SEQUENCE ALIGNMENT:
INVESTIGATING AN INFLUENZA OUTBREAK

BIO 300/CMPSC 300
Dr. Kristen Webb
Spring 2016
Descent with Modification

- DNA replication ensures a mostly faithful passing of the genome to progeny

- What would be the consequence of 100% accurate replication?
Descent with Modification

How does descent with modification happen?
Descent with Modification

How does descent with modification happen?

- **Mutation**
  - A change in a DNA sequence
  - Results from errors in replication or repair
  - Mutation is the ultimate source of genetic variation

Domestic Dog – *Canis lupus familiaris*
All descendent from the grey wolf
All the same species (sub-species)
Breeds – variation within the species
Sequence Variations

- sequences may have diverged from a common ancestor through various types of mutations:
  - substitutions (ACGA $\rightarrow$ AGGA)
  - insertions (ACGA $\rightarrow$ ACCGGAGA)
  - deletions (ACGGAGA $\rightarrow$ AGA)
Tracking Infectious Disease – 2009 H$_1$N$_1$ Influenza Pandemic
**H₁N₁ Influenza Virus**

**H₁N₁ Influenza Virus**

**Hemagglutinin** - protein the virus uses to attach to the host cells

**Neuraminidase** - enables the virus to be released from the host cell
Viral Evolution

• Viruses evolution very quickly
  • Some of the highest mutation rates known
    • Arms race with immune system
  • Viruses – mutation rate 0.0001 - 0.000001 mutations per base per generation
    • One mutation every 10,000 – 1,000,000 nucleotides
    • Influenza genome size = ~14,000 nucleotides
    • Entire genome is coding regions (genes)

• Humans – 0.00000001 mutations per base per generation
  • One mutation every 100,000,000 nucleotides
  • Human genome size 3 billion nucleotides
  • Only 1.5% of genome is coding regions (genes)
Pairwise Alignment
Similarity and Relatedness

Alignment of a gene from two closely related viruses
Hemagglutinin gene from virus A: ATGAACGCAATACTCGTAGTT...
Hemagglutinin gene from virus B: ATGAAGGCAATACTAGTAGTT...

Alignment of a gene from two distantly related viruses
Hemagglutinin gene from virus A: ATGAACGCAATACTCGTAGTT...
Hemagglutinin gene from virus C: ATGCACGAAATGCTCGGACCT...
Tracing an Infection to a Source - HIV

• 1990 – CDC report that a woman in Florida had contracted HIV from her dentist
  • Dentist diagnosed with HIV in 1986, developed AIDS in 1987
  • Patient had no other risk factors and had not been in contact with other HIV-positive persons
  • Patient had had an invasive dental procedure

• Tested dentist’s other patients - 10 tested positive for HIV
• Did they contract HIV from the dentist?
### Table 23.10

<table>
<thead>
<tr>
<th>Person</th>
<th>Sex</th>
<th>Known risk factors</th>
<th>Average differences in DNA sequences (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dentist</td>
<td>M</td>
<td>Yes</td>
<td>3.4</td>
</tr>
<tr>
<td>Patient A</td>
<td>F</td>
<td>No</td>
<td>4.4</td>
</tr>
<tr>
<td>Patient B</td>
<td>F</td>
<td>No</td>
<td>3.4</td>
</tr>
<tr>
<td>Patient C</td>
<td>M</td>
<td>No</td>
<td>4.4</td>
</tr>
<tr>
<td>Patient E</td>
<td>F</td>
<td>No</td>
<td>3.4</td>
</tr>
<tr>
<td>Patient G</td>
<td>M</td>
<td>No</td>
<td>4.9</td>
</tr>
<tr>
<td>Patient D</td>
<td>M</td>
<td>Yes</td>
<td>13.6</td>
</tr>
<tr>
<td>Patient F</td>
<td>M</td>
<td>Yes</td>
<td>10.7</td>
</tr>
</tbody>
</table>

Source: After C. Ou et al., Science 256(1992):1165-1171, Table I.
Sequence Alignment

- substitutions (ACGA → AGGA)
- insertions (ACGA → ACCGGAGA)
- deletions (ACGGAGA → AGA)
Global Pairwise Alignment

- **Dynamic Programming**
  - Divide a problem into a series of smaller subproblems
  - Solve each subproblem
  - Use the solutions to build the solution to the original problem

- **Needleman-Wunsch Algorithm**
  - Creates a matrix of partial alignment scores
  - Backtracks along a path to the best possible alignment
Needleman-Wunsch Algorithm

- Create $N \times M$ matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1$^{st}$ row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + match/mismatch score
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.
**Example**

Alignment score = 0

Let:
- Match = +1
- Mismatch = 0
- Gap = -1

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C</strong></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></td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The alignment score is 0, indicating no matches or mismatches.

[Red box highlighting the bottom right cell with a score of 0]
**Example**

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</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>G</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**CACGTAT**

**---CGCA---**
Example

Let:
Match = +1
Mismatch = 0
Gap = -1

Alignment score = 0

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</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>G</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>C</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>A</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-1</td>
</tr>
</tbody>
</table>
Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

CACGTAT
CGC--A--
Example

Alignment score = 0

Let:
Match = +1
Mismatch = 0
Gap = -1

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</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>G</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
</tr>
<tr>
<td>C</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>C</td>
<td>-4</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>