CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 14

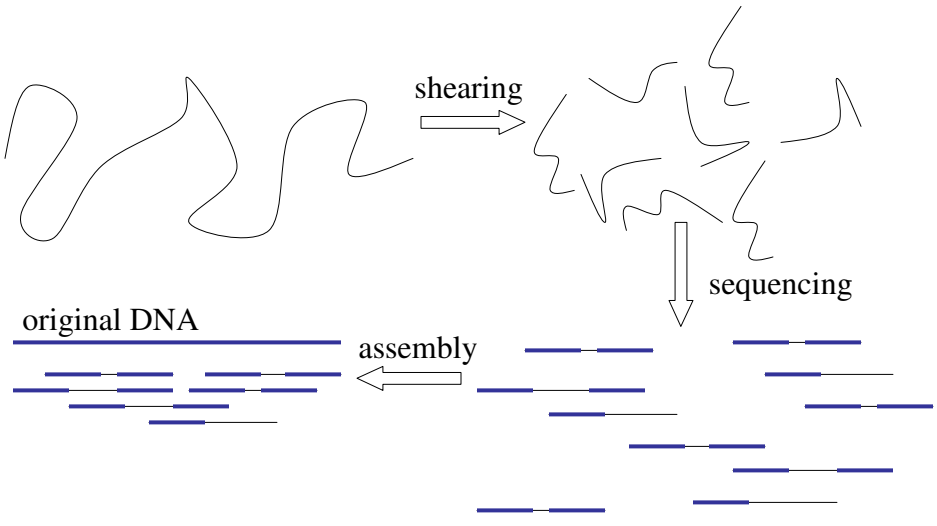
Genome assembly

Administrativia

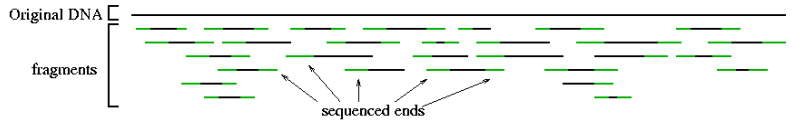
- CMSC423 forum on CS forums
  http://forum.cs.umd.edu/
- Project questions?

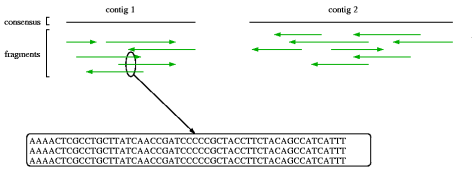
Homework 3 answer

Shotgun sequencing



Unifying view of assembly





Assembly

Scaffolding

Shortest common superstring problem

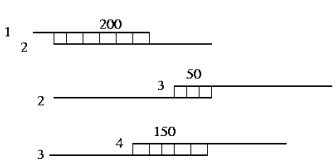
Given a set of strings, $\Sigma=(s_1, ..., s_n)$, determine the shortest string S such that every s_i is a sub-string of S.

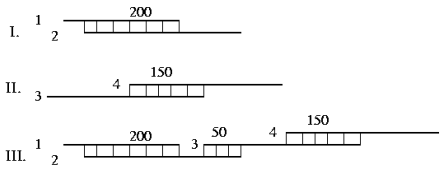
NP-hard

approximations: 4, 3, 2.89, ...

```
...ACAGGACTGCACAGATTGATAG
          ACTGCACAGATTGATAGCTGA...
```

Greedy algorithm





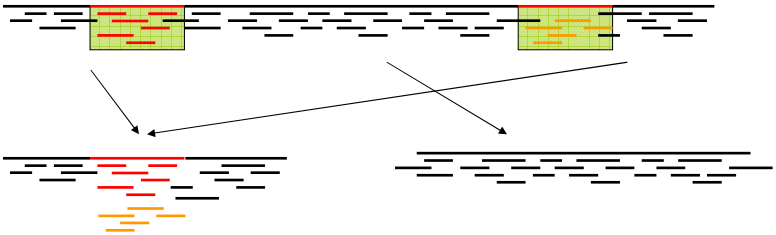
phrap, TIGR Assembler, CAP

Repeats

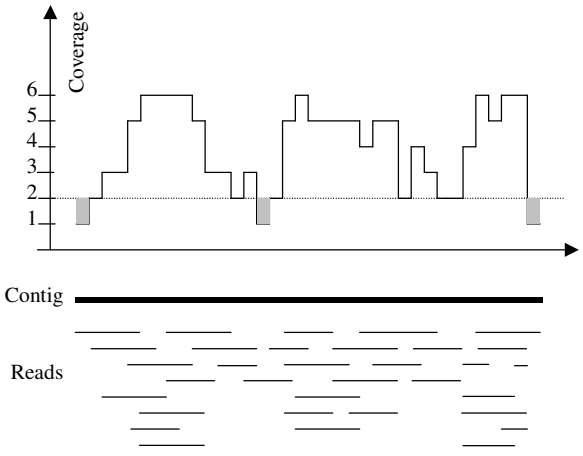
AAAAAAAAAAAAAAAAAAAA

AAAAAA AAAAAA AAAAAA
AAAAAA AAAAAA
AAAAAA AAAAAA

AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA
AAAAAA



Typical contig coverage

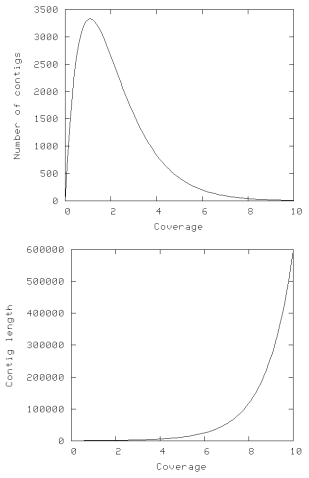


Imagine raindrops on a sidewalk

Lander-Waterman statistics

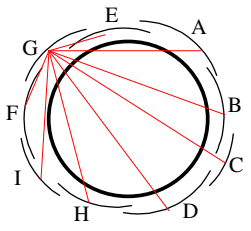
L = read length
T = minimum overlap
G = genome size
N = number of reads
c = coverage (NL / G)
$\sigma = 1 - T/L$

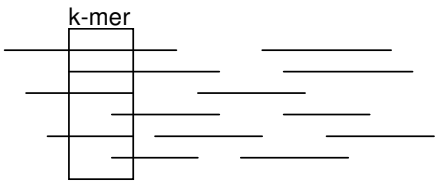
$E(\#islands) = Ne^{-c\sigma}$
$E(island\ size) = L(e^{c\sigma} - 1) / c + 1 - \sigma$
contig = island with 2 or more reads



All pairs alignment

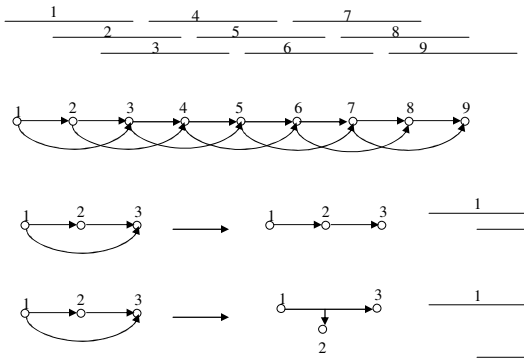
- Needed by the assembler
- Try all pairs – must consider ~ n^2 pairs
- Smarter solution: only n x coverage (e.g. 8) pairs are possible
  - Build a table of k-mers contained in sequences (single pass through the genome)
  - Generate the pairs from k-mer table (single pass through k-mer table)





Overlap-layout-consensus

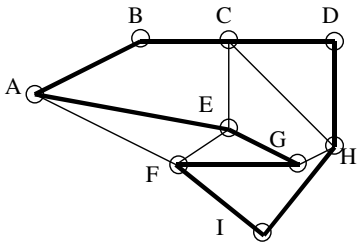
Main entity: read
Relationship between reads: overlap

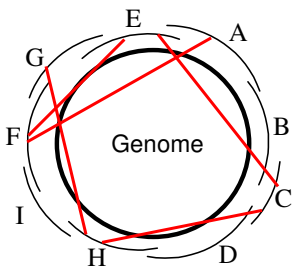


ACCTGA
ACCTGA
AGCTGA
ACCAGA

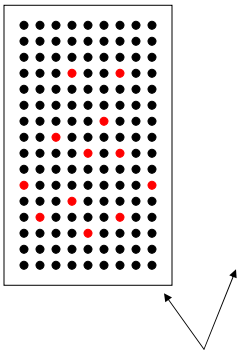
Paths through graphs and assembly

- Hamiltonian circuit: visit each node (city) exactly once, returning to the start





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
```

Assembling SBH data

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

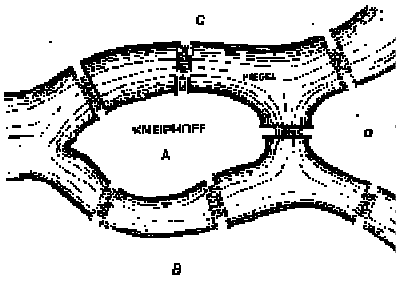
ACCTGATGCCAATTGCACT...

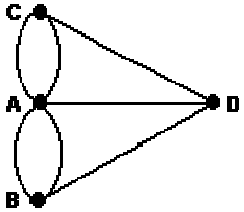
CTGAT follows CCTGA (they share 4 nucleotides: CTGA)

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