**NAME: BASHIR AL-AMIN MOHAMMED**

**COURSE CODE/NAME: CSC 262/ INFORMATION AND BIOINFORMATICS**

**DEPARTMENT: PHYSIOLOGY**

**MATRIC NUMBER: 18/MHS01/108**

**ASSIGNMENT**

**1(A) CRITERION OF DATABASE SEARCHING:**

1. Sensitivity
2. Selectivity/Specificity
3. Speed

* **SENSITIVITY:** This refers to the ability to find as many correct hits as possible. It is measured by the extent of inclusion of correctly identified sequence members of the same family. These correct hits are considered “true positives”**.**
* **SELECTIVITY/SPECIFICITY:** This is the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered “false positives.”
* **SPEED:** The third criterion, which is the time it takes to get results from database searches. Depending on the size of the database, speed sometimes can be a primary concern. Ideally, one wants to have the greatest sensitivity, selectivity, and speed in database searches.

**1(B) BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST):**

The heuristicalgorithm of BLAST locates all common three-letter words between the sequence of interest and the hit sequence or sequences from the database. This result will then be used to build an alignment. After making words for the sequence of interest, the rest of the words are also assembled.

**2(A)** **DIFFERENCE BETWEEN PAM AND BLOSUM MATRICES**

**POINT ACCEPTED MUTATION (PAM) MATRIX**

A point accepted mutation, also known as a PAM is the replacement of a single amino acid in the primary structure of a protein with another single amino acid, which is accepted by the processes of natural selection.

A PAM matrix is a matrix where each column and row represents one of the twenty standard amino acids. In bioinformatics, PAM matrices are regularly used as substitution matrices to score sequence alignments for proteins. Each entry in a PAM matrix indicates the likelihood of the amino acid of that row being replaced with the amino acid of that column through a series of one or more point accepted mutations during a specified evolutionary interval, rather than these two amino acids being aligned due to chance. Different PAM matrices correspond to different lengths of time in the evolution of the protein sequence. While on the other hand;

**BLOCKS SUBSTITUTION MATRIX (BLOSUM)**

This matrix is a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences. They are based on local alignments. BLOSUM matrices were first introduced in a paper by Steven Henikoff and Jorja Henikoff. They scanned the BLOCKS database for very conserved regions of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of amino acids and their substitution probabilities. Then, they calculated a log-odds score for each of the 210 possible substitution pairs of the 20 standard amino acids. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins like the PAM Matrices.

**2(B)** **HEURISTIC DATABASE SEARCHING:**

This is a computational strategy to find an empirical or near optimal solution by using rules of thumb. Although accurate and reliable, is too slow and impractical when computational resources are limited. The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming. There are two major heuristic algorithms for performing database searches: BLAST and FASTA.

**3(A) DEFINE THE FOLLOWING:**

1. **SEQUENCE HOMOLOGY:**

When two sequences are descended from a common evolutionary origin, they are said to have a homologous relationship or share homology.

Sequence homology is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity.

1. **SEQUENCE SIMILARITY:**

It is a measure of an empirical relationship between sequences. Its common objective is establishing the likelihood for sequence homology i.e chance that sequences has evolved from a common ancestor.

1. **SEQUENCE IDENTITY:**

Sequence identity is the amount of characters which match exactly between two different sequences. A similarity score is therefore aimed to approximate the evolutionary distance between a pair of nucleotide or protein sequences.

**3(B)** (i)**DOT MATRIX METHOD:** It is a graphical way of comparing two sequences in a two dimensional matrix. In a dot matrix, two sequences to be compared are written in the horizontal and vertical axes of the matrix.

(ii) **DYNAMIC PROGRAMMING METHOD**: is an exhaustive and quantitative method to find optimal alignments. This method effectively works in three steps. It first produces a sequence versus sequence matrix. The second step is to accumulate scores in the matrix. The last step is to trace back through the matrix in reverse order to identify the highest scoring path. This scoring step involves the use of scoring matrices and gap penalties.

(iii) **THE WORD METHOD:** This method is used in fast database similarity searching.

**3(B) (II) PAIRWISE SEQUENCE ALIGNMENTS:**

This is the process of aligning two sequences and is the basis of database similarity searching and multiple sequence alignments; it is also the fundamental component of many bioinformatics applications.

Pairwise sequence alignment is used to identify regions of similarity that may indicate functional, structural and or/ evolutionary relationships between two biological sequences e.g. say: (protein or nucleic acid).

**4(A)** **GLOBAL ALIGNMENT** deals with two sequences to be aligned that are assumed to be generally similar over their entire length, gaps are added to each until the end of one is reached. Also, in global alignment both sequences are all taken into consideration when finding alignment.

While on the other hand, **LOCAL ALIGNMENT** finds local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions. They do not assume that the two sequences have similarity over the entire length.

**4(B) (I) DISTINGUISHING SEQUENCE HOMOLOGY AND SEQUENCE SIMILARITY**

**Sequence homology** is an inference or a conclusion about a common ancestral relationship drawn from sequence similarity comparison when the two sequences share a high enough degree of similarity.

**Sequence similarity** refers to a direct result of observation from the sequence alignment, doesn’t imply homology. It is the likeness between two sequences and also the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity.

**(II)DISTINGUISHING SEQUENCE SIMILARITY AND SEQUENCE IDENTITY**

Sequence similarity and sequence identity are synonymous for nucleotide sequences. For protein sequences, however, the two concepts are very different.

In a protein sequence alignment; **Sequence similarity** refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other. **Sequence identity** refers to the percentage of matches of the same amino acid residues between two aligned sequences.

On another note, it is a measure of an empirical relationship between sequences and the amount of characters which match exactly between two different sequences **respectively.**

**GOD SAVEL US ALL FROM COVID-19**