Sequence alignment

Bioinformatics MTAT.03.239
22.02.2018

Priit Adler
This lecture

- Reference genome
- Genomic variation
- Sequence alignment
  - mapping reads to reference!
• How long is human DNA?
• How many “genes” do we have?
• Describe the “Central dogma of molecular biology”
Biology milestones

1950
- DNA structure reported by Watson and Crick (1953)

1960
- DNA genetic code deduced for 20+ codons and their respective amino acids and reported by M. Niremberg, H. Mathaei and S. Ochoa (1966)

1970
- First recombinant DNA molecules were created (P. Berg 1972)
- Di-deoxy sequencing reported by F. Sanger (1977)

1980
- Polymerase chain reaction reported (Mullis 1983)

1990
- Human genome project initiated in public sector (1990)
- First completed genome sequence in public sector for M. genitalium (Fraser et al. 1995)
- Human genome project initiated in private sector (Venter 1998)

2000
- Initial drafts of human genome reported simultaneously in private and public sectors (2001)

2004
- Finished drafts of euchromatic sequence of human genome reported in public sector (2004)
- Rough draft (5x coverage) chimpanzee reported (2005)

2010
- Massively parallel sequencing technologies now being widely utilized about 2007 (technology made publicly available in 2004)

http://imihumangenomproject.blogspot.com.ee/2012/12/genome-sequencing.html
Estimate the number of genes in Human genome
Genomic data

http://www.futuretimeline.net/blog/2014/01/16.htm#VfsvUZ2qPBc
Genomic data

Growth of GenBank and WGS

Analysis of sequences

- Sequence alignment
- Gene prediction
- Genome assembly
- Protein structure / domains
A reference genome (also known as a reference assembly) is a digital nucleic acid sequence database, assembled by scientists as a representative example of a species' set of genes.

https://en.wikipedia.org/wiki/Reference_genome
Genome

• Is the entirety of an organism’s hereditary information

• The genome includes both the genes and non-coding sequences of DNA/RNA

• In 1995, Haemophilus influenzae or was the first genome of a living organism to be sequenced in July 1995

  • 1,830,140 base pairs of DNA in single circular chromosome that contains 1,740 protein-coding gene, 58 transfer RNA genes and 18 other RNA genes
Whole Genomes

Drosophila
C. elegans
Rat
Human
Mouse
Rice
Mosquito
Yeast
E. coli
H. influenza
Arabidopsis
## Genome sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Completion date</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>phage phiX174</td>
<td>1978</td>
<td>5,368 bp</td>
<td>1st viral genome</td>
</tr>
<tr>
<td>human mtDNA</td>
<td>1980</td>
<td>16,571 bp</td>
<td>1st organelle genome</td>
</tr>
<tr>
<td>lambda phage</td>
<td>1982</td>
<td>48,502 bp</td>
<td>important virus model</td>
</tr>
<tr>
<td>HIV</td>
<td>1985</td>
<td>9,193 bp</td>
<td>AIDS retrovirus</td>
</tr>
<tr>
<td>H. influenzae</td>
<td>1995</td>
<td>1,830 Kb</td>
<td>1st bacterial genome</td>
</tr>
<tr>
<td>M. genitalium</td>
<td>1995</td>
<td>580 Kb</td>
<td>smallest bacterial genome</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>1996</td>
<td>12.5 Mb</td>
<td>1st eukaryotic genome</td>
</tr>
<tr>
<td>E. coli K12</td>
<td>1997</td>
<td>4.6 Mb</td>
<td>bacterial model organism</td>
</tr>
<tr>
<td>C. trachomatis</td>
<td>1998</td>
<td>1,042 Kb</td>
<td>internal parasite of eukaryotes</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>2000</td>
<td>180 Mb</td>
<td>fruit fly, model insect</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>2000</td>
<td>125 Mb</td>
<td>thale cress, model plant</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>2001</td>
<td>3,000 Mb</td>
<td>human</td>
</tr>
<tr>
<td>SARS</td>
<td>2003</td>
<td>29,751 bp</td>
<td>coronavirus</td>
</tr>
</tbody>
</table>

*Introduction to Computational Biology, Nello Christiani and Matthew W. Hahn*
“Completely” sequenced genomes
Human genome

One cell
- 23 pairs of chromosomes

DNA
- 3 billion pairs of DNA bases

RNA
- ≈21,000 to 23,000 genes

Protein
- ≈100,000 different proteins

Human body
- $10^{14}$ cells
  - (100 trillion)
Human full genome: 3234.8 Mb

Tallinn - Jõgeva - Misso: 320 km

ATGCTCGTAC = 1mm
DNA

• Protein coding genes cover only 1.5% of human genome

• Basepair variation between 2 genomes <~ 1%

• Structural variation accounts for more…

• What does the rest do?
Graph genome

https://www.sevenbridges.com/graph/
MCF7 (cancer model)
genomic rearrangement
Genomic variation

- SNPs — single short nucleotide polymorphisms
- Indels — insertions / deletions
- CNVs — copy number variations
- Genomic rearrangements
DNA sequencing

- Read length
- Single reads
- Paired end reads

Questions

• Name sources of genetic variance

• Is human genome complete?

• What is the typical sequencing read length?
Gene expression

DNA

preRNA

mRNA
DNA vs RNA sequencing

DNA seq

RNA seq

reference genome
DNA complementarity

3' - ATGC GG TAGGACGGCTAATGCCA - 5'

5' - TACGCCATCCTTGCCGATTACG GT - 3'
DNA reverse complementarity

3’ - ATGCGGTAGGACGGCTAATGCCA - 5’

is same as

TGGCATTAAGCCGTCCTACCGCAT
Alignment problem

Find best fitting matching position from reference genome to a sequence read
Alignement problem

- Exact matching
- Edit distance
- sequence alignment
Sequence alignment

dynamic programming

http://avatar.se/lectures/molbioinfo2001/dynprog/dynamic.html
Sequence alignment

dynamic programming

http://avatar.se/lectures/molbioinfo2001/dynprog/dynamic.html
Sequence alignment

dynamic programming

http://avatar.se/lectures/molbioinfo2001/dynprog/dynamic.html
Sequence alignment

dynamic programming

http://avatar.se/lectures/molbioinfo2001/dynprog/dynamic.html
Sequence alignment

dynamic programming

http://avatar.se/lectures/molbioinfo2001/dynprog/dynamic.html
Sequence alignment

Global alignment

Local alignment

Fitting alignment (global - local alignment)
Rosalind glossary

Global alignment - [http://rosalind.info/glossary/alignment/](http://rosalind.info/glossary/alignment/)

Local alignment - [http://rosalind.info/glossary/local-alignment/](http://rosalind.info/glossary/local-alignment/)

Fitting alignment (global - local alignment) - [http://rosalind.info/glossary/fitting-alignment/](http://rosalind.info/glossary/fitting-alignment/)
HTS aligners timeline
https://www.ebi.ac.uk/~nf/hts_mappers/
HISAT sequence aligner

• aligner specifically developed for RNA-seq data sequencing. Optimised for reads generated by Illumina sequencer (read length 75bp - 150bp, can handle up to 300bp)

• Successor of Bowtie and TopHat

• Uses Hierarchical Graph FM index
Burrows-Wheeler transform

BWT for sequence ‘acaacg’

Retrieve originals sequence from BWT

Given BWT ‘gc$aaac’ we can always get the first column by sorting
Retrieve originals sequence from BWT

Match string ‘aac’ using Last-to-Fist mapping

Consider string "abracadabra$"

\[
\text{LF}(i) = C[L[i]] + \text{Occ}(L[i], i)
\]

```
C[c] of "ard$rcaaaabb"

<table>
<thead>
<tr>
<th>c</th>
<th>$</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>C[c]</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

Occ(c, k) of "ard$rcaaaabb"

<table>
<thead>
<tr>
<th>a</th>
<th>r</th>
<th>d</th>
<th>$</th>
<th>r</th>
<th>c</th>
<th>a</th>
<th>a</th>
<th>a</th>
<th>a</th>
<th>b</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>r</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
```
Consider string "abracadabra$"

Match ‘bra’ using LF mapping

1. $\{C[a] + 1..C[a+1]\} = [2..6]$
2. $\{C[r] + \text{Occ}(r, \text{start}-1) + 1..C[r] + \text{Occ}(r, \text{end})\} = [10 + 0 + 1..10 + 2] = [11..12]$
3. $\{C[b] + \text{Occ}(b, \text{start}-1) + 1..C[b] + \text{Occ}(b, \text{end})\} = [6 + 0 + 1..6 + 2] = [7..8]$

$$LF(i) = C[L[i]] + \text{Occ}(L[i], i)$$
Graph BWT

Prefix-sorted graph
Last-First (LF) mapping

<table>
<thead>
<tr>
<th>Node rank</th>
<th>First letter</th>
<th>Outgoing edge(s)</th>
<th>Node rank</th>
<th>Last letter</th>
<th>Incoming edge(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td></td>
<td>1</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td></td>
<td>2</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td></td>
<td>3</td>
<td>Z</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>G</td>
<td></td>
<td>4</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>G</td>
<td></td>
<td>5</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>G</td>
<td></td>
<td>6</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>G</td>
<td></td>
<td>7</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>T</td>
<td></td>
<td>8</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>T</td>
<td></td>
<td>9</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>T</td>
<td></td>
<td>10</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Z</td>
<td></td>
<td>11</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>