**Name**:Omodele Morayo

**Matric Number**:18/mhs01/295

**Department**:Anatomy

**Course code**:CSC 262

**Criterion for database searching**

1a) I) The first criterion is sensitivity, which 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.

II) The second criterion is selectivity, also called specificity, which refers to the ability to exclude incorrect hits. These incorrect hits are unrelated sequences mistakenly identified in database searching and are considered “false positives.”

III) The third criterion is speed, 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.

**BLAST in database similarity searching**

1b)In bioinformatics,BLAST is an algorithm and program used for comparing primary biological sequence information, such as the amino acid sequence of proteins or nucleotides of DNA and RNA sequences. A blast search enables a researcher to compare a subject protein or nucleotide sequence with a database of sequences and identify sequences that resemble the query sequence. When working with genes, BLAST can locate common genes in two related species, and can be used to map annotations from one organism to another.

2a) **DIFFERENCE BETWEEN PAM AND BLOSUM MATRICES**

**PAM matrices BLOSUM matrices**

|  |  |
| --- | --- |
| Based on global alignment | Based on local alignment |
| Alignment has high similarities | Alignment has a low similarities |
| Used to score alignment between closely related protein sequences | Used to score alignments between evolutionarily divergent protein sequences |
| Higher numbers in the PAM matrix naming denotes great evolutionary distance. | Higher numbers in the BLOSUM matrix naming denotes higher sequences similarity and smaller evolutionary distance. |

2b. **HEURISTIC DATABASE SEARCHING**

Searching a large database using the dynamic programming methods, such as the Smith–Waterman algorithm, although accurate and reliable, is too slow and impractical when computational resources are limited

• Thus, speed of searching became an important issue. To speed up the comparison, heuristic methods have to be used.

• The heuristic algorithms perform faster searches because they examine only a fraction of the possible alignments examined in regular dynamic programming.

• Currently, there are two major heuristic algorithms for performing database searches: BLAST and FASTA.

3a) I) **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

II)**sequence similarity** is the percentage of aligned residues that are similar in physiochemical properties such as size, charge, and hydrophobicity

III)**Sequence Identity** is the amount of characters which match exactly between two different sequences

3b)I). **Alignment algorithms**

* Dot matrix methods
* The dynamic programming method
* The word method

II)**Pairwise sequence** alignment is the process of aligning two sequences and is the basis of database similarity searching and multiple sequence alignment

• Pairwise sequence alignment is 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 sequence e.g. say: (protein or nucleic acid)

• It is extremely useful in structural, functional, and evolutionary analyses of sequences. Pairwise sequence alignment provides inference for the relatedness of two sequences.

4a) **Difference between global and local alignment**

**Global alignment** **Local alignment**

|  |  |
| --- | --- |
| Two sequences to be aligned are assumed to be generally similar over their entire length | Does not assume that the two sequences in question have similarity over the 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. | 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 |

4b) I) **Difference between sequence homology and sequence similarity**

**Sequence homology. Sequence Similarity**

|  |  |
| --- | --- |
| It refers to the shares ancestry | It refers to the likeness or% identity between 2 sequences |
| Usually implies similarity | Doesn’t imply homology |
| 2 sequences are homologous if they are derived from a common ancestral sequence | They share a statistically significant number of amino acid or base |

4b) II) **Difference between sequence similarity and sequence identity**

**sequence similarity** **sequence identity**

|  |  |
| --- | --- |
| Similarity refers to the percentage of aligned residues that have similar physicochemical characteristics and can be more readily substituted for each other. | In a protein sequence alignment, sequence identity refers to the percentage of matches of the same amino acid residues between two aligned sequences. |
| Sequence similarity is a measure of an empirical relationship between sequences. Its common objective is establishing the likelihood for sequence homology ie chance that sequences has evolved from a common ancestor | 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. |