**Ezeano NNaemeka Prince**

**16/sci01/012**

**Csc262**

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

Database Search Criteria Mathematical Criteria > More Than or After >100 for example < Less Than or Before = More Than or Equal To >=100 for example <= Less Than or Equal To <=100 for example Logical Operators Between And Between 1980 And 2000 for example Or January Or April for example Not Used to exclude data – Not 2000 for example Like Criteria Like \* \* Contains the words…… - Like \*kite\* for example Like a\* Fields starting with a for example Like \*a Fields ending with a for example Finding fields with no data Null Used to display fields that contain no data Searching fields that contain a certain number of characters “?????” Used to search fields for particular number of characters (in this case 5) NOTE: This is the only criteria where you have to type in the quotation marks (“ “) yourself. For the other criteria you should let Access add the quotation marks for you.

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

The BLAST program scans the database sequences for the remaining high-scoring word, such as PEG, of each position. If an exact match is found, this match is **used** to seed a possible un-gapped **alignment** between **the** query and database sequences. Extend **the** exact matches to high-scoring segment pair (HSP). Since the discovery of the [genetic code](https://www.nature.com/scitable/topicpage/Reading-the-Genetic-Code-1042), biological research has undergone a sea of change in the way it is performed. Until the early twentieth century, biology focused on the processes of living organisms and almost always involved experiments in laboratories and in the field. The growth of molecular biology during the twentieth century moved research into the test tube, where biological systems could be painstakingly dissected and reassembled. Then, beginning in the 1970s, scientists started accumulating DNA and protein sequence data at an exponential rate; in fact, researchers currently have approximately 97 billion bases sequenced and over 93 million records. Amazingly, this sequence data doubles every 18 months!

But how do investigators make sense of this massive amount of data? How can they identify the functions of newly cloned genes? And is it possible to estimate the evolutionary relationships between genes or proteins just by examining their nucleotide or amino acid sequences? To address these important issues, researchers must first tease out the relationships between different species that are descended from a common ancestor. Any sequence similarity can then be used to infer function and evolutionary relationships. In fact, one common method for examining and comparing genes is to search for similarities between newly sequenced DNA and databases of gene sequences that have already been described. By identifying related .genes or gene families with known functions, scientists can infer the functions and evolutionary relationships of newly cloned genes or even whole genomes.

As gene and protein sequence databases grew at the end of the twentieth century, scientists turned to computers to help analyze this abundant and ever-growing amount of data. Today, one of the most common tools used to examine DNA and protein sequences is the Basic Local Alignment Search Tool, also known as BLAST (Altschul *et al*., 1990). BLAST is a computer algorithm that is available for use online at the [National Center for Biotechnology Information (NCBI) website](http://blast.ncbi.nlm.nih.gov/Blast.cgi), as well as many other sites. BLAST can rapidly align and compare a query DNA sequence with a database of sequences, which makes it a critical tool in ongoing genomic research. In fact, the initial paper describing the program, published in the *Journal of Molecular Biology*and entitled "[Basic Local Alignment Search Tool](http://www.ncbi.nlm.nih.gov/pubmed/2231712)," was the most highly cited publication of the 1990s (Taubs, 2000). In recent years, the parallel development of large-scale sequencing projects and bioinformatic tools like BLAST has enabled scientists to study the genetic blueprint of life across many species, and it has also helped connect biology and computer science in the maturing field of bioinformatics.

**Similarity searches** are an essential component of most bioinformatic applications. They form the bases of structural motif identification, gene identification, and insights into functional associations. With the rapid increase in the available genetic data through a wide variety of databases, similarity searches are an essential tool for accessing these data in an informative and productive way. In this chapter, we provide an overview of similarity searching approaches, related databases, and parameter options to achieve the best results for a variety of applications. We then provide a worked example and some notes for consideration.

The standard for identifying a nucleotide sequence record is by anaccession.version system where the accession number is an identifierof two letters followed by six digits and the version is an incre-mental number indicating the number of changes that have been made to the sequence since it was first submitted. Locus names(see Note 1) are older, less standardized identifiers whose originalpurpose was to group entries with similar sequences (10). Theoriginal locus format was intended to hold information about theorganism and other common group characteristics (such as geneproduct). That ten-character format is no longer able to hold such information for the large number and variety of sequences nowavailable, so the locus has become yet another unique identifieroften set to be the same value as the accession number. Databaseidentifiers are simply two- or three-character strings that serve to indicate which database originally received and stored the infor-mation. The database identifier is the first value listed in theFASTA identifier syntax (Table 1.1).When a sequence is first submitted to GenBank, it is submittedwith several defined features associated with the sequence. Some include CDS (coding sequence), RBS (ribosome binding site),rep\_origin (origin of replication), and tRNA (mature transfer RNA) information. A translation of protein coding nucleotide sequences into amino acids is provided as part of the features section. Likewise, labeling of different open reading frames,introns, etc., are all part of the table of features. A list of features and their descriptions, formats, and conventions that were agreed

upon by INSDC can be found in the Feature

2A)

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



**PAM matrices** are used to score alignments **between** closely related protein sequences. **BLOSUM matrices** are used to score alignments **between** evolutionarily divergent protein sequences. Reference : **BLOSUM** - A **matrix**; derived from ungapped alignments.

The **PAM** matrices are based on scoring all amino acid positions in related sequences, whereas the **BLOSUM** matrices are based on substitutions and conserved positions in blocks, which represent the most-alike common regions in related sequences.

**2B) Explain briefly Heuristic Database Searching**

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. For example, it may approximate **the** exact solution. 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.

Heuristics may produce results by themselves, or they may be used in conjunction with optimization algorithms to improve their efficiency (e.g., they may be used to generate good seed values).

Results about [NP-hardness](https://en.wikipedia.org/wiki/NP-hard) in theoretical computer science make heuristics the only viable option for a variety of complex optimization problems that need to be routinely solved in real-world applications.

Heuristics underlie the whole field of Artificial Intelligence and the computer simulation of thinking, as they may be used in situations where there are no known [algorithms](https://en.wikipedia.org/wiki/Algorithm).

**3A)   Define the following (i) Sequence Homology    (ii) Sequence Similarity  (iii) Sequence Identity**

1. **Sequence homology** is the biological homology between DNA, RNA, or protein sequences, defined in terms of shared ancestry in the evolutionary history of life. (Science: molecular biology) Strictly, [refers](https://www.biologyonline.com/dictionary/refers) to the [situation](https://www.biologyonline.com/dictionary/situation) where [nucleic acid](https://www.biologyonline.com/dictionary/nucleic-acid) or [protein](https://www.biologyonline.com/dictionary/protein) [sequences](https://www.biologyonline.com/dictionary/sequences) are [similar](https://www.biologyonline.com/dictionary/similar) because they have a common evolutionary [origin](https://www.biologyonline.com/dictionary/origin). Often used loosely to [indicate](https://www.biologyonline.com/dictionary/indicate) that sequences are very similar. Sequence similarity is observable, [homology](https://www.biologyonline.com/dictionary/homology) is an [hypothesis](https://www.biologyonline.com/dictionary/hypothesis) based on [observation](https://www.biologyonline.com/dictionary/observation).
2. **Sequence Similarity** Searching is a method of searching **sequence** databases by using alignment to a query **sequence**. By statistically assessing how well database and query **sequences** match one can infer homology and transfer information to the query **sequence**.
3. **Sequence identity** is the amount of characters which match exactly between two different **sequences**. Hereby, gaps are not counted and the measurement is relational to the shorter of the two **sequences**.

**3B)** (i)         **Give any three (3) methods of Alignment Algorithm  (ii)  Discuss briefly Pairwise Sequence Alignment**

The three primary methods of producing pairwise alignments are dot-matrix methods, dynamic programming, and word methods; however, multiple sequence alignment techniques can also align pairs of sequences..

**Discuss briefly Pairwise Sequence Alignment**

What is **Pairwise Sequence Alignment**? The goal of pairwise sequence alignment is to come up with the best possible alignment of two sequences. Given a scoring system for matches, mismatches and gaps (See Figure 4), the best possible alignmen**t** is defined as the alignment that optimizes the total summed score.

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

A general **global alignment** technique is the Needleman–Wunsch algorithm, which is based on dynamic programming. **Local alignments** are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar **sequence** motifs within their larger **sequence** context.

****

**4B) Distinguish between the following**

1. **Sequence Homology and Sequence Similarity**
2. **Sequence Similarity and Sequence Identity**

**Similarity**: Degree of likeness **between** two **sequences**, usually expressed as a percentage of similar (or identical) residues over a given length of the alignment. ... **Homology**: Statement about common evolutionary ancestry of two **sequences**. Can only be true or false The term "homology" pertains to comparative studies. Homology indicates an ancient common origin and temporal evolution and refers to structural characteristics. In comparative anatomy, it is used to compare structures in different animal species.

In comparative protein biochemistry, "homology" retains the original meaning of "having a common evolutionary origin" and is used to evolutionarily define two or more proteins by locating common structural characteristics and common spatial distribution of, for instance, beta strands, helices, and folds. Accordingly, homologous protein structures are defined by spatial analyses. Measuring structural homology involves computing the geometric–topological features of a space. One approach used togenerate and analyze three-dimensional (3D) protein structures is homology modeling (also called comparative modeling or knowledge-based modeling). Homology modeling works by finding similar sequences on the basis of the obvious fact that 3D similarity reflects 2D similarity. Nonetheless, it is important to note that homologous structures do not imply sequence similarity as a necessary condition.

Sequence identity is the amount of characters which match exactly between two different sequences. Hereby, gaps are not counted and the measurement is relational to the shorter of the two sequences.

This has the effect that sequence identity is not transitive, i.e. if sequence A=B and B=C then A is not necessarily equal C (in terms of the identity distance measure) :

A: AAGGCTT

B: AAGGC

C:AAGGCAT

Here identity(A,B)=100% (5 identical nucleotides / min(length(A),length(B))).

Identity(B,C)=100%, but identity(A,C)=85% ((6 identical nucleotides / 7)). So 100% identity does not mean two sequences are the same.

Sequence similarity is first of all a general description of a relationship but nevertheless its more or less common practice to define similarity as an optimal matching problem (for sequence alignments or unless defined otherwise).

Hereby, the optimal matching algorithm finds the minimal number of edit operations (inserts, deletes, and substitutions) in order to transform the one sequence into an exact copy of the other sequence being aligned (edit distance). Using this, the percentage sequence similarity of the examples above are sim(A,B)=60%, sim(B,C)=60%, sim(A,C)=86% (semi-global, sim=1-(edit distance/unaligned length of the shorter sequence)). But there are other ways to define similarity between two objects (e.g. using tertiary strucure of proteins).

(II)

So 100% **identity** does not mean two **sequences** are the same. **Sequence similarity** is first of all a general description of a relationship but nevertheless its more or less common practice to define **similarity** as an optimal matching problem (for **sequence** alignments or unless defined otherwise).