Genome Assembly and de Bruijn Graphs

6.874 Spring 2011
**de novo** whole-genome shotgun assembly

Challenges for short read assembly

- Genomic regions that share perfect repeats can be indistinguishable (depending on read length)
- Separation of inexact repeats is confounded by sequencing error
- Low coverage induces gaps in assemblies
Three common approaches to short read assembly

• Greedy graph-based assemblers
  – repeat:
    – given any read or contig, add one more read or contig
  – examples: SSAKE, SHARCGS

• Overlap/Layout/Consensus assemblers
  – 3 steps:
    • all-against-all pair-wise read comparison
    • construct an overlap graph to approximate read layout
    • multiple sequence alignment determines the precise layout
  – examples: Newbler, CABOG, Edena, Shorty

• de Bruijn graph-based assemblers
  – 3 steps:
    • construct k-mer graph
    • process graph to deal with sequencing error
    • trace Eulerian path in graph
  – examples: Euler, Velvet, ABySS, AllPaths, SOAPdenovo
A k-mer graph is a form of de Bruijn graph

- every k-mer is represented with a node
  - hash table lookup allows for efficient construction
  - choice of k is a trade-off:
    - larger k: less false overlaps
    - smaller k: more true overlaps

- an edge connects nodes that overlap by k-1 bases

GACTGGGACTCC

![Diagram of k-mer graph with nodes and edges representing overlaps.](image-url)
The graph can be compressed by simplifying chains of nodes.
Error Correction

– Errors at end of read
  • Trim off ‘dead-end’ tips

– Errors in middle of read
  • Pop Bubbles

– Chimeric Edges
  • Clip short, low coverage nodes

Figure adapted from presentation by Michael Schatz
Tips are caused by errors at the ends of reads

Criteria for tip removal:
1. <2k
2. “minority count” (the tip has smaller multiplicity than other outgoing edges)
The Tour Bus algorithm removes bubbles caused by internal read errors

• Conduct a breadth-first search of the graph
• When a previously visited node is encountered:
  – backtrack and find closest common ancestor
  – extract and align sequences
  – if sequences are similar enough, merge them
An example of bubble removal
Low coverage nodes left after Tour Bus are removed

• Most unique regions have been simplified and will have coverage close to average
• Short nodes that are not erroneous are likely to be present in the genome multiple times and are not likely to have low coverage
• Thus, low coverage nodes are chimeric connections with high probability
The Breadcrumb algorithm uses paired end reads to resolve repeats

Two long contigs produced after error correction, A and B, are joined by several paired reads (red and blue arcs). The path between the two can be broken up because of a repeat internal to the connecting sequence, because of an overlap with a distinct part of the genome, or because of some unresolved errors. The small square nodes represent either nodes of the path between A and B, or other nodes of the graph connected to the former. Finding the exact path in the graph from A to B is not straightforward because of all the alternate paths that need to be explored. However, if we mark all the nodes that are paired up to either A or B (with a blue circle), we can define a subgraph much simpler to explore. Ideally, only a linear path connects both nodes.