NAME: KOLAWOLE OLUREMI GRACE

DEPARTMENT: HUMAN ANATOMY

MATRIC NUMBER: 18/MHS03/006

COURSE CODE: CSC 262

COURSE TITLE: BIOINFORMATICS II

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

1. Sensitivity
2. Selectivity
3. Speed

1. Sensitivity: Is 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.
2. Selectivity: also called specificity, is the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered false positives.
3. Speed: this is the time it takes to get results from database searches.

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

BLAST is an example of heuristic algorithms used for performing database searches. This type of algorithm performs faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming. BLAST is an algorithm and program for comparing primary biological sequence information, such as the [amino-acid](https://en.wikipedia.org/wiki/Amino_acid) sequences of proteins or the [nucleotides](https://en.wikipedia.org/wiki/Nucleotide) of [DNA](https://en.wikipedia.org/wiki/DNA_sequence) and/or [RNA](https://en.wikipedia.org/wiki/RNA) sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (called a query) with a library or [database](https://en.wikipedia.org/wiki/Database) of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Different types of BLASTs are available according to the query sequences and the target databases. For example, following the discovery of a previously unknown gene in the [mouse](https://en.wikipedia.org/wiki/Mus_musculus), a scientist will typically perform a BLAST search of the [human genome](https://en.wikipedia.org/wiki/Human_genome) to see if humans carry a similar gene; BLAST will identify sequences in the human genome that resemble the mouse gene based on similarity of sequence.

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

1. 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.
2. 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.
3. 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.
4. 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.

2b) Explain briefly heuristic database searching

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 and they are BLAST and FASTA.

3a) Define the following: i) Sequence Homology ii) Sequence similarity iii) Sequence identity

1. Sequence 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.
2. Sequence similarity: Sequence similarity 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
3. 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.

3b) (i) Give any three (3) methods of Alignment Algorithm

(ii) Discuss briefly Pairwise Sequence Alignment

1. a) Dot matrix method

b) Dynamic programming method

c) The word method

1. The pairwise sequence alignment is the fundamental component of many bioinformatics applications. It is used to identify regions of similarity that may indicate functional, structural and or evolutionary relationships between two biological sequence. It is 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

In Global Alignment, two sequences to be aligned are assumed to be generally similar over their entire length. Alignment is carried out from beginning to end of both sequences to find the best possible alignment across the entire length between the two sequences. You take entirety of both sequences into consideration when finding alignment while Local Alignment does not assume that the two sequences in question have similarity over the entire length. It only 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. This approach can be used for aligning more divergent sequences with the goal of searching for conserved patterns in DNA or protein sequences.

4b) Distinguish between the following:

1. Sequence Homology and Sequence Similarity
2. Sequence Similarity and Sequence Identity
3. 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 while similarity is a direct result of observation from the sequence alignment. Sequence similarity can be quantified using percentages; homology is a qualitative statement. For example, one may say that two sequences share 40% similarity but It is incorrect to say that the two sequences share 40% homology. They are either homologous or non-homologous
4. In a protein sequence alignment, sequence identity refers to the percentage of matches of the same amino acid residues between two aligned sequences while similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other. Sequence similarity is a measure of an empirical relationship between sequences while Sequence identity is the amount of characters which match exactly between two different sequences.