2012/02/23 CSE 427 guest lecture

A brief introduction to applying de Bruijn graphs for de novo genome assembly

Made by: Elizabeth Tseng (lachesis@cs.washington.edu)

NOTE: most of the images are from reference [2] and [3].

Worksheet p.1
Konigsberg in Prussia (modern day Russia). 7 bridges connecting 4 parts of the city. Can one stroll through every part of the city by walking across each bridge exactly one?

Euler (1735) formulated it as a graph problem. Define Eulerian paths & cycles.

Euler's Theorem A strongly connected, [directed]* graph G contains a Eulerian cycle if and only if every node is balanced.

1. Prove theorem in both directions.
2. So how is this related to genome assembly? first, introduce concept of de bruijn graphs.

Worksheet p.2
1. Don't think about reads yet. Just how given a circular genome, a number k=3, you would convert it to a de Bruijn graph.
2. Show how to make nodes and connect edges
3. Show that with this construction the graph is always balanced

Worksheet p.4
1. (students) try to find a Eulerian path! (p.4)
2. Show that a particular Eulerian path reconstructs the genome
3. Show that as opposed to (3), if one or more k-mer (ex: CGT) is missing, no Eulerian cycle exists
4. Show that an alternative path to (4) does NOT give the right genome ➔ instead, we need the reads themselves to help us!

Worksheet p.3
1. Given a bunch of reads sequenced from the genome, represent them on the graph as a set of read paths P.
2. **Eulerian Superpath Problem** Given a Eulerian graph G and a set of paths P, find an Eulerian cycle that contains all paths in P as subpaths.
3. How do we solve (2)? A series of equivalent transformations! See paper [2].
4. Work through the example first, then talk about the rules of transformation. (p.5 as guidance)

Worksheet p.6

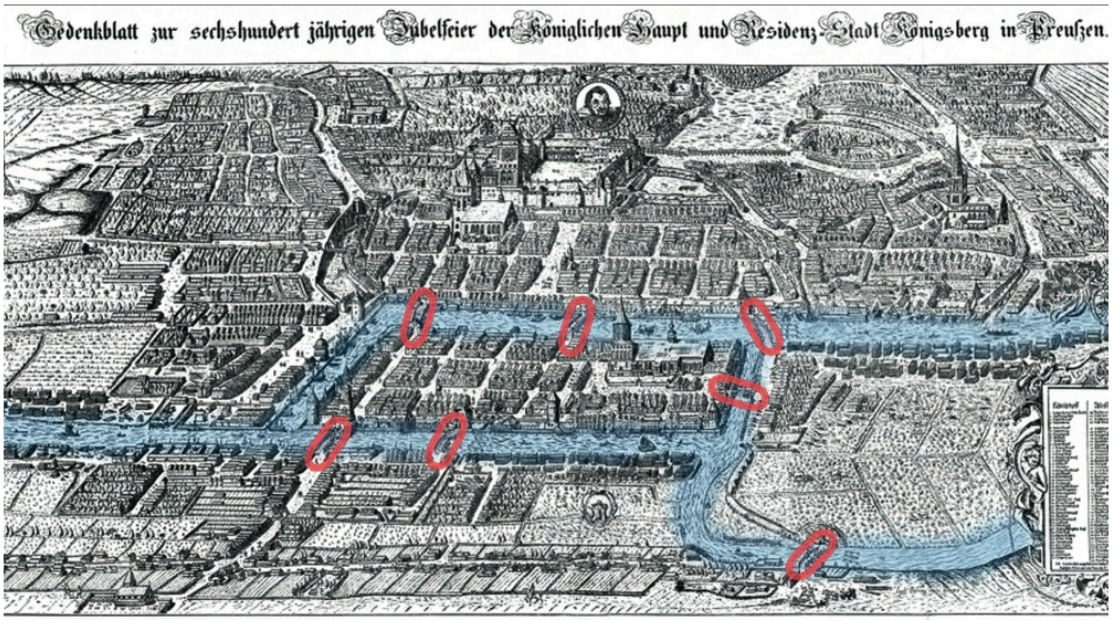
1. The simple no-multidge de Bruijn graph doesn't work for this example when *k*=3 because of multiplicity
2. Show how to add the multiplicity in
3. Show how to do the transformation with correct updating of paths

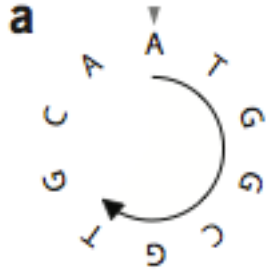
Show a printed list of assembly papers? (brief summary? cite the review?)
(opt) Talk about different sequencing technologies (show WT animations?)
(opt) Talk about the old OLC approach and why not good (Hamilton NP-complete)

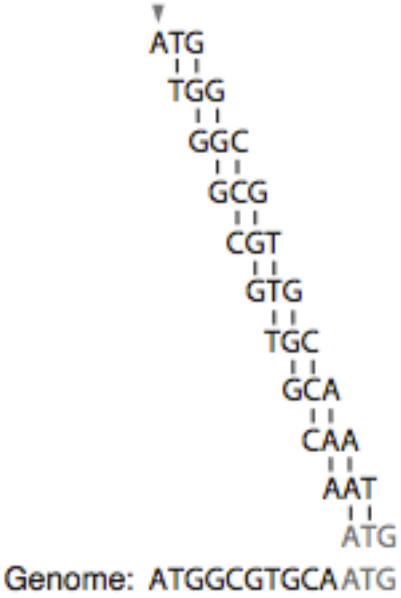
Other topics if more time:

Problems in real sequencing

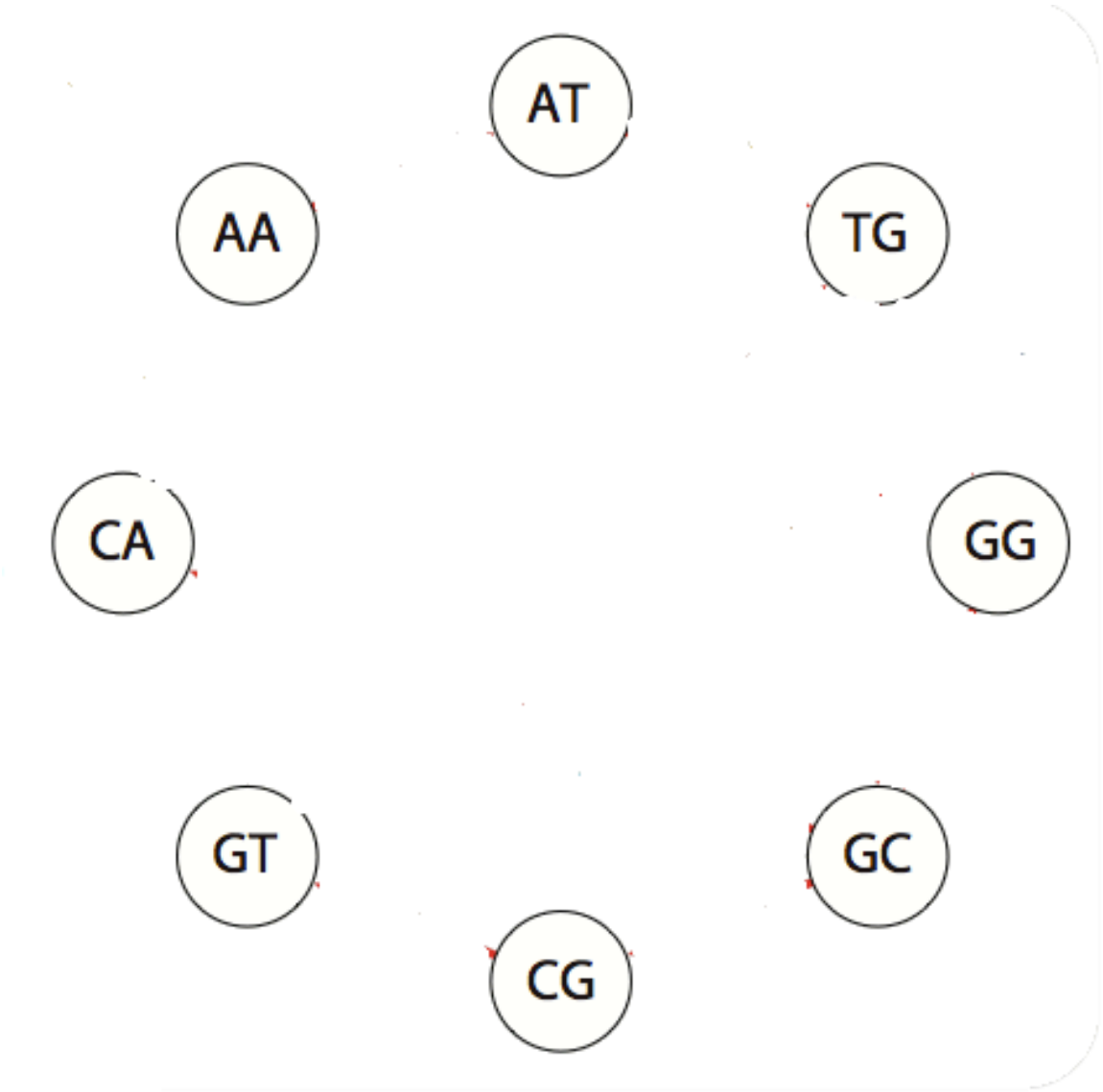
1. Sequencing error & polymorphisms. Worksheet p.7 (a) seq err at read end; (b) seq err middle of read OR polymorphisms. **Solution:** error correction through spectral alignment (EULER) and m-count cutoffs (VELVET) and finally, collapse near-identical sequences.
2. Repeats. Worksheet p.7 (c). **Solution:** consult reads and mate pairs.
3. Palindrome (seq = its own rev comp). **Solution:** (VELVET) require that k is odd.

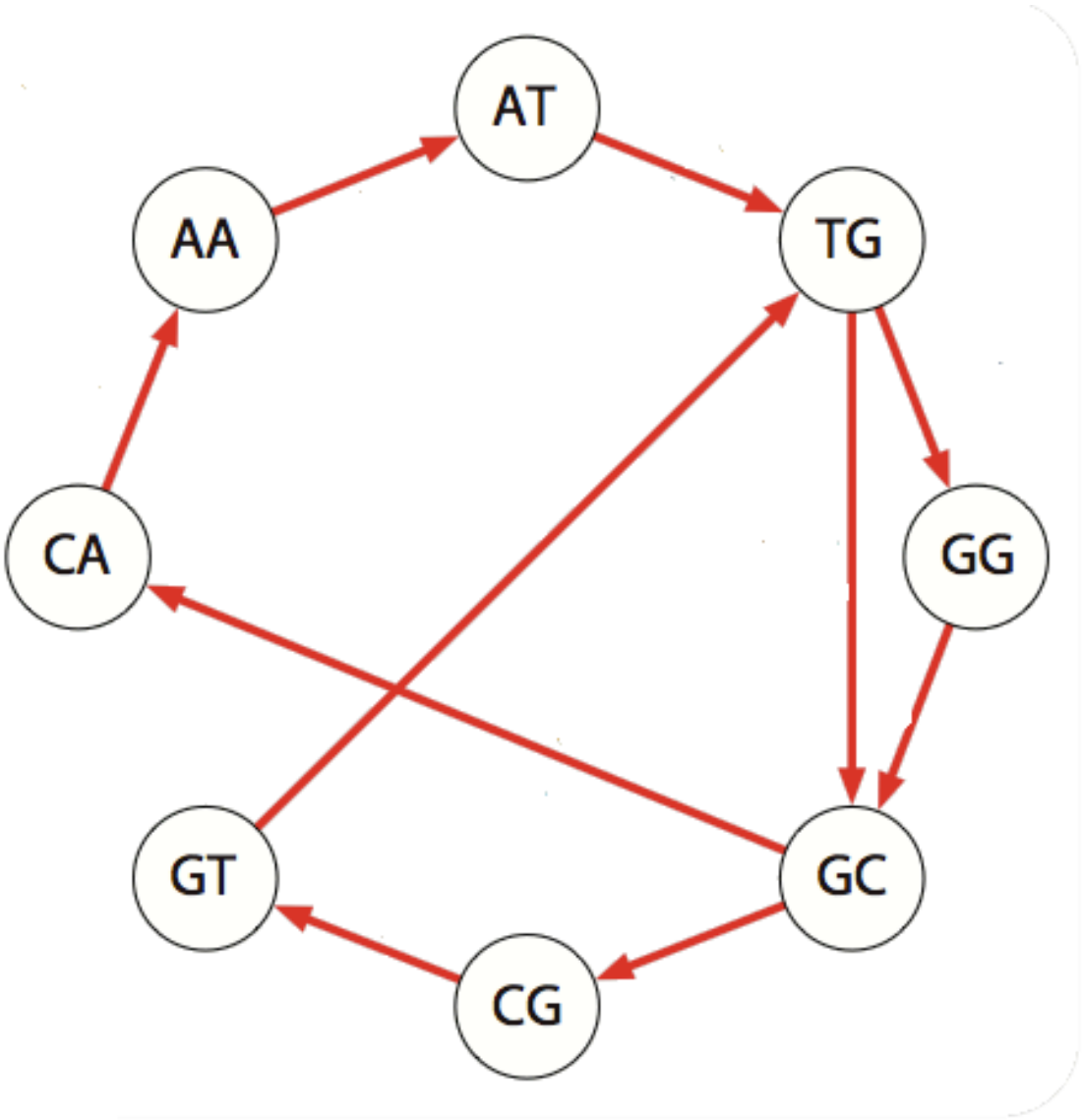

Gedenkblatt zur sechshundert jährigen Jubelfeier der Königlichen Haupt und Residenz-Stadt Königsberg in Preußen.


a
A
T
G
G
C
T
G
C
A


ATG
TGG
GGC
GCG
CGT
GTG
TGC
GCA
CAA
AAT
ATG
Genome: ATGGCGTGCAATG

Real genome:
ATGGCGTGCA
Sequenced reads:
ATGGCGT
GGCGTGC
CGTGCAA
TGCAATG
CAATGGC



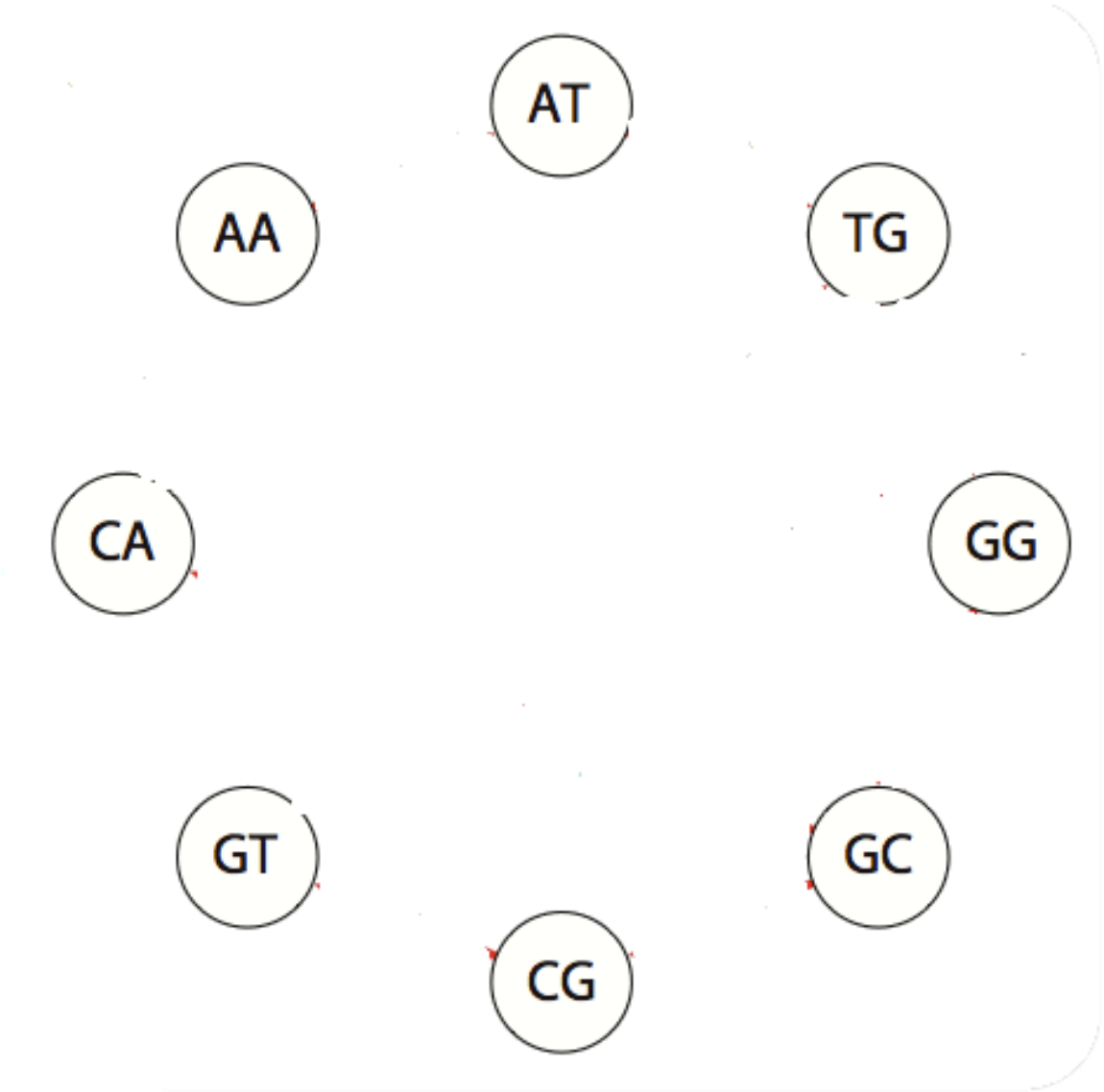

AT
AA
TG
CA
GG
GT
GC
CG

Original:
ATG → TGG → GGC → GCG → CGT
GGC → GCG → CGT → GTG → TGC
CGT → GTG → TGC → GCA → CAA
TGC → GCA → CAA → AAT → ATG
CAA → AAT → ATG → TGG → GGC

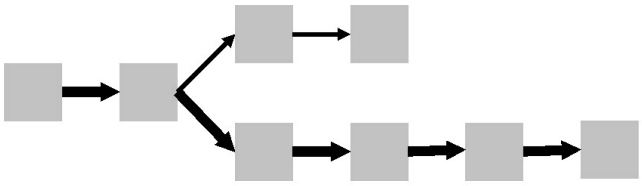
Detach TGG → GGC:
ATG → TGGC → GCG → CGT
TGGC → GCG → CGT → GTG → TGC
CGT → GTG → TGC → GCA → CAA
TGC → GCA → CAA → AAT → ATG
CAA → AAT → ATG → TGGC

Detach GTG → TGC:
ATG → TGGC → GCG → CGT
TGGC → GCG → CGT → GTGC
CGT → GTGC → GCA → CAA
GTGC → GCA → CAA → AAT → ATG
CAA → AAT → ATG → TGGC

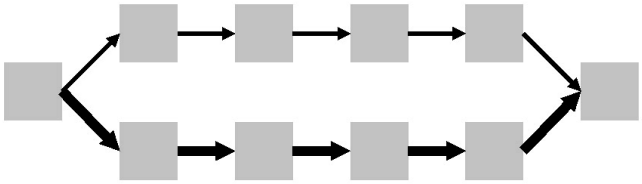
Real genome:
ATGCGGTGCGTGGCA
Sequenced reads:
ATGCGGT
GCGGTGC
GGTGCGT
GCGTGGC
CGTGGCA
GGCAATG



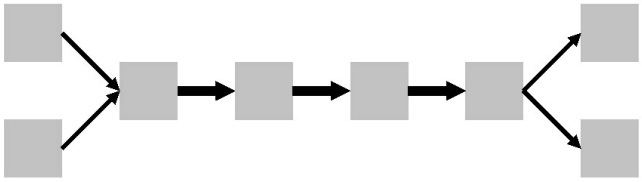
(a)



(b)



(c)



References[1234]

1. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327 (2010).
2. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98**, 9748 (2001).
3. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology* **29**, 987–991 (2011).
4. DNA animations | Wellcome Trust. at <http://www.wellcome.ac.uk/Education-resources/Teaching-and-education/Animations/DNA/>