Pairwise & Multiple sequence alignments

Urmila Kulkarni-Kale
Bioinformatics Centre
University of Pune, Pune 411 007

urmila@bioinfo.ernet.in
Basis for Sequence comparison

• Theory of evolution:
  – gene sequences have evolved/derived from a common ancestor
  – trace history of mutations/evolutionary changes

• Proteins that are similar in sequence are likely to have similar structure and function
WHAT IS ALIGNMENT?

Alignments are useful organizing tools because they provide pictorial representation of similarity / homology in the protein or nucleic acid sequences.
Sample Alignment

• SEQ_A: GDVEKGGKIFIMKCSQ
• SEQ_B: GCVEKGGKIFINWCSQ

There are two possible linear alignments

1. GDVEKGGKIFIMKCSQ
   \[\begin{array}{c}
   \text{SEQ_A} \\
   \text{SEQ_B}
   \end{array}\]
2. GDVEKGGKIFIMKCSQ
   \[\begin{array}{c}
   \text{SEQ_A} \\
   \text{SEQ_B}
   \end{array}\]
The optimal alignment

GDVEKGGKKIFIMKCSQ

\[ \text{GCVEKGK} - \text{IFINWCSQ} \]

Insertion of one break maximizes the identities.
Theoretical background

• Alignment is the method based on the theoretical view that the two sequences are derived from each other by a number of elementary transformations –
  – Mutations (residue substitution)
  – Insertion/deletion
  – Slide function
Transformations
Substitution, Addition/deletion, Slide function

- The most homologous sequences are those which can be derived from one another by the smallest number of such transformations.
- How to decide “the smallest number of transformation?”
- Therefore alignments are an optimization problem.
Terminology

• Identity

• Similarity

• Homology
Identity

• Objective and well defined
• Can be quantified
  – Percent
  – The number of identical matches divided by the length of the aligned region
What is Similarity?

Protein similarity could be due to –

- Evolutionary relationship
- Similar two or three dimensional structure
- Common Function

- Can be quantified
  - Percent
  - The number of identical + similar matches divided by the length of the aligned region
What is Homology?

Homologous proteins may be encoded by-

- Same genes in different species
- Genes that have transferred between the species
- Genes that have originated from duplication of ancestral genes.
Difference between Homology & Similarity

• Similarity does not necessarily imply Homology.
• Homology has a precise definition: having a common evolutionary origin.
• Since homology is a qualitative description of the relationship, the term “% homology” has no meaning.
• Supporting data for a homologous relationship may include sequence or structural similarities, which can be described in quantitative terms.
  – % identities, rmsd
Global Alignment
Local Alignment
An optimal alignment

AALIM

AAL-M

A sub-optimal alignment

AALIM

AA-LM
Needleman & Wunsch algorithm

- Maximizes the number of amino acids of one protein that can be matched with the amino acids of other protein while allowing for optimum deletions/insertions.
- Based on theory of random walk in two dimensions
Random walk in two dimensions

- 3 possible paths
  - Diagonal
  - Horizontal
  - Vertical

- Optimum path
  - Diagonal
N & W Algorithm

• The optimal alignment is obtained by maximizing the similarities and minimizing the gaps.

GLOSSARY

1. PROTEINS The words composed of 20 letters
2. LETTER is an element other than NULL
3. NULL is an symbol “-” i.e. the GAP
4. GAPS Run of nulls which indicates the deletion(s) in one sequence and insertion(s) in other sequence
Contd../

5. SCORING MATRIX

Assigns a value to each possible pair of Amino acids. Examples of matrices are UN, MD, GCM, CSW, UP.

6. PENALTY

There are two types of penalties.

- Matrix Bias: is added to every cell of the scoring matrix and decides the size of the break. Also called Gap continuation penalty.

- Break Penalty: Applied every time a gap is inserted in either sequence.
Unitary Matrix

• Simplest scoring scheme

• Amino acids pairs are classified into 2 types:
  – Identical
  – Non-identical

• Identical pairs are scored 1

• Non-identical pairs are scored 0

• Less effective for detection of weak similarities

```
  A  1  0  0  0
  R  0  1  0  0
  N  0  0  1  0
  D  0  0  0  1

...```

MAT(i,j) = SM(A_i, B_j) + max(x, y, z) where
X = row max along the diagonal – penalty
Y = column max along the diagonal – penalty
Z = next diagonal: MAT(i+1,j+1)

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>D</th>
<th>V</th>
<th>E</th>
<th>K</th>
<th>G</th>
<th>K</th>
<th>K</th>
<th>I</th>
<th>F</th>
<th>I</th>
<th>M</th>
<th>K</th>
<th>C</th>
<th>S</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>1</td>
<td>40</td>
<td>37</td>
<td>34</td>
<td>31</td>
<td>29</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>35</td>
<td>37</td>
<td>35</td>
<td>32</td>
<td>29</td>
<td>27</td>
<td>25</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>3</td>
<td>32</td>
<td>32</td>
<td>35</td>
<td>33</td>
<td>30</td>
<td>27</td>
<td>25</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>E</td>
<td>4</td>
<td>29</td>
<td>29</td>
<td>29</td>
<td>32</td>
<td>31</td>
<td>28</td>
<td>25</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>K</td>
<td>5</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
<td>29</td>
<td>29</td>
<td>26</td>
<td>23</td>
<td>19</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>K</td>
<td>7</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>23</td>
<td>23</td>
<td>23</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td>8</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>14</td>
<td>11</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>19</td>
<td>13</td>
<td>11</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>I</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>13</td>
<td>16</td>
<td>11</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>W</td>
<td>12</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>13</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>9</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>S</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Q</td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</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>
• Real Score (R)
  – Similarity Score of real sequences

• Mean Score (M)
  – Average similarity score of randomly permuted sequences

• Standard deviation (Sd)
  – Standard deviation of the similarity scores of randomly permuted sequences.

• Alignment Score (A)
  – A = (R-M)/sd
  – Alignment score is expressed as number of standard deviation units by which the similarity score for real sequences (R) exceeds the average similarity score (M) of randomly permuted sequences.
Trace back

GDVEKGGKKIFIMKCSQ
| | | | | | | |
GCVEKGGK-IFINWCSQ
Sample output
Evolutionary process
Orthologues

- A single Gene X is retained as the species diverges into two separate species
- Genes in two species are Orthologues
Evolutionary process

Paralogues: genes that arise due to duplication

- Single gene X in one species is duplicated
- As each gene gathers mutations, it may begin to perform new function or may specialize in carrying out functions of ancestral genes
- These genes in a single species are paralogues
- If the species diverges, the daughter species may maintain the duplicated genes, therefore each species contain an Orthologue and a Paralogue to each gene in other species
Homologous/Orthologous/Paralogous sequences

- Orthologous sequences are homologous sequences in different species that have a common origin.
- Distinction of Orthologues is a result of gradual evolutionary modifications from the common ancestor.
- Perform same function in different species.

- Paralogous sequences are homologous sequences that exists within a species.
- They have a common origin but involve gene duplication events to arise.
- Purpose of gene duplication is to use sequence to implement a new function.
- Perform different functions.
Local Sequence Alignment Using Smith-Waterman Dynamic Programming Algorithm
Significance of local sequence alignment

• In global alignment, an attempt is made to align the entire sequences, as many characters as possible.

• In local alignment, stretches of sequence with the highest density of matches are given the highest priority,

• generating one or more islands of matches in the aligned sequences.

Applications: locating common domains in proteins

Example: transmembrane proteins, which might have different ends sticking out of the cell membrane, but have common 'middleparts'

For comparing long DNA sequences with a short one
Comparing a gene with a complete genome

For detecting similarities between highly diverged sequences which still share common subsequences (that have little or no mutations).
Local sequence alignment

• Performs an exhaustive search for optimal local alignment

• Modification of Needleman-Wunsch algorithm:
  
  • Negative weighting of mismatches
  
  • Matrix entries non-negative
  
  • Optimal path may start anywhere (not just first / last row/column)

• After the whole path matrix is filled, the optimal local alignment is simply given by a path starting at the highest score overall in the path matrix, containing all the contributing cells until the path score has dropped to zero.
Smith-Waterman Algorithm

The score at $s(i,j)$ is the maximum of:

- $s(i-1,j-1) + s(a,b)$
- $s(i,j-1)$ - gap penalty
- $s(i-1,j)$ - gap penalty
Example of local alignment

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>E</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>W</th>
<th>G</th>
<th>H</th>
<th>E</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</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>5</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>18</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>2</td>
<td>16</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>18</td>
<td>28</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>21</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>13</td>
<td>18</td>
<td>12</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

From Burbin et al. 1998

AWGHE
AW-HE

University of Pune, Pune
### Scoring the alignment using BLOSUM50 matrix

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>E</th>
<th>A</th>
<th>G</th>
<th>A</th>
<th>W</th>
<th>G</th>
<th>H</th>
<th>E</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>-3</td>
<td>0</td>
<td>-2</td>
<td>-1</td>
</tr>
<tr>
<td>W</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>15</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

**Gap penalty:** -8