Whole Genome Alignment

Adam Phillippy
University of Maryland, Fall 2012
Motivation
Breast cancer karyotypes
Goal of whole-genome alignment

- For two genomes, A and B, find a mapping from each position in A to its corresponding position in B.

Megabase-sized sequences cannot be aligned with an $O(n^2)$ algorithm like dynamic programming.
Global vs. Local alignments

- Global pairwise alignment
  
  ```
  . . . AAGCTTGGGCTTAGCTGCTAGGGCTTTGGA . . .
  . . . AAGCTGGGCTTAGTTGCTAG . . . TAGGCTTTTGG . . .
  ```

- Whole genome alignment
Alignment Visualization
Global visualization

- Gene model conservation across 3 *Plasmodium* species
Genome alignment visualization

- How can we visualize whole genome alignments?

- With an alignment dot plot
  - $N \times M$ matrix
    - Let $i =$ position in genome $A$
    - Let $j =$ position in genome $B$
    - Fill cell $(i,j)$ if $A_i$ shows similarity to $B_j$

- A perfect alignment between $A$ and $B$ would completely fill the positive diagonal
Translocation  Inversion  Insertion

http://mummer.sourceforge.net/manual/AlignmentTypes.pdf
The look similar, what about their genomes?
Drosophila shuffling

Multiple alignment visualization

Open problem for many genomes
MUMmer

Aligning two genomes in under a minute
Nucmer algorithm

1. Find exact match seeds (MUMmer Suffix Tree)
2. Cluster significant matches (Union-Find)
3. Extend and combine alignments (Smith-Waterman)
4. Filter repeats (Dynamic programming)
Suffix trees

- $O(n)$ construction
- $O(n)$ space
- $O(n+m)$ Longest common substring
- $O(n+m+k)$ Find all $k$ maximal matches
**MUMmer**

- **Maximal Unique Matcher (MUM)**
  - **match**
    - exact match of a minimum length
  - **maximal**
    - cannot be extended in either direction without a mismatch
  - **unique**
    - occurs only once in both sequences (MUM)
    - occurs only once in a single sequence (MAM)
    - occurs one or more times in either sequence (MEM)
Is it a MEM, MAM or MUM?

**MUM** : maximal unique match

**MAM** : maximal almost-unique match

**MEM** : maximal exact match

MUMs inherently avoid repetitive regions, which do not make good seeds.
Clustering

cluster length = $\sum m_i$

gap distance = C

indel difference = $|B - A|$
Banded dynamic programming

Match score 0, Edit score +1, Max edits 2

<table>
<thead>
<tr>
<th></th>
<th>^</th>
<th>T</th>
<th>T</th>
<th>G</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>^</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3*</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3*</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3*</td>
<td>2</td>
<td>2</td>
<td>3*</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3*</td>
<td>2</td>
</tr>
</tbody>
</table>

Linear time and space dynamic programming, also see Divide and Conquer
Extending

Match score +3, Edit score -7

break length = A
break point = B

score ~70%

7 matches and 3 edits over a 10 bp window = 7*3-3*7 = 0
L. monocytogenes alignment

18-mer seeds

alignments

Why isn't it a single diagonal?
Microbial Genomics
Comparative genomics

- Study genomic content and function across different taxa

- Why?
  - study evolution
  - link phenotype with genotype
  - reveal genomic organization and function
  - transfer functional annotation

- How?
  - genome sequencing and alignment

Observation and comparison yield tremendous insight.
Microbes are underappreciated

- They’re everywhere
- Harmful
  - disease, spoiling
- Beneficial
  - human microbiota
  - bio-energy, bio-remediation
  - synthetic genomics
- Easy to work with
  - rapid generation time
  - small genomes
  - extremely efficient
  - simpler models

$10^{14}$ bacterial cells vs. $10^{13}$ human cells
Listeria monocytogenes

- *Listeria monocytogenes*
  - Important foodborne pathogen (cheese, lunch meat, etc.)
  - 3 Mbp genome, 3 primary lineages (I, II, III*)
Bacteria have sex

- A few mechanisms of “horizontal gene transfer”
  - **Transformation**: the genetic alteration of a cell resulting from the introduction, uptake, and expression of foreign genetic material (DNA or RNA).
  - **Transduction**: the process in which bacterial DNA is moved from one bacterium to another by a virus (e.g. phage).
  - **Bacterial conjugation**: a process in which genetic material is transferred to another cell by cell-to-cell contact.
Pan-genomics

- **Core genome**
  - minimal gene set necessary for survival
  - defining characteristics of the species
  - orthologs, gene groups

- **Accessory genome**
  - mediate adaptation to different environments
  - e.g. stress and antibiotic resistance, nutrient metabolism

- **Pan genome**
  - union of core and accessory genes (non-redundant)
  - defines total genetic diversity of the species
How big is a pan-genome?

- How many new genes will be discovered in sequencing the $k^{th}$ genome?
  - For all $k!$ possible permutations of $k$ genomes
    - how many new genes are found in the $k^{th}$ genome?
  - Perform regression on the average values

FOR $k = 1$ to $N$
  FOR each random sample
    Randomly generate an ordered set of $k$ genomes
    Compute # unique genes in the $k^{th}$ genome
  END FOR
END FOR
L. monocytogenes pan genome

- Power law — non-linear least squares fit to means

Heap’s law. “Open” pan-genome? Consequences for antibiotic resistance?
L. monocytogenes core genome

- Exponential decay — non-linear least squares fit to means

Draft genomes missing ~5 genes per genome