Reading DNA Sequences:
18-th Century Mathematics for 21-st Century Technology

Michael Waterman
University of Southern California
Tsinghua University
DNA

• Genetic information of an organism

• Double helix, complementary base pairs

• A pairs T; G pairs C

• *E. coli*, a bacterium, has 5 million base pairs

• Humans have 3 (6) billion base pairs per nucleus; equal to about 1 yard of DNA molecule
Outline

I. 20-th Century DNA Sequencing History

II. DNA Sequence Assembly
   Shotgun Assembly

III. New Generation Sequencing Technologies
   Technologies
   Shotgun Assembly
I. DNA SEQUENCING HISTORY

- Sanger receives a 1958 Nobel Prize for sequencing insulin, a protein

- Sanger and Gilbert receive the 1980 Nobel prize for DNA sequencing methods
...history repeats itself: sequencing insulin

Fred Sanger
1958 (!) Nobel prize for sequencing insulin by Edman degradation

Average read length = 5 aa!
History of DNA Sequencing

- 1870: Miescher: Discovers DNA
- 1940: Avery: Proposes DNA as ‘Genetic Material’
- 1953: Watson & Crick: Double Helix Structure of DNA
- 1965: Holley: Sequences Yeast tRNA^{Ala}
- 1970: Wu: Sequences λ Cohesive End DNA
- 1977: Sanger: Dideoxy Chain Termination
- 1977: Gilbert: Chemical Degradation
- 1980: Messing: M13 Cloning
- 1990: Cycle Sequencing
- Improved Sequencing Enzymes
- Improved Fluorescent Detection Schemes

Efficiency (bp/person/year)
- 1
- 15
- 150
- 1,500
- 25,000
- 50,000
- 200,000
• Clone and amplify the target DNA to obtain a large number of identical molecules

• Attach primer to one end of single stranded DNA

• Polymerase extends from labeled primer

• Four separate reactions, for each letter A,T,G,C

• Extension proceeds as normal until a chain terminating dideoxynucleotide is incorporated
Chain-terminating (dideoxy) nucleotides

DNA polymerase I
+ 4 dNTPs
+ ddATP

Labeled primer

3' DNA strand

5' DNA strand
Gearing Up

- Automated sequencing machines

- Label became attached fluorescent dyes, one color for each terminating base

- Reaction could be run in the same chamber and only use one column of the gel

- Detection of bases became image processing

- Caltech origins: Lee Hood, Lloyd Smith, Hunkapilar brothers
# GeneBank Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Bp</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>680338</td>
<td>606</td>
</tr>
<tr>
<td>1983</td>
<td>2274029</td>
<td>2427</td>
</tr>
<tr>
<td>1984</td>
<td>3368765</td>
<td>4175</td>
</tr>
<tr>
<td>1985</td>
<td>5204420</td>
<td>5700</td>
</tr>
<tr>
<td>1986</td>
<td>9615371</td>
<td>9978</td>
</tr>
<tr>
<td>1987</td>
<td>15514776</td>
<td>14584</td>
</tr>
<tr>
<td>1988</td>
<td>23800000</td>
<td>20579</td>
</tr>
<tr>
<td>1989</td>
<td>34762585</td>
<td>28791</td>
</tr>
<tr>
<td>1990</td>
<td>49179285</td>
<td>39533</td>
</tr>
<tr>
<td>1991</td>
<td>71947426</td>
<td>55627</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Bp</th>
<th>Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>101008486</td>
<td>78608</td>
</tr>
<tr>
<td>1993</td>
<td>157152442</td>
<td>143492</td>
</tr>
<tr>
<td>1994</td>
<td>217102462</td>
<td>215273</td>
</tr>
<tr>
<td>1995</td>
<td>384939485</td>
<td>555694</td>
</tr>
<tr>
<td>1996</td>
<td>651972984</td>
<td>1021211</td>
</tr>
<tr>
<td>1997</td>
<td>1160300687</td>
<td>1765847</td>
</tr>
<tr>
<td>1998</td>
<td>2008761784</td>
<td>2837897</td>
</tr>
<tr>
<td>1999</td>
<td>3841163011</td>
<td>4864570</td>
</tr>
<tr>
<td>2000</td>
<td>11101066288</td>
<td>10106023</td>
</tr>
<tr>
<td>2001</td>
<td>15849921438</td>
<td>14976310</td>
</tr>
</tbody>
</table>
DNA Sequencing History

Shotgun sequencing

1995
TIGR (The Institute of Genomic Research)
The first complete DNA sequence of the genome of a free living organism --- the bacterium *Haemophilus influenzae* (1.8Mbp).

1996
International Consortium
The first complete DNA sequence of the genome of a eukaryote --- the yeast *Saccharomyces cerevisiae* (12Mbp).

1998
International Consortium
The first complete DNA sequence of the genome of a multicellular organism --- the roundworm *Caenorhabditis elegans* (97Mbp).

1999
Celera Genomics
The entire genomic sequence of the fruitfly *Drosophila melanogaster* (137Mbp).

2000
International Consortium & Celera
The first draft of the sequence of the entire human genome is published.
II. DNA SEQUENCE ASSEMBLY

When rapid DNA sequencing technology came from Sanger’s laboratory at Cambridge in the MRC in 1976, Sanger also recruited Roger Staden who created the first assembly program.
In *shotgun sequencing*, whole genomes are sequenced by making clones, breaking them into small pieces, and trying to put the pieces together again based on overlaps.

Note that the fragments are *randomly* sampled, and thus no positional information is available.
Whole Genome Assembly: problem description

• The goal is to reconstruct an unknown source sequence (the genome) on \{A, C, G, T\} given many random short segments from the sequence, the shotgun reads.

• A read is a sequence of nucleotides of length 30-800, taken from a random place in the genome.

• Reads contain two kinds of errors: base substitutions and indels. Base substitutions occur with a frequency from 0.5 – 2%. Indels occur roughly 10 times less frequently.

• Strand orientation is unknown.
Assembly is challenging

The most natural notion of assembly is to order the fragments so as to form the shortest string containing all of them.

```
ABRAC
ACADA
ADABR
DABRA
RACAD

ABRACADABRA

ABRAC
RACAD
ACADA
ADABR
DABRA
```

However, the problem of finding the shortest common superstring of a set of strings is NP-complete.
Shortest Superstring

- Merge two strings with largest overlap, continue (greedy algorithm)
- A nice computer science problem
- Conjecture: GREEDY IS AT WORST 2 TIMES OPTIMAL
- The best known constant is 2.5
- Analysis using word periodicity--not simple
Even worse...

- We must deal with significant errors in the sequence reads.
- The orientation of each read is unknown.
- Genomes have many repeats (approximate copies of the same sequence), which are very hard to identify and reconstruct.
- Gaps due to low coverage
- The size of the problem is very large. Celera’s Human Genome sequencing project contained roughly 26.4 million reads, each about 550 bases long.
• **Repeats**: A major problem for fragment assembly
• > 50% of human genome is repeats:
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)
Overlap-Layout-Consensus

Assembler programs:

ARACHNE, PHRAP, CAP, TIGR, CELERA

Common Approach:
Overlap-Layout-Consensus

Assembler programs: ARACHNE, PHRAP, CAP, TIGR, CELERA

Common Approach:
Overlap: find potentially overlapping reads
Overlap-Layout-Consensus

Assembler programs: ARACHNE, PHRAP, CAP, TIGR, CELERA

Common Approach:

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs
Overlap-Layout-Consensus

Assembler programs: ARACHNE, PHRAP, CAP, TIGR, CELERA

Common Approach:

Overlap: find potentially overlapping reads

Layout: merge reads into contigs and contigs into supercontigs

Consensus: derive the DNA sequence and correct read errors
Overlap computation in more detail

Overlaps include mismatched letters as well as missing ones (indels).

Each pair of reads has the possibility they are “read” the same direction or not.

Careful comparisons take computing time proportional to \((\text{length}(a) \times \text{length}(b))\)
Human Genome example

- The Celera project
- $26.5 \times 10^6$ reads of length 550
- Pairs of reads $= 7 \times 10^{14}$
- One pair takes $3 \times 10^5$ units
- Total resources $= 2 \times 10^{20}$
SUMMARY

• Computer assembly of genomes is challenging
• The computer programs used in the HGP were sophisticated upgrades of Staden’s approach
• Much work and ingenuity went into these computational projects
IV. NEW

• Novel 21-st Century technologies, highly parallel

• The era of $1000$ genomes is coming!
New Generation Sequencing

Cost of Sequencing

Capacity of One DNA Sequencer

<table>
<thead>
<tr>
<th>Company</th>
<th>Format</th>
<th>Read Length (bases)</th>
<th>Expected Throughput MB (million bases)/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>454 Life Sciences</td>
<td>Parallel bead array</td>
<td>100</td>
<td>96</td>
</tr>
<tr>
<td>Agencourt Bioscience</td>
<td>Sequencing by ligation</td>
<td>50</td>
<td>200</td>
</tr>
<tr>
<td>Applied Biosystems</td>
<td>Capillary electrophoresis</td>
<td>1000</td>
<td>3–4</td>
</tr>
<tr>
<td>Microchip Biotechnologies</td>
<td>Parallel bead array</td>
<td>850–1000</td>
<td>7</td>
</tr>
<tr>
<td>NimbleGen Systems</td>
<td>Map and survey microarray</td>
<td>30</td>
<td>100</td>
</tr>
<tr>
<td>Solexa</td>
<td>Parallel microchip</td>
<td>35</td>
<td>500</td>
</tr>
<tr>
<td>LI-COR</td>
<td>Electronic microchip</td>
<td>20,000</td>
<td>14,000</td>
</tr>
<tr>
<td>Network Biosystems</td>
<td>Biochip</td>
<td>800+</td>
<td>5</td>
</tr>
<tr>
<td>VisiGen Biotechnologies</td>
<td>Single molecule array</td>
<td>NA</td>
<td>1000</td>
</tr>
</tbody>
</table>
Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies1, Michael Egholm1,*, William E. Altman1, Said Attiya1, Joel S. Bader1, Lisa A. Bemben1, Jan Berka1, Michael S. Braverman1, Yi-Ju Chen1, Zhoutao Chen1, Scott B. Dewell1, Lei Du1, Joseph M. Fierro1, Xavier V. Gomes1, Brian C. Godwin1, Wen He1, Scott Helgesen1, Chun He Ho1, Gerard P. Iryuki1, Szilveszter C. Jando1, Maria L. I. Alenguer1, Thomas P. Jarvie1, Kishama S. Jirage1, Jong-Bum Kim1, James R. Knight1, Janna R. Lanza1, John H. Leamon1, Steven M. Lefkowitz1, Ming Lei1, Jing Li1, Kenton L. Lohman1, Hong Lu1, Vinod B. Makhijani1, Keith E. Mcdaide1, Michael P. McKenna1, Eugene W. Myers2, Elizabeth Nickerson2, John R. Nobile1, Ramona Plant1, Bernard P. Puc1, Michael T. Ronan1, George T. Roth1, Gary J. Sarks1, Jan Fredrik Simons1, John W. Simpson1, Maithreyan Srinivasan1, Karrie R. Tartaro1, Alexander Tomasz1, Kari A. Vogt1, Greg A. Volkmer1, Shally H. Wang1, Yong Wang1, Michael P. Weiner1, Pengguang Yu1, Richard F. Begley1 & Jonathan M. Rothberg1

- > 20 million bases
- 100 bp reads
- 200,000+ clonal reads
- Single 5-hour run
PicoTiterPlates™

• Multiple optical fibers are fused to form an optical array

• Selective removal of core material leaves wells that serve as ‘test tubes’

• Reactions occurring in the ‘test tubes’ can be monitored optically, through the remaining fiber

• Well diameter: 44μ

• Current plate contains 1.6M wells
Process Overview

1) Prepare Adapter Ligated ssDNA Library

2) Clonal Amplification on 28 μ beads

3) Load beads and enzymes in PicoTiter Plate™

4) Perform Sequencing by synthesis on the 454 Instrument
454 Technology - Sequencing Instrument
Typical Run Results

- 30MB per run
- \(~300,000\) reads
- \(~100\) bp per read
- 1-2% avg. error
- \(~50\%\) error-free

*10+ includes partial and unmapped reads
DNA Sequencing with Solexa® Technology

Generating one billion bases of high quality DNA sequence per run at less than 1% of the cost of capillary-based methods, the Illumina Genome Analyzer is designed to enable researchers to dramatically improve the efficiency and speed of current applications. Now an expanded scale of research that was previously unimaginable with other technology platforms is possible with the Genome Analyzer.
**FIGURE 2: SEQUENCING TECHNOLOGY OVERVIEW**

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.

4. **FRAGMENTS BECOME DOUBLE STRANDED**
   - The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. **DENATURE THE DOUBLE-STRANDED MOLECULES**
   - Denaturation leaves single-stranded templates anchored to the substrate.

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

10. IMAGE SECOND CHEMISTRY CYCLE
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

12. ALIGN DATA
Align data, compare to a reference, and identify sequence differences.
THE SEQUENCING RACE

The increasingly crowded market for genome-sequencing machines includes new entrants looking to push the boundaries in both speed and accuracy.

- **Complete Genomics** ($0.002)
  - Projected for June 2009
- **Helicos** ($0.0005)
- **Applied Biosystems** ($0.002)
- **Illumina** ($0.002)
- **Pacific Biosciences** (N/A)
  - Projected for June 2010
- **454/Roche** ($0.05)
- **ABI capillary sequencer** ($1)

Estimated costs are given per thousand base pairs.
Overlap--Layout-Consensus

- Even the overlap step requires an impossible number of pairwise comparisons

For example, \((10^9) \times (10^9) = 10^{18}\), for a single Illumina machine in a single day
Overlap-Layout-Consensus with short reads

- Finding consensus is NP hard
- Number of short reads is large, and tiling of each overlap is very small => tremendous runtime.

1. Replace all reads with a vertex.
2. Create the directed edge \((u, w)\) if the end of read \(u\) overlaps with the beginning of read \(w\).
3. Find the shortest path that visits every vertex once.

This corresponds to the Hamiltonian Path Problem (NP). There is no known efficient solution.
IV. EULER’S GRAPHS

We begin with Euler’s original 1736 insight.
The Bridges of Konigsberg
The Konigsberg Bridges Graph
The Konigsberg Bridges
Bridges are joined by edges if they can be reached by walking on land.
If the nodes are bridges, it is a Hamilton tour problem to visit each vertex exactly once and is NP hard.

Instead Euler in 1736 turned the problem inside out!

He made bridges into edges and his problem is to visit all edges (once and only once) returning to the starting location.
• An Eulerian path is one which visits each edge once and only once.

• An Eulerian circuit is an Eulerian path which starts and ends at the same vertex. This gets Euler back home.
The Konigsberg Bridges
The 7 bridges are the edges, the land bodies are the vertices.
• There is at least one Eulerian circuit if and only if the graph is connected and, for every vertex v, the degree is even (and positive).

• For directed graphs, in(v) = out(v).

• There is beautiful combinatorics to give the number of Eulerian paths (and for uniqueness).
Every vertex is of odd degree—no Eulerian circuits.
Algorithms to produce an Eulerian circuit are quite elementary:

- Start anywhere

- Follow any edge, marking edges as you go

- If you return to a vertex with no remaining unmarked edges, just go to any unexplored edge. If none exist you are finished.
IVb. De Bruijn Sequences

Our alphabet will be \{A, G, C, T\}

A De Bruijn sequence is a cyclic sequence where all \(4^k\) \(k\)-words appear exactly once.

They exist in the shortest possible length sequence \textit{a la} Euler as follows.
• Verticies are all sequences of length $k-1$

• Directed edges exist where the $k-2$ suffix of one vertex is the $k-2$ prefix on the other vertex

• Edges correspond to $k$-words

• Every vertex has 4 incoming edges and 4 outgoing edges, $\text{in}(v)=4=\text{out}(v)$
An Example

\[ \text{ACCTGA} \rightarrow \text{CCTGAT} \]

- $k=7$, $k-1=6$, and $k-2=5$
- The $k$-word or edge is ACCTGAT
- Many Eulerian circuits exist
De Bruijn sequences go back to Sainte-Marie 1894

- Each vertex is a \((k-1)\)-word
- Each vertex has 4 edges entering and 4 leaving
- The number of DeBruijn sequences is

\[ 4! \frac{4^{(k-1)}}{4^k} \]

(for \(k=3\), this is \(2*10^{20}\))
Sequence Graphs

For a specific sequence build the graph as above.
Example: DNA = ATGTGCGCCGCA, \( k=3 \)

Determining the (or a) sequence from the graph is equivalent to finding an Eulerian path.
Sequence Graphs

For a specific sequence build the graph as above. Example: DNA = ATGTGCGGCA, $k=3$

\[
\begin{array}{c}
AT \\
\rightarrow \\
TG
\end{array}
\]

Determining the (or a) sequence from the graph is equivalent to finding an Eulerian path.
Sequence Graphs

For a specific sequence build the graph as above. Example: DNA = ATGTGCCGCA, $k=3$

Determining the (or a) sequence from the graph is equivalent to finding an Eulerian path.
Sequence Graphs

For a specific sequence build the graph as above. Example: DNA = ATGTGCGCGCA, $k=3$

Determining the (or a) sequence from the graph is equivalent to finding an Eulerian path.
ATG
ATG!C
ATG!C
ATG!C!A
ATG!CCA
ATG! CC-GA
ATG!CCGCA
AT → TG → GC → CA

ATGTCCGCA
ATGTGCGGA
ATGTCGCA
IVc. SEQUENCING BY HYBRIDIZATION

• This method was proposed in 1988 and 1989.

• While never achieving its initial goal, the arrays which were created are in widespread usage today.
Sequencing By Hybridization (SBH)

- DNA array -- all possible oligonucleotides of length $k$ ($k$-tuples)
- target sequence -- labeled single stranded DNA fragment
- hybridization -- target DNA fragment hybridizes with oligonucleotides complementary to $k$-tuples of the target fragment
- sequence reconstruction -- reconstruct the sequence of the target DNA from its spectrum ($k$-tuple content)
Example: DNA=ATGTGCCGCA, $k=3$
Hamiltonian path approach

Lysov et al. (1988), Drmanac et al. (1989):

Transform spectrum S into a graph H
Vertices in H -- $k$-tuples in the spectrum
S Edges in H -- overlapping $k$-tuples.
Sequence reconstruction (Hamiltonian)

Represent each $k$-tuple by a vertex and draw a directed edge from vertex $a_1 \ldots a_l$ to vertex $b_1 \ldots b_l$ iff $a_2 \ldots a_l = b_1 \ldots b_{l-1}$.

Example: DNA = ATGTGCCGCA, $l=3$

Determining the sequence is equivalent to finding a Hamiltonian tour (NP-hard).
Sequence reconstruction
(Eulerian path)

Represent $l$-tuples by edges and $(l-1)$-tuples by vertices such that an edge $a_1...a_l$ is directed from the vertex $a_1...a_{l-1}$ to the vertex $a_2...a_l$.

Example: DNA = ATGTGCGCGCA, $l=3$

Determining a sequence consistent with the data is equivalent to finding an Eulerian circuit. (Pevzner)
IVd. EULERIAN ASSEMBLY

• Take the reads
• Break up reads into overlapping $k$-words
  (k=25, say)
• Merge identical $k$-words
• Find Eulerian paths
Eulerian path approach

Sequence reconstruction is the search for Eulerian paths in graphs from the $k$-word data.

Idury and Waterman (1995)

Edge in graph $G$ is a $k$-word in one or more reads
Euler Approach to Assembly

Vertices: \((k-1)\)-words in each read

Edges: \(k\)-words in each read
Euler approach: advantages

- Linear time
- *No* pairwise alignment
- *No* layout
- *No* consensus or multiple alignment
- One letter indels are no harder to handle than mismatches
• Orientation: the usual difficulty of deciding a consistent orientation for each fragment is handled by putting in the reverse complement for each fragment.

• Computation time is double that if one knew exactly the orientation of each read.
  (A big advantage over an exponential # of possibilities!)
Euler approach: difficulties

• Erroneous edges

• Tangled graph

• Storage requirements (huge)
Erroneous edges

error

Erroneous edges
erroneous l-tuples in the sequencing read

sequencing read

erroneous l-tuples in the complementary sequencing read

error
Detecting Chimerical Reads

Multiplicity

$k$-word position
How can we simplify the tangled Eulerian graph?
Given an Eulerian graph and a collection of paths in this graph, find an Eulerian path in this graph that contains all these paths as subpaths.
Solving Eulerian Superpath Problem: Equivalent Transformations

- Simplify the system of path with the goal of transforming the Eulerian Superpath Problem into the Eulerian Path Problem.

- Equivalent transformation of the repeat graph: there exists a one-to-one correspondence between Eulerian superpaths before and after the transformation.

- Make a series of equivalent transformations

\[
(G, P) \rightarrow (G_1, P_1) \rightarrow \ldots \rightarrow (G_n, P_n)
\]

that lead to a system of paths with every path being a single edge in the repeat graph.
Repeats

• **Repeats**: A major problem for fragment assembly

• > 50% of human genome is repeats:
  - over 1 million *Alu* repeats (about 300 bp)
  - about 200,000 LINE repeats (1000 bp and longer)
Overview of Eulerian assembly

1. The *repeat graph* of a genome is constructed from reads, where each vertex is a repeat.

2. The repeat graph is *transformed* so that an Eulerian path may be found (the genome is assembled), or a simplified repeat structure is determined (with edges as contigs in the assembly).
TECHNOLOGY CONTINUES TO EVOLVE

• Every new technology can make extinct a set of computational methods.

• Every new technology is likely to create a new suite of problems.
Collaborators:

Ramana Idury (1995)

Pavel Pevzner and Haixu Tang (2001)

Xiaoman Li (2003)

THANKS FOR LISTENING!