MATLAB CIE 1 QUIZ ANSWERS

1. The **toolbox** is a collection of functions built on the MATLAB® technical computing environment.
2. What is the purpose of Matlab Bioinformatics Toolbox

**Bioinformatics Toolbox** provides sequence analysis and visualization tools for genomic and proteomic sequence data. You can perform a variety of analyses, including multiple sequence alignments and the building and interactive viewing and manipulation of phylogenetic trees.

1. The latest version of Gene Ontology database, available from http://www.geneontology.org is downloaded over the Web into MATLAB using the function **geneont** when the LIVE parameter is true
2. Write the commands for downloading a file from the nucleotide databases GenBank and EMBL and the protein database GenPept

**For GenBank, S = getgenbank('M10051')**

**For EMBL, out = getembl('X00558','ToFile','rat\_protein.txt')**

**For GenPept, Seq = getgenpept('AAA59174')**

1. Describe these commands - nt2aa, aa2nt, dna2rna & rna2dna with an example

**nt2aa - Convert nucleotide sequence to amino acid sequence.**

**protein1 = nt2aa(ND1gene,'GeneticCode', 2);**

**aa2nt - Convert amino acid sequence to nucleotide sequence**

**aa2nt(seq,'GeneticCode',2)**

**dna2rna - Convert DNA sequence to RNA sequence**

**rna = dna2rna(dna)**

**rna2dna - Convert RNA sequence to DNA sequence**

**rna2dna('ACGAUGAGUCAUGCUU')**

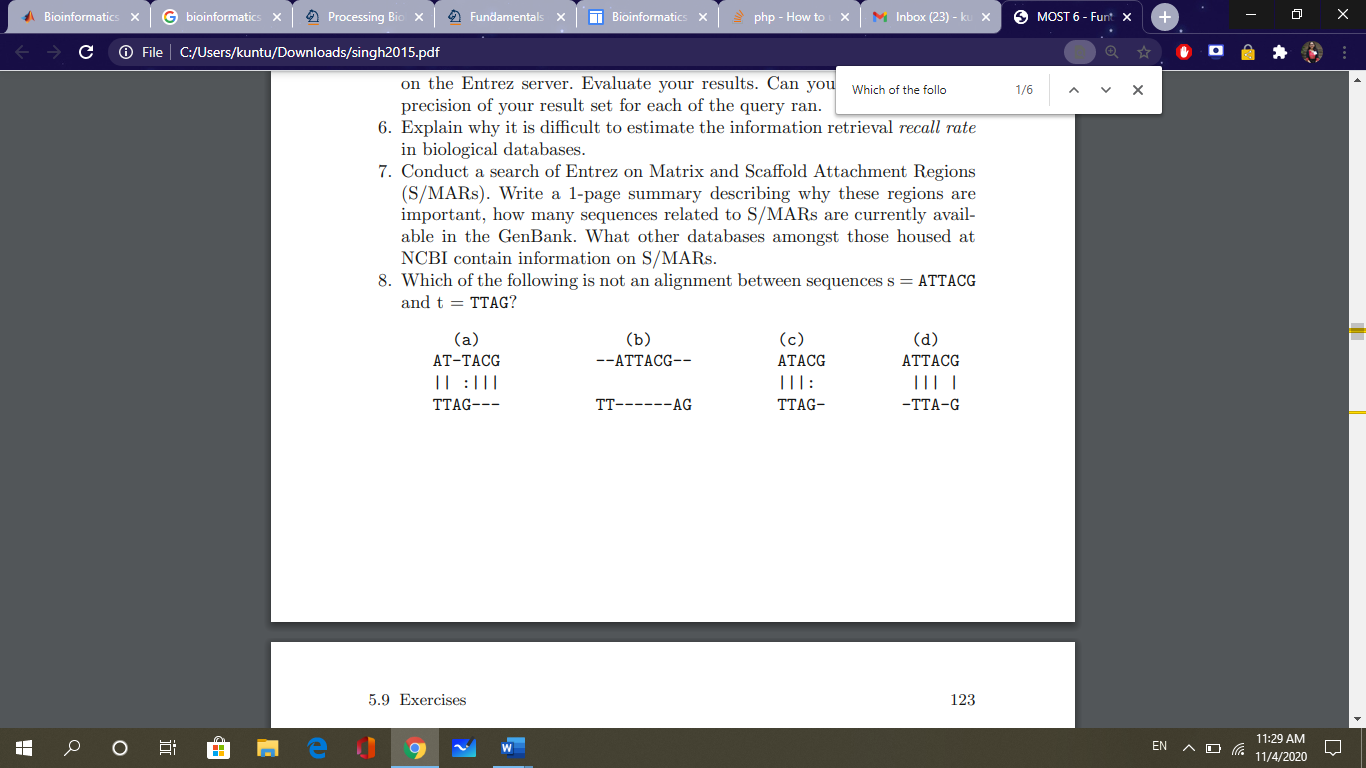
1. Use matlab bioinformatics toolbox function to download the GenBank sequence with accession number U15422. Extract out the raw sequence data from the above sequence object and perform a display of the nucleotides.

**s = getgenbank('U15422');**

**s.Sequence**

1. Which of the following is not an alignment between sequences s = ATTACG and t = TTAG?

**Option (b)**



1. Perfrom a BLAST query for a nucleotide sequence (GCGCGTCTGTTCTGTGGAACAGGAGGCAGTTGTTTTCCGTCCGGCT) in the nr database using the blastn program and retrieve the results from NCBI.

**result =**

**blastncbi('GCGCGTCTGTTCTGTGGAACAGGAGGCAGTTGTTTTCCGTCCGGCT', 'blastn', 'Database', 'nr')**

**result = 'WY3JWUVX016'**

1. Give a brief description of BLOSUM with Matlab

**blosum returns the BLOSUM scoring matrix. Matrix = blosum(Identity) returns a BLOSUM (Blocks Substitution Matrix) scoring matrix with a specified percent identity.**

1. Phylogenetics is **Option (B)**

A) The grouping of organisms by their physical characteristics

B) The study of evolutionary relationships in organisms

C) The study of gene expression in organisms

D) The extraction of phylo from genetic sequences

1. What is the correct order of the steps the ClustalW algorithm uses for Multiple Sequence Alignment: **Option (A)**

1. ClustalW constructs a distance matrix of N(N-1)/2 pairs of sequences by pairwise alignment of the sequences

2. ClustalW builds a guide tree from the distance matrix using the clustering method (neighbor-joining) by Saitou and Nei

3. ClustalW will convert the similarity scores to evolutionary distances based on the model by Kimura

Correct order of processing steps:

A) 1,2,3

B) 2,3,1

C) 1,3,2

D) 2,1,3

1. Consider the following piece of matlab code:

data = {’German\_Neanderthal’ ’AF011222’; ’Russian\_Neanderthal’ ’AF254446’; ’European\_Human’ ’X90314’ ; ’Mountain\_Gorilla\_Rwanda’ ’AF089820’; ’Chimp\_Troglodytes’ ’AF176766’; };

for ind = 1:5

seqs(ind).Header = data{ind,1};

seqs(ind).Sequence = getgenbank(data{ind,2},... ’sequenceonly’, true);

end

distances = seqpdist(seqs,’Method’,’Jukes-Cantor’,’Alphabet’,’DNA’);

tree = seqlinkage(distances,’UPGMA’,seqs)

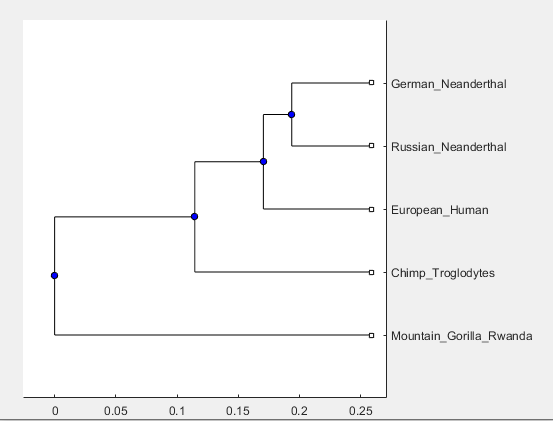
Answer the following questions:

1. Comment on what is stored in the data array.

**In the data array, we have the Species Description in column 1 and GenBank Accession ID in column 2. It is of the class ‘cell’ and is a 5x2 cell.**

1. Run the program and examine the output.

**The output can be seen below. The German and the Russian Neanderthal are closely related in terms of sequence. The Mountain Gorilla Rwanda specie is the most distantly related in terms of sequence.**



1. What is the output of excuting the following matlab code:

>> cellseq = { ’TTGGTT’ , ’TTTGGG’ , ’TGGTTGGT’ , ’GGGTTT’ } ;

>> comp = r e g e x pi ( c e l l s e q , ’TTT’ ) ;

>> ind = fi n d ( c e l l f u n ( ’ isempty ’ , comp ) ) ;

>> cellseq (ind )

**ans =**

**1×2 cell array**

**{'TTGGTT'} {'TGGTTGGT'}**

1. What is the purpose of Microarray Box Plot?

**The box plot is a common tool in exploratory data analysis to determine if a factor has a significant effect on the response variable1 with respect to either location or variation. In gene expression data analysis, the response variable is the expression level, and represented by the vertical axis. The horizontal axis represents the factor of interest, which is an experiment or sample, in gene expression data analysis. It is also helpful in summarizing large quantities of data.**

1. Show the value of the variable seqs upon executing the following matlab commands: >> seq 1 = ’ATTA’ ;

>> seq 2 = ’ATTTA’ ;

>> seq 3 = ’ATTTAA’ ;

>> seqs = char ( seq1 , seq2 , seq3 );

**seqs =**

**3×6 char array**

**'ATTA '**

**'ATTTA '**

**'ATTTAA'**