Next gen sequencing and assembly
DNA

...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTTAGATACCTGAC
TGATTTTTAAAAAAATATT...
PHASE TWO: INTERPRETATION

I THINK I FOUND A CORNER PIECE.
Genome Assembly

Input: Multiple copies of the genome

Process: Randomly fragment each copy
“Traditional” genome project

Slide from Mihai Pop and Michael Schatz
Review

• Early days:
  – “Sanger” sequencing using gels

• “good old” days:
  – Sanger sequencing using capiliaries

• Future
  – Pyrosequencing
  – Single molecule reading
DNA chromatogram
Sequencing technology

Sanger sequencing

Cost per finished bp:

- 1975: $10.00
- 1980: $1.00
- 2000: $0.10
- 2008: $0.01

Read length:
- 15 – 200 bp
- 500 – 1,000 bp

Throughput:
- "grad-student years"
- $2 \cdot 10^6$ bp/day

Source: bioinformatics.org
Sequencing technology
Next-generation sequencing

Read length: 450 bp  “short reads”
Throughput: 600-800 Mb/day
Cost: ~ 20,000 bp/$
De novo: yes

Genome Sequencer / FLX
454 Sequencing

Roche (454) GSFLX Workflow:
- Library construction
- Emulsion PCR
- PTP loading

Pyrosequencing reaction:
- DNA capture bead containing millions of copies of a single clonally amplified fragment
- Sulfurylase
- Luciferase
- Polymerase
- APS
- PPi
- Annealed primer
- ATP
- Luciferin
- Light + Oxy Luciferin
Nanopore Sequencing

The Envisioned Device:
A Solid State Nanopore with Embedded Nanotube Sensor
Goal

• Sequencing a human genome
$985 deCODEme (November 2007)

$399 Personal Genome Service (November 2007)

$2,500 Health Compass service (April 2008)

Genetic Information Nondiscrimination Act (May 2008)

$\sim1,000$ Whole-genome sequencing (2013)
Demand for more sequencing

Sequencing technology improvement

New sequencing applications

Increase in sequencing data output

Demand for more sequencing
## Applications

<table>
<thead>
<tr>
<th>Whole-genome sequencing</th>
<th>Genotyping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative genomics</td>
<td>Population genetics</td>
</tr>
<tr>
<td>Genome resequencing</td>
<td>Migration studies</td>
</tr>
<tr>
<td>Structural variation analysis</td>
<td>Ancestry inference</td>
</tr>
<tr>
<td>Polymorphism discovery</td>
<td>Relationship inference</td>
</tr>
<tr>
<td>Metagenomics</td>
<td>Genetic screening</td>
</tr>
<tr>
<td>Environmental sequencing</td>
<td>Drug targeting</td>
</tr>
<tr>
<td>Gene expression profiling</td>
<td>Forensics</td>
</tr>
<tr>
<td>Technology</td>
<td>Read length (bp)</td>
</tr>
<tr>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Sanger</td>
<td>1,000</td>
</tr>
<tr>
<td>454</td>
<td>250</td>
</tr>
<tr>
<td>Solexa/ABI</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Application</th>
<th>Sanger</th>
<th>454</th>
<th>Solexa/ABI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial sequencing</td>
<td>yes</td>
<td>yes</td>
<td>maybe</td>
</tr>
<tr>
<td>Mammalian sequencing</td>
<td>yes</td>
<td>sort of</td>
<td>probably no</td>
</tr>
<tr>
<td>Mammalian resequencing</td>
<td>Lots of $$</td>
<td>Lots of $$</td>
<td>yes</td>
</tr>
</tbody>
</table>
Shortest Superstring Problem

• **Problem**: Given a set of strings, find a shortest string that contains all of them

• **Input**: Strings \(s_1, s_2, \ldots, s_n\)

• **Output**: A string \(s\) that contains all strings \(s_1, s_2, \ldots, s_n\) as substrings, such that the length of \(s\) is minimized

• **Complexity**: NP – complete

• **Note**: this formulation does not take into account sequencing errors
Shortest Superstring Problem: In class

The Shortest Superstring problem

Set of strings:  {000, 001, 010, 011, 100, 101, 110, 111}

Concatenation  000 001 010 011 100 101 110 111
Superstring      000 001 010 011 100 101 110 111
Shortest Superstring Problem: Example

The Shortest Superstring problem

Set of strings: \{000, 001, 010, 011, 100, 101, 110, 111\}

Concatenation
Superstring

Shortest superstring
Overlap-Layout-Consensus
Assemblers: ARACHNE, PHRAP, CAP, TIGR, CELERA

**Overlap:** find potentially overlapping reads

**Layout:** merge reads into contigs and contigs into supercontigs

**Consensus:** derive the DNA sequence and correct read errors
Triazzle: A Fun Example

The puzzle looks simple

BUT there are repeats!!!

The repeats make it very difficult.

Try it – $10.99 at www.triazzle.com

iPhone version too!
Compressed de Bruijn Graph

It was the best of times, it

of times, it was the

it was the worst of times, it

the age of foolishness

it was the age of

the age of wisdom, it was the
The Bridge Obsession Problem

Find a tour crossing every bridge just once

*Leonhard Euler, 1735*

*Bridges of Königsberg*
Challenges in Fragment Assembly

• Repeats: A **major** problem for fragment assembly
• > 50% of human genome are repeats:
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)

Green and blue fragments are interchangeable when assembling repetitive DNA