sequence

alignment

Team:

Abdelrhman Salah Salem	211001698
Youssef Salem Hassan	211000582
Khadija Hossam El-din	211001375
Nourhan Sayed	211001780
Pola Alaa	211001681

Dr. Shereen

Sequence alignment definition

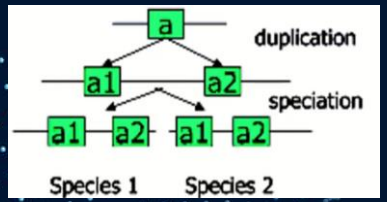
Sequence alignment involves comparing biological sequences like DNA, RNA, or protein sequences to identify similarities and differences.

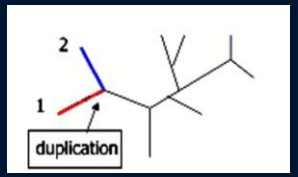
Origin of identical sequences

01

duplication

mechanism that copies a sequence within a genome, allowing one copy to continue its original function while the other to evolve new functions.

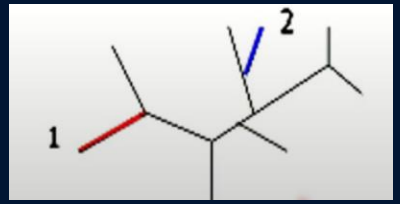




02

Convergence

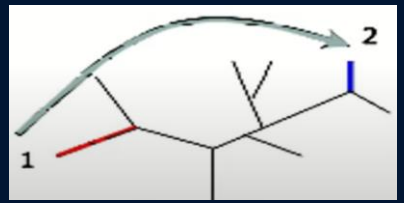
process of similar sequences evolving independently in different organisms..



03

Transfer

movement of genetic material between organisms through processes such as horizontal gene transfer, viral infection, or hybridization, which can introduce new or similar sequences into the recipient genome.



Comparing of sequence

Alignment of the sequences is performed and mutations that have occurred since the divergence of the common ancestor are identified.

Comparing of sequence:

Insertion:

when one sequence has an additional residue or compared to the other sequence

Deletion:

is a situation where one sequence is missing a residue or a gap compared to the other sequence.

Scoring technique

Match: (score)

occurs when two residues or nucleotides in different sequences are identical in terms of their chemical identity.

Mismatch

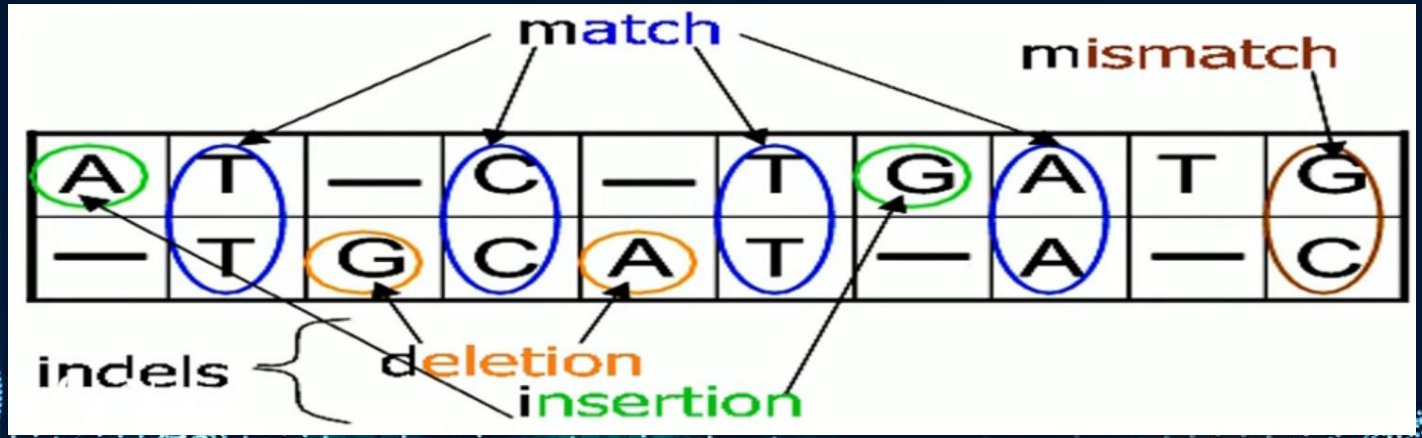
occurs when two residues or nucleotides in different sequences are different in terms of their chemical identity

Gap

refers to a position where one sequence has an insertion or deletion of one or more residues or nucleotides compared to the other sequence and represented by a dash ("-") or a dot (".") symbol in the alignment

Example:
There are two sequences

TGCATAC ATCTGATG



Homologs:

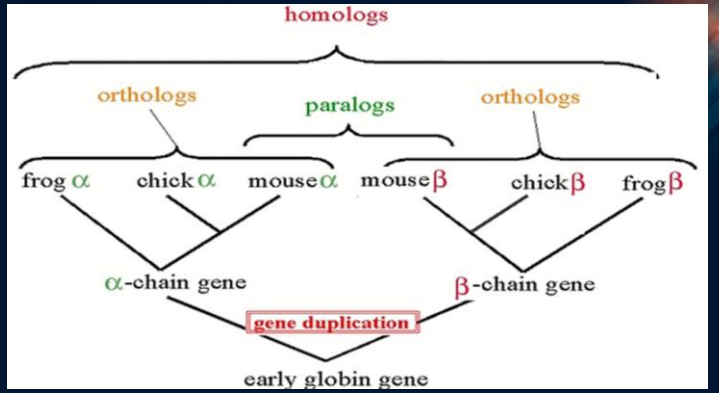
genes or proteins share a common ancestor and have similar sequences and functions. The degree of sequence similarity and evolutionary relationship can be used to classify them into different types of homologs

Orthologs:

homologous genes or proteins that are present in different species and have diverged through speciation. They typically have similar functions and can be used to infer evolutionary relationships between species.

Paralogs

homologous genes or proteins that are present within the same species and have diverged through gene duplication. Paralogs can have similar or different functions and can contribute to the evolution of new gene functions or gene families.



Pairwaise alignment

Global

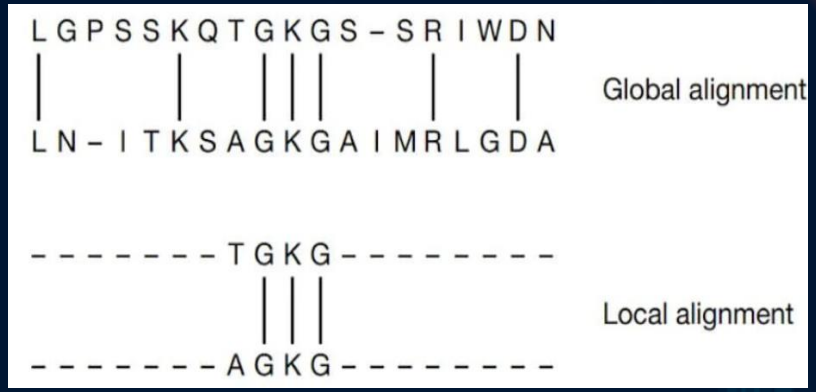
aligns two entire sequences from start to end, by comparing their entire lengths and introducing gaps as necessary to achieve the best overall alignment. It is useful for comparing sequences that have similar lengths and high degree of similarity.

- **Used by Needleman-Wunsch algorithm**
- **EX: clustal**

local

identifies a region of similarity within two sequences. The algorithm searches for a region of the two sequences that has the highest similarity score, even if the rest of the sequences do not match well. The alignment can contain gaps or mismatches as long as the region of similarity is maximized.

- **Used by Smith-Waterman algorithm.**
- **EX: Blast**



brute force approach:

involves comparing every possible alignment of two sequences and selecting the optimal alignment based on a scoring system. This approach is computationally expensive and is not practical for aligning long sequences.

- This approach has an exponential.
- Time complexity of **O(k^(n+m))**

1) computationally expensive.
2) not practical for aligning long sequences

Greedy approach

- is a simple, intuitive algorithm that is used in optimization problems. The algorithm makes the optimal choice at each step as it attempts to find the overall optimal way to solve the entire problem. At each step, a greedy algorithm chooses the locally optimal option based on some criteria and without considering the full problem.
- Complexity is o(n)

A greedy algorithm may align them as: ATCG and ATCA and want to align them

A|A (A matches with a high score so align them)
T|T
C|C
G|-
(G doesn't match C, so leave a gap)
This results in an alignment score of 3 for the matched characters.However, the globally optimal alignment would be:
A|A
T|T
C|C
G|A
With an alignment score of 4.
So the greedy approach produces a suboptimal solution because it only optimizes the alignment at each position.

Dynamic Programing :

The algorithm finds the alignment by searching for the highest scores in the matrix. Dynamic programming solves the original problem by breaking it down into smaller, independent sub-problems. This approach is used in many different aspects of computer science. The Needleman-Wunsch and Smith-Waterman algorithms for sequence alignment are defined using dynamic programming.

Steps of the dynamic programming matrix:

1.Initialization of the matrix with the scores possible.
2.Matrix filling with maximum scores.
3.Trace back the residues for appropriate alignment.

Time complexity=o(mn)

where m and n are the lengths of the two sequences being aligned. This is because the dynamic programming algorithm involves filling in an m x n matrix of scores to compute the optimal alignment score.

The Needleman-Wunsch and Smith-Waterman algorithms use a scoring system for optimal sequence alignment. Scoring matrices are used, which are relatively simple for nucleotide sequences because the mutation frequency for all bases is equal. Positive values are assigned for matches and negative values for mismatches. These assumption-based scores can be used for scoring matrices. Predefined matrices like PAM and BLOSUM are commonly used for amino acid substitutions.

Scoring matrices

- **PAM Matrices:**
  The PAM matrix quantifies evolutionary distances between protein sequences. One PAM unit (PAM1) represents a 1% change in amino acid residues, meaning that 99% of the sequence remains the same.

- **BLOSUM:**
  uses conserved regions to calculate substitution scores for amino acids. These matrices are actual percentage identity values and depend on similarity. For example, BLOSUM 62 indicates 62% similarity between sequences.

- **Gap score or gap penalty:**
  defines the penalty given to an alignment when an insertion or deletion occurs. In cases where continuous gaps are present, a linear gap penalty may not be appropriate. Therefore, gap open and gap extension penalties have been introduced for continuous gaps (five or more). The gap open penalty is applied at the start of the gap, followed by a gap extension penalty that is less severe than the gap open penalty. Typical values for gap open and gap extension penalties are -12 and -4, respectively.

The best alignment:

Has

High Match

Low Mismatch

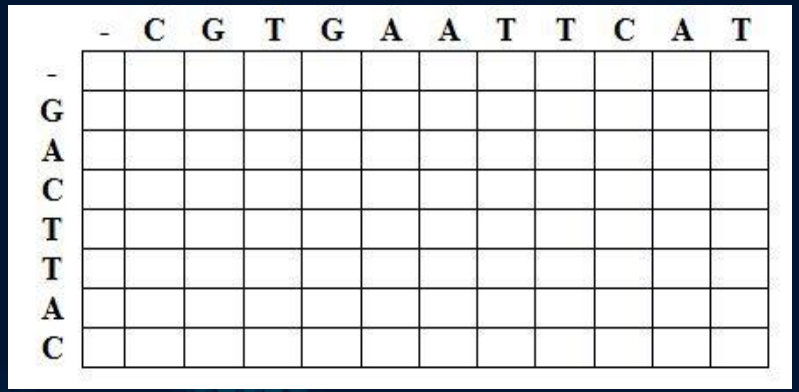
Low Gap

Needleman -Wunsch Algorithm

consider the two given sequences:

CGTGAATTCAT (sequence #1) , GACTTAC (sequence #2)

- There are two sequences, one with 11 nucleotides/amino acids and the other with 7. To create a matrix for alignment, an extra row and column are added to accommodate gaps at the beginning of the matrix. The resulting matrix has A+1 columns and B+1 rows.
- Once the initial matrix is created, a scoring schema must be defined. The simplest schema assigns a score of 1 if two nucleotides or amino acids at the same position in the two sequences match (S(i,j)= 1), and a score of -1 if they do not match (S(i,j)= -1). The penalty for introducing a gap is assumed to be -1.



1.Initialization Step

- When a gap penalty is assumed, the first row and first column of the matrix can be filled with 0.
  The gap score can be added to the previous cell of the row or column if a gap is introduced.

	-	C	G	T	G	A	A	T	T	C	A	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11
G	-1											
A	-2											
C	-3											
T	-4											
T	-5											
A	-6											
C	-7											

2.Matrix Fill Step

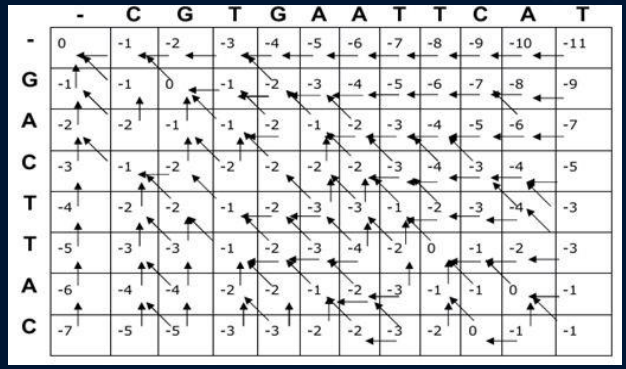
1. The second step of the algorithm involves filling the matrix by calculating the maximum score for each cell. To do this, the neighboring scores (diagonal, left, and right) of the current position are needed. The match or mismatch score is added to the diagonal value, while the gap score is added to the other neighboring values. The maximum value among the three is then taken and used to fill the ith and jth position with the score obtained. The equation for calculating the maximum score can be expressed as [M(i-1,j-1)+S(i,j),M(i,j-1)+w,M(i-1,j)+w].

$$M_{i,j} = \text{Maximum}\left[M_{i-1,j-1} + S_{i,j}, M_{i,j-1} + W, M_{i-1,j} + W\right]$$

2. To score the first position of the matrix (M1,1), the formulae described earlier can be used. If the first nucleotides or amino acids in the two sequences are 'G' and 'C', which are mismatching residues, then the score (S(i,j)) would be -1.

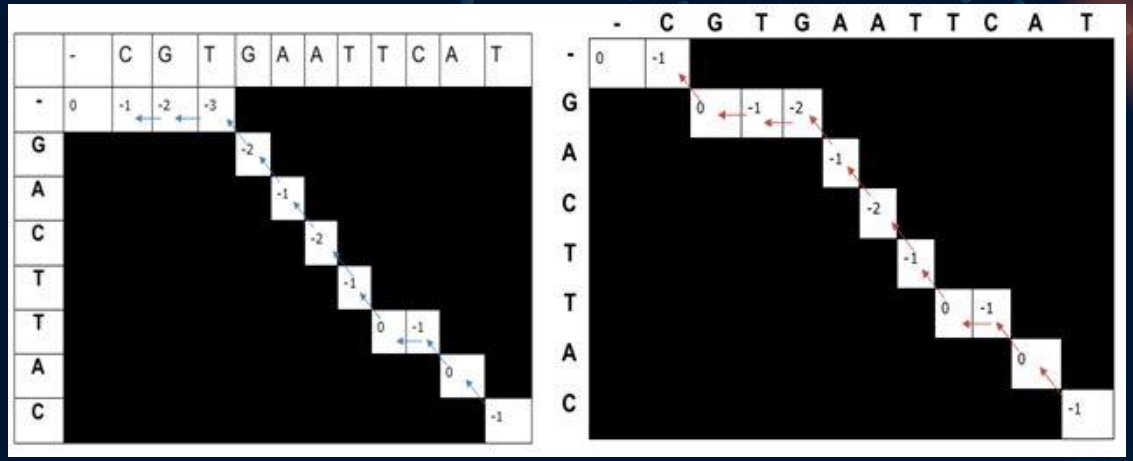
$$\begin{aligned} M_{1,1} &= Max\left[M_{0,0} + S_{1,1}, M_{1,0} + W, M_{0,1} + W\right] \\ &= Max\left[0+(-1),\ 0+(-1), 0+(-1)\right] \\ &= Max\left[-1,-1,-1\right] \\ &= -1 \end{aligned}$$

3. The score of -1 obtained in the previous step is placed in position i,j (1,1) of the scoring matrix. Using the same method and equation, the remaining rows and columns are filled. Back pointers are placed in each cell to indicate the predecessor cell from which the maximum score was obtained .

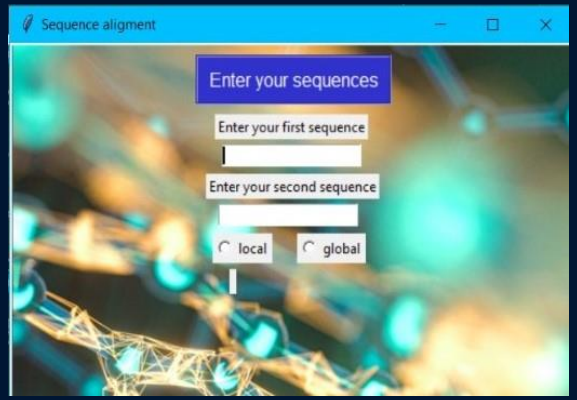


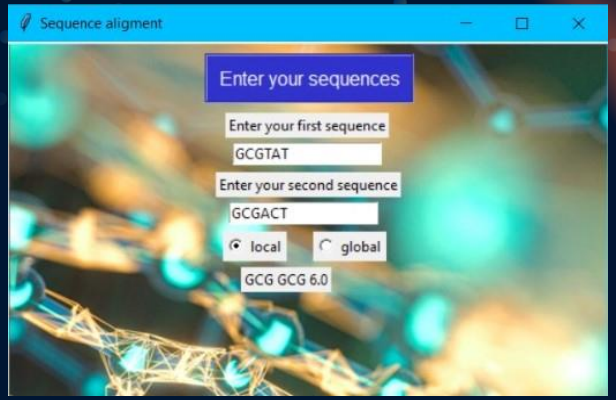
3.Trace back Step

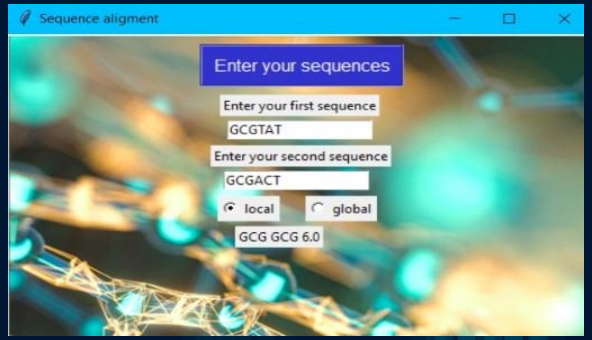
- The final step in the algorithm is to trace back for the best alignment.
- It's important to note that there may be multiple possible alignments between the two sequences.
- By continuing the trace back step using the method described earlier, one can reach the 0th row and 0th column.
- The alignment of the two sequences can be found by following the steps described above.
- The best alignment among the possible alignments can be identified by using the maximum alignment score, which may be user-defined based on the match, mismatch, and gap penalties.

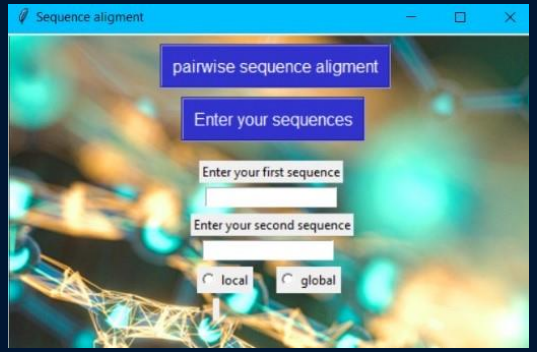


Sample output









References:

https://brilliant.org/wiki/greedy-algorithm/

https://medium.com/@hasini.dbv/pairwise-sequence-alignment-global-and-local-alignments-5ebacf83c752

THANK YOU☺🧡