Questions from the last midterm are at
http://gogarten.uconn.edu/mcb5472_2010/midterm2010_with_answers.pdf

Sample questions:

You just downloaded all contigs of a nearly finished genome and it turns out they are in 48 separate contig files ending in .fna but all the programs you want to work with require them to be in one file. Aside from copying and pasting them into one file what is an easy unix command to put them into one file?

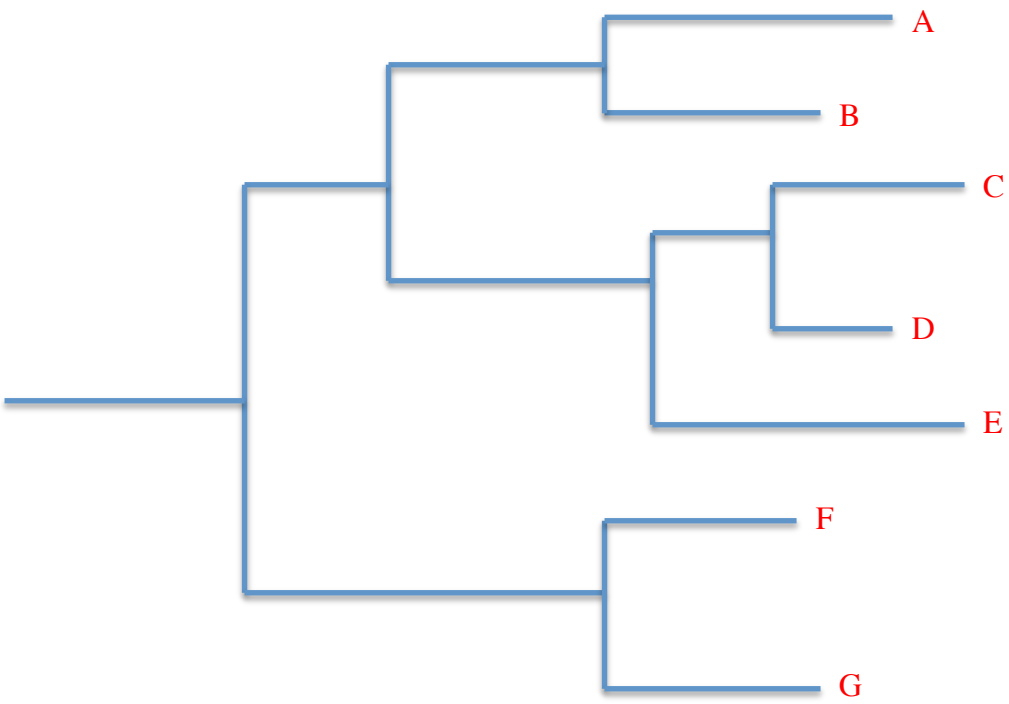
A friend sent you a file to do an analysis with and every time you go to use it a "permission denied" error pops up what is the likely problem and how would you fix it?
What do the following unix commands do ?
ls, cd, mv, mkdir, rm

In perl, which beginning symbols denote variables that contain Scalars, Arrays, Hashes?

In perl, which variable name is use by default in a `foreach (@sequence){}` loop to successively hold the entries from the array?

What is the difference between the most recent common ancestor (MRCA) and last universal common ancestor (LUCA)?

In the following tree mark the MRCA for C and E and for A,B, C ,D, E,



What differences between the clustalw and muscle alignment programs cause one sometimes to be preferred over the other, and which is the one results in more 'publishable' alignments?

In Perl, when using strict, what must be put in front of a variable denoting it a new variable?

What is the advantage of using variables that are valid only in part of prescript?

Why may maximum likelihood reconstruction produce more reliable phylogenetic reconstructions as compared to parsimony?

Under which conditions can the concept of a synapomorphy applied to the analysis of molecular data?

Define the following terms:
Orthologs
Paralogs
Xenologs
Synologs

Which of these - synolog, xenolog, ortholog, analog, and paralogs - are homologs?

What is the difference between a synolog and a xenolog?

Can two genes (gene A and gene A') in genome I both be orthologs to a single gene B in genome II?

What is the difference between paraphyletic and holphyletic?

Which of these are considered monophyletic sensu Hennig?

What is the difference between a paraphyletic and holophyletic taxonomic group?

What is bootstrap support and how do you calculate it?

What is the difference in parametric and non-parametric bootstrapping?

Which of these is applied routinely in many approaches to phylogenetic reconstruction?

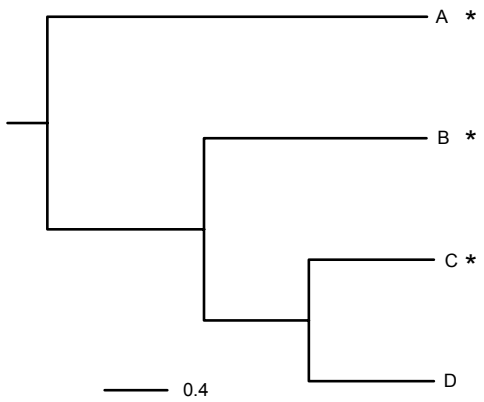
What is long-branch attraction?

How might you mitigate its effects in your phylogenetic analysis?

How are support values for branches calculated in a Bayesian analysis using MCMCMC?

How are these support values usually called?

In this tree, A, B, C, and D represent a monophyletic group. Which taxa are sister to each other?



Would it be correct to label A as an older species?

This tree shows that B, C, and A all share the "star" trait but D does not.

a. What is the most parsimonious explanation for these results? Draw on the tree where this trait would have been acquired.

b. What is an alternative explanation?

c. This tree was built with nucleotide data, what does the 0.4 mean next to the horizontal line?

d. Approximately how different are Lineage A and B?

Are genealogies the same thing as phylogenies? Explain why or why not.

Which type of characters are useful for taxonomic grouping, synapomorphies or sympleisiomorphies?

Do symplesiomorphies define paraphyletic groups?

List and describe at least three different types of BLAST searchings methods.

Describe the query and database for each of the following BLAST searches: blastp, blastn, blastx, tblastn, tblastx

You want to find all integrons containing an integrase in a genome (you have both the nucleotide sequence and the aa sequences encoded by the genome available as searchable databanks). You are also interested in pseudogenes of the integrase. Which blast search algorithm could you use, (blastn, blastp, blastx, tblastn, tblastx)

Why would one use PSI BLAST rather than BLAST?

Why is an E-value of 10^{-40} in the 4^{th} iteration of a PSI blast search not considered proof of homology?

If the integrases are not highly conserved in sequence, what could you do to improve your chances for finding all integrases present in the genome?

Draw a geneplot of two hypothetical genomes that have two inversions between them.

Name three alignment programs

What is the difference between a local and a global alignment program?

What is the GC skew of a genome?

G pairs with C, how can it be that there are more Gs than Cs in a DNA?

In depiction of the cumulative GC skew, one often find a single peak and a minimum. What do these extrema usually identify?

How does the plot of cumulative GC skew change as consequence of the recent inversion?

How would a wrongly annotated origin of replication reveal itself in a GC skew analysis?

How does a gene plot change as consequence of the recent inversion in one of the genomes?

How does a gene plot change as consequence of a wrongly annotated origin of replication in one of the genomes?

List three different approaches to phylogenetic reconstruction. How do they work?

Why has phylogenetic reconstruction using Bayesian statistics become popular over the last few years?

What are the names of the three domains of life?

Who is credited with having discovered the Archaea?

What can be used as an outgroup to root a phylogeny that includes homologs from all living cellular organisms?

What other approaches can you use to root a molecular phylogeny?

Why is percent identity not a good parameter to assess the significance of a match in a blast search?

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

Homologous sequences almost always show significant similarity

Proteins with moderately complex amino acid composition that show significant similarity in a BLAST search almost always are homologous.

Give a short definition of E-value and P-value:

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
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEMGLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRR
PLS \n

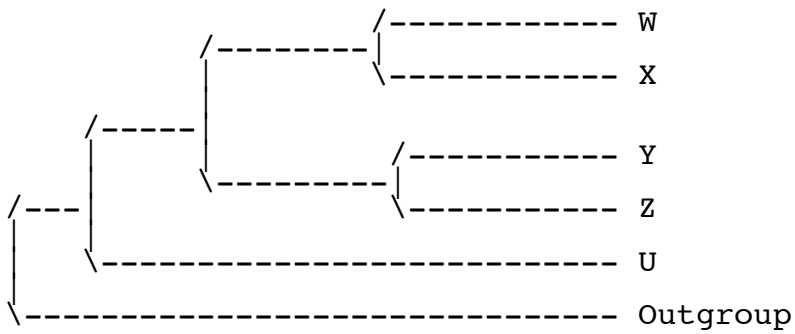
>gi|12643370|sp|Q9P997.2|VATA_THEAC this is a test sequence RecName: Full=Tac atpA intein \n
MGKIIRISGPVVVAEDVEDAKMYDVVKVGEM \n
GLIGEIIKIEGNRSTIQVYEDTAGIRPDEKVENTRRPLS \n

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

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

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

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



In the rooted tree depicted above the group W+X represents a:
A) Clade; B) Paraphyletic group; C) Monophyletic group; D) A and C

E) None of the above.

In the rooted tree depicted above the group U+Z+Y represents a:
A) Clade; B) Paraphyletic group; C) Monophyletic group
D) A and C; E) None of the above.

If the output of a consensus tree analysis says

```
Species in order :
Outgroup
U
W
X
Y
Z
```

What branches are indicated by the following lines:

```
**..**  88
****..  70
*.**..  10
```

You compare gene families form the same set of species. List the reasons why the phylogenies of individual families might not agree with one another?

Practical portion

Write a script to calculate multiple sequence alignments for the sequences located in each of the *.faa files; and to calculate a maximum likelihood tree using phyml (assume that phyml is installed on your computer)

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

Which annotated protein in the *Thermus thermophilus* HB8 genome has the highest bitscore when compared to the genome of *Thermotoga maritima*.
(give the GI number and annotation line for the match).