Sequence Alignment Practical

Presented by
Kirill Bessonov
Nov 12, 2013
Talk Structure

• Introduction to sequence alignments
• Methods / Logistics
  – Global Alignment: Needleman-Wunsch
  – Local Alignment: Smith-Waterman
• Illustrations of two types of alignments
  – step by step local alignment
• Computational implementation of alignment
  – Retrieval of sequences using R
  – Alignment of sequences using R
• Homework – HW2
Sequence Alignments

Comparing two objects is intuitive. Likewise sequence pairwise alignments provide info on:

- evolutionary distance between species (e.g. homology)
- new functional motifs / regions
- genetic manipulation (e.g. alternative splicing)
- new functional roles of unknown sequence
- identification of binding sites of primers / TFs
- \textit{de novo} genome assembly
  
  - alignment of the short “reads” from high-throughput sequencer (e.g. Illumina or Roche platforms)
Comparing two sequences

• There are two ways of pairwise comparison
  – Global using Needleman-Wunsch algorithm (NW)
  – Local using Smith-Waterman algorithm (SW)

• Both approaches use similar methodology, but have completely different objectives
  – Global alignment (NW)
    • tries to align the “whole” sequence
    • more restrictive than local alignment
  – Local alignment (SW)
    • tries to align portions (e.g. motifs) of given sequences
    • more flexible as considers “parts” of the sequence
    • works well on highly divergent sequences
Global alignment (NW)

- Sequences are aligned end-to-end along their entire length
- Many possible alignments are produced
  - The alignment with the highest score is chosen
- Naïve algorithm is very inefficient \(O^{exp}\)
  - To align sequence of length 15, need to consider
    - Possibilities \# = (insertion, deletion, gap)^{15} = 3^{15} = 1.4 \times 10^7
  - Impractical for sequences of length >20 nt
- Used to analyze homology/similarity of entire:
  - genes and proteins
  - assess gene/protein overall homology between species
Methodology of global alignment (1 of 4)

• Define scoring scheme for each event
  – mismatch between $a_i$ and $b_j$
    • $s(a_i, b_j) = -1$ if $a_i \neq b_j$
  – gap (insertion or deletion)
    • $s(a_i, -) = s(-, b_j) = -2$
  – match between $a_i$ and $b_j$
    • $s(a_i, b_j) = +2$ if $a_i = b_j$

• Provide no restrictions on minimal score

• Start completing the alignment $M \times N$ matrix
Methodology of global alignment (2 of 4)

• The matrix should have extra column and row
  – M+1 columns, where M is the length of sequence M
  – N+1 rows, where N is the length of sequence N

• Initialize the matrix by introducing \textit{gap penalty} at every initial position along rows and columns

• Scores at each cell are \textit{cumulative}

\[
\begin{array}{|c|c|c|c|c|}
\hline
& W & H & A & T \\
\hline
\text{W} & 0 & -2 & -2 & -2 & -2 & -8 \\
\text{H} & -2 & -4 & -6 & -8 \\
\text{Y} & -6 \\
\hline
\end{array}
\]
Methodology of global alignment (3 of 4)

- For each cell consider all three possibilities
  1) Gap (horiz/vert)  
  2) Match (W-W diag.)  
  3) Mismatch (W-H diag)

- Select the maximum score for each cell and fill the matrix
Methodology of global alignment (4 of 4)

- Select the most **very bottom right** cell
- Consider different path(s) going to **very top left** cell
  - Path is constructed by finding the **source cell** w.r.t. the current cell
  - How the current cell value was generated? From where?

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>H</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
</tr>
<tr>
<td>H</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>Y</td>
<td>-4</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Y</td>
<td>-6</td>
<td>-2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

**WHAT**

**WHY**

- Overall score = 1

- Select the best alignment(s)
Local alignment (SW)

- Sequences are aligned to find **regions** where the **best** alignment occurs (i.e. highest score)
- Assumes a **local** context (aligning parts of seq.)
- Ideal for finding short motifs, DNA binding sites
  - **helix-loop-helix (bHLH)** - motif
  - TATAAT box (a famous promoter region) – DNA binding site
- Works well on **highly divergent** sequences
  - Sequences with highly variable introns but highly conserved and sparse exons
Methodology of local alignment (1 of 4)

• The scoring system is similar with one exception
  – The minimum possible score in the matrix is zero
  – There are no negative scores in the matrix

• Let’s define the same scoring system as in global
  1) mismatch between \( a_i \) and \( b_j \)
     \[
     s(a_i, b_j) = -1 \text{ if } a_i \neq b_j
     \]
  2) gap (insertion or deletion)
     \[
     s(a_i, -) = s(-, b_j) = -2
     \]
  3) match between \( a_i \) and \( b_j \)
     \[
     s(a_i, b_j) = +2 \text{ if } a_i = b_j
     \]
Methodology of local alignment (2 of 4)

• Construct the $M \times N$ alignment matrix with $M+1$ columns and $N+1$ rows

• Initialize the matrix by introducing gap penalty at 1$^{st}$ row and 1$^{st}$ column
Methodology of local alignment (3 of 4)

• For each subsequent cell consider all possibilities (i.e. motions)
  1) Vertical 2) Horizontal 3) Diagonal

• For each cell select the highest score
  – If score is negative $\rightarrow$ assign zero

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>H</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Methodology of local alignment (4 of 4)

- Select the initial cell with the highest score(s)
- Consider different path(s) leading to score of zero
  - Trace-back the cell values
  - Look how the values were originated (i.e. path)

\[
\begin{array}{c|cccc}
 & W & H & A & T \\
\hline
\text{B} & 0 & 0 & 0 & 0 \\
\text{W} & 0 & 2 & 0 & 0 \\
\text{H} & 0 & 0 & 4 & 2 \\
\text{Y} & 0 & 0 & 2 & 3 \\
\end{array}
\]

- Mathematically

\[
M(A, B) = \max\{S(I, J) : I \subset A, J \subset B\}
\]

where \( S(I, J) \) is the score for sub-sequences \( I \) and \( J \)
Local alignment illustration (1 of 2)

• Determine the best **local** alignment and the maximum alignment score for

• **Sequence A:** ACCTAAGG
• **Sequence B:** GGCTCAATCA

• Scoring conditions:
  - \( s(a_i, b_j) = +2 \) if \( a_i = b_j \),
  - \( s(a_i, b_j) = -1 \) if \( a_i \neq b_j \) and
  - \( s(a_i, -) = s(-, b_j) = -2 \)
Local alignment illustration (2 of 2)

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>G</th>
<th>C</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Local alignment illustration (3 of 3)

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

CTCAA           GGCTCAATCA
CT-AA            ACCT-AAGG

Best score: 6

in the whole seq. context
Aligning proteins
Globally and Locally
Protein Alignment

• Protein local and global alignment follows the same rules as we saw with DNA/RNA

• Differences
  – alphabet of proteins is 22 residues long
  – special scoring/substitution matrices used
  – conservation and protein proprieties are taken into account
  • E.g. residues that are totally different due to charge such as polar Lysine and apolar Glycine are given a low score
Substitution matrices

• Since protein sequences are more complex, matrices are collection of scoring rules
• These are 2D matrices reflecting comparison between sequence A and B
• Cover events such as
  – mismatch and perfect match
• Need to define gap penalty separately
• Popular **BLOcks SUbstitution Matrix (BLOSUM)**
BLOSUM-x matrices

- Constructed from aligned sequences with specific x% similarity
  - matrix built using sequences with no more than 50% similarity is called **BLOSUM-50**

- For highly mutating / dissimilar sequences use
  - BLOSUM-45 and lower

- For highly conserved / similar sequences use
  - BLOSUM-62 and higher
### BLOSUM 62

|     | C | S | T | P | A | G | N | D | E | Q | H | R | K | M | I | L | V | F | Y | W |
| C   | 9 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| S   | -1 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| T   | -1 | 1 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| P   | -3 | -1 | -1 | 7 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| A   | 0 | 1 | 0 | -1 | 4 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| G   | -3 | 0 | -2 | -2 | 0 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
| N   | -3 | 1 | 0 | -2 | -2 | 0 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |   |
| D   | -3 | 0 | -1 | -1 | -2 | -1 | 1 | 6 |   |   |   |   |   |   |   |   |   |   |   |   |
| E   | -4 | 0 | -1 | -1 | -2 | 0 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |   |
| Q   | -3 | 0 | -1 | -1 | -2 | 0 | 0 | 2 | 5 |   |   |   |   |   |   |   |   |   |   |   |
| H   | -3 | -1 | -2 | -2 | -2 | -2 | 1 | -1 | 0 | 0 | -8 |   |   |   |   |   |   |   |   |   |
| R   | -3 | -1 | -1 | -2 | -1 | -2 | 0 | -2 | 0 | 1 | 0 | 5 |   |   |   |   |   |   |   |   |
| K   | -3 | 0 | -1 | -1 | -2 | 0 | -1 | 1 | 1 | 1 | -1 | 2 | 5 |   |   |   |   |   |   |   |
| M   | -1 | -1 | -1 | -2 | -1 | -3 | -2 | -3 | -2 | 0 | -2 | -1 | -1 | 5 |   |   |   |   |   |   |
| I   | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -3 | -3 | -3 | -3 | -3 | -1 | 4 |   |   |   |   |   |   |
| L   | -1 | -2 | -1 | -3 | -1 | -4 | -3 | -4 | -3 | -2 | -3 | -2 | 2 | 2 | 4 |   |   |   |   |   |
| V   | -1 | -2 | 0 | -2 | 0 | -3 | -3 | -3 | -2 | -2 | -3 | -3 | -2 | 1 | 3 | 1 | 4 |   |   |
| F   | -2 | -2 | -2 | -4 | -2 | -3 | -3 | -3 | -3 | -1 | -3 | -3 | -3 | 0 | 0 | 0 | -1 | 6 |   |
| Y   | -2 | -2 | -2 | -3 | -2 | -3 | -2 | -3 | -2 | -1 | 2 | -2 | -2 | -1 | -1 | -1 | -1 | 3 | 7 |   |
| W   | -2 | -3 | -2 | -4 | -3 | -2 | -4 | -4 | -3 | -2 | -2 | -3 | -3 | -1 | -1 | -3 | -2 | -3 | 1 | 2 | 11 |

- **What diagonal represents?** Perfect match between a.a.
- **What is the score for substitution E → D (acid a.a.)?** Score = 2
- **More drastic substitution K → I (basic to non-polar)?** Score = -3
Practical problem:
Align following sequences both globally and locally using BLOSUM 62 matrix with gap penalty of -8

Sequence A: AAEKKLAAA
Sequence B: AARRIA
Aligning globally using BLOSUM 62

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>E</th>
<th>E</th>
<th>K</th>
<th>K</th>
<th>L</th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-8</td>
<td>-16</td>
<td>-24</td>
<td>-32</td>
<td>-40</td>
<td>-48</td>
<td>-56</td>
<td>-64</td>
<td>-72</td>
<td>-80</td>
</tr>
<tr>
<td>A</td>
<td>-8</td>
<td>4</td>
<td>-4</td>
<td>-12</td>
<td>-20</td>
<td>-28</td>
<td>-36</td>
<td>-44</td>
<td>-52</td>
<td>-60</td>
</tr>
<tr>
<td>A</td>
<td>-16</td>
<td>-4</td>
<td>8</td>
<td>0</td>
<td>-8</td>
<td>-16</td>
<td>-24</td>
<td>-32</td>
<td>-40</td>
<td>-48</td>
</tr>
<tr>
<td>R</td>
<td>-24</td>
<td>-12</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>-6</td>
<td>-14</td>
<td>-22</td>
<td>-30</td>
<td>-38</td>
</tr>
<tr>
<td>R</td>
<td>-32</td>
<td>-20</td>
<td>-8</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>-4</td>
<td>-12</td>
<td>-20</td>
<td>-28</td>
</tr>
<tr>
<td>I</td>
<td>-40</td>
<td>-28</td>
<td>-16</td>
<td>-8</td>
<td>0</td>
<td>5</td>
<td>-1</td>
<td>-2</td>
<td>-10</td>
<td>-18</td>
</tr>
<tr>
<td>A</td>
<td>-48</td>
<td>-36</td>
<td>-24</td>
<td>-16</td>
<td>-8</td>
<td>-1</td>
<td>4</td>
<td>-2</td>
<td>2</td>
<td>-6</td>
</tr>
</tbody>
</table>

**Score:** -14

Other alignment options? Yes

AAEEKKLAAA

AA--RRIA--
Aligning locally using BLOSUM 62

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>E</th>
<th>E</th>
<th>K</th>
<th>K</th>
<th>L</th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<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>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

**KKLA**

**RRIA**

**Score: 10**
Using R for:

- Sequence Retrieval and Analysis
Protein database **UniProt**

- UniProt database ([http://www.uniprot.org/](http://www.uniprot.org/)) has high quality protein data **manually** curated
- It is manually curated
- Each protein is assigned **UniProt ID**
Retrieving data from **UniProt**

- In search field one can enter either use **UniProt ID** or common protein name
  - **example:** myelin basic protein

- We will use retrieve data for **P02686**
Understanding UniProt fields

• Information is divided into categories

P02686 (MBP_HUMAN) Reviewed, UniProtKB/Swiss-Prot
Last modified October 3, 2012. Version 154. History...

- Clusters with 100%, 90%, 50% identity
- Documents (4)
- Third-party data
- Names • Attributes • General annotation • Ontologies • Alt products • Sequence annotation • Sequences • References • Web links

• Click on ‘Sequences’ category and then FASTA
FASTA format

- FASTA format is widely used and has the following parameters
  - Sequence name start with > sign
  - The fist line corresponds to protein name

```
>sp|P02686|MBP_HUMAN Myelin basic protein OS=Homo sapiens GN=MBP PE=1 SV=3
MGNHAGKRELAEKASTNSETNRGESEKKNRLGELSRTTSEDENEVFGEADANQNNQGTSQQ
DTAVTDSKRTADPKNAWQDAHPADPGSRPHLIRLSRDAPGREDNTFKDRPSESDELQRT
QEDSAAATSESLDVMASQKRPSPQRHGSKYLATASTMDHARHGFLPRHRDTGILDSIGRFFG
GDRGAPKRSGKDSHHPARTAHYGSLPKSHGRTQDENPVHFFKNIVTPRTPSSQGKK
RGLSLSRFSWGAEGRPCRPGFYGYGRASDYKSAHKGFKGVDQGLSETYKIFKGDDRSRSGSP
MARR
```
Retrieving protein data with R and SeqinR

- Can “talk” programmatically to UniProt database using R and seqinR library
  - seqinR library is suitable for
    - “Biological Sequences Retrieval and Analysis”
    - Detailed manual could be found [here](#)

- Install this library in your R environment
  ```
  install.packages("seqinr")
  library("seqinr")
  ```

- Choose database to retrieve data from
  ```
  choosebank("swissprot")
  ```

- Download data object for target protein **(P02686)**
  ```
  query("MBP_HUMAN", "AC=P02686")
  ```

- See sequence of the object **MBP_HUMAN**
  ```
  MBP_HUMAN_seq = getSequence(MBP_HUMAN)
  ```
Dot Plot (comparison of 2 sequences) (1of2)

• 2D way to find regions of similarity between two sequences
  – Each sequence plotted on either vertical or horizontal dimension
  – If two a.a. from two sequences at given positions are identical the dot is plotted
  – matching sequence segments appear as diagonal lines (that could be parallel to the absolute diagonal line if insertion or gap is present)
Dot Plot (comparison of 2 sequences) (2of2)

- Let’s compare two protein sequences
  - Human MBP (Uniprot ID: P02686)
  - Mouse MBP (Uniprot ID: P04370)

- Download 2nd mouse sequence
  query("MBP_MOUSE", "AC=P04370");
  MBP_MOUSE_seq = getSequence(MBP_MOUSE);

- Visualize dot plot
  dotPlot(MBP_HUMAN_seq[[1]], MBP_MOUSE_seq[[1]], xlab="MBP - Human", ylab = "MBP - Mouse")

- Is there similarity between human and mouse form of MBP protein?
- Where is the difference in the sequence between the two isoforms?
Using R and **Biostrings** library for:

- Pairwise **global** and **local** alignments
Installing Biostrings library

• Install library from Bioconductor
  
  ```
  source("http://bioconductor.org/biocLite.R")
  biocLite("Biostrings")
  library(Biostrings)
  ```

• Define substitution martix (e.g. for DNA)
  
  ```
  DNA_subst_matrix = nucleotideSubstitutionMatrix(match = 2,
  mismatch = -1, baseOnly = TRUE)
  ```

• The scoring rules
  
  - Match: \( s(a_i, b_j) = 2 \) if \( a_i = b_j \)
  - Mismatch: \( s(a_i, b_j) = -1 \) if \( a_i \neq b_j \)
  - Gap: \( s(a_i, -) = -2 \) or \( s(-, b_j) = -2 \)
Global alignment using R and Biostrings

• Create two sting vectors (i.e. sequences)
  ```
  seqA = "GATTA"
  seqB = "GT TA"
  ```

• Use `pairwiseAlignment()` and the defined rules
  ```
  globalAlignAB = pairwiseAlignment(seqA, seqB,
         substitutionMatrix = DNA_subst_matrix, gapOpening = -2,
         scoreOnly = FALSE, type="global")
  ```

• Visualize best paths (i.e. alignments)
  ```
  globalAlignAB
  ```

  Global PairwiseAlignedFixedSubject (1 of 1)
  pattern: [1] GATTA
  subject: [1] G-TTA
  score: 2
Local alignment using R and Biostrings

• Input two sequences
  seqA = "AGGATTTTAAAA"
  seqB = "TTTT"

• The scoring rules will be the same as we used for global alignment
  globalAlignAB = pairwiseAlignment(seqA, seqB,
                                    substitutionMatrix = DNA_subst_matrix, gapOpening = -2,
                                    scoreOnly = FALSE, type="local")

• Visualize alignment
  globalAlignAB
  Local PairwiseAlignedFixedSubject (1 of 1)
  pattern: [5] TTTT
  subject: [1] TTTT
  score: 8
Aligning protein sequences

• Protein sequences alignments are very similar except the substitution matrix is specified
  
  ```r
  data(BLOSUM62)
  BLOSUM62
  ```

• Will align sequences
  
  ```r
  seqA = "PAWHEAE"
  seqB = "HEAGAWGHEE"
  ```

• Execute the global alignment
  
  ```r
  globalAlignAB <- pairwiseAlignment(seqA, seqB,
                                     substitutionMatrix = "BLOSUM62", gapOpening = -2,
                                     gapExtension = -8, scoreOnly = FALSE)
  ```
Summary

• We had touched on practical aspects of
  – Global and local alignments
• Thoroughly understood both algorithms
• Applied them both on DNA and protein seq.
• Learned on how to retrieve sequence data
• Learned on how to retrieve sequences both with R and using UniProt
• Learned how to align sequences using R
Resources

• Online Tutorial on Sequence Alignment

• Graphical alignment of proteins

• Pairwise alignment of DNA and proteins using your rules:

• Documentation on libraries
  – SeqinR: [http://seqinr.r-forge.r-project.org/seqinr_2_0-7.pdf](http://seqinr.r-forge.r-project.org/seqinr_2_0-7.pdf)
Homework – HW2
Homework 2 – literature style (type 1)

You are asked to **analyze critically** by writing a report and **present one** of the following papers in a group:

   - A review paper on popular NGS under the context of genetics of complex diseases

   - A more technical paper on how deep sequencing can help in association studies of rare variants to disease phenotypes under context of statistical genetics

   - An overview paper describing on how NGS technology can be used in the context of epigenetic research. NGS technology described in detail

   - This paper describes on how NGS could be interpreted and contrasted to GWAS. The paper focuses on functional interpretation of genetic variants found in the data
Homework 2 – computer style (type 2)

• You would implement the Needleman–Wunsch global alignment algorithm in R
  – Follow the pseudo-code provided
  – Will translate it into R
  – Will understand alignment in-depth
  – Provide copy of your code and write a short report
    • Report should contain information on scoring matrix and rules used
    • Example sequences used for alignment
    • In code use comments (# comment)
Homework 2 – Q&A style (type 3)

• Here you would need to answer questions
  – Complete the local and global alignment of DNA and protein sequences graphically
  – Use seqinR library to retrieve protein sequences
  – Use Biostrings library to do alignment of sequences
  – Complete missing R code
  – Copy output from R as a proof
  – Calculate alignment scores
Feedback on HW1
HW 1a feedback

• Some almost confused the name of the disease abbreviation with the disease associated genes (e.g. HDL syndromes has no HDL1 gene but PRNP gene is associated with HDL1)

• Some printed the whole genome sequence around the disease gene, but your were asked to print only the protein coding region (CDS)

• Would be nice to get more screen snapshots and see the search query used to find articles
  – From HW1a: “Provide below the search key words used to obtain the results”
HW 2b feedback

• Computer style (type 2):
  – Good analysis on gene level with literature searches
  – Could of addressed results variation before and after cleaning data. What is overlap in results before and after QC?
  – Would be nice to have top 10 SNPs and corresponding p-values before and after cleaning
  – Overall, well done

• Q&A style (type 2)
  – The issue of loading *.phe and *.raw files
    • Set working directory in R where these files are located via
      – setwd()
    • Check current location by getwd()