CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfF18.html

More on NGS Assembly

Basic Assembler

- **Read**: sequenced fragment; **Contig**: contiguous segment. How to assemble a contig?

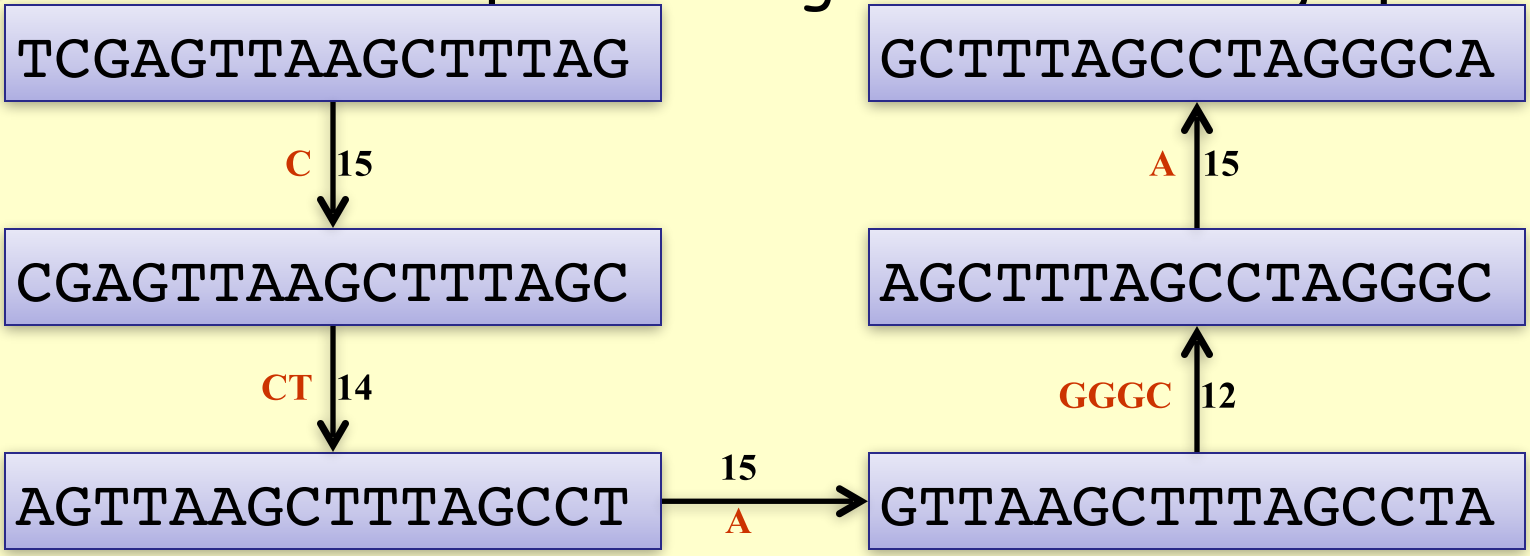
```
TCGAGTTAAGCTTTAG
 CGAGTTAAGCTTTAGC
   AGTTAAGCTTTAGCCT
    GTTAAGCTTTAGCCTA
       AGCTTTAGCCTAGGGC
         GCTTTAGCCTAGGCAG
              …
```

```
AGCTTTAGCCTAGGGC
AGTTAAGCTTTAGCCT
CGAGTTAAGCTTTAGC
GCTTTAGCCTAGGCAG
GTTAAGCTTTAGCCTA
TAAGCTTTAGCCTAGG
TCGAGTTAAGCTTTAG
```

Problem: Need to try every pair of reads!

Reduce to Graph Problem

- How to assemble a contig?
  - Node ↔ Read
  - Edge between Nodes ↔ Overlapping Reads
  - **Problem**: Find a path through each node in graph.



TCGAGTTAAGCTTTAG

C 15

CGAGTTAAGCTTTAGC

CT 14

AGTTAAGCTTTAGCCT

15 A

GTTAAGCTTTAGCCTA

GGGC 12

AGCTTTAGCCTAGGGC

A 15

GCTTTAGCCTAGGGCA

Issues: Problem is NP-Complete

# nodes = # reads

# of edges < k(# nodes)

String graph

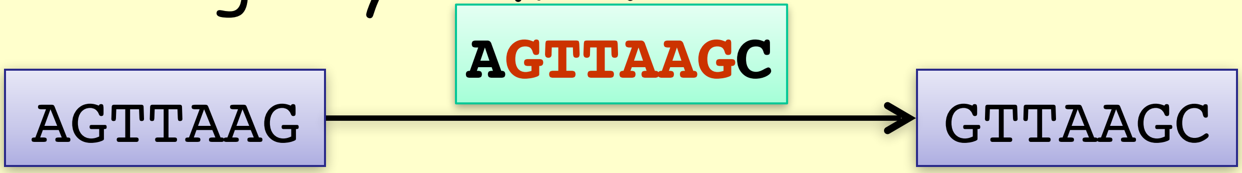
- Combine nodes that form paths into strings

A better solution

- Take each read and chop it into k-mers.
- Represent k-mers by nodes in a graph and edges between k-mers that overlap in k-1 bases.
- **Consequence**:
  - Number of nodes = 4^k ;
  - Number of edges = $k4^k$;
- **Issues**:
  - Problem (i.e., find path through all vertices) remains NP-Complete

A more efficient solution: de Bruijn Graphs

- ❑ Represent every possible (k-1)-mer by a node.
- ❑ Edges connect 2 nodes if they share k-2 bases.
- ❑ Label each edge by k-mer.



- ❑ Problem:
  - Find a path through each edge in the graph
- ❑ The Eulerian path problem is **NOT** NP-Complete. It can be solved in linear time!

Pevzner, PA, l-tuple DNA sequencing: Computer analysis. Journal of Biomolecular Structure and Dynamics 7(1), 63-73, 1989.

Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- **REPEATS**
- Combinations of the above

How to deal with REPEAT Regions

- If no errors or repeat regions, then the graph has a unique path through all the edges.
- The de Bruijn graph method quickly deteriorates with sequencing errors
  - Either correct reads before assembly OR
  - Correct de Bruijn graph for spurious edges
- **Problem**: REPEAT regions cause branching in graph. If no errors in reads, then the graph has a unique path through all edges, but with some edges traversed more than once.
- How to identify REPEAT regions:
  - Higher coverage of repeat regions
  - Branching of nodes

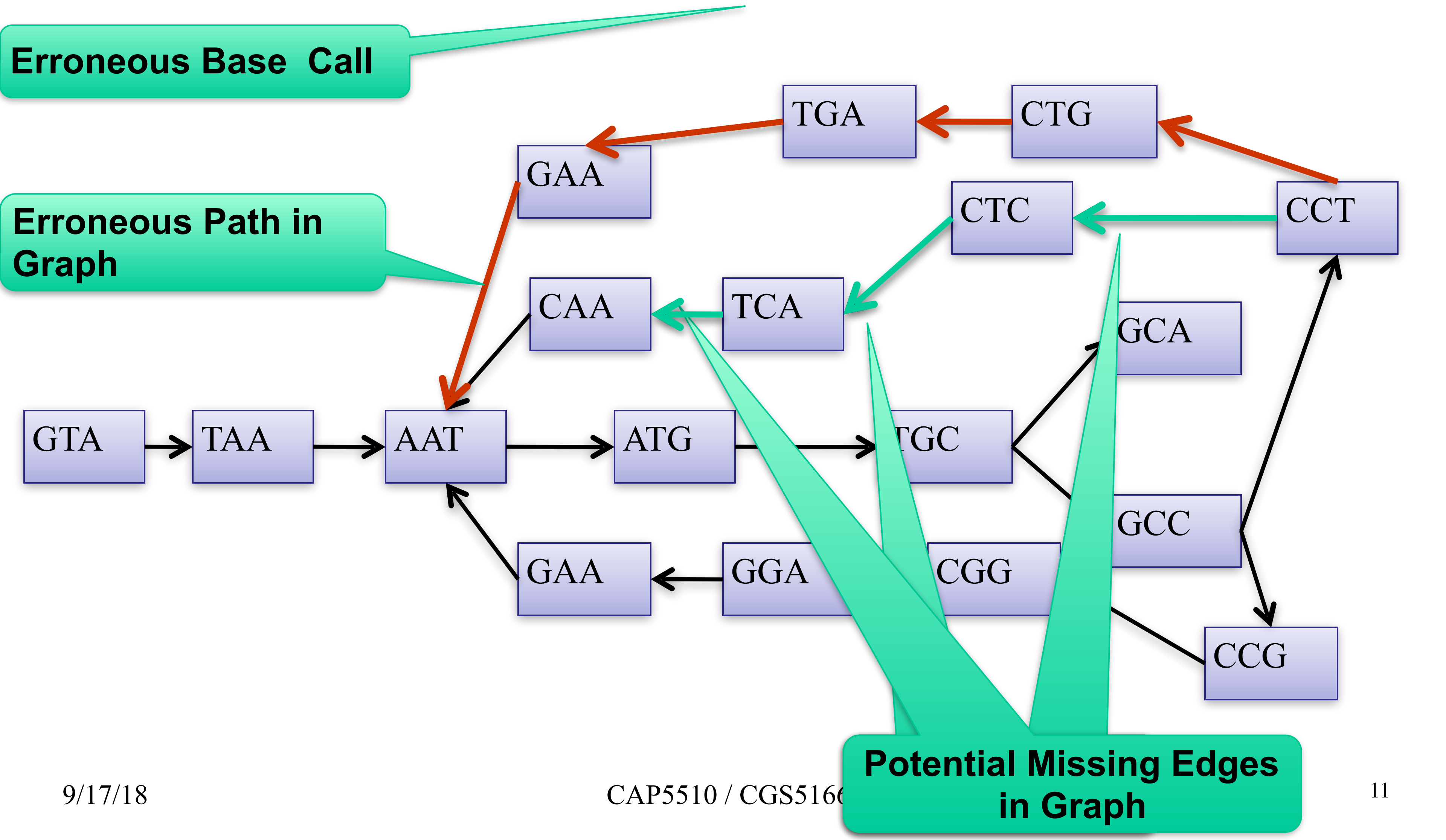
Sources of Assembly Errors

- Errors in reads – caused by technology
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)
- Missing reads – sequencing bias
- Read orientation error
  - One or both orientations may occur
  - Not told which ones are present
- Sequence Variations – mixed sample study
  - SNP, cancer, metagenomics studies
- Combinations of the above

Handling Read Error

GTAATGCCTCAATGCCGGAATGCA

CTGAA



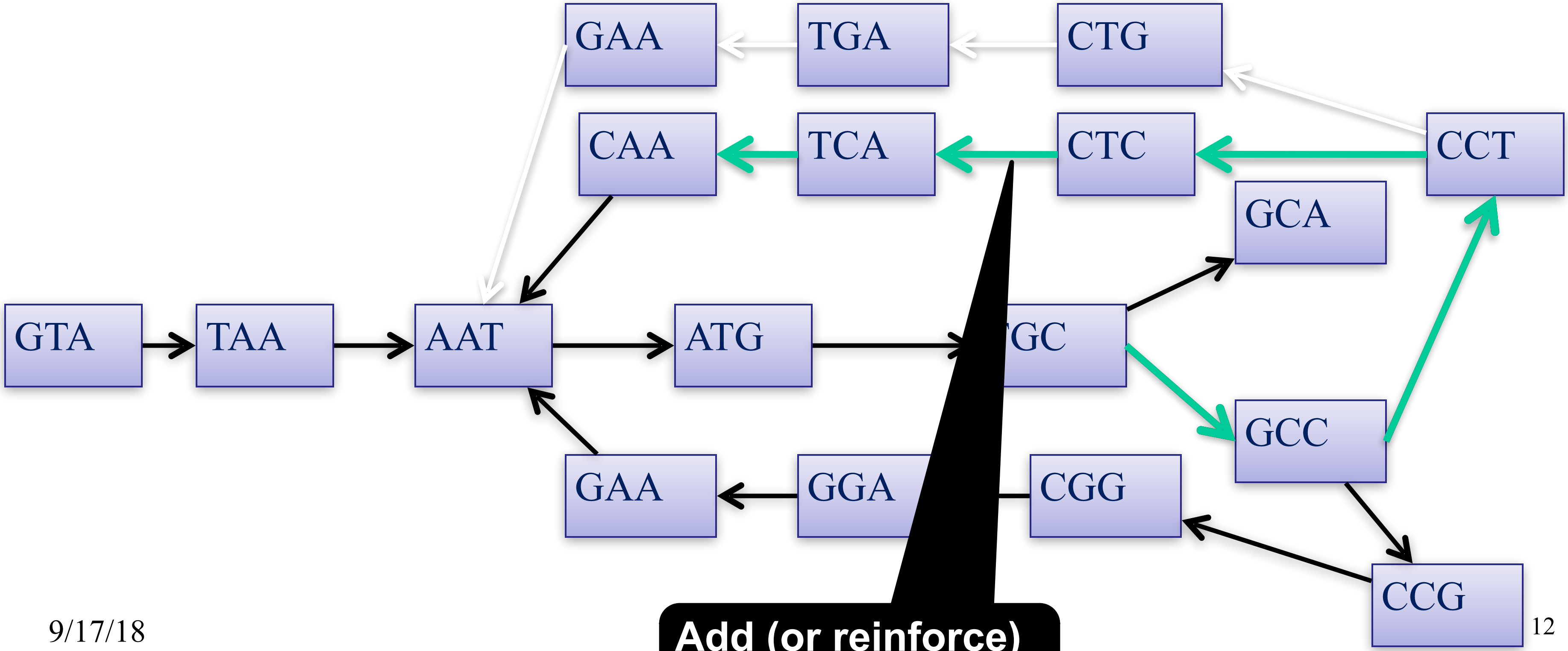
Well conserved regions in related genomes

TGCCTCAA

TGCCTCAA

GTAATGCCTCAATGCCGGAATGCA

CTGAA



Add (or reinforce) path in graph

Issues and Ideas

- Small k gives rise to many spurious edges
- Large k makes the graph sparse
- Start with k-mer graph or string graph or overlap graph or contig (Velvet) graph
  - Advantages/disadvantages of each?
- Place highly conserved reads or regions on this graph
- Identify missing nodes/edges/paths
- Paired de Bruijn graphs incorporated paired reads directly into graph when the distance between the pairs are fixed
- Pathset de Bruijn graphs do the same when distance between pairs are variable
- Positional de Bruijn graphs incorporate positional information about k-mers
- Colored de Bruijn graphs are used to analyze genetic variants

When is a genome assembly done?

- Almost never perfectly! Great cost in time, effort, and money.
  - Currently 92% of human genome is done to 99.99% accuracy [Schmutz et al., Nature 429, 365-368]
  - More likely to complete with bacterial and viral genomes, but they evolve much faster.
- Hard part with bacterial genomes are genomic rearrangements
- Often enough to get gene content to perform comparative genomics
- Tools to compare gene content
  - CEGMA – Eukaryote
  - CheckM – Bacterial; https://peerj.com/preprints/554.pdf
- Useful papers
  - Salzberg et al., Genome Res, 2012
  - Vezzi et al., PLoS ONE, 2012, DOI: 10.1371/journal.pone.0031002
  - Gurevich et al., Bioinformatics, 29(8): 1072-75, 2013
  - Shengguan et al., PLoS ONE, 2013, DOI: 10.1371/journal.pone.0069890

N50 measure

- https://www.broad.harvard.edu/crd/wiki/index.php/N50
- Statistical measure of “average length” of a set of sequences.
- Used widely in evaluating assemblies.
- N50 length is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N.
- N50 is a weighted median statistic such that 50% of entire assembly is contained in contigs or scaffolds equal to or larger than this value
- Given list of lengths L. Create another list L' , which is identical to L, except that every element n in L has been replaced with n copies of itself. Then the median of L' is the N50 of L.
- **Example**:
  - Let L = {2, 2, 2, 3, 3, 4, 8, 8},
  - L' consists of six 2's, six 3's, four 4's, and sixteen 8's; the N50 of L is the median of L' , which is 6.
  - Alternatively, sum = 32, halfSum = 16. You need the two 8’s to sum up to 16

How much of a genome is unsequenced?

- Assumption: fragments are independently and uniformly distributed across genome
  - R = Depth of Coverage
  - N = Genome length
- Fraction of genome not sequenced is Ne^{-R}
- "Law of diminishing returns": doubling sequencing depth from R to 2R reduces unsequenced portion of genome by a factor of e^{-R}

- Lander, Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis" Genomics 2(3):231–239, 1988
- Roach, "Random subcloning" Genome Research 5(5):464–473, 1995

Important Papers

- Kent, Haussler, "Assembly of the working draft of the human genome with gigassembler", Genome Research 11(9):1541–1548 (2001)
  - GIGASSEMBLER was used by the Human Genome Project to assemble about 30,000 clones. It used BAC end sequencing along with
    - genome-wide physical map,
    - radiation hybrid map,
    - Genetic map,
    - YAC-STS map, and
    - cytogenetic map,
  - GIGASSEMBLER used the "overlap-layout-consensus" approach:
    - Detect prefix-suffix overlaps between BAC contigs to build an overlap graph,
    - Removed edges in graph that can be transitively inferred, and
    - Find paths in graph to generate contigs
- Bao, Jiang and Girke, "AlignGraph: algorithm for secondary de novo genome assembly guided by closely related references", Bioinformatics (2014).