DNA sequencing

How we obtain the sequence of nucleotides of a species

...ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTTAGATACCTGAC
TGATTTTTTTTTTTTTTTTATTT...
DNA Sequencing

GENERAL CONCEPTS AND CAPILLARY (SANGER) 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
DNA Sequencing: History

**Sanger method** (1977):
labeled ddNTPs terminate DNA copying at random points.

**Gilbert method** (1977):
chemical method to cleave DNA at specific points (G, G+A, T+C, C).

Both methods generate labeled fragments of varying lengths that are further electrophoresed.
## History of DNA Sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1870</td>
<td>Miescher:</td>
<td>Discovers DNA</td>
</tr>
<tr>
<td>1940</td>
<td>Avery:</td>
<td>Proposes DNA as ‘Genetic Material’</td>
</tr>
<tr>
<td>1953</td>
<td>Watson &amp; Crick:</td>
<td>Double Helix Structure of DNA</td>
</tr>
<tr>
<td>1953</td>
<td>Holley:</td>
<td>Sequences Yeast tRNA&lt;sub&gt;Ala&lt;/sub&gt;</td>
</tr>
<tr>
<td>1965</td>
<td>Wu:</td>
<td>Sequences λ Cohesive End DNA</td>
</tr>
<tr>
<td>1970</td>
<td>Sanger:</td>
<td>Dideoxy Chain Termination</td>
</tr>
<tr>
<td>1977</td>
<td>Gilbert:</td>
<td>Chemical Degradation</td>
</tr>
<tr>
<td>1977</td>
<td>Messing:</td>
<td>M13 Cloning</td>
</tr>
<tr>
<td>1980</td>
<td>Hood et al.:</td>
<td>Partial Automation</td>
</tr>
<tr>
<td>1986</td>
<td>• Cycle Sequencing</td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>• Improved Sequencing Enzymes</td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>• Improved Fluorescent Detection Schemes</td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>• Next Generation Sequencing</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>• Improved enzymes and chemistry</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>• New image processing</td>
<td></td>
</tr>
</tbody>
</table>

*Adapted from Eric Green, NIH; Adapted from Messing & Llaca, *PNAS* (1998)*
Sequencing by Hybridization (SBH): History

• **1988**: SBH suggested as an alternative sequencing method.

• **1991**: Light directed polymer synthesis developed by Steve Fodor and colleagues.

• **1994**: Affymetrix develops first 64-kb DNA microarray

  - First microarray prototype (**1989**)

  - First commercial DNA microarray prototype w/16,000 features (**1994**)

  - 500,000 features per chip (**2002**)

How SBH Works

- Attach all possible DNA probes of length $l$ to a flat surface, each probe at a distinct and known location. This set of probes is called the DNA array.
- Apply a solution containing fluorescently labeled DNA fragment to the array.
- The DNA fragment hybridizes with those probes that are complementary to substrings of length $l$ of the fragment.
How SBH Works (cont’d)

- Using a spectroscopic detector, determine which probes hybridize to the DNA fragment to obtain the \( l \)-mer composition of the target DNA fragment.

- Apply the combinatorial algorithm (below) to reconstruct the sequence of the target DNA fragment from the \( l \) – mer composition.
Hybridization on DNA Array

Universal DNA Array

DNA target TATCCGTTT (complement of ATAGGCAAAA) hybridizes to the array of all 4-mers:

ATAGGCAAAA
ATAG
TAGG
GGCA
GCAA
CAA
\textbf{\textit{l}-mer composition}

- \textit{Spectrum (s, l)} - unordered multiset of all possible \((n - l + 1)\) \textit{l}-mers in a string \textit{s} of length \textit{n}.
- The order of individual elements in \textit{Spectrum (s, l)} does not matter.
- For \(s = \text{TATGGTGC}\) all of the following are equivalent representations of \textit{Spectrum (s, 3)}:
  \begin{align*}
  \{\text{TAT, ATG, TGG, GGT, GTG, TGC}\} \\
  \{\text{ATG, GGT, GTG, TAT, TGC, TGG}\} \\
  \{\text{TGG, TGC, TAT, GTG, GGT, ATG}\}
  \end{align*}
Different sequences – the same spectrum

- Different sequences may have the same spectrum:
  
  \[
  \text{Spectrum(GTATCT,2)} = \text{Spectrum(GTCTAT,2)} = \{\text{AT, CT, GT, TA, TC}\} 
  \]
The SBH Problem

- **Goal**: Reconstruct a string from its \( l \)-mer composition

- **Input**: A set \( S \), representing all \( l \)-mers from an (unknown) string \( s \)

- **Output**: String \( s \) such that \( \text{Spectrum} \ (s, l) = S \)
1-mer composition

- **Spectrum \((s, l)\)** - unordered multiset of all possible \((n – l + 1)\) \(l\)-mers in a string \(s\) of length \(n\)
- The order of individual elements in **Spectrum \((s, l)\)** does not matter
- For \(s = \text{TATGGTGC}\) all of the following are equivalent representations of **Spectrum \((s, 3)\)**:
  - \(\{\text{TAT, ATG, TGG, GGT, GTG, TGC}\}\)
  - \(\{\text{ATG, GGT, GTG, TAT, TGC, TGG}\}\)
  - \(\{\text{TGG, TGC, TAT, GTG, GGT, ATG}\}\)
SBH: Hamiltonian Path Approach

\[ S = \{ \text{ATG, AGG, TGC, TCC, GTC, GGT, GCA, CAG} \} \]

Path visited every VERTEX once
SBH: Hamiltonian Path Approach

A more complicated graph:

\[ S = \{ \text{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT} \} \]
SBH: Hamiltonian Path Approach

\[ S = \{ \text{ATG, TGG, TGC, GTG, GGC, GCA, GCG, CGT} \} \]

Path 1:

\[ \text{ATGCGTGCA} \]

Path 2:

\[ \text{ATGGCGTGCA} \]
SBH: Eulerian Path Approach

\[ S = \{ \text{ATG, TGC, GTG, GGC, GCA, GCG, CGT} \} \]

Vertices correspond to \((l-1)\)-mers: \{AT, TG, GC, GG, GT, CA, CG\}

Edges correspond to \(l\)-mers from \(S\)

Path visited every EDGE once
SBH: Eulerian Path Approach

$S = \{ \text{AT, TG, GC, GG, GT, CA, CG} \}$ corresponds to two different paths:

$\text{ATGGCGTGCA}$

$\text{ATGCGTGGCA}$
Some Difficulties with SBH

- **Fidelity of Hybridization:** difficult to detect differences between probes hybridized with perfect matches and 1 or 2 mismatches.

- **Array Size:** Effect of low fidelity can be decreased with longer \( l \)-mers, but array size increases exponentially in \( l \). Array size is limited with current technology.

- **Practicality:** SBH is still impractical.

- **Practicality again:** Although SBH is still impractical, it spearheaded expression analysis and SNP analysis techniques.
1. Start at primer (restriction site)

2. Grow DNA chain

3. Include dideoxynucleotide (modified a, c, g, t)

4. Stops reaction at all possible points

5. Separate products with length, using gel electrophoresis
Capillary (Sanger) sequencing

Capillary sequencing (Sanger): Can only sequence ~1000 letters at a time
Electrophoresis diagrams
Challenging to Read Answer
Reading an electropherogram

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

PHRED – PHil’s Revised EDitor (by Phil Green)
   Based on dynamic programming

Several better methods exist, but labs are reluctant to change
Output of PHRED: a read

A read: ~1000 nucleotides

ACGAATCAG...A
16 18 21 23 25 15 28 30 32 ...21

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

“FASTQ format”: ASCII character that corresponds to $q+33$ (or 64)

(I = 73; 73-33 = 40 = q; q40 $>$ 0.01% error)

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

Double-barreled (paired-end, matepair) sequencing:
Both leftmost & rightmost ends are sequenced
Traditional DNA Sequencing

DNA

Shear

DNA fragments

Vector
Circular genome (bacterium, plasmid)

+ =

Known location (restriction site)
Double-barreled sequencing

Get two reads from each segment (paired-end)

~1000 bp  ~1000 bp

cut many times at random (Shotgun)

genomic segment

~1000 bp
Need to cover region with >7-fold redundancy (7X) if you use Sanger technology
Overlap reads and extend to reconstruct the original genomic region

Reconstructing The Sequence
Definition of Coverage

Length of genomic segment:  \( L \)
Number of reads:  \( n \)
Length of each read:  \( l \)

**Definition:**  Coverage  \( C = \frac{n \cdot l}{L} \)

How much coverage is enough?

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

- Sequencing errors
  \(~0.1\% of bases are wrong\)

- Repeats

- Computation: \(~O(N^2)\) where \(N = \#\) reads
Sanger sequencing

- **Advantages**
  - Longest read lengths possible today (>1000 bp)
  - Highest sequence accuracy (error < 0.1%)
  - Clone libraries can be used in further processing

- **Disadvantages**
  - The most expensive technology
    - $1500 per Mb
  - Building and storing clone libraries is hard & time consuming
NEXT GENERATION SEQUENCING
WGS revisited

Test genome

Random shearing and Size-selection

Paired-end sequencing

Read mapping

Reference Genome (HGP)

Maps to Forward strand

Maps to Reverse strand
WGS revisited

Test genome

Random shearing and Size-selection

Paired-end sequencing

Read mapping

Reference Genome (HGP)

Maps to Forward strand
Maps to Reverse strand
NGS Technologies

- 454 Life Sciences: the first, acquired by Roche
  - *Pyrosequencing*
- Illumina (Solexa): current market leader
  - GAIIx, HiSeq2000, MiSeq, HiSeq2500
  - *Sequencing by synthesis*
- Applied Biosystems:
  - SOLiD: “color-space reads”
Features of NGS data

• Short sequence reads
  – ~500 bp: 454 (Roche)
  – 35 – 150 bp Solexa(Illumina), SOLiD(AB)

• Huge amount of sequence per run
  – Gigabases per run (600 Gbp for Illumina/HiSeq2000)

• Huge number of reads per run
  • Up to billions

• Bias against high and low GC content (most platforms)
  • GC% = (G + C) / (G + C + A + T)

• Higher error (compared with Sanger)
  – Different error profiles
Next Gen: Raw Data

- Machine Readouts are different

- Read length, accuracy, and error profiles are variable.

- All parameters change rapidly as machine hardware, chemistry, optics, and noise filtering improves
Current and future application areas

Genome re-sequencing: somatic mutation detection, organismal SNP discovery, mutational profiling, structural variation discovery

De novo genome sequencing

Sequencing is becoming an alternative to microarrays for:
- DNA-protein interaction analysis (CHiP-Seq)
- novel transcript discovery
- quantification of gene expression
- epigenetic analysis (methylation profiling)
Fundamental informatics challenges

1. Interpreting machine readouts – base calling, base error estimation

2. Data visualization

3. Data storage & management
Gzip compressed raw data for one human genome > 100 GB
Informatics challenges (cont’d)

4. SNP, indel, and structural variation discovery

5. De novo Assembly
### What can we use them for?

<table>
<thead>
<tr>
<th></th>
<th>SANGER</th>
<th>454</th>
<th>Solexa</th>
<th>AB SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>De novo assembly</strong></td>
<td>Fragmented</td>
<td>Fragmented</td>
<td>Heavily Fragmented</td>
<td>Heavily Fragmented</td>
</tr>
<tr>
<td>SNP Discovery</td>
<td>Yes</td>
<td>Yes</td>
<td>&gt;95% of human</td>
<td>&gt;95% of human</td>
</tr>
<tr>
<td>Larger events</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Transcript profiling (rare)</td>
<td>No</td>
<td>Maybe</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Week 3, Lectures 2-3

CURRENT PLATFORMS & DATA COMPRESSION