**“BLAST” wordsearch**

BLAST uses an algorithm to find and score matches between sequences. The program can be performed very rapidly over a specific sequence or a database of millions of sequences. In this assignment, you will perform a ‘word’ search similar to what the BLAST algorithm can do, but you will need to perform this manually. BLAST can search for different lengths of ‘words’ (often 3-7 characters).

Overview: You will be manually performing a small BLAST search. BLAST does not work well with short sequences, but it would be too tedious to manually perform a BLAST type of search with a long (even 20 nucleotides) sequence.

|  |  |
| --- | --- |
| 1. How often would you expect to find the sequence GGCTA in a random sequence?
 | 1 in 45 or 1 in 1096 bases |
| 1. The sequence in Box 1 is 560 bp in length. How many times do you expect to find the 5-base sequence GGCTA (sequence in question 1)?
 | Maybe 0 or 1 times is expected |

In this part, you will be searching for the sequence GGCTA using 3-letter words. Using the sequence in Box 1, which is presented in the FASTA format, search for the ‘words’ below. Highlight the ‘words’ in Box 1.

In addition to the sequence below, you should have three different highlighters available to mark the sequence as you find it.

Box 1 – partial sequence of *Bacillus subtilis* 16S rRNA as a FASTA sequence

>NR\_112116\_partial\_B\_subtilis\_16S\_rRNA

TTATCGGAGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGGA

CAGATGGGAGCTTGCTCCCTGATGTTAGCGGCGGACGGGTGAGTAACACGTGGGTAACCTGCCTGTAAGA

CTGGGATAACTCCGGGAAACCGGGGCTAATACCGGATGGTTGTTTGAACCGCATGGTTCAAACATAAAAG

GTGGCTTCGGCTACCACTTACAGATGGACCCGCGGCGCATTAGCTAGTTGGTGAGGTAACGGCTCACCAA

GGCAACGATGCGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGGCCCAGACTCCTA

CGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGA

AGGTTTTCGGATCGTAAAGCTCTGTTGTTAGGGAAGAACAAGTACCGTTCGAATAGGGCGGTACCTTGAC

GGTACCTAACCAGAAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCAAGCGTTG

|  |
| --- |
| Using a color highlighter, search the Box 1 sequence and highlight any occurrence of the 3-character ‘word’ GGC. |
| How many GGC sequences do you identify? | 16 |
| Using a second color, highlight any occurrences of the 3-character ‘word’ GCT. |
| How many GCT sequences do you identify? | 8 |
| How many overlap with the previously highlighted GGC sequence? | 5 |
| Now, using a third color, highlight any occurrence of the 3-character ‘word’ CTA. |
| How many CTA sequences do you identify? | 7 |
| How many overlap with the GCT sequence (second color)? | 2 |
| How many overlap both the GCC (first color) and GCT (second color)? | 1 |

The 5-base sequence (GGCTA) used in this manual ‘BLAST’ search is too short for use in BLAST. However, adding two bases (now GGCTAAT) will allow you to perform a search with this sequence. BLAST can perform this analysis in under 30 seconds (generally). How long did it take you to do the three steps (#3-5) above against a single sequence?