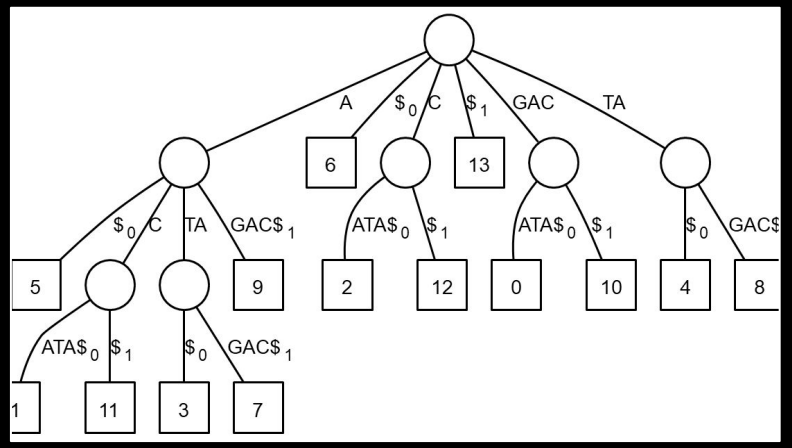
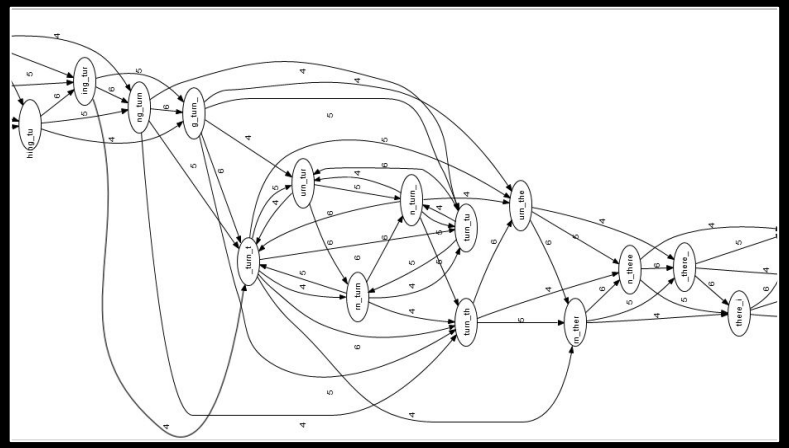
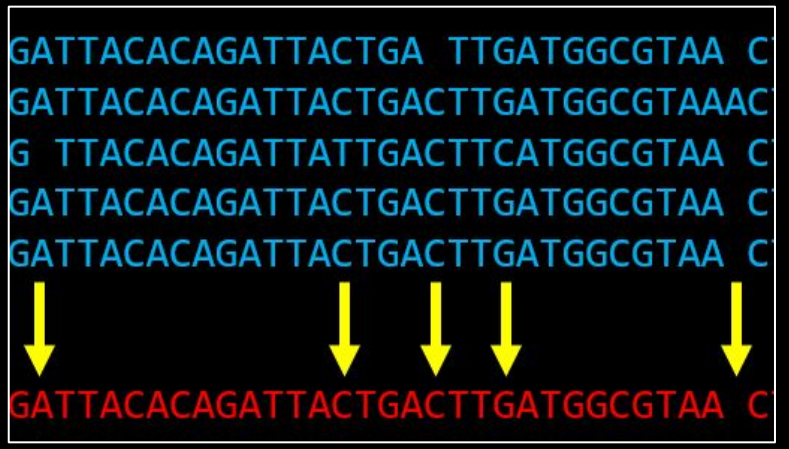
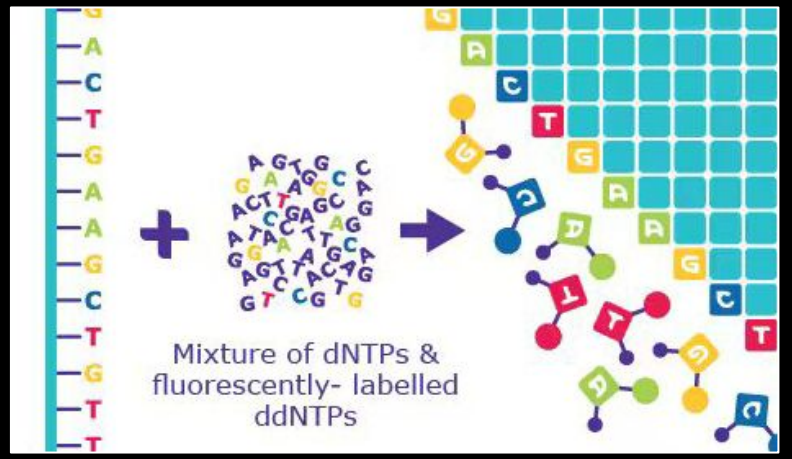
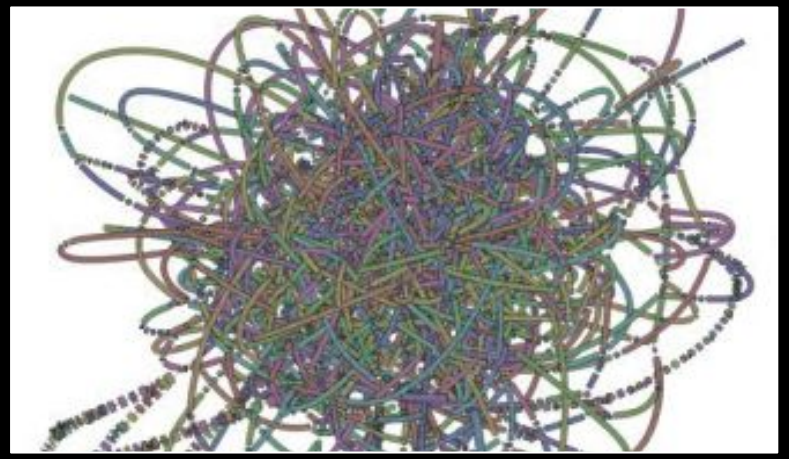
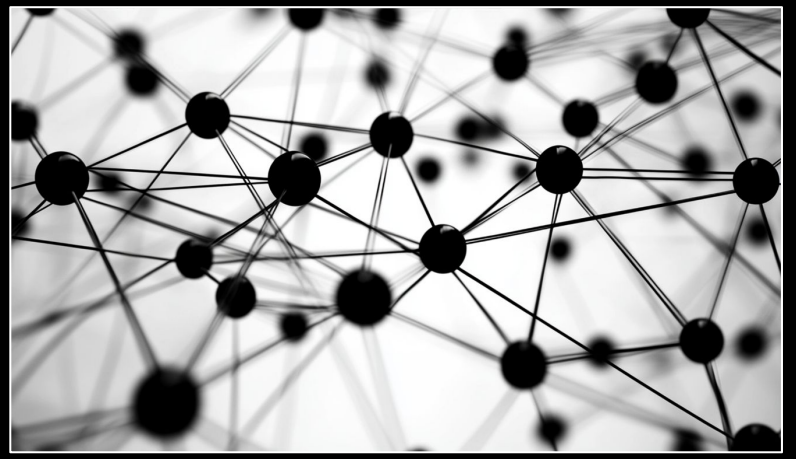




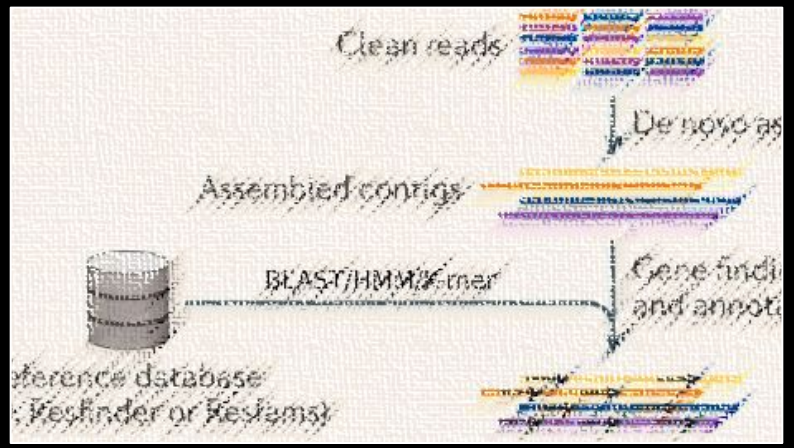








Module 2



Assembly

Module Summary

- The DNA molecules we want to sequence can be long (>>> 1 million nucleotides)

- DNA sequencing technology is great for this, but:
  - Most methods only give us short pieces
  - Merging these short pieces is not trivial
  - Sequencing can be error prone

- We need sequence assembly algorithms to handle this
  - Typically graph based: overlap-consensus (today), de Bruijn (next)

Slides adapted from Fin Maguire
Who adapted most slides from Ben Langmead's Teaching Materials
(www.langmead-lab.org/teaching.html)

Assembly: Overlap Layout Consensus

- or -

Genomic Frustration == Algorithmic Opportunities

Overview

- DNA sequencing technologies
- The assembly problem
- Graph representations
- The overlap-layout-consensus method

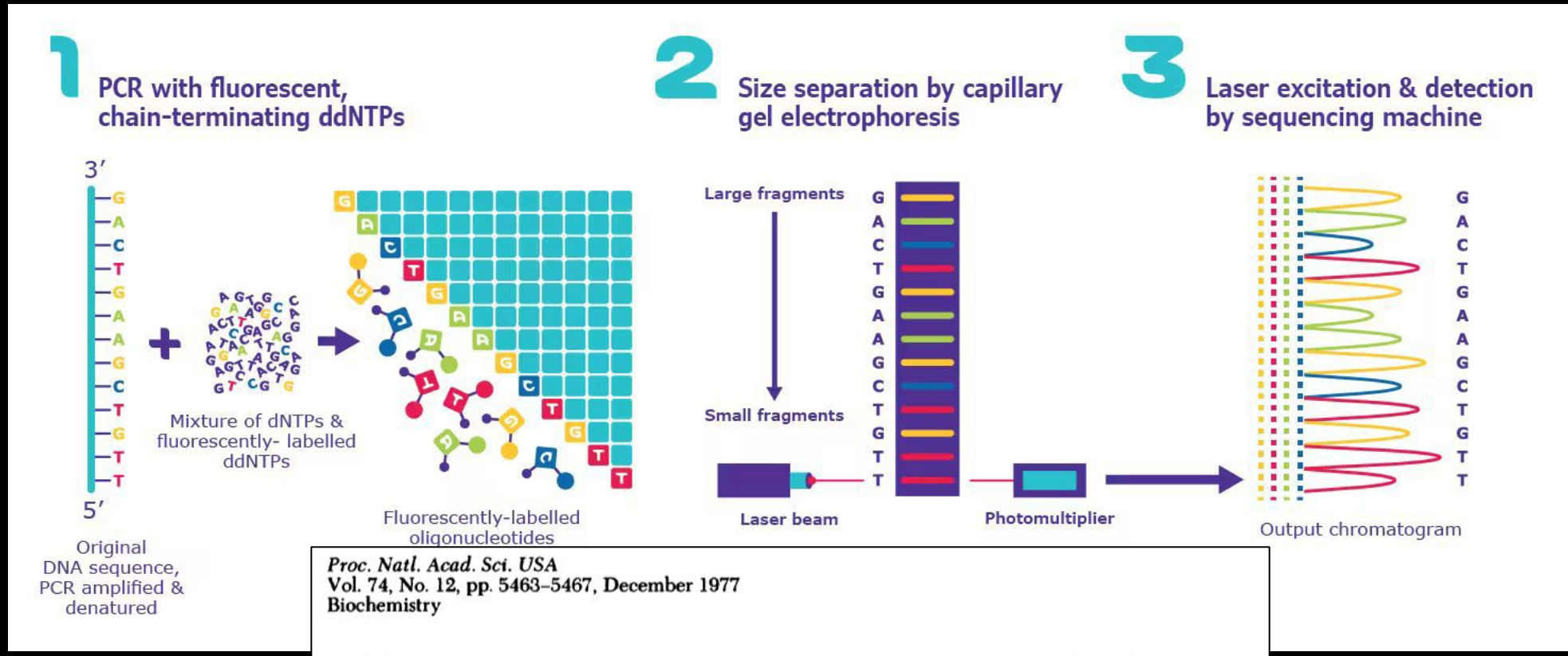
What Do You Want to Sequence?

- One gene?
- A few specific genes?
- A section of a chromosome?
- An entire chromosome?
- The entire genome?
- Just the parts that are being transcribed?

- AND -

- How much money do you have?
- How accurate does the sequence need to be?

Great moments in DNA Sequencing:
#1 - Sanger chain-termination sequencing (1977)



Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

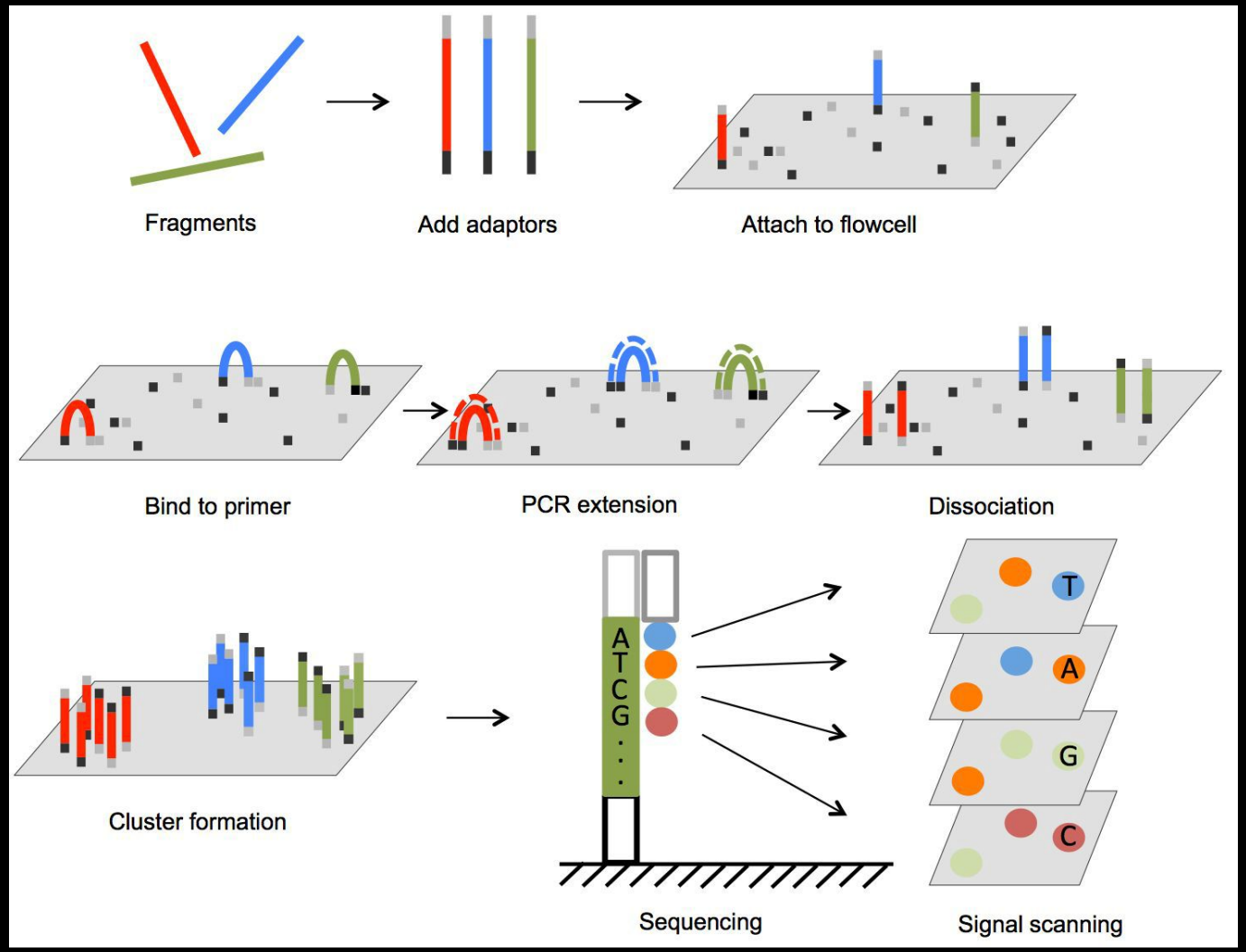
F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

https://www.sigmaaldrich.com/CA/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

Great moments in DNA Sequencing:
#2 – Sequencing by synthesis (<2008)



Vol 456|6 November 2008|doi:10.1038/nature07517 nature

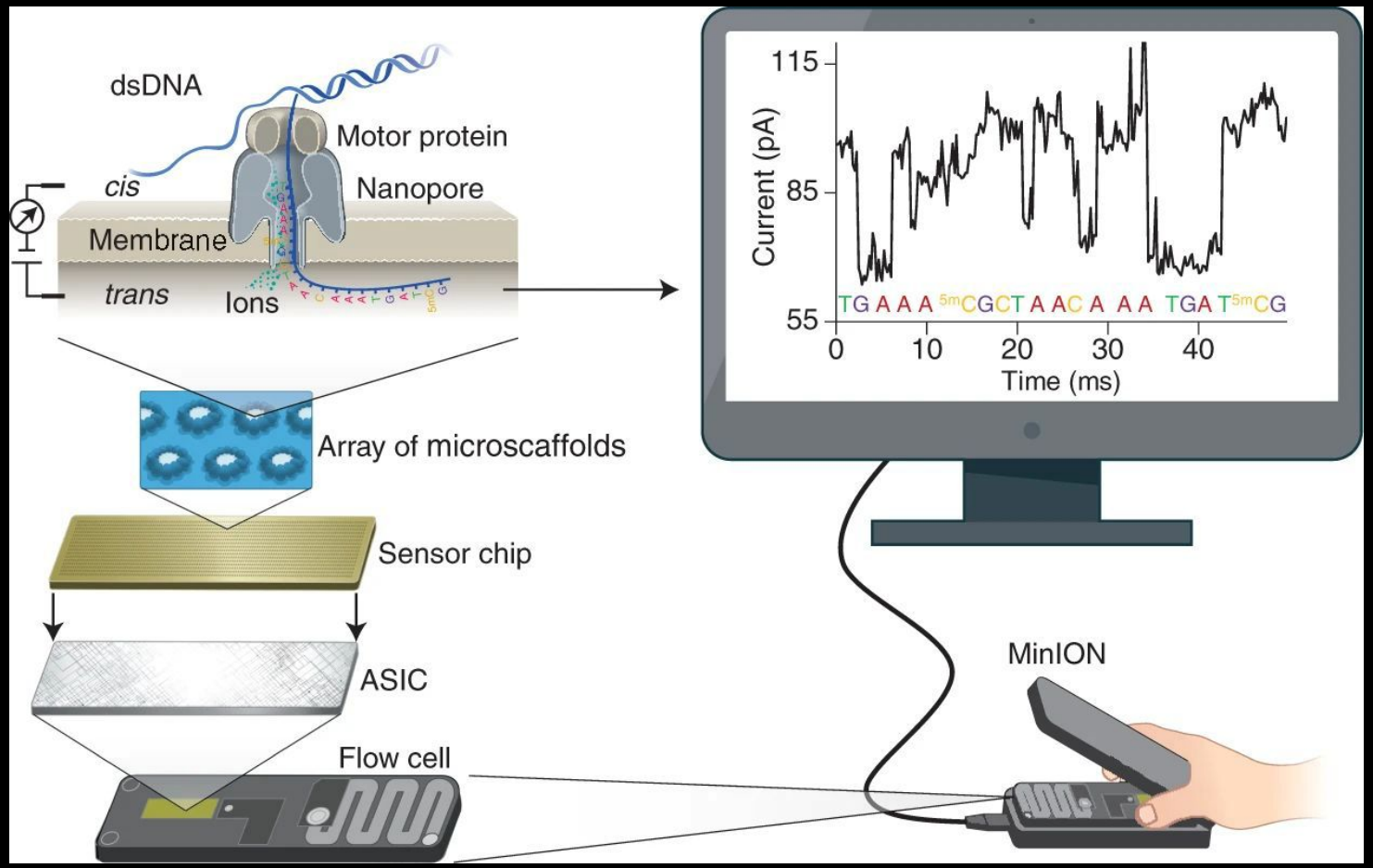
ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

A list of authors and their affiliations appears at the end of the paper

https://www.intechopen.com/chapters/49419

Great moments in DNA Sequencing:
#3 – Nanopore sequencing (2009)



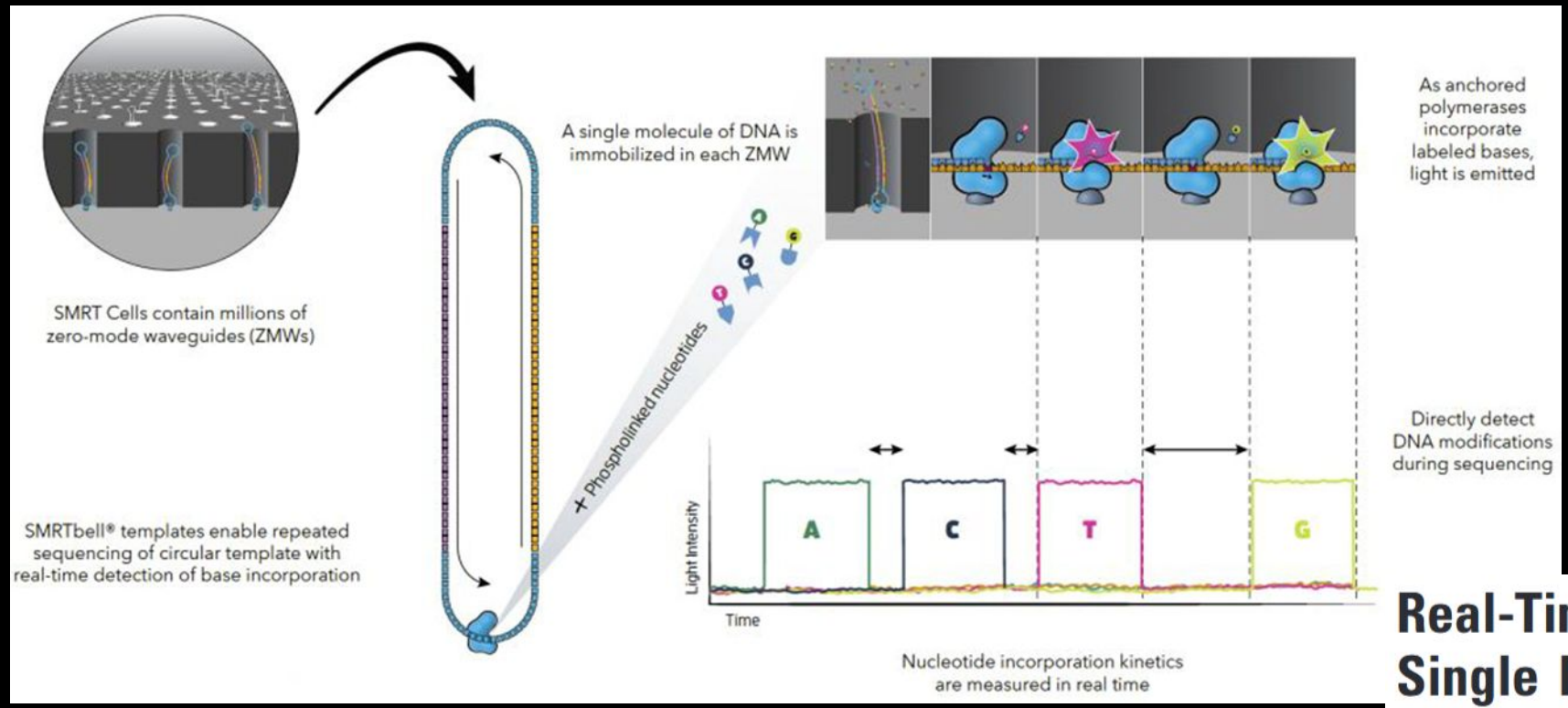
nature nanotechnology
ARTICLES
PUBLISHED ONLINE: 22 FEBRUARY 2009 | DOI: 10.1038/NNANO.2009.12

Continuous base identification for single-molecule nanopore DNA sequencing

James Clarke[1], Hai-Chen Wu[2], Lakmal Jayasinghe[1,2], Alpesh Patel[1], Stuart Reid[1] and Hagan Bayley[2]*

https://www.nature.com/articles/s41587-021-01108-x/figures/1

Great moments in DNA Sequencing:
#4 – Single-molecule real time (SMRT) sequencing (2009)



Real-Time DNA Sequencing from Single Polymerase Molecules

John Eid,* Adrian Fehr,* Jeremy Gray,* Khai Luong,* John Lyle,* Geoff Otto,* Paul Peluso,* David Rank,* Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach,† Stephen Turner†

Eid et al. (2009) *Science*
https://www.nature.com/articles/s41587-021-01108-x/figures/1

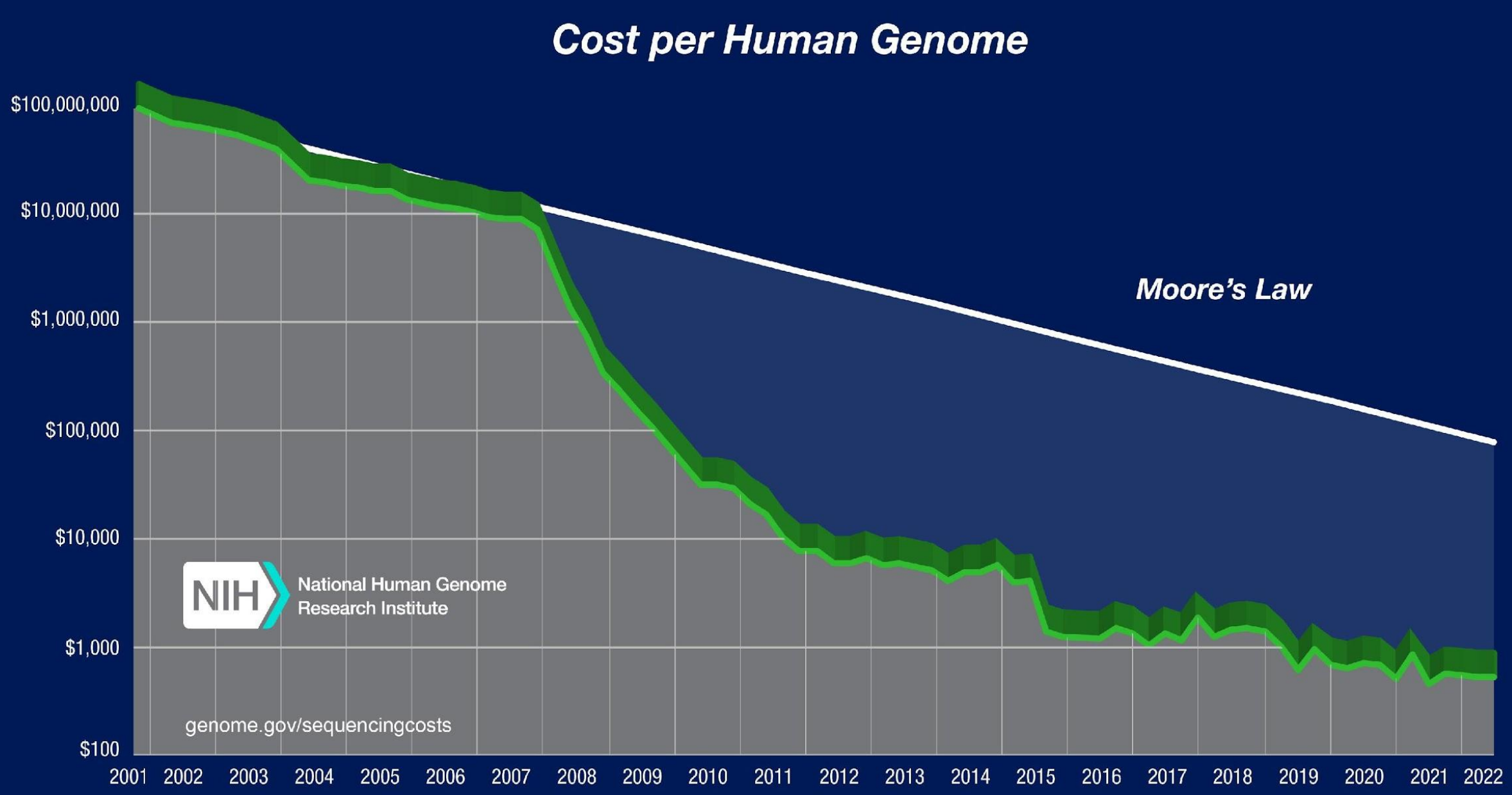
These (+ several others) are still in use

Why?

- Cost per base
- Accuracy
- Length of sequences produced at a time

all differ.

(Plus: legacy reasons)


Cost per Human Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
$100
Moore's Law
NIH
National Human Genome
Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

The Problem

- DNA sequencing techniques give you sequence reads. These are typically shorter (potentially much shorter!) than the DNA molecule you'd like to sequence

- Sanger sequencing: ~1000 bp
- Illumina sequencing: 150-300 bp
- Nanopore: Varies widely (record is over 2 million bp!)
- Pac Bio SMRT: "15,000 to 20,000 base pairs or more"

https://www.pacb.com/blog/long-read-sequencing/

Shotgun Sequencing: Fragmentation

Original population of DNA sequences:

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT

Fragmented sequences:

GGC GTCTA TATCTCGG CTCTAGGCCCTC ATTTTTT
G GCGTCTATAT CTCGGCTC TAGGCCCT CATTTTTT
GGCGTCTAT ATC TCGGCTCT AGGCCCTCATT TTTT
GGCGTCT ATA TCTCGG CTCTAGGCCCTCATTTT TT

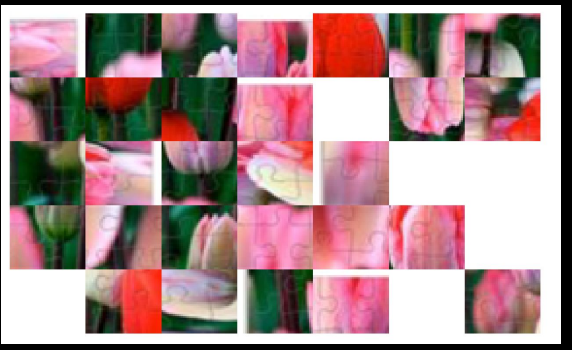
Sequence a subset of fragments at random

Assembling Shotgun Sequences

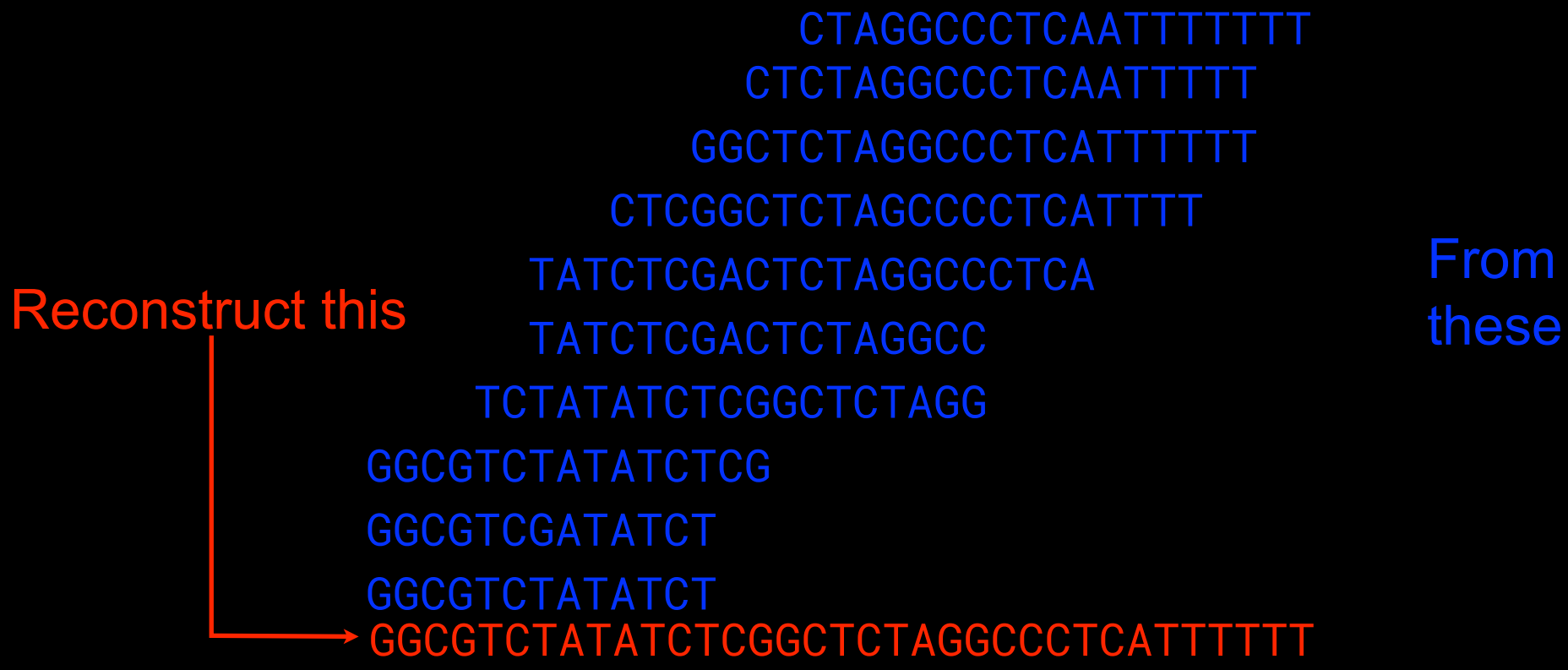
Input DNA

Small pieces

Reassemble into correct full-length sequence



Assembling the Sequence Fragments



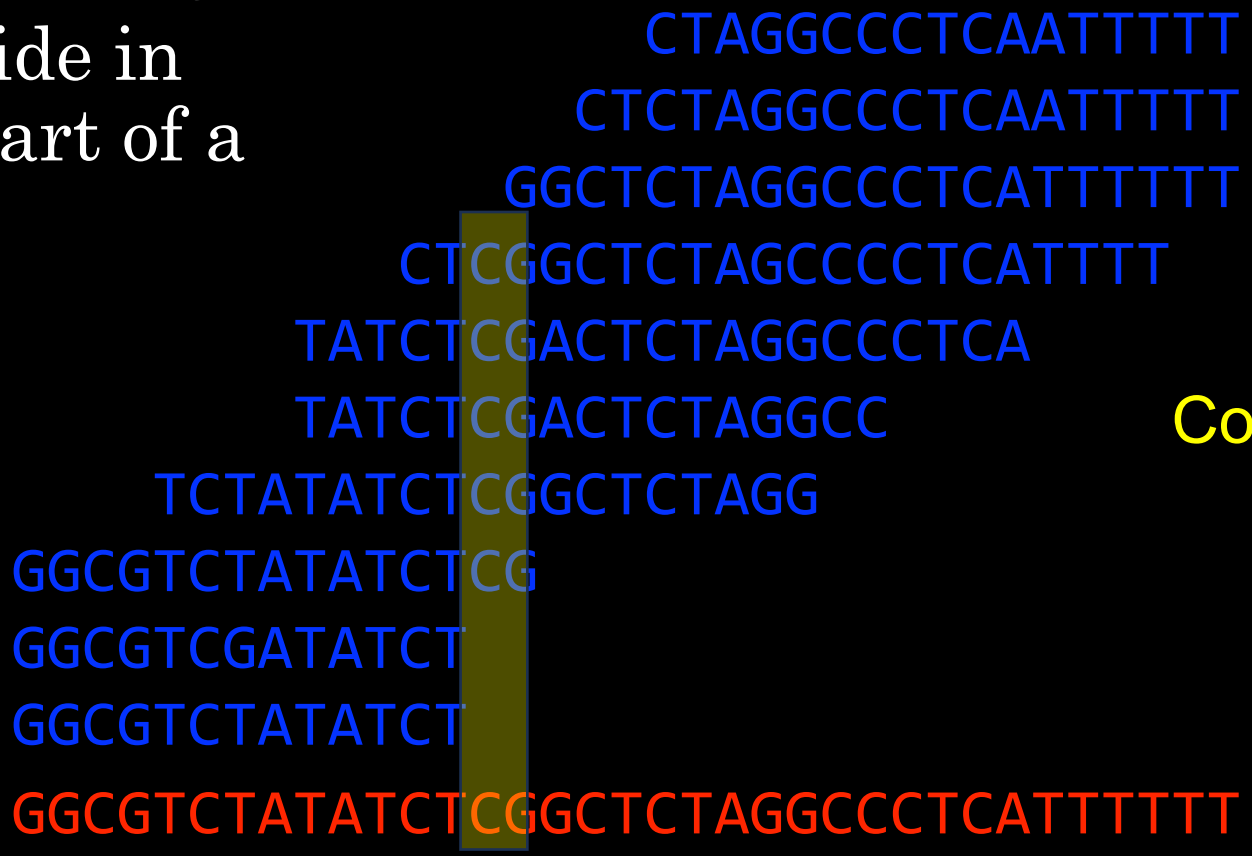
Easy!

Assembling the Sequence Fragments

maybe not so easy

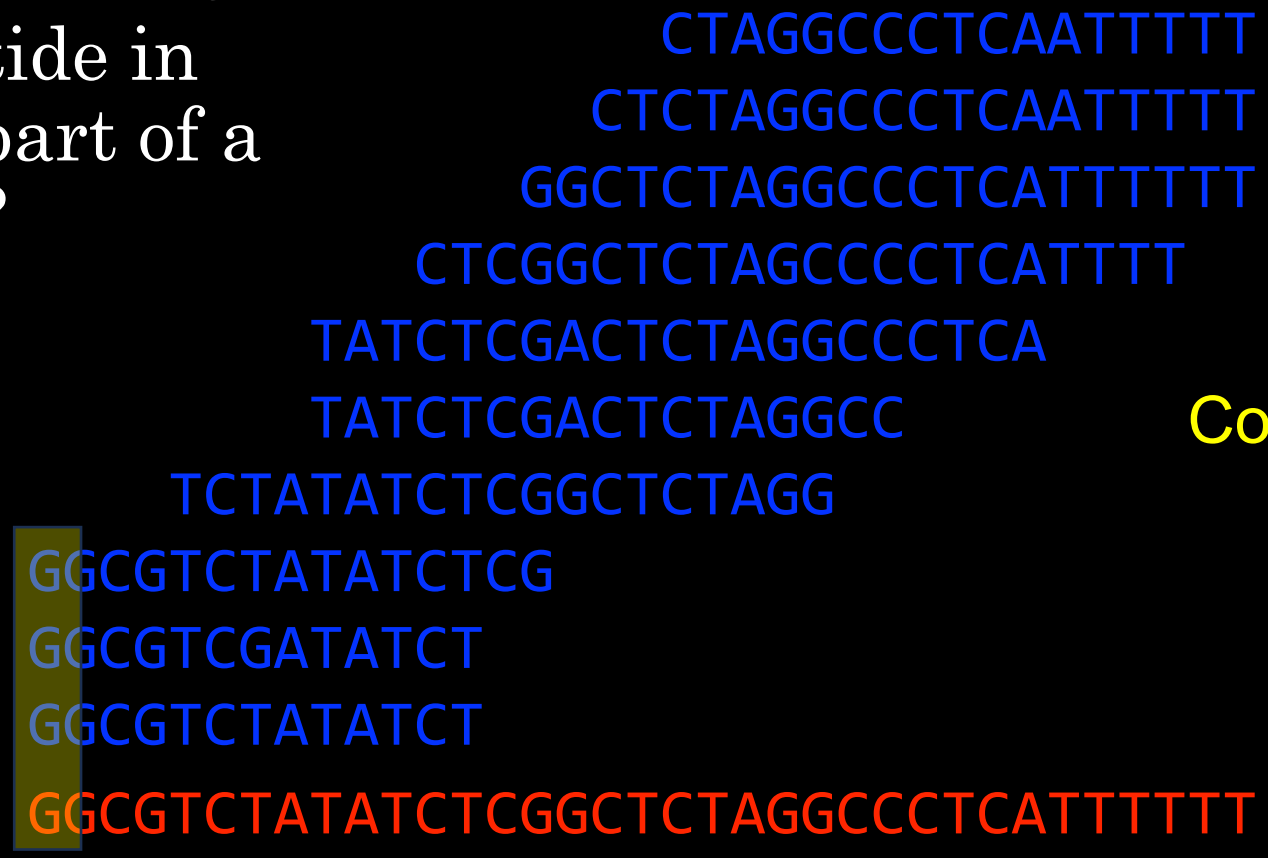
Coverage

- What is the average number of times each nucleotide in the original DNA is part of a sequenced fragment?



Coverage

- What is the average number of times each nucleotide in the original DNA is part of a sequenced fragment?



Coverage = 3

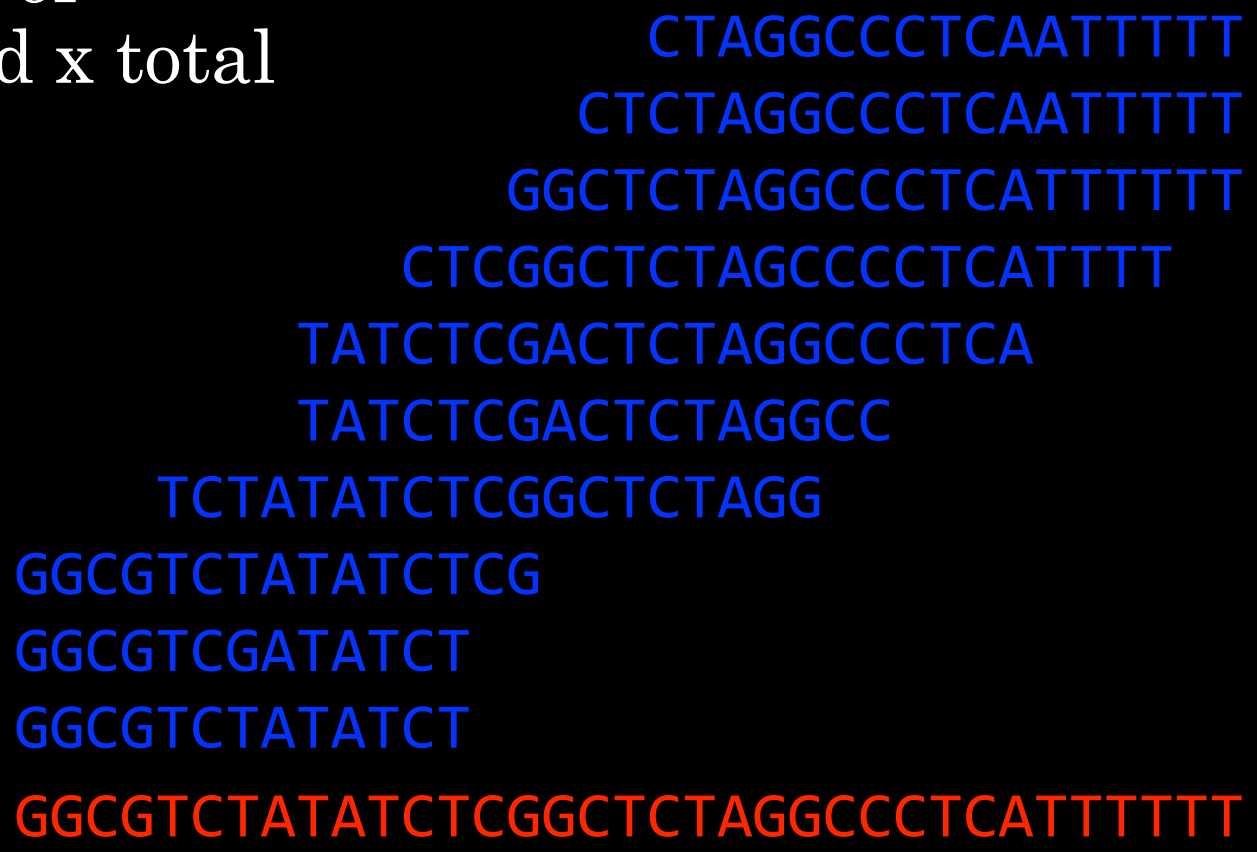
Coverage

- Average coverage = # of nucleotides sequenced x total length

Total sequenced = 177

Total length = 36

Average coverage = 4.92



Coverage

- Coverage is not uniform, so a given coverage level C will have some genomic regions with a depth > C, and others with depth < C (including some that may not be covered at all!)

- IF coverage is unbiased, then read depth will fit a Poisson distribution

- Formula for % of genome sequenced at least once:

$$P = 1 - e^{-C}$$

Coverage

Formula for % of genome sequenced at least once:

$$P = 1 - e^{-C}$$

Average coverage

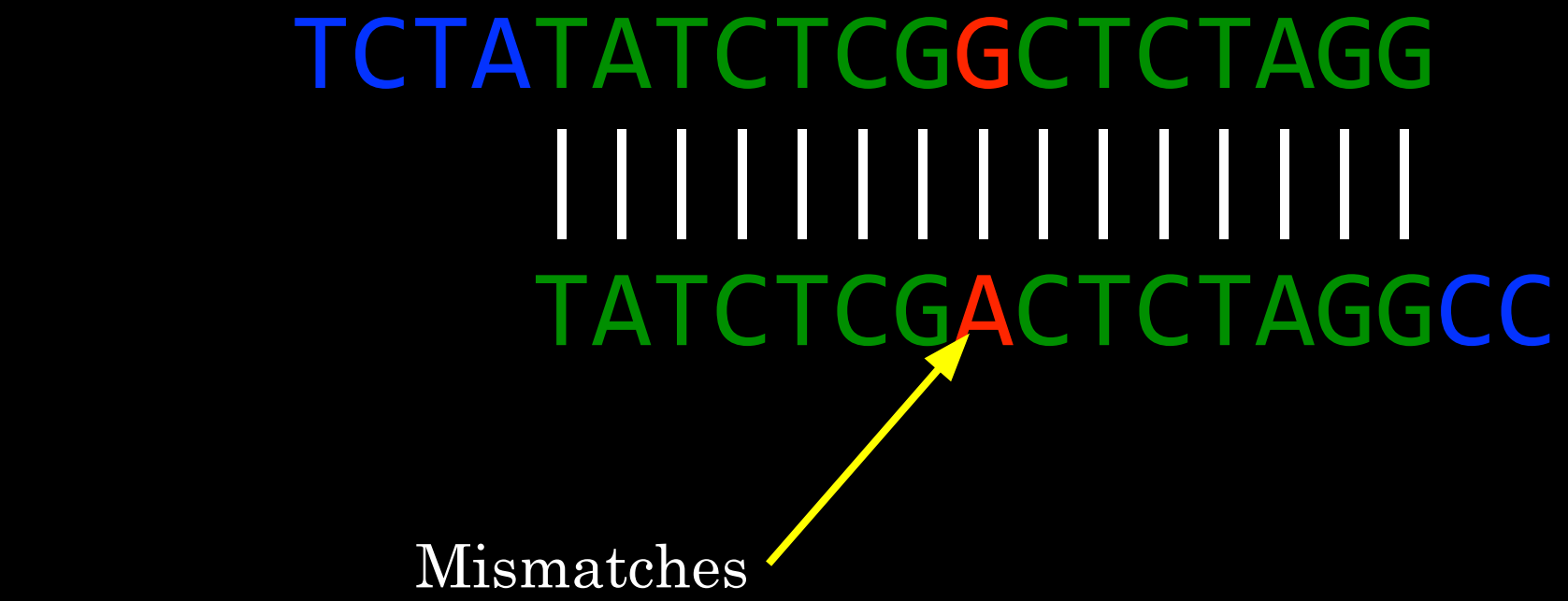
Percentage

Euler's constant: 2.718…

C = 1: P ≈ 63.2%
C = 5: P ≈ 99.3%
C = 10: P ≈ 99.995%

Typically, coverage needs to be higher so we can deal with sequencing errors, non-random read depth, assembly weirdness, …

Overlapping Reads



Overlapping Reads

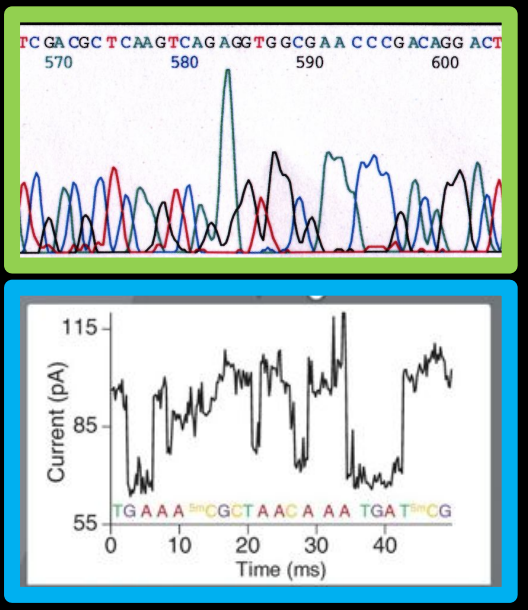
```
TCTATATCTCGGCTCTAGG
   |||||||||||||||||
    TATCTCGACTCTAGGCC
```

Why mismatches?

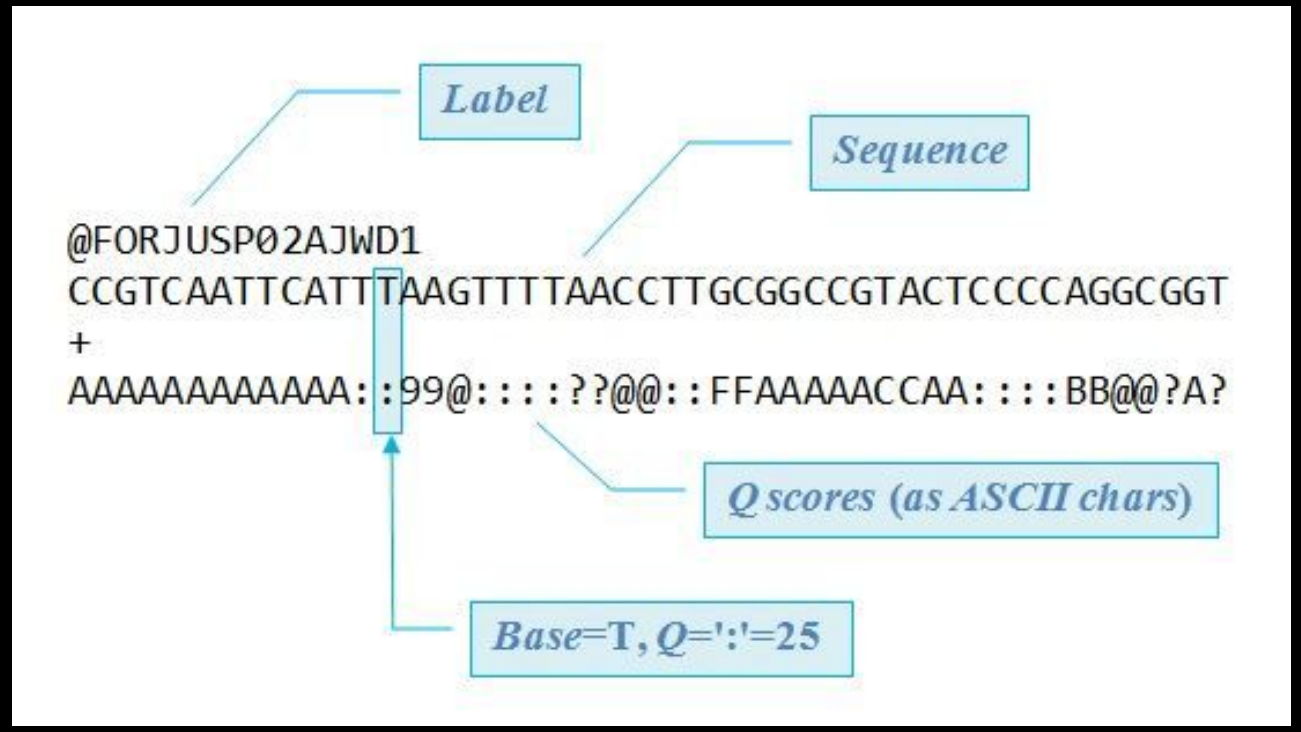
- Sequencing errors
- Variation in copies of the source DNA (ploidy)

Why Sequencing Errors?

- Misincorporation of a base
- Misreading of a base through fluorescence, pore signal, etc
- Difficulty with runs: GGGGGG → GGGGGGG
- Different sequencing technologies have different error rates and error profiles



Capturing measurement error: The FASTQ Format



Quality value Q is an integer representation of the probability p that a corresponding base call is incorrect

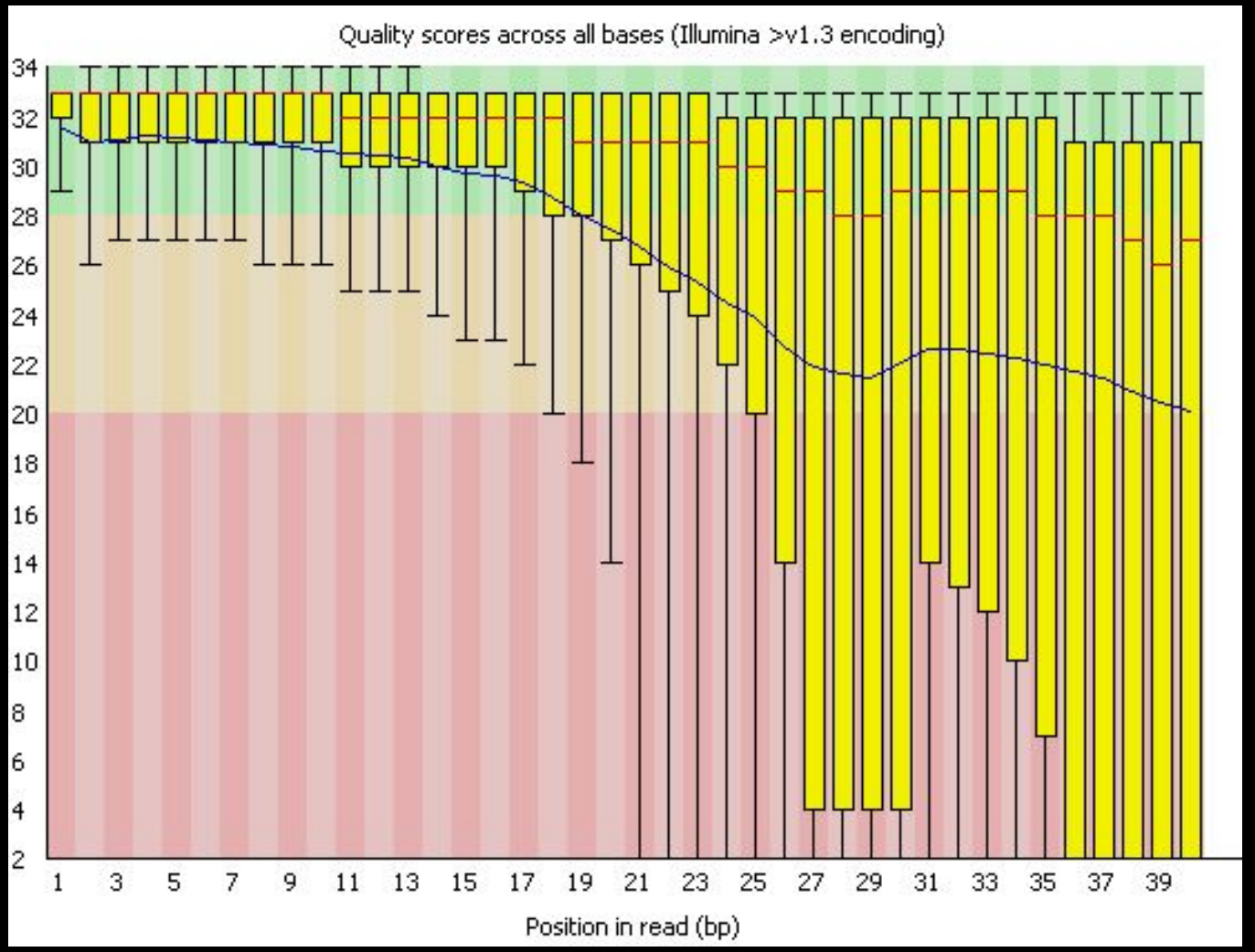
$$Q = -10 \log_{10} P \quad \Rightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

https://www.drive5.com/usearch/manual/fastq_files.html

https://learn.gencore.bio.nyu.edu/ngs-file-formats/quality-scores/

Accuracy across a sequence read



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

First Law of Assembly

If a suffix of read A is similar to a prefix of read B...

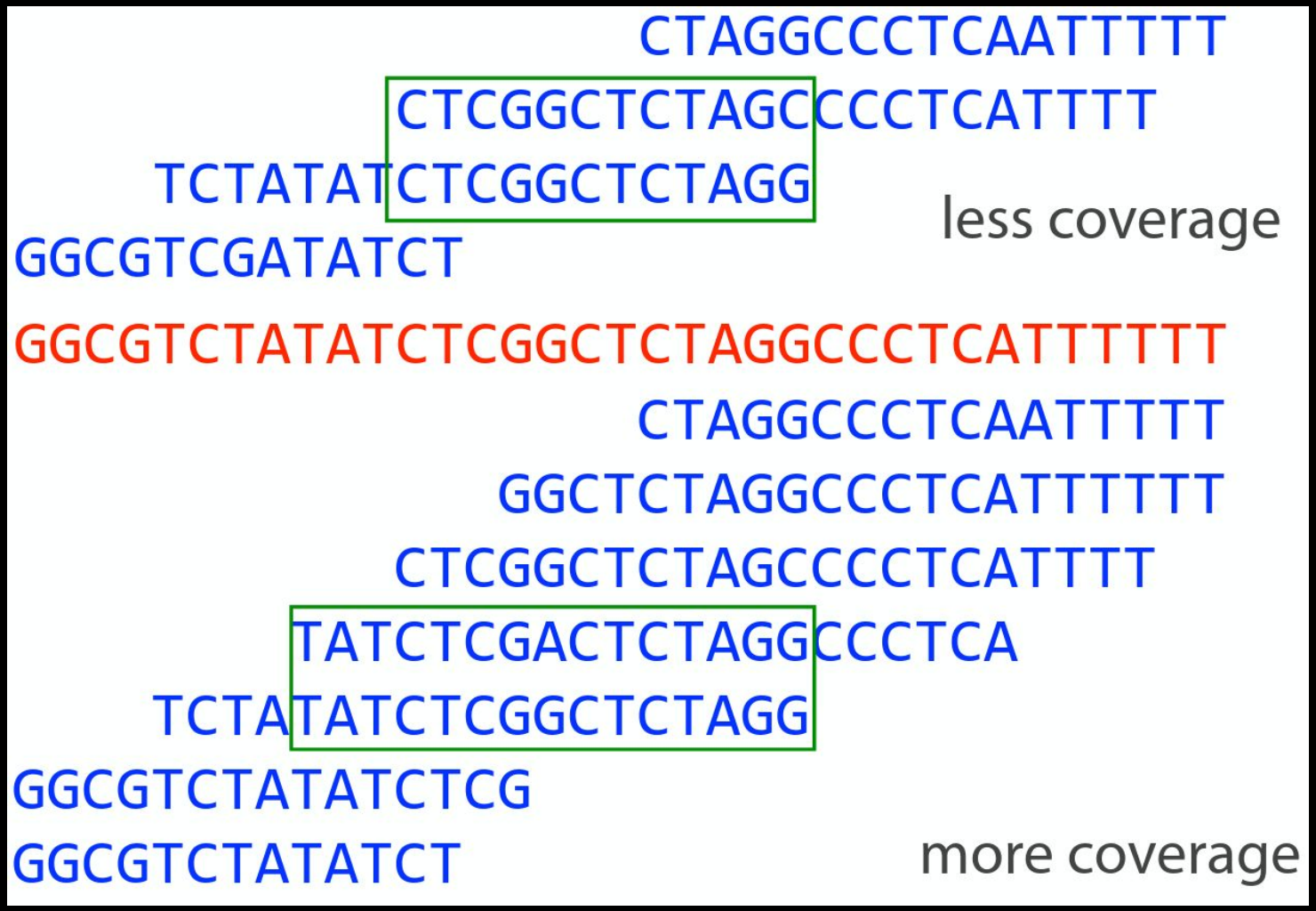
```
A: TCTATATCTCGGCTCTAGG
B:      TATCTCGACTCTAGGCC
```

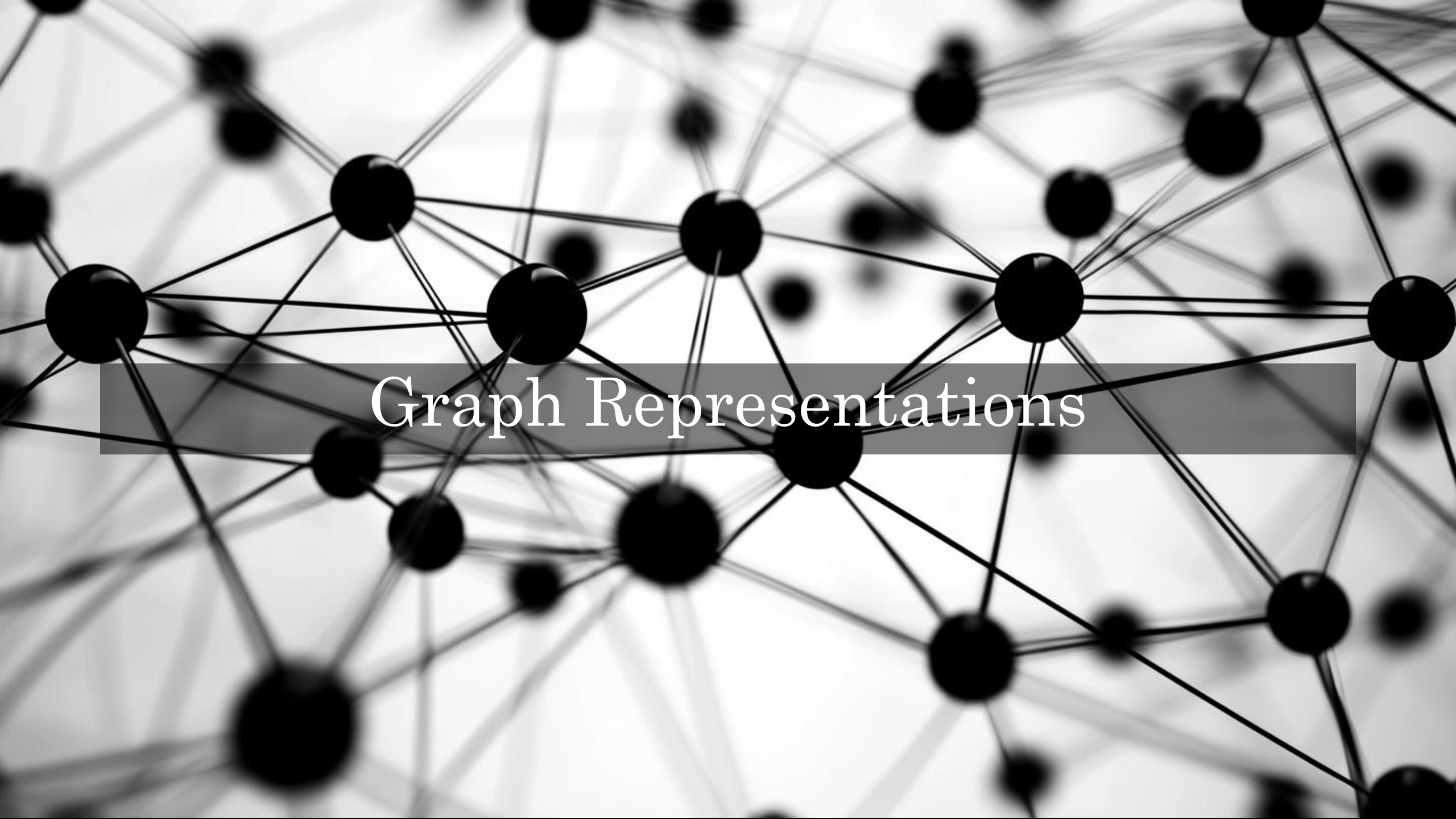
...then A and B might overlap in the genome

```
    TCTATATCTCGGCTCTAGG
GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
      TATCTCGACTCTAGGCC
```

Second Law of Assembly

More coverage leads to more and longer overlaps




Graph Representations

EXTRA! THE ONION EXTRA!

Wednesday, August 5, 1914 | The Best Source of News in Our Great Republic. | Price Two Cents.

WAR DECLARED BY ALL

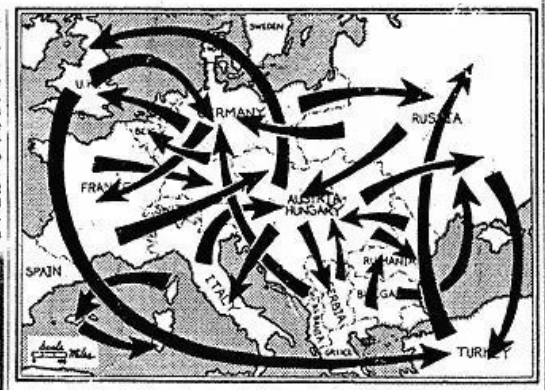
AUSTRIA DECLARES WAR ON SERBIA DECLARES WAR ON GERMANY DECLARES WAR ON FRANCE DECLARES WAR ON TURKEY DECLARES WAR ON RUSSIA DECLARES WAR ON BULGARIA DECLARES WAR ON BRITAIN

OTTOMAN EMPIRE ALMOST DECLARES WAR ON ITSELF

NATIONS STRUGGLE TO REMEMBER ALLIES

From the London and Washington Bureaus, Aug 4 —After weeks of unbearable tension following the assassination of the heir to the Austro-Hungarian throne, the great European powers have declared war on one another in a bitter struggle for something that is sure to be determined by war's end. Enmeshed in a confusing string of rivalries and alliances, diplomats, politicians, and military leaders

President Wilson, who has vowed to maintain U.S. apathy.



mighty Triple Entente alliance, which no one can really put his finger on, either.

CABLE NOTIFIES WILSON

President Woodrow Wilson received the fateful news via a trans-Atlantic cable com- not exist before the June 18th assassination of Austrian Archduke Franz Ferdinand in Sarajevo, but some of the fast-mobilizing nations seem skeptical about crediting his death as the cause of this new and all European nations who wish to obliterate each other from the face of the Earth."

'WAR TO END ALL EUROPE'

American observers are calling this great, world-wide war the "War to End All Europe," as new techno-

ASSASSINATION OF ARCHDUKE SPREADS FEAR AT ARCHDUKE CONVENTION

The Hague, Netherlands, Aug. 4.—Eurppean archdukes attending their annual convention expressed alarm at the assassination of fellow aristocrat Archduke Franz Ferdinand, whose end came at the hand of a Serbian nationalist in Sarajevo several weeks ago.

Archdukes attending the convention expressed concern for the issue of nobleman safety, once considered a birthright. Most also claimed they had no knowledge of the Serbian demand for independence from Austria. "Could it be that the oppressed minority peoples in our own provinces are weary of centuries of authoritarian rule, as well?" asked Archduke Karl von Krembelz, who cut short a month-long fox-hunting jaunt to attend the convention.

The convention, held yearly

AREA DRUNKARD DECLARES WAR ON IRELAND

ALE-HOUSE PEERS FALL IN AS ALLIES

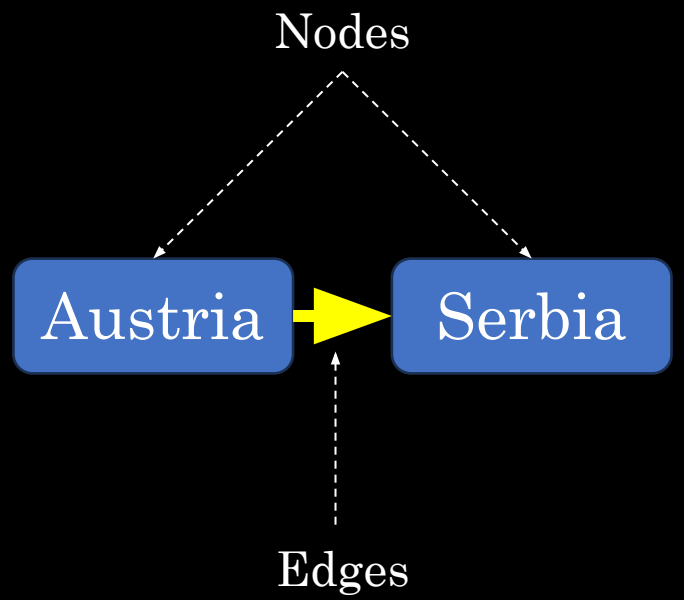
Davenport, Iowa, Aug 4. — As bloody conflict rages throughout Europe, Orvald Brunvald, a drunkard of many years' experience, declared war on Ireland Friday evening.

A RESOUNDING DECLARATION

"Consarn ever' one of 'em to damnation," Brunvald stated over a mug of porter at Quigle's Ale House in Davenport. "I'll lick 'em all to death!" he said. "Filthy potato-eating Papists, they are."

https://theonion.com/august-5-1914-1819588242/

Directed Graphs



Directed Graphs

EXTRA! THE ONION EXTRA!

Wednesday, August 5, 1914 — The Best Source of News in Our Great Republic. — Price Two Cents.

WAR DECLARED BY ALL

AUSTRIA DECLARES WAR ON SERBIA DECLARES WAR ON GERMANY DECLARES WAR ON FRANCE DECLARES WAR ON TURKEY DECLARES WAR ON RUSSIA DECLARES WAR ON BULGARIA DECLARES WAR ON BRITAIN

OTTOMAN EMPIRE ALMOST DECLARES WAR ON ITSELF

NATIONS STRUGGLE TO REMEMBER ALLIES

From the London and Washington Bureaus, Aug 4 —After weeks of unbearable tension following the assassination of the heir to the Austro-Hungarian throne, the great European powers have declared war on one another in a bitter struggle for something that is sure to be determined by war's end. Enmeshed in a confusing string of rivalries and alliances, diplomats, politicians, and military leaders

President Wilson, who has vowed to maintain U.S. apathy.

mighty Triple Entente alliance, which no one can really put his finger on, either.

CABLE NOTIFIES WILSON

President Woodrow Wilson received the fateful news via a trans-Atlantic cable com-

not exist before the June 18th assassination of Austrian Archduke Franz Ferdinand in Sarajevo, but some of the fast-mobilizing nations seem skeptical about crediting his death as the cause of this new and

all European nations who wish to obliterate each other from the face of the Earth."

'WAR TO END ALL EUROPE'

American observers are calling this great, world-wide war the "War to End All Europe," as new techno-

ASSASSINATION OF ARCHDUKE SPREADS FEAR AT ARCHDUKE CONVENTION

The Hague, Netherlands, Aug. 4.—Eurppean archdukes attending their annual convention expressed alarm at the assassination of fellow aristocrat Archduke Franz Ferdinand, whose end came at the hand of a Serbian nationalist in Sarajevo several weeks ago.

Archdukes attending the convention expressed concern for the issue of nobleman safety, once considered a birthright. Most also claimed they had no knowledge of the Serbian demand for independence from Austria. "Could it be that the oppressed minority peoples in our own provinces are weary of centuries of authoritarian rule, as well?" asked Archduke Karl von Krembelz, who cut short a month-long fox-hunting jaunt to attend the convention.

The convention, held yearly

AREA DRUNKARD DECLARES WAR ON IRELAND

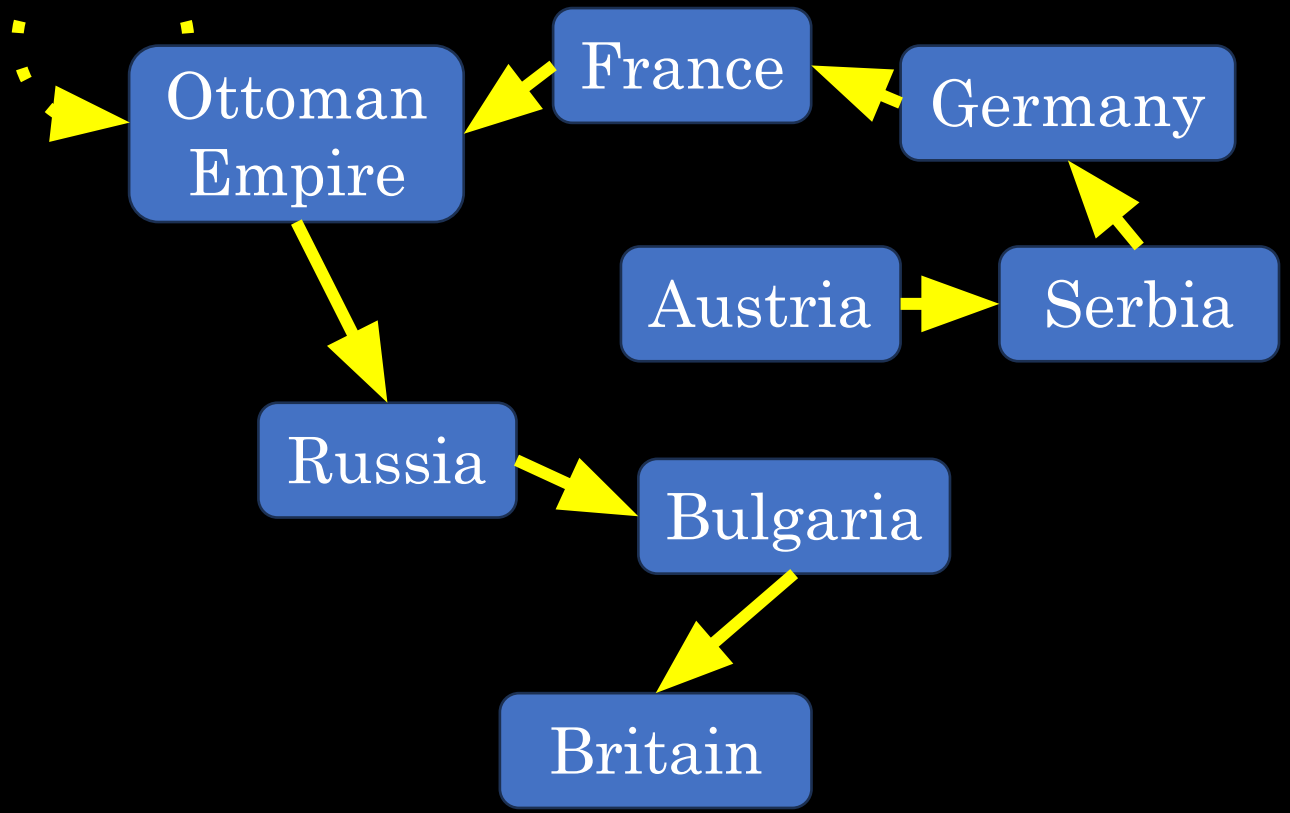
ALE-HOUSE PEERS FALL IN AS ALLIES

Davenport, Iowa, Aug 4. — As bloody conflict rages throughout Europe, Orvald Brunvald, a drunkard of many years' experience, declared war on Ireland Friday evening.

A RESOUNDING DECLARATION

"Consarn ever' one of 'em to damnation," Brunvald stated over a mug of porter at Quigle's Ale House in Davenport. "I'll lick 'em all to death!" he said. "Filthy potato-eating Papists, they are."

https://theonion.com/august-5-1914-1819588242/

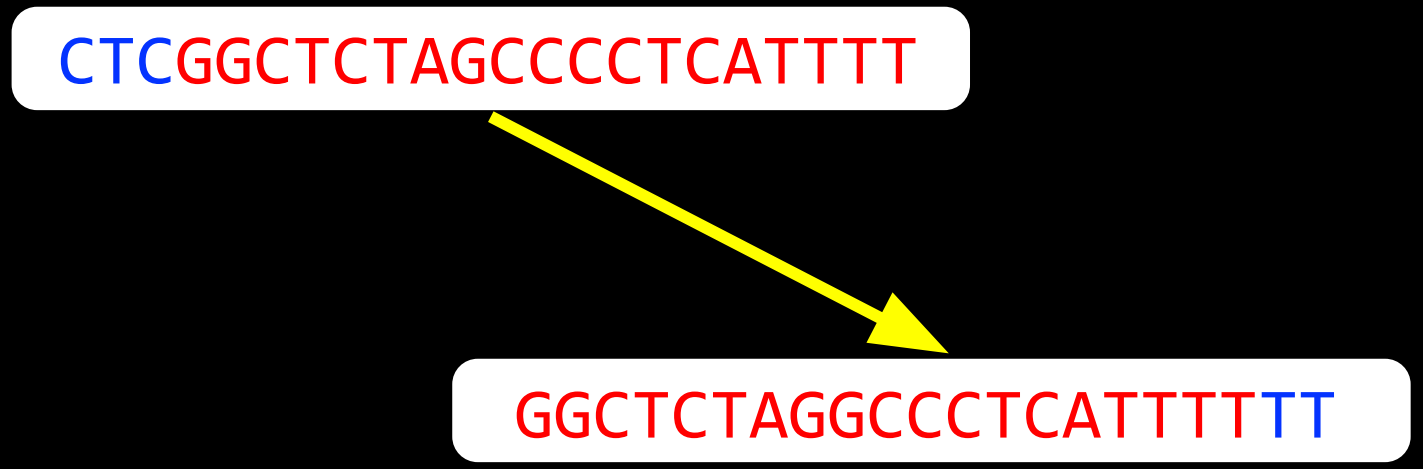


Overlap Graph

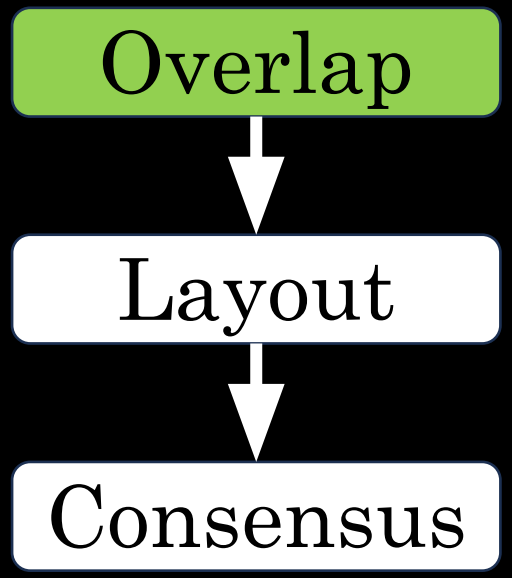
Each node is a read

CTCGGCTCTAGCCCCTCATTTT

Draw edge A -> B when suffix of A overlaps prefix of B



OLC: Building Longer Contiguous Sequences (contigs) From Reads



Build overlap graph

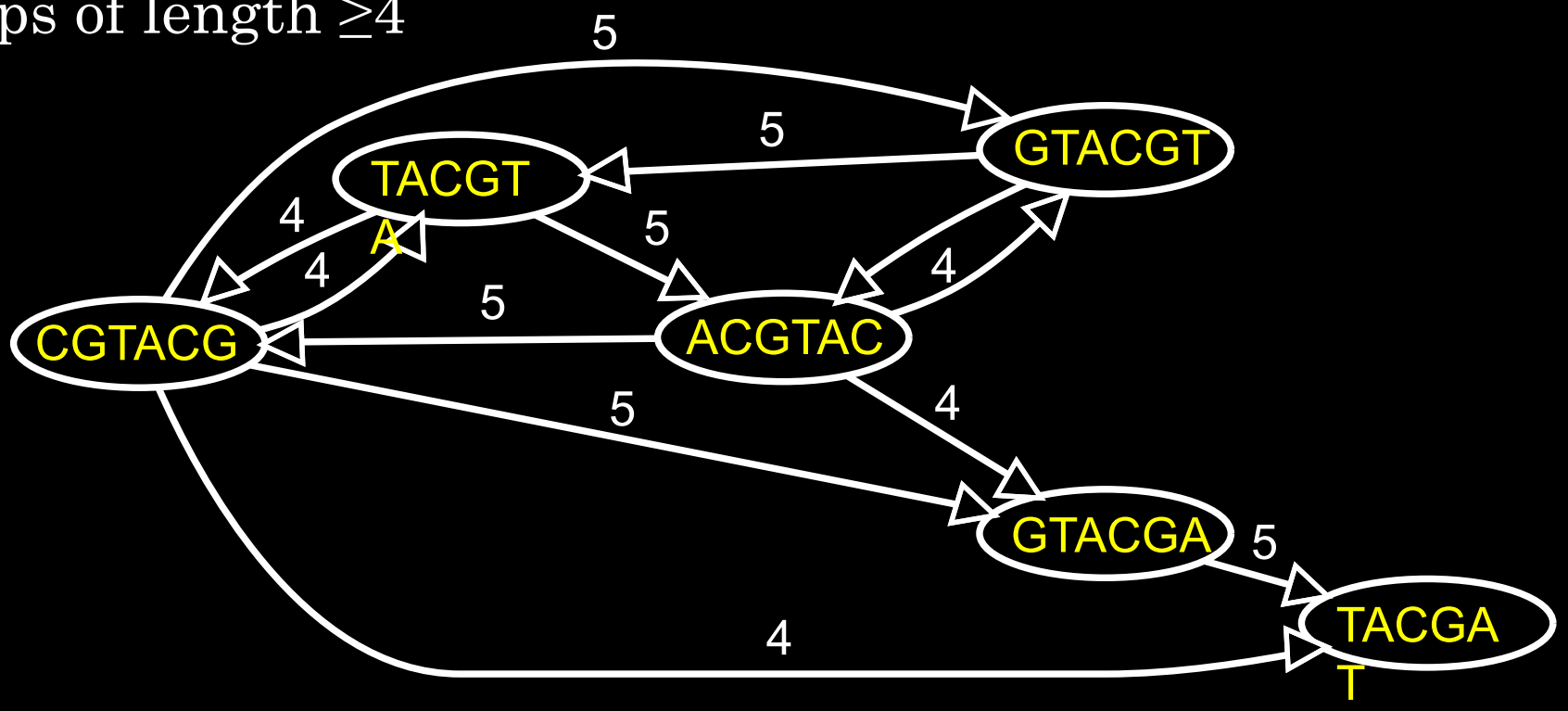
Coalesce paths into contigs

Pick nucleotide sequence for each contig

Overlap Graph

Nodes: all 6-mers from GTACGTACGAT

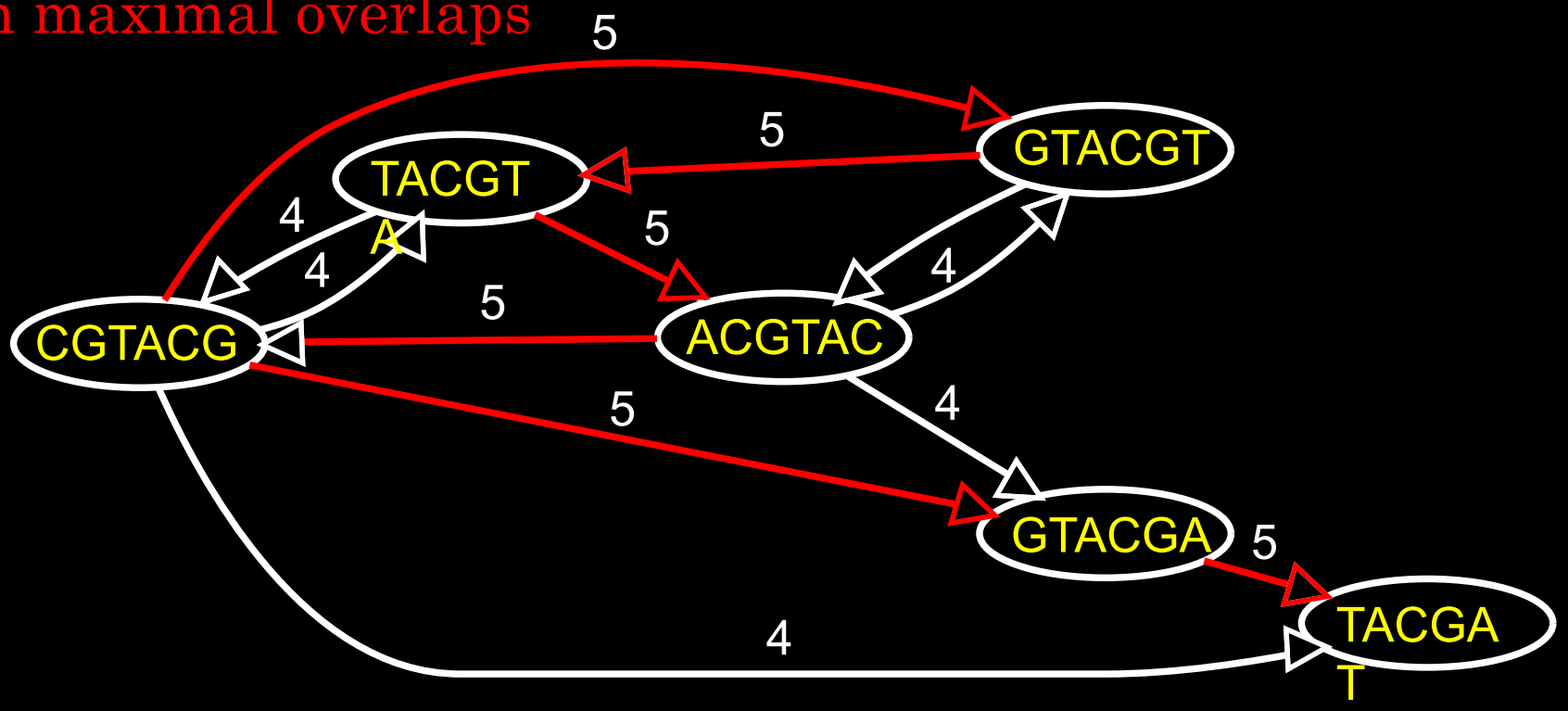
Edges: overlaps of length ≥4



Overlap Graph

Nodes: all 6-mers from GTACGTACGAT

Path based on maximal overlaps



Finding Overlaps

Overlap: Suffix of X of length $\geq l$ matches prefix of Y; l is given

Naïve: Look in X for occurrences of Y's length-l prefix. Extend matches to the right to confirm whether entire suffix of X matches.

Finding Overlaps – Suffix Tree

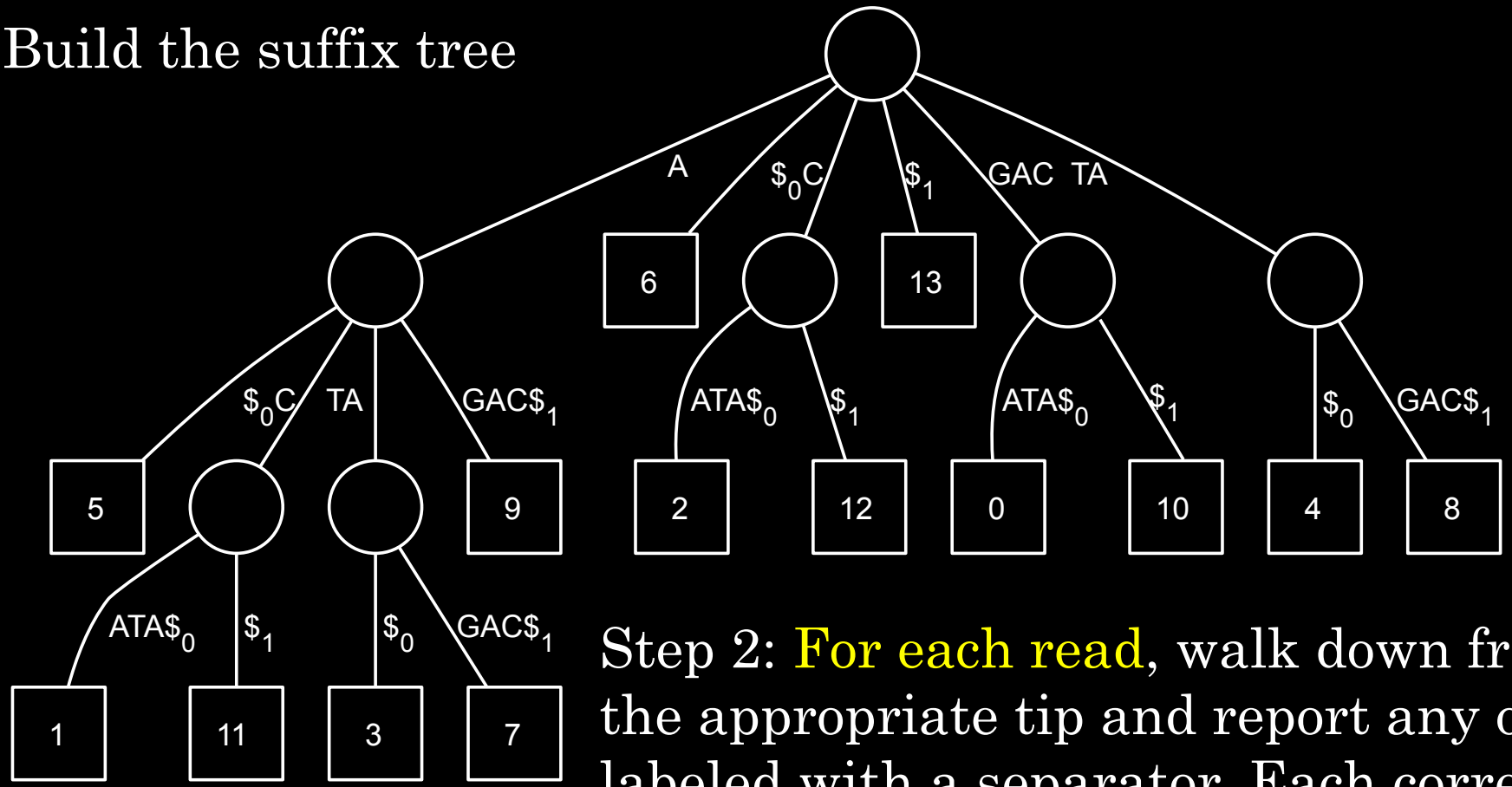
- Given a collection of strings S, for each string X in S,
- Find all overlaps involving a prefix of X and a suffix of another string Y
- Build a suffix tree from all reads; put a terminator character (remember $?) at the end of each read

Concatenated representation of { “GACATA”, “ATAGAC” }:

$GACATA\$_0ATAGAC\$_1$

GACATA$$_0$$ATAGAC$$_1$$

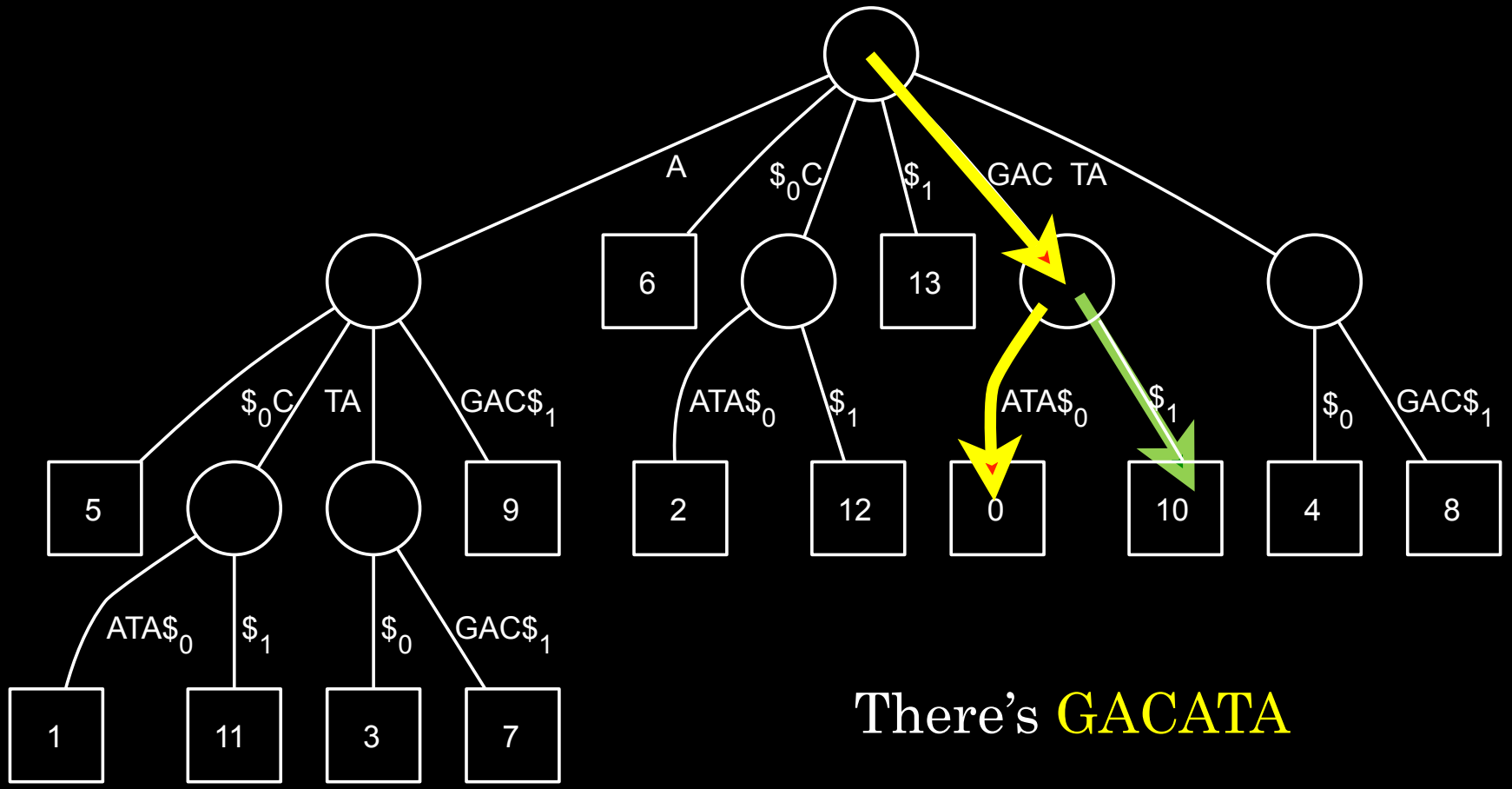
Step 1: Build the suffix tree



Step 2: For each read, walk down from the root to the appropriate tip and report any outgoing edge labeled with a separator. Each corresponds to a prefix/suffix match of query read and suffix of another read that ends in the separator.

GACATA$$_0$$ATAGAC$$_1$$

Let's start by searching for GACATA



There's GACATA

A terminator exists after GAC!

Read #1 ends with GAC

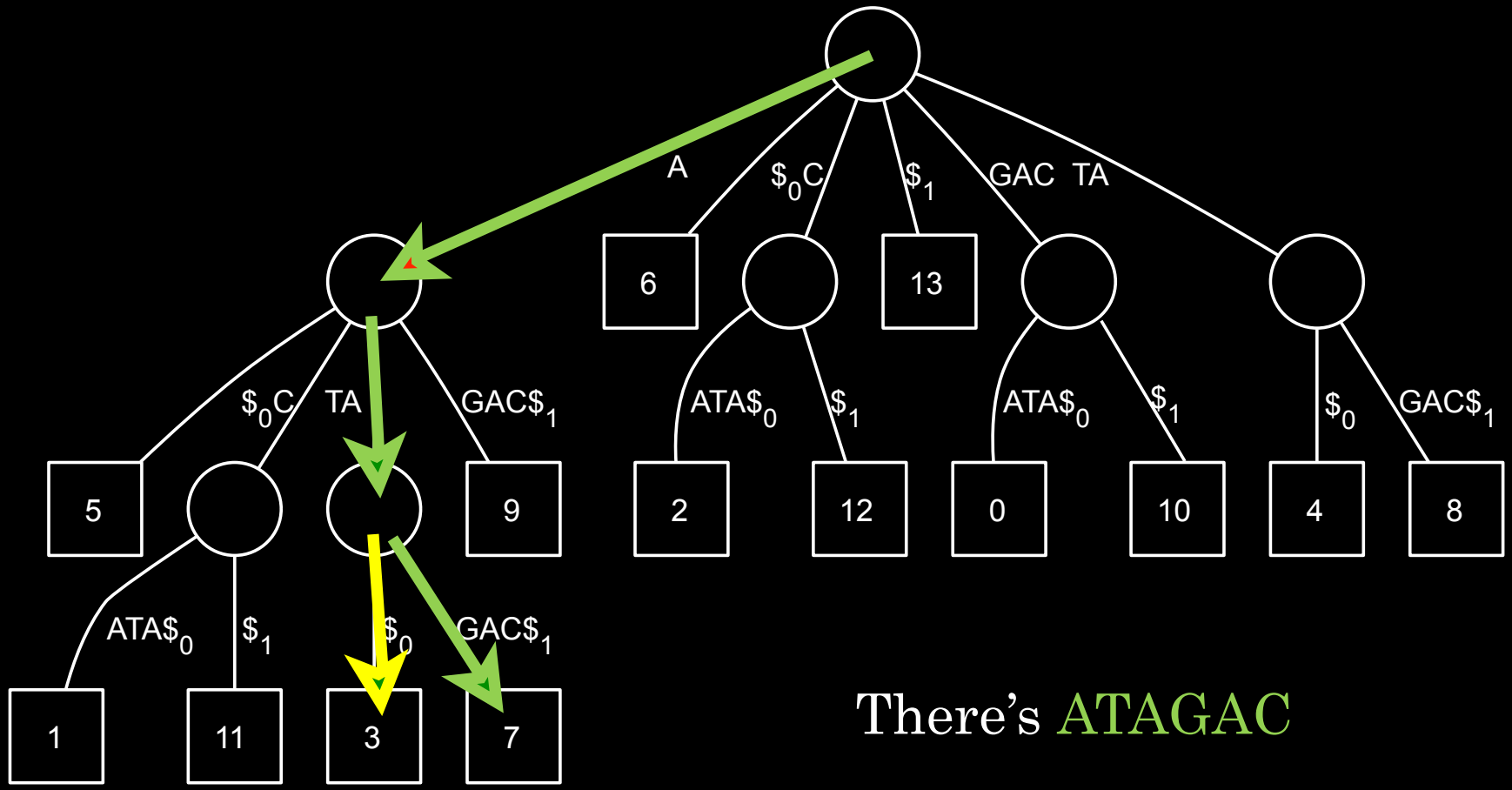
ATAGAC

| | |

GACATA

GACATA$$_0$$ATAGAC$$_1$$

Let's start by searching for ATAGAC



There's ATAGAC

A terminator exists after A+TA!

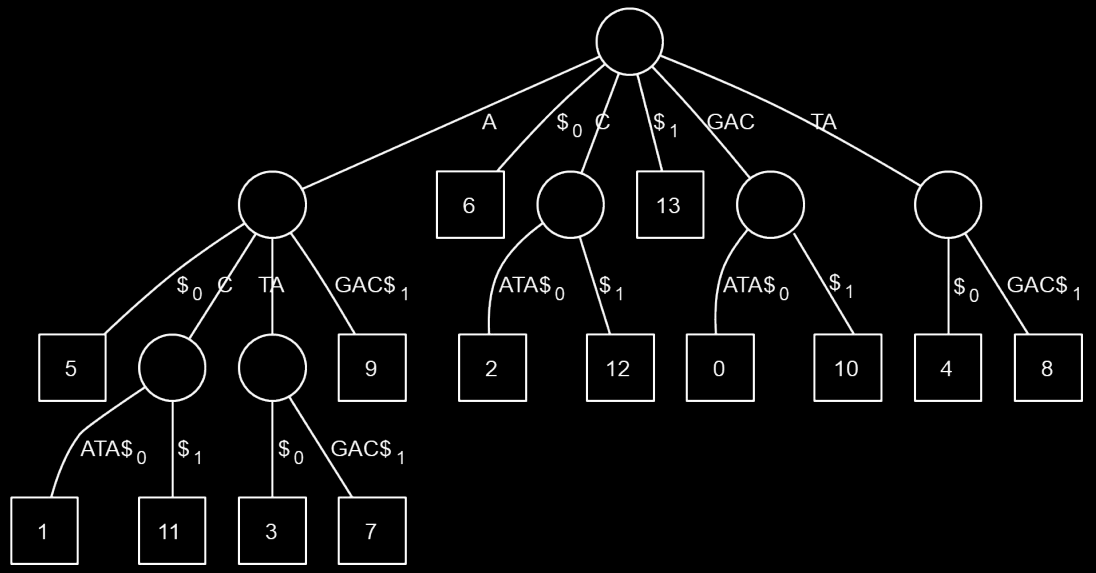
Read #0 ends with ATA

ATAGAC

|||

GACATA

Finding Overlaps - Complexity



Say there are d reads of length n, total length $N = dn$, and α = # read pairs that overlap

Assume for given read pair we report only the longest suffix/prefix match

Time to build suffix tree: $O(N)$

... to walk down red paths: $O(N)$

... to find & report overlaps (green): $O(\alpha)$

Overall: $O(N + \alpha)$

About Those Sequencing Errors….

Exact matching is no longer good enough!

```
CTCGGCCCTAGG
|||  |||||
  GGCTCTAGGCCC
```

The solution is to do dynamic programming

Finding Overlaps with Dynamic Programming

```
CTCGGCCCTAGG
|||  |||||
  GGCTCTAGGCCC
```

Use global alignment and score function

Match / mismatch scores s (smaller is better!)

	A	C	G	T	-
A	0	4	2	4	8
C	4	0	4	2	8
G	2	4	0	4	8
T	4	2	4	0	8
-	8	8	8	8	

$$D[i,j] = \min \begin{cases} D[i\text{-}1,j] + s(x[i\text{-}1], \text{gap}) & \text{Insertion in read } j \\ D[i,j\text{-}1] + s(\text{gap}, y[j\text{-}1]) & \text{Insertion in read } i \\ D[i\text{-}1,j\text{-}1] + s(x[i\text{-}1], y[j\text{-}1]) & \text{Match} \end{cases}$$

How do we force it to find prefix / suffix matches?

Finding Overlaps with Dynamic Programming

How to initialize first row & column so suffix of X aligns to prefix of Y?

First column gets 0s
(any suffix of X is possible)

First row gets ∞s
(must be a prefix of Y; leading gaps not allowed)

Fill the matrix with scores

Backtrace from last row

Finding Overlaps with Dynamic Programming

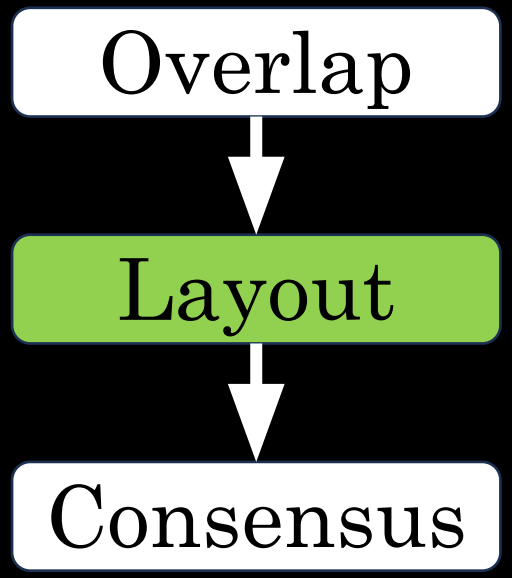
Say there are *d* reads of length *n*, total length *N* = *dn*, and *a* is total number of pairs with an overlap

# overlaps to try:	$O(d^2)$
Size of each DP matrix:	$O(n^2)$
Overall: $O(d^2n^2)$, or	$O(N^2)$

$O(N^2)$ is worse that exact-match suffix tree: $O(N + a)$

Real-world overlappers mix the two; index filters out vast majority of non-overlapping pairs, dynamic programming used for remaining pairs

OLC

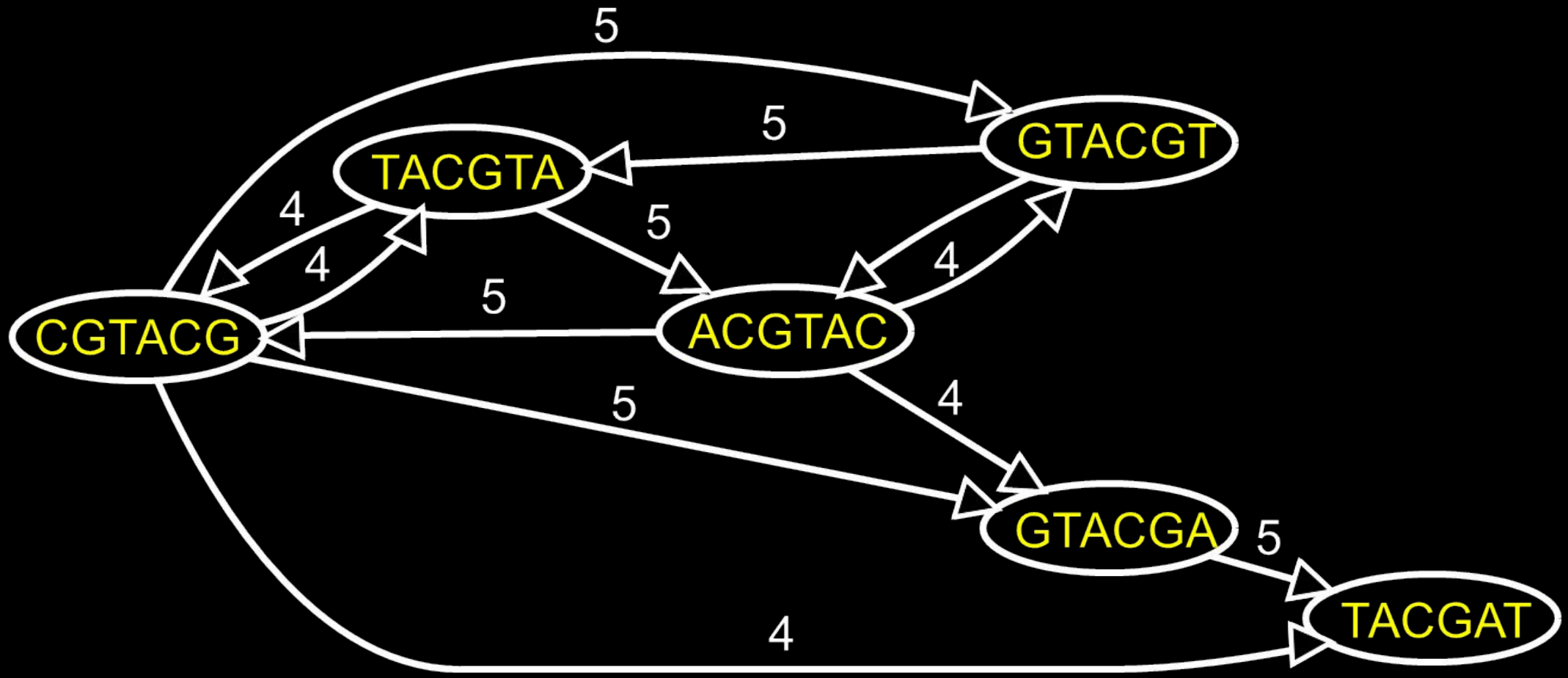


Build overlap graph

Coalesce paths into contigs

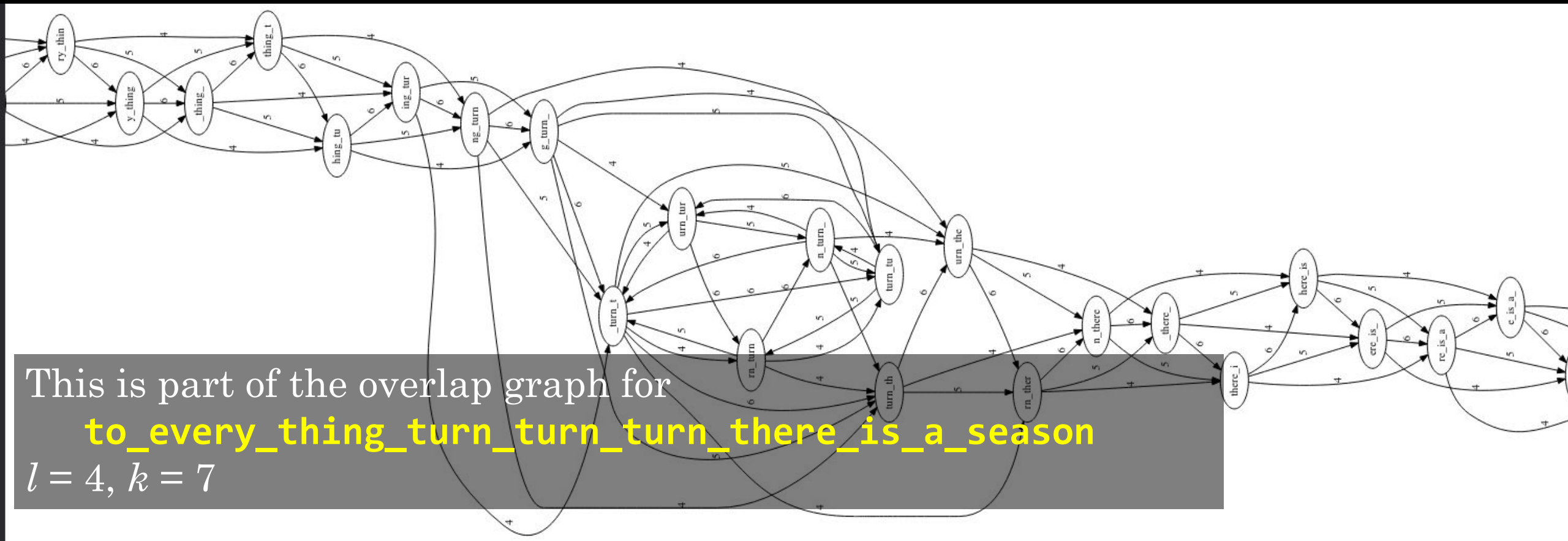
Pick nucleotide sequence for each contig

Remember: Overlap Graph



Layout

Try to make sense of this.
Overlap graph is big and messy. Contigs don't "pop out" at us.



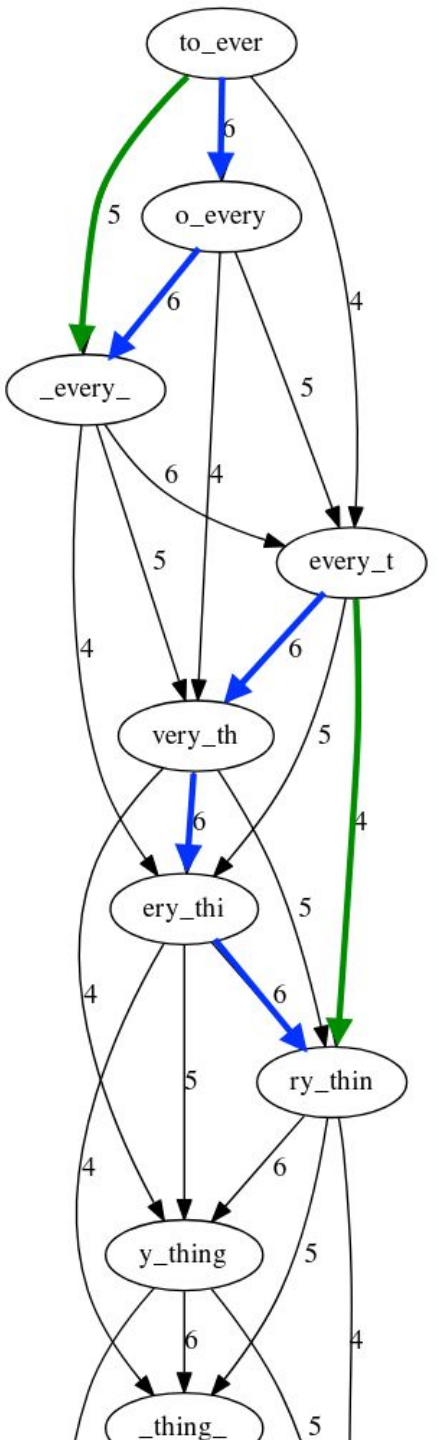
This is part of the overlap graph for

to_every_thing_turn_turn_turn_there_is_a_season

$l = 4$, $k = 7$

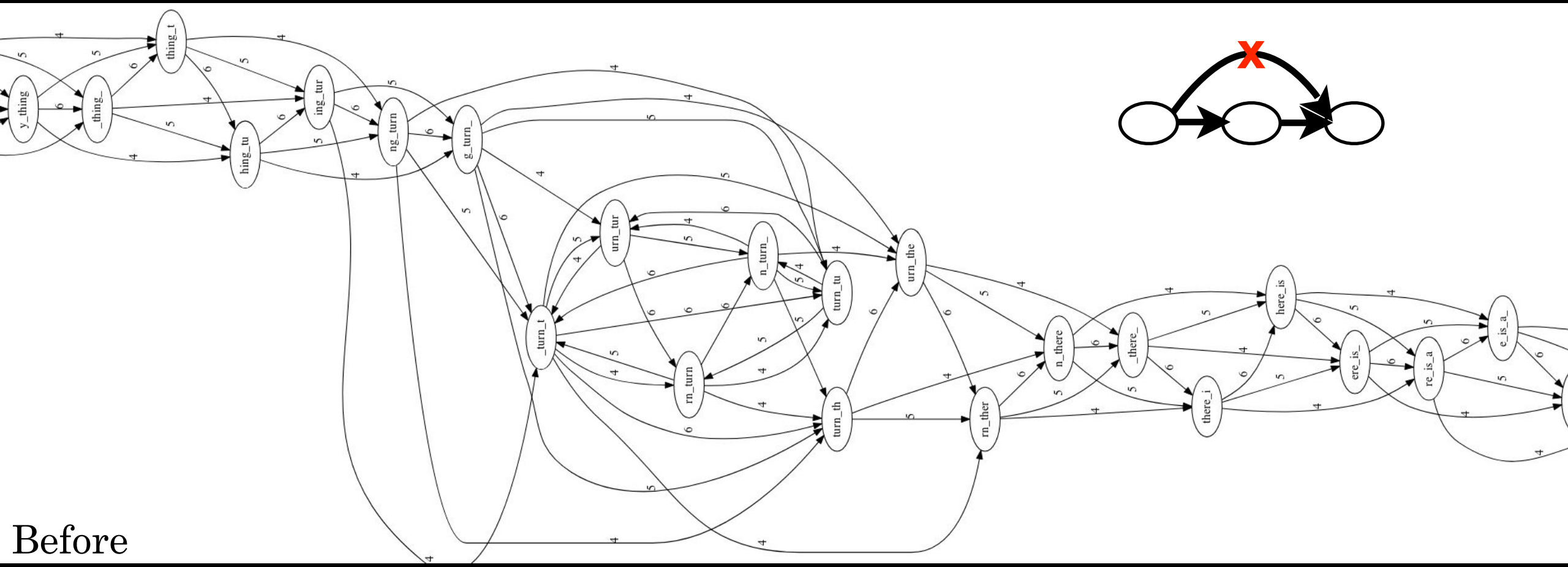
Layout

- We can simplify by removing redundancy from the overlap graph
- Some edges can be inferred (transitively) from other edges
- E.g. green edge can be inferred from blue



Layout

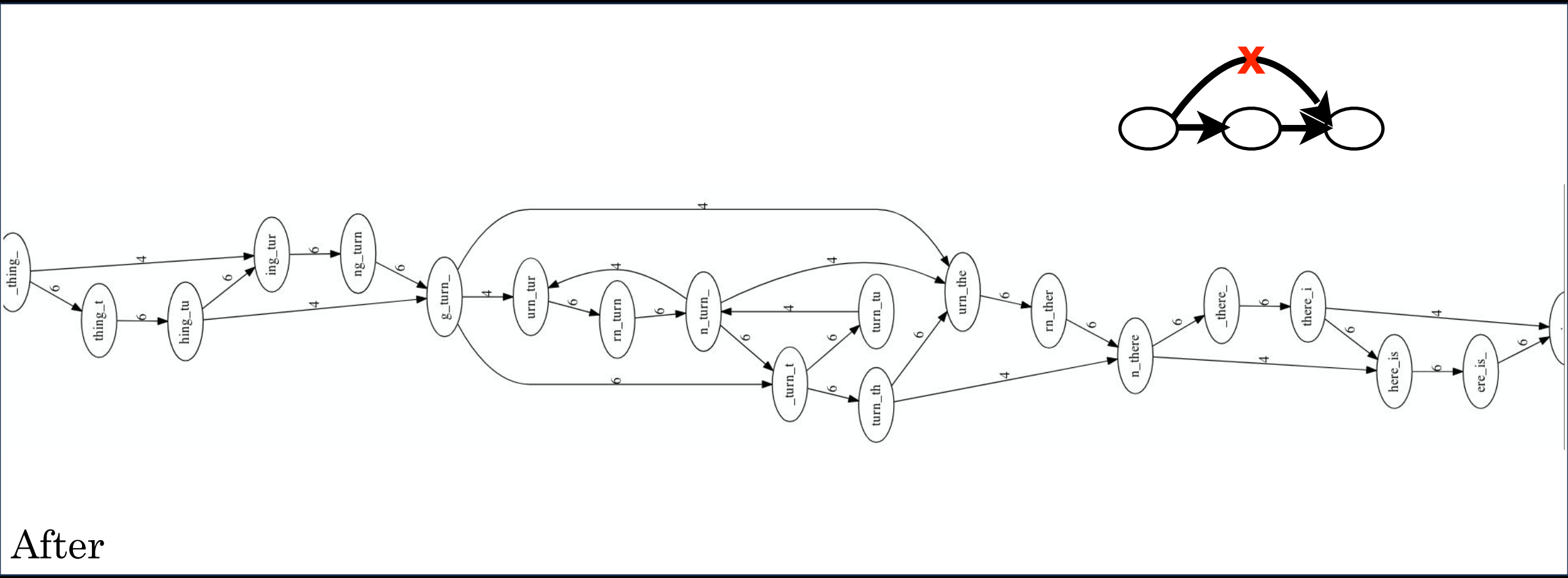
Remove transitively inferable edges, starting with edges that skip one node:



Before

Layout

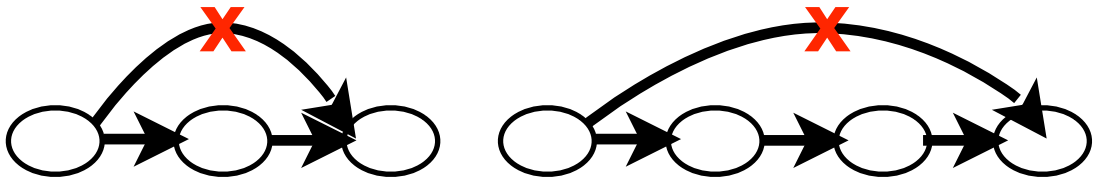
Remove transitively inferable edges, starting with edges that skip one node:

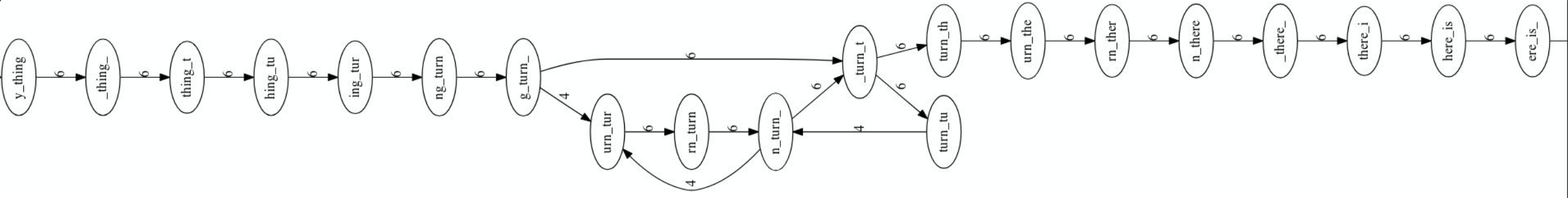


After

Layout

Now remove edges that skip one or two nodes:

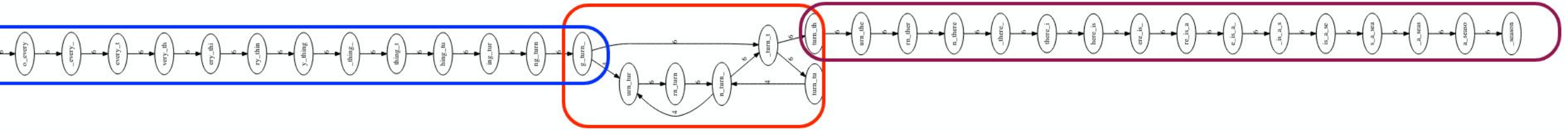




After

Layout

Every non-branching stretch is a contig



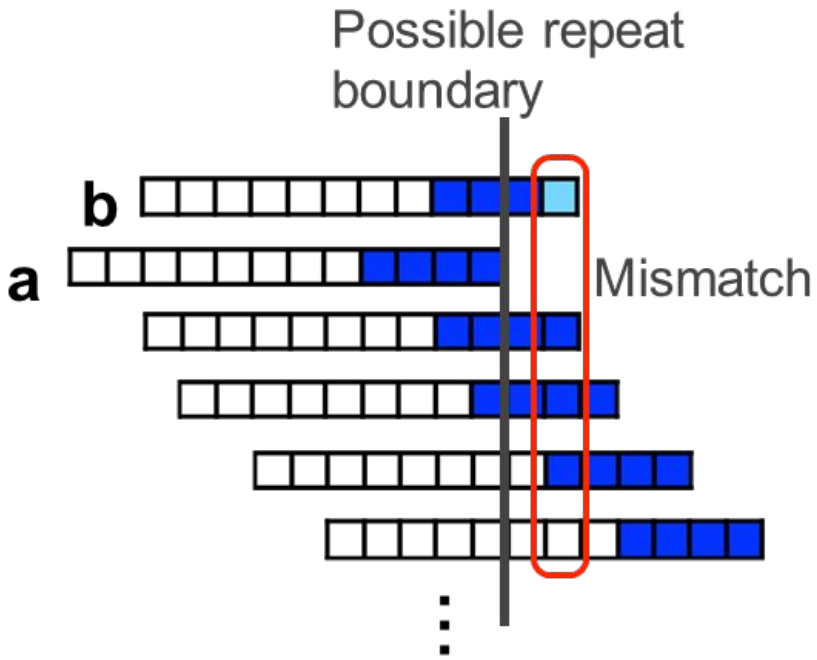
`to_every_thing_turn_`

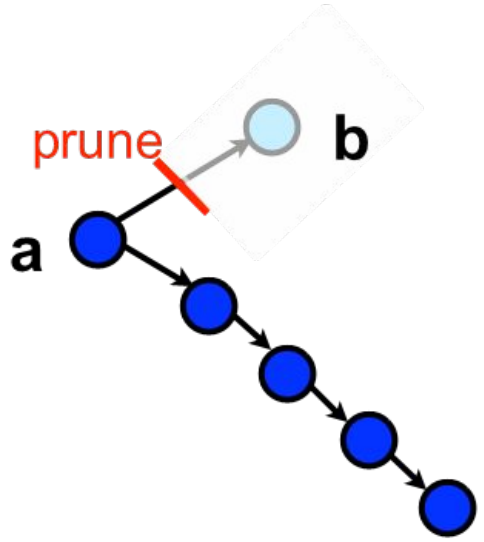
`turn_there_is_a_season`

Unresolvable branching mess

Layout

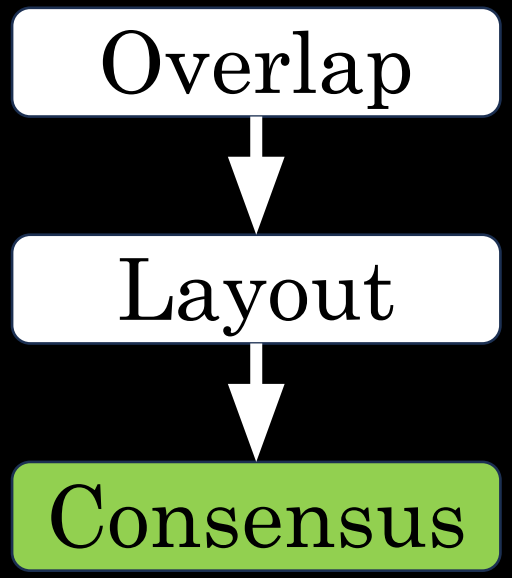
Spurious weird stuff gets pruned





Forking pattern from (a), early termination at (b). Lies??

OLC

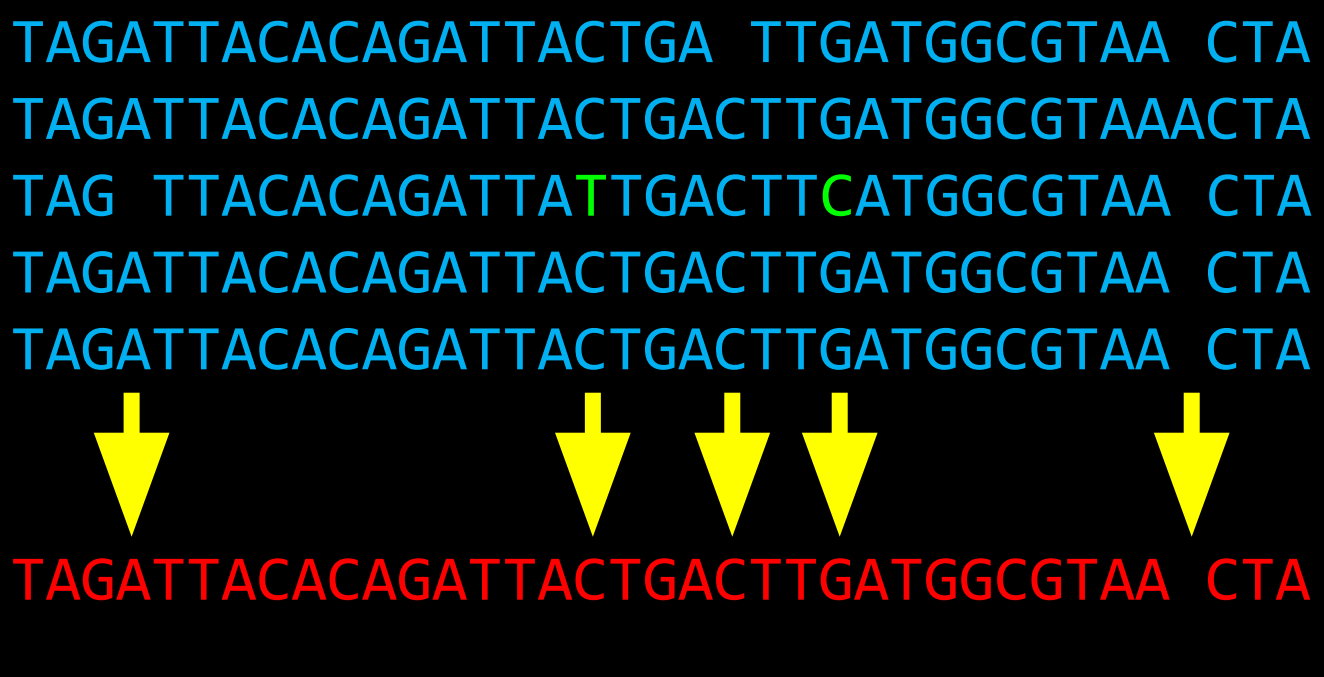


Build overlap graph

Coalesce paths into contigs

Pick nucleotide sequence for each contig

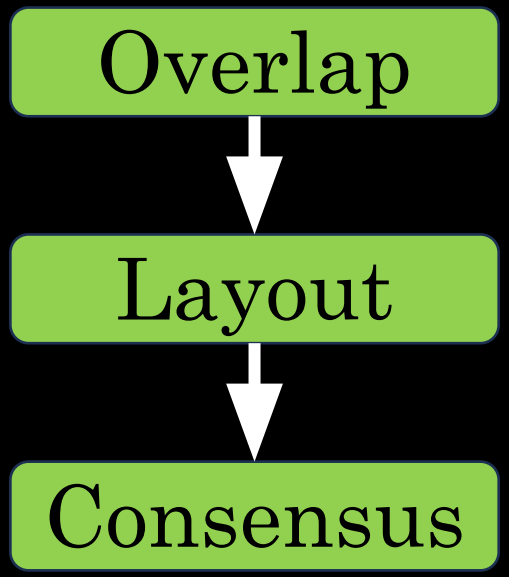
Consensus



Line up the reads associated with a contig

Take the consensus (majority vote?)

OLC



Build overlap graph

Coalesce paths into contigs

Pick nucleotide sequence for each contig

Problem solved!!!

- EXCEPT -

- Building overlap graph is slow. $O(N + a)$, $O(N^2)$
- Overlap graph is huge: one node per read, # edges can grow superlinearly with # reads
- Sequencing datasets are ~ 100s of millions or billions of reads

OLC is Generally Not the Approach

- Then why did we just spend a lecture talking about it?

- OLC is more intuitive and covers many of the key concepts we'll see next time
  - Graphs
  - Overlap
  - Weird paths
  - Complex, messy structures
  - Sequencing error