**NAME: AMADI CHIMENUMHAOMA ALMA.**

**MATRIC NUMBER: 18/MHS05/003.**

**DEPARTMENT: PHYSIOLOGY**

**CSC 262**

**Assignment**

1(a) List and discuss any three (3) criterion of database searching

 Answer.

* Sensitivity: This criterion 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" in the database searching exercise.
* Selectivity: This can also be known as specificity. It refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered to be "false positives".
* Speed: This can be described as the time it takes to get results from database searching. Depending on the size of a database, speed can sometimes be a primary concern.

 1(b) Explain briefly Basic Local Alignment Search Tool (BLAST) as used in database similarity searching

 Answer.

* BLAST is a pairwise local alignment search tool that is designed to operate maore quickly than exact methods, but without a guarantee of finding the best possible alignment. At present BLAST is the preferred tool for searching large sequence databases. Searching a database with a query sequence can be considered to be a problem in local alignment. Local alignment tools seek a region of alignment between two sequences where the match between the sequences is better than some threshold (which is often set to be the expected match between random sequences of similar length and composition). It is possible to calculate the expected match score for two random sequences that match the composition of the query and target sequences. Several variants of BLAST are now available, reflecting both progressive improvement of the algorithm and tuning for specific purposes.

2(a) Briefly discuss the differences between Dayhoff PAM Matrices and BLOSUM Matrices

 Answer.

* PAM matrices are based on an explicit evolutionary model (i.e. replacements are counted on the branches of a phylogenetic tree), whereas the BLOSUM matrices are based on an implicit model of evolution.
* The PAM matrices are based on mutations observed throughout a global alignment, this includes both highly conserved and highly mutable regions. The BLOSUM matrices are based only on highly conserved regions in series of alignments forbidden to contain gaps.
* The method used to count the replacements is different: unlike the PAM matrix, the BLOSUM procedure uses groups of sequences within which not all mutations are counted the same.
* Higher numbers in the PAM matrix naming scheme denote larger evolutionary distance, while larger numbers in the BLOSUM matrix naming scheme denote higher sequence similarity and therefore smaller evolutionary distance. Example: PAM150 is used for more distant sequences than PAM100; BLOSUM62 is used for closer sequences than BLOSUM50.

2(b) Explain briefly Heuristic Database Searching

 Answer.

* Heuristic Database Searching is a technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution. This is achieved by trading optimality, completeness, accuracy, or precision for speed. In a way, it can be considered a shortcut. A heuristic function, also called simply a heuristic, is a function that ranks alternatives in search algorithms at each branching step based on available information to decide which branch to follow. The objective of a heuristic is to produce a solution in a reasonable time frame that is good enough for solving the problem at hand. This solution may not be the best of all the solutions to this problem, or it may simply approximate the exact solution. But it is still valuable because finding it does not require a prohibitively long time.

3(a) Define the following (i) Sequence Homology (ii) Sequence Similarity

(iii) Sequence Identity

 Answer.

* Sequence homology: This 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: This is a measure of an empirical relationship between sequences. It's common objective is establishing the likelihood for sequence homology i.e chance that sequences have evolved from a common ancestor.
* Sequence Identity: This 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) Give any three (3) methods of Alignment Algorithm

 Answer.

* Dot Matrix Method.
* Dynamic Programming Method.
* Word Method.

(ii) Discuss briefly Pairwise Sequence Alignment

 Answer.

* Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid). Pairwise alignment is one of the most fundamental tools of bioinformatics and underpins a variety of other, more sophisticated methods of annotation. Pairwise alignment in its most rigorous form uses a method called ‘dynamic programming’, which is highly accurate, but also incredibly costly to compute.In order to align anything other than an exact alphabetic match, the algorithm has to know what it is looking for and how it can evaluate the worth of what it finds. To this end, ‘comparison matrices’ have been created which define a score for every possible match possibility—an effective tally of how well the computational alignment is doing. The software will search for the highest score available. The final score is relevant only with its resulting alignment and cannot be used outside this context.

4(a) Differentiate between Global Alignment and Local Alignment

 Answer.

4b Distinguish between the following

1. Sequence Homology and Sequence Similarity

Answer.

* Sequence similarity refers to the likeness or % identity between two sequences while sequence homology refers to shared ancestry.
* Sequence similarity means sharing a statistically significant number of bases of amino acids while in sequence homology, two sequences are homologous if they are derived from a common ancestral sequence.
* Sequence similarity does not imply homology but sequence homology implies similarity.
1. Sequence Similarity and Sequence Identity

Answer.

For protein sequences:

Sequence Identity refers to the percentage of matches of the same amino acids residues between two aligned sequences while sequence similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other.