Matric No: 18/MHS03/014

NAME: Chike-osegbue Gabriella Ifeanyi

COURSE CODE: CSC 262

*ASSIGNMENT

1. List and discuss any three criterion of database searching*

-Sensitivity

-Specificity/Selectivity

-Speed

A. Sensitivity: It is a statistical measure of the performance of binary classification test. It is also called the true positive rate, the recall or probability of detection. It is measured by the extent of inclusion of correctly identified sequence members of the same family.

B. SELECTIVITY: (Also called true negative rate) measures the proportion of actual negatives that are correctly identified as such. It refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered "false positives".

NOTE: The terms "positive" and "negative" don't refer to the value of the condition of interest, but to its presence or absence; the condition itself could be a disease, so that "Positive" might mean "Diseased" while "Negative" might mean "Healthy".

C. This is the time it takes to get results from databases, speed can be a primary concern but the keywords you provide, the higher the probability that the specified data will be found with a faster speed.

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

It is a sequence similarity search program that can be used to quickly search a sequence database for matches to a query sequence. BLAST provides an exact value, statistical information about the significance of each alignment.

Its main function is to compare a sequence of interest, the query sequence, to sequences in a large database, BLAST then reports the best matches, or "hits" found in the database.

2A. *Briefly discuss the differences between Dayhoff PAM Matrices and BLOSUM* *Matrices*

Dayhoff PAM(Point Accepted Mutations) was created by Margaret Dayhoff. PAM Matrices are regularly used as substitution matrices to score sequence alignments, between closely related protein sequences. They are based on global alignments, mutations in global alignments are very significant. Their alignments have high similarity than BLOSUM alignments. Different PAM Matrices correspond to different lengths of time in the evolution of protein sequence. Higher numbers in the PAM Matrix naming denotes greater evolutionary distance. Example: PAM 250 is used for more distant sequences than PAM 120.

WHILE

BLOSUM (Blocks substitution Matrix) which was created by Steven and Henikoff is a substitution matrix used to score alignments between evolutionary divergent protein alignments. They are based on local alignments. Their alignments have low similarity than PAM alignments. Based on highly conserved stretches of alignments. All BLOSUM Matrices are based on observed alignments, they are not extrapolated from comparisons of closely related proteins like the PAM matrices. Higher numbers in the BLOSUM matrix naming denotes higher sequence similarity and smaller evolutionary distance.

Example: BLOSUM 80 is used for closely related sequences than BLOSUM 62.

2B. *Explain briefly Heuristic database searching.*

Heuristic database searching refers to a search strategy that attempts to optimize a problem by iteratively improving the solution based on a given heuristic function or a cost measure. Heuristic algorithms have been designed to reduce the time required to build an alignment that has a reasonable chance to be the best one. Such algorithms have been implemented as fast and efficient programs (Blast, FastA) available in different types to address different kinds of problems. The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming.

3A. *Define the following: [1] Sequence Homology [2] sequence similarity [3] sequence identity*

I. *Sequence Homology:* it is an inference or a conclusion about a common ancestral relationship. When two sequences are descended from a common evolutionary origin, they are said to have a Homologous relationship or share Homology. Homology sequences usually have the same, or very similar functions so new sequences can be relatively assigned functions sequences with known functions can be identified.

ii. *Sequence similarity:* Sequence similarity is a measure of an empirical relationship between sequences. Its common objective is establishing the likelihood for sequence Homology. It is the likeness (resemblance) between two sequences in comparison.

iii. *Sequence Identity:* sequence identity is the amount of characters which match exactly between two different sequences. The gaps are not counted and the measurement is relational to the shorter of the two sequences. It is the extent to which two (nucleotide or amino acid) sequences have the same residues at the same positions in an alignment, often expressed as a percentage.

3B. *[A] Give any three methods of alignment algorithm [B] Discuss briefly pairwise sequence alignment.*

[AI] Dot matrix method

[Aii] Dynamic programming method

[Aiii] Word method.

[Bi] *Pairwise Sequence Alignment:*Pairwise sequence alignment methods are used to find the best matching local or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision such as searching a database for sequences. Extremely useful in structural, functional and evolutionary analyses of sequences. Pairwise sequence alignment provides inference for the relatedness of two sequences.

4A. *Differentiate between Global Alignment and Local Alignment*

Global Alignment is made to align the entire sequence (end to end alignments). It has the same length and it is suitable for aligning two closely related sequences usually done for comparing homologous gene. Example is the Needleman-Wunsch algorithm.

While

Local Alignment find local regions with the highest level of similarity between the two sequences, finds stretches of sequence with High level of matches, suitable for Aligning more distantly related sequences. It is used for finding out conserved patterns of DNA. Example is the Smith-Waterman algorithm.

4B. *Distinguish between the following:

1. Sequence Homology and Sequence similarity

2. Sequence similarity and Sequence Identity*

1 *. Sequence Homology* is an inference or conclusion about a common ancestral relationship. Presence of similar features because of statement common decent. It is a qualitative statement. Homology usually implies similarity.

While

Sequence similarity is a direct result of observation from sequence alignment. It quantifies a likeness or % identity. It can be quantified using percentage. Similarity does not imply Homology.

2. *Sequence similarity*is the extent to which nucleotide or protein sequences are related. It is the likeness(resemblance) between two sequences in comparison.

While

*Sequence Identity*is the extent to which two (nucleotide or amino acid) sequences are invariant. Number of characters that match exactly between two different sequences.