Outline

• Mate-pairs, scaffolding and copy number
• Diploid sequencing: Intro. genetics
• Resequencing: Structural variation.
EULER Fragment Assembly Approach

- **Input**: Reads $s_1, \ldots, s_N$
- Further subdivide reads into $k$-mers (e.g. $k = 20$)
- Build repeat graph on resulting $k$-mers
- Each read is path in resulting graph.
- Solve Eulerian Superpath Problem.  
  
  *Eulerian superpath*: an Eulerian path that contains set of paths (reads) as subpaths.

Eulerian Superpath

Example: $S = \text{ATGGCGTGCA}$

Reads $= \{\text{ATGCC}, \text{GGCGTG}, \text{GTGCA}\}$

$3$-mers $= \{\text{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT}\}$

*Eulerian superpath*: an Eulerian path that contains all these paths as subpaths.
Using Mate-Pair Information

Repeats and other ambiguities lead to **tangles** in repeat graph

1 → 3 and 2 → 4
OR
1 → 4 and 2 → 3?

Using Mate-Pair Information

Mate-pair \((r_1, r_2)\) gives pair of positions in G.

\(l(r_1, r_2)\) length of mate pair

Find path \(P\) in G from \(r_1\) to \(r_2\).

\(d(r_1,r_2) = \text{length of this path.}\)

If **unique** path \(P\) with \(d(r_1, r_2) = l(r_1, r_2)\) length of mate pair, then use \(P\) as “long read” in superpath algorithm
How many Eulerian paths?

BEST (de Bruijn, van Aardenne-Ehrenfest, Smith and Tutte) Theorem:
Number of Eulerian cycles in an Eulerian graph $G = (V,E)$ is

$$C(G) \prod_{v \in V} (d^+(v) - 1)!$$

$d^+(v) = \text{outdegree of } v$

$C(G) = \text{number of directed spanning trees rooted at } v_0$

Which Eulerian path to choose?

Stop assembly and output contig when ambiguity is reached.

1 \rightarrow 3 and 2 \rightarrow 4
OR
1 \rightarrow 4 and 2 \rightarrow 3 ?

Two types of contigs:
G-contigs: caused by gaps in coverage (on at least one side)
R-contigs: caused by repeats at sides
Comparison between Methods

In Overlap-Layout-Consensus approach, repeats were removed at first step.

Fill gaps in supercontigs with paths of repeat contigs

Use length/distance constraints

Link Contigs into Supercontigs

Find all links between unique contigs

[Diagram showing the process of linking contigs into supercontigs]
Using Mate-Pair Information

Scaffolding: Joining contigs

Using Mate-Pair Information

Sequence

Contigs generated by overlap-layout-consensus assembler

Incorrect scaffold

Graph generated by Euler

Correct scaffold generated by Euler

6
Copy number problem

Let $d(v) = \text{indegree} - \text{outdegree}$

**Balanced graph**: $d(v) = 0$ for all $v$.

With real genome sequencing data:
- Do not get exact count of each $k$-mer in genome,
- only an approximation based on random sampling of reads.
- Resulting deBruijn graphs are not balanced.

Copy number problem

Goal: Introduce multiplicities on edges so that graph is balanced.

Optimization criterion: Use as few extra edges as possible.

Balance each vertex by adding edge multiplicities (weights)

Assign flow $f(e)$ to each edge such that $d(v) = 0$ for all vertices.
Copy number problem

Let \( d(v) = \text{indegree} - \text{outdegree} \)

**Balanced graph**: \( d(v) = 0 \) for all \( v \).

Graph \( G = (V, e, w) \).

**Copy Number Problem** (Pevzner & Tang 2001): For an edge \( e \) in \( G \), find a flow minimizing the multiplicity \( f(e) \) of \( e \) and satisfying \( f(e) \geq 1 \).

**Min-flow Max-cut Theorem**: For a directed acyclic graph \( G = (V, E, w) \) with lower capacity bounds on each edge: \( \min \text{ flow from } v \text{ to } w = \text{capacity of the maximum cut separating } v \text{ from } w \).
Copy number problem

**Copy Number Problem** (Pevzner & Tang 2001): For an edge $e$ in $G$, find a flow minimizing the multiplicity $f(e)$ of $e$.

Min-cost circulation (See Myers 2005): Assign cost $c(e) = 1$ to each edge.

$$\min \sum c(e) f(e) \text{ such that } f(e) \geq w(e) \text{ for all } e.$$  
$$d(v) = 0 \text{ for all vertices.}$$

---

Fragment Assembly Summary

Two major approaches

1) **Overlap-Layout-Consensus**
   - **Graph**: Vertices are reads. Edges are overlaps.
   - **Assembly**: Hamiltonian path. (Uses each vertex once).
   - **Repeats**: Removed/masked before assembly.

2) **deBruijn graph (Euler)**
   - **Graph**: Vertices are (k-1)-mers. Edges are k-mers. Reads split into k-mers
   - **Assembly**: Eulerian path
   - **No removal of repeats.**
New applications of sequencing

• Personal genomics
  – Find genome sequence for an individual.
  – Find individual differences from reference genome.

New applications of sequencing

• Metagenomics: Simultaneous sequencing of all organisms in an environment
  – Sample of ocean water.
  – Acidic water from mine.
  – Human gut.
  – New York air.
New applications of sequencing

• Cancer genomics
  – Find genome sequence for an individual.
  – Find individual differences from reference genome.

Personal Genomics

Humans are diploid organisms: two copies (paternal and maternal) of each chromosome

[Image: Karyotype diagram]
Genetics 101

• Thus, far we have **haploid** genome assembly (one copy of each chromosome).

• Most difference b/w chromosomes (individuals) are **single nucleotide polymorphisms (SNPs)**

  ![Reference genome vs. individual genome]

  • Individual has ~4 million SNPs.

Genetics 101

• **Locus**: Region on a chromosome (gene, nucleotide, etc.)

• **Allele**: “Value” at a locus.

• SNP is a single nucleotide locus.

• Most SNPs are biallelic: two possible states. (We will assume biallelic).

  ![Reference genome vs. individual genome]

• **Major (minor) allele**: allele at locus with highest (lowest) frequency in population.
Genetics 101

• Can represent biallelic SNPs in binary (0/1)
  ...CCTAACTAGACTACGGAACGCT... Reference genome
  ...CCGAACTAGCTACGGAATGCT... Chromosome of individual

• Represent a chromosome as binary string.
• Suppose 0 → minor allele. 1 → major allele and all major alleles in reference.
  – Chromosome above is 0010

Genetics 101

• **Genotype**: Pair of alleles (maternal and paternal) at locus

• Three possible genotypes for a biallelic SNP:
  00 homozygous (minor)
  01 heterozygous
  11 homozygous (major)
Genetics 101

**Haplotype**: Alleles of loci on *same* chromosome (maternal *or* paternal)

**Genotypes** \{01,01,11,01,00,01,11\}
(unordered) pairs of alleles

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Maternal chromosome</th>
<th>Paternal chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>...011001...</td>
<td>...101111...</td>
<td></td>
</tr>
</tbody>
</table>

DNA Sequencing of Diploids

Whole genome shotgun sequencing yields genotypes.

<table>
<thead>
<tr>
<th>Reads</th>
<th>Maternal chromosome</th>
<th>Paternal chromosome</th>
</tr>
</thead>
<tbody>
<tr>
<td>...CTAACTAGTTACCCGAACGCT...</td>
<td>...CTAACTAGTTACCCGAACGCT...</td>
<td></td>
</tr>
<tr>
<td>...CCAACTAGCTACCCGAATGCT...</td>
<td>...CCAACTAGCTACCCGAATGCT...</td>
<td></td>
</tr>
</tbody>
</table>

Can we recover the haplotypes during assembly?

Bansal and Bafna paper.
Sources

• Serafim Batzoglou
  http://ai.stanford.edu/~serafim/CS262_2006/ (Sequencing slides)
• http://bioalgorithms.info (Euler slides)
• Euler assembler. Pevzner, Tang, and Waterman (PNAS 2001): on website