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 distrbuted 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) terminate DNA copying at random points.

Fredrick Sanger

Gilbert method (1977):

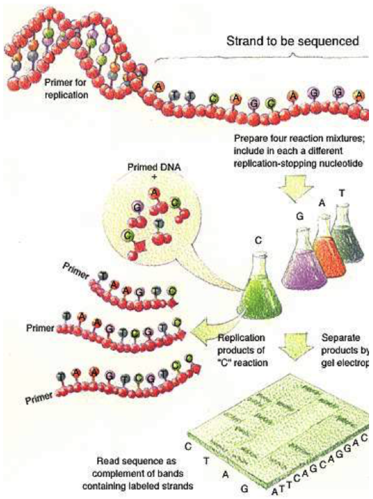
Used various chemicals (Dimethyl Sulfate, Hydrasine) to modify and then cleave DNA at specific points (G, G+A, T+C, C).

Walter Gilbert

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

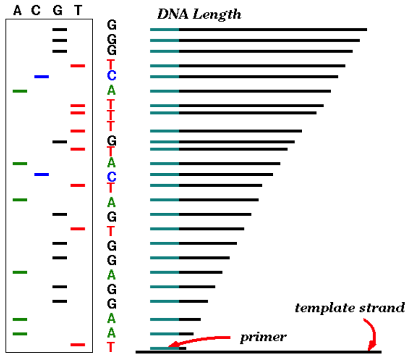
Sanger Method

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 inititate 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 *custom* primer, sort of like our the *origin of replication*, we inititate 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
- Redundacy 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

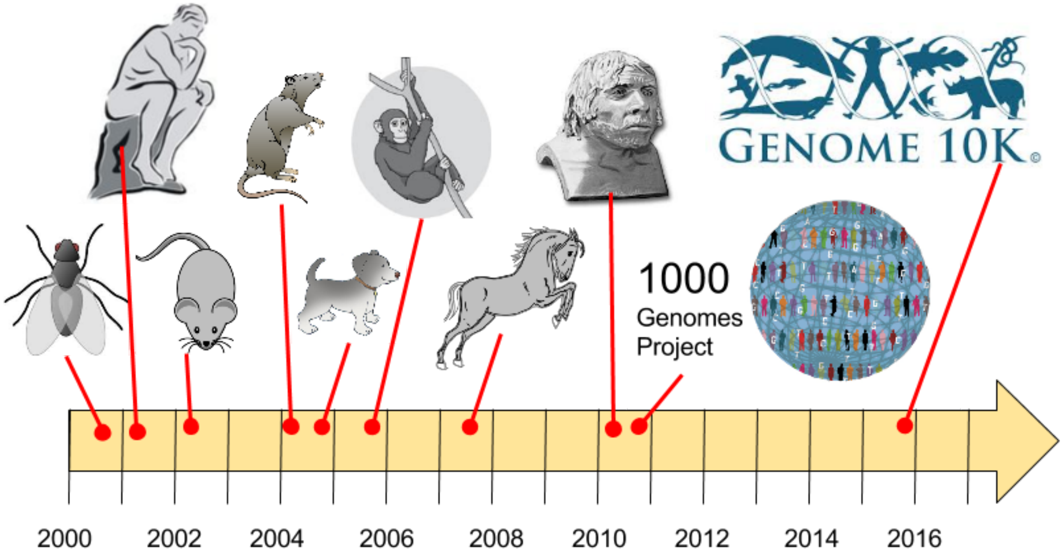
A $3 billion dollar NIH funded public effort led by Francis Collins with a 15-year plan. It would distribute the work across several labs in a community effort by assigning primers to groups on a first-come basis. New sequencing results yielded new primers, so the project required a central coordination.

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

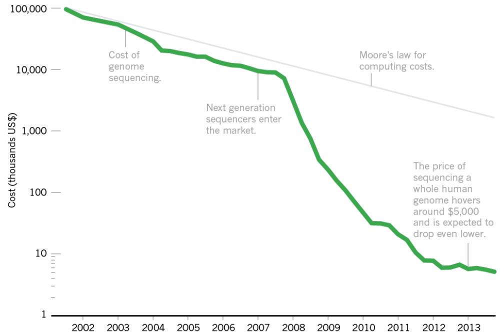
Since the Human Genome project there have been an explosion of genomes sequenced. Initially, the focus was on model organisms, then favorites, then all of human diversity, and finally a catalog of life's diversity.



The secret behind this explosion of genomes

Next generation sequencing machines have revolutionized the DNA sequencing process. They work in various ways including massiviely-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.

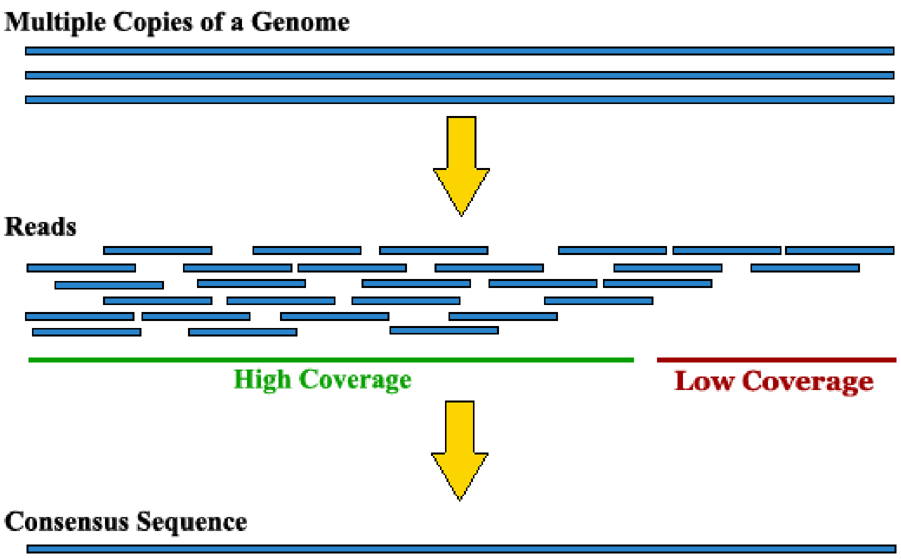
In a way, the *genome moonshot* was far more successful than the real moonshot. The rate at which genomes can be sequenced, and the cost per base has seen unprecented improvements. Faster than even Moore's Law.



Falling fast

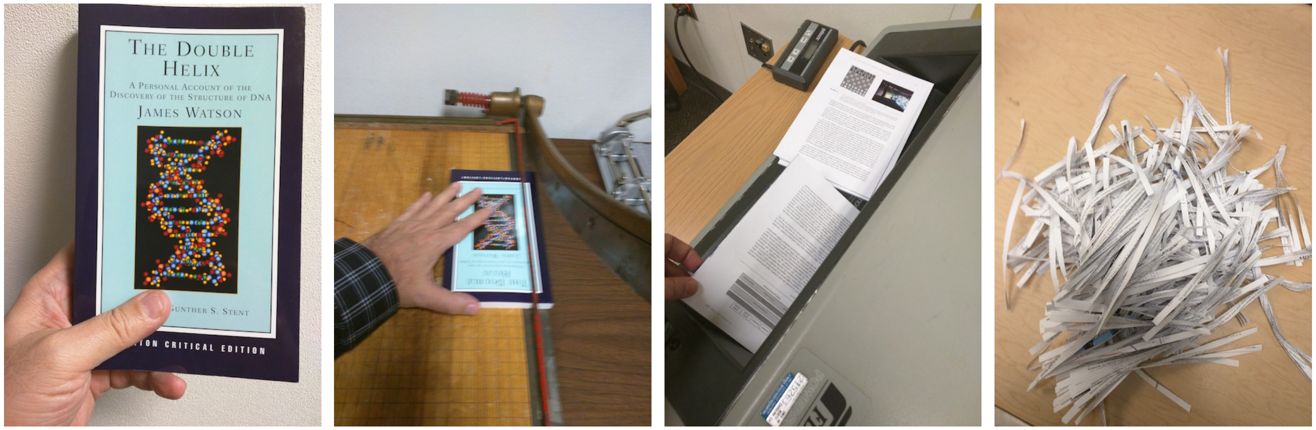
In the first few years after the end of the Human Genome Project, the cost of genome sequencing roughly followed Moore's law, which predicts exponential declines in computing costs. After 2007, sequencing costs dropped precipitously.

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

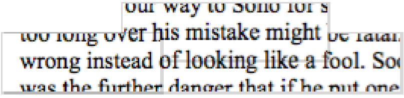
- A better analogy would have been to shred 100's of books
- Shuffle the pages before shredding
- Oh yeah, my book has approximately 850,000 characters.
- The entireity of Encyclopedia Britannica is approximately 250,000,000 characters. Your genome is approximately 12 times larger

How would you Reassemble our Book?

Each paper shred is like a DNA read.

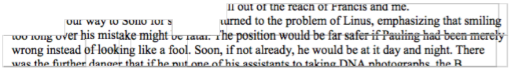
First Previous Next Notebook

Searching for overlaps



our way to Soho for s
too long over his mistake might be fatal.
wrong instead of looking like a fool. So
was the further danger that if he put one

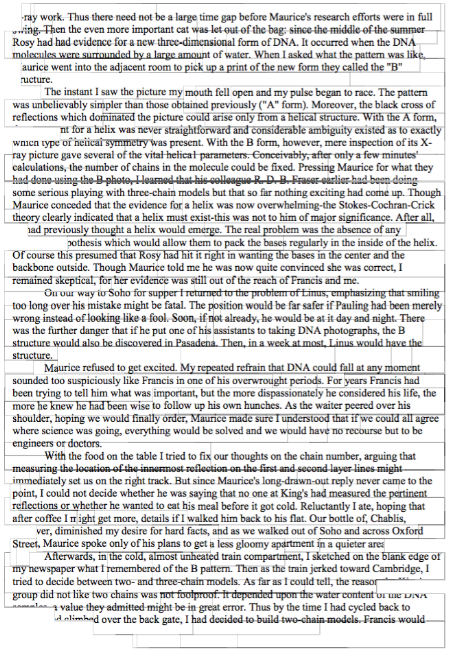
You'd look for fragments that fit together based on some *overlapping context* that they share.



ll out of the reach of Francis and me.
our way to Soho for s
turned to the problem of Linus, emphasizing that smiling
too long over his mistake might be fatal. The position would be far safer if Pauling had been merely
wrong instead of looking like a fool. Soon, if not already, he would be at it day and night. There
was the further danger that if he put one of his assistants to taking DNA photographs, the B

And then, build upon those to assemble a more complete picture

Until finally you assemble a nearly complete version



-ray work. Thus there need not be a large time gap before Maurice's research efforts were in full wing. Then the even more important cat was let out of the bag: since the middle of the summer Rosy had had evidence for a new three-dimensional form of DNA. It occurred when the DNA molecules were surrounded by a large amount of water. When I asked what the pattern was like, aurice went into the adjacent room to pick up a print of the new form they called the "B" ructure.

The instant I saw the picture my mouth fell open and my pulse began to race. The pattern was unbelievably simpler than those obtained previously ("A" form). Moreover, the black cross of reflections which dominated the picture could arise only from a helical structure. With the A form, nt for a helix was never straightforward and considerable ambiguity existed as to exactly which type of helical symmetry was present. With the B form, however, mere inspection of its X-ray picture gave several of the vital helical parameters. Conceivably, after only a few minutes' calculations, the number of chains in the molecule could be fixed. Pressing Maurice for what they had done using the B photo, I learned that his colleague R. D. B. Fraser earlier had been doing some serious playing with three-chain models but that so far nothing exciting had come up. Though Maurice conceded that the evidence for a helix was now overwhelming-the Stokes-Cochran-Crick theory clearly indicated that a helix must exist-this was not to him of major significance. After all, ad previously thought a helix would emerge. The real problem was the absence of any othesis which would allow them to pack the bases regularly in the inside of the helix. Of course this presumed that Rosy had hit it right in wanting the bases in the center and the backbone outside. Though Maurice told me he was now quite convinced she was correct, I remained skeptical, for her evidence was still out of the reach of Francis and me.

On our way to Soho for supper I returned to the problem of Linus, emphasizing that smiling too long over his mistake might be fatal. The position would be far safer if Pauling had been merely wrong instead of looking like a fool. Soon, if not already, he would be at it day and night. There was the further danger that if he put one of his assistants to taking DNA photographs, the B structure would also be discovered in Pasadena. Then, in a week at most, Linus would have the structure.

Maurice refused to get excited. My repeated refrain that DNA could fall at any moment sounded too suspiciously like Francis in one of his overwrought periods. For years Francis had been trying to tell him what was important, but the more dispassionately he considered his life, the more he knew he had been wise to follow up his own hunches. As the waiter peered over his shoulder, hoping we would finally order, Maurice made sure I understood that if we could all agree where science was going, everything would be solved and we would have no recourse but to be engineers or doctors.

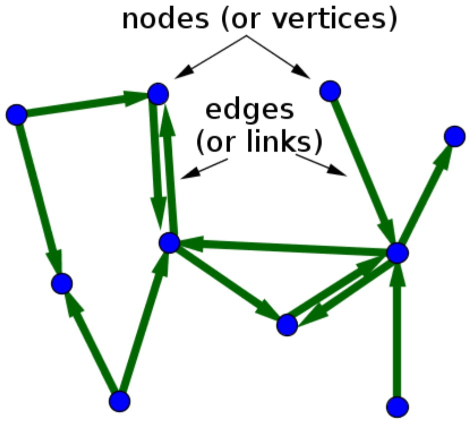
With the food on the table I tried to fix our thoughts on the chain number, arguing that measuring the location of the innermost reflection on the first and second layer lines might immediately set us on the right track. But since Maurice's long-drawn-out reply never came to the point, I could not decide whether he was saying that no one at King's had measured the pertinent reflections or whether he wanted to eat his meal before it got cold. Reluctantly I ate, hoping that after coffee I might get more, details if I walked him back to his flat. Our bottle of, Chablis, ver, diminished my desire for hard facts, and as we walked out of Soho and across Oxford Street, Maurice spoke only of his plans to get a less gloomy apartment in a quieter area

Afterwards, in the cold, almost unheated train compartment, I sketched on the blank edge of my newspaper what I remembered of the B pattern. Then as the train jerked toward Cambridge, I tried to decide between two- and three-chain models. As far as I could tell, the reasor group did not like two chains was not foolproof. It depended upon the water content of the DNA a value they admitted might be in great error. Thus by the time I had cycled back to d over the back gate, I had decided to build two-chain models. Francis would

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 representaions for, and algorithms that operate on, graphs.

- For example, you can find the shortest path between to 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 ssembly 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
GACGG
 ACGGC
  CGGCG
   GGCGG
    GCGGC
     CGGCG
      GGCGC             - The complete set of 16 5-mers
       GCGCA
        CGCAC
         GCACG
          CACGG
           ACGGC
            CGGCG
             GGCGC
              GCGCA
               CGCAA
```

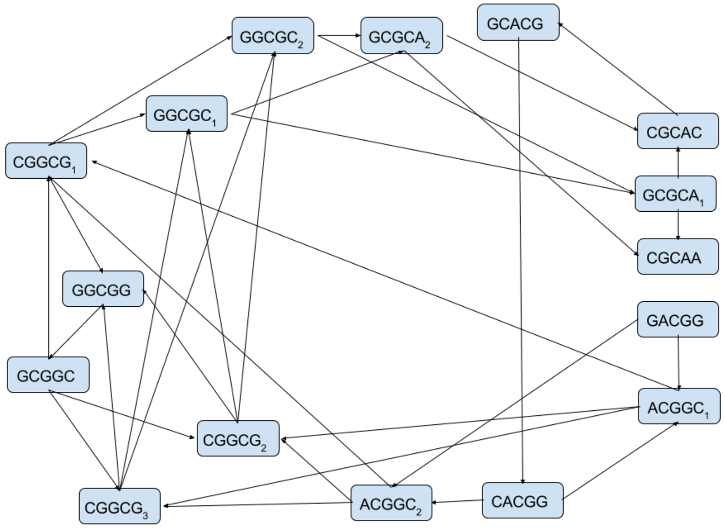
Now we can construct a graph where:

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:

`GACGGCGGCGCACGGCGCAA`

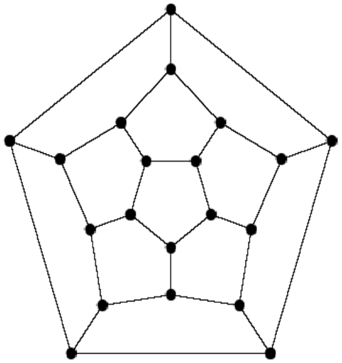


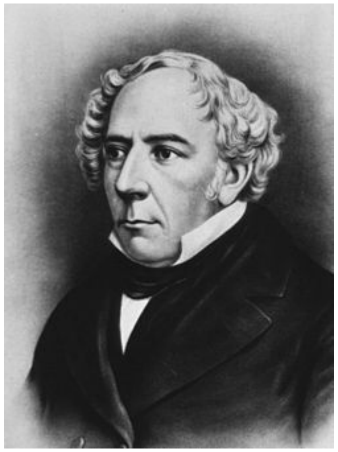
The problem is ***How to infer the original sequence from this graph?***

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:



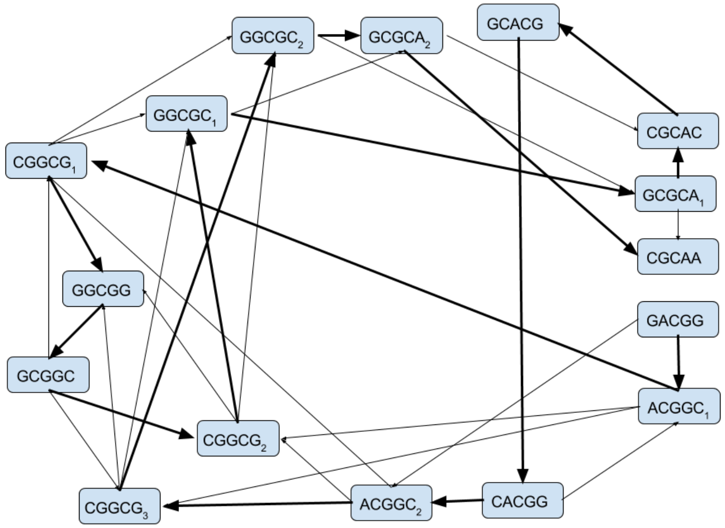


Finding a Hamiltonian Path in a graph

Our desired sequence:

`GACGGCGGCGCACGGCGCAA`

is indeed a path in this graph

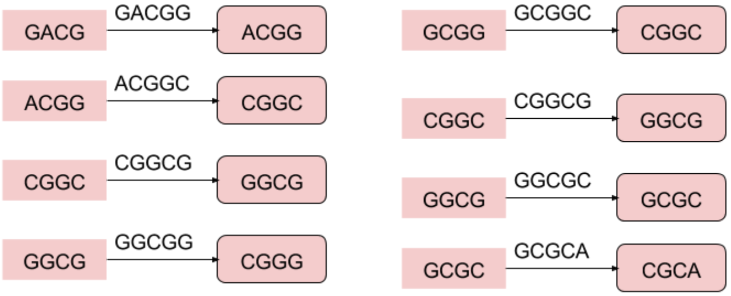


How can we write a program to solve Hamilton's puzzle?

Is the solution unique?

Another way that to represent our k-mers in a graph

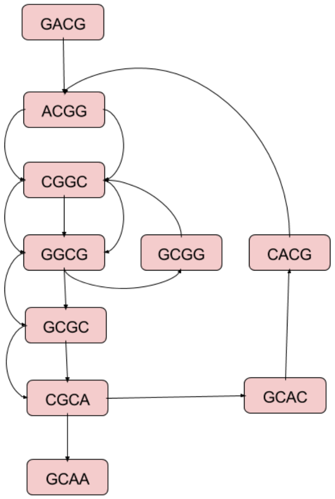
- Rather than making each k-mer a node, let's try making them an edge
- That seems odd, but it is related to the overlap idea
    - The 5-mer `GACGG` has a prefix `GACG` and a suffix `ACGG`
    - Think of the k-mer as the edge connecting a prefix to a suffix
    - This leads to a series of simple graphs



- 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.



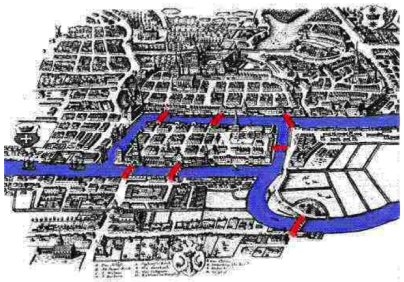
The problem is ***How to infer the original sequence from this graph?***

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

Leonhard Euler

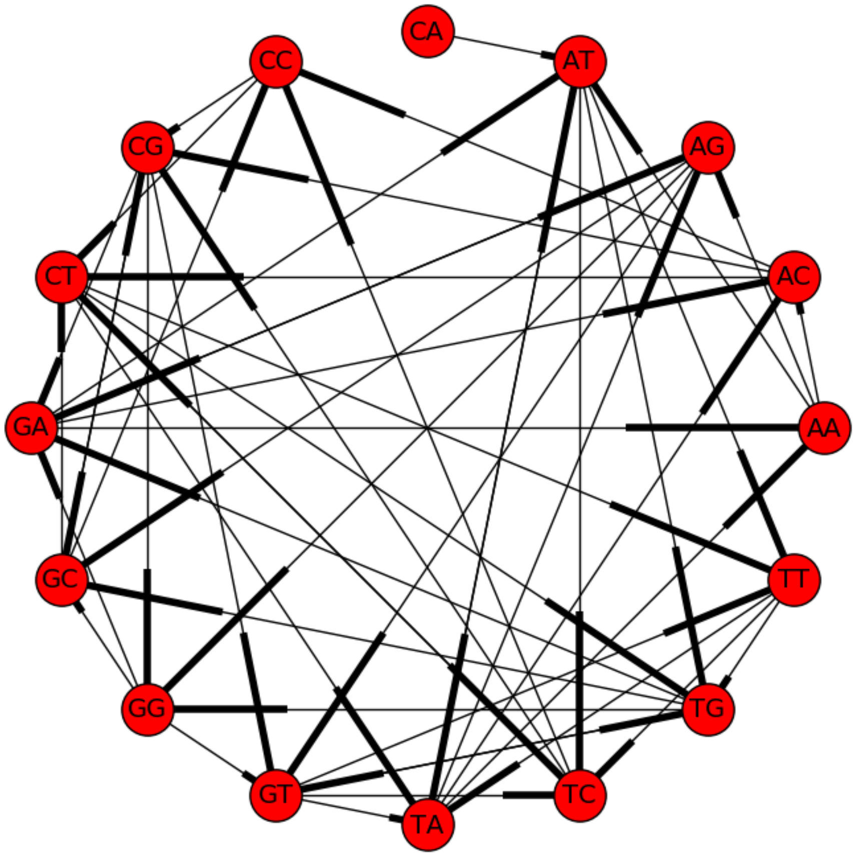
A version of Euler's game:



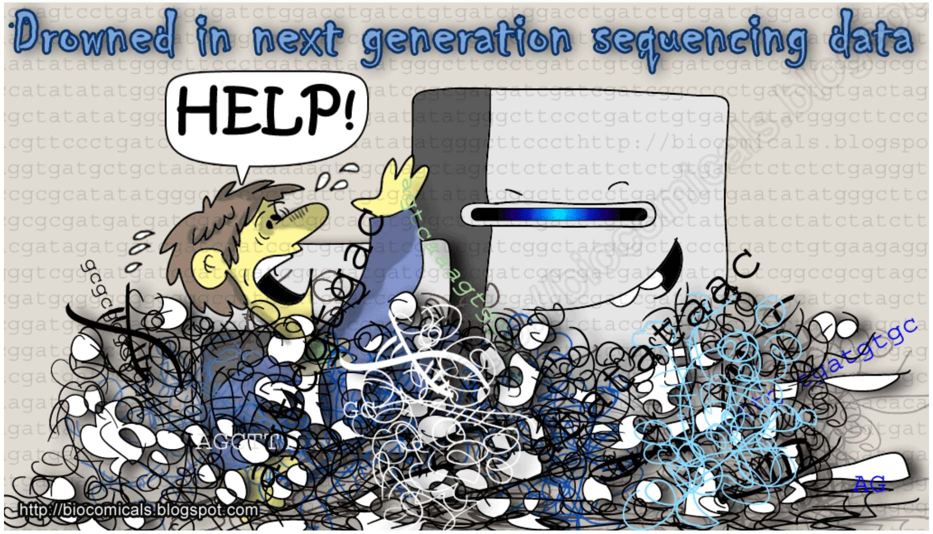
Bridges of Königsberg
Find a city tour that crosses
every bridge just once

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 simplier
- Consider which code is Faster