Assembling a Genome

- An introduction to Graph algorithms
What we know about Genomes

- DNA sequences are a biological system's *hard drive*
  - They contain an *operating system* with all the low-level support for growing, dividing, and reproducing
  - They contain *application programs* for making cells that move our bodies, remember our mother's face, and store energy for use in lean times
  - They are robust. They have programs for repairing and replicating themselves. They even have backups!
- DNA sequences vary in size
  - Human nuclear DNA is composed of roughly 6 billion base-pairs distributed over 46 pairs of chromosomes
  - These 6 billion bases are comprised of 2 nearly identical copies
  - One of these copies is called a *haplotype* and its sequence is called a *genome*
  - Among humans, any two haplotypes are are 99.9% identical
- How can we read off the sequence of DNA?
DNA Sequencing History

- DNA sequencing was one of the most significant breakthroughs of the 20th century
- This was so inherently obvious it was awarded a Noble prize only 3 years after its development

**Sanger method (1977):**

Uses labeled dideoxynucleotide-triphosphates (ddNTPs) to terminate DNA copying at random points.

Fredrick Sanger

**Gilbert method (1977):**

Used various chemicals (Dimethyl Sulfate, Hydrazine) to modify and then cleave DNA at specific points (G, G+A, T+C, C).

Walter Gilbert

Both methods generate labeled fragments of varying lengths that are further electrophoresed
Sanger Method

1. Use the polymerase chain reaction (PCR) to make billions of copies of a DNA sequence
2. Starting at custom primer, sort of like our the origin of replication, we initiate one last replication
3. Include chemically altered and fluorescently labelled nucleotides, called dideoxynucleotide-tri-phosphates (ddNTPs)
4. If a ddNTP gets incorporated into a sequence it stops further replication
5. Separate replication products by length, using gel electrophoresis
6. Good for 500-1000 bases, then the error rates grow and extension rate slows
7. About 10 bases-per-second or 9.5 years to read an entire genome if we could do it from beginning to end
Sanger Method

1. Use the polymerase chain reaction (PCR) to make billions of copies of a DNA sequence.
2. Starting at a custom primer, sort of like our the origin of replication, we initiate one last replication.
3. Include chemically altered and fluorescently labelled nucleotides, called dideoxynucleotide-tri-phosphates (ddNTPs).
4. If a ddNTP gets incorporated into a sequence it stops further replication.
5. Separate replication products by length, using gel electrophoresis.
6. Good for 500-1000 bases, then the error rates grow and extension rate slows.
7. About 10 bases-per-second or 9.5 years to read an entire genome if we could do it from beginning to end.
Assembling the Human Genome

In 1990, a moon-shot-like project was begun to sequence the entire Human Genome.

- It would require 30x coverage to provide enough sequences
- Recall there are sequence differences-- Approximately 1:1000 bases
- Redundancy was needed to find the majority base from 16 different individuals (32 genomes)
- Also needed the extra coverage to assure that there is enough overlap to assemble the 500 base-pair reads

A $3 billion dollar NIH funded public effort led by Francis Collins with a 15-year plan. It would distribute the work across several labs in a community effort by assigning primers to groups on a first-come basis.
New sequencing results yielded new primers, so the project required a central coordination.

In 1997 a private company, Celera, lead by Craig Venter, suggested they could beat the public effort by dispensing with primers. They’d just randomly fragment DNA and sequence each with no idea of the how sequenced fragments would fit together. In other words, they were going to rely on computer science to assemble their reads algorithmically.

The result was that, despite tensions, the groups ended up sharing data and technologies. And the competition led to a completed draft 5 years ahead of schedule.
The Sequencing Race

Since the Human Genome project there have been an explosion of genomes sequenced. Initially, the focus was on model organisms, then favorites, then all of human diversity, and finally a catalog of life's diversity.
The secret behind this explosion of genomes

Next generation sequencing machines have revolutionized the DNA sequencing process. They work in various ways including massively-parallel single-base extension methods, to captured Dnases whose motions suggest a the base being replicated, to microholes that only a single DNA molecule can pass through, and the bases are determined by detectable charge differences.

In a way, the genome moonshot was far more successful than the real moonshot. The rate at which genomes can be sequenced, and the cost per base has seen unprecedented improvements. Faster than even Moore's Law.
How does it all work?

It is as if we must first smash a grecian urn in order to completely see it.
An Analogy

Some important differences

- A better analogy would have been to shred 100's of books
- Shuffle the pages before shredding
- Oh yeah, my book has approximately 850,000 characters.
- The entirety of Encyclopedia Britannica is approximately 250,000,000 characters. Your genome is approximately 12 times larger
How would you Reassemble our Book?

Each paper shred is like a DNA read.
Searching for overlaps

You'd look for fragments that fit together based on some *overlapping context* that they share.

And then, build upon those to assemble a more complete picture.
Until finally you assemble a nearly complete version
Key idea: Find links between fragment pairs

This leads us to a computational analogy called a **graph**

- A graph is composed of *nodes*, which can represent entities, in our case read fragments
- Nodes are connected by *edges* that represent some relationship between a pair of nodes
- The edges of a graph can be directed

One can devise both representations for, and algorithms that operate on, graphs.

- For example, you can find the shortest path between two nodes in a graph. Your GPS solves this problem, where addresses or locales are nodes, and roads are edges.
- You can find a minimal set of edges that maintains that keeps the graph *connected*

Let's rethink our DNA assembly problem as a graph problem.
The graph of a sequence

For the moment let's imagine that reads are like k-mers from a sequence, as they do tend to be uniform in length.

```
GACGGCGGCGCACGGCGCAA - Our toy sequence
GACG
ACGGC
CGGCG
GGCGG
CGGCG
GGCGC
GCGGC
CGGCG
GGCGC
GCGCA
CGCAA
```

- The complete set of 16 5-mers

Now we can construct a graph where:

1. Each 5-mer is a node
2. There is a directed edge from a k-mer that shares its (k-1)-base suffix with the (k-1)-base prefix of another k-mer
A read-overlap graph

The read-overlap graph for the 5-mers from:

GACGGCGGGCAGCAGCGCAA

The problem is *How to infer the original sequence from this graph?*
The rules of our game

- Every node, k-mer, can be used exactly once
- The object is to find a path along edges that visits every node one time
- This game was invented in the mid 1800's by a mathematician called Sir William Hamilton

A version of Hamilton's game:
Finding a Hamiltonian Path in a graph

Our desired sequence:

GACGGCGGCGCACGGCGCAA

is indeed a path in this graph.

How can we write a program to solve Hamilton's puzzle?

Is the solution unique?
Another way that to represent our k-mers in a graph

- Rather than making each k-mer a node, let's try making them an edge
- That seems odd, but it is related to the overlap idea
  - The 5-mer GACGG has a prefix GACG and a suffix ACGG
  - Think of the k-mer as the edge connecting a prefix to a suffix
  - This leads to a series of simple graphs

Then combine all nodes with the same Label
A De Bruijn Graph

This rather odd graph is called the "De Bruijn" graph, was named after a famous mathematician.

The problem is *How to infer the original sequence from this graph?*
The rules of our new game

- Every edge, k-mer, can be used exactly once
- The object is to find a path in the graph that uses each edge only one time
- This game was invented in the late 1700's by a mathematician called Leonhard Euler

A version of Euler's game:

Bridges of Königsberg
Find a city tour that crosses every bridge just once
Let's do a warm up exercise!

What is the shortest DNA sequence that starts with the subsequence "cat" and contains all possible dimers?

**Hint:** It's a graph problem!
Next Time

- Code that solves our graph problems
- Consider which code is simpler
- Consider which code is Faster