COMSATS University
Park Road, Chak Shahzad, Islamabad Pakistan

BIO310 – Introduction to Bioinformatics

Assignment 2

Sequence Alignment

By

Waleed Butt CU/SP18-BCS-170/ISB

Instructor

Dr. Muhammad Sajjad

Class/Section: BCS-8B

Submission Date: 11/10/2021

Bachelor of Science in Computer Science (2018-2022)

1 Sequences Alignment

In order to get useful information from the biological sequences, the sequences are compared to show optimal similarity. This process is called sequence alignment. DNA sequence alignments, RNA sequence alignments and protein sequence alignments are routinely performed. Sequence Alignment is very important as many researchers are storing new biological sequences in databases. This exponential rise in dataset makes comparison essential for functional and evolutionary inference of new protein with protein already existing in database. Sequence Alignment is relatively straightforward computational problem. There can be many possible alignments. There may exist more than one solution.

There are multiple methods for sequence alignment. One method of sequence alignment is pairwise alignment. The sequences are shifted relative to other to find the position where maximum matches are found. Global Alignment or Local Alignment are many of the two strategies often used. In global alignment the sequences of almost equal length are aligned from beginning to end to find the best possible alignment. While in case of local alignment, local regions of sequence are aligned ignoring rest of the sequence.

In order to align three or more than three sequences, Multiple Sequence Alignment (MSA) is used. MSA can be done by hand, Dot Plot (Software), Heuristically using BLAST and FASTA or dynamic programming.

Scoring System is used to decide the best alignment. Symbol comparison table is used to assign a numerical value to each symbol pairing. There are penalties that reduce the score base on gaps. Opening penalties are the cost for introducing a gap and Extension for increasing the gap. Dot Plot gives an overview of all possible alignments. Each Sequence is placed on respective axis and common points are plotted like a graph. There are many software to create dot plots like ANACON, D-Genies, Dotlet and Dotmatcher etc.

In Protein Scoring system a table of values that describe the probability of a residue pair occurring in alignment, is used commonly known as scoring matrix. There are two famous scoring matrices PAM (Percent Accepted Mutation) Matrix and BLOSUM (Blocks Substitution Matrix). In the PAM matrix the as the number increases so does evolutionary distance while it is the reverse it the BLOSUM. The PAM-1 matrix reflects an average change of 1% of all amino acid positions. PAM matrices for larger evolutionary distances can be extrapolated from the PAM-1 matrix by multiplication. Greater numbers mean bigger evolutionary distance. Clusters are counted as a single sequence. Different BLOSUM matrices differ in the percentage of sequence identity used in clustering. The number in the matrix name (e.g. 62 in BLOSUM62) refers to the percentage of sequence identity used to build the matrix. Greater numbers mean smaller evolutionary distance.

Generally, BLOSUM matrices perform better than PAM matrices for local similarity searches (Henikoff & Henikoff, 1993). When comparing closely related proteins one should use lower PAM or higher BLOSUM matrices, For distantly related proteins higher PAM or lower BLOSUM matrices. For database searching the commonly used matrix is BLOSUM62.

One other approach for sequence alignment is Dynamic Programming. In dynamic programming algorithms are written that give the optimal alignment. The algorithm creates a alignment path matrix. It then stepwise calculates the score values. And then backtracks to find the optimal path. Needleman-Wunsch (1970) provided the first automatic method using D-P to find global alignment.

The score of the best possible alignment that ends at a given pair of positions (i,j) in the two sequences is the score of the best alignment previous to those two positions PLUS the score for aligning those two positions. The algorithms are different for global and local alignments. Global alignment algorithms

start at the beginning of two sequences and add gaps to each until the end of one is reached. Local alignment algorithms finds the region (or regions) of highest similarity between two sequences and build the alignment outward from there. Global algorithms are often not effective for highly diverged sequences i.e., do not reflect the biological reality that two sequences may only share limited regions of conserved sequence. Sometimes two sequences may be derived from ancient recombination events where only a single functional domain is shared. Global methods are useful when you want to force two sequences to align over their entire length