CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 14

Genome assembly

Administrativia

- CMSC423 forum on CS forums
  http://forum.cs.umd.edu/

- Project questions?
Homework 3 answer

Shotgun sequencing

original DNA → shearing → sequencing → assembly
Unifying view of assembly

Shortest common superstring problem

Given a set of strings, \( \Sigma = (s_1, \ldots, s_n) \), determine the shortest string \( S \) such that every \( s_i \) is a sub-string of \( S \).

NP-hard

approximations: 4, 3, 2.89, ...

\[ \ldots \text{ACAGGACACATGATTGATAG} \]
\[ \text{ACTGACAGATTGATAGCTGA} \ldots \]

Greedy algorithm

phrap, TIGR Assembler, CAP
Repeats

```
AAAAAAA AAAAAA AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
AAAAAA  AAAAAA
```

Typical contig coverage

Imagine raindrops on a sidewalk
Lander-Waterman statistics

- $L$ = read length
- $T$ = minimum overlap
- $G$ = genome size
- $N$ = number of reads
- $c$ = coverage ($NL / G$)
- $\sigma = 1 - T/L$

- $E(\#\text{islands}) = Ne^{-c\sigma}$
- $E(\text{island size}) = \frac{L(e^{c\sigma} - 1)}{c} + 1 - \sigma$
- contig = island with 2 or more reads

All pairs alignment

- Needed by the assembler
- Try all pairs – must consider $\sim n^2$ pairs
- Smarter solution: only $n \times$ coverage (e.g. 8) pairs are possible
  - Build a table of k-mers contained in sequences (single pass through the genome)
  - Generate the pairs from k-mer table (single pass through k-mer table)
Overlap-layout-consensus

Main entity: read
Relationship between reads: overlap

Paths through graphs and assembly

• Hamiltonian circuit: visit each node (city) exactly once, returning to the start
Sequencing by hybridization

probes - all possible k-mers

AACAGTAGCTAGATG
AACATAGCAGAT
ACAGAGCTGATG
CAGTAGCTA
AGTAGCTAG
GTAGTAG

Assembling SBH data

Main entity: oligomer (overlap)
Relationship between oligomers: adjacency

ACCTGATGCCAATTGCACT...

CTGAT follows CCTGA (they share 4 nucleotides: CTGA)

Problem: given all the k-mers, find the original string

In assembly: fake the SBH experiment - break the reads into k-mers
Eulerian circuit

- Eulerian circuit: visit each edge (bridge) exactly once and come back to the start