PLNT4610/7690 BIOINFORMATICS

FINAL EXAMINATION

Wednesday December 13, 2017 13:30 to 15:30 Engineering E2-160, seats 1 - 22

Answer any combination of questions totalling to exactly 100 points. The questions on this exam total to 120 points. If you answer questions totalling more than 100 points, answers will be discarded at random until the total points equal 100. This exam is worth 20% of the course grade.

Hand in the question sheets along with your exam booklet. All questions must be answered in the exam book. The question sheets will be shredded after the exam.

Ways to write a readable and concise answer:
i. Just answer the question. Save time by specifically addressing what is asked. Don't give irrelevant background if it doesn't contribute to the question that was asked.
ii. Avoid stream of consciousness. Plan your answer by organizing your key points, and then write a concise, coherent answer. Make your point once, clearly, rather than repeating the same thing several times with no new information.
iii. Point form, diagrams, tables, bar graphs, figures are welcome. Often they get the point across more clearly than a long paragraph.
iv. Your writing must be legible. If I can't read it, I can't give you any credit.

---

1. (10 points)

a) Briefly describe the aspects of High Performance Computing (HPC) systems that distinguish them from desktop computing systems.

b) What is the distinction between serial computing and parallel computing?

2. (10 points) Suppose you were trying to construct a phylogenetic tree for an enzyme such as phenylalanine ammonia lyase (PAL). When you search for protein sequences to do the alignment, it is often the case that the identical protein has been sequenced two or more times by different projects. Aside from the trivial reason that a smaller dataset takes fewer computational resources, why is it important to remove duplicate copies of a protein from the dataset?
At which point does it make the most sense to eliminate duplicates: prior to doing the multiple alignment, or before constructing the phylogenetic tree from the alignment?

3. (10 points) Below is a generalized statement of Bayes theorem. When used for phylogenetic analysis, explain in words the meaning of Model and Data. In other words, what do they represent in phylogeny? Next, when applied to a phylogenetic tree inferred from a multiple sequence alignment, explain the meaning of each of the four probability terms in the equation.

$$P(Model|Data) = \frac{P(Data|Model)\ P(Model)}{P(Data)}$$

4. (10 points) Explain how the following equation is used to determine the number of reads necessary to sequence a genome. Make sure to define each of the terms N, C, P and f.

$$N = C \frac{\ln(1-P)}{\ln(1-f)}$$

5. (10 points) The N50 value is the most common parameter for evaluating a genome assembly.

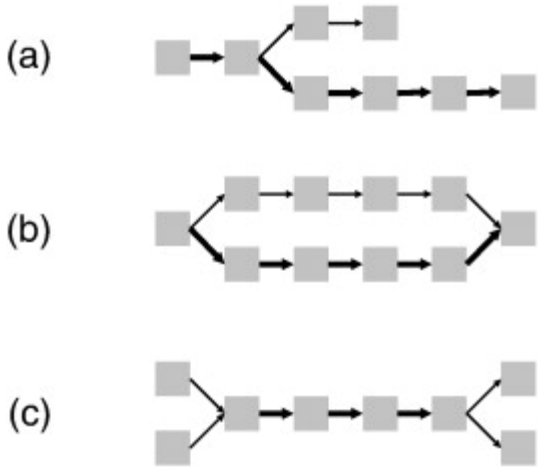
a) Define N50.

b) In some cases, an N50 decreases if, prior to doing the assembly, we correct errors in reads using a programs such as Quake or Pollux. What does the decrease in N50 value tell us?

6. (15 points)

i) Draw a de Bruijn graph for assembly of a short contig (eg. 5 k-mers) in which the input reads had no errors, and the assembly was perfect.

ii) The three graphs at right illustrate three possible de de Bruijn graphs. Briefly explain what these graphs indicate.



iii) How do sequence assembly programs generally solve the assembly problems seen in the above graphs?

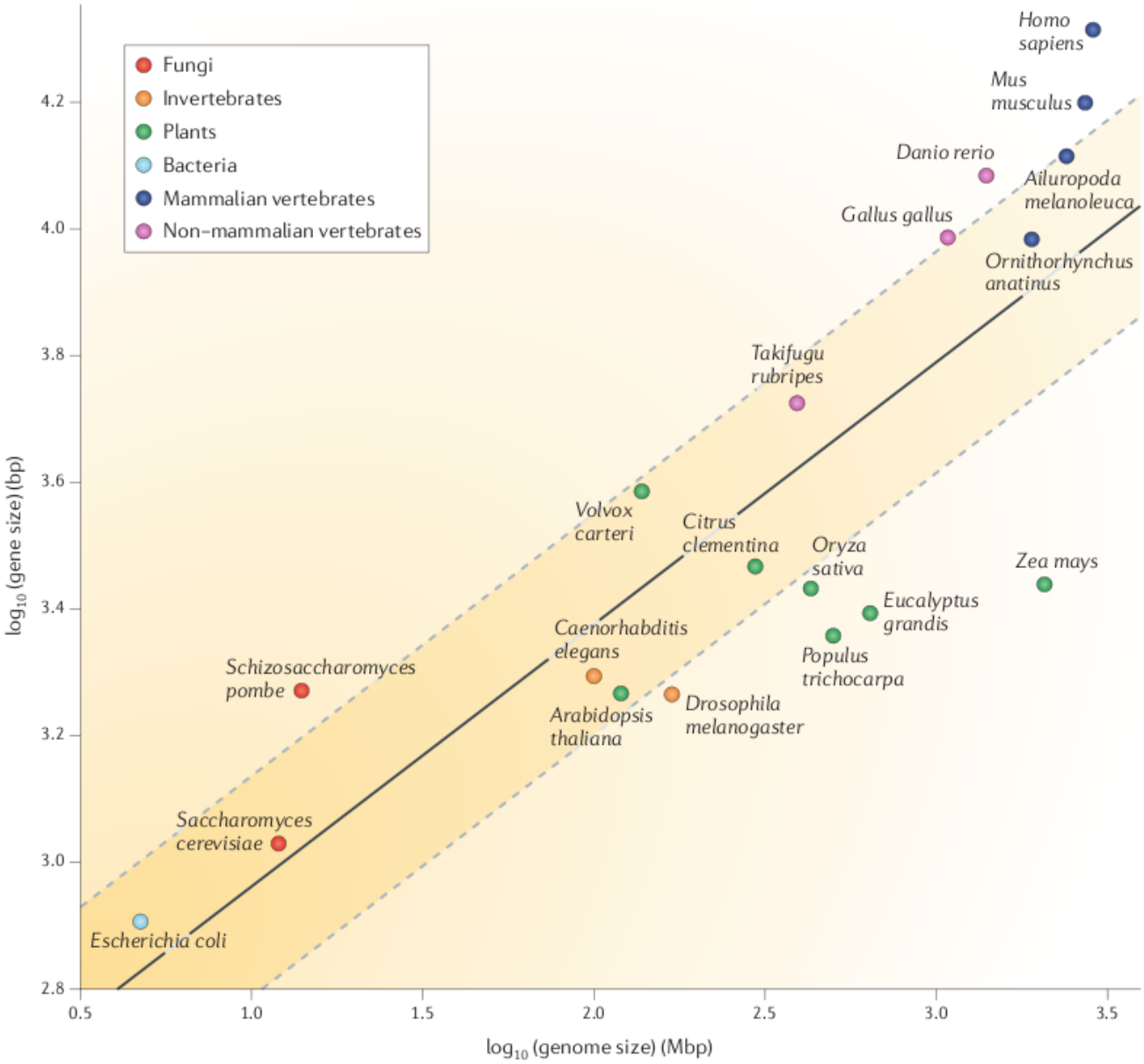
7. (5 points) Although it is possible to annotate a genome for a eukaryotic species, without also having transcriptome data, how does transcriptome data improve the gene annotation process?

8. (5 points) Before constructing libraries for RNAseq transcriptomics experiments, RNA samples are purified on an oligo-dT column to eliminate rRNA, resulting in purified mRNA. What would be the problem with measuring gene expression, if this step were omitted?

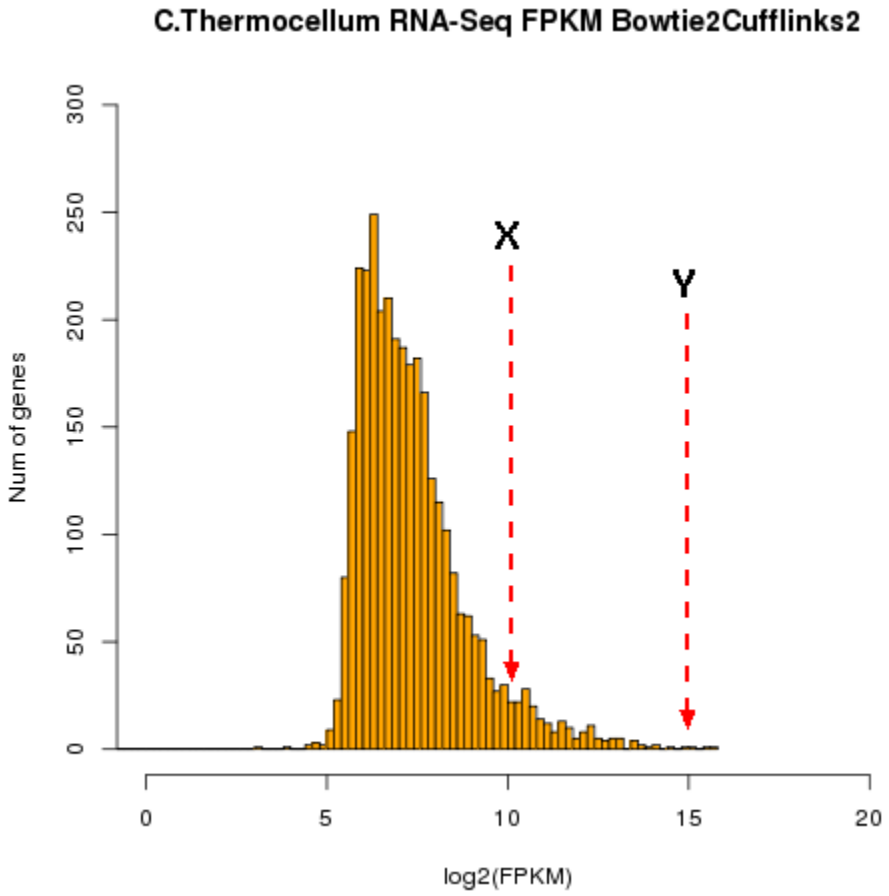
9. (15 points)

a) Explain why genome assembly is so much more difficult for eukaryotic genomes than for prokaryotic genomes.

b) Based on data in the figure below, explain why genome annotation is more challenging for eukaryotic genomes, compared to prokaryotic genomes.

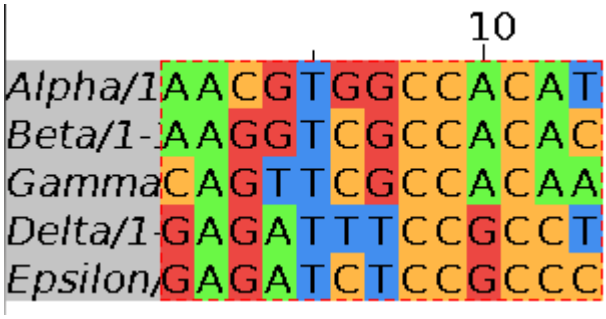


10. (5 points) Two groups of genes in an RNA-Seq experiment are pointed to by X and Y. What is the difference in expression levels between X and Y? For full credit, you need to specify a numerical ratio between X and Y, rather than just saying that one is expressed at a higher level than another.



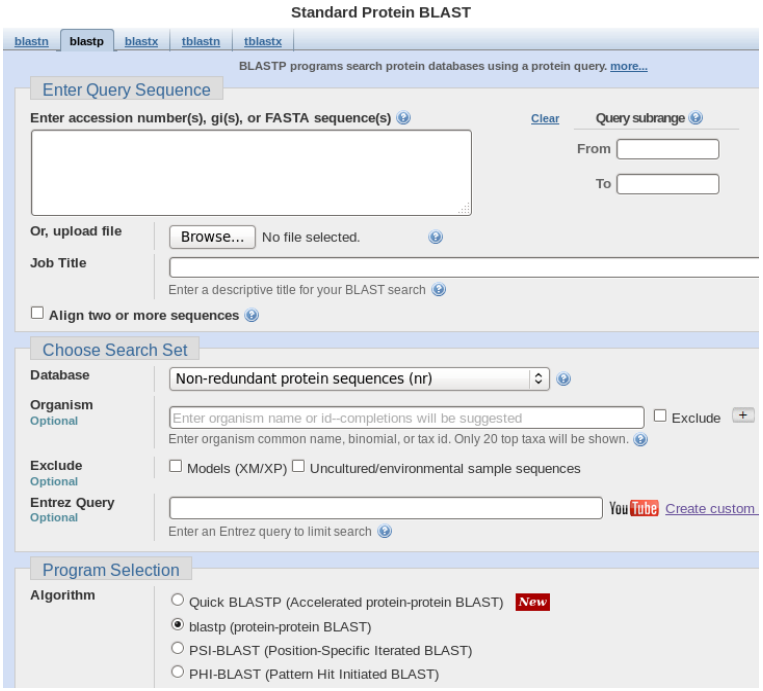
11. (5 points) What is the distinction between a program and an algorithm. Give an example.

12. (10 points) A multiple sequence alignment is shown for five DNA sequences, each 13 nt long. When considering construction of a phylogenetic tree, which positions are the most informative? Which positions are the least informative. Explain your reasoning.



13. (10 points) In class we talked about two examples of client/server programs: NCBI BLAST, and the Jmol 3D protein structure viewer at the PDB web site. Screenshots for each are shown below. Both use the client-server model, in which some steps of the task run on the user's compter, and some run on the server. In each case, describe which of the major steps run on the user's computer, and which steps are run on the server. Explain the reason for why these steps are allocated to either the client or server, in each of the two cases.

NCBI BLAST



JMOL