PLNT4610/7690 BIOINFORMATICS

FINAL EXAMINATION

Tuesday December 15, 2020 13:30 to 15:30 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. (10 points) Genome assembly is usually done *de-novo*, meaning that each assembly starts with sequencing reads, and no other input. Some assembly programs also allow you to read in a genome from a close relative, to be used as a guide in the assembly process. For example, genomic sequences are available for *Brassica rapa*, *B. nigra*, and *B. oleracea*. If we wanted to assemble the genome for *Brassica repanda*, one might be tempted to use one of the other *Brassica* genomes as a reference genome to guide the assembly. What are the reasons that this approach would be a bad idea, compared to *de-novo* assembly?
Hint: Consider the many mechanisms by which genomes evolve.

2. (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?

3. (5 points) The goal of genome annotation is to identify every gene in the genome. Why is this harder than one might initially think? In other words, how does choice of mRNA population(s) used for RNA sequencing affect the completeness of the transcriptome, and consequently, the annotated genome?

4. (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) 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.

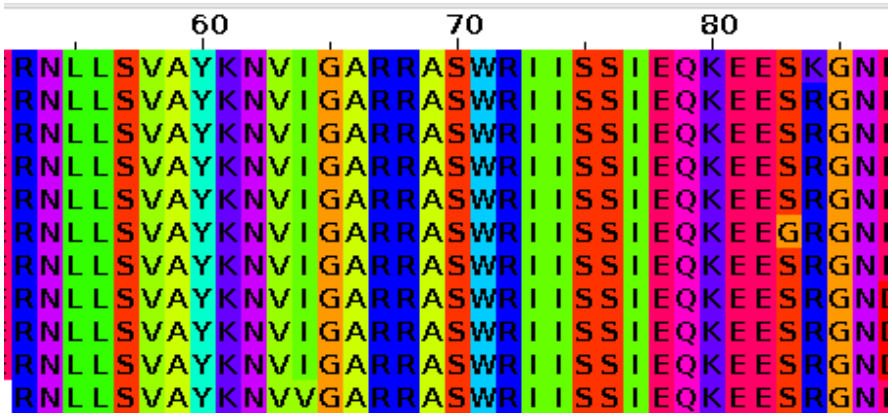
Outline an experimental strategy to determine how many searches you can run at 1 time on each server.

5. (10 points) Is XML object-oriented? Why or why not? With reference to the example below, explain your answer.

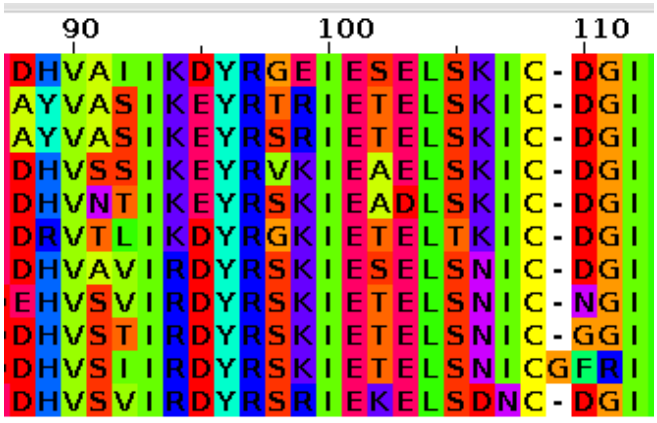
```
−<uniprot xsi:schemaLocation="http://uniprot.org/uniprot http://www.uniprot.org/support
 /docs/uniprot.xsd">
 −<entry dataset="Swiss-Prot" created="1990-01-01" modified="2008-11-04" version="43">
    <accession>P13240</accession>
    <name>DR206_PEA</name>
  +<protein></protein>
  +<gene></gene>
  −<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>
```

6. (10 points) Three regions from a multiple sequence alignment of plant 14-3-3 proteins are shown below. For A, B and C, briefly state what each region will contribute to construction of a phylogenetic tree. What are the differences between A,B and C, and how do those differences lead to a better tree, introduce ambiguity into the tree, or have little or no effect on the construction of the tree.

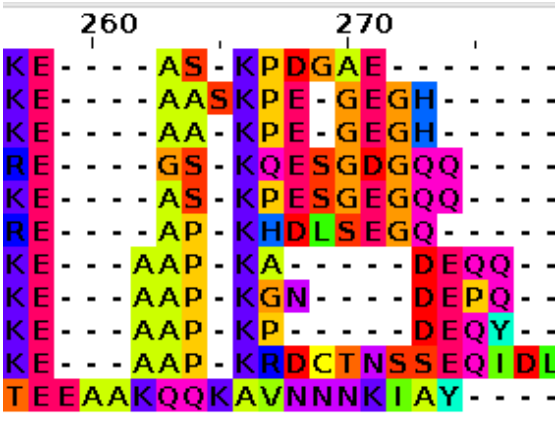
A



B



C

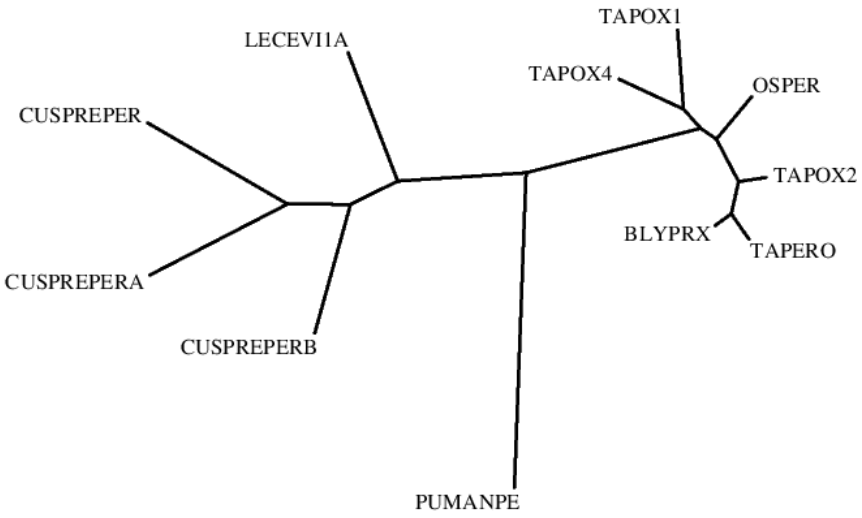


7. (10 points) Below is an excerpt from a GFF3 file, describing features in a genome. Briefly explain the distinction between items labeled mRNA0001, mRNA0002 and mRNA0003.

```
 0  ##gff-version   3
 1  ##sequence-region   ctg123 1 1497228
 2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
 3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

 4  ctg123 . mRNA            1050  9000  .  +  .
ID=mRNA00001;Parent=gene00001;Name=EDEN.1
 5  ctg123 . five_prime_UTR  1050  1200  .  +  .  Parent=mRNA0001
 6  ctg123 . CDS             1201  1500  .  +  0  Parent=mRNA0001
 7  ctg123 . CDS             3000  3902  .  +  0  Parent=mRNA0001
 8  ctg123 . CDS             5000  5500  .  +  0  Parent=mRNA0001
 9  ctg123 . CDS             7000  7600  .  +  0  Parent=mRNA0001
10  ctg123 . three_prime_UTR 7601  9000  .  +  .  Parent=mRNA0001

11  ctg123 . mRNA            1050  9000  .  +  .
ID=mRNA00002;Parent=gene00001;Name=EDEN.2
12  ctg123 . five_prime_UTR  1050  1200  .  +  .  Parent=mRNA0002
13  ctg123 . CDS             1201  1500  .  +  0  Parent=mRNA0002
14  ctg123 . CDS             5000  5500  .  +  0  Parent=mRNA0002
15  ctg123 . CDS             7000  7600  .  +  0  Parent=mRNA0002
16  ctg123 . three_prime_UTR 7601  9000  .  +  .  Parent=mRNA0002

17  ctg123 . mRNA            1300  9000  .  +  .
ID=mRNA00003;Parent=gene00001;Name=EDEN.3
18  ctg123 . five_prime_UTR  1300  1500  .  +  .  Parent=mRNA0003
19  ctg123 . five_prime_UTR  3000  3300  .  +  .  Parent=mRNA0003
20  ctg123 . CDS             3301  3902  .  +  0  Parent=mRNA0003
21  ctg123 . CDS             5000  5500  .  +  2  Parent=mRNA0003
22  ctg123 . CDS             7000  7600  .  +  2  Parent=mRNA0003
23  ctg123 . three_prime_UTR 7601  9000  .  +  .  Parent=mRNA0003
```

Columns 3 and 4 list start and finish coordinates for features. Column 5 indicates the strand (+ or -) of the feature.

8. (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.

9. (5 points)

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

a) cannot reference records from external sources
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

10. (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:

- none
- Neighbor joining
- all except Neighbor Joining
- all distance methods
- all character methods
- all phylogeny methods
- Maximum Likelihood
- Bayesean phylogeny

11. (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?