PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Tuesday, October 24, 2017

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 12 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. (5 points) The BLAST database services at NCBI must process over 100,000 BLAST searches per day. Researchers at NCBI realized that the most critical bottleneck in the process was the simple matter of reading in all the sequence data when comparing a query sequence with all sequences in a database. What solution was found to solve this problem?

2. (15 points) If you wanted to design an oligonucleotide as a hybridization probe, you want to ensure that the oligo sequence is unique within the genome ie. it is not likely to occur by random chance. To help in your calculations, a table is given with some relevant information.

n	4^n	2 x 4^n
10	1.05E+06	2.10E+06
11	4.19E+06	8.39E+06
12	1.68E+07	3.36E+07
13	6.71E+07	1.34E+08
14	2.68E+08	5.37E+08
15	1.07E+09	2.15E+09
16	4.29E+09	8.59E+09
17	1.72E+10	3.44E+10
18	6.87E+10	1.37E+11

a) How big would an oligo probe have to be for use with haploid yeast, *Saccharomyces cerevisiae*, (1N = 1.2 x 10^7 bp)? That is, how long does the oligo have to be to ensure that it is not likely to occur in the genome due to random chance?

b) Yeast also go through a diploid phase. If you were hybridizing to DNA extracted from diploid yeast, would you need to use a longer oligo? Explain.

c) Most eukaryotic genomes, especially for higher organisms, are largely composed of middle repetitive sequences such as the AluI family in mammals. How would this affect our estimates of the likelihood of finding a particular oligonucleotide in a eukaryotic genome?

3. (10 points) The dot-matrix plot below shows a comparison of two SAR8 proteins. What are the most obvious differences or mutations between these two sequences?

```
P1HOM         Version  5/13/91
X-axis: XP_009786495
Y-axis: XP_009588711
SIMILARITY RANGE:  10      MIN.PERCENT SIMILARITY:  30
SCALE FACTOR:    0.90      COMPRESSION:              2

               20        40        60        80

       V        .         .         .         .         .         .         .
        Q       .         .         .         .         .         .         .
         L      .         .         .         .         .         .         .
          Ij    .         .         .         .         .         .         .
           Hj   .         .         .         .         .         .         .
            D   .         .         .         .         .         .         .
             D  .         .         .         .         .         .         .
              C .         .         .         .         .         .         .
               C.         .         .         .         .         .         .
    20 .........C............................................................
                .A        .         .         .         .         .         .
                . C       .         .         .         .         .         .
                .  D      .         .         .         .         .         .
                .   E     .         .         .         .         .         .
                .    G    .         .         .         .         .         .
                .     I   .         .         .         .         .         .
                .      L  .         .         .         .         .         .
                .       L .         .         .         .         .         .
                .        L.         .         .         .         .         .
    40 ...................L..................................................
                .         .K        .         .         .         .         .
                .         . J       .         .         .         .         .
                .         .  H      .         .         .         .         .
                .         .   G     .         .         .         .         .
                .         .    G    .         .         .         .         .
                .         .     G   .         .         .         .         .
                .         .      I  .     j   .         .         .         .
                .         .       I .jj    j  .         .         .         .
                .         .        J. j       .         .         .         .
    60 .............................J........................................
                .         .         .L        .         .         .         .
                .         .         . L       .         .         .         .
                .         .         .  L      .         .         .         .
                .         .         .   M     .         .         .         .
                .         .     h   .  i Q    .         .         .         .
                .         .      f f.   d T   .         .         .         .
                .         .         g    V a  .         .         .         .
                .         .        f.i    Q d .         .         .         .
                .         .         i      L  .         .         .         .
    80 .....................................I................................
                .         .         .     i  F.         .         .         .
                .         .         .     i   C         .         .         .
                .         .         .         .C        .         .         .
                .         .         .         . A       .         .         .
                .         .         .         .  A      .         .         .
                .         .         .         .   D     .         .         .
                .         .         .         .    H    .         .         .
                .         .         .         .     L   .         .         .
                .         .         .         .      R  .         .         .
   100 ......................................................................
                .         .         .     b   .         .         .         .
                .         .         .      U  .         .         .         .
                .         .         .       O .         .         .         .
                .         .         .        J.         .         .         .
                .         .         .         F         .         .         .
                .         .         .         .C        .         .         .
                .         .         .         . C       .         .         .
                .         .         .         .  C      .         .         .
                .         .         .         .   F     .         .         .
   120 ............................................J.........................
                .         .         .         .     O   .         .         .
                .         .         .         .      X  .         .         .
                .         .         .         .         .         .         .
```

4. (5 points) It has been demonstrated that it is impractical to extend the dynamic programming algorithm of Needleman and Wunsch/Smith Waterman to constructing multiple alignments for more than a few sequences. If there are k sequences of length n, give a simple expression that tells how does the problem scales, in terms of n and k?

5. (10 points) The shell script below accepts a FASTA file as input. What is written to the output? An example of a FASTA file containing 3 sequences is shown in the box.

```
#!/bin/bash
infile=$1
outfile=$2
cat $infile | grep '>' | cut -c2- | cut -f1 -d " " > $outfile
```

```
>BLYTHNA 137 bp
MATNKSIKSVVICVLILGLVLEQVQVEAKSCCKNTTGRNCYNACRFAGGSRPVCATACGC
KIISGPTCPRDYPKLNLLPESGEPNATEYCTIGCRTSVCDNMDNVSRGQEMKFDMGLCSN
ACARFCNDGEVIQSVEA
>TATTH20MR 131 bp
MGGGQKGLESAIVCLLVLGLVLEQVQVEGVDCGANPFKVACFNSCLLGPSTVFQCADFCA
CRLPAGLASVRSSDEPNAIEYCSLGCRSSVCDNMINTADNSTEEMKLYVKRCGVACDSFC
KGDTLLASLDD
>TGTHI13 107 bp
MMVVVILGLVVAQTQVEAKSCCRNTTARNCYNVCRLPGTPRPVCAATCDCKIISSGKCPP
GYEKLGFSDVADEALDVAEEVMKEAVERCNNACSEVCTKGSYAVVTA
```

Notes:
-c2- tells the cut command to print all columns after and including column 2.
-d " " tells the cut command to use a blank space as a field delimiter

6. (15 points) A number of alternative genetic codes have been discovered. Examples are found in mitochondria, plastids, bacteria and archaea. In all of the alternative genetic codes seen so far, most of the codons code for the same amino acids as in the Standard Genetic Code, with a few codons differing. For example in some cases, a stop codon codes for an amino acid, or a codon for an amino acid is used as a stop codon. In other cases, one or two codons are reassigned to a different amino acid.

Type of search	NCBI	FASTA
a)DNA vs. DNA database	blastn	fasta3 ssearch3 (slow, full Smith-Waterman alignment)
b) protein vs. protein database	blastp	fasta3 ssearch3 (slow, full Smith-Waterman alignment)
c) protein vs. translated DNA database	tblastn	tfasta3
d) translated DNA vs. translated DNA database	tblastx	tfastx3, tfasty3
e) translated DNA vs. protein database	blastx	fastx3, fasty3 (especially well-suited for cDNAs, which often contain frameshift errors)

Yeast mitochondria use a non-standard genetic code. Suppose you had the seqeunces for a yeast mitochondrial gene, and its corresponding protein, and wished to find homolgues in other species.

How would the difference in genetic codes affect Needleman-Wunsch similarity scoreseach for the types of searches listed above? Indicate: Increase, Decrease, or No-effect. Also, state in a single sentence or phrase the reason.

7. (5 points) Sequence database search programs such as the FASTA and BLAST family of programs do not read database files that include annotation for each sequence, as would be found in GenBank or Uniprot entries. Rather, they read files in FASTA or similar formats, which include just a name and definition for each sequence, along with each sequence itself. What is the advantage, when doing a sequence database search, of eliminating the annotation?

8. (10 points) Two antifreeze proteins were aligned using both GGLSEARCH and GLSEARCH.

a) Which of the two alignments is deemed to be more statistically significant?

b) Why does the GGSEARCH alignment have a long gap, followed by a phenylalanine (F) at the end of ISP2_H? How does that gap contribute to the difference in Needleman-Wunsch (n-w) scores?

GGSEARCH

```
Algorithm: Global/Global affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL62 matrix (11:-4), open/ext: -11/-1

>>ISP2_OSMMO 175 bp                                       (175 aa)
 n-w opt: 315  Z-score: 295.7  bits: 61.1 E(1): 1.3e-133
global/global (N-W) score: 315; 39.1% identity (65.4% similar) in 179 aa overlap (1-163:1-175)

               10        20        30        40        50
ISP2_H MLTVSLLVCAMMALTQA-NDDKILKGTATEAGPVSQRAPPNCPAGWQPLGDRCIYYETTA
       ::. .::::::.:::.: : :   ... : .   : .   .::. :. .. ::. ..
ISP2_O MLA-ALLVCAMVALTRAANGDTGKEAVMTGS---SGKNLTECPTDWKMFNGRCFLFNPLQ
                10        20        30           40        50

      60        70        80        90          100       110
ISP2_H MTWALAETNCMKLGGHLASIHSQEEHSFIQTLN-AGVV--WIGGSACLQAGAWTWSDGTP
       . :: :. .::: :..:::::: ::..:.. :. ::..  ::::: :  .  : : :.:
ISP2_O LHWAHAQISCMKDGANLASIHSLEEYAFVKELTTAGLIPAWIGGSDCHVSTYWFWMDSTS
         60        70        80        90       100       110

        120       130       140       150       160
ISP2_H MNFRSWCSTKPDDVLAACCMQMTAAADQCWDDLPCPASHKSVCAMT------------F
       :.: .::...:: .:. ::.:..... .::.: ::   : ::::              .
ISP2_O MDFTDWCAAQPDFTLTECCIQINVGVGKCWNDTPCTHLHASVCAKPATVIPEVTPPSIM
        120       130       140       150       160       170
```

GLSEARCH

```
Algorithm: Global/Local affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL62 matrix (11:-4), open/ext: -11/-1

>>ISP2_OSMMO 175 bp                                       (175 aa)
 n-w opt: 336  Z-score: 328.6  bits: 67.2 E(1): 4e-171
global/local score: 336; 41.9% identity (69.5% similar) in 167 aa overlap (1-163:1-163)

               10        20        30        40        50
ISP2_H MLTVSLLVCAMMALTQA-NDDKILKGTATEAGPVSQRAPPNCPAGWQPLGDRCIYYETTA
       ::. .::::::.:::.: : :   ... : .   : .   .::. :. .. ::. ..
ISP2_O MLA-ALLVCAMVALTRAANGDTGKEAVMTGS---SGKNLTECPTDWKMFNGRCFLFNPLQ
                10        20        30           40        50

      60        70        80        90          100       110
ISP2_H MTWALAETNCMKLGGHLASIHSQEEHSFIQTLN-AGVV--WIGGSACLQAGAWTWSDGTP
       . :: :. .::: :..:::::: ::..:.. :. ::..  ::::: :  .  : : :.:
ISP2_O LHWAHAQISCMKDGANLASIHSLEEYAFVKELTTAGLIPAWIGGSDCHVSTYWFWMDSTS
         60        70        80        90       100       110

        120       130       140       150       160
ISP2_H MNFRSWCSTKPDDVLAACCMQMTAAADQCWDDLPCPASHKSVCAMTF
       :.: .::...:: .:. ::.:..... .::.: ::   : ::::
ISP2_O MDFTDWCAAQPDFTLTECCIQINVGVGKCWNDTPCTHLHASVCAKPATVIPEVTPPSIM
        120       130       140       150       160       170
```

9. (10 points) Draw a dot-matrix plot (eg. DXHOM) of this sequence, compared with itself. You can assume that the only repeats of any significance are the ones documented in this GenBank entry.

```
LOCUS       AY966016                 380 bp    DNA     linear   PLN 14-NOV-2005
DEFINITION  Aspergillus flavus isolate NPL GA3-3 hexose transporter-like (hexA)
            gene/telomere breakpoint junction.
ACCESSION   AY966016
VERSION     AY966016.1  GI:67944627
KEYWORDS    .
SOURCE      Aspergillus flavus
  ORGANISM  Aspergillus flavus
            Eukaryota; Fungi; Dikarya; Ascomycota; Pezizomycotina;
            Eurotiomycetes; Eurotiomycetidae; Eurotiales; Aspergillaceae;
            Aspergillus.
REFERENCE   1  (bases 1 to 380)
  AUTHORS   Chang,P.K., Horn,B.W. and Dorner,J.W.
  TITLE     Sequence breakpoints in the aflatoxin biosynthesis gene cluster and
            flanking regions in nonaflatoxigenic Aspergillus flavus isolates
  JOURNAL   Fungal Genet. Biol. 42 (11), 914-923 (2005)
   PUBMED   16154781
REFERENCE   2  (bases 1 to 380)
  AUTHORS   Chang,P.-K.
  TITLE     Direct Submission
  JOURNAL   Submitted (17-MAR-2005) Food and Feed Safety, Southern Regional
            Research Center, 1100 Robert E. Lee Boulevard, New Orleans, LA
            70124, USA
FEATURES             Location/Qualifiers
     source          1..380
                     /organism="Aspergillus flavus"
                     /mol_type="genomic DNA"
                     /isolate="NPL GA3-3"
                     /db_xref="taxon:5059"
                     /note="type: L"
     gene            62..244
                     /gene="hexA"
     misc_feature    62..244
                     /gene="hexA"
                     /note="similar to hexose transporter"
     misc_recomb     244^245
                     /gene="hexA"
                     /note="hexA-telomere breakpoint junction; recombination
                     results in deletion of aflatoxin gene cluster"
     repeat_region   245..376
                     /note="telomeric repeat"
                     /rpt_unit_seq="tcaacattaggg"
ORIGIN
        1 gtctttcccg ccaacttgaa gtccagcagt atccttaaca gtaccctttg ttactgacac
       61 catggttgct ggcggtggag ttgttccttc atccggtatg gatgcatacc gggccctgcc
      121 aaacaatacg aactcgaact ggttcaagga caagggcctc cggcgtctga atttcggcct
      181 catgcttatg tttgcatccg ctgcagcaaa tgggtatgat ggggctttga tgaatgggct
      241 cctgtcaaca ttagggtcaa cattagggtc aacattaggg tcaacattag ggtcaacatt
      301 agggtcaaca ttagggtcaa cattagggtc aacattaggg tcaacattag ggtcaacatt
      361 agggtcaaca ttagggtcaa
//
```

10. (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 11, when answering this question.*)

11. (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	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
	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys

12. (5 points) Why is it important to eliminate duplicate sequences before doing a multiple protein alignment?

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

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		