with the rest. As a result, quality tends to decrease the longer into the strand.

- **Quality decreases as a function of length!**