Sequence Assembly

Fall 2016

BMI/CS 576

www.biostat.wisc.edu/bmi576/

Irene Ong

Irene.ong@wisc.edu

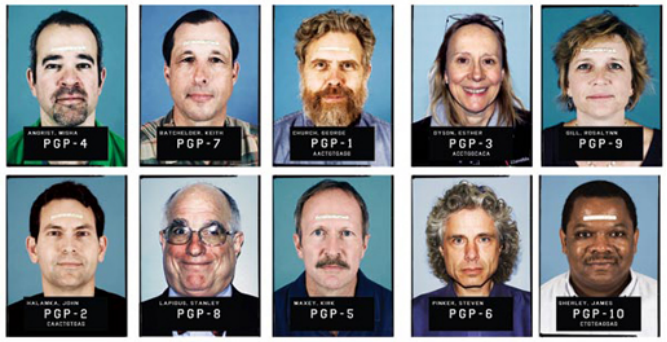
Acknowledgement

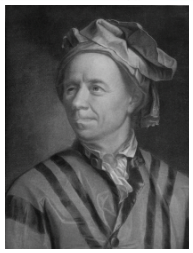
- Some slides courtesy of Phillip Compeau and Pavel Pevzner's Bioinformatics Algorithms
- Please do not share outside of class.

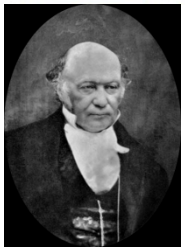
Outline

- What Is Genome Sequencing?
- Exploding Newspapers
- The String Reconstruction Problem
- String Reconstruction as a Hamiltonian Path Problem
- String Reconstruction as an Eulerian Path Problem
- Similar Problems with Different Fates
- De Bruijn Graphs
- Euler's Theorem
- Assembling Read-Pairs
- De Bruijn Graphs Face Harsh Realities of Assembly

Who Are These People?







Euler
1707-1783

Hamilton
1805-1865

De Bruijn
1918-2012

The human genome is a three billion nucleotide long "book" written in A, C, G, T alphabet.

Some genomes are 200 X larger than the human genome:

Amoeba dubia

japonica

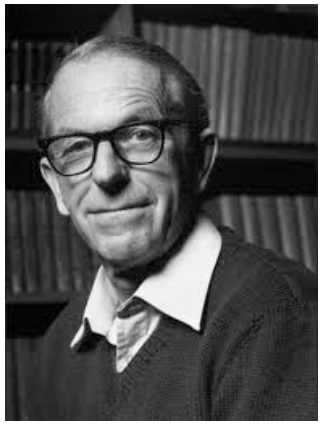
Why Do We Sequence 1000s of Species?

- Applications in **medicine** (genomes of fungi-producing bacteria), **agriculture** (oil palm genome), **biotechnology** (genomes of energy-producing cyanobacteria), etc., etc., etc.

Brief History of Genome Sequencing

- **1977**: Walter Gilbert and Frederick Sanger develop independent DNA sequencing methods.
- **1980**: They share the Nobel Prize.
- Still, their sequencing methods were too expensive ($3 billion to sequence the human genome).

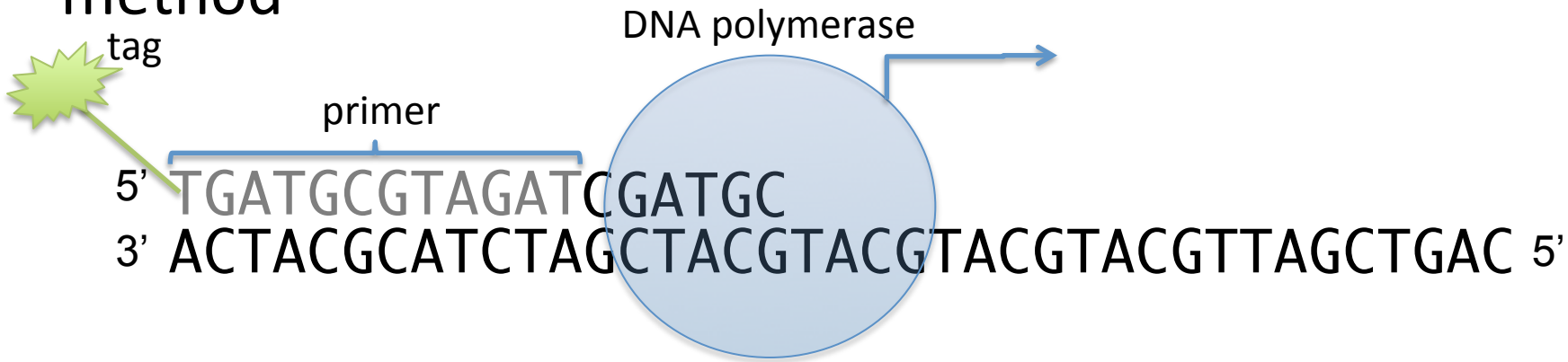
Walter Gilbert



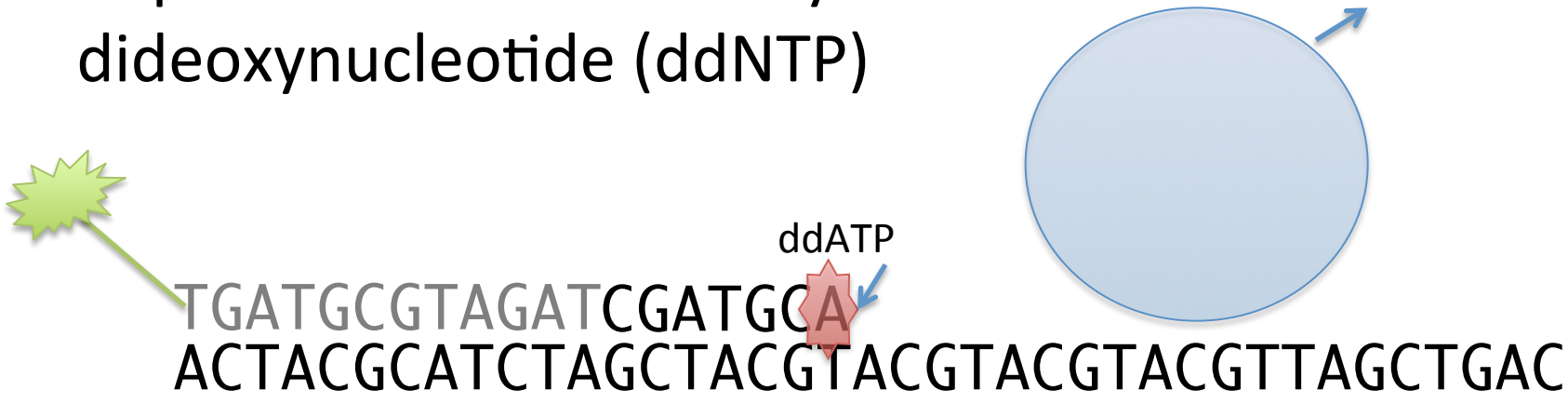
Frederick Sanger

Sanger sequencing

- Classic sequencing technique: "Chain-termination method"

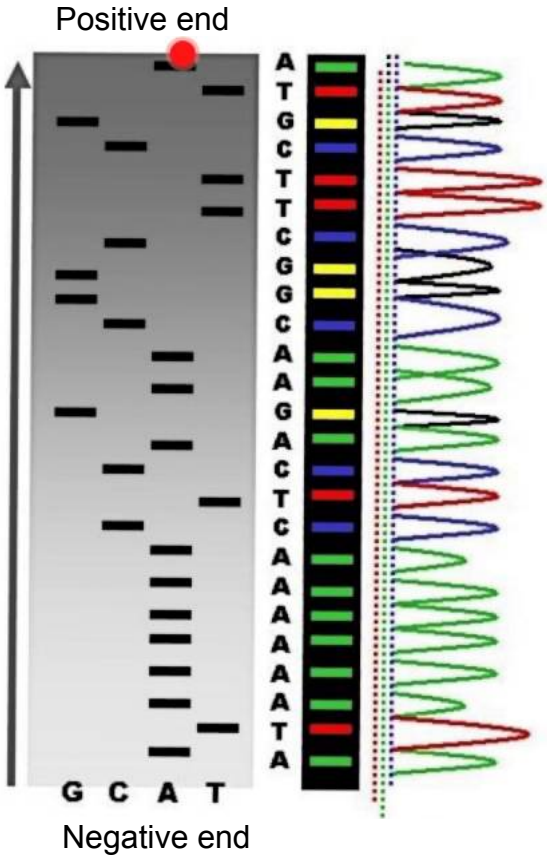


- Replication terminated by inclusion of dideoxynucleotide (ddNTP)



Sequencing gels

- Run replication in four separate test tubes
  - Each with one of some concentration of either ddATP, ddTTP, ddGTP, or ddCTP
- Depending on when ddNTP is included, different length fragments are synthesized
- Fragments separated by length with electrophoresis gel
- Sequence can be read from bands on gel



The sequencing problem

- Determine the base pairs of a genome
- No way to "read" the genome sequence from beginning to end
- Can read *short* pieces (substrings) of DNA
  - Sanger sequencing: 500-700 bp/read
  - Hybridization arrays: 8-30bp/probe

The Race to Sequence the Human Genome

- **1990**: The public Human Genome Project, headed by Francis Collins, aims to sequence the human genome by 2005.

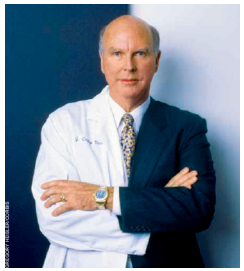
Francis Collins

The Race to Sequence the Human Genome

- **1990**: The public Human Genome Project, headed by Francis Collins, aims to sequence the human genome by 2005.

Francis Collins

- **1997**: Craig Venter founds Celera Genomics, a private firm, with the same goal.



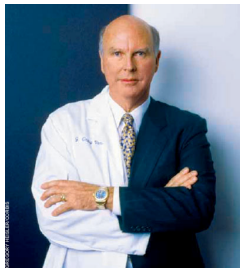
Craig Venter

The Race to Sequence the Human Genome

- **1990**: The public Human Genome Project, headed by Francis Collins, aims to sequence the human genome by 2005.

Francis Collins

- **1997**: Craig Venter founds Celera Genomics, a private firm, with the same goal.

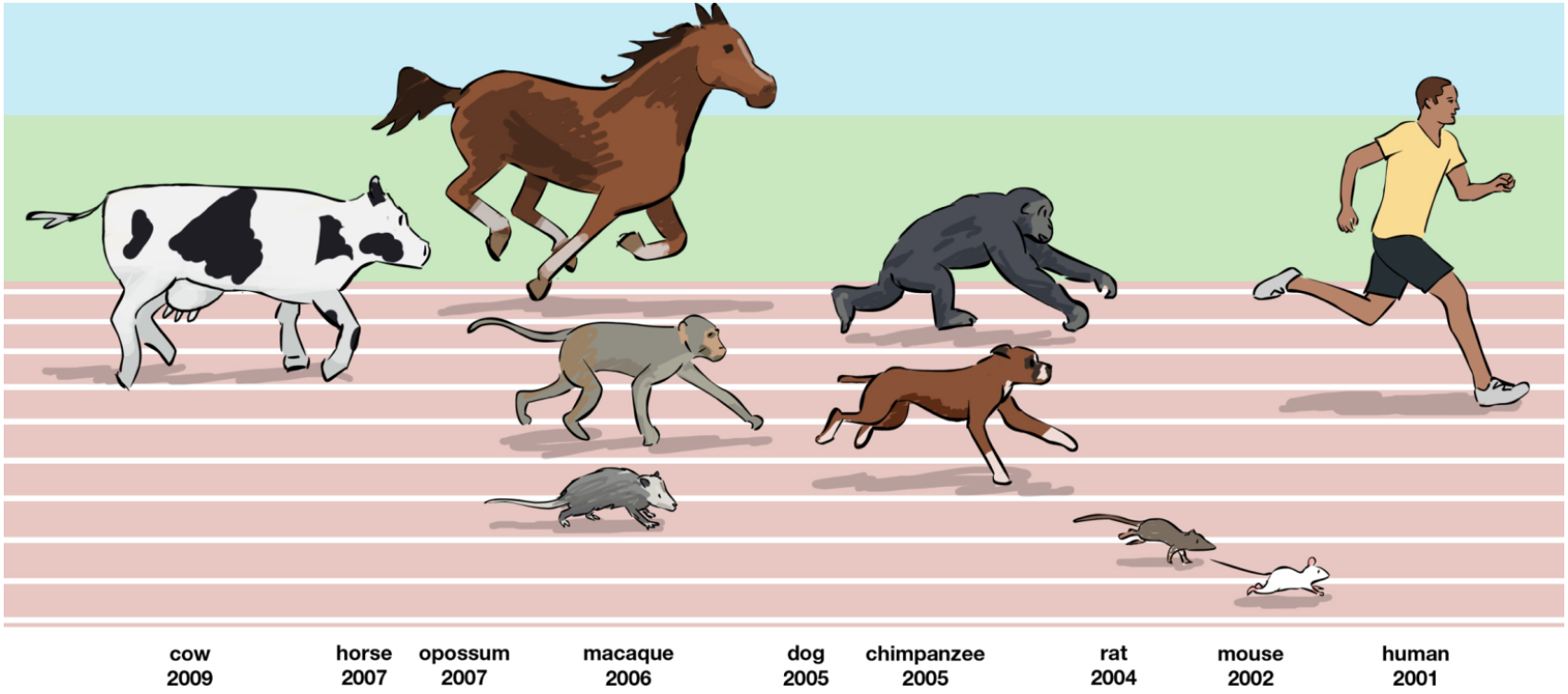


Craig Venter

- **2000:**

From Human to Mouse to Rat to …

Early 2000s: Many more mammalian genomes were sequenced using the Sanger sequencing method, but new technology was needed for further progress.



Latest technologies

- 454:
  - “Sequencing by synthesis”
  - Light emitted and detected on addition of a nucleotide by polymerase
  - 400-600 Mb / 10 hour run
- Illumina
  - Also “sequencing by synthesis”
  - ~100 Gb/day on one machine
  - Uses fluorescently-labeled reversible nucleotide terminators
  - Like Sanger, but detects added nucleotides with laser after each step

Latest technologies

- Pacific Biosciences:
  - "Sequencing by synthesis"
  - Single molecule sequencing
  - Detects addition of single fluorescently-labeled nucleotides by an immobilized DNA polymerase
  - Real-time: reads bases at the rate of DNA polymerase
  - 4 hours for sequencing with reads up to 60kb long
  - video

Oxford Nanopore

- Emerging technology
- Pocket-sized
- High error rate

Next Generation Sequencing Technologies

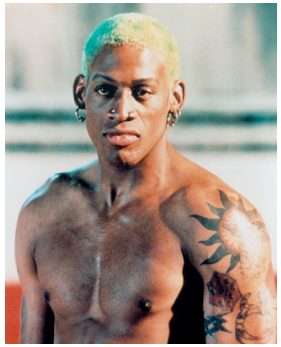
- **Late 2000s**: The market for new sequencing machines takes off.
  - Illumina reduces the cost of sequencing a human genome from $3 billion to $10,000.
  - Complete Genomics builds a genomic factory in Silicon Valley that sequences hundreds of genomes per month.
  - Beijing Genome Institute orders hundreds of sequencing machines, becoming the world's largest sequencing center.

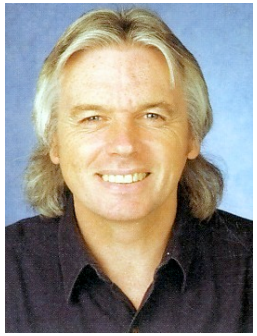
illumina

Complete genomics

华大基因 BGI

Personal Genome Sequencing





Few Mutations Can Make a Big Difference…

- Different people have slightly different genomes: on average, roughly 1 mutation in 1000 nucleotides.

- The 1 in 1000 nucleotides difference accounts for height, high cholesterol susceptibility, and 1000s of genetic diseases.

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGGGACTATTA
TCGACTACAGATAAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGA
TCAGCTACAACATCGTAGCTACGATGCATTAGCAAGCTAT
CGATCGATCGATCGATTATCTACGATCGATCGATCGATCA
CTATACGAGCTACTACGTACGTACGATCGCGTGACTATTA
TCGACTACAGATGAAACATGCTAGTACAACAGTATACATA
GCTGCGGGATACGATTAGCTAATAGCTGACGATATCCGAT

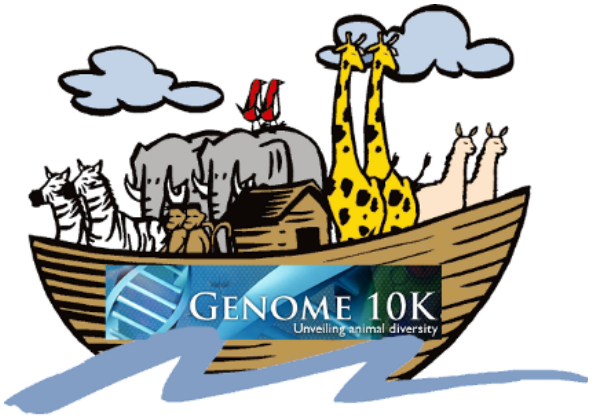
Why Do We Sequence Personal Genomes?

- **2010:** Nicholas Volker became first human being to be saved by genome sequencing.

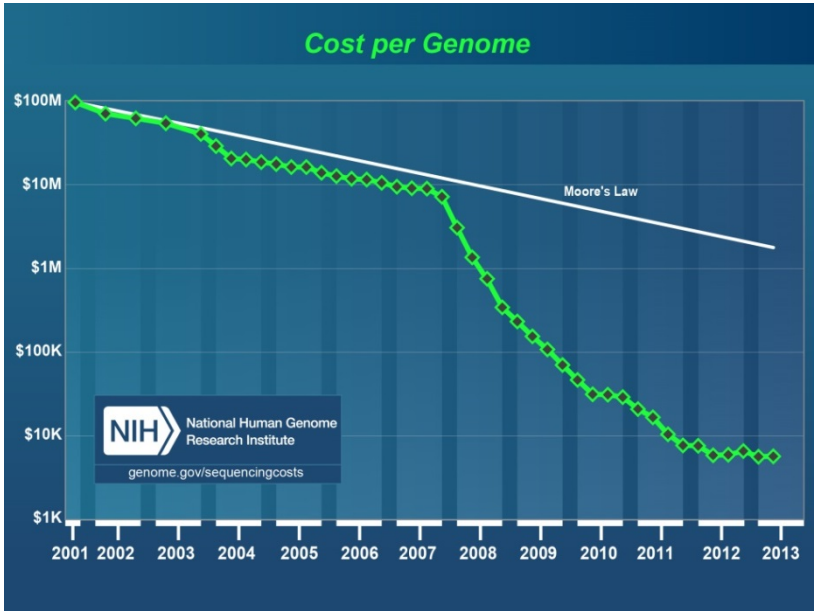
  - Doctors could not diagnose his condition; he went through dozens of surgeries.
  - Sequencing revealed a rare mutation in a *XIAP* gene linked to a defect in his immune system.
  - This led doctors to use immunotherapy, which saved the child.

10,000 Genomes and Beyond

- **2010**: Scientists launch a project to sequence 10,000 vertebrate genomes.



- **Now**: Human genome sequencing costs just a few thousand dollars and under $1,000 human genomes may arrive any day now.



Outline

- What Is Genome Sequencing?
- Exploding Newspapers
- The String Reconstruction Problem
- String Reconstruction as a Hamiltonian Path Problem
- String Reconstruction as an Eulerian Path Problem
- Similar Problems with Different Fates
- De Bruijn Graphs
- Euler's Theorem
- Assembling Read-Pairs
- De Bruijn Graphs Face Harsh Realities of Assembly

The Newspaper Problem

stack of NY Times, June 27, 2000

The Newspaper Problem

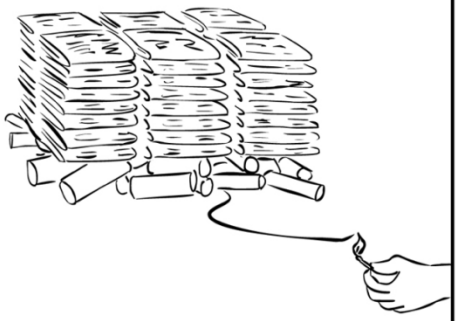
stack of NY Times, June 27, 2000

stack of NY Times, June 27, 2000
on a pile of dynamite

The Newspaper Problem

stack of NY Times, June 27, 2000

stack of NY Times, June 27, 2000
on a pile of dynamite

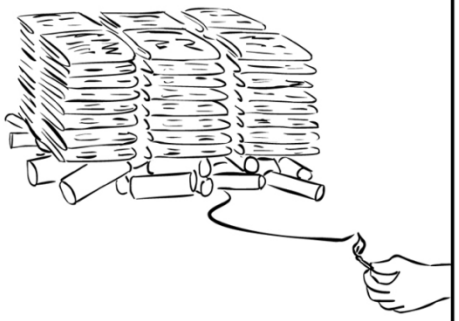


this is just hypothetical

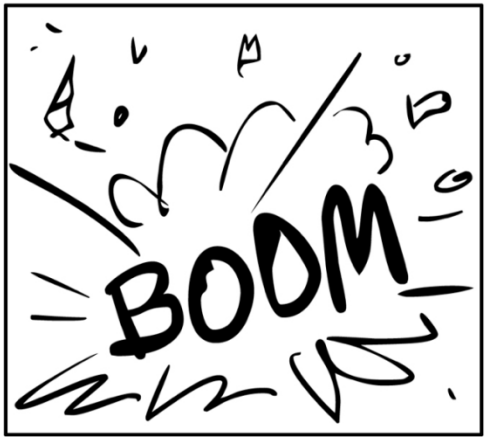
The Newspaper Problem

stack of NY Times, June 27, 2000

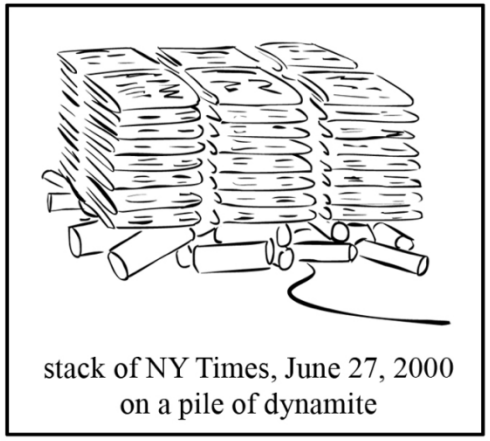
stack of NY Times, June 27, 2000
on a pile of dynamite

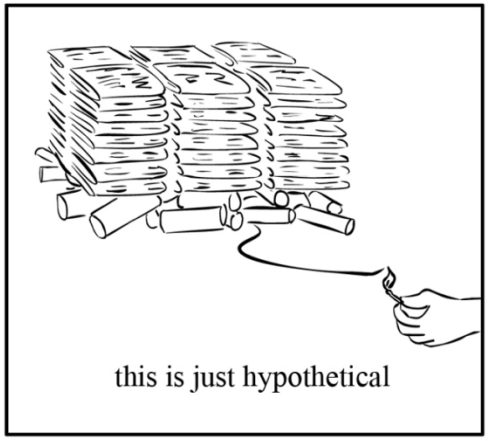


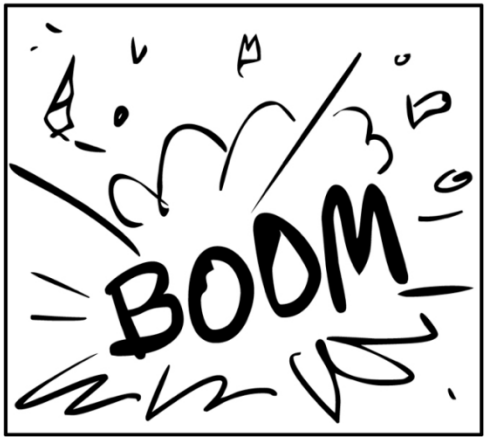
this is just hypothetical

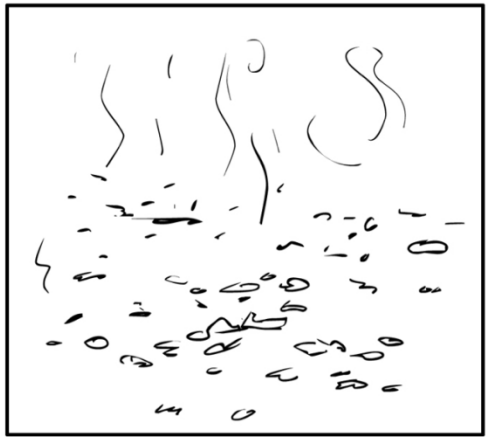


The Newspaper Problem





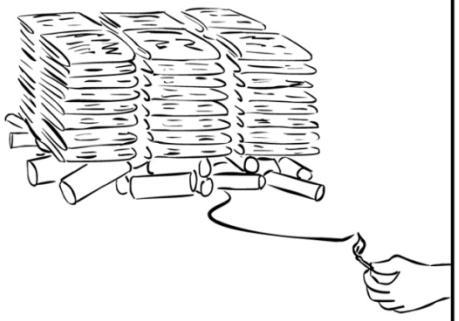




The Newspaper Problem

stack of NY Times, June 27, 2000

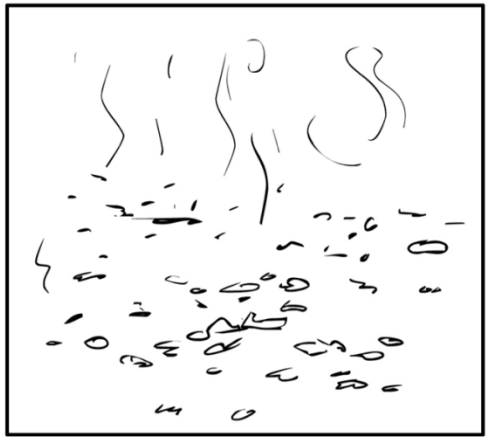
stack of NY Times, June 27, 2000
on a pile of dynamite

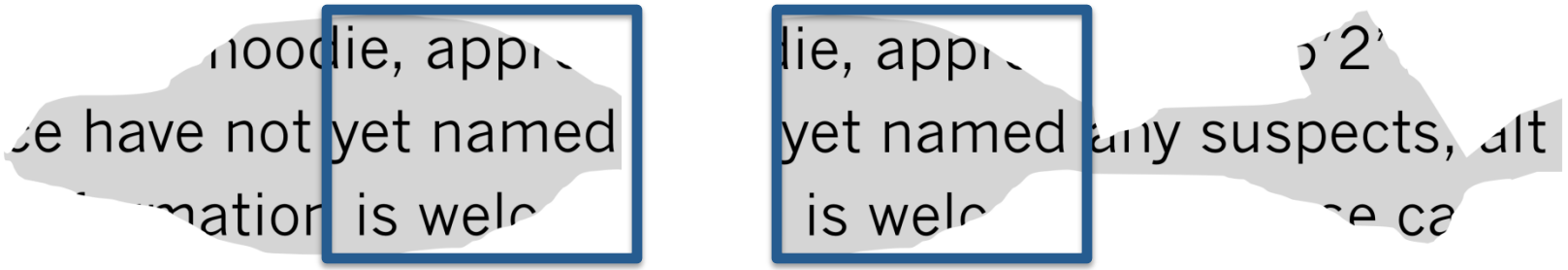


this is just hypothetical

so, what did the June 27, 2000 NY
Times say?

The Newspaper Problem as an Overlapping Puzzle





The Newspaper Problem as an Overlapping Puzzle

hoodie, appr 2'
e have not yet name spects, alt
mation is welc e ca

Multiple Copies of a Genome (Millions of them)

stack of NY Times, June 27, 2000

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

Breaking the Genomes at Random Positions

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

Generating “Reads”

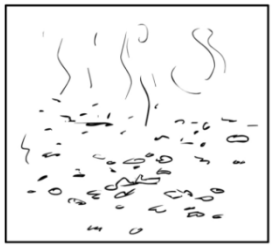
CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC

CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC

CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC

CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

“Burning” Some Reads



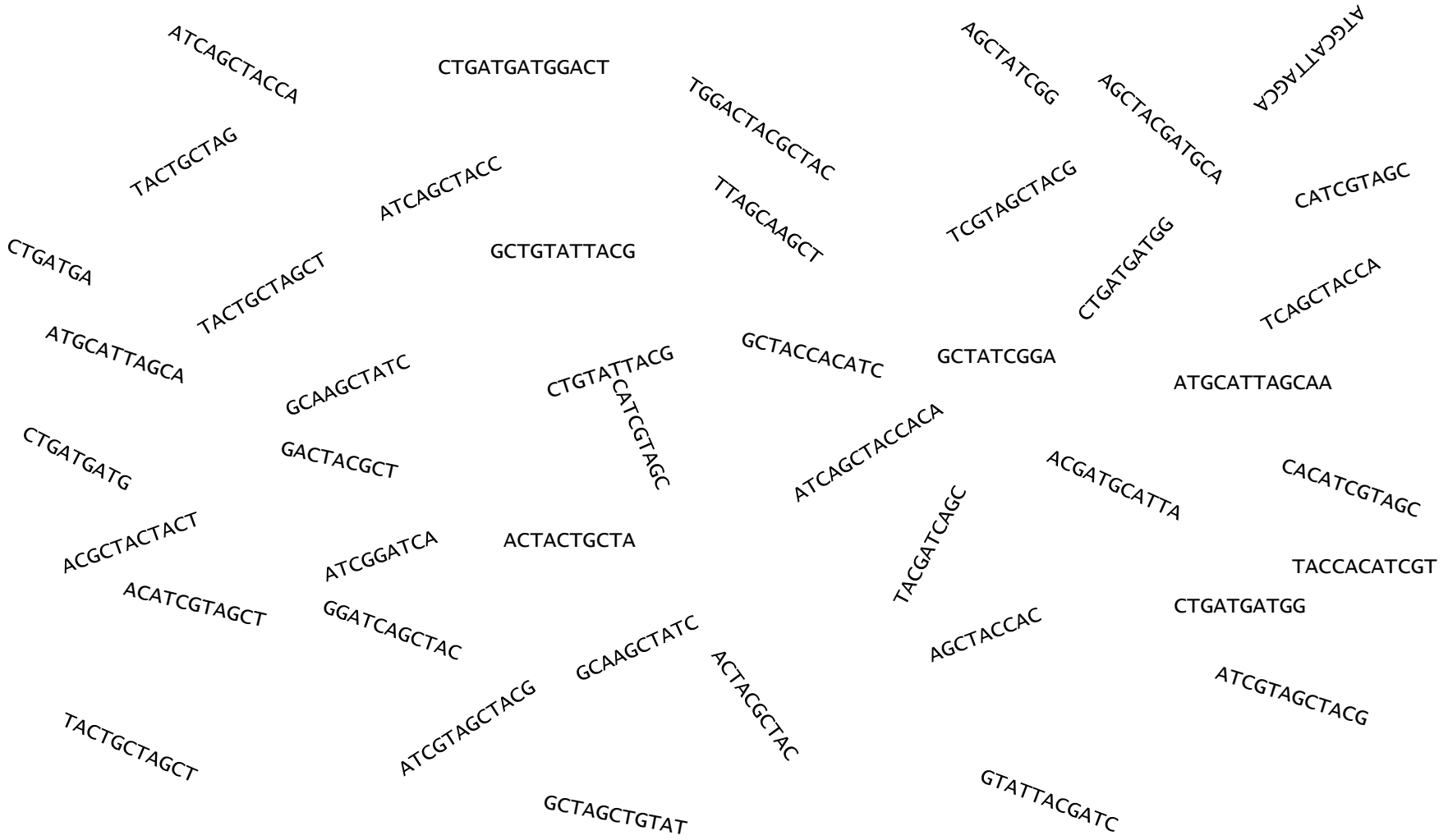
CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC

CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC

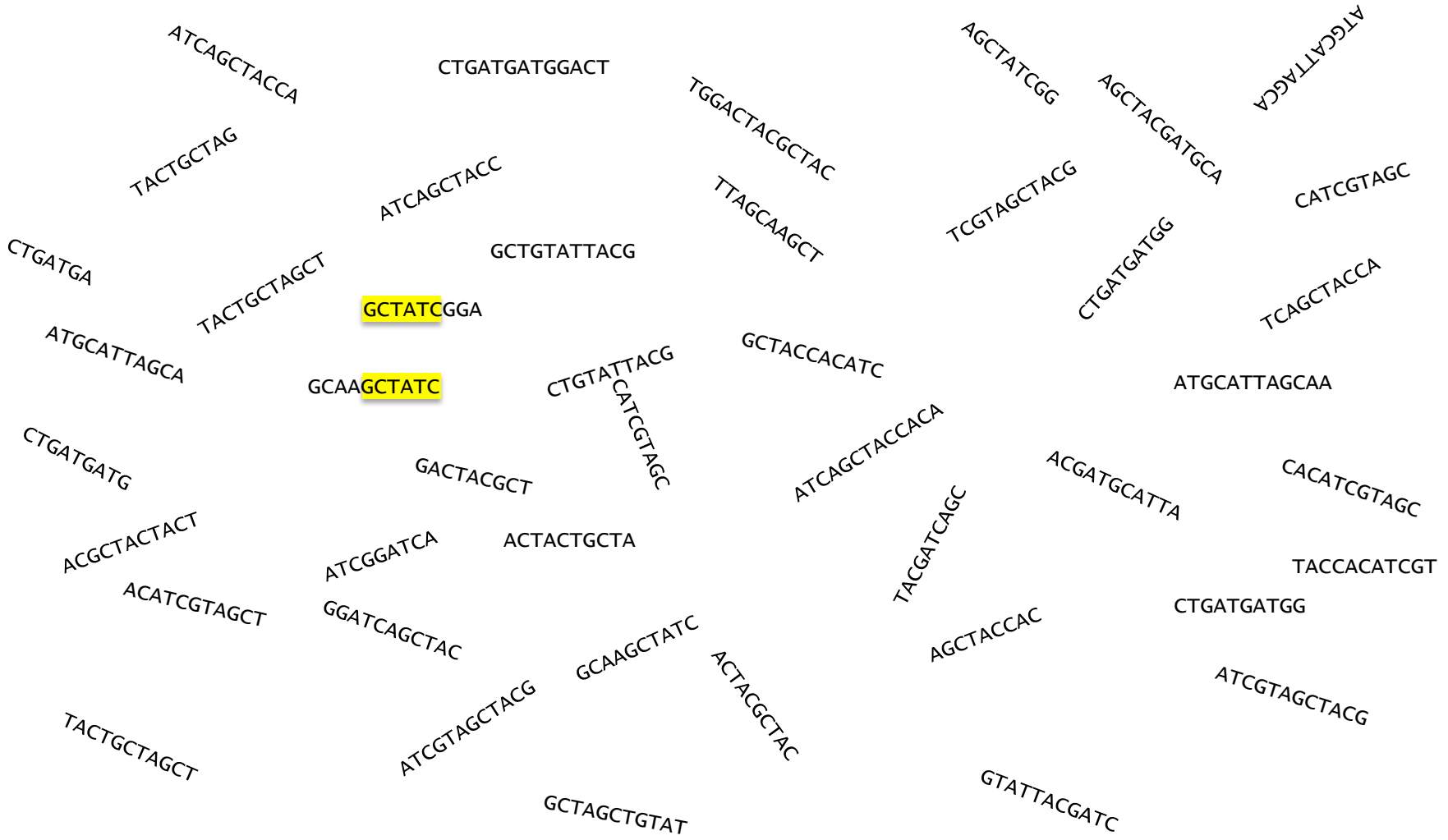
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC

CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

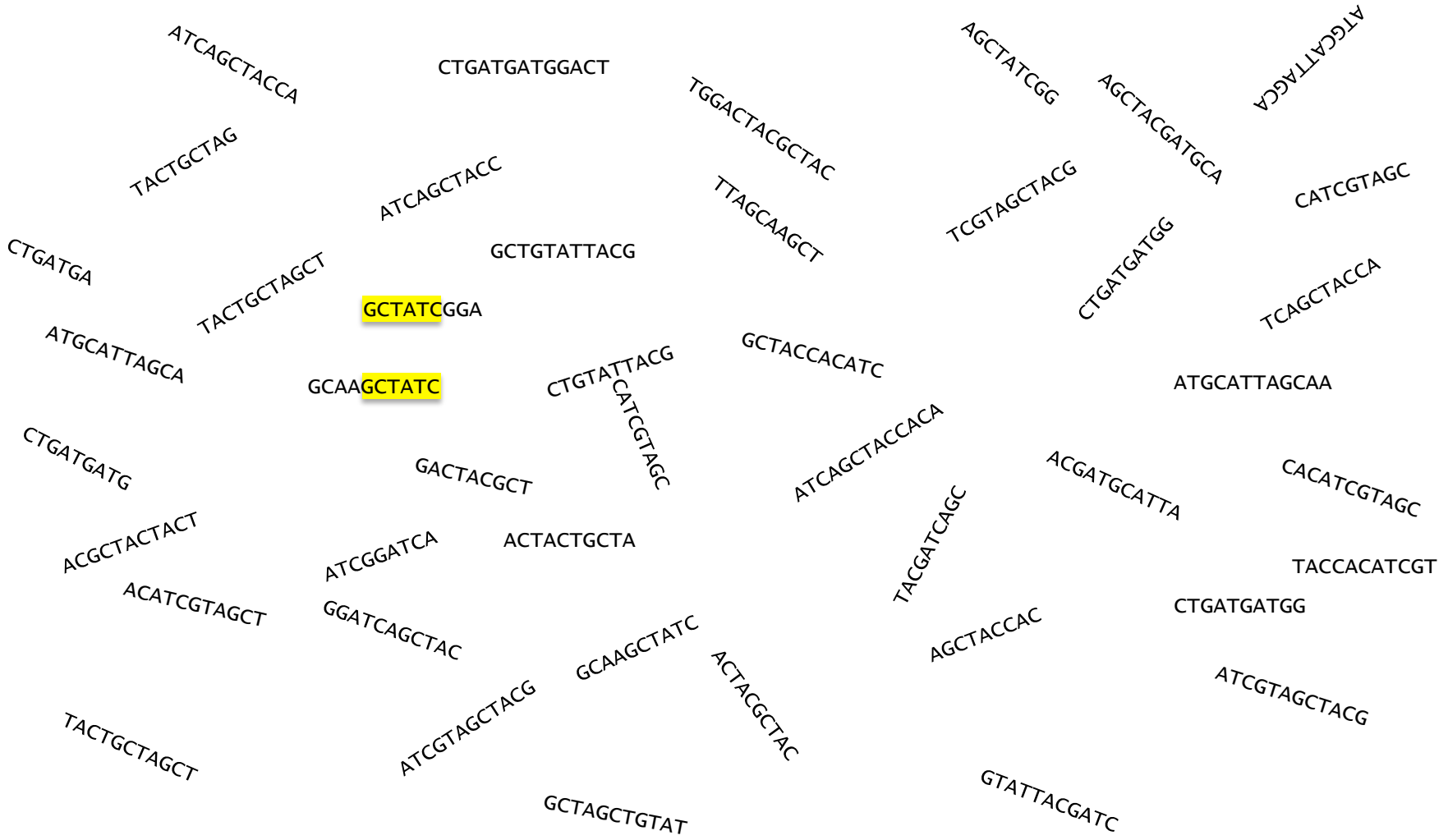
No Idea What Position Every Read Comes From



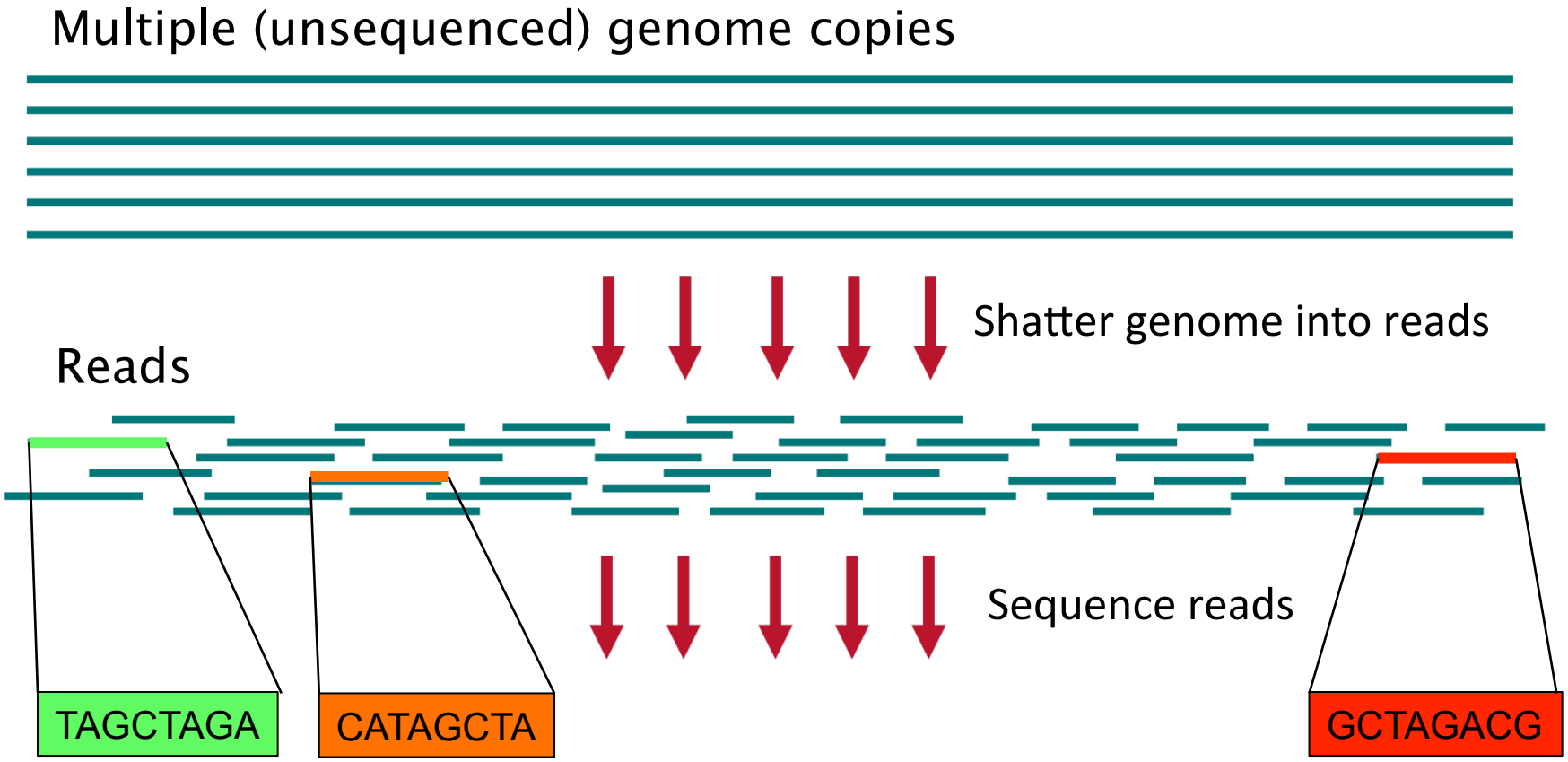
No Idea What Position Every Read Comes From



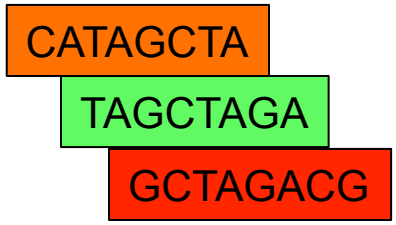
No Idea What Position Every Read Comes From



From Experimental to Computational Challenges



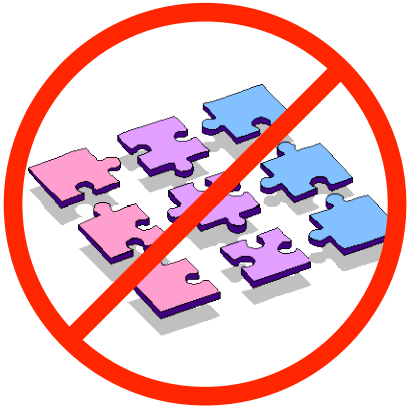
Assemble genome using overlapping reads



What Makes Genome Sequencing Difficult?

- Modern sequencing machines cannot read an entire genome one nucleotide at a time from beginning to end (like we read a book)

- They can only shred the genome and generate short **reads**.
- The genome assembly is not the same as a jigsaw puzzle: we must use *overlapping* reads to reconstruct the genome, a giant **overlap puzzle**!



Outline

- What Is Genome Sequencing?
- Exploding Newspapers
- The String Reconstruction Problem
- String Reconstruction as a Hamiltonian Path Problem
- String Reconstruction as an Eulerian Path Problem
- Similar Problems with Different Fates
- De Bruijn Graphs
- Euler's Theorem
- Assembling Read-Pairs
- De Bruijn Graphs Face Harsh Realities of Assembly

A Naive String Reconstruction Approach
Shortest superstring example

- Reads:

  {ACG, CGA, CGC, CGT, GAC, GCG, GTA, TCG}

- Shortest superstring (length 10)

  Composition_3 **(TCGACGCGTA)**

  TCG
  CGA
  GAC
  ACG
  CGC
  GCG
  CGT
  GTA

A Naive String Reconstruction Approach
Algorithms for shortest superstring problem

- Simple *greedy* strategy:

```
while # strings > 1 do
    merge two strings with maximum overlap
loop
```

- This problem turns out to be *NP*-complete
- Conjectured to give string with

  length ≤ 2 × minimum length
- "2-approximation"
- Other algorithms will require *graph theory*…

Graph Basics

- A graph (*G*) consists of vertices (*V*) and edges (*E*)

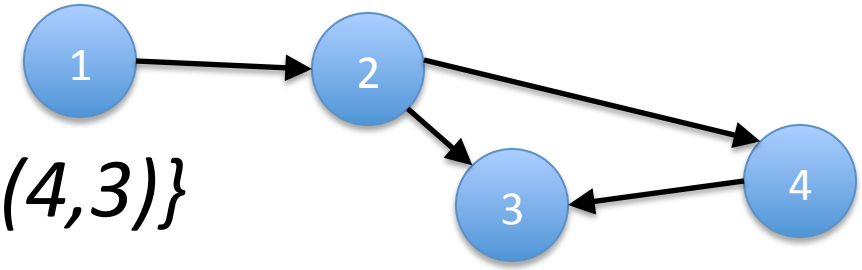
$$G = (V,E) \quad V = (v_1, v_2, v_3, v_4)$$

$$E = \{(v_1, v_2), (v_2, v_3), (v_2, v_4), (v_4, v_3)\}$$

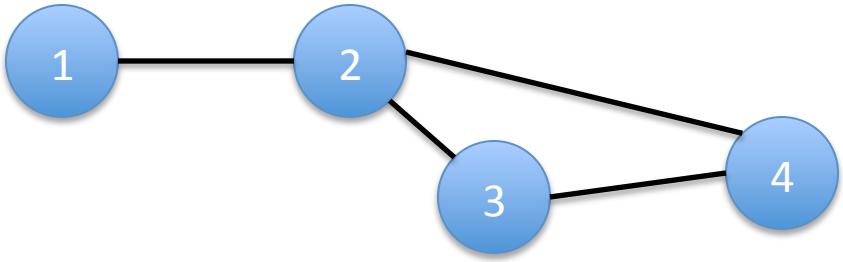
- Edges can either be *directed* (*directed graphs)*

V = (1,2,3,4)

E = {(1,2),(2,3),(2,4),(4,3)}

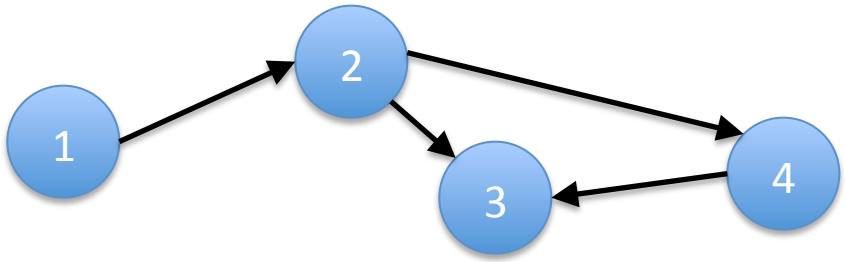


- or *undirected*



Vertex degrees

- The *degree* of a vertex: the # of edges incident to that vertex
- For directed graphs, we also have the notion of
  - *indegree:* The number incoming edges
  - *outdegree:* The number of outgoing edges



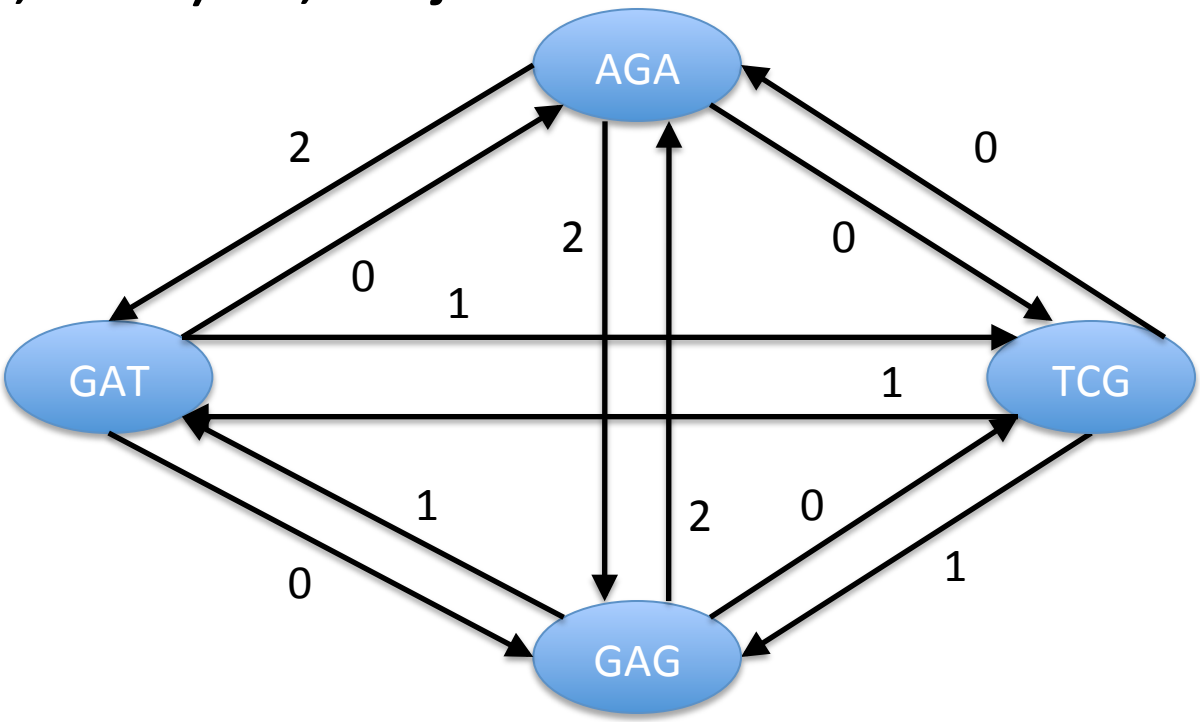
Degree(2) = 3
Indegree(2) = 1
Outdegree(2) = 2

Overlap graph

- For a set of sequence reads S, construct a directed weighted graph *G = (V,E,w)*
  - with one vertex per read (v_i corresponds *to* s_i)
  - edges between all vertices (a *complete* graph)
  - weights = $w(v_i,v_j)$ = *overlap*(s_i,s_j) = length of longest suffix of s_i that is a prefix of s_j

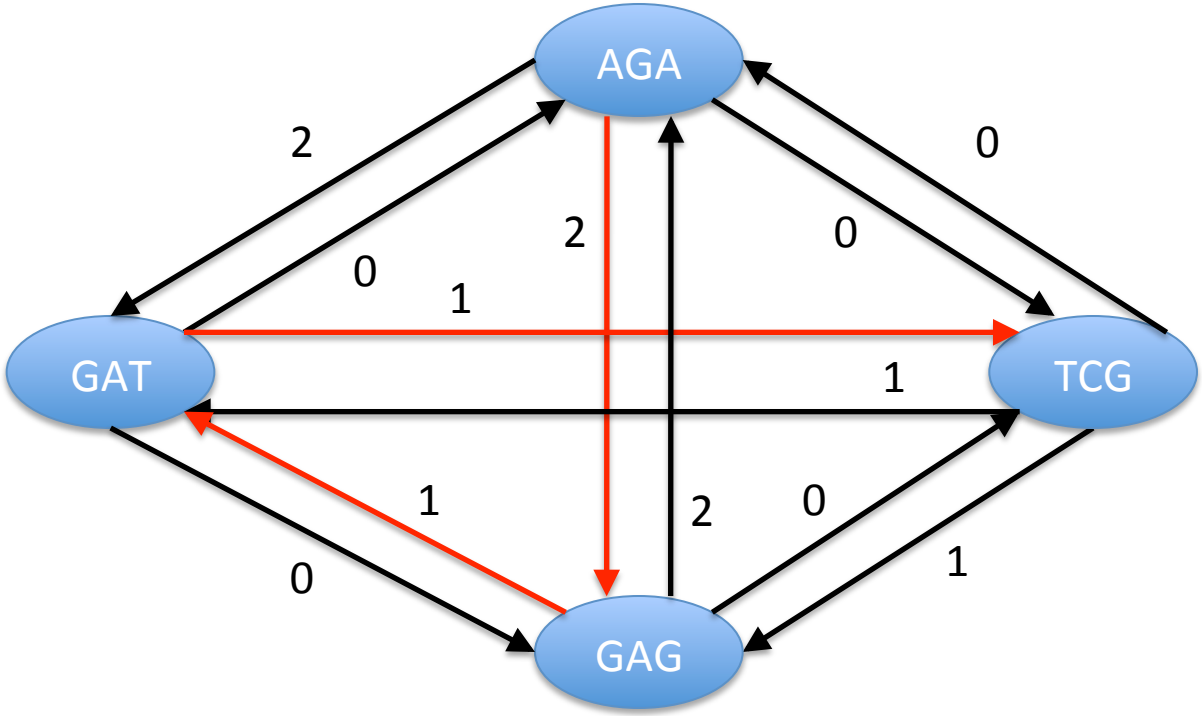
Overlap graph example

- Let S = {AGA, GAT, TCG, GAG}
- Weights of graph G: {w(AGA,GAT)=2, w(GAT,AGA)=0, ⋯}



Assembly as Hamiltonian Path

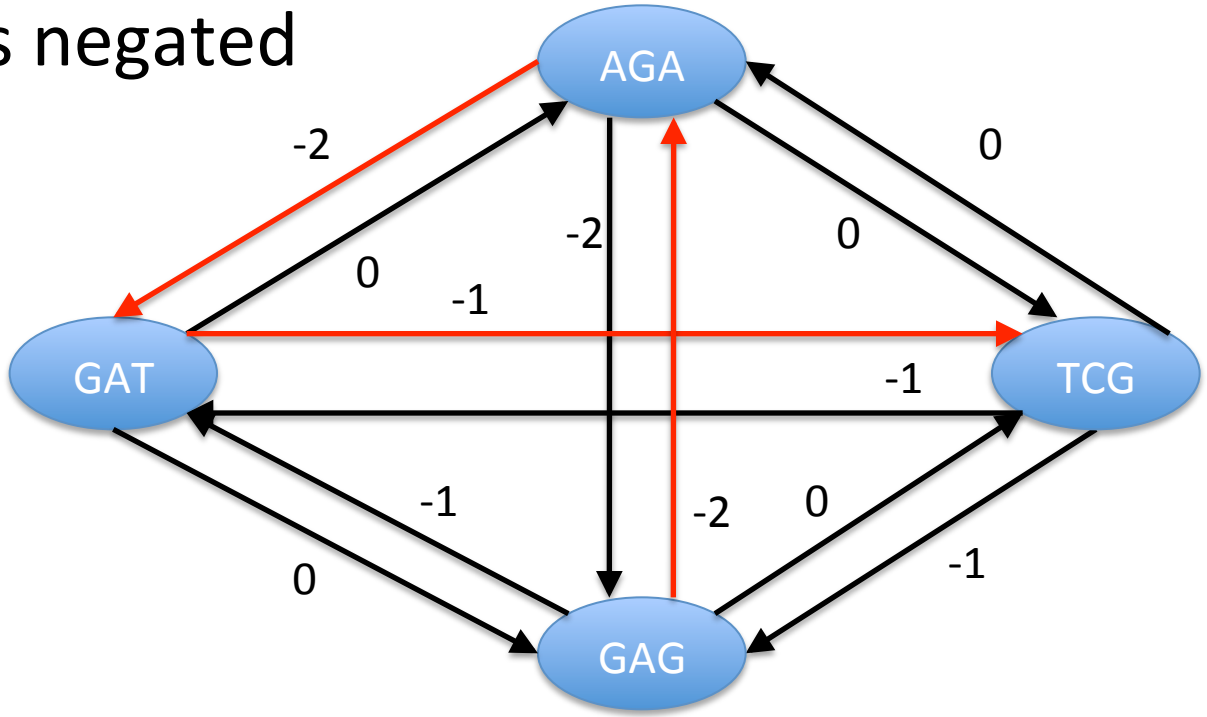
- *Hamiltonian Path*: path through graph that visits each vertex exactly once



Path: AGA->GAG->GAT->TCG Sequence: AGAGATCG

Shortest superstring as TSP

- minimize superstring length ➔minimize hamiltonian path length in overlap graph with edge weights negated



Path:
GAGATCG
Path length: -5
String length: 7

- This is essentially the Traveling Salesman Problem (also *NP*-complete)

The Genome Sequencing Problem

Genome Sequencing Problem. Reconstruct a genome from reads of similar sizes.

- Input. A collection of strings Reads.
- Output. A string Genome reconstructed from Reads.

Genome assembly is more difficult than Newspaper Problem

- DNA is double-stranded
- Sequencing errors
- Regions with no reads coverage

For now assume all reads come from same strand, have no errors, and have perfect coverage so every k-mer substring is generated

What Is k-mer Composition?

***Composition*$_3$(TAATGCCATGGGATGTT)=**

TAA
AAT
ATG
TGC
GCC
CCA
CAT
ATG
TGG
GGG
GGA
GAT
ATG
TGT
GTT

k-mer Composition

$Composition_3$(TAATGCCATGGGATGTT) =

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

=

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

e.g., lexicographic order (like in a dictionary)

Reconstructing a String from its Composition

String Reconstruction Problem. Reconstruct a string from its k-mer composition.

- Input. A collection of k-mers.
- Output. A Genome such that $\text{Composition}_k(\text{Genome})$ is equal to the collection of k-mers.

A Naive String Reconstruction Approach

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TAA TGC TGG TGT

A Naive String Reconstruction Approach

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA

A Naive String Reconstruction Approach

AAT ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA

A Naive String Reconstruction Approach

ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT

A Naive String Reconstruction Approach

ATG ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT

A Naive String Reconstruction Approach

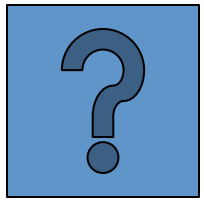
ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT
ATG

A Naive String Reconstruction Approach

ATG ATG CAT CCA GAT GCC GGA GGG GTT

TGC TGG TGT



TAA
AAT
ATG

A Naive String Reconstruction Approach

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG TGT

TAA
AAT
ATG

A Naive String Reconstruction Approach

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG

TAA
AAT
ATG
TGT

A Naive String Reconstruction Approach

ATG ATG CAT CCA GAT GCC GGA GGG GTT TGC TGG

TAA
AAT
ATG
TGT

What's Next?

ATG ATG CAT CCA GAT GCC GGA GGG TGC TGG

TAA
AAT
ATG
TGT
GTT

Outline

- What Is Genome Sequencing?
- Exploding Newspapers
- The String Reconstruction Problem
- String Reconstruction as a Hamiltonian Path Problem
- String Reconstruction as an Eulerian Path Problem
- Similar Problems with Different Fates
- De Bruijn Graphs
- Euler's Theorem
- Assembling Read-Pairs
- De Bruijn Graphs Face Harsh Realities of Assembly

Representing a Genome as a Path

Composition_3(TAATGCCATGGGATGTT)=

TAA AAT ATG TGC GCC CCA CAT ATG TGG GGG GGA GAT ATG TGT GTT

Representing a Genome as a Path

Composition_3(TAATGCCATGGGATGTT)=

Representing a Genome as a Path

Composition_3(TAATGCCATGGGATGTT)=

Can we construct this genome path without knowing the genome TAATGCCATGGGATGTT, only from its composition?

Representing a Genome as a Path

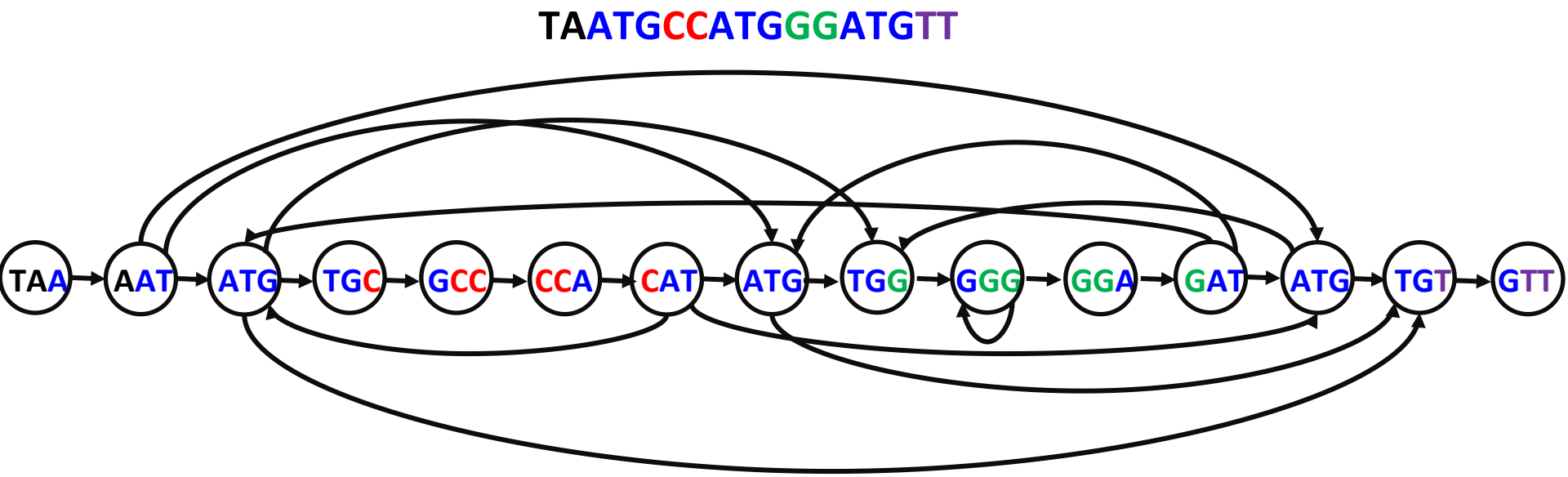
$\text{Composition}_3(\text{TAATGCCATGGGATGTT}) =$

Can we construct this genome path without knowing the genome TAATGCCATGGGATGTT, only from its composition?

Yes. We simply need to connect k-mer_1 with k-mer_2 if suffix(k-mer_1)=prefix(k-mer_2).

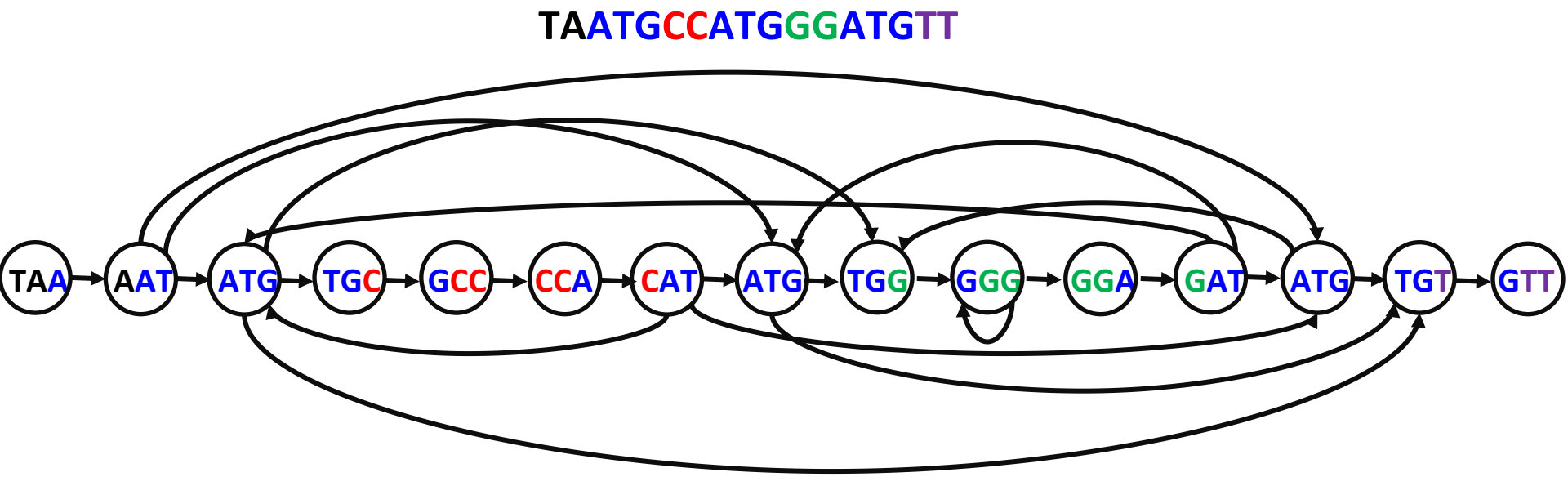
E.g. TAA → AAT

A Path Turns into a Graph



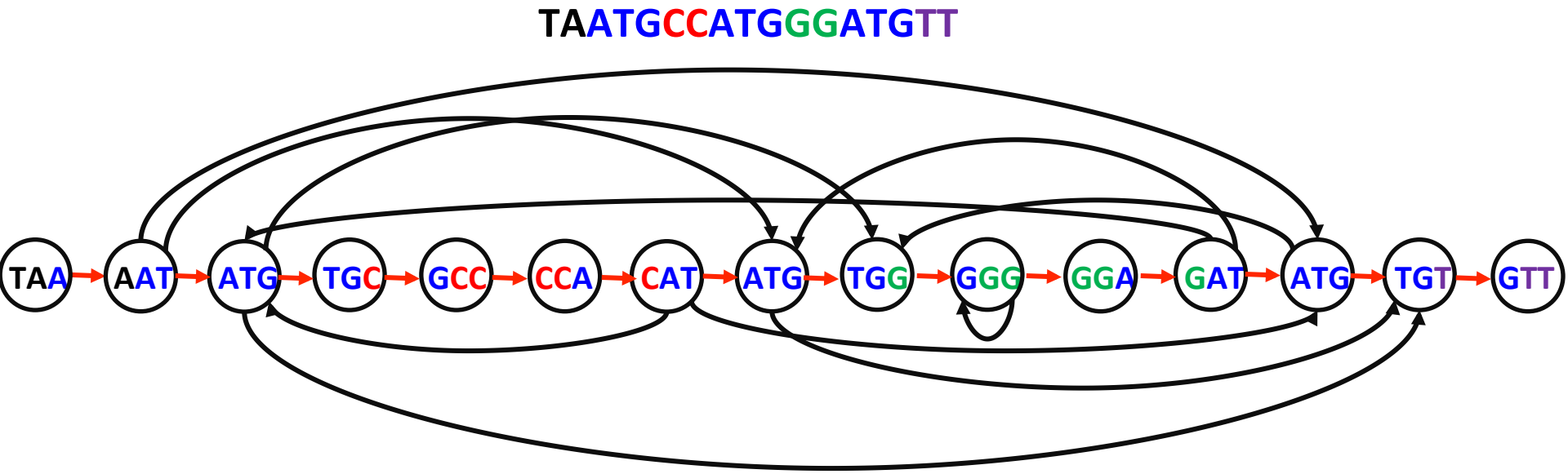
Yes. We simply need to connect k-mer$_1$ with k-mer$_2$ if suffix(k-mer$_1$)=prefix(k-mer$_2$).
E.g. TAA → AAT

A Path Turns into a Graph



Can we still find the genome path in this graph?

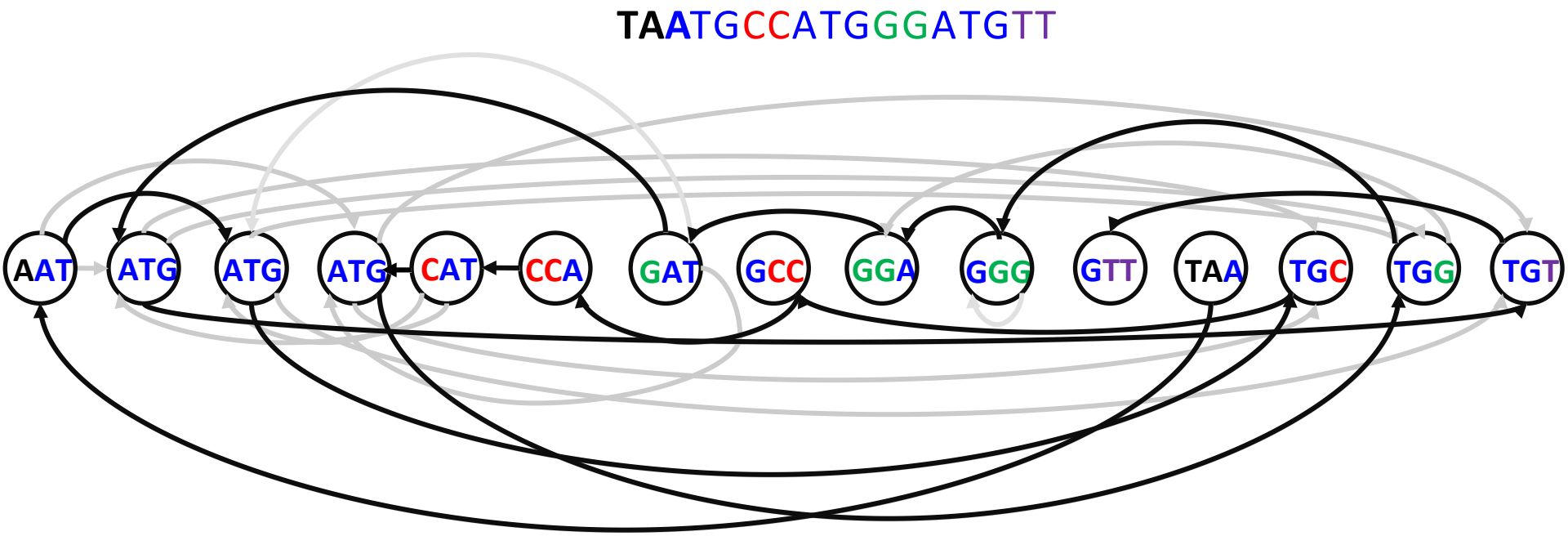
A Path Turns into a Graph



Can we still find the genome path in this graph?

Where Is the Genomic Path?

A **Hamiltonian path:** a path that visits each node in a graph exactly once.



What are we trying to find in this graph?

Hamiltonian Path Problem

Hamiltonian Path Problem. Find a Hamiltonian path in a graph.

- Input. A graph.
- Output. A path visiting every node in the graph exactly once.

Does This Graph Have a Hamiltonian Path?

Hamiltonian Path Problem. Find a Hamiltonian path in a graph.
Input. A graph.
Output. A path visiting every node in the graph exactly once.

Icosian game (1857)

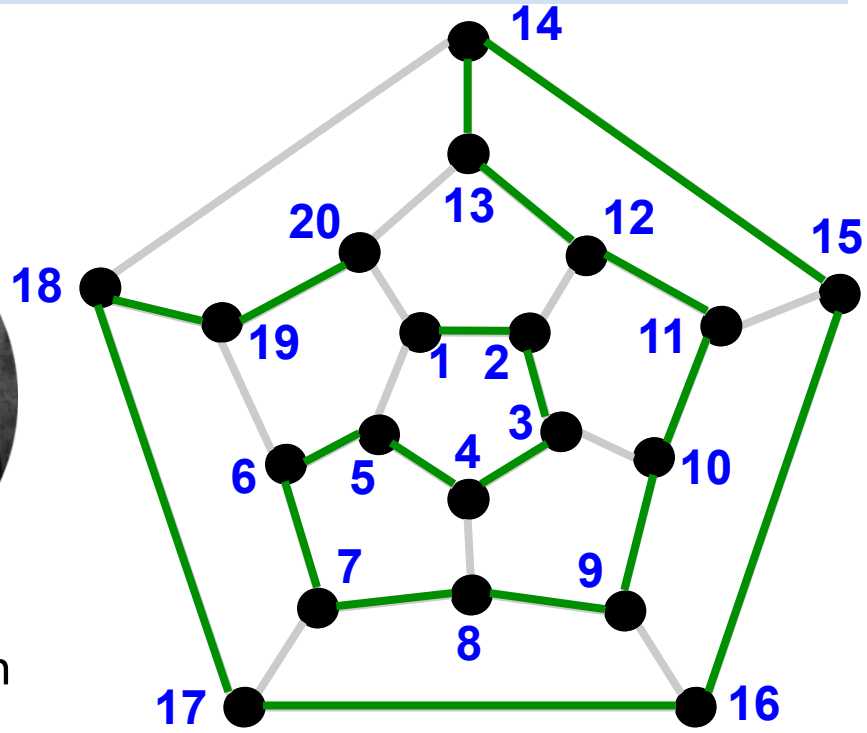
William Hamilton

Does This Graph Have a Hamiltonian Path?

Hamiltonian Path Problem. Find a Hamiltonian path in a graph.
Input. A graph.
Output. A path visiting every node in the graph exactly once.

Icosian game (1857)

William Hamilton



Undirected graph