De Novo Repeat Classification and Fragment Assembly (Pevzner, et. al. 2004)

Presenter: Taher Mun

Introduction

Tool presented: RepeatGluer

Concepts introduced: A-bruijn graph

Motivation: Need a good way to represent repeats in genomes while **preserving their mosaic structure** and **repeat-boundaries**

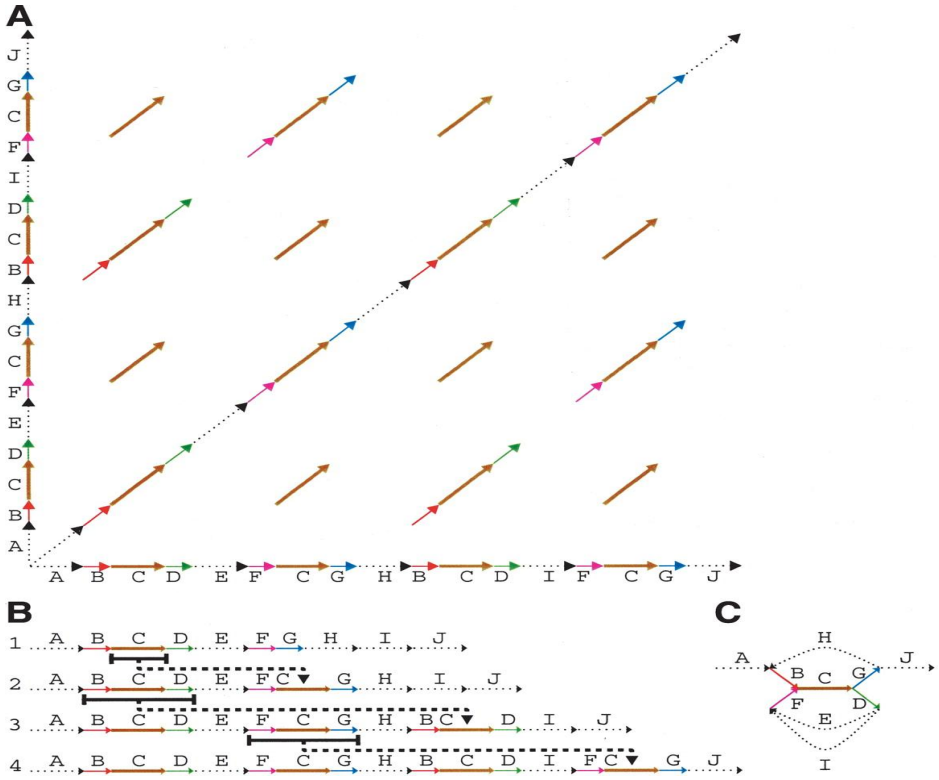
Why is repeat representation important?

- Many problems require boundaries of sub-repeats to be defined:
  - Finding mobile elements
  - Determining nature of segmental duplication - evolution
  - Assembly from short reads
  - (in my case) - alignment to repeat regions

Previous work

- RepeatMasker, MaskerAid
  - Use repeat libraries defined by RepBase
- RepeatMatch, REPuter
  - Simply lists repeats
  - Doesn’t show any underlying structures

(A) Genomic dot-plot of an imaginary sequence with repeats containing sub-repeats.

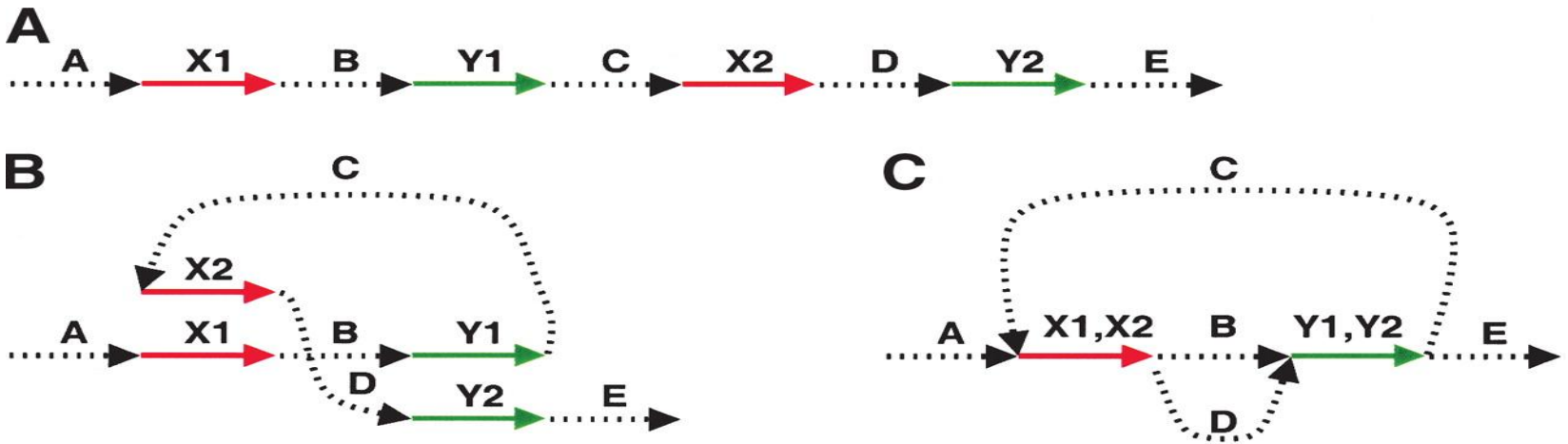


Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Figure 1

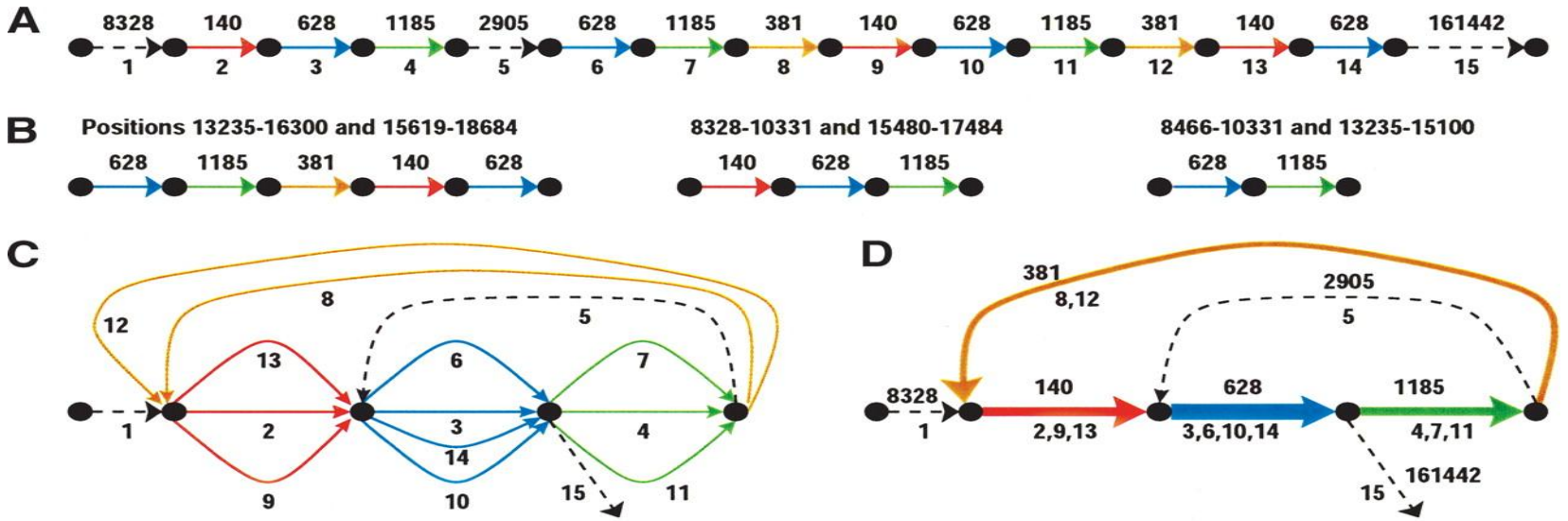
- Sub-repeats within repeats
  - Determining repeat boundaries?
  - Determining repeat structure?
- De-bruijn graph good way to represent (1C)
  - sub-repeats only occur once per edge
  - Though, only good for “perfect” repeats
  - Real genomic repeats are not perfect - indels, mismatches
- How to allow for imperfect pairwise alignments?
  - A-bruijn graph

(A) A hypothetical DNA sequence with unique regions A, B, C, D, E and repeats X (appearing twice as X1 and X2) and Y (appearing twice as Y1 and Y2).



Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Mosaic repeat organization of BAC from human Chromosome Y. For purposes of illustration, only sufficiently long and very conservative repeats are shown.



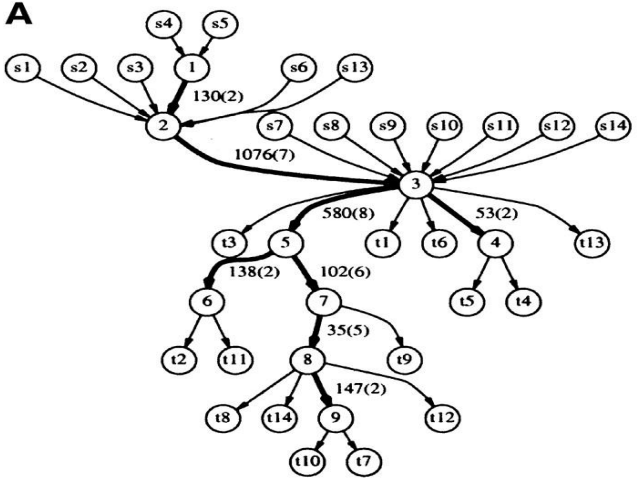
Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Figure 3

- Real world example - BAC from Chromosome Y
- Repeats + subrepeats present (3A)
- REPuter doesn’t capture sub-repeats (3B)
- A-bruijn graph approach does (3CD)

RepeatGluer representation of a 14-copy transposase IS30 repeat family in the N. meningitidis genome as a mosaic of eight sub-repeats >30 bp (shown by bold edges).

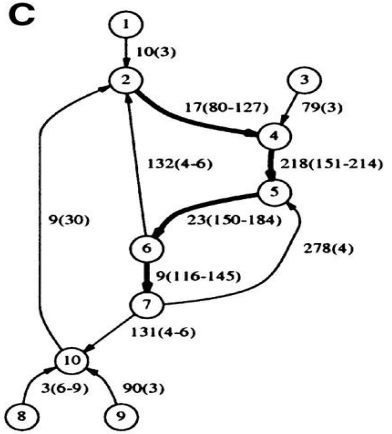
A



B

	Repeats		Sub-repeats w/in a repeat	
	# of repeat copies	Maximum length of repeat copy	# sub-repeats longer than 30bp	Maximum length of a sub-repeat
1	20	4371	37	1695
2	14	1794	8	1076
3	4	6317	3	6091
4	4	740	1	740
5	3	1475	3	641
6	3	620	4	180
7	3	838	4	387
8	2	2775	1	2775
9	2	2700	1	2700
10	2	523	1	523
11	2	494	1	494
12	2	474	1	474
13	2	456	1	456
14	2	446	1	446
15	2	420	1	420
16	2	415	1	415
17	2	410	1	410
18	2	406	1	406
19	2	402	1	402

C



Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Figure 4

- *N. meningitis* genome
- Evolutionary history of duplication captured in A-bruijn graph (4A)
  - One repeat family shown - each member contains 1-4 subrepeats
- No. of repeat families and their subrepeats (4B)
- Alu repeats in ChrX (4C)

A-bruijn graph

- Generalization of DBG for imperfect repeats
- Terms:

  Let S = genome of length n

  Let A = (a_ij) = n x n similarity matrix

  Let α = all pairwise local alignments from S to S

  a_ij = 1 iff pos. i and j are aligned somewhere in α, 0 otherwise.

A-bruijn graph, cont’d

A-graph is graph based on adjacency matrix formed by A

V = set of all connected components in A-graph

A-bruijn graph is G(V, E) -

- V - all the connected components collapsed into single vertices
- E - Eulerian path from 1->n in A-graph.

A-bruijn graph, cont’d

- α can be any arbitrary set of pw alignments
    - Ex. a DBG is an ABG where α is all k-mers in S (!)
    - Allows ABGs to be extended in cools ways (explored later)
- Here, edges represent multiple sequence alignments.
- Analysis + simplification complicated by Whirls and Bulges

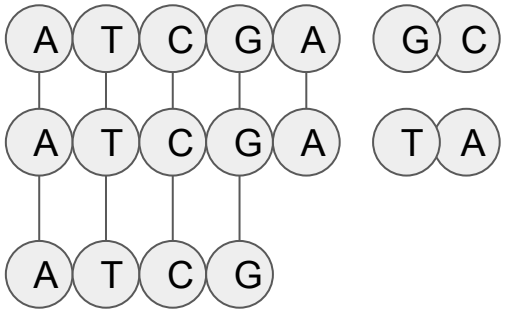
A-bruijn graph example

ATCGAGC**ATCGA**TA**ATCG**

Alignments:

ATCGA	ATCGA	ATCGA
ATCGA	ATCG	ATCG

Create A: (edges only between 1s in A matrix)



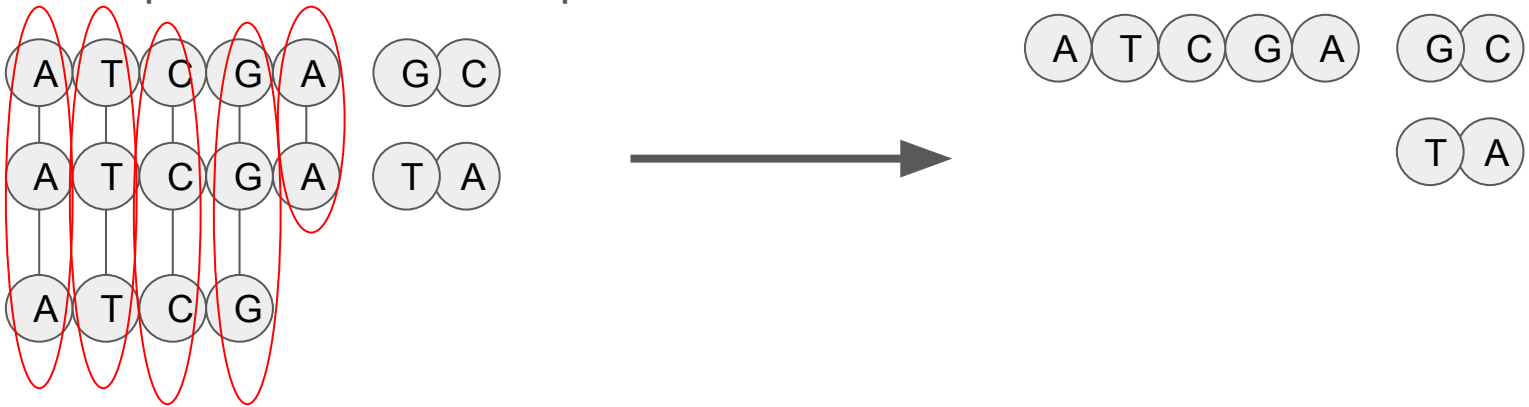
A-bruijn graph example

ATCGAGC**ATCGA**TA**ATCG**

Alignments:

ATCGA	ATCGA	ATCGA
ATCGA	ATCG	ATCG

Collapse connected components



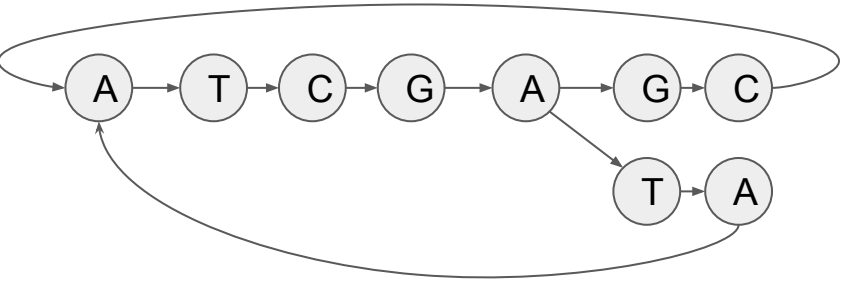
A-bruijn graph example

ATCGAGC**ATCGA**TA**ATCG**

Alignments:

ATCGA	ATCGA	ATCGA
ATCGA	ATCG	ATCG

Create edges



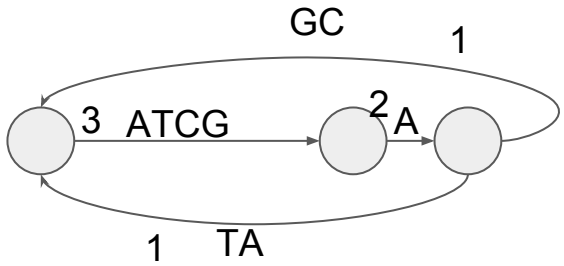
A-bruijn graph example

ATCGAGC**ATCGA**TA**ATCG**

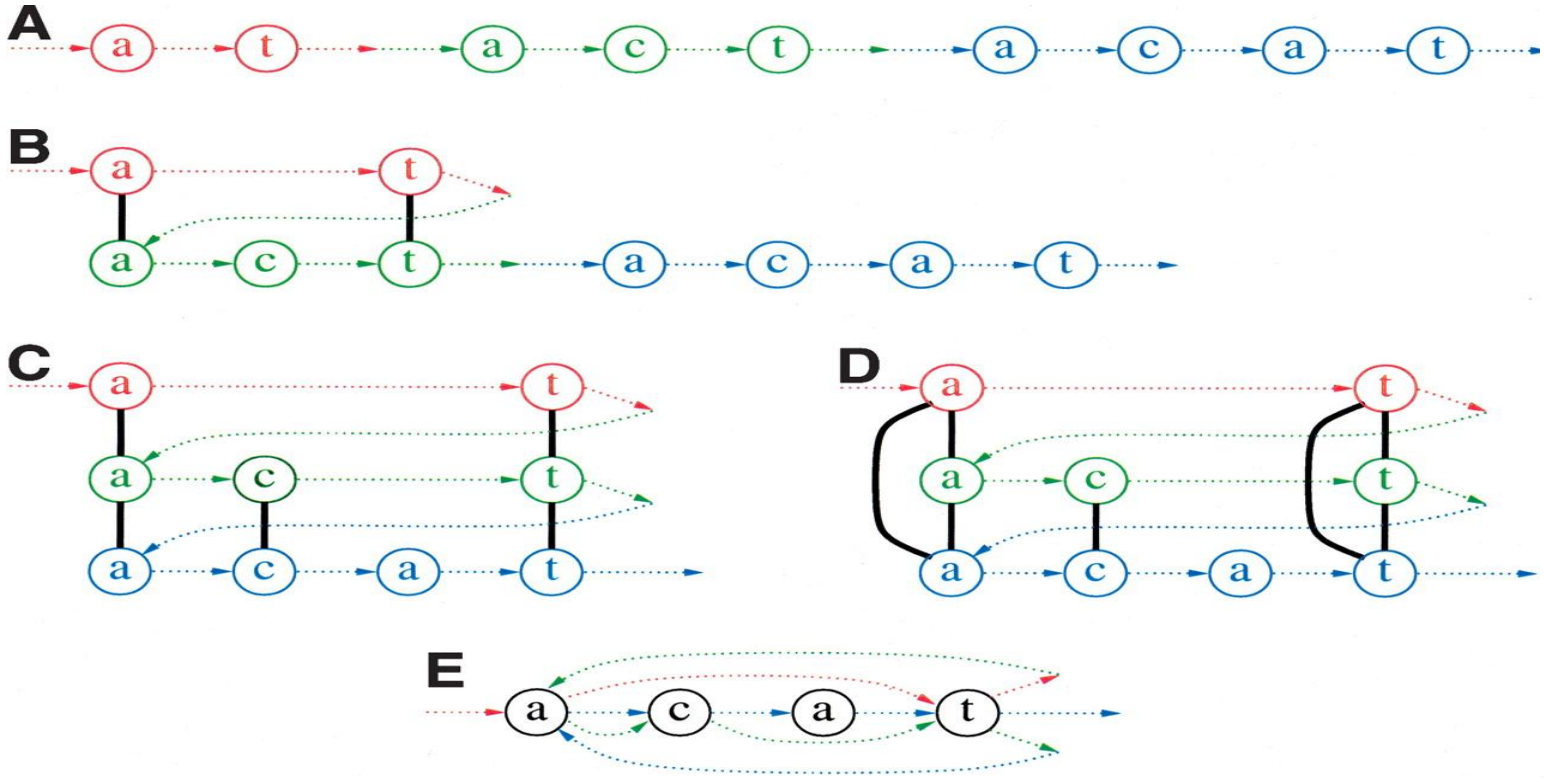
Alignments:

ATCGA	ATCGA	ATCGA
ATCGA	ATCG	ATCG

simplify



(A) Construction of the A-graph from the sequence...at...act...acat by applying three pairwise alignments (B) a-t versus act, (C) act versus acat, and (D) a-t versus acat.

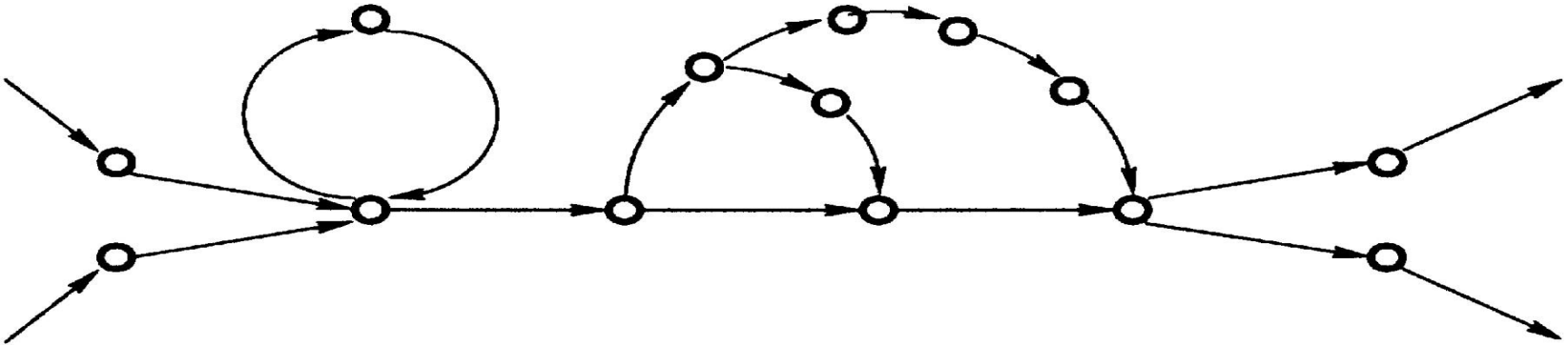


Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Whirls and Bulges

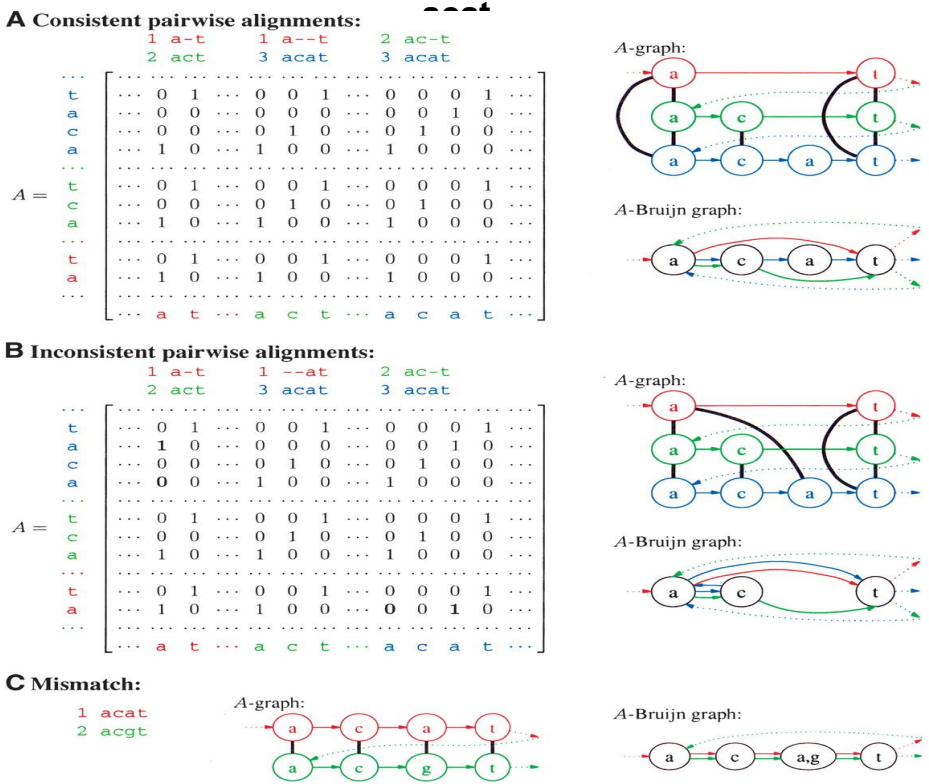
- Types of short cycles
- Whirl - edges all oriented the same dir.
  - Caused by inconsistent MSAs, tandem repeats
- Buldge - edges oriented in both dirs
  - Caused by gaps - gap of length g -> bulge of length g + 2

A repeat region in an A-Bruijn graph in which alignment inconsistencies have caused a whirl and a network of bulges.



Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Construction of A-Bruijn graphs from (A) consistent pairwise alignments and (B) inconsistent pairwise alignments, for the genomic sequence...at...act...acat..with a repeat represented by three copies: at, act, acat



Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Removing whirls

- 1 vertex represents a set of positions (P(v))
- A composite vertex contains positions that are “near” in genome
- Get max-multiplicity edge M connecting composite to non-composite vs
- Split composite v according to positions defined by M, delete edges from A accordingly.
- Repeat until no composite vertices exist
- This process separates “close” repeats and removes whirls

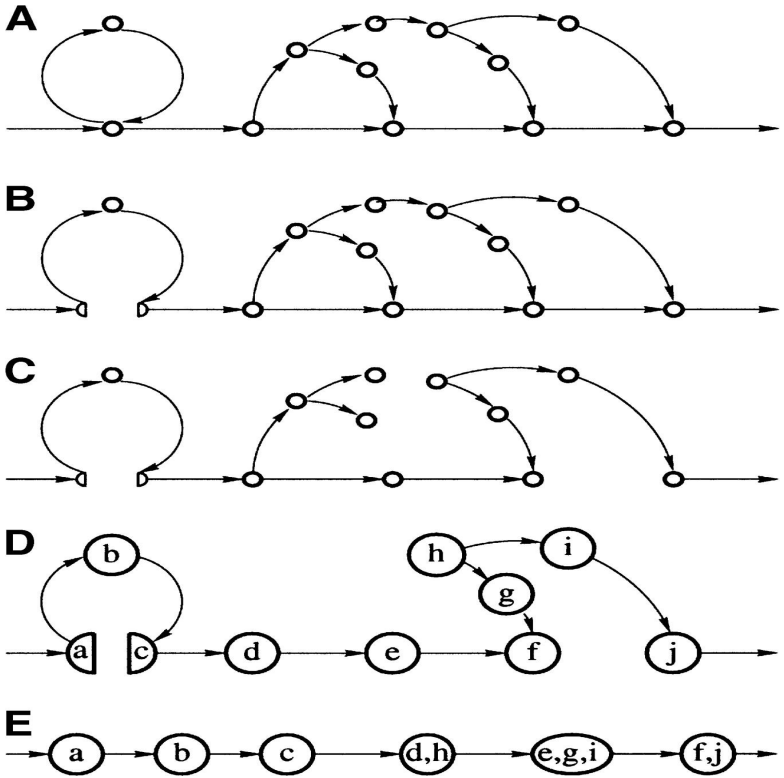
Removing bulges

- Fixed by removing a single edge
  - Which edge do we remove?
- Solve the Maximum Subgraph with Large Girth problem
  - (MST problem with arbitrary minumum cycle length)
  - This is hard, so they use an approximation
- Find a MST T, and add largest multiplicity edges possible, such that no short cycles are formed

Other clean-up procedures

- Erosion - remove leaves left by bulge removal
- Straighten zigzags
  - Erosion -> long simple paths.
  - These paths might not be unidirectional (zigzag)
  - Fix this by uniting vertices w/ same consensus nucleotide
- Threading
  - Graph now disjoint
  - “Thread” graph together using positions in S represented in each vertex v
  - This creates a “consensus” sequence for each edge

(A) Initial A-Bruijn graph (weighted graph representation instead of multigraph).



Paul A. Pevzner et al. Genome Res. 2004;14:1786-1796

Applications:

- Multiple alignments - concatenate all sequences, create de-bruijn graph
  - Partial order alignment
  - Order-independent
  - Raphael, B., Zhi, D., Tang, H., and Pevzner, P. (2004). A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res. 14, 2336–2346.
- Fragment assembly - constructing ABG without similarity matrix
  - Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19, 455–477.
- Structural Variation Discovery - by alignment to A-bruijn graph
  - Lee, H., Popodi, E., Foster, P.L., and Tang, H. (2014). Detection of Structural Variants Involving Repetitive Regions in the Reference Genome. Journal of Computational Biology 21, 219–233.

Code location

https://github.com/COL-IU/RepGraph