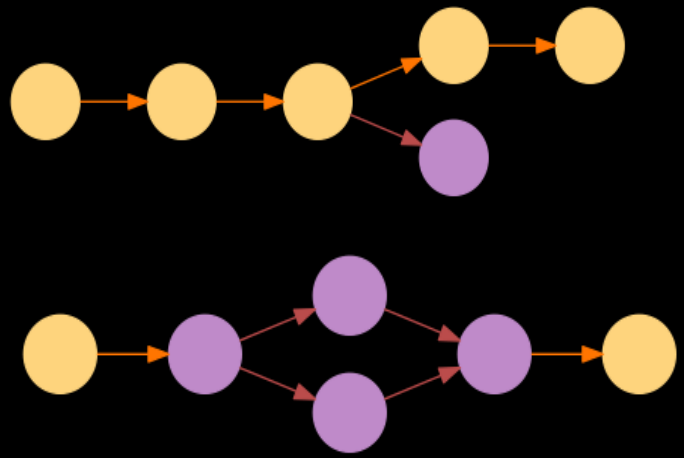
Development of algorithms for Extension index data structure

Malova Anna
Advisor: Anton Bankevich

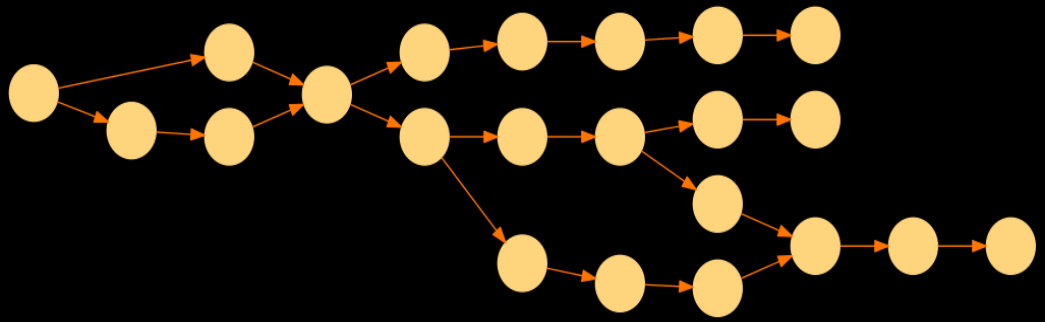
Errors in de Bruijn graph

- Tips
  - erroneous reads
- Bulges
  - mismatches
  - indels
- Erroneous connections
  - have no specific topological structure
  - removed by coverage arguments



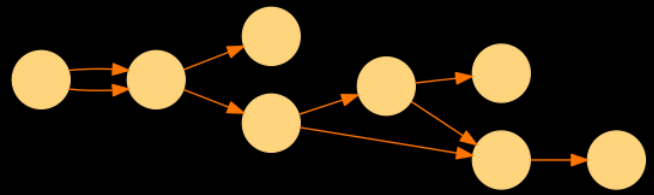
Example (without simplification)

Uncondensed graph



23 edges

Condensed graph



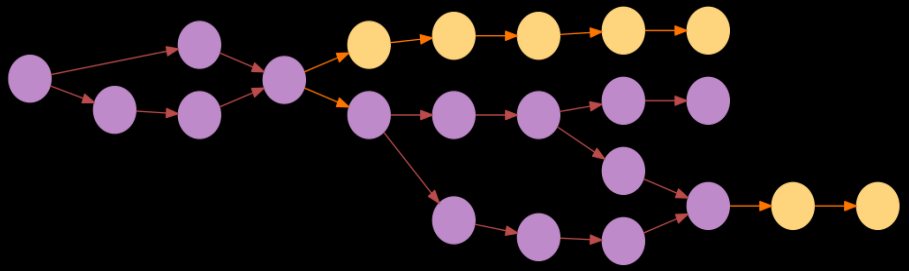
9 edges

Motivation

- Storing the unsimplified condensed graph is highly memory ineffective
- Reduce the number of edges in condensed graph
  - Only remove errors from graph
  - Don’t do anything that harms stored information

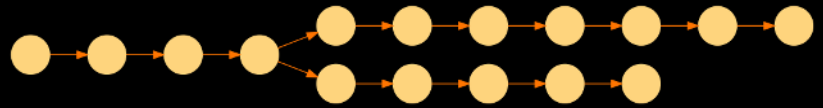
Example (simplification)

Uncondensed graph



23 edges

Simplified graph

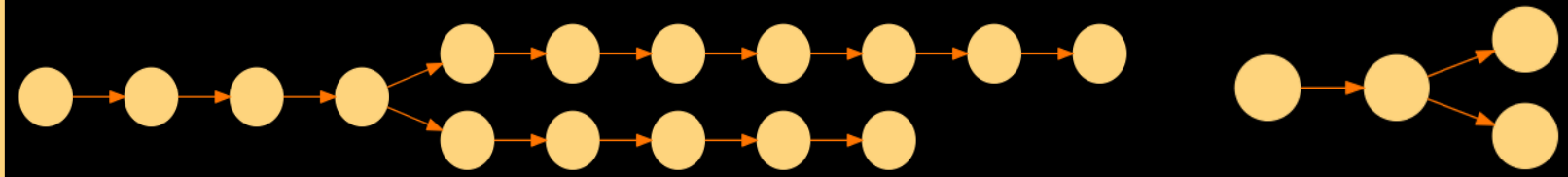


15 edges

Example (with simplification)

Simplified graph

Condensed graph



15 edges

3 edges

Extension index

- Perfect hash maps distinct elements to set of integers from 0 to N with no collision
  - Keys - all k-mers
  - Storing 8 bit information about adjacents

Project goals

- Implement bulge remover
  - Done
  - Use bfs for finding bulges
- Add storing information about coverage
  - Done
- Implement erroneous connections removing
  - Done

Results for bulge removal

dataset E.Coli is220

value of K	Number of edges before removing bulges	Number of edges after removing bulges
21	6631358	6554794
33	6096894	6084686
55	5543520	5543460

Questions

Thank you for your attention!
Questions?