PLNT4610/7690 BIOINFORMATICS

FINAL EXAMINATION

Wednesday December 19, 2018 13:30 to 15:30 Agriculture 130

Answer any combination of questions totalling to <u>exactly</u> 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.

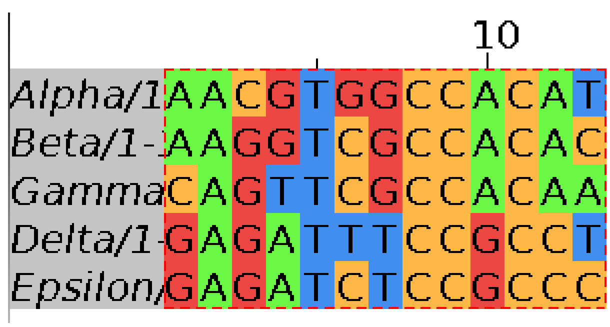
Hand in the question sheets along with your exam booklet. All questions must be answered in the exam book. The question sheets will be shredded after the exam.

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.

---

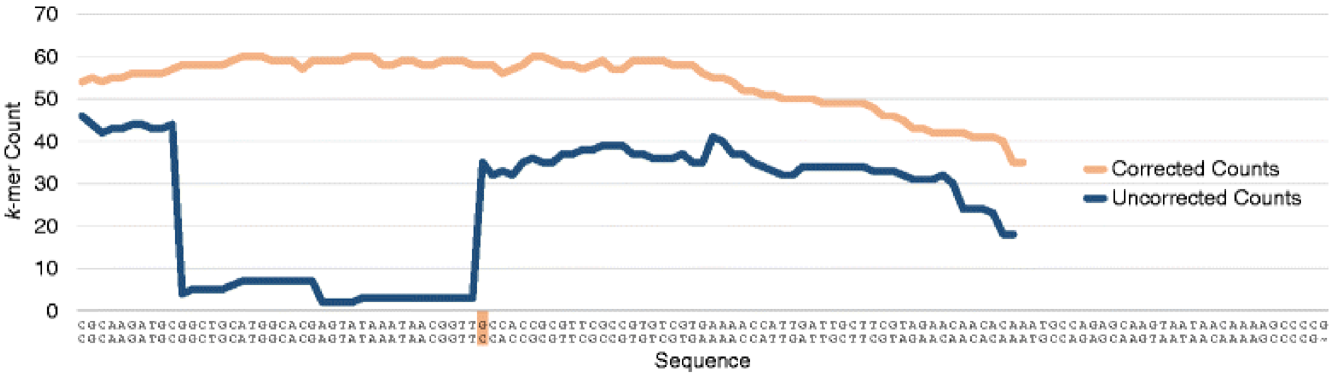
1. (5 points) Before constructing libraries for RNAseq transcriptomics experiments, RNA samples are purified on an oligo-dT column to eliminate rRNA, resulting in purified mRNA. What would be the problem with measuring gene expression, if this step were omitted?

2. (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.



3. (5 points) In a _______a________ database, all data are represented in records, forcing the data into the structure of a single data type. In a relational database, all data are organized into ________b_______. Links between tables are referred to as ________c_______. In an object-oriented database, data is organized as classes, which have _______d________ and _______e________.

4. (10 points) Using the accompanying figure as a guide, describe how the Pollux algorithm finds errors in sequencing reads and corrects them.



5. (10 points) In RNAseq experiments, gene expression is typically calculated by the equation

FPKM (Fragments Per Kilobase per Million mapped fragments) = F/CN

Expression of two genes was measured as follows

	control cells (5 million mapped fragments)	hormone-treated cells (7 million mapped fragments)
	number of fragments	number of fragments
geneA (length =1.8 kb)	15,000	20,000
geneB (length =5 kb)	8000	11,000

a) Calculate the FPKM values for genes A and B in control and hormone-treated cells. In a table similar to the one below, show the calculations and the FPKM values.

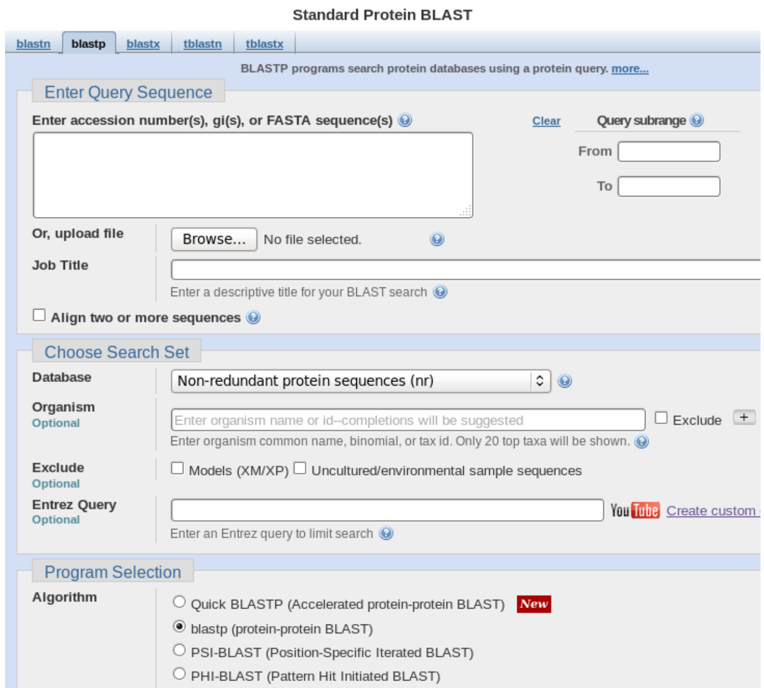
	FPKM	
	control	hormone-treated cells
geneA		
geneB		

b) What do these results tell you about the response of genes A and B to hormone treatment?

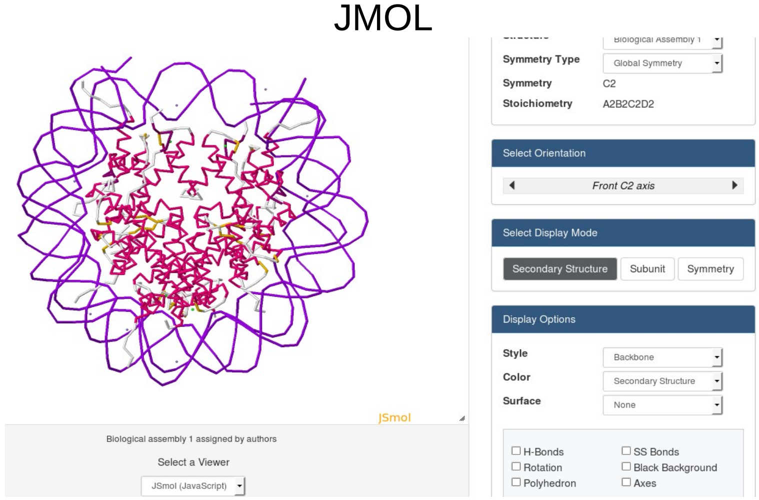
6. (5 points) In RNAseq experiments, when deciding whether genes are differentially expressed between two conditions, what is the distinction between the Power of the statistical test, and the False Discovery Rate?

7. (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



8. (10 points) For **n** sequences, the number of possible tree topologies is given in the table.

a) In principle, one could construct phylogenies using an exhaustive algorithm that considered every possible tree topology. How much more computing time would it take to calculate a phylogeny for 50 sequences, compared to 10 sequences?

b) Suppose that you had access to a High Performance Computer cluster with 100,000 CPUs. Discuss whether or not that HPC cluster could be used for performing exhaustive phylogenetic analysis. Put another way, to what extent would the speedup from 100,000 CPUs help for these types of problems? Show calculations to support your conclusions.

# of sequences **n**	# of trees $\prod_{i=3}^{n} (2i-5)$
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
50	2.8×10^{74}

9. (10 points) In your exam booklet, create a table similar to the one below, and fill in the appropriate values eg. faster or slower, yes or no. For the example in c), give the name of one distance method, and one character method.

	Phylogenetic Method	
	Distance	Character
a) speed (faster/slower)		
b) accuracy (more accurate/less accurate)		
c) example (name of method)		
d) can detect homoplasies (yes/no)		
e) can reconstruct ancestral sequences (yes/no)		

10. (10 points) The main goal of genome annotation is to create models of features in a genome. In particular, annotation pipelines try to create "gene models" for each gene in the genome.

a) Briefly describe the types of information annotation pipelines use to create gene models,
b) Briefly describe what information a gene model contains.

11. (10 points) Volcano plots show, for every gene, the log of the fold change between two conditions or treatments on the X-axis, and the log of the p value on the Y axis. In other words, the X-axis shows the degree to which the expression of that gene has either increased or decreased between the two treatments, and the Y axis shows how significant the change is for that gene, based on the change seen in different replicates for that gene.

What is different about the two experiments shown in Fig. 1 and Fig. 2 below?

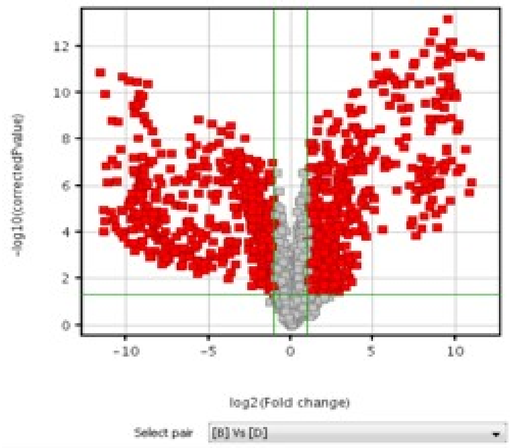


Figure 1

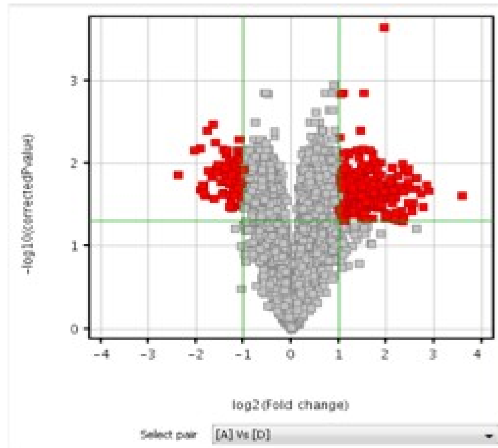


Figure2

12. (5 points) A series of random DNA sequences was constructed, each with a different percentage of AT bases. and sequences were compared using several phylogeny methods. In each case, 100 bootstrap replicates were done. The results are presented in the table below.

ran20 - 20% AT ran35 - 35% AT ran50 - 50% AT ran65 - 65% AT ran80 - 80% AT	Parsimony

```
            +-----------------------ranAT35
  +-------|
  |       |       +---------------ranAT50
  |       +-100.0-|
  |               |       +-------ranAT80
  |               +-100.0-|
  |                       +-------ranAT65
  |
  +-------------------------------ranAT20
```

Consider the parsimony tree. Explain why the sequences group as they do in this tree.

13. (10 points) In order to get the best possible genome assembly for the fungus *Rhodosporidium toruloides*, the same set of sequencing reads were assembled several times using different assembly programs. Two of the assemblies are shown below. Which is the better assembly for this genome? Cite evidence to support your conclusion

Rhodosporidium toruloides genome assemblies			
SRR516824		SRR516830	
Assembly	contigs	Assembly	contigs
# contigs (>= 0 bp)	2056	# contigs (>= 0 bp)	2930
# contigs (>= 1000 bp)	610	# contigs (>= 1000 bp)	2155
# contigs (>= 5000 bp)	374	# contigs (>= 5000 bp)	1247
# contigs (>= 10000 bp)	318	# contigs (>= 10000 bp)	707
# contigs (>= 25000 bp)	231	# contigs (>= 25000 bp)	143
# contigs (>= 50000 bp)	139	# contigs (>= 50000 bp)	8
Total length (>= 0 bp)	20261352	Total length (>= 0 bp)	20156200
Total length (>= 1000 bp)	19959191	Total length (>= 1000 bp)	19822018
Total length (>= 5000 bp)	19440025	Total length (>= 5000 bp)	17409372
Total length (>= 10000 bp)	19020132	Total length (>= 10000 bp)	13480241
Total length (>= 25000 bp)	17590721	Total length (>= 25000 bp)	4804497
Total length (>= 50000 bp)	14282511	Total length (>= 50000 bp)	477339
# contigs	790	# contigs	2452
Largest contig	538739	Largest contig	69119
Total length	20084923	Total length	20037245
GC (%)	61.95	GC (%)	61.85
N50	80570	N50	14509
N75	44935	N75	8031
L50	71	L50	423
L75	156	L75	880

14. (10 points) Databases turn data into knowledge. Using the example at right, describe how this database turns data into knowledge. Make sure to use precise terminology associated with databases in your description.

Note: The Type field enumerates the host response to the pathogen, in this case Pattern Triggered Immunity.