Two Sequence Alignment & Scoring Matrices

BME 110: CompBio Tools
Todd Lowe
April 7, 2011
Admin

• TA Hours posted on class syllabus:
  – Thur 9-10:30 & Fri 2:30-4pm

• Reading for Today:
  – Chapter 8
  – Short Paper
    "Where did the BLOSUM62 alignment score matrix come from?" by S.R. Eddy

• Homework #1 due Friday, Apr 15 @ 9pm
• eCommons site is up, let me know if there are any problems
Dot-plot: Most Basic Sequence Comparison

• Put sequence on X & Y axes
• Mark "X" for nucleotide match
• Self-sequence comparison:
  – Find repeats, insertions, deletions, palindromes
• Two-sequence analysis
  – Find similarity, inversions, transversions, on large scale

Many tools out there: Dotlet is good java app
Full Genome
Dot-Plot
Multiple Genome Alignment
Dot Plots

P. arsenaticum
P. calidifontis
P. islandicum
P. aerophilum

P. aerophilum
P. aerophilum
P. aerophilum

P. calidifontis
P. islandicum
P. islandicum

P. arsenaticum
P. arsenaticum
P. calidifontis
Full-genome Comparisons

Each dot is a gene match

From Zivanovic et al., NAR 30: 1902-10
Pair-wise Sequence Comparison

- Basis for relating biological information from a well-studied gene to a new sequence
- Many programs exist for pairwise comparison
- Some specialize in fast database searching and get “good” alignments
  - One sequence v. many thousands:
    - BLAST or FASTA
- Some are much slower, but guarantee the “optimal alignment”
  - Smith-Waterman is the de facto standard
Dot-plots: Dotlet

http://myhits.isb-sib.ch/cgi-bin/dotlet

- Example: In Archaeal Genome browser, bring up *Pyrobaculum aerophilum*
- Select CRISPR2 region (chr:45,423-46,754) to compare to CRISPR6-7 region (chr:1,898,656-1,899,678)
- Get DNA, paste into Dotlet, one at a time, giving descriptive labels, Zoom 1:5,
- Are there direct or inverted repeats in each CRISPR (against itself?)
- Relative to each other, are these direct or inverted repeats?
What is Optimal??

• How do we get an “optimal” alignment
• Optimal to who?
• Optimal based on **scoring model**:
  – Substitution scoring matrix
  – Insertion / deletion scoring (penalties)
• Caution: Just because it is optimal for a given scoring scheme, doesn’t mean it is biologically correct!!
Dynamic Programming

• Fancy term for type of algorithm used to get the “optimal” or best possible alignment between two sequences

• Needleman and Wunsch (1970) most basic method
  – Gives the “global” (end to end) best alignment

• Smith-Waterman based closely on this algorithm, but allows for “local” alignments (best subsequence match only)
Basic Example

• Find best global alignment of two sequences:

\[
\begin{align*}
  &G A T C \\
  &G T G C
\end{align*}
\]
Which is better?

Match +1, Mismatch −1, Gap -2

G A T C

|     |    OR    (Score = 0)
G T G C

G A T - C

|   |   |
|   |   |

G - T G C

(GScore = -1)

+1−1−1+1

+1−2+1−2+1
Which is better?
Match +1, Mismatch −1, Gap -1

\[
\begin{align*}
G & & A & & T & & C & & +1 & & -1 & & -1 & & 1 & & 1 \\
| & & | & & & & & & \text{OR} & & \text{(Score} & = & 0) \\
G & & T & & G & & C
\end{align*}
\]

\[
\begin{align*}
G & & A & & T & & - & & C & & +1 & & -1 & & 1 & & -1 & & 1 & & 1 \\
| & & | & & | & & & & & & \text{(Score} & = & 1) \\
G & & - & & T & & G & & C
\end{align*}
\]
Moral: Scoring Model Matters!!

• For DNA, model can be very simple:
  • +1 match, -1 mismatch

• However, not all mutations have equal likelihood:
  • Transition: $A \leftrightarrow G$ or $C \leftrightarrow T$
    – more likely
  • Transversion: $A \leftrightarrow C$ or $G \leftrightarrow T$
    – less likely
Kimura Two-parameter Scoring Matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>0.1</td>
<td>0.6</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>G</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>T</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Actual values not important, only values relative to each other.
Same Matrix (*10)

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
</tbody>
</table>

Actual values not important, only values relative to each other
Protein Matrices, Same Idea

- Original: Dayhoff matrix aka PAM
- PAM = Percent accepted mutations
- Based on small number of *correctly* aligned proteins
- Simply count how often each amino acid is substituted for another
- Frequency of substitutions based on properties of amino acids relative to each other
Newer “Version” of Protein Matrices: BLOSUM

- By Henikoff & Henikoff (1992), based on a much larger group of aligned protein sequences in the Blocks database
- BLOSUM = Blocks substitution matrix
- Used most commonly today
- Closer two amino acids are, more similar in properties
Versions of Matrices

- Should use different substitution matrices based on expected evolutionary distance between two sequences
- PAM1 original matrix
- Can derive matrices of varying evolutionary distances from original PAM 1 matrix (alignment of protein sequences where, on average, one change per 100 amino acids)
  - PAM250 = on average, 250 substitutions per 100 amino acids
  - PAM1 < PAM120 < PAM250
# Dayhoff PAM 250 Scoring Matrix

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>Cys</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>P</td>
<td>-3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>N</td>
<td>-4</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>D</td>
<td>-5</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>E</td>
<td>-5</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Q</td>
<td>-5</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>-4</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>K</td>
<td>-5</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>M</td>
<td>-5</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>I</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>L</td>
<td>-6</td>
<td>-3</td>
<td>-2</td>
</tr>
<tr>
<td>V</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>F</td>
<td>-4</td>
<td>-3</td>
<td>-5</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>-3</td>
<td>-5</td>
</tr>
<tr>
<td>W</td>
<td>-8</td>
<td>-2</td>
<td>-5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>C</th>
<th>T</th>
<th>P</th>
<th>A</th>
<th>G</th>
<th>N</th>
<th>D</th>
<th>E</th>
<th>Q</th>
<th>H</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-8</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-8</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>P</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>G</td>
<td>-3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>N</td>
<td>-4</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>-6</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>D</td>
<td>-5</td>
<td>0</td>
<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>-1</td>
<td>-2</td>
</tr>
<tr>
<td>E</td>
<td>-5</td>
<td>0</td>
<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>-1</td>
<td>-2</td>
</tr>
<tr>
<td>Q</td>
<td>-5</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>H</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-5</td>
<td>-6</td>
<td>-7</td>
<td>-1</td>
<td>-2</td>
</tr>
<tr>
<td>R</td>
<td>-4</td>
<td>0</td>
<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>-1</td>
<td>-2</td>
</tr>
<tr>
<td>K</td>
<td>-5</td>
<td>0</td>
<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>-1</td>
<td>-2</td>
</tr>
</tbody>
</table>

Blossum62 matrix - [link]