4. MULTIPLE SEQUENCE ALIGNMENT

BIOINFORMATICS COURSE
MTAT.03.239

02.10.2012
ALIGNMENT

- Alignment is the task of locating “**equivalent**” regions of **two or more** sequences to **maximize their similarity**.
- **Homology**: similarity due to descent from a **common ancestor**.
- Often we can detect **homology** by the **similarity** of the sequences.
- We can **deduce structure and function** from the aligned sequences, for selective pressure of evolution results from the need to conserve a function.
EVOLUTION OF THE GLOBINS

http://www.els.net/WileyCDA/ElArticle/refId-a0005134.html

"Sequence analysis"
Bioinformatics Course
EVOLUTIONARY CONSERVATION


"Sequence analysis"
Bioinformatics Course
ISSUES IN SEQUENCE ALIGNMENT

- The sequences are of different lengths.
- Only a relatively small region in the sequences match against each other.
- Some amino acid substitutions are more likely to occur during evolution than others.
- Sequence changes more rapidly in evolution than does structure and function.
QUESTIONS WHEN ALIGNING

➢ What do we align?
  ➢ Global alignment – finds the best match of sequences over its entire length.
  ➢ Local alignment – finds best subsequence matches.

➢ How do we score the alignment?
  ➢ gap penalty – cost of a gap on sequence alignment
  ➢ substitution matrix – cost of character substitutions

➢ How do we find the best alignment?
  ➢ Compute and score all possible alignments (e.g. 2 sequences of length 100 will have \(10^{77}\) possible alignments).
SCORING THE ALIGNMENT

The score of the alignment is the sum of the scores of pairs of aligned characters plus the scores for gaps.

BIOINFORMATICS

|||     |||||
BIORS---MATICS

Score would be:
\[ s(B,B) + s(I,I) + s(O,O) + s(I,R) + s(N,S) + 3g + \ldots \]
EQUALLY OPTIMAL ALIGNMENTS

- One sequence comparison can produce several optimal alignments.
- One can use preference ordering over paths when doing traceback.

- *highroad* and *lowroad* alignments show the two most different optimal alignments.
## Equally Optimal Alignments

### Highroad Alignment
- Alignment: AG - C
- Sequence: AAAC

### Lowroad Alignment
- Alignment: -AGC
- Sequence: AAAC
GLOBAL VS LOCAL ALIGNMENT

- Global alignment tries to align the entire sequence, using as many characters as possible, up to both ends of each sequence.
- In local alignment, stretches of sequences with the highest density of matches are aligned, generating one or more subalignments in the aligned sequences.

```
LGPSSKQTGKGS-SRIWDN
\ | | | | | | |
LN-ITKSAGKGAIMRLGDA  

----------TGKG--------
\ | | | |  
----------AGKG--------
```

Global alignment

Local alignment
MULTIPLE SEQUENCE ALIGNMENT (MSA) INTRODUCTION

- DNA sequences of different organisms are often related.
- Similar genes are conserved across widely divergent species, often performing a similar or even identical function.
- Through simultaneous alignment of the sequences of these genes, sequence patterns that have been subject to alteration may be analyzed.
- MSA is the starting point for phylogenetic analysis.
- The most successful is the alignment that most closely represents the evolutionary history of sequences.
### MULTIPLE SEQUENCE ALIGNMENT

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Alignment</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q5E940</td>
<td><strong>ROF IN</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_HUMAN</td>
<td><strong>HUMAN</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_MOUSE</td>
<td><strong>MOUSE</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_RAT</td>
<td><strong>RAT</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_CHICK</td>
<td><strong>CHICK</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_RANSA</td>
<td><strong>RANSA</strong></td>
<td>76</td>
</tr>
<tr>
<td>Q7ZUG3</td>
<td><strong>BRARE</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_ICTPU</td>
<td><strong>ICTPU</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_DROME</td>
<td><strong>DROME</strong></td>
<td>76</td>
</tr>
<tr>
<td>RLA0_DICDI</td>
<td><strong>DICDI</strong></td>
<td>75</td>
</tr>
<tr>
<td>Q541P0</td>
<td><strong>DICDI</strong></td>
<td>75</td>
</tr>
</tbody>
</table>

**Sequence analysis**

Bioinformatics Course
WHY TO DO MSA?

- To characterize protein families, identify shared regions of homology in MSA.
- We can observe which regions of a gene are susceptible to mutations and which are not.
- Help predictions of secondary and tertiary structures of new sequences.
- The presence of several domains similar to domains in previously characterized sequences, can imply similar structure or function.
- To identify motifs.
WHY TO DO MSA?

➢ To characterize protein families, identify shared regions of homology in MSA.

➢ We can observe which regions of a gene are susceptible to mutations and which are not.

➢ An alignment provides a bird’s eye view of the underlying evolutionary, structural, or functional constraints characterizing a protein family in a concise, visually intuitive format.

➢ The presence of several domains similar to domains in previously characterized sequences can imply similar structure or function.

➢ To identify motifs.

“Sequence analysis" Bioinformatics Course
WHICH SEQUENCES TO ALIGN?

- The more sequences to align the better.
- Don’t include too similar sequences.
- Try to include sequences not too far apart in length.
  - Trim sequences down, so as to only use regions that have been deemed similar by BLAST for example.
- Good starting point: use sequences that are 30-70% similar to most of the other sequences in the data set.
TO IDENTIFY MOTIFS

“Sequence analysis”
Bioinformatics Course
HOW TO REPRESENT A PATTERN

\[
\begin{align*}
\text{Phe or Tyr} & \quad \text{Cys} & \quad \text{not Val or Ala} & \quad \text{three His} \\
\rightarrow & \quad \rightarrow & \quad \rightarrow & \quad \rightarrow \\
[FY] & -x-C-x(2) & \{VA\} & -x-H(3) \\
\rightarrow & \quad \rightarrow & \quad \rightarrow & \quad \rightarrow \\
\text{any amino acid} & \quad \text{any two amino acids} & \quad \text{any amino acid} \\
\end{align*}
\]
PROFILE REPRESENTATION OF MSA

- Can we align a sequence against a profile?
- Can we align a profile against a profile?

bix.ucsd.edu/bioalgorithms/presentations/Ch06_MultAlign.ppt
PROFILE REPRESENTATION OF MSA

- Allows us to identify consensus sequence.
- Depend upon patterns and motifs containing conserved residues.
- Profile can be used in database searches.
- Used to find new sequences that match the profile.
- Can be used to compute multiple sequence alignments heuristically.
- Allow for the analysis of distantly related proteins.
CONSENSUS SEQUENCE

- A single sequence that represents the most common character specific to every position

sequenceC
-VLSPADKTNVKAAGKVGAGHAGYGAELERMFSLSFPTTKTYFPFHDLSH 50

sequenceD
-VLSAADKTNVKAASSKVGAGHAGYGAELERMFSLGFPTTKTYFPFHDLSH 50

sequenceA
VHLTPEEKSATALWGNKN--VDEVGGEALGRLLVYPWTQRFFESFGDSLST 50

sequenceB
VQLSGGEKAAVLALWGDKN--EEEVGEALGRVLVYPWTQRFFDSGISL 50


VVLSPADKTNVKAAWGKVGAHAHEYGAELERMFSLVPTTKTYFPFHDLSH 50

EE A L NG V G LLV Y W QRF S
MSA APPROACHES

- **Optimal Global Alignments** – Dynamic programming.
  - Generalization of Needleman-Wunch.
  - Find alignment that maximizes a score function.
- **Progressive Alignments** – Match closely related sequences first using a guide tree.
- **Iterative Methods** – Multiple re-building attempts to find best alignment.
- **Probabilistic approaches**
Dynamic programming alignment of $k$ strings of length $m$ would take time proportional to $O(m^k)$. This is not computable for more then a few sequences. Each alignment is a path through the dynamic programming matrix.

\[
\begin{align*}
O(c) & \quad \text{utopian} \\
O(\log k) & \quad \text{excellent} \\
O(k) & \quad \text{very good} \\
O(k^2) & \quad \text{not so good} \\
O(k^3) & \quad \text{pretty bad} \\
O(c^k) & \quad \text{disaster}
\end{align*}
\]
PROGRESSIVE ALIGNMENT

- The most practical and widely used method in MSA is the hierarchical extensions of pairwise alignment methods.
- Employs multiple pairwise alignments in a series of three steps:
  - Estimate alignment scores between all possible pairwise combinations of sequences in the set.
  - Build a “guide tree” determined by the alignment score.
  - Align the sequences on the basis of the guide tree.
- Best known programs employing this approach is the CLUSTAL family.
Gene evolution is commonly assumed to take place along a tree

Interdependence of alignment and phylogeny. (A) Sequences are related through an unknown evolutionary tree; mutations occur along the branches that connect the sequences at the nodes. (B) Progressive alignment methods try to backtrack the evolutionary process and build the multiple alignment from sequential pairwise alignments performed according to a guide tree. (C) Sequence alignment is an inference of homology: each column represents the descendants of one ancestral character. (D) Phylogenetic inference is based on the alignment (C), which may be affected by the guide tree used to build it (B). The method of Liu et al. reduces this effect by alternating between steps (B) and (D) until their likelihood criterion no longer improves.

http://www.cbrg.ethz.ch/education/Biolnf2/MSA.key.pdf
Löytynoja and Goldman, Science 2009
PROGRESSIVE ALIGNMENT

 Align two sequences at a time.
 Perform cluster analysis by gradually building up multiple sequence alignment by merging larger and larger sub-alignments based on their similarity.
 Uses protein scoring matrices and gap penalties to calculate alignments having the best score.
 Major advantages of method:
   Generally fast.
   Alignments generally of high quality.

http://www.genome.gov/12514288
PROGRESSIVE ALIGNMENT

>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST
>sequence B
VQLSGEEKAAVLALWDKVNEEVEVGGEALGRLLVVYPWTQRFFDSFGDSLN
>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDSLH
>sequence D
VLSAADKTNVKAAWSKVGHHAGEYGAEALERMFLGFPTTKTYFPHFDSLH

http://www.genome.gov/12514288
PROGRESSIVE ALIGNMENT

- Calculate a similarity score (percent identity) between every pair of sequences to drive the alignment

How many calculations required for $n$ sequences?

http://www.genome.gov/12514288
PROGRESSIVE ALIGNMENT

- Calculate a similarity score (percent identity) between every pair of sequences to drive the alignment

How many calculations required for \( n \) sequences?

\[
\left[ n \times (n - 1) \right] / 2
\]

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
</tr>
<tr>
<td>25</td>
<td>300</td>
</tr>
<tr>
<td>50</td>
<td>1,225</td>
</tr>
<tr>
<td>100</td>
<td>4,950</td>
</tr>
</tbody>
</table>
PROGRESSIVE ALIGNMENT

>sequence A
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLST

>sequence B
VQLSGEEKAAVLALWDKVNEEEVGGEGALGRLLVVYPWTQRFFDSSFGDSDLN

>sequence C
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTTYFPHFDLSH

>sequence D
VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTTYFPHFDLSH

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>80</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>44</td>
<td>40</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>40</td>
<td>92</td>
<td>100</td>
</tr>
</tbody>
</table>

In this matrix, sequence A is 80% identical to sequence B
PROGRESSIVE ALIGNMENT

- Derive a dendrogram – a guide tree - based on the pairwise comparisons to determine the “order” of sequences.
- Based on the tree A and B share greater similarity with each other than with either C or D.

<table>
<thead>
<tr>
<th>%ID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>80</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>44</td>
<td>40</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>40</td>
<td>40</td>
<td>92</td>
<td>100</td>
</tr>
</tbody>
</table>

http://www.genome.gov/12514288
Starting with a star tree (A), the Q matrix is calculated and used to choose a pair of nodes for joining, in this case f and g. These are joined to a newly created node, u, as shown in (B). The part of the tree shown as dotted lines is now fixed and will not be changed in subsequent joining steps. This process is then repeated, using a matrix of just the distances between the nodes, a, b, c, d, e, and u, and a Q matrix derived from it. In this case u and e are joined to the newly created v, as shown in (C). Two more iterations lead first to (D), and then to (E), at which point the algorithm is done, as the tree is fully resolved.
PROGRESSIVE ALIGNMENT

- Align A with B → alignment AB (fixed).
- Align C with D → alignment CD (fixed).
- Represent alignments AB and CD as single sequences.
PROGRESSIVE ALIGNMENT

- Align “sequence” **AB** with “sequence” **CD**
- Continue following the branching order of the tree from the **tips to the root**, merging each new pair of “sequences”.

http://www.genome.gov/12514288
PROGRESSIVE ALIGNMENT

➢ Start by aligning the two most similar sequences.
➢ Following the guide tree, add in the next sequences, aligning to the existing alignment.
➢ Insert gaps if necessary.

Calculate:
step 1 = alignment(A, B)
step 2 = alignment(C, D)
step 3 = alignment((C, D), E)
step 4 = alignment((A, B), ((C, D), E))
CLUSTAL 2.1 Multiple Sequence Alignments

Sequence type explicitly set to Protein
Sequence format is Pearson
Sequence 1: sequenceA 50 aa
Sequence 2: sequenceB 50 aa
Sequence 3: sequenceC 50 aa
Sequence 4: sequenceD 50 aa
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score: 76
Sequences (1:3) Aligned. Score: 38
Sequences (1:4) Aligned. Score: 34
Sequences (2:3) Aligned. Score: 36
Sequences (2:4) Aligned. Score: 36
Sequences (3:4) Aligned. Score: 92
Guide tree file created: [clustalw2-I20121001-152141-0431-66276267-oy.dnd]

There are 3 groups
Start of Multiple Alignment
Aligning...
Group 1: Sequences: 2 Score: 1050
Group 2: Sequences: 2 Score: 956
Group 3: Sequences: 4 Score: 702
Alignment Score 881

CLUSTAL-Alignment file created [clustalw2-I20121001-152141-0431-66276267-oy.aln]
PROGRESSIVE ALIGNMENT
CLUSTALW

"Sequence analysis"
Bioinformatics Course
ClustalW calculates the genetic distance as the number of mismatched positions in an alignment divided by the total number of matched positions (positions opposite a gap are not scored).

\[
\text{genetic distance} = \frac{\text{no. of mismatches in the alignment}}{\text{no. of matches in the alignment}}
\]
PROGRESSIVE ALIGNMENT
CLUSTALW
PROGRESSIVE ALIGNMENT

- To construct a multiple alignment, one may have to introduce gaps in sequences at positions where there were no gaps in the corresponding pairwise alignment.
- Multiple alignments often contain more gaps than the original pairwise alignments.
- Gap-free blocks probably correspond to regions of secondary structure.
- Gap-rich blocks probably correspond to unstructured or loop regions.
PROGRESSIVE ALIGNMENT
DISADVANTAGES

➢ Errors in the initial alignments of the most closely related sequences are propagated to the MSA.

➢ Once an alignment is “fixed” it is not reconsidered, so any errors in the early alignments may propagate through subsequent alignments.

➢ The right choices of suitable scoring matrices and gap penalties?
ITERATE METHODS

- Similar to progressive alignment.
- Rectify the mistake in alignment by iteration.
- The objective is to improve the overall alignment score, such as a sum of pairs score.
- Iterations are performed till no further improvement.
- Examples: MUSCLE, IterAlign, Praline, MAFFT
ITERATE METHODS

**Step 1.** Create an initial alignment.

**Step 2.** Take out one sequence.

**Step 3.** Realign the one sequence to the profile of the remaining sequences.

**Step 4.** Choose another sequences and realign.

**Step 5.** Perform step 4 until the alignment score converges.

“Sequence analysis"
Bioinformatics Course
ITERATE METHODS
MUSCLE

“Sequence analysis”
Bioinformatics Course
SCORING MSA

- Number of matches (multiple longest common subsequence score)
- Entropy score
- Sum of pairs (SP)
LCS SCORE

- A column is a “match” if all the letters in the column are the same.
- Only good for very similar sequences.

AAA
AAA
AAT
ATC
ENTROPY

- Calculate frequencies of each letter in each column of the multiple sequence alignment.
- Calculate entropy of each column.
- Entropy for MSA is the sum of entropies of its columns.

\[-\sum_{X=A,T,G,C} p_x \log p_x\]

A
AAA
A
AAA
AAT
ATC

\[
\text{entropy} \begin{pmatrix} A \\ A \\ A \\ A \end{pmatrix} = 0 \quad \text{Best case}
\]

\[
\text{entropy} \begin{pmatrix} A \\ T \\ G \\ C \end{pmatrix} = -\sum \frac{1}{4} \log \frac{1}{4} = -4 \left( \frac{1}{4} \ast -2 \right) = 2 \quad \text{Worst case}
\]
ENTROPY

- Calculate frequencies of each letter in each column of the multiple sequence alignment.
- Calculate entropy of each column.
- Entropy for MSA is the sum of entropies of its columns.

```r
> a = -( 1*log2(1) )
> b = -( .25*log2(.25) + .75*log2(.75) )
> c = -( .25*log2(.25) + .25*log2(.25) + .25*log2(.25) + .25*log2(.25) )
> entropy = a + b + c
> entropy
[1] 2.811278
```
**SUM OF PAIRS (SP)**

- The SP score is the sum of all pairwise scores for all residues in the alignment.
- Pairs of gaps score 0.

```r
library(Biostring)
library(Biostrings)
data(BLOSUM62)
BLOSUM62
```

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

* gap penalty -4

**Sequence 1:** ELVIS  
**Sequence 2:** LIVES  
**Sequence 3:** ALIVE

Total = \([s(E,L) + s(E,A) + s(L,A)] + [s(S,S) + s(S,E) + s(S,E)] \)  
\((-3+1-1)+(2+4+2)+(4+3+3) + (-3+3-2)+(4+0+0) = 15\)
ALN format was originated in the alignment program ClustalW2. The file starts with the word "CLUSTAL" and then some information about which clustal program was run and the version of clustal used.

E.g. "CLUSTAL W (2.1) multiple sequence alignment"
The type of clustal program is "W" and the version is 2.1.
The alignment is written in blocks of 60 residues.
Every block starts with the sequence names, obtained from the input sequence, and a count of the total number of residues is shown at the end of the line.
The information about which residues match is shown below each block of residues:

"*" means that the residues or nucleotides in that column are identical in all sequences in the alignment.
"." means that conserved substitutions have been observed.
"." means that semi-conserved substitutions are observed.
Fig. 1. Diagram of the basic steps in a prototypical modern multiple sequence alignment program: computation of matrix of distances between all pairs of input sequences; estimation of phylogenetic guide tree based on distance matrix; progressive alignment according to guide tree; guide tree reestimation and realignment; iterative refinement; and postprocessing and visualization.

http://ai.stanford.edu/~chuongdo/papers/alignment_review.pdf
LINKS

➢ ClustalW
http://www.ebi.ac.uk/Tools/msa/clustalw2/

➢ MUSCLE
http://www.ebi.ac.uk/Tools/msa/muscle/

➢ T-Coffee
http://www.ebi.ac.uk/Tools/msa/tcoffee/