CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

Giri Narasimhan
ECS 254; Phone: x3748
giri@cis.fiu.edu
www.cis.fiu.edu/~giri/teach/BioinfS11.html
Other sequencing methods

- Sanger Method (70Kbp/run)
- Sequencing by Hybridization (SBH)
- Dual end sequencing
- Chromosome Walking (see page 5-6 of Pevzner’s text)
- 454 Sequencing (60Mbp/run)
- Solexa Sequencing (600Mbp/run) [Illumina]
454 Sequencing: New Sequencing Technology

- 454 Life Sciences, Roche
- Fast (20 million bases per 4.5 hour run)
- Low cost (lower than Sanger sequencing)
- Simple (entire bacterial genome in days with one person -- without cloning and colony picking)
- Convenient (complete solution from sample prep to assembly)
- PicoTiterPlate Device
  - Fiber optic plate to transmit the signal from the sequencing reaction
- Process:
  - Library preparation: Generate library for hundreds of sequencing runs
  - Amplify: PCR single DNA fragment immobilized on bead
  - Sequencing: “Sequential” nucleotide incorporation converted to chemilluminiscent signal to be detected by CCD camera.
(a) Fragment, (b) add adaptors, (c) “1 fragment, 1 bead”, (d) emPCR on bead, (e) put beads in PicoTiterPlate and start sequencing: “1 bead, 1 read”, and (f) analyze
emPCR

**Figure 8**

- **DNA Library Preparation**: 4.5 hours
  - Anneal sstDNA to an excess of DNA Capture Beads

- **emPCR**: 8 hours
  - Emulsify beads and PCR reagents in water-in-oil microreactors
  - Clonal amplification occurs inside microreactors

- **Sequencing**: 7.5 hours
  - Break microreactors enrich for DNA-positive beads

**gDNA** → **sstDNA Library**
Sequencing

FIGURE 9

- DNA Library Preparation
- emPCR
- Sequencing

4.5 HOURS
8 HOURS
7.5 HOURS

- Well diameter: average of 44μm
- 400,000 reads obtained in parallel
- A single cloned amplified ssDNA bead is deposited per well

Amplified ssDNA library beads

Quality filtered bases
Sequencing

**FIGURE 10**

- **DNA Library Preparation**: 4.5 hours
- **emPCR**: 8 hours
- **Sequencing**: 7.5 hours

- 4 bases (TACG) cycled 100 times
- Chemiluminescent signal generation
- Signal processing to determine base sequence and quality score

**DNA Capture Bead**

- Containing millions of copies of a single clonal fragment

**Amplified ssDNA library beads**

- Quality filtered bases
Solexa Sequencing

1. PREPARE GENOMIC DNA SAMPLE
   - Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE
   - Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION
   - Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Solexa Sequencing

4. Fragments become double-stranded

5. Denature the double-stranded molecules

6. Complete amplification

The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denaturation leaves single-stranded templates anchored to the substrate.

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
Solexa Sequencing

7. DETERMINE FIRST BASE

First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
Solexa Sequencing

10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment at a single base at a time.
Assemblers

- TIGR Assembler (TIGR)
- Phrap (U Washington)
- Celera Assembler (Celera Genomics)
- Arachne (Broad Institute of MIT & Harvard)
- Phusion (Sanger Center)
- Atlas (Baylor College of Medicine)
Applications of Sequencing

- Sequencing
- Resequencing
- SNP detection
- RNA-Seq
- CHiP-Seq
- Metagenomics
Basic Assembler

- Read: sequenced fragment; Contig: contiguous segment. How to assemble a contig?

TCGAGTTAAGCTTTAG
CGAGTTAAGCTTTAGC
AGTTAAGCTTTAGCCT
GTAAAGCTTTAGCCTA
AGCTTTAGCCTAGGGC
GCTTTAGCCTAGGCAG

Problem: Need to try every pair of reads!
Reduce to Graph Problem

- **How to assemble a contig?**
  - Node → Read
  - Edge between Nodes ← Overlapping Reads
  - **Problem**: Find a path through each node in graph.

**Issues**: Problem is NP-Complete

- # nodes = # reads
- # of edges ≤ k(# nodes)
String graph

- Combine nodes that form paths into strings
A better solution

- Take each read and chop it into k-mers.
- Represent k-mers by nodes in a graph and edges between k-mers that overlap in k-1 bases.

**Consequence:**
- Number of nodes = $4^k$;
- Number of edges = $k4^k$;

**Issues:**
- Problem (i.e., find path through all vertices) remains NP-Complete
A more efficient solution

- Represent every possible (k-1)-mer by a node.
- Edges connect 2 nodes if they share k-2 bases.
- Label each edge by k-mer.

Problem:
- Find a path through each edge in the graph
- The Eulerian path problem is NOT NP-Complete. It can be solved in linear time!
Sources of Assembly Errors

- **Errors in reads - caused by technology**
  - Error in base calls, color calls (SOLID Technology), or repeated base calls (454 Technology)

- **Missing reads - sequencing bias**

- **Read orientation error**
  - One or both orientations may occur
  - Not told which ones are present

- **Sequence Variations - mixed sample study**
  - SNP, cancer, metagenomics studies

- **REPEATS**

- **Combinations of the above**
How to deal with REPEAT Regions

- If no errors or repeat regions, then the graph has a unique path through all the edges.

- **Problem**: REPEAT regions cause branching in graph. If no errors in reads, then the graph has a unique path through all edges, but with some edges traversed more than once.

- How to identify REPEAT regions:
  - Higher coverage of repeat regions
  - Branching of nodes