Introduction to Bioinformatics

Course NR. 22111

Pairwise alignment

Stefan Olevinskiy
s246026

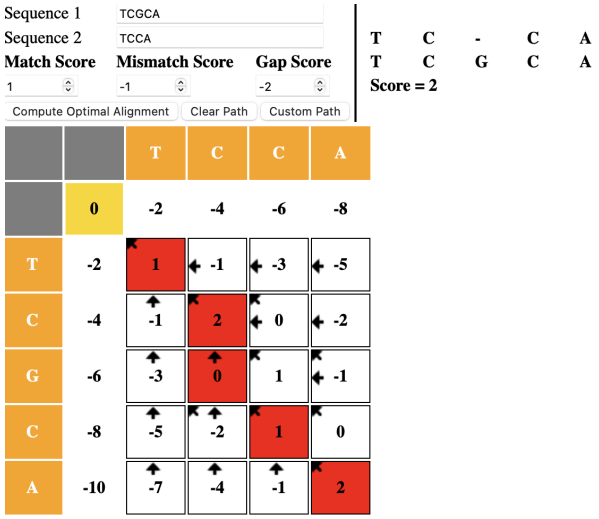
Polina Krasikova
s245850

Part 0 — interactive demos

Question 0:

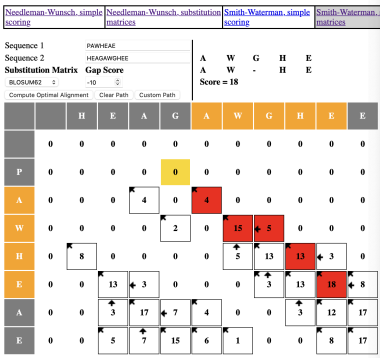
1. **Try replacing the two sequences with the sequences we used in the example in the lecture (slides 36-43). Does the website give you the same alignment matrix? Insert a screenshot of the resulting alignment matrix in your answer.**

   Matrix is the same:



2. **In "Smith-Waterman with substitution matrices", try replacing the two amino acid sequences with the sequences we used in the Smith-Waterman example in the lecture (slide 46). Does the website give you the same alignment matrix? Insert a screenshot of the resulting alignment matrix in your answer.**

   Resulting alignment is the same, but matrices are not identical, they have different numbers.



Part 1 — basic use

Question 1:

Which sequence format are the two sequences listed in?

Fasta

Question 2:

- Alignment score: 860.5
- Alignment length: 361
- % and fraction Identity (The value reported for "Identity" includes perfect matches only): 176/361 (48.8%)
- % and fraction Similarity (The value reported for "Similarity" includes perfect matches + "close" mismatches): 214/361 (59.3%)

Question 3:

Report the same values as above (Alignment score etc). Consider the alignments produced by the two different approaches: do YOU think one of them is more biologically relevant than the other, or do both contribute valuable information?

- *Local alignment (Smith–Waterman):* Length = 269, Identity = 176/269 (65.4%), Similarity = 214/269 (79.6%), Gaps = 0/269 (0.0%), Score = 916.0.
- *Global alignment (Needleman–Wunsch):* Length = 361, Identity = 176/361 (48.8%), Similarity = 214/361 (59.3%), Gaps = 92/361 (25.5%), Score = 860.5.

The local alignment highlights the conserved core region with high identity and no gaps, which is important for identifying functional or structural homology. The global alignment, on the other hand, shows overall sequence architecture, including the extra ~90 amino acids at the N-terminus of one protein. Therefore, both alignments provide valuable information: the local one for functional conservation, and the global one for full-length comparison.

Question 4:

P29600 is derived from the 3D structure of the mature protease, so the sequence starts directly with the active enzyme and lacks signal and pro-peptides. P41363, in contrast, is translated from the gene and protein sequencing, therefore it includes the signal peptide (1–24) and pro-peptide (25–93) before the mature protease. Both proteins are secreted enzymes. The difference reflects precursor vs. mature form of the same type of protease.

Question 5:

Based on what you've learned about the P41363 protein from the alignment to Savinase and from the data on the Uniprot site: do you think this could be used as an enzyme in washing powder? (Why? / why not?).

Yes — likely a good candidate. P41363 is a serine protease (S8 family), thermally stable and highly similar to Savinase, so it should have the right catalytic properties; however, its pH optimum and compatibility with detergent components must be confirmed or optimized before industrial use.

Part 2 — about gaps and dubious alignments

Question 6:

```
Length: 1256
Identity:     110/1256 ( 8.8%)
Similarity:   154/1256 (12.3%)
Gaps:         994/1256 (79.1%)
Score: -244.0
```

A large part of the alignment consists of gaps, which indicates that the two sequences cannot be aligned in a meaningful way. The negative alignment score (–244.0) shows that this result is worse than random and therefore not biologically relevant. In other words, the sequences are too different to produce a useful alignment. (Alignment not shown)

Question 7:

```
Length: 1290
Identity:      73/1290 ( 5.7%)
Similarity:   131/1290 (10.2%)
Gaps:        1062/1290 (82.3%)
```

```
Score: 158.5
```

Although an alignment was produced, most of it consists of gaps and the overall similarity is extremely low. This suggests that there is no real biological relationship between the proteins. (Alignment not shown)

Question 8:

```
# Length: 296
# Identity:      71/296 (24.0%)
# Similarity:   129/296 (43.6%)
# Gaps:          73/296 (24.7%)
# Score: 173.0
```

Question 9:

Do you think local or global alignment is best for finding similar parts of distantly related proteins? Why?

Local alignment is best for distantly related proteins because it focuses on the conserved functional core, ignoring unrelated regions. An alignment is “true” if it reflects a real evolutionary relationship.

Question 10:

How do the local alignments look? (What are the ranges of Alignment score, Alignment length, Identity, Similarity, and gap percentage)?

Shuffled Once:

```
# Length: 144
# Identity:      33/144 (22.9%)
# Similarity:    61/144 (42.4%)
# Gaps:          21/144 (14.6%)
# Score: 43.0
```

Shuffled Twice:

```
# Length: 181
```

```
# Identity:      39/181 (21.5%)
# Similarity:    60/181 (33.1%)
# Gaps:          46/181 (25.4%)
# Score: 47.0
```

Shuffled Thrice:

```
# Length: 170
# Identity:      39/170 (22.9%)
# Similarity:    56/170 (32.9%)
# Gaps:          47/170 (27.6%)
# Score: 50.5
```

Question 11:

Comparing the Savinase/shuffled alignment to the previous Savinase/Human Peptidase alignment - how will you judge the alignment with human peptidase now? (More/Less confidence in relation between the sequences?).

Clearly the previous score of 173 is a lot higher than either one of the three shuffled ones. Hence, we can be more confident in the relation between Savinase and Human peptide sequences.

Part 3 — about parameters

Question 12:

- **What are the alignment results (Length, score, gaps, identity, similarity)?**
- **How do alignment length and % identity depend on the BLOSUM number (compare also to your answer to question 8)?**

BLOSUM90:

```
# Length: 279
# Identity:      73/279 (26.2%)
# Similarity:   107/279 (38.4%)
# Gaps:          91/279 (32.6%)
# Score: 147.5
```

BLOSUM30:

```
# Length: 326
# Identity:      76/326 (23.3%)
# Similarity:   149/326 (45.7%)
# Gaps:          88/326 (27.0%)
# Score: 342.5
```

BLOSUM30 produces a much higher score at a higher length with lower gap % and higher similarity % at a similar identity.

This is because BLOSUM90 is built from alignments that are at least 90% similar, so it favors similarities and punishes dissimilarities heavily, whilst BLOSUM30 is built from alignments that are at least 30% similar and hence gives a higher score for distantly related sequences.

Question 13:

- **How do the quality parameters look this time (Length, score, gaps, identity, similarity)?**
- **Is this alignment biologically meaningful at all?**

```
# Length: 1254
# Identity:     192/1254 (15.3%)
# Similarity:   228/1254 (18.2%)
# Gaps:        1010/1254 (80.5%)
# Score: 896.576
```

Well, gap % is extremely high, as expected. Score and length are extremely high, also as expected. Similarity is lower and identity is lower. Obviously, the alignment is not biologically meaningful, despite the high score.

Question 14:

The gap is “ALIGNE” :)

```
GLB7A_CHITH        1 MKFFAVLALCIVGAIASPLSADQAALVKSTWAQVRNSEVEILAAVFTAYP     50
                     |||. :|||||:  |.||.||.||..||:||:.:|:...|.||.|||.|.|
GLBE_CHITH         1 MKFI-ILALCV--AAASALSGDQIGLVQSTYGKVKGDSVGILYAVFKADP     47
```

```
GLB7A_CHITH       51 DIQARFPQFAGKDVASIKDTGAFATHAGRIVGFVSEIIALIGNESNAPAV    100
                     .|||.||||.|||:.:||....|:||||||||||:..:|      .:.|.:
GLBE_CHITH        48 TIQAAFPQFVGKDLDAIKGGAEFSTHAGRIVGFLGGVI------DDLPNI     91

GLB7A_CHITH      101 QTLVGQLAASHKARGISQAQFNEFRAGLVSYVSSNVAWNAAAESAWTAGL    150
                     ...|..|.|:||.||::.||||.|||..::|:..:|.:.||.|:||.|..
GLBE_CHITH        92 GKHVDALVATHKPRGVTHAQFNNFRAAFIAYLKGHVDYTAAVEAAWGATF    141

GLB7A_CHITH      151 DNIFGLLFAAL    161
                     |..||.:||.:
GLBE_CHITH       142 DAFFGAVFAKM    152
```