Genome assembly paradigms

Mihai Pop

Recap

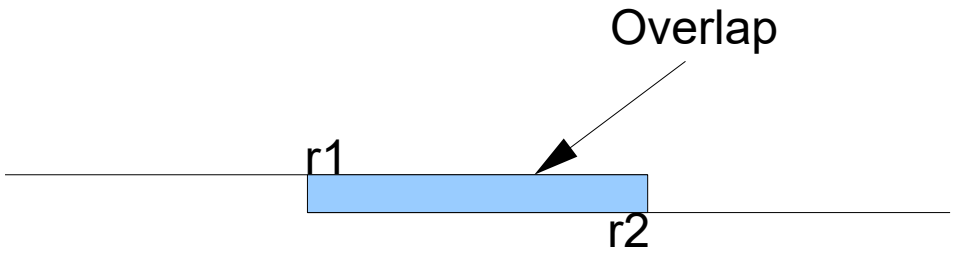
- A simple greedy algorithm can solve the assembly problem reasonably well.
- The greedy algorithm gets stuck in repeats - graph based approaches may address this problem

Genome assembly paradigms

- Greedy algorithm
  - easy to implement
  - relatively efficient
  - but... can make mistakes because it is greedy (only takes into account local information)
- How can you "reason" about repeats?
- Graph theory can help: 2 paradigms
  - Overlap-Layout-Consensus: nodes=reads, edges= reads overlap
  - deBruijn/repeat graph: nodes = k-mers, edges = k+1-mers (extracted from the reads).
- Both translate into: find a constrained path within a graph

Overlap-layout-consensus

- Essentially an extension/refinement of the greedy approach
- Given the set of reads, what can we infer about the genome?

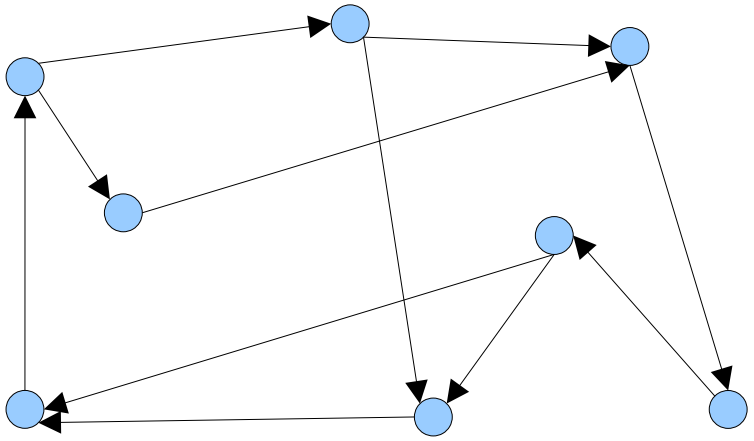


size/fidelity of overlap
=
strength/likelihood of edge

- All reads (nodes) must be used exactly once
- Algorithm?

Overlap Layout Consensus

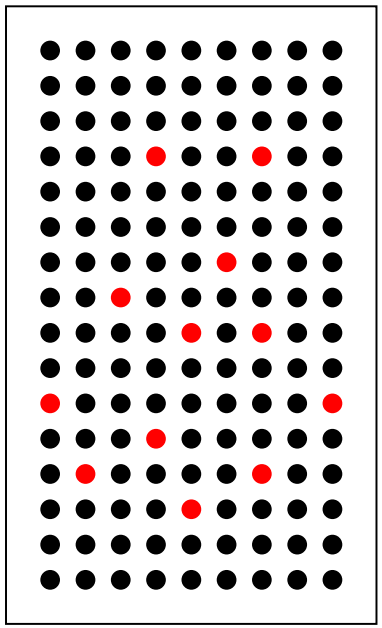
- Build a graph
- Traverse it such that each node is seen exactly once



- Hamiltonian path/cycle – NP-hard

De Bruijn graph (Eulerian) formulation

Inspiration: sequencing by hybridization



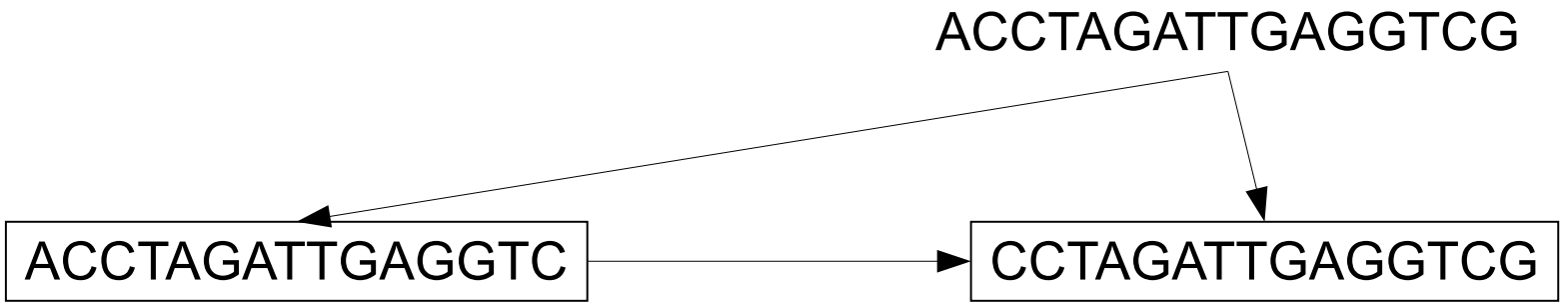
AAAA
AAAC
AAAG
AAAT
AACA
AACG
AACT
AAGA
...

probes - all possible k-mers

AACAGTAGCTAGATG

```
AACA TAGC AGAT
 ACAG AGCT GATG
  CAGT GCTA
   AGTA CTAG
    GTAG TAGA
```

De Bruijn graph formulation

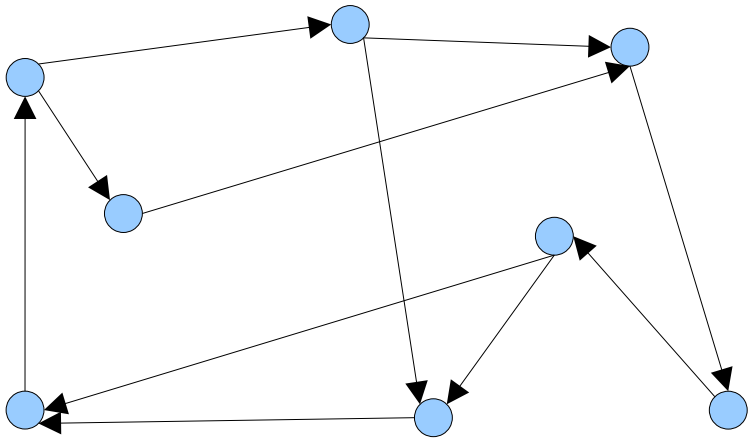


- (segment of) read = pair of k-mers overlapping by k-1 bp = edge

- Need to use all k+1-mers in the genome (the reads), i.e., all edges

de Bruijn graph

- Traverse a graph such as each edge is visited



- Exactly once – Eulerian path/cycle
- At least once (but least amount necessary) – Postman/route inspection path/cycle

- Both can be solved efficiently

Aside: graph traversals

- Hamiltonian path: visit every single node of a graph EXACTLY once (NP-hard)
- Eulerian path: visit every edge of a graph EXACTLY once (polynomial time)
- Postman/route inspection: find the shortest path in a graph that visits all the edges (i.e. Eulerian path where you allow a minimum number of edges to be reused)

- Note: a Hamiltonian path or an Eulerian path are not guaranteed to exist. A postman path can always be constructed

Eulerian circuit

The 7 bridges of Koenigsburg