Introduction to Bioinformatics – AS 250.265
Assignment 9

The "post-genomic" era has ushered in a deluge of new bioinformatics tools that are able to perform tasks that were inconceivable to scientists from 100 years ago. In your lab, you explored the how one can construct a phylogram using the PAUP software. In this assignment, we will explore additional concepts in genomics and molecular evolution. You will get practice thinking about the size of the human genome as well as explore a simple algorithm for constructing phylogenetic trees. These concepts will be useful in the final part in the course when we examine how all of the information we have learned can be used in studying and curing human disease.

This assignment is worth 100 points toward the homework portion of the class. The laboratory questions you do in class constitute 35 of those points, and this handout makes up the remainder. Both parts are due on April 28 at the start of class. Electronic portions of the assignment should be submitted through the course website with appropriate filenames (e.g. `hw09-q4a.gif`). The written portions of the assignment may be submitted by hand or uploaded as a Word document or PDF file.

Question 1: Laboratory Number 6 (35 points)

Submit your solutions the laboratory questions on gene trees that you completed in class.

Question 2: Exploring the Human Genome (15 points)

In this part of the assignment, you will examine some practical repercussions of the size of the human genome.

a. An order of magnitude rate of DNA polymerase is 500 bp per second. Given that S-phase (the part of the cell cycle where the DNA is replicated) in human somatic cells takes approximately 12 hours, how many replication origins are needed? (5 points)

b. Currently, the human genome is thought to have approximately 25,000 genes. Estimate the percentage of DNA that is not expressed (in introns). Write down the assumptions you make in your calculation. (5 points)

c. How many times would we have to sequence the human genome in order for the error rate to be 1 in 20,000 bases? (5 points)

Question 3: DNA Sequencing (10 points)

Answer the following questions about determining the sequence of DNA using Sanger's dideoxy nucleotide method.

a. What is a dideoxy nucleotide? Why is it important in the Sanger method for DNA sequencing. (2 points)

b. You have applied shotgun sequencing to a DNA of interest and have identified the sequences below. Piece these sequences together into the original sequence. (5 points)

```
gacctgtttagg
ctagaccaagca
aacggacaacta
agcacattataa
gacaactagacc
ttataacggaca
```

c. What problems would you have to consider if you wanted to implement a computer program to piece together sequencing as you did above? Remember that real experimental data is rarely as ideal as data given on problem sets. (3 points)

Question 4: Phylogenetic Trees and Transitional Forms (20 points)

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a protein involved in metabolism and is central to the glycolytic pathway of many different organisms. Because it is nearly ubiquitous in higher organisms, it is possible to construct a phylogram using GAPDH sequences similar to the tree shown on page 400 of your book.

a. Using the ClustalW program on the Biology Workbench website, perform a phylogenetic analysis of glyceraldehyde-3-phosphate dehydrogenase. In your analysis, use the following organisms:

Achlya bisexualis
Dictyostelium discoideum
Encephalitozoon cuniculi
Entamoeba histolytica
Giardia intestinalis
Homo sapiens
Mus musculus
Naegleria andersoni
Paramecium tetraurelia
Porphyra yezoensis
Rattus norvegicus
Trichomonas vaginalis
Zea mays

You can obtain the sequences for these organisms on the NCBI or Biology Workbench website. Not all of them will have RefSeq protein sequences, but you should use them were possible. When you have finished, you can import them (individually, or by pasting them into a single MFASTA file) into Biology Workbench. Name your sequences by the genus (e.g. *Mus*)

Submit the image of your unrooted tree to the course webserver. You may use the default ClustalW options. (5 points)

b. Compare your phylogram to the eukarotic branch of the phylogram on page 400 of your book (figure 12.1). What aspects are different between the two trees? Given that the tree on page 400 was generated from small subunit rRNA, why do you think this tree might be different? (5 points)

c. On the assignment 9 website, you will a FASTA file `artificial.fasta` that has been constructed to be an artificial "ancestor" of both mouse and rat. It is a sequence that is transitional between the two proteins in the sense that it is exactly four residues different from both.

Regenerate the tree from part (a), including this artificial sequence. You may omit the following phyla: *Giardia, Trichomonas, Ecephalitozoon,* and *Entamoeba*. Submit your new unrooted tree to the course web server. (4 points)

d. If, in the future, we were able to obtain the DNA sequences for *bona fide* transitional forms, how do you think the present-day phylogeny software would handle the sequences? Why is this? (6 points)

Question 5: Unweighted Pair Group Method with Arithmetic Mean Trees (20 points)

The algorithm for the UPGMA method is given in your book on pages 380-382. For this question, you should create a tree using the UPGMA method for the distance matrix below. Your final solution should include not only the final tree itself (drawn approximately to scale), but also the distance matrix and partial tree after each iteration of the algorithm. Your distance matrices should be labeled so that the merged OTUs are clearly identifiable (one idea would be using letters for the sequences and numbers for the branch points). As an example of building up partial trees, see figure 11.17 on page 382.

In the distance matrix below, A, B, C, D, and E are hypothetical species.

	A	**B**	**C**	**D**	**E**
A	0	5	1	4	8
B	5	0	4	2	1.5
C	1	4	0	3	3
D	4	2	3	0	4
E	8	1.5	3	4	0