EE550
Computational Biology

Week 6 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Sequence alignment
  - – Dynamic programming
  - – BLAST
  - – Multiple sequence alignment

Sequence Alignment

- Similarity between gene and protein sequences has dramatic implications in terms of evolutionary and functional relationships
  - Similarity implies simply that sequences “look alike”
  - Similar sequences carry homology, and homology implies evolutionary relationship
- Assessment of similarity between pairs or sets of sequences is obtained via alignment
  - The sequences “slide” with respect to another
  - They are “anchored” at sites where strong matching is observed
    - Strong matching indicates correspondence between the respective sites
  - The remaining regions are padded with gaps to provide the most likely pairing
- The “best” way of aligning two sequences requires making certain assumptions on the probabilistic nature of mismatches and gaps
  - Different “optimal” alignments depending on the optimality criterion of choice

Example: Aligning Two Amino Acid Sequences

- Cell division inhibitor MciZ
  - Bacillus subtilis (strain 168)
  - Sequence:

  `MKVHRMPKGVVLVGKAWEIRAKLKEYGRTFQYVKDWISKP`
- Uncharacterized protein DUF3936
  - Bacillus sp. YR335
  - Sequence

  `MGFLIMKIYRLEKGIVLIGKAWEIRTKLKEYHRTYATVNEWLHDEKTLKQSSKVSKQ`

```
MKVHRMPKGVVLVGKAWEIRAKLKEYGRTFQYVKDWISKP
|    |       |      |
MGFLIMKIYRLEKGIVLIGKAWEIRTKLKEYHRTYATVNEWLHDEKTLKQSSKVSKQ
```

Amino Acid Sequence Alignment

- Strong similarities between amino acid sequences of two proteins indicate
  - a close evolutionary relationship,
  - shared domains and sequence motifs, and hence,
  - functional similarity
- Aligning two amino acid sequences identifies the conserved sequence segments between them
  - Alignment establishes the **correspondence between the sites of the two sequences**

Amino Acid Sequence Alignment

- Amino acid sequence alignment models an unknown hypothetical sequence of origin
  - The two sequences are presumed to have been obtained from the hypothetical sequence via **deletions** and **substitutions**
  - Alignment consists of identifying
    - the **sequence patterns that overlap** up to only a few amino acid replacements, and
    - the **sequence gaps** in the two sequences that shift the overlapping segments onto one another

```
     MKVHRMPKGVVLVGKAWEIRAKLKEYGRTFQYVKDWISKP
   ? ||? |? ||?||?|||||||?|||||?|| ? |? |      ?
MGFLIMKIYRLEKGIVLIGKAWEIRTKLKEYHRTYATVNEWLHDEKTLKQSSKVSKQ
```

Example: Aligning Two Amino Acid Sequences

```
_____MKVHRMPKGVVLVGKAWEIRAKLKEYGRTFQYVKDWISKP____________
     ||  |  || || ||||||| ||||| ||   |  |
MGFLIMKIYRLEKGIVLIGKAWEIRTKLKEYHRTYATVNEWLHDEKTLKQSSKVSKQ
```

?

```
_____MKVHRMPKGVVLVGKAWEIRAKLKEYGRTF______________QYVKDWISKP
     ||  |  || || ||||||| ||||| ||               |  |   ||
MGFLIMKIYRLEKGIVLIGKAWEIRTKLKEYHRTYATVNEWLHDEKTLKQSSKV__SKQ
```

Alignment via Dynamic Programming

- The sequence alignment problem can be solved optimally using dynamic programming
  - Costs are assigned to replacements and gaps
  - The overall similarity is maximized via the optimal alignment
- Setup:
  - Given two amino acid sequences P and Q of respective lengths N and M
  - Define the similarity function

$$S_{i,j} = \begin{cases} \alpha & \text{if } P_i = Q_j \\ \beta & \text{otherwise} \end{cases}$$

for $i = 1,2,\ldots,N$ and $j = 1,2,\ldots,M$, and where

α measures the improved similarity by a perfect match, and

β the cost of replacing one with the other

  - Assign a gap penalty ω for moving along one sequence while remaining stationary in the other

Alignment via Dynamic Programming

- Procedure:
  - Start with an empty $(N+1) \times (M+1)$ matrix A
    - except for the first row and the first column which are all zero
  - Gradually fill the matrix using the rule

$$A_{i,j} = \max\{A_{i-1,j-1} + S_{i,j}, A_{i-1,j} + \omega, A_{i,j-1} + \omega\}$$

and keep track of which entry produced the maximum

  - Once all the matrix is filled, backtrack the maximum-producing steps starting from the maximal element of A on the last row or the last column
  - The traced steps reveal the optimal alignment

Example

- Two amino acid sequences are given

$$P\text{: MGLESKDSPLDGRE}$$

$$Q\text{: MGERSDSPGSDERAWE}$$

with $N = 14$ and $M = 16$

- These sequences are to be aligned using

$$\alpha = 2$$

$$\beta = -1$$

$$\omega = -2$$

and dynamic programming

Example

		M	G	E	R	S	D	S	P	G	S	D	E	R	A	W	E
M																	
G																	
L																	
E																	
S																	
K																	
D																	
S																	
P																	
L																	
D																	
G																	
R																	
E																	

Example

- Starting with $A_{1,1}$:

$$A_{1,1} = \max\{A_{0,0} + S_{1,1}, A_{0,1} + \omega, A_{1,0} + \omega\}$$

  - Evaluating the three options:
    - $A_{0,0} = 0, S_{1,1} = 2 \rightarrow A_{0,0} + S_{1,1} = 2$
    - $A_{0,1} = 0, \omega = -2 \rightarrow A_{0,1} + \omega = -2$
    - $A_{1,0} = 0, \omega = -2 \rightarrow A_{1,0} + \omega = -2$
  - Hence,

$$A_{1,1} = 2$$

  - Furthermore, $A_{1,1}$ is best reached from $A_{0,0}$

Example

- Calculating $A_{2,1}$:

$$A_{2,1} = \max\{A_{1,0} + S_{2,1}, A_{1,1} + \omega, A_{2,0} + \omega\}$$

  - Evaluating the three options:
    - $A_{1,0} = 0, S_{2,1} = -1 \rightarrow A_{1,0} + S_{2,1} = -1$
    - $A_{1,1} = 2, \omega = -2 \rightarrow A_{1,1} + \omega = 0$
    - $A_{2,0} = 0, \omega = -2 \rightarrow A_{2,0} + \omega = -2$
  - Hence,

$$A_{2,1} = 0$$

  - Furthermore, $A_{2,1}$ is reached best from $A_{1,1}$

Example

- Calculating $A_{1,2}$:

$$A_{1,2} = \max\{A_{0,1} + S_{1,2}, A_{0,2} + \omega, A_{1,1} + \omega\}$$

  - Evaluating the three options:
    - $A_{0,1} = 0, S_{1,2} = -1 \to A_{0,1} + S_{1,2} = -1$
    - $A_{0,2} = 0, \omega = -2 \to A_{0,2} + \omega = -2$
    - $A_{1,1} = 2, \omega = -2 \to A_{1,1} + \omega = 0$
  - Hence,

$$A_{1,2} = 0$$

  - Furthermore, $A_{1,2}$ is reached best from $A_{1,1}$

Example

- Calculating $A_{2,2}$:

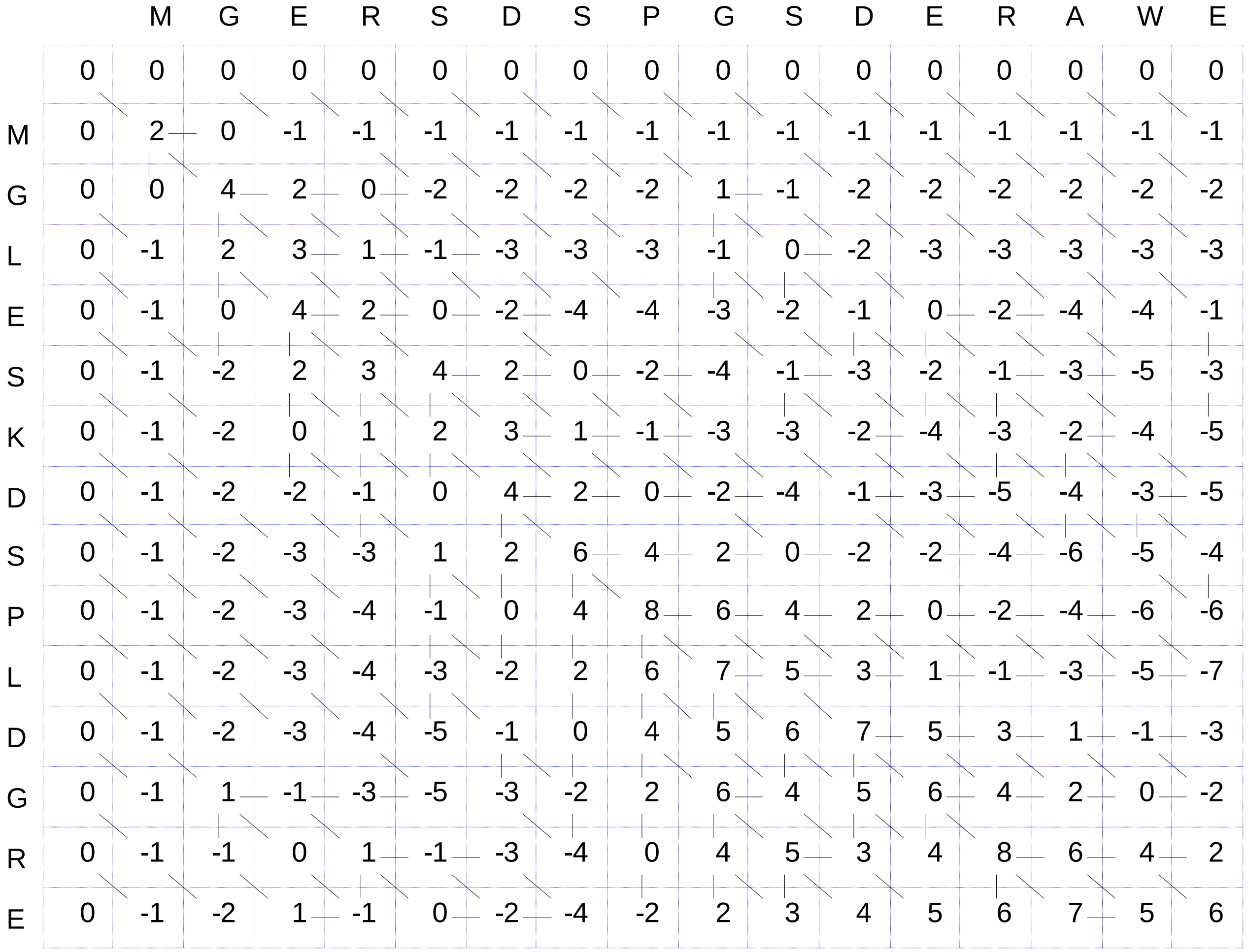
$$A_{2,2} = \max\{A_{1,1} + S_{2,2}, A_{1,2} + \omega, A_{2,1} + \omega\}$$

  - Evaluating the three options:
    - $A_{1,1} = 2, S_{2,2} = 2 \rightarrow A_{1,1} + S_{2,2} = 4$
    - $A_{1,2} = 0, \omega = -2 \rightarrow A_{1,2} + \omega = -2$
    - $A_{2,1} = 0, \omega = -2 \rightarrow A_{2,1} + \omega = -2$
  - Hence,

$$A_{2,2} = 4$$

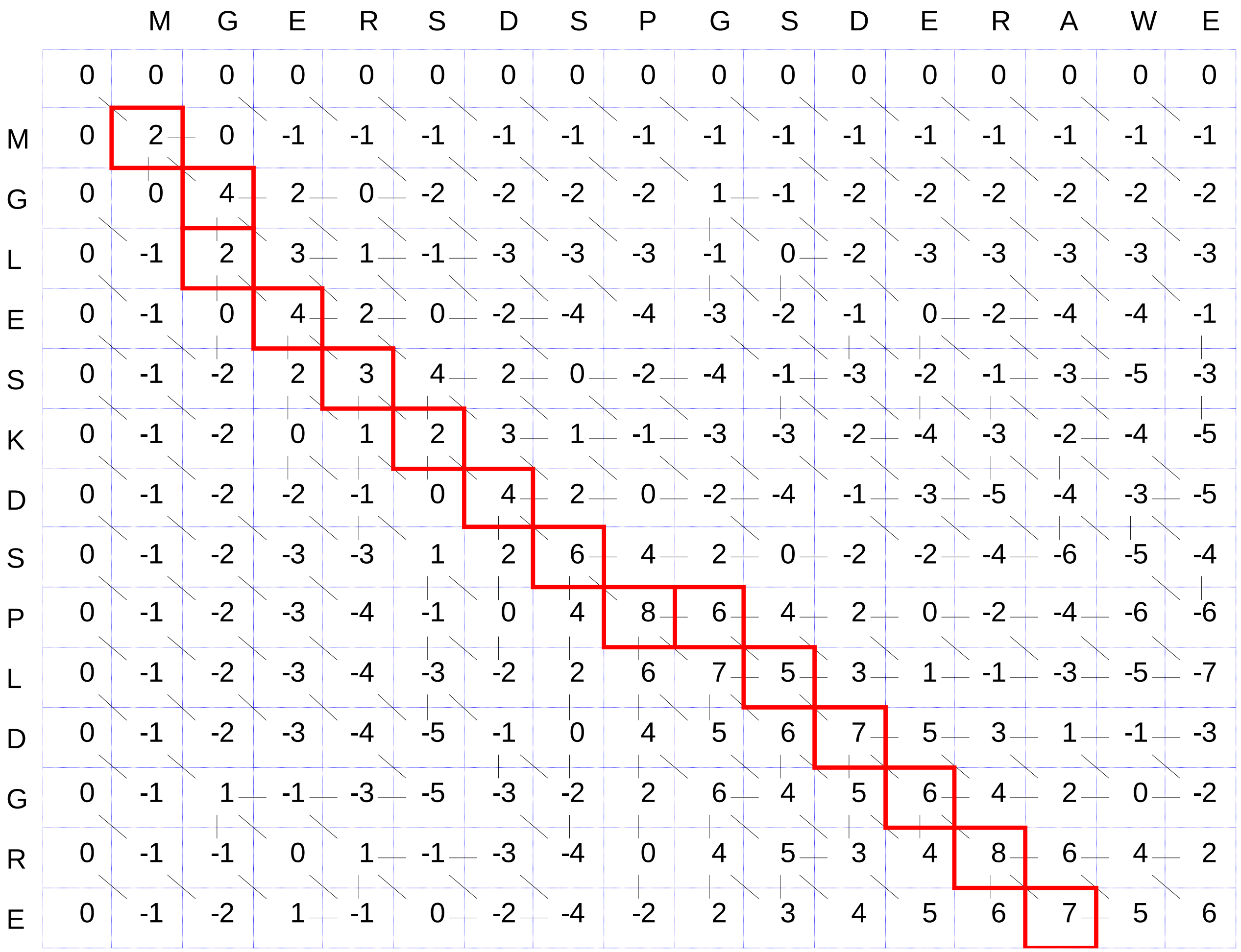
  - Furthermore, $A_{2,2}$ is reached best from $A_{1,1}$

Example



		M	G	E	R	S	D	S	P	G	S	D	E	R	A	W	E
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	2	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	0	0	4	2	0	-2	-2	-2	-2	1	-1	-2	-2	-2	-2	-2	-2
L	0	-1	2	3	1	-1	-3	-3	-3	-1	0	-2	-3	-3	-3	-3	-3
E	0	-1	0	4	2	0	-2	-4	-4	-3	-2	-1	0	-2	-4	-4	-1
S	0	-1	-2	2	3	4	2	0	-2	-4	-1	-3	-2	-1	-3	-5	-3
K	0	-1	-2	0	1	2	3	1	-1	-3	-3	-2	-4	-3	-2	-4	-5
D	0	-1	-2	-2	-1	0	4	2	0	-2	-4	-1	-3	-5	-4	-3	-5
S	0	-1	-2	-3	-3	1	2	6	4	2	0	-2	-2	-4	-6	-5	-4
P	0	-1	-2	-3	-4	-1	0	4	8	6	4	2	0	-2	-4	-6	-6
L	0	-1	-2	-3	-4	-3	-2	2	6	7	5	3	1	-1	-3	-5	-7
D	0	-1	-2	-3	-4	-5	-1	0	4	5	6	7	5	3	1	-1	-3
G	0	-1	1	-1	-3	-5	-3	-2	2	6	4	5	6	4	2	0	-2
R	0	-1	-1	0	1	-1	-3	-4	0	4	5	3	4	8	6	4	2
E	0	-1	-2	1	-1	0	-2	-4	-2	2	3	4	5	6	7	5	6

Example



		M	G	E	R	S	D	S	P	G	S	D	E	R	A	W	E
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	2	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
G	0	0	4	2	0	-2	-2	-2	-2	1	-1	-2	-2	-2	-2	-2	-2
L	0	-1	2	3	1	-1	-3	-3	-3	-1	0	-2	-3	-3	-3	-3	-3
E	0	-1	0	4	2	0	-2	-4	-4	-3	-2	-1	0	-2	-4	-4	-1
S	0	-1	-2	2	3	4	2	0	-2	-4	-1	-3	-2	-1	-3	-5	-3
K	0	-1	-2	0	1	2	3	1	-1	-3	-3	-2	-4	-3	-2	-4	-5
D	0	-1	-2	-2	-1	0	4	2	0	-2	-4	-1	-3	-5	-4	-3	-5
S	0	-1	-2	-3	-3	1	2	6	4	2	0	-2	-2	-4	-6	-5	-4
P	0	-1	-2	-3	-4	-1	0	4	8	6	4	2	0	-2	-4	-6	-6
L	0	-1	-2	-3	-4	-3	-2	2	6	7	5	3	1	-1	-3	-5	-7
D	0	-1	-2	-3	-4	-5	-1	0	4	5	6	7	5	3	1	-1	-3
G	0	-1	1	-1	-3	-5	-3	-2	2	6	4	5	6	4	2	0	-2
R	0	-1	-1	0	1	-1	-3	-4	0	4	5	3	4	8	6	4	2
E	0	-1	-2	1	-1	0	-2	-4	-2	2	3	4	5	6	7	5	6

Example

- The optimal alignment is therefore

```
MGLESKDSP_LDGRE
|| |  |||  | |
MG_ERSDSPGSDERAWE
```

- The score of the optimal alignment is 7
  - $A_{14,14} = 7$
  - This corresponds to an alignment with 8 matches, 5 mismatches and 2 gaps resulting in an alignment score of

$$8 \cdot 2 + 5 \cdot (-1) + 2 \cdot (-2) = 16 - 5 - 4 = 7$$

- Note:
  - The optimal alignment is not necessarily unique
  - Different alignments can be obtained with other choices of α, β, and ω

Embedding Prior Information into Sequence Alignment

- The alignment requires specification of two basic parameters:
  - Matching or replacement scores between two amino acids
  - Gap penalties
- Choices for these parameters can incorporate the rates at which the associated events are observed in real sequences
- The resulting alignment then provides the “most likely” correspondence given an observed sequence data
  - In terms of the evolutionary process via selection
  - This is tantamount to embedding the statistical structure of point mutations and indels into the alignment algorithm

Embedding Prior Information into Sequence Alignment

- Matching score between two amino acids:
  - In an ideal alignment, the two sequences would be identical, and all amino acids would pair up
  - Deviations from this ideal situation pairs up different amino acids
  - The objective of the alignment is
    - to match up the amino acids that are the most strongly preserved, and
    - to avoid the situations in which two amino acids that are highly unlikely to replace one another are aligned
  - The rates at which different amino acids can replace one another are quantified by amino acid scoring matrices
    - PAM
    - BLOSUM
    - …
  - Such scoring systems can be incorporated into the alignment algorithm by modifying the similarity function $S_{i,j}$:

$$S_{i,j} = PAM250_{P_i,Q_j}$$

Embedding Prior Information into Sequence Alignment

- Gap penalties:
  - The initial impetus is to penalize each and every missing amino acid equally
    - The contribution of each gap block is a linear function of its length L:
      $$W(L) = \omega L$$
  - Other alternatives, however, may be more adequate to capture the stochastic nature of mutations causing sequence indels
    - Consider the affine alternative given by
      $$W(L) = \omega_{open} + \omega_{ext}(L-1)$$
      where ω_{open} is the cost of the creation of the gap, and ω_{ext} is the cost of extending the gap over one more site
    - A more general gap cost function would start at 0 with no gaps, jump to a high cost with the opening, and incur decreasing costs with each extension
      $$W(L) = \omega \log_B(L+1)$$

BLAST: Basic Local Alignment Search Tool

- The dynamic programming strategy is guaranteed to provide the optimal alignment
  - For the choice of matching scores and gap penalties
- But it is costly to carry out in a multiple alignment setting where the sequences are long and many
- The BLAST algorithm provides a feasible alternative (Altschul et al., 1990)
  - Local matches to a word of fixed length are sought
  - When found, the match is extended until the similarity score starts to fall
  - The search is then repeated with the next word
- It provides very fast alignment with similarity scores close to the optimal alignment for those sequences that are similar to one another beyond a statistically meaningful threshold
  - Note that alignments are usually sought between evolutionarily related sequences anyway

Multiple Sequence Alignment

- Pairwise sequence alignment aligns any two sequences by placing gaps in most suitable locations
  - Suitable in the sense that
    - well-preserved nucleotides/amino acids are aligned, and
    - those that are not substituted for one another are not
- The alignment problem becomes complicated over a set of sequences
  - The most straightforward strategy is to pick one of the sequences as the reference and align all the rest to it
  - Then, all the gaps introduced onto the reference in all pairwise alignments can be effectuated in all the others
  - However,
    - This strategy depends on the choice of the reference sequence
    - The resulting alignments are highly likely to be suboptimal

Multiple Sequence Alignment

- Example:
  - Consider the following sequences:
    - Sequence 1: `MGLSKDSLDGE,`
    - Sequence 2: `MGLSKDSPLGR,`
    - Sequence 3: `MGLESKPLDE,`
    - Sequence 4: `MGLESKSPLGRE`
  - Pairwise alignments using the first sequence as the reference with $\alpha = 2$, $\beta = -1$, and $\omega = -1$ ➔

Sequences 1 and 2:

```
MGLSKDS_LDGE
||||||| | |
MGLSKDSPL_GR
```

Sequences 1 and 3:

```
MGL_SKDSLDGE
||| ||  || |
MGLESK_PLD_E
```

Sequences 1 and 4:

```
MGL_SKDS_LDG_E
||| || | | | |
MGLESK_SPL_GRE
```

Multiple Sequence Alignment

- Example (continued):
  - The first alignment stands as

```
MGLSKDS_LDGE
MGLSKDSPL_GR
```

  - Introducing the gaps in the second alignment provides

```
MGL_SKDS_LDGE
MGL_SKDSPL_GR
MGLESK_P_LD_E
```

  - Finally, with the gaps in the final alignment, we obtain

```
MGL_SKDS_LDG_E
MGL_SKDSPL_G_R
MGLESK_P_LD__E
MGLESK_SPL_GRE
```

Multiple Sequence Alignment

- A more viable alternative is provided by the progressive alignment strategy
  - Instead of picking a reference sequence, all sequence pairs are aligned, and the alignment scores obtained
  - The sequences are then **clustered** according to their alignment scores
    - A common approach is to convert alignment scores to distances via

$$d = -100 \log\left(\frac{S - S_{\mathrm{rand}}}{S_{\mathrm{ident}} - S_{\mathrm{rand}}}\right)$$

where S is the observed alignment score, S_{rand} is the expected alignment score for random sequences of the same length, and S_{ident} is the alignment score for identical sequences

    - Otherwise, any monotonically decreasing function of S can be used as well

Multiple Sequence Alignment

- When two separate clusters of sequences are joined in a single cluster, an **alignment algorithm for two sequence clusters** must be used
  - The only difference when aligning clusters instead of single sequences is the measurement of similarity scores $S_{i,j}$
  - A typical choice for scoring the alignment at a given site is to use the average of alignment scores between all possible combinations of amino acids observed in the two clusters at that site
- In addition, merging sequence clusters after sequence cluster alignment propagates the gaps obtained for clusters onto their respective sequences

Multiple Sequence Alignment

- Example:
  - Consider again the sequences `MGLSKDSLDGE`, `MGLSKDSPLGR`, and `MGLESKPLDE`
  - Pairwise alignment of these three sequences with one another provides the alignment scores
  - The alignment scores are converted to distances via

$$S_{\text{ident}} = \alpha L_{\text{aligned}} + \omega L_{\text{gap}}$$

and

$$S_{\text{rand}} = \left(\frac{\alpha}{N} + \frac{\beta(N-1)}{N}\right) L_{\text{aligned}} + \omega L_{\text{gap}}$$

  - Clusters are formed by merging the most similar sequences and sequence clusters

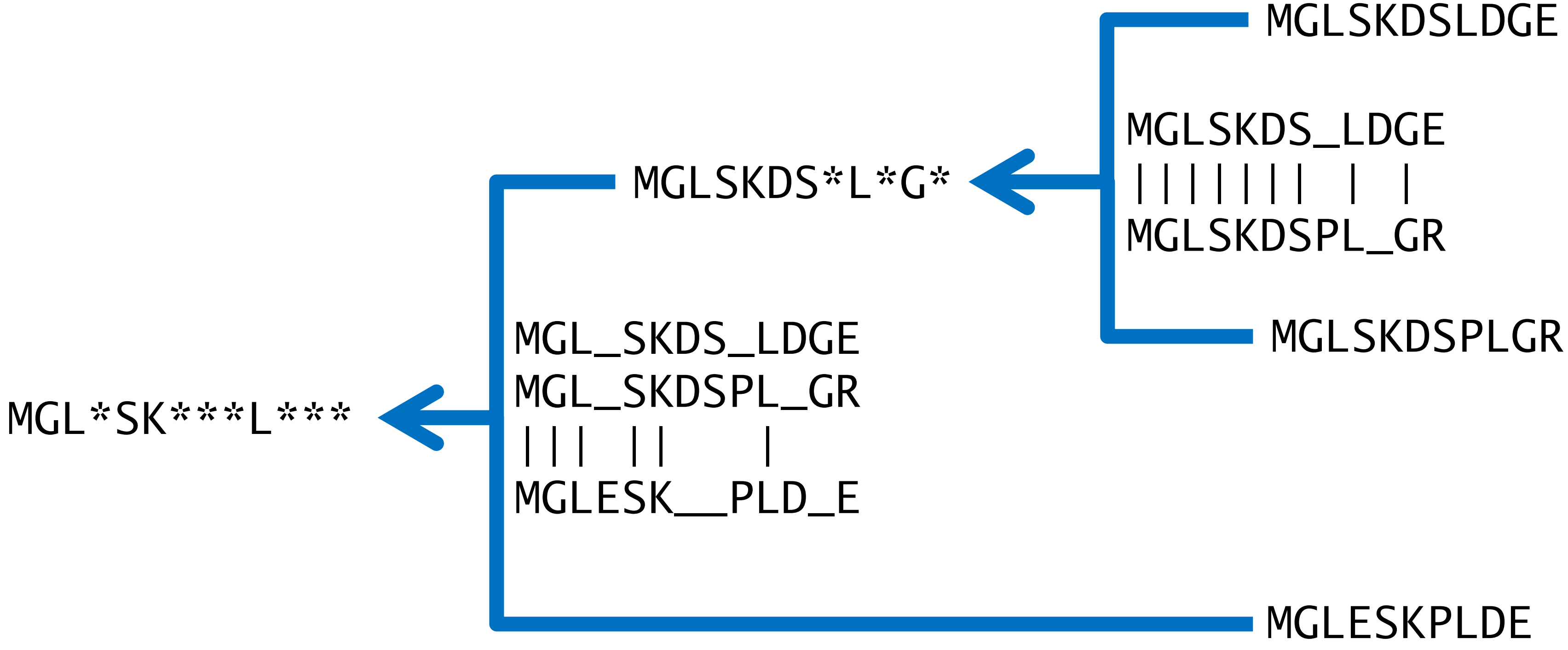
Alignment scores:

	SQ_1	SQ_2	SQ_3
SQ_1	22	15	12
SQ_2	15	22	9
SQ_3	12	9	20

Converted distances:

	SQ_1	SQ_2	SQ_3
SQ_1	0	3.8466	4.3203
SQ_2	3,8466	0	9.6928
SQ_3	4.3203	9.6928	0

Multiple Sequence Alignment



```
MGLSKDSLDGE
```

```
MGLSKDS_LDGE
||||||| | |
MGLSKDSPL_GR
```

```
MGLSKDSPLGR
```

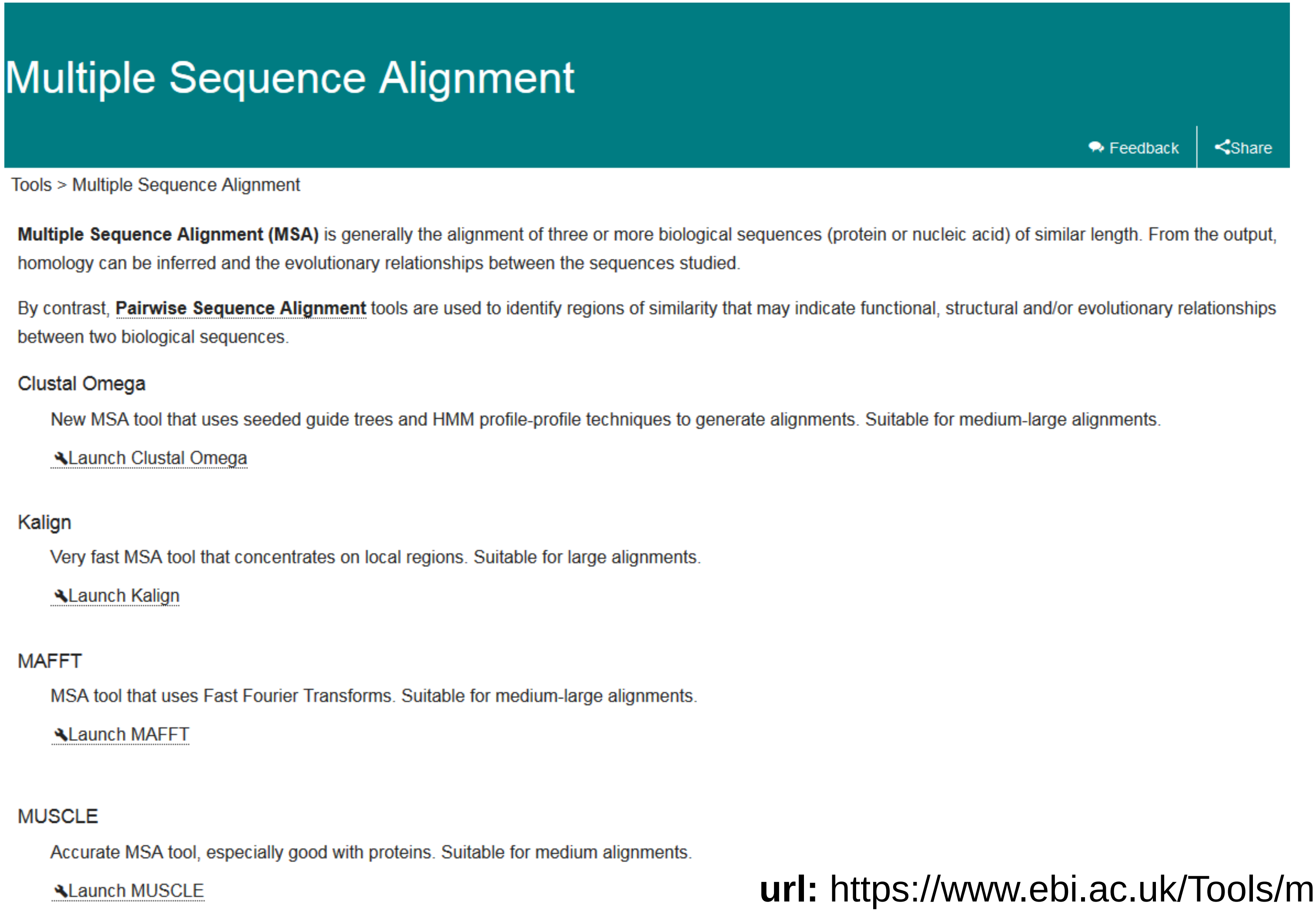
```
MGLSKDS*L*G*
```

```
MGL_SKDS_LDGE
MGL_SKDSPL_GR
||| ||   |
MGLESK__PLD_E
```

```
MGLESKPLDE
```

```
MGL*SK***L***
```

Multiple Sequence Alignment at EBI



Multiple Sequence Alignment

Tools > Multiple Sequence Alignment

Multiple Sequence Alignment (MSA) is generally the alignment of three or more biological sequences (protein or nucleic acid) of similar length. From the output, homology can be inferred and the evolutionary relationships between the sequences studied.

By contrast, **Pairwise Sequence Alignment** tools are used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences.

Clustal Omega

New MSA tool that uses seeded guide trees and HMM profile-profile techniques to generate alignments. Suitable for medium-large alignments.

Launch Clustal Omega

Kalign

Very fast MSA tool that concentrates on local regions. Suitable for large alignments.

Launch Kalign

MAFFT

MSA tool that uses Fast Fourier Transforms. Suitable for medium-large alignments.

Launch MAFFT

MUSCLE

Accurate MSA tool, especially good with proteins. Suitable for medium alignments.

Launch MUSCLE

url: https://www.ebi.ac.uk/Tools/msa/

Summary

- Similarity of sequences is essential when evaluating the evolutionary origins and relationships of different sequence fragments
  - Gene sequences
  - Amino acid sequences
- Sequence alignment algorithms determine how to superpose pairs or groups of sequences in a way to maximize an alignment score
  - The alignment algorithms use a priori parameters for matching characters as well as mismatch and gap penalties
  - The algorithms then determine where to put gaps so that the alignment score is optimized
    - Where the gaps are placed in respective sequences determines which letters/sites are aligned
- Multiple sequence alignments also provide a hierarchical grouping among the set of sequences
  - Most similar sequences are clustered together first
  - Eventually all sequences are merged into a single cluster