Lectures 18, 19: Sequence Assembly

Fall 2018
Nov 13, 15, 2018
Outline

• Introduction

• Sequence Assembly Problem

• Different Solutions:
  ◦ Overlap-Layout-Consensus Assembly Algorithms
  ◦ De Bruijn Graph Based Assembly Algorithms

• Resolving Repeats

• Introduction to Single-Cell Sequencing
Whole Genome Shotgun Sequencing

- Frederick Sanger (and others) shared a Nobel Prize in Chemistry in 1980 for developing a method to sequence short regions of DNA.

- There is no current technology to simply read the whole genome sequence from one end to the other.

- The human genome is 3 billion nucleotides long. Sequencing it requires breaking it into little pieces, sequencing the pieces separately, and fitting them back together, like a jigsaw puzzle.
DNA Sequencing

- Shear DNA into millions of small fragments
- Read 500 – 700 nucleotides at a time from the small fragments (Sanger method)
Whole Genome Shotgun Sequencing

Start with many copies of genome. Bacterial genome length: ~5 million.

Fragment them and sequence reads at both ends. Read length: 35 to 1000 bp.

Find overlapping reads.

ACGTA\ldots\text{TAGAATCGACCATG}\ldots

\ldots\text{AACATA}\ldots\text{GTTGACGTAGAATC}

Merge overlapping reads into contigs.

\ldots\text{AACATA}\ldots\text{GTTGACGTAGAATCGACCATG}\ldots

Contig \quad \text{Gap} \quad \text{Contig} \quad \text{Gap} \quad \text{Contig}

Coverage at this location=2
Sequencing Coverage

Number of reads: ~28 million, read length: 100 bp, genome size: 4.6 Mbp, coverage: ~600x

Sequencing by Hybridization (SBH): History

• 1988: SBH suggested as an alternative sequencing method. Nobody believed it will ever work.

• 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 ATAGGCAAA)
hybridizes to the array of all 4-mers:

ATAGGCAAAA
ATAG
TAGG
GGCA
GCAA
CAA
/-mer composition

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

Different sequences may have the same spectrum:

\[
\begin{align*}
\text{Spectrum}(\text{GTATCT}, 2) &= \{\text{AT, CT, GT, TA, TC}\} \\
\text{Spectrum}(\text{GTCTAT}, 2) &= \{\text{AT, CT, GT, TA, TC}\}
\end{align*}
\]
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 $Spectrum\ (s,l) = S$
Some Applications of Sequencing

- **1000 Human Genomes Project**
  
  An international effort to map variability in the genome


- **Prostate Cancer Genomics**


- **Genome 10K Project**
  
  
  - An international effort to sequence, *de novo* assemble, and annotate 10,000 vertebrate genomes; 300+ species to be started in 2011.

  Genome 10K Community of Scientists, *J Heredity* (Sep 2009) 100 (6): 659-674
**De Novo Genome Assembly**

*Problem:* given a collection of reads, i.e. short subsequences of the genomic sequence in the alphabet “A, C, G, T”, completely reconstruct the genome from which the reads are derived.

*Challenges:*

- Repeats in the genome
  
  …ACCCAGTТGACTGGGGATCCTTTTTAAAGACTGGGGATTTTAACGCG…
  
  CAGTTGACTG
  
  \[ \begin{align*}
  &\text{ACTGGGGATCC} \\
  &\text{GACTGGGGATT}
  \end{align*} \]
  
  Sample reads

- Sequencing errors: substitutions, insertions, deletions, and others.
  
  TTTTTTATA\textit{GA} (substitution), CCTT—TAAACG\textit{G} (deletion and insertion)

- Size of the data, e.g. 1.5 billion reads in 110GB FASTA file.
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
Repeat Types

- **Low-Complexity DNA**  (e.g. ATATATATACATA…)

- **Microsatellite repeats**  \((a_1…a_k)^N\) where \(k \sim 3-6\)  
  (e.g. CAGCAGTAGCAGCACCAG)

- **Transposons/retrotransposons**
  - **SINE**  Short Interspersed Nuclear Elements  
    (e.g., \(Alu\): \(\sim\)300 bp long, \(10^6\) copies)
  - **LINE**  Long Interspersed Nuclear Elements  
    \(\sim\)500 - 5,000 bp long, 200,000 copies
  - **LTR retroposons**  Long Terminal Repeats \(\sim\)700 bp at each end

- **Gene Families**  genes duplicate & then diverge

- **Segmental duplications**  \(\sim\)very long, very similar copies
Triazzle: A Fun Example

The puzzle looks simple
BUT there are repeats!!!
The repeats make it very difficult.
Try it
De Novo Genome Assembly

Current solutions

- Overlap-layout-consensus (*Celera, Newbler*)
  - Suitable for low coverage, long reads
  - Highly parallelizable

- De Bruijn graph construction (*ALLPATHS-LG, ABySS, Velvet, SOAPdenovo, EULER-SR, SPAdes, and HyDA*)
  - Suitable for high coverage, short reads
  - Fast but memory-intensive
  - Sensitive to sequencing errors
  - Mathematically elegant repeat classification
Overlap-Layout-Consensus Assembly
Overlap-Layout-Consensus

Assemblers: SGA, 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..
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*

- 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
- Extend to full alignment – throw away if not $>95\%$ similar
Overlapping Reads and Repeats

- A $k$-mer that appears $N$ times, initiates $N^2$ 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

TAGATTACACAGATTACTGA
TAGATTACACAGATTACTGA
TAG ATTACACAGATTATTGA
TAGATTACACAGATTACTGA
TAGATTACACAGATTACTGA
TAGATTACACAGATTACTGA
TAG ATTACACAGATTATTGA
TAGATTACACAGATTACTGA
TAGATTACACAGATTACTGA
Finding Overlapping Reads (cont’d)

- Correct errors using multiple alignment

\[
\begin{array}{cccc}
\text{C:} & 20 & \text{C:} & 20 \\
\text{C:} & 35 & \text{C:} & 35 \\
\text{T:} & 30 & \text{C:} & 0 \\
\text{C:} & 35 & \text{C:} & 35 \\
\text{C:} & 40 & \text{C:} & 40 \\
\end{array}
\]

- Score alignments

- Accept alignments with good scores
Repeats are a major challenge.

Do two aligned fragments really overlap, or are they from two copies of a repeat?

Solution: repeat masking – hide the repeats!!!

Masking results in high rate of misassembly (up to 20%).

Misassembly means alot more work at the finishing step.
Merge Reads into Contigs

repeat region

Merge reads up to potential repeat boundaries
Repeats, Errors, and Contig Lengths

- Repeats shorter than read length are OK.
- Repeats with more base pair differences than sequencing error rate are OK.
- To make a smaller portion of the genome appear repetitive, try to:
  - Increase read length.
  - Decrease sequencing error rate.
De Bruijn Graph Based Assembly
De Bruijn Graph Example
Shred reads into k-mers (k = 3)

Read 1
G G A C T A A A
G G A
G A C
A C T
C T A
T A A
A A A

Read 2
G A C C A A A T
G A C
A C C
C C A
C A A
A A A
A A T

De Bruijn Graph Example
Merge vertices labeled by identical k-mers

Read 1:
- GGA (1x)
- GAC (1x)
- ACT (1x)
- CTA (1x)
- TAA (1x)
- AAA (1x)

Read 2:
- GAC (1x)
- ACC (1x)
- CCA (1x)
- CAA (1x)
- AAA (1x)
- AAT (1x)

Resulting Graph:
- GGA (1x)
- GAC (2x)
- ACT (1x)
- CTA (1x)
- TAA (1x)
- AAA (2x)
- AAT (1x)
- ACC (1x)
- CCA (1x)
- CAA (1x)
Another Example
Constructing the graph \((k = 4)\)

Sequencing errors are normally detected by a coverage cutoff threshold

A branching vertex is caused by either a repeat in the original sequence or a sequencing error
Example
After condensation

AAGTCGA  CGAG  CGACGC

GAGGCT  GCTCTAG  GCTTTTAG  TAGA  AGAG  GAGACAA

AGATCCGATGAG
Example
After error removal

AAGTCGA → CGAG → GAGGCT → GCTTTTAG → TAGA → AGAG → GAGACAA

AGATCCGATGAG
Example
After recondensation

AGATTCGATGAG

Any non-branching path in this graph corresponds to a contig in the original sequence.

Taking the risk of following arbitrary branching paths may create chimeric species.

Source: Serafim Batzoglou
Resolving Repeats
Using paired reads

Genome

Read 1

Read 2

Insert size: a design parameter
Resolving Repeats
Equivalent transformations

Genome: … S₁ REPEAT S₂ ……………… S₃ REPEAT S₄ …

Matches the distance in the graph,
Longer than repeat length

Resolving Repeats

Failure

**Mate pair transformation** (*Velvet, ABysS, EULER-SR*)

- Find a unique path between mates in the graph.
- *When multiple paths match the distance between mate-pairs, mate pair transformation fails.*

To resolve a repeat, insert size must be larger than the repeat length and smaller than the length of potential conjugate paths (same length paths passing through the repeat).
Single Cell Sequencing
Whole genome amplification

Start with a single copy of genome.

Amplify (copy) the genome using multiple displacement amplification (MDA) technique invented by Roger Lasken at J. Craig Venter Institute.


Fragment them and sequence reads at both ends.
Sequencing Coverage
Normal multicell vs. single cell

Green regions are blackout

Number of reads: ~28 million, read length: 100 bp, genome size: 4.6 Mbp, coverage: ~600x

A cutoff threshold will eliminate about 25% of valid data in the single cell case, whereas it eliminates noise in the normal multicell case.