MCB 5472 Midterm

Your Name

In case of multiple choice questions, more than one answer might be correct.

1) Why is %identity not a good parameter to assess the significance of a match in a blast search?
% identity does not take the lengths of a match into consideration (a 100% match of three residues frequently occurs by chance)

2) Two sequences found in two different organisms are homologous if and only if they evolved from the same ancestral sequence that existed in some organism in the past. ***Correct – Incorrect***

3) Homologous sequences almost always show significant similarity. ***Correct - Incorrect***
(most homologs have diverged beyond recognition)

4) Proteins with moderately complex amino acid composition that show significant similarity in a BLAST search almost always are homologous. ***Correct – Incorrect***

5) In a BLASTP search an E-value for a match is reported as 3 e-62 . Which of the following statements is correct.
A) A value of this magnitude is not sufficient to suggest homology. Convergent evolution frequently gives rise to sequence matches with very small E-values.
B) This means that a match of this quality due to chance would be obtained $3\ 10^{-62}$ times. A match of this quality demonstrates homology beyond reasonable doubts.
C) This means that a match of this quality due to chance would be obtained $3\ e^{-62}$ times (e denotes Euler's number, about 2.72).

6) Give a short definition of E-value and P-value:
E-value: Number of matches expected due to chance with a score equal or better than the one obtained.
P-value: probability to obtain a match with a score equal or better than the one obtained due to chance alone.

7) Which of the following are correctly formed fasta sequences? (new line symbols are given as **\n;** tabulator symbols as **\t**

```
>gi|12643370|sp|Q9P997.2|VATA_THEAC this is a test sequence RecName: Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEMGLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n
```

```
>gi|12643370|sp|Q9P997.2|VATA_THEAC this is a test sequence RecName: Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEM \n GLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n
```

```
\n >gi|12643370|sp|Q9P997.2|VATA_THEAC this is a test sequence RecName: Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEMGLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n
space following the first \n would be a problem
```

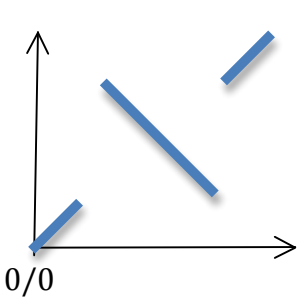
```
>gi|12643370|sp|Q9P997.2|VATA_THEAC \t Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEMGLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n
```

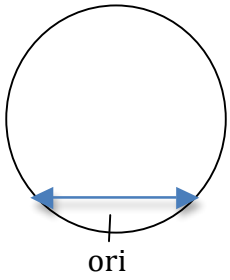
```
>gi|12643370|sp|Q9P997.2|VATA_THEAC \n Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEMGLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n
```

8) What steps does clustal performs to calculate a multiple sequence alignment?
A) calculate all possible pairwise alignments, create bootstrap samples, calculate guide tree, calculate cluster alignment
B) calculate all possible pairwise alignments, calculate master slave alignment using the most conserved sequence as master
C) calculate all possible pairwise alignments, calculate guide tree, calculate cluster alignment

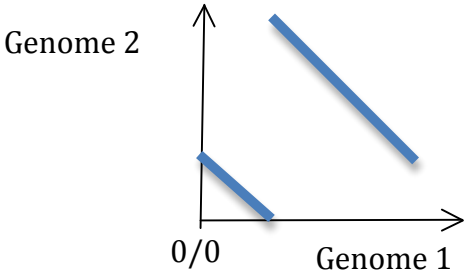
Your name

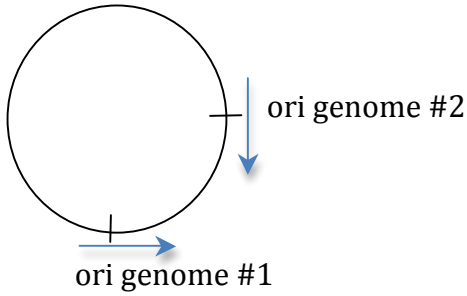
9) Assume that the origin of replications are at (0/0) in the coordinate system on the left (below), and that you compare two closely related genomes. The organisms in question have circular genomes. The sketch below on the right indicates where a rearrangement event took place in one of the genomes. Draw a sketch of the expected gene plot, assuming that except for the single recombination event the two genomes are syntenic.





10) Assume that you compare two syntenic genomes in a gene plot comparison. One of the genomes has the origin of replication assigned to the wrong place. The direction in which the ORFs are listed in the genomes is given by the arrow. What result do you expect for the gene plot?





11) Regarding types of error in a normal BLAST search, which is correct:
A) False positives (i.e. non-homologous proteins in the databank that are identified as matches) occur frequently
B) False negatives (i.e. homologous proteins that are not identified as matches) occur frequently.
C) Most false negatives can be identified, if one turns on the filter for low complexity

12) What are the names given to the three domains of life.
A) Prokaryotes, Eukaryotes, Bacteria
B) Prokaryota, Eukaryota, Archaea
C) Archaea, Bacteria, Eukaryotes
D) Plants, Animals, Protozoa

13) The leg-hemoglobin in plants is a homolog to the myo- and hemoglobin molecules found in animals. If all these proteins evolved from the same ancestral molecule that was present in the most recent common ancestor of plants and animals, then, according to Walter Fitch's terminology,
A) the plant globin is an ortholog to the animals myglobin genes and an ortholog to animals hemoblobin genes.
B) the plant homolog is a paralog to the animal hemoglobin genes
C) hemoglobin genes are parlogs to the myoglobin genes

14) In perl, which beginning symbols denote variables that contain Scalars, Arrays, Hashes?
$, @, %

Your name

Practical:

To which order and phylum does *Geodermatophilus obscurus* belong?

order: Actinomycetales
phyum(kingdom): Actinobacteria

Which annotated protein in the *Geodermatophilus obscurus* genome has the highest bitscore when compared to the genome of Frankia CCI3? (give the GI number and annotation line for the match)

the highest scoring pair was
gi|284991635|ref|YP_003410189.1| gi|86738724|ref|NC_007777.1|

both are listed as encoding a Glutamate synthase (and you get the same result using tblastn or using either genome as query).

Use a combination of BLAST and multiple sequence alignment to pinpoint the beginning and end of the inserted domain in 255732025. In your answer, briefly describe your approach and give the sequence of the additional domain.

do blast search, or use blink, get sequences, align in clustalx, cut out inserted sequence:

CLSEDMRVQTNKGFLGLDEVK
DQWRDLKFANYNPETKQIQYLPASNFILKDAANHKMVEFSDYDINSDAHGSFSLFVTDNHDMYVQTGRVD
KEAGDINRIVYEENSEFSKVEASQLVGSGKGIRFTTTAPNGIDIASVASYKQVVSENQQQTFLELYGYWV
GNADKVGETGVTFTAANEANSAWLSKAISELEGKVDGTTITDSKLSALFNGSEQSFAEWVWDLAKDELRS
VVHGFARASGDENKKIYTSSVILRDELVRVLLHAGYTSRFELNATKGWEITYVEDVAQCVNPVLYSDKNV
KVVDDYFGRVWCVTVPTGLIIVQRVVKNAEDVVVKASRPTIVGNC

(from the alignment it is not clear if the first or the last C is part of the insertion sequence)

At http://gogarten.uconn.edu/seqfiles.zip is an archive that contains several folders with FASTA formatted sequence files. Your task is to write a perl script that does two things:

A) Calculate multiple sequence alignments for the sequences located in each of the *.faa files.

B) Count how often the different amino acids occur in each of the multiple sequence faa files. The program should write the results into a nice, human readable tab delimited table.

Do not forget to incorporate some comments into your script. If you are running out of time, sketch the outline of the program in pseudo-code.

Copy the folder seqfiles.zip to the cluster and unpack the archive using the command

```
uncompress seqfiles.zip
```

Include the annotated program with your answers.
(email to jpgogarten@gmail.com Subject MCB5472 EXAM)

While the question asks for a single program, in writing the script, it is a good idea to break down the problem into parts that can be tested independently. If not, everytime you run the script, you need to wait till the multiple sequence alignment is done … .

Your name

For part A:

```
#! \usr\bin\perl
use warnings;
while(defined($file=glob("*.faa"))){
    system("clustalw -align -infile=$file -type=protein");
}
```

For part B:

```
#!\usr\bin\perl
#MCB 5472 Exam Part 2
use warnings;
my %amino_acids;
#We will open each .faa file in the directory
#We will groom the file to remove GI numbers and new lines
#then we will set up an array that goes through the file character by character

while (defined($file=glob("*.faa")))
{
  open(IN, "< $file") or die "cannot open $file:$!";
        my $inseq = "";
        while(defined(my $line=<IN>)){
                chomp($line);
                $line =~ s/\s//g; #removes blank spaces
                if ($line=~/^>/) {} #ignores annotation lines
                else {$inseq .= $line}; #transfers file to $inseq
        }
        close(IN);
        my @seqarray = split(//,$inseq); #makes an array from $inseq
        foreach (@seqarray){
 #Now we will compare each character to all the keys %amino_acids
                                $amino_acids{$_} +=  1;

        }


# Now we must print out the hash

        print "\n\nThe number of times each amino acid occurs in $file is:\n";

        my @keys = keys (%amino_acids);
        my @vals = values (%amino_acids);

        foreach my $key (@keys){
                print"\t$key\t$amino_acids{$key}\n";
                $amino_acids{$key}=0 #Sets value back to 0 for next pass thru
        }

};
```

Your name

```
#MCB 5472 Exam Part 2
use warnings;
my %amino_acids;
#We will open each .faa file in the directory
#We will groom the file to remove GI numbers and new lines
#then we will set up an array that goes through the file character by character
while (defined($file=glob("*.faa")))
{
  open(IN, "< $file") or die "cannot open $file:$!";
        my $inseq = "";
        while(defined(my $line=<IN>)){
                chomp($line);
                $line =~ s/\s//g; #removes blank spaces
                if ($line=~/^>/) {} #ignores annotation lines
                else {$inseq .= $line}; #transfers file to $inseq
        }
        close(IN);
        my @seqarray = split//,$inseq); #makes an array from $inseq
        foreach (@seqarray){
 #Now we will compare each character to all the keys %amino_acids
                                $amino_acids{$_} +=  1;

        }


# Now we must print out the hash

        print "\n\nThe number of times each amino acid occurs in $file is:\n";

        my @keys = keys (%amino_acids);
        my @vals = values (%amino_acids);

        foreach my $key (@keys){
                print"\t$key\t$amino_acids{$key}\n";
                $amino_acids{$key}=0 #Sets value back to 0 for next pass thru
        }

};
```