CSCI2950-C
DNA Sequencing and Fragment Assembly

Lecture 2: Sept. 15, 2009
http://cs.brown.edu/courses/csci2950-c/

DNA sequencing

How we obtain the sequence of nucleotides of a species
Outline

• DNA Sequencing Technology: Old and New
• Fragment Assembly Problem
  • Overlap-Layout-Consensus
• Sequencing by Hybridization
  • Eulerian and Hamiltonian Graphs

How to Sequence DNA?

• Co-opt machinery of the cell.
• Every time a cell divides it copies its DNA sequence.
DNA Replication

- One strand used as template to make a copy of other strand.
DNA Replication

Suppose we could:
1) Stop this reaction and read the last nucleotide inserted.
2) Measure the length of DNA molecules.

DNA Sequencing

1. Start at fixed location (primer)
2. Grow DNA chain
3. Include mixture of normal nucleotides (A,C,T,G) and modified (fluorescent/radioactive) nucleotides.
4. Modified nucleotides stop reaction at all possible points
5. Separate products with length, using gel electrophoresis

Sanger sequencing (F. Sanger 1975 -- Nobel prize 1980)
Technical Limitations

1. Need a lot of DNA
2. Reaction only works for 500-1000bp

Solutions

1. Biology
2. Computer Science

DNA Sequencing – Cloning

DNA

Fragment

DNA fragments

Vector
Circular genome (bacterium, plasmid)

+ =

Known location (restriction site)

Many host cells $\rightarrow$ DNA amplification
Different types of vectors

<table>
<thead>
<tr>
<th>VECTOR</th>
<th>Size of insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasmid (bacteria)</td>
<td>2,000-10,000</td>
</tr>
<tr>
<td></td>
<td>Can control the size</td>
</tr>
<tr>
<td>Fosmid (bacteria)</td>
<td>40,000</td>
</tr>
<tr>
<td>BAC (Bacterial Artificial Chromosome)</td>
<td>70,000-300,000</td>
</tr>
<tr>
<td>YAC (Yeast Artificial Chromosome)</td>
<td>&gt; 300,000</td>
</tr>
<tr>
<td></td>
<td>Not used much recently</td>
</tr>
</tbody>
</table>

Disadvantages of Traditional (Sanger) Sequencing

1. Cloning process is laborious. Grow (bacterial) cells, each containing a single fragment.
2. Only one fragment sequenced at a time.
**Polony Sequencing**

1. **PREPARE GENOMIC DNA**
   - 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.

4. **FRAGMENT COMPLEMENTARY DNA**
   - The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. **COMPLETE AMPLIFICATION**
   - Several million dense clusters of double-stranded DNA are generated at each channel of the flow cell.

**Illumina Sequencing**

1. **PREPARE GENOMIC DNA**
   - 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.

4. **FRAGMENT COMPLEMENTARY DNA**
   - The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. **COMPLETE AMPLIFICATION**
   - Several million dense clusters of double-stranded DNA are generated at each channel of the flow cell.
Illumina Sequencing

DNA Sequencing

Goal:
Find the complete sequence of A, C, G, T’s in DNA

Challenge:
There is no machine that takes long DNA as an input, and gives the complete sequence as output

Can only sequence:
500-1000 letters at a time (Sanger sequencing)
25-150 letters (Next-gen sequencing)
Issues

• How to process data from sequencing machine?
  – Base-calling

• How to obtain sequence for longer DNA molecules?

Sequence Trace
Challenging to read answer

Reading an sequence trace

1. Filtering
2. Smoothening
3. Correction for length compressions
4. A method for calling the letters – PHRED

PHRED – PHil’s Read EDitor (by Phil Green)

Better methods exist, but labs are reluctant to change.
New methods for different sequencing technologies.
Output of PHRED: a read

A read: 500-700 nucleotides

A C G A A T C A G A
16 18 21 23 25 15 28 30 32 ... 21

Quality scores: $-10 \times \log_{10} \text{Prob(Error)}$

Reads can be obtained from leftmost, rightmost ends of the insert

---

Sequencing Longer Regions

Cover region with ~7-fold redundancy (7X)

Overlap reads and extend to reconstruct the original genomic region
Questions

1. How much to sequence/oversample?
2. How to assemble the sequence?

Definition of Coverage

\[ c = \frac{n \cdot L}{G} \]

Length of genomic segment: \( G \)
Number of reads: \( N \)
Length of each read: \( L \)

**Definition:** Coverage \( c = \frac{n \cdot L}{G} \)

How much coverage is enough?

**Lander-Waterman model:**
Assuming uniform distribution of reads, \( C=10 \) results in 1 gapped region /1,000,000 nucleotides
Lander-Waterman Statistics

**Given:** \( N \) reads of length \( L \) from a genome of size \( G \).

Assume left end of reads uniformly distributed on genome.

\[
P(\text{read } R \text{ starts at } \zeta) = \frac{1}{G}
\]

\[
P(\zeta \text{ covered by read } R) = \frac{L}{G}
\]

\[\zeta\]

\[
P(\zeta \text{ covered by read}) = 1 - (1 - \frac{L}{G})^N
\]

\[\approx 1 - e^{-c}, \text{ where } c = \frac{NL}{G} \text{ is coverage}
\]

**Poisson approximation**

Rescale \( g = \frac{G}{L}, l = \frac{L}{L} \).

Left ends of reads are Poisson process with rate \( c = \frac{NL}{G} \).

\[
\Pr[k \text{ reads in interval of length } I] = e^{-ct} (ct)^k / k!
\]

<table>
<thead>
<tr>
<th>( c )</th>
<th>( P(\text{covered}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.632</td>
</tr>
<tr>
<td>2</td>
<td>0.864</td>
</tr>
<tr>
<td>4</td>
<td>0.982</td>
</tr>
<tr>
<td>8</td>
<td>0.999</td>
</tr>
</tbody>
</table>
Questions

1. How much to sequence/oversample?
2. How to assemble the sequence?

Fragment Assembly

**Computational Challenge**: assemble individual short fragments (reads) into a single genomic sequence ("superstring").

Objective function?
Shortest Common Superstring Problem (SCS)

- **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
  - Has minimum length.

**Note**: this formulation does not take into account sequencing errors

---

**Shortest Common Superstring Problem: Example**

<table>
<thead>
<tr>
<th>Set of strings: {000, 001, 010, 011, 100, 101, 110, 111}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenation Superstring</td>
</tr>
<tr>
<td>000 001 010 011 100 101 110 111</td>
</tr>
<tr>
<td>[000]</td>
</tr>
<tr>
<td>[001]</td>
</tr>
<tr>
<td>[010]</td>
</tr>
<tr>
<td>[011]</td>
</tr>
<tr>
<td>[100]</td>
</tr>
<tr>
<td>[101]</td>
</tr>
<tr>
<td>[110]</td>
</tr>
<tr>
<td>[111]</td>
</tr>
<tr>
<td>Shortest superstring</td>
</tr>
<tr>
<td>0 0 0 1 1 1 0 1 0 0</td>
</tr>
<tr>
<td>[000]</td>
</tr>
<tr>
<td>[001]</td>
</tr>
<tr>
<td>[001]</td>
</tr>
<tr>
<td>[110]</td>
</tr>
<tr>
<td>[101]</td>
</tr>
<tr>
<td>[110]</td>
</tr>
</tbody>
</table>
Greedy Algorithm for SCS

1. Find pair of strings $s_i, s_j$ with longest overlap.
2. Merge($s_i, s_j$)
3. Recurse

How “good” is the greedy algorithm?

Algorithm Analysis

$S = \{s_1, s_2, \ldots, s_n\}$
Opt($S$) = length of SCS

**Claim:** Greedy solution $\leq 4 \text{ Opt}(S)$.

**Conjecture:** Greedy solution $\leq 2 \text{ Opt}(S)$.
[Best known bound 2.5]

**Theorem:** SCS is NP–complete
SCS and Overlap Graph

Build directed graph $G = (V,E)$
$V = \{s_1, s_2, \ldots, s_n\}$
$e = (s_i, s_j)$ if prefix of $s_j$ matches suffix of $s_i$
$w(s_i, s_j) = \text{length of overlap b/w } s_i, s_j$

Goal: Find a maximum weight path visiting every VERTEX exactly once in the OVERLAP graph:

SCS to TSP: An Example

$S = \{ \text{ATC, CCA, CAG, TCC, AGT} \}$
Fragment Assembly

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 yellow fragments are interchangeable when assembling repetitive DNA
Repeat Types

Repeat types:

- **Low-Complexity DNA** (e.g. ATATATATACATA…)
- **Microsatellite repeats** \((a_1 \ldots a_k)^n\) where \(k \approx 3-6\) (e.g. CAGCAGTAGCAGCACCAG)
- **Transposons**
  - **SINE** (Short Interspersed Nuclear Elements)
    e.g., ALU: \(\sim 300\)-long, \(10^6\) copies
  - **LINE** (Long Interspersed Nuclear Elements)
    \(\sim 4000\)-long, \(200,000\) copies
  - **LTR retroposons** (Long Terminal Repeats (~700 bp at each end)
    cousins of HIV
- **Gene Families**
  genes duplicate & then diverge (paralogs)
- **Recent duplications** \(\sim 100,000\)-long, very similar copies

Triazzle: A Fun Example

The puzzle looks simple

**BUT** there are repeats!!!

The repeats make it very difficult.

Try it – only $7.99 at www.triazzle.com
Sequencing and Fragment Assembly

3x10^9 nucleotides

50% of human DNA is composed of repeats

Glued together two distant regions

Double-barreled sequencing:
(1990)

Both leftmost & rightmost ends are sequenced, reads are paired

~500 bp ~500 bp
Sequencing and Fragment Assembly

3x10^8 nucleotides

Strategies for whole-genome sequencing

1. Hierarchical – Clone-by-clone
   i. Break genome into many long pieces
   ii. Map each long piece onto the genome
   iii. Sequence each piece with shotgun

   Example: Yeast, Worm, Human, Rat

2. Online version of (1) – Walking
   i. Break genome into many long pieces
   ii. Start sequencing each piece with shotgun
   iii. Construct map as you go

   Example: Rice genome

3. Whole genome shotgun

   One large shotgun pass on the whole genome

   Example: Drosophila, Human (Celera), Neurospora, Mouse, Rat, Dog

Until late 1990s the shotgun fragment assembly of human genome was viewed as intractable problem.
Whole Genome Shotgun Sequencing

- genome
- cut many times at random (shotgun)
- plasmids (2 – 10 Kbp)
- fosmids (40 Kbp)
- known dist
- forward-reverse paired reads (mate pair)
- ~500 bp

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
..ACGATTACAATAGGTT..
What can we do about repeats?

Two main approaches:

• Cluster the reads

• Link the reads

Sequencing and Fragment Assembly

3x10^9 nucleotides

ARB, CRD

or

ARD, CRB?
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

Overlap

- Find the best match between the suffix of one read and the prefix of another

- Due to sequencing errors, need to use dynamic programming to find the optimal overlap alignment $O(N^2 L^2)$ for $N$ reads of length $L$

- Apply a filtration method to filter out pairs of fragments that do not share a significantly long common substring
Overlapping Reads

- Sort all $k$-mers in reads  \((k \sim 24)\)
- Find pairs of reads sharing a $k$-mer via hashing
- Extend to full alignment – throw away if not $>95\%$ similar

Overlapping Reads and Repeats

- A $k$-mer that appears $M$ times, initiate $M^2$ comparisons
- For an $Alu$ that appears $10^6$ times $\rightarrow 10^{12}$ comparisons – too much
- **Solution:**
  Discard all $k$-mers that appear more than $t \times \text{Coverage}, (t \sim 10)$
Finding Overlapping Reads

Create local multiple alignments from the overlapping reads

TAGGTTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA
TAGATTACGCAAGATTACTGA

Sources

• Genomes sequenced: http://www.genome.gov/10002154
• Serafim Batzoglou http://ai.stanford.edu/~serafim/CS262_2006/ (Sequencing slides)