Experiment 5.1: Pairwise sequence alignment using EMBOSS Needle (global alignment) and EMBOSS Water (local alignment)

AIM

To study pairwise sequence alignment using EMBOSS Needle (global alignment) and EMBOSS Water (local alignment) tools, and to understand how substitution matrices and gap penalties affect alignment outcomes for DNA and protein sequences.

THEORY

Introduction to Sequence Alignment

Sequence alignment is a fundamental technique in bioinformatics that involves arranging DNA, RNA, or protein sequences to identify regions of similarity that may indicate functional, structural, or evolutionary relationships between sequences. The underlying principle of sequence alignment is based on the assumption that sequences with significant similarity share common ancestry, known as homology, or have similar biological functions due to convergent evolution. This technique has become indispensable in modern biology for applications ranging from identifying disease-causing mutations to understanding evolutionary relationships between species and predicting protein structure and function.

Global versus Local Alignment Algorithms

The two primary approaches to pairwise sequence alignment are global and local alignment, each serving different biological purposes. Global alignment, implemented through the Needleman-Wunsch algorithm developed in 1970, attempts to align sequences from end to end across their entire length. This algorithm, available in EMBOSS Needle, uses dynamic programming with a time complexity of O(mn) where m and n represent the lengths of the two sequences being aligned. Global alignment is most appropriate when comparing sequences of similar length that are expected to be homologous across their full length, such as orthologous proteins from different species or allelic variants of the same gene. The algorithm forces alignment of all residues in both sequences, which may introduce numerous gaps when sequences have different lengths or contain non-homologous regions.

In contrast, local alignment, implemented through the Smith-Waterman algorithm developed in 1981, identifies and aligns only the best matching subsequences between two sequences. This algorithm,

available in EMBOSS Water, is a modification of the Needleman-Wunsch algorithm where negative scores are set to zero, allowing the alignment to start and end at any position where positive scores are obtained. Local alignment is particularly useful when searching for conserved domains or motifs within larger sequences, identifying regions of similarity between sequences of different lengths, or when comparing sequences that may contain non-homologous regions. The ability to find optimal local matches without forcing full-length alignment makes this approach invaluable for database searches and domain identification.

Scoring Systems in Sequence Alignment

The quality and biological relevance of sequence alignments depend critically on the scoring systems used to evaluate matches, mismatches, and gaps. For DNA sequences, scoring is relatively straightforward, typically using simple identity matrices where matches receive positive scores and mismatches receive negative scores. Some models also distinguish between transitions (purine to purine or pyrimidine to pyrimidine changes) and transversions (purine to pyrimidine changes), as transitions occur more frequently in evolution.

Protein sequence alignment requires more sophisticated scoring systems due to the biochemical properties of amino acids and their substitution patterns in evolution. Two major families of substitution matrices are widely used: PAM and BLOSUM matrices. The PAM (Point Accepted Mutation) matrices, developed by Margaret Dayhoff in 1978, are based on evolutionary models derived from closely related proteins. PAM matrices are extrapolated from observed mutations in proteins that are 85% identical, with the number indicating the amount of evolutionary distance; PAM1 represents 1% accepted mutations, PAM120 represents 120% mutations (multiple substitutions at the same site), and PAM250 represents 250% mutations suitable for distantly related sequences. Higher PAM numbers indicate greater evolutionary distance, making PAM250 appropriate for sequences that diverged long ago, while PAM30 or PAM70 are better for closely related sequences.

The BLOSUM (BLOcks SUbstitution Matrix) series, developed by Henikoff and Henikoff in 1992, represents an alternative approach based on observed substitution frequencies in conserved protein blocks from related sequences without evolutionary extrapolation. The number in each BLOSUM matrix indicates the clustering percentage used in its construction; BLOSUM80 is derived from sequences clustered at 80% identity or higher, making it suitable for comparing very similar sequences or those that have diverged recently. BLOSUM62, constructed from sequences clustered at approximately 62% identity, serves as the default matrix for most applications as it performs well across a moderate range of evolutionary distances. BLOSUM45 is designed for distantly related sequences with less than 50% identity, where detecting weak signals of homology requires allowing more diverse substitutions. The key difference between PAM and BLOSUM is that PAM matrices are based on an evolutionary model and extrapolated to longer times, while BLOSUM matrices are based on observed alignments at different similarity levels. In practice, BLOSUM62 is roughly equivalent to PAM160, and BLOSUM45 corresponds approximately to PAM250, though BLOSUM matrices generally perform better for database searches and local alignments.

Gap Penalties and Their Biological Significance

Gaps in sequence alignments represent insertion or deletion events (collectively called indels) that have occurred during evolution. The penalty system for gaps significantly influences alignment quality and biological interpretation. The affine gap penalty model, which is standard in most alignment programs, uses two components: the gap opening penalty (GOP) represents the cost of starting a new gap and typically ranges from 10 to 15 for protein sequences, while the gap extension penalty (GEP) represents the cost of extending an existing gap and is usually much lower, ranging from 0.5 to 2. This model reflects the biological reality that a single mutational event can delete or insert multiple consecutive residues, making it more likely to have fewer, longer gaps than many scattered single-residue gaps.

The total gap penalty is calculated as GOP + (GEP × gap_length), and the choice of these parameters profoundly affects alignment outcomes. High gap penalties, such as 15 for opening and 2 for extension, minimize the number of gaps introduced, forcing the algorithm to align residues even when they are dissimilar. This approach works well for very similar sequences where few indels are expected. Conversely, low gap penalties, such as 5 for opening and 0.1 for extension, allow more gaps to be introduced, accommodating sequences that have undergone numerous insertion and deletion events. This setting is more appropriate for divergent sequences or when aligning sequences with known structural variations.

EXERCISE : Part A: DNA Sequence Alignment

Sequences in fasta format

>Gene_X_Human (150 bp)

ATGGCAAGCCTGACCCTGAAGCAGATCCGCATCCACACCAAGGGCAAGCAGGAGGTC
AGCCTGAGCCAGCGCCGCGGCAAGAGCAAGCCCAAGGCCAGCAGCCAGGTCACCAT
CGAGAACCTGCAGCTGGATGTCAGCGCCCTGGAGAAG

>Gene_X_Mouse (148 bp)

ATGGCAAGCCTGACCCTGAAGCAGATCCGCATCCACACCAAGGGCCAGCAGGAGGTC
AGCCTGAGCCAGCGCCGGGGCAAGAGCAAGCCCAAGGCCAGCAGCCAGGTCACCAT
TGAGAACCTGCAGCTGGATGTCAGCGCCCTGGAGAAG

>Gene_X_Zebrafish (155 bp)

ATGGCGTCCCTGACCCTCAAACAGATCCGGATCCACACCAAGGGCAAGCAGGAGGTG
TCTCTGAGTCAGCGCAGAGGCAAGAGCAAGCCGAAGGCCAGCAGCCAGGTCACCAT
TGAGAACCTGCAGCTGGATGTGTCCGCTCTGGAGAAGACCTTC

>Partial_Homolog (75 bp)

AAGGGCAAGCAGGAGGTCAGCCTGAGCCAGCGCCGCGGCAAGAGCAAGCCCAAGG
CCAGCAGCCAGGTCACCATC

Task A1: DNA Global Alignment with Gap Penalty Testing

1. **Go to:** https://www.ebi.ac.uk/jdispatcher/psa/emboss_needle (if this doesn't work, just google "EMBOSS Needle)
2. **Select:** "DNA" for sequence type

A1.1. Default Parameters

- **Input:** Gene_X_Human vs Gene_X_Zebrafish
- **Parameters:** Gap open: 10, Gap extend: 0.5 (check parameters by clicking on more options)
- **Matrix:** DNAfull
- **Submit and record:** Score, Identity%, Gaps and Length in a tabular form. Also, provide the screenshot of the alignment

A1.2. High Gap Penalty: Gap open: 25, Gap extend: 5

A1.3. Low Gap Penalty: Gap open: 5, Gap extend: 0.1

Record the data and image for A1.2 and A1.3 as well. (Use same sequences as in A1.1.)

Task A2: DNA Local Alignment - Finding Conserved Regions

1. **Go to:** https://www.ebi.ac.uk/Tools/psa/emboss_water/

2. **Input:**
   Sequence 1: Gene_X_Zebrafish (full length)
   Sequence 2: Partial_Homolog (75 bp fragment)

3. **Use default parameters**

Questions for Discussion for Part A

Q1: In the global alignment, which gap penalty setting produced the most biologically meaningful alignment? Why?

Q2: What region of the zebrafish gene does the partial homolog match? (Positions ____ to ____)

Q3: Would Needle or Water be better for finding where a PCR primer matches in a gene?

- Needle (global)
- Water (local)

EXERCISE: Part B: Protein Sequence Alignment

Sequences in fasta format

>Cytochrome_C_Human (105 aa)

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK
GIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

>Cytochrome_C_Rhesus (105 aa)

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNK
GIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

>Cytochrome_C_Mouse (105 aa)

MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAAGFSYTDANKNK
GITWGEDTLMEYLENPKKYIPGTKMIFAGIKKKGERADLIAYLKKATNE

>Cytochrome_C_Chicken (104 aa)

MGDIEKGKKIFVQKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAEGFSYTDANKNK
GITWGEDTLMEYLENPKKYIPGTKMIFAGIKKTEREDLIAYLKKATNE

>Cytochrome_C_Yeast (108 aa)

MTEFKAGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDA
NIKKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKACE

>Unknown_Fragment (45 aa)

CSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIW

Task B1: Testing Substitution Matrices

Using EMBOSS Needle

1. **Select:** "Protein"
2. **Input:** Cytochrome_C_Human vs Cytochrome_C_Mouse
3. **Parameters:**
   Matrix: BLOSUM62
   Gap penalties: 10/0.5 (default)
4. **Record:** Score, Identity% and Similarity% in a tabular form. Also, provide the screenshot of the alignment
5. **Repeat this with the following Matrices as well:**
   BLOSUM80
   BLOSUM45
   PAM250

   Discussion Question: Which matrix gave the highest score for these closely related sequences? Why?

Task B2: Local Alignment for Protein Domain Finding

Task B2.1 Compare the Unknown_Fragment against ALL species (Domain conservation across species)

1. Run Water with Unknown_Fragment vs each Cytochrome_C sequence
2. Use BLOSUM62 for all comparisons
3. Record:

Species Compared	Score	Matched Region	Identity %	Similarity %
vs Human		Positions ____ to ____		
vs Mouse		Positions ____ to ____		
vs Chicken		Positions ____ to ____		
vs Yeast		Positions ____ to ____		

Task B2.2 Matrix effect on domain detection

1. Switch to EMBOSS Water
2. Input:

   Sequence 1: Cytochrome_C_Yeast (full)

   Sequence 2: Unknown_Fragment
3. Test with three matrices: BLOSUM62, BLOSUM45, and PAM250

4. Record:

Matrix	Score	Matched Region	Identity %
BLOSUM62		Positions _____ to _____	
BLOSUM45		Positions _____ to _____	
PAM250		Positions _____ to _____	

Discussion Question:

Is the fragment found at the same position in all species? What does this suggest about its functional importance?

Would Needle (global alignment) be appropriate for this domain-finding task or not? Explain.