PLNT4610/7690 BIOINFORMATICS

FINAL EXAMINATION

Monday December 19, 2022 09:00 to 11:00 Frank Kennedy Gold Gym seats 227 - 242

Answer any combination of questions totalling to exactly 100 points. There are 12 questions on this exam totaling 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.

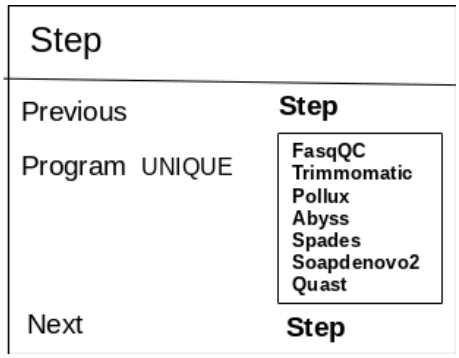
Hand in the question sheets along with your exam booklet. All questions must be answered in the exam book. The question sheets will be shredded after the exam.

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 schema for a database describing the succession of steps in bioinformatics workflows includes a class called Step, illustrated at right. What is the obvious flaw in this class as it is defined? Design a better Step class (and other classes if necessary) that solve that problem.



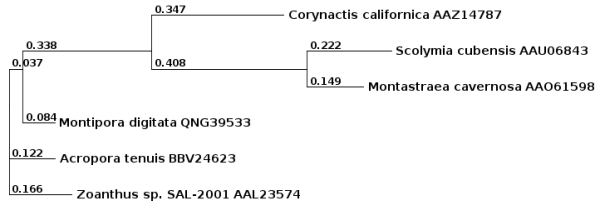
2. (5 points) What is the distinction between a program and an algorithm. Give an example.

3. (5 points) Pollux detects errors in DNA sequencing reads based by only including "trusted" k-mers in a read. Trusted k-mers are k-mers which appear at roughly the same frequency in the genome as the coverage. When scanning along a read, any sudden dip in k-mer frequency will mark the position of a sequencing error. Explain why this strategy cannot be used in correcting RNA sequencing reads.

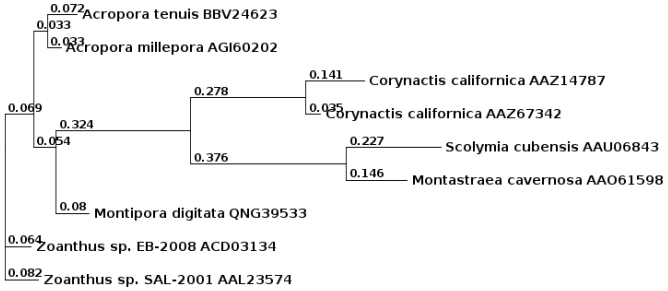
4. (15 points) A dataset of 9 red fluorescent proteins (RFP) has been chosen for phylogenetic analysis using the maximum likelihood method implemented in PROML. Several variations of the phylogeny workflow were:

i) cd-hit ---> MAFFT ---> Gblocks ---> PROML
ii) cd-hit ---> MAFFT ---> PROML
iii) MAFFT ---> Gblocks ---> PROML
iv) MAFFT ---> PROML

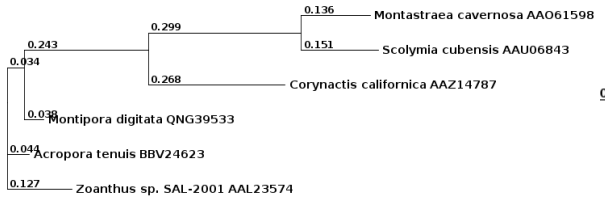
A



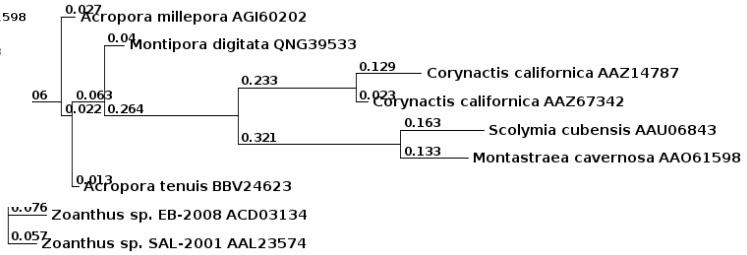
B



C



D



Branch lengths are indicated on the trees. Line lengths are not necessarily to scale.

a) For each workflow i - iv, indicate the letter (A-D) of the tree produced by that workflow.

i)
ii)
iii)
iv)

b) Why would it not be valid to make a maximum likelihood tree omitting the MAFFT step?

5. (10 points) The output from top is shown on two different machines, venus and cc11. What are the differences between the machines, with respect to which is most busy, free RAM, users, and programs that employ parallel processing? Cite evidence from the top output to support your conclusions.

venus

```
top - 10:45:39 up 69 days,  3:38, 21 users,  load average: 0.33, 0.21, 0.24
Tasks: 1403 total,   1 running, 1401 sleeping,   1 stopped,   0 zombie
%Cpu(s):  0.1 us,  0.2 sy,  0.2 ni, 99.6 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 26394057+total, 11791680 free, 17318692 used, 23483020+buff/cache
KiB Swap:  8388604 total,  8388596 free,        8 used. 24320516+avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
29089 root      20   0  240652  12328   2308 S   7.8  0.0 204:20.82 python-thi+
45836 frist     24   4  442524 280048  56504 S   7.5  0.1 117:42.74 Xvnc
11517 malhotr3  24   4 9217080 159212  33620 S   2.6  0.1 355:34.09 spssengine
51177 lutze     24   4 7964404 241088  80548 S   1.6  0.1 363:31.65 gnome-shell
17031 frist     24   4  174040   3860   1796 R   1.3  0.0   0:00.44 top
46587 frist     24   4 3557992 379128 106312 S   1.3  0.1  27:22.76 thunderbird
38418 beheshti  24   4   73.5g   2.7g 345892 S   1.0  1.1  55:33.08 MATLAB
51663 lutze     24   4   36.6g   1.8g 283444 S   1.0  0.7 150:14.43 MATLAB
11393 malhotr3  24   4 9168428 537084  30136 S   0.7  0.2  57:54.58 STATISTICS
52439 lutze     24   4   35.2g   2.2g 339916 S   0.7  0.9 170:20.23 MATLAB
    9 root      20   0       0      0      0 S   0.3  0.0  42:52.86 rcu_sched
 8167 frist     24   4   38.6g 298204 107852 S   0.3  0.1   0:25.98 soffice.bin
```

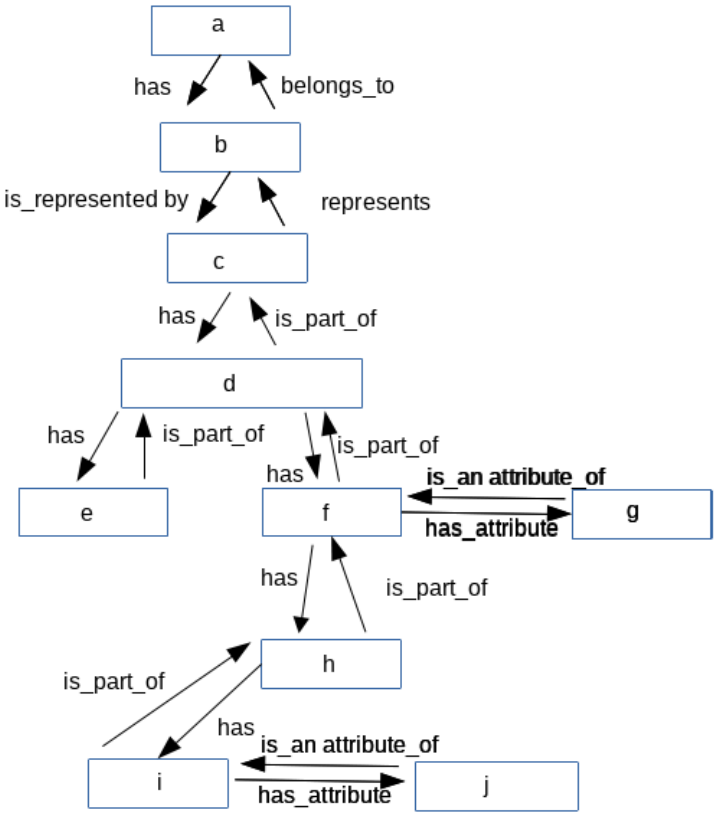
cc11

```
top - 10:46:53 up 80 days, 18:41,  3 users,  load average: 59.48, 59.59, 57.66
Tasks: 705 total,  63 running, 642 sleeping,   0 stopped,   0 zombie
%Cpu(s):  1.1 us,  2.9 sy, 90.6 ni,  5.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem : 26394057+total, 20284611+free, 38164912 used, 22929560 buff/cache
KiB Swap:  8388604 total,  8388604 free,        0 used. 22345259+avail Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND
49126 fangx     24   4 1105320 605784  26956 R 100.0  0.2  52:09.38 dns_input
49133 fangx     24   4 1105556 606536  26976 R 100.0  0.2  52:14.08 dns_input
49135 fangx     24   4 1105296 605416  26980 R 100.0  0.2  52:12.42 dns_input
49139 fangx     24   4 1104524 604628  26960 R 100.0  0.2  52:13.40 dns_input
49147 fangx     24   4 1101992 602132  26980 R 100.0  0.2  52:08.98 dns_input
49152 fangx     24   4 1102592 601476  26952 R 100.0  0.2  52:13.06 dns_input
49156 fangx     24   4 1102064 600956  26964 R 100.0  0.2  52:11.16 dns_input
49158 fangx     24   4 1101784 602620  26964 R 100.0  0.2  52:10.76 dns_input
49165 fangx     24   4 1100060 600472  26940 R 100.0  0.2  52:10.37 dns_input
49168 fangx     24   4 1101056 600028  26972 R 100.0  0.2  52:12.03 dns_input
49169 fangx     24   4 1100896 600928  26972 R 100.0  0.2  52:11.87 dns_input
49176 fangx     24   4 1098652 598892  26952 R 100.0  0.2  52:13.26 dns_input
```

6. (10 points) What is the distinction between a spreadsheet and a database? In other words, why is a spreadsheet not a database?

7. (10 points) Matching - An ontology for genome assembly is shown. For each box in the DAG, choose the appropriate term.



read
species
scaffold
spacer (Ns)
contig
sequence
quality scores
genome
orientation
assembly

a)
b)
c)
d)
e)
f)
g)
h)
i)
j)

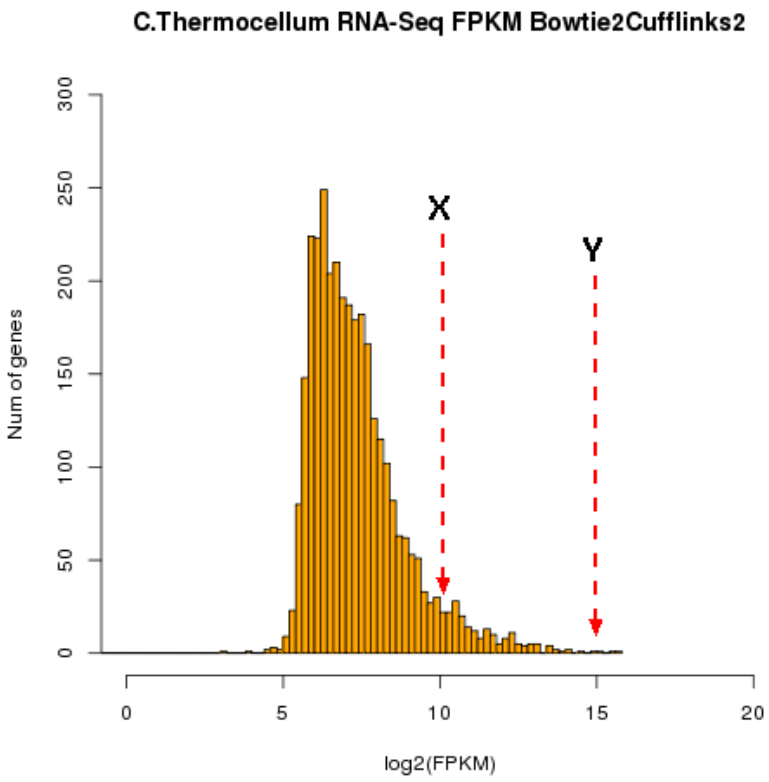
8. (10 points) Statistics for a genome assembly and a transcriptome assembly are compared for a fungus. A - E, choose the most reasonable column heading from the list below.

	A	B	C	D	E
genome	5 x 10^7	1,000,000	100	500,000	5000
transcriptome	1 x 10^7	16,000	55,000	3500	200

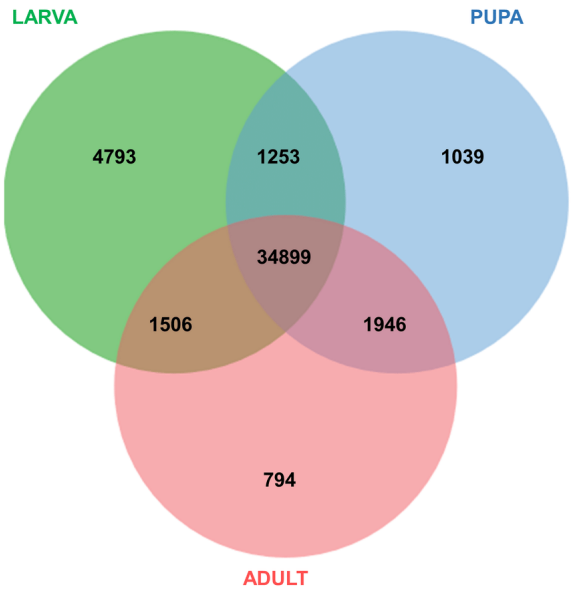
number of contigs smallest contig largest contig N50 total size (bp)

A)
B)
C)
D)
E)

9. (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.



10. (10 points) For each of the three main stages in the life cycle of the firefly (*Sclerotia aquatilis)* the number of distinct transcripts from RNA sequencing are shown.



One or more of the following statements is incorrect or misleading. Others are correct. For each of the incorrect statements, briefly explain why it is incorrect. For statements that are correct, simply state "correct".

a) The total number of genes in the *S. aquatilis* genome is the sum of the numbers in the Venn diagram, or 46,230.

b) Only 794 genes are expressed in adults.

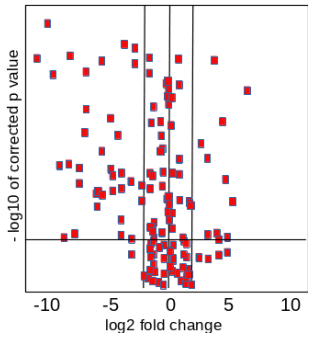
c) The vast majority of genes (34899) are transcribed at the same level in all 3 stages.

d) 1253 distinct transcripts are found in Larvae and Pupae, but not in adults.

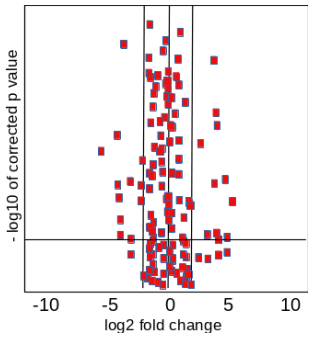
e) 794 distinct transcripts are found in Adults, that are not seen in Larvae and Pupae

11. (10 points) Match each volcano plot with one of the five statements.

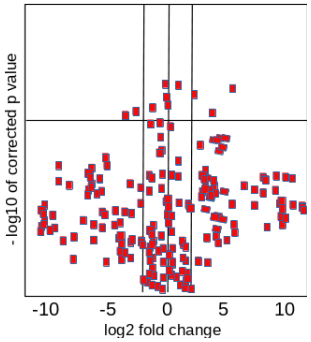
A



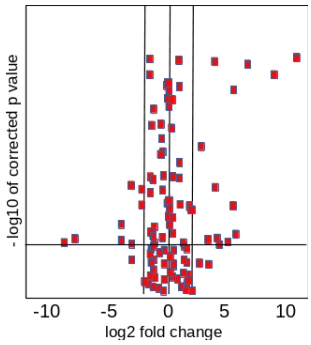
B



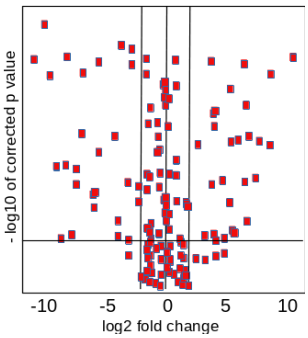
C



D



E

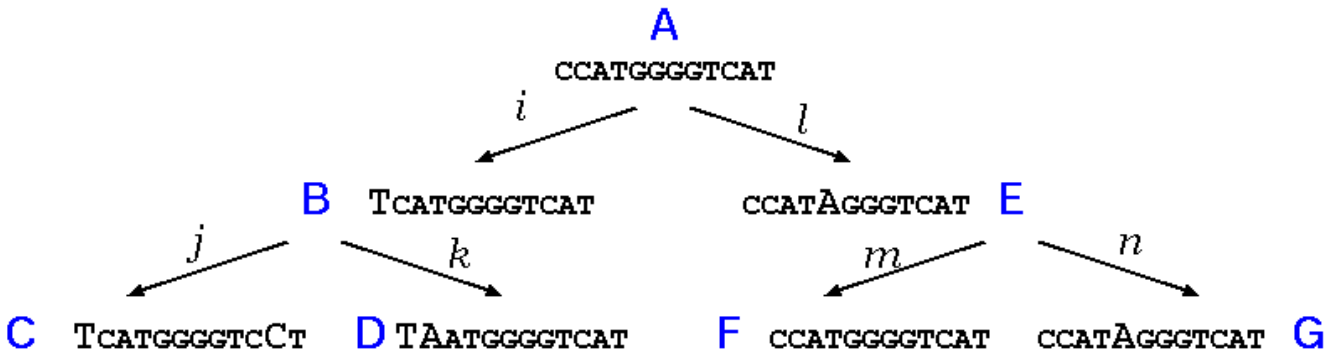


i) Roughly an equal number of genes increase and decrease.
ii) Most changes in gene expression are due to up-regulation of a few genes.
iii) There was so much experimental variation that nothing can be concluded from this data.
iv) The two conditiona are almost identical, with respect to gene expression.
v) Most of the change in gene expression is due to down-regulation.

Answer in order

i)
ii)
iii)
iv)
v)

12. (20 points) The diagram below illustrates the evolution of a DNA sequence through several speciation events. Point mutations that differ from ancestor A are shown as larger letters, compared to the rest. Branch lengths are labeled i - n.



a) Although the complete tree is shown above, in a real world situation, we would never know the ancestral sequences A, B and E. Fill in the distance matrix below with pairwise distances for C,D, F and G, where pairwise distances are simply the number of mutations needed to convert one sequence into the other.

	C	D	F	G
C				
D				
F				
G				

b) Redraw the tree above, but instead of the letters i - n, write in the number of mutations to convert one sequence to the next eg. the A to B distance would be 1. Do NOT waste your time writing out each sequence. Just use A - G to represent the nodes in the tree.

c) Using the tree you have drawn, recalculate the D to F distance. Why is it different from the distance calculated by pairwise comparisons above? What effect would this have on construction of a distance tree?