**ASIRA DARRELL**

**18/MHS01/088**

**ANATOMY**

**CSC 262 (Bio Informatics)**

1A.) Database similarity search is based upon sequence alignment methods also used in pairwise sequence comparison. Sequence alignment can be global (whole sequence alignment) or local (partial sequence alignment) and there are algorithms to find the optimal alignment given particular comparison criteria. There are unique requirements for implementing algorithms for sequence database searching. They are:

a. 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.

b. 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.”

c. *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.

1B.) BLAST (basic local alignment search tool) is an [algorithm](https://en.wikipedia.org/wiki/Algorithm) and program for comparing [primary](https://en.wikipedia.org/wiki/Primary_structure) biological sequence information, such as the [amino-acid](https://en.wikipedia.org/wiki/Amino_acid) sequences of [proteins](https://en.wikipedia.org/wiki/Protein) 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. BLAST is based on a heuristic algorithm. A heuristic algorithm is an algorithm that provides almost the correct answer or a solution for some instances of the problem. Through a heuristic approach, BLAST identifies homologous sequences by locating short matches between the two sequences being compared. This process is referred to as seeding, and it is after this initial match that BLAST begins to make local alignments. During the process of seeding, BLAST tries to locate all common three-letter words between the sequence of interest and the hit sequence, or sequences, from the database. In this context, a word is simply defined as a number of letters. For example, for blast , the default word size is 3 W=3. If a query sequence has ABCDE, the searched words are ABC, BCD, CDE . After synthesizing words for a given sequence of interest, neighborhood words are also assembled. Once both words and neighborhood words are organized, they are compared with the database sequences in order to find matches. The alignment, which is normally 3 residues long, is extended in either direction by the BLAST algorithm. Each extension increases or decreases the score of the alignment, and should the score be higher than a pre-determined threshold, the alignment will be included in the results given by BLAST.

2a.) PAM matrices :( Point Accepted Mutation)  is a [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) where each column and row represents one of the twenty standard amino acids. In [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics), PAM matrices are regularly used as [substitution matrices](https://en.wikipedia.org/wiki/Substitution_matrix) to score [sequence alignments](https://en.wikipedia.org/wiki/Sequence_alignment) for proteins. PAM matrices are also used as a scoring matrix when comparing DNA sequences or protein sequences to judge the quality of the alignment. This form of scoring system is utilized by a wide range of alignment software including [BLAST](https://en.wikipedia.org/wiki/BLAST).

While

**BLOSUM :(**BLOcks SUbstitution Matrix) matrix is a [substitution matrix](https://en.wikipedia.org/wiki/Substitution_matrix) used for [sequence alignment](https://en.wikipedia.org/wiki/Sequence_alignment) of [proteins](https://en.wikipedia.org/wiki/Protein). BLOSUM matrices are used to score alignments between [evolutionarily divergent](https://en.wikipedia.org/wiki/Evolutionary_divergence) protein sequences. They are based on local alignments. They scanned the [BLOCKS database](https://en.wikipedia.org/w/index.php?title=BLOCKS_database&action=edit&redlink=1) for very [conserved regions](https://en.wikipedia.org/wiki/Conserved_sequence) of protein families (that do not have gaps in the sequence alignment) and then counted the relative frequencies of [amino acids](https://en.wikipedia.org/wiki/Amino_acids) and their substitution probabilities.

2b**.) HEURISTIC SEARCH** 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. A heuristic search method does not always guarantee to find an optimal or the best solution, but may instead find a good or acceptable solution within a reasonable amount of time and memory space. 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.

3ai.) **Sequence homology** is the [biological homology](https://en.wikipedia.org/wiki/Homology_(biology)) between [DNA](https://en.wikipedia.org/wiki/DNA_sequence), [RNA](https://en.wikipedia.org/wiki/RNA_sequence), or [protein sequences](https://en.wikipedia.org/wiki/Protein_primary_structure), defined in terms of shared ancestry in the [evolutionary history of life](https://en.wikipedia.org/wiki/Evolutionary_history_of_life). Two segments of DNA can have shared ancestry because of three phenomena: either a [speciation](https://en.wikipedia.org/wiki/Speciation) event (orthologs), or a [duplication event](https://en.wikipedia.org/wiki/Gene_duplication) (paralogs), or else a [horizontal (or lateral) gene transfer](https://en.wikipedia.org/wiki/Horizontal_gene_transfer) event (xenologs).

3aii.) **Sequence Similarity** is a method of searching sequence databases by using alignment to a query sequence. Sequence similarity is a concept from computational biology and computer science. Sequence similarity is a number that shows how much two sequences are similar. Sequence similarity is sometimes, but [not always](https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm), defined via sequence distance: the smaller the distance, the more similar the sequences.

3aiii.) **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.

3bi.)

Dot-matrix methods: The dot-matrix approach, which implicitly produces a family of alignments for individual sequence regions, is qualitative and conceptually simple, though time-consuming to analyze on a large scale. In the absence of noise, it can be easy to visually identify certain sequence features—such as insertions, deletions, repeats, or inverted repeats from a dot-matrix plot.

Dynamic programming:The technique of dynamic programming is theoretically applicable to any number of sequences; however, because it is computationally expensive in both time and [memory](https://en.wikipedia.org/wiki/Computer_memory), it is rarely used for more than three or four sequences in its most basic form. This method requires constructing the *n*-dimensional equivalent of the sequence matrix formed from two sequences, where *n* is the number of sequences in the query.

Word method: Word methods, also known as *k*-tuple methods, are [heuristic](https://en.wikipedia.org/wiki/Heuristic) methods that are not guaranteed to find an optimal alignment solution, but are significantly more efficient than dynamic programming. These methods are especially useful in large-scale database searches where it is understood that a large proportion of the candidate sequences will have essentially no significant match with the query sequence. Word methods are best known for their implementation in the database search tools [FASTA](https://en.wikipedia.org/wiki/FASTA) and the [BLAST](https://en.wikipedia.org/wiki/BLAST) family.

3bii) **PAIRWISE SEQUENCE ALIGNMENT** : are used to find the best-matching piecewise (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 with high similarity to a query). 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. Although each method has its individual strengths and weaknesses, all three pairwise methods have difficulty with highly repetitive sequences of low [information content](https://en.wikipedia.org/wiki/Information_content)  especially where the number of repetitions differ in the two sequences to be aligned. One way of quantifying the utility of a given pairwise alignment is the 'maximum unique match' (MUM), or the longest subsequence that occurs in both query sequences. Longer MUM sequences typically reflect closer relatedness.

4a.) Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. (This does not mean global alignments cannot start and/or end in gaps.) A general global alignment technique is the [Needleman–Wunsch algorithm](https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm), which is based on dynamic programming.  A global alignment would attempt to force the alignment to extend beyond the region of overlap, In that case, the short sequence should be globally (fully) aligned.

While

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. A local alignment might not fully cover the region of overlap, in that case only a local (partial) alignment is desired for the long sequence.

4bi.) 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

Sequence similarity is a direct result of observation from the sequence alignment and can be quantified using percentages. If the sequence similarity level is high enough, a common evolutionary relationship can be inferred.

4bii.) 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.

While

Sequence identity is the amount of characters which match exactly between two different sequences.