PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Thursday, October 18, 2018

Answer any combination of questions totalling to exactly 100 points. If you answer questions totalling more than 100 points, answers will be discarded at random until the total points are less than or equal to 100. There are 11 questions to choose from, totaling 120 points. This exam is worth 20% of the course grade.

Hand in this question sheet along with your exam book. All questions must be answered in the exam book. The exam sheets will be shreded 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. (20 points) Define the following:

a) similar
b) homologous
c) analogous
d) orthologous
e) paralogous

Draw an ontology diagram, showing the relationships between these concepts.

2. (10 points) The table below gives the algorithms for two methods of the MAFFT program for multiple sequence alignment.

A	pre-calculate all pairwise alignments between each pair of sequences to create a library of aligned sequence pairs for each aligned sequence pair in the library calculate all possible alignemts with sequence c choose the highest-scoring three-way alignment

B	pre-calculate all pairwise alignments between each pair of sequences to create a library of aligned sequence pairs
	repeat
	  for each aligned sequence pair in the library
	  calculate all possible alignemts with sequence c
	  choose the highest-scoring three-way alignment
	
	until (improvement in the alignment score negligible)

a) Which algorithm is faster, A or B? Explain your reasoning.

b) Which algorithm is more likely to give the best alignment, A or B? Explain your reasoning.

3. (10 points) Explain the rationale behind the fact that in multiple sequence alignments, a match between two gaps is scored as 0. When considering this question, it may help to review the Blosum45 matrix, shown below.

Blosum 45 Amino Acid Similarity Matrix

	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys
Gly	7																			
Pro	-2	9																		
Asp	-1	-1	7																	
Glu	-2	0	2	6																
Asn	0	-2	2	0	6															
His	-2	-2	0	0	1	10														
Gln	-2	-1	0	2	0	1	6													
Lys	-2	-1	0	1	0	-1	1	5												
Arg	-2	-2	-1	0	0	0	1	3	7											
Ser	0	-1	0	0	1	-1	0	-1	-1	4										
Thr	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
Ala	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
Met	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
Val	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
Ile	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
Leu	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
Phe	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Tyr	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
Trp	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
Cys	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

4. (10 points) A pairwise alignment between two superoxide dismutases, NPSODM and PSSODI, is shown below. Calculate the similarity score, using the BLOSUM45 scoring matrix provided in the previous question. Show your work.

```
NPSODM GEDGTASFTL
          .:.:. :.
PSSODI NAEGVAEATI
```

5. (10 points) Describe how thin clients such as Thinlinc divide tasks between the user's desktop computer and the remote server. How does this division of tasks make it possible for "Any user can do any task from anywhere".

6. (20 points) Fill in the blanks.

Over the course of sequence evolution, some positions undergo base or amino acid substitutions, and bases or amino acids can be inserted or deleted. Any measurement of similarity must therefore be done with respect to the best possible alignment between two sequences. Because insertion/deletion events are __a__ compared to base substitutions, it makes sense to penalize gaps __b__ than mismatches when calculating a similarity score. As an example, a very simple scoring scheme would add +1 for each match, -1 for each mismatch, and -2 for each gap inserted. That is, the larger the gap, the more we subtract. The similarity between two sequences would then be

Similarity = _____c_____

From an evolutionary point of view, the gap penalty scheme in the simple Needleman-Wunsch algorithm is highly unrealistic. Although single point insertions or deletions ("indels") are probably more common than large indels, it is not obvious that any sort of linear relation exists between frequency of indels and length. That is, it's just as easy to delete 4 bases as to delete 2. Most alignment programs deal with this problem using __d__ gap penalties. __d__ gap penalties consist of a __e__ penalty, and an __f__ penalty for each subsequent gap character inserted into the alignment. Typically, the __e__ penalty is __g__ negative than the __f__ penalty. Unfortunately, there is no empirical data to guide the choice of values for these penalties.

__d__ gap penalties are calculated by the formula:

penalty = _____h_____

As mentioned above, there are no good theoretical criteria for choosing __e__ and __f__ penalties for proteins. That being said, we know that insertion/deletion events are less frequent than amino acid substitutions. Therefore, it makes sense that a __e__ penalty should be more negative than the

_____i_____. For example, with the Blosum45 matrix, the _____i_____ is a Trp (W) - Cys (C) substitution, which gives a score of -5. Therefore, it would be reasonable that for even a single gap that the _e__ penalty should be _____j_____ negative than the _____i_____.

(*It may be useful to consider the Blosum45 scoring matrix in question 3, when answering this question.*)

7. (10 points) You wish to design an oligonucleotide probe that would identify genes encoding the Superoxidase dismutase protein. Given the following amino acid sequence from the SOD protein

G F H I H A

use the genetic code table and the ambiguity code table (both found on the last page of this question sheet) to design a degenerate oligonucleotide that should recognize SOD genes containing this protein motif, and would recognize all possible DNA sequences for this hexameric sequence. How many distinct DNA sequences would this degenerate oligonucleotide represent if you synthesized 17-mer oligos? Show your work.

8. (5 points) Why is it important to eliminate duplicate sequences before doing a multiple protein alignment?

9. (10 points) The script below, testprot_answer.sh, was part of the tutorial on Basic Shell Scripting.

```
#!/bin/bash

# Test whether a fasta file is nucleic acid or protein

# Read arguments from the command line, and set variables to
# represent the arguments
infile=$1
outfile=$2

# process the input file

result=`cat $infile | grep -v '^>' | grep -i -e [FPEJLZOIQ*X] |wc -l`
echo $result

if (($result > 0))
then
    msg="$infile contains protein."
else
    msg="$infile contains DNA."
fi

# output the result
echo $msg > $outfile
```

Modify this script so that instead of testing whether a fasta file contains nucleic acid or proteins, the new script instead prints out the number of sequences in a fasta file. For example, if there was a fasta file called pro.fsa, it might contain the following sequences

```
>CUSPREPERB:CDS1 323 bp
MAASSKVIVSLVLCLMMAVSVRSQLSSTFYDTTCPNVSSIVHGVMQQALQSDDRAGAKII
RLHFHDCFVDGCDGSVLLEDQDGITSELGAPGNGGITGFNIVNDIKTAVENVCPGVVSCA
DILALGSRDAVTLASGQGWTVQLGRRDSRTANLQGARDRLPSPFESLSNIQGIFRDVGLN
DNTDLVALSGAHTFGRSRCMFFSGRLNNNPNADDSPIDSTYASQLNQTCQSGSGTFVDLD
PTTPNTFDRNYYTNLQNNQGLLRSDQVLFSTPGASTIATVNSLASSESAFADAFAQSMIR
MGNLDPKTGTTGEIRTNCRRLN*
>PUMANPE:CDS1 364 bp
MVSCLGDKDGNANGLGFLFLLALSLLFISSQLYVSATYSTVPAVKGLEYNFYHSSCPKLE
TVVRKHLKKVFKEDVGQAAGLLRLHFHDCFVQGCDASVLLDGSASGPSEQDAPPNLSLRS
KAFEIIDDLRKLVHDKCGRVVSCADLTALAARDSVHLSGGPDYEVPLGRRDGLNFATTEA
TLQNLPAPSSNADSLLTALATKNLDATDVVALSGGHTIGLSHCSSFSDRLYPSEDPTMDA
EFAQDLKNICPPNSNNTTPQDVITPNLFDNSYYVDLINRQGLFTSDQDLFTDTRTKEIVQ
DFASDQELFFEKFVLAMTKMGQLSVLAGSEGEIRADCSLRNADNPSFPASVVVDSDVESK
SEL*
>TAPOX4:CDS1 320 bp
MAMAMASSLSVLLLLCLAAPSSAQLSPRFYARSCPRAQAIIRRGVAAAVRSERRMGASLL
RLHFHDCFVQGCDASILLSDTATFTGEQGAGPNAGSIRGMNVIDNIKAQVEAVCTQTVSC
ADILAVAARDSVVALGGPSWTVPLGRRDSTTASLSLANSDLPPPSFDVANLTANFAAKGL
SVTDMVALSGAHTIGQAQCQNFRDRLYNETNIDTAFATSLRANCPRPTGSGDSSLAPLDT
TTPNAFDNAYYRNLMSQKGLLHSDQVLINDGRTAGLVRTYSSASAQFNRDFRAAMVSMGN
ISPLTGTQGQVRLSCSRVN*
```

If the new script was called countseq.sh, the command

```
countseq.sh pro.fsa
```

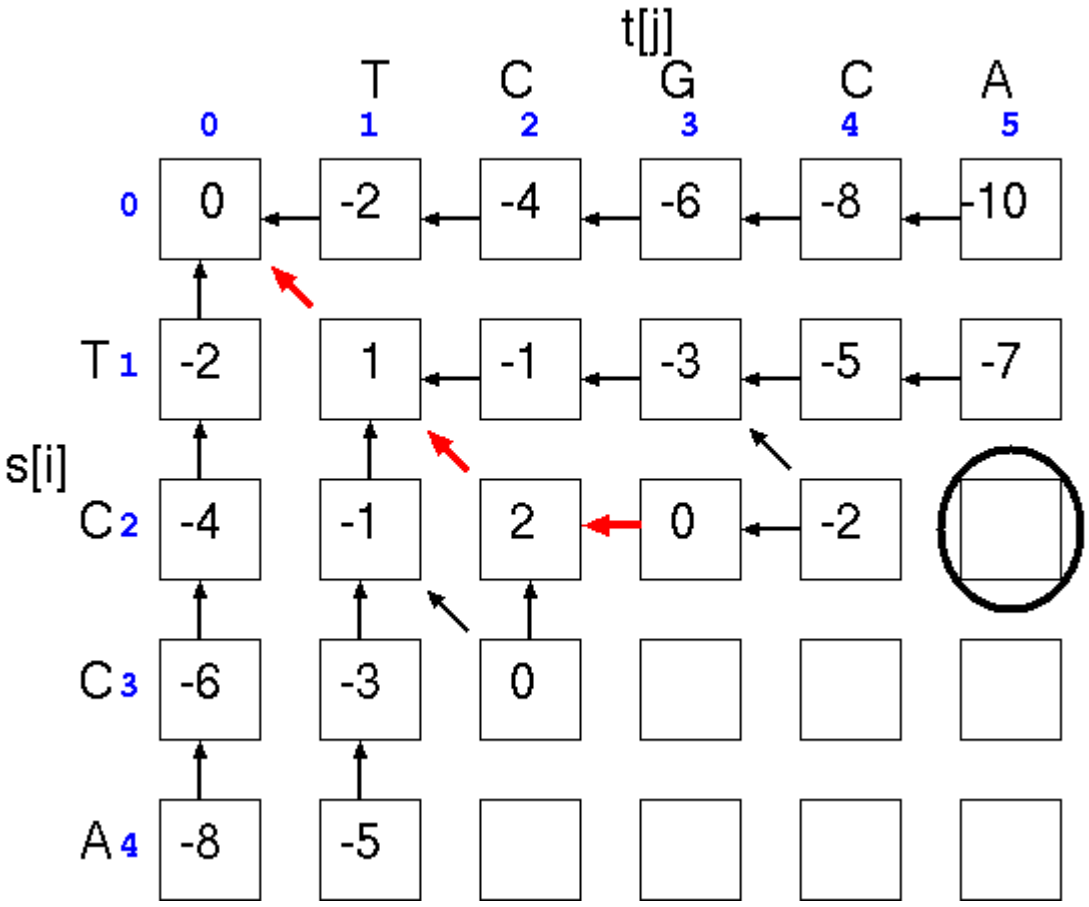
would result in the output

```
3
```

That is, if the number of sequences would be written to the terminal, and not to a file.

Hint: The resultant script will be <u>much simpler</u> than the original script.

10. (10 points)



For the cell circled in the dynamic programming alignment, evaluate a[i,j].

$$a[i,j] = \max \begin{cases} a[i,j-1] - 2 \\ a[i-1,j-1] + p(i,j) \\ a[i-1,j] - 2 \end{cases}$$

11. 5 (points) The histogram below shows the distribution of similarity scores from a tfasta database search using a plant lipid transfer protein as the query. There is a distinct peak for sequences with z-scores greater than 120. What does this peak represent?

```
 one = represents 428 library sequences
 for inset = represents 8 library sequences

z-opt        E()
< 20   273     0 :=
  22     1     0 :=
  24     4     0 :=
  26     7     5 :*
  28    16    52 :*
  30    75   319 :*
  32   384  1232 := *
  34  1515  3341 :====   *
  36  4706  6861 :===========     *
  38 10955 11338 :=========================*
  40 16069 15816 :====================================*=
  42 23280 19333 :================================================*=========
  44 25655 21326 :================================================*==========
  46 25330 21721 :================================================*=========
  48 25551 20795 :================================================*==========
  50 21174 18976 :============================================*=====
  52 17154 16683 :====================================*==
  54 13351 14250 :================================ *
  56  9609 11903 :=======================    *
  58  6904  9772 :=================     *
  60  4873  7916 :============      *
  62  3531  6346 :=========     *
  64  3072  5047 :========   *
  66  2424  3989 :======   *
  68  2088  3138 :=====  *
  70  1736  2459 :=====*
  72  1633  1921 :====*
  74  1404  1498 :===*
  76  1224  1166 :==*
  78  1153   906 :==*
  80  1079   704 :=*=
  82   923   538 :=*=
  84   835   426 :*=
  86   698   330 :*=
  88   520   255 :*=
  90   427   198 :*
  92   376   153 :*          :==================*====================
  94   281   118 :*          :==============*======================
  96   257    92 :*          :===========*=====================
  98   189    71 :*          :========*===============
 100   165    55 :*          :======*==============
 102   143    42 :*          :=====*============
 104   118    33 :*          :====*==========
 106    91    25 :*          :===*========
 108    59    20 :*          :==*=====
 110    60    15 :*          :=*======
 112    36    12 :*          :=*===
 114    28     9 :*          :=*==
 116    32     7 :*          :*===
 118    29     5 :*          :*===
>120   206     4 :*          :*=========================
72490335 residues in 38566 sequences
 statistics extrapolated from 20000 to 231238 sequences
 results sorted and z-values calculated from opt score
 15333 scores better than 54 saved, ktup: 2, variable pamfact
```

The IUPAC-IUB symbols for nucleotide nomenclature [Cornish-Bowden (1985)Nucl. Acids Res. 13: 3021-3030.] are shown below:

Symbol	Meaning	Symbol	Meaning
G	Guanine	K	G or T
A	Adenine	S	G or C
C	Cytosine	W	A or T
T	Thymine	H	A or C or T
U	Uracil	B	G or T or C
R	Purine (A or G)	V	G or C or A
Y	Pyrimidine (C or T)	D	G or T or A
M	A or C	N	G or A or T or C

The Universal Genetic Code							
UUU UUC UUA UUG	phe leu	UCU UCC UCA UCG	ser	UAU UAC UAA UAG	tyr stop stop	UGU UGC UGA UGG	cys stop trp
CUU CUC CUA CUG	leu	CCU CCC CCA CCG	pro	CAU CAC CAA CAG	his gln	CGU CGC CGA CGG	arg
AUU AUC AUA AUG	ile met	ACU ACC ACA ACG	thr	AAU AAC AAA AAG	asn lys	AGU AGC AGA AGG	ser arg
GUU GUC GUA GUG	val	GCU GCC GCA GCG	ala	GAU GAC GAA GAG	asp glu	GGU GGC GGA GGG	gly

3-letter	1-letter	3-letter	1-letter	3-letter	1-letter
Phe	F	Leu	L	Ile	I
Met	M	Val	V	Ser	S
Pro	P	Thr	T	Ala	A
Tyr	Y	His	H	Gln	Q
Asn	N	Lys	K	Asp	D
Glu	E	Cys	C	Trp	W
Arg	R	Gly	G	STOP	*
Asx	B	Glx	Z	UNKNOWN	X
Xle (Leu/Ile)	J	Pyl (pyrrolysine)	O		