PLNT4610/7690 BIOINFORMATICS

FINAL EXAMINATION

Saturday January 22, 2022 16:00 to 18:00 ONLINE

Answer any combination of questions totalling to exactly 100 points. There are 11 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. (5 points) The first step in preprocessing of sequencing reads is to trim the reads. How would the genome assembly be affected if this step were not done?

2. (20 points)

You have been given the task of designing a genome annotation pipeline. The first phase of genome annotation is to do a blast search of each contig against a database to identify protein coding regions.

a) (10 points) To search for protein coding sequences in contigs, which BLAST program (blastn, blastx, tblastn, tblastx) would you use, and which database would you search? Explain the reasons for choice of program and choice of database.

b) (10 points) Assume that you will need to compare hundreds of contigs against the database. Instead of doing them 1 at a time on a single computer, you decide to use the CC Linux cluster. There are 15 compute nodes, each of which has 256 Gb of RAM and 64 CPUs. All of these mount the BLAST databases from the same filesystem on the file server.

Obviously, with 15 servers, you could run 15 searches simultaneously. However, you'd like to find out if you can run more than 1 search per server at the same time. Outline an experimental strategy to determine how many searches you can run at 1 time on each server.

3. (10 points) Is XML object-oriented? Why or why not? With reference to the example below, explain your answer.

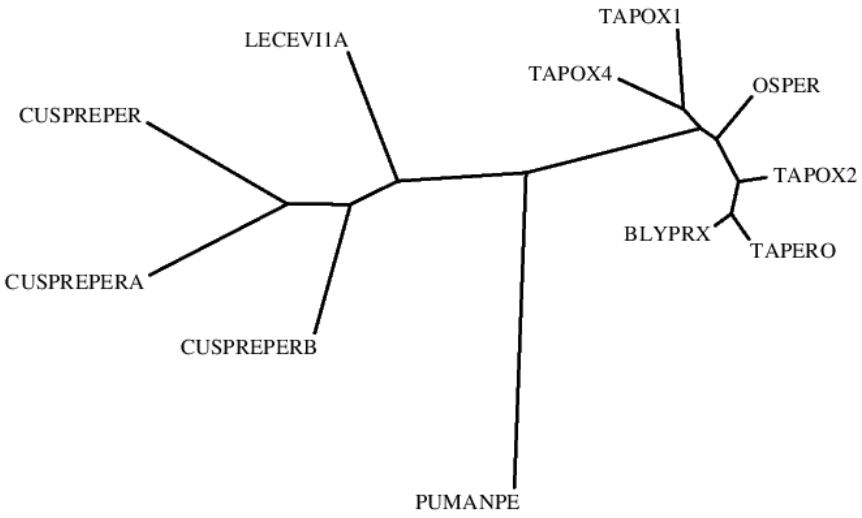
```
−<uniprot xsi:schemaLocation="http://uniprot.org/uniprot http://www.uniprot.org/support
 /docs/uniprot.xsd">
 −<entry dataset="Swiss-Prot" created="1990-01-01" modified="2008-11-04" version="43">
    <accession>P13240</accession>
    <name>DR206_PEA</name>
  +<protein></protein>
  +<gene></gene>
  −<organism key="1">
     <name type="scientific">Pisum sativum</name>
     <name type="common">Garden pea</name>
     <dbReference type="NCBI Taxonomy" id="3888" key="2"/>
   +<lineage></lineage>
   </organism>
  −<reference key="3">
   −<citation type="journal article" date="1995" name="Plant Physiol." volume="107" first="301"
    last="302">
    −<title>
       Molecular characterization of disease-resistance response gene DRR206-d from Pisum sativum (L.).
     </title>
    −<authorList>
       <person name="Culley D.E."/>
       <person name="Horovitz D."/>
       <person name="Hadwiger L.A."/>
     </authorList>
     <dbReference type="MEDLINE" id="95175620" key="4"/>
     <dbReference type="PubMed" id="7870833" key="5"/>
     <dbReference type="DOI" id="10.1104/pp.107.1.301" key="6"/>
   </citation>
```

4. (10 points) Statistics for a genome assembly and a transcriptome assembly are compared for a fungus. Fill in the table with the most reasonable column heading from the list below.

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

i) number of contigs
ii) smallest contig
iii) largest contig
iv) N50
v) total size (bp)

5. (20 points) A maximum likelihood tree was constructed for 11 plant peroxidase proteins.



BLYPRX	Hordeum vulgare	barley	monocot
CUSPREPER	Cucumis sativus	cucumber	dicot
CUSPREPERA	Cucumis sativus	cucumber	dicot
CUSPREPERB	Cucumis sativus	cucumber	dicot
LECEVI1A	Lycopersicon esculentum	tomato	dicot
OSPER	Oryza sativa	rice	monocot
PUMANPE	Petroselinum crispum	parsley	dicot
TAPERO	Triticum aestivum	wheat	monocot
TAPOX1	Triticum aestivum	wheat	monocot
TAPOX2	Triticum aestivum	wheat	monocot
TAPOX4	Triticum aestivum	wheat	monocot

With reference to the data shown, choose all that apply:

a) Peroxidases have diverged more in monocots than in dicots.
b) Peroxidases are more highly conserved in dicots than in monocots.
c) TAPERO is orthologous to TAPOX4, TAPOX1 and TAPOX2.
d) TAPOX4 and TAPOX1 are ortholgous.
e) TAPOX4 and TAPER0 are ortholgous.
f) TAPOX4 and TAPER0 are paralogous.
g) CUSPREPER, CUSPREPERA,CUSPREPERB represent a multigene family that arose from a single gene since cucumber diverged from parsley and tomato.
i) Of all the peroxidases shown, OSPER is most similar to the common ancestral copy of those genes in monocots.
j) The wheat genome contains at least 4 peroxidase genes.
k) The barley genome contains only a single copy of the peroxidase gene.

6. (5 points)

Which of the following is true, regarding flat file databases (check all that apply):

a) can be viewed in a text editor
b) time required to find a record is proportional to the size of the database
c) adding or removing a record requires reading and writing the full database
d) minimize redundancy
e) records in a database can represent many different types (eg. classes) of data

7. (10 points) Match each phylogenetic analysis concept with a phylogeny method.

concept
a) requires a multiple sequence alignment as input
b) reconstructs ancestral sequences
c) considers alternative tree topologies
d) samples at random the solution space of all possible tree topologies
e) samples the solution space of possible tree topologies in a thorough and heuristic (trial and error) fashion
f) considers all possible tree topologies
g) the most practical method for very large numbers of sequences
h) branch lengths are underestimated because of homoplasies
i) converges on a tree that is close to the best tree
j) only considers one tree

Methods:

1 none
2 Neighbor joining
3 all except Neighbor Joining
4 all distance methods
5 all character methods
6 all phylogeny methods
7 Maximum Likelihood
8 Bayesean phylogeny

8. (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 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.

9. (15 points) The following database objects are examples of bad database practices. Explain what the problem is in each object, and what would be a better way to implement it.

a) The class is fine. What's wrong with the object, and how would you fix the problem?

CLASS		OBJECT	
Author		Schmidlup CT et al.	
Publication	?**Publication**	Publication	**Population diversity in Ulm...**
			Evidence for balancing sele...

b) A class was designed that would record the conditions for testing plants for disease resistance to microbial pathogens. In the example, two canola lines (Westar and Glacier) were inoculated with two different strains of the blackleg fungus (PG1, PG2), for a total of four experiments. The same concentration of inoculum was used in all experiments. Westar is resistant to PG1 but susceptible to PG2, while Glacier is resistant to both.

CLASS		OBJECT	
Experiment		GN285	
Pathogen	?**Strain**	Pathogen	**PG1**
Host	?**Plant_line**		**PG2**
Inoculum [sp/ml]	Float	Host	**Westar**
Disease score UNIQUE	Resistant		**Glacier**
	Susceptible	Inoculum [sp/ml]	10e6

What is wrong with Experiment object GN285? Using the same Experiment class with no changes, how would you create objects that more accurately describe the four experiments?

10. (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
```

11. (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