CSE5095 – Topics in Big Data Analytics
Date: 04/29/2014

Professor Rajasekaran
Note taker: Wei Zeng

Applications

1. Material Genomics

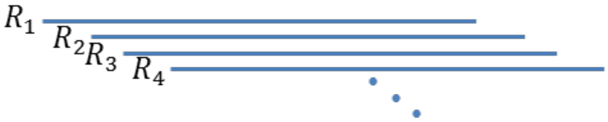
Build a repository of materials -> Text Mining

2. Sequence Assembly

Input: A set of Reads.

Sequencers output random substrings of G, each such substring is a read.

Output: A close approximation to G.



We utilize overlaps among reads.

Challenges:

1) There could be errors in the reads.
2) There could be repeats.

- **Sanger**:

  Read length≃ a few thousands.

- **Next Generation Sequencing (NGS)**:

  Has read lengths in a few tens.

Coverage: the expected number of reads that cover any position in G.

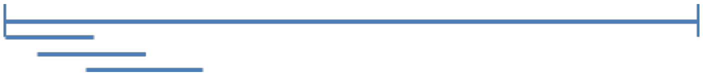
Basic Idea:

Construct a directed graph $G(V,E)$

$V \rightarrow Reads$

$(R_1, R_2) \in E$ if a suffix of R_1 of length $\geq \lambda$ is the same as a prefix of R_2.

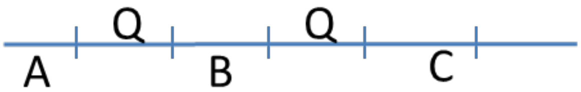
Example assemblers: VELVET, ABySS, SGA, GSA, Leap

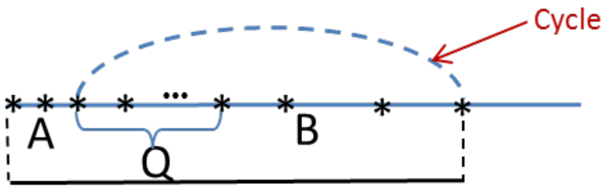


Do an appropriate walk in the graph to identify long paths and output them. Each such path is a “CONTIG”.

Observation:

Repeats cause cycles in the graph.

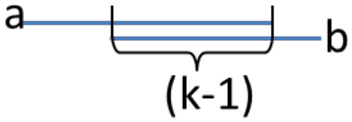




When cycle happens, just cut the path AQB and call it a CONTIG.

1) Overlap graph -> Each read is a node
2) De Bruijn Graph:

```
For every read
    Generate k-mers;
If the read length is r,
Then there will be (r-k+1) k-mers from every read.
Construct a graph $G(V,E)$ where $V$ →k-mers, $(a,b) \in E$ if $a$ & $b$ overlap by (k-1)
```



Performance Measure:

N50 value:

Sort the contigs in terms of lengths;

Let $C_1, C_2, \dots, C_N$ be the sorted sequence in nondecreasing order;

Let $\sum_{i=1}^{N} C_i = Q$;

If q is the least index such that $\sum_{i=1}^{q} C_i \geq \frac{1}{2} Q$,

Then $|C_q|$ is the N50 value

Scaffolding:

Input: A set of contigs.

Output: an ordering among the contigs.

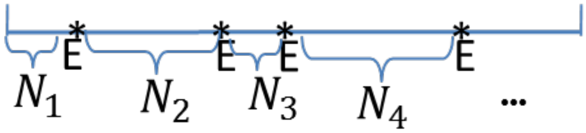
One approach is to use optical restriction maps.

Start with a restriction enzyme => a small string

(possibly generated randomly)

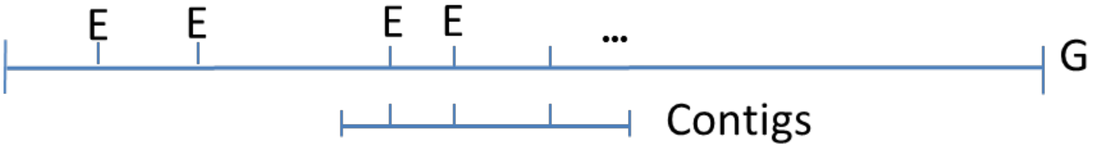
e.g., gaactat =E

Identify where E occurs in each of the contigs.



Output: $N_1, N_2, \ldots,$ ← Optical Restriction Map

We pose the problem of scaffolding as an optimization problem.



Objective function: The sum of all distance discrepancies for the contigs should be minimum.

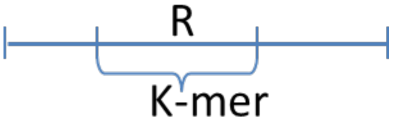
Pop et al. used dynamic programming to solve it.

Error Correction:

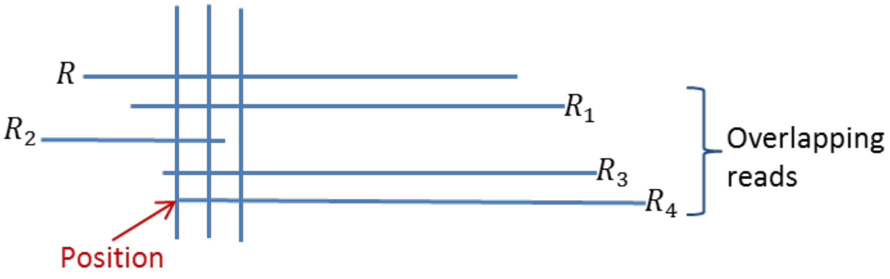
Input: A set of reads.

Output: Corrected reads.

Idea: Let R be any read



Algorithms: Reptile, Coral, RACER (improves HiTec)



Consensus in any position is used to correct that position.

Analysis:

Let ϵ be the error rate;

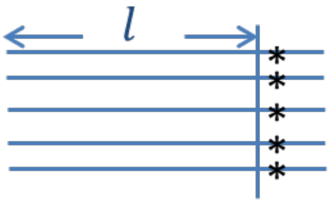
Let q be the number of reads overlapping in any column;

Let X be the number of errors in the column

$$X \rightarrow B(q, \epsilon)$$

Prob. of an incorrect correction is $= Prob\left[X \geq \frac{1}{2}q\right] = \sum_{i=\frac{q}{2}}^{q} \binom{q}{i} \epsilon^{i} (1-\epsilon)^{q-i}$

RACER: generates l-mers.



Check the next character of each read, find the majority of the character and replace the others with that character.