PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Tuesday, October 21, 2014

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. 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.

---

1. (5 points)

XhoII recognizes the following restriction site: 5'R^GATCY3'

Re-write the following as a double-stranded sequence, showing where cuts occur, and for each 5' and 3' end in the restriction site, label the coordinate of each end.

```
          10        20
      |    |    |    |
5'CGGGATCGATAGATCCGGAATTC3'
```

R = purine, Y = pyrimidine

2. (15 points) Given four sequences, show the steps that T-COFFEE would perform to create a multiple protein alignment.

```
A) CMEKEKYE
B) CVKKHKKI
C) CVKKHKKI
D) CIQEDKFE
```

a) Draw the guide tree, based on visual inspection of the four sequences for similarity.

b) Write out the pairwise alignments, based on the guide tree

c) Write out the complete alignment, based on the pairwise alignments and the guide tree.

To make your job easier, just score alignments by considering perfect amino acid matches, rather than taking into account a scoring matrix. Remember, the goal of an optimal alignment is to maximize the similarity scores while minimizing the number of gaps added.

3. (10 points) The PAM matrices were constructed using a set of protein alignments that had been done on proteins representing fairly distant evolutionary relationships. Many of these alignments required gaps to construct optimal alignments. The BLOSUM matrices were constructed using a dataset of protein domains that required no gaps for alignment, but were probably more closely-related than the proteins used for the PAM matrices. Discuss the tradeoffs between the two approaches. In other words, what is the perceived advantage of one versus the other, and what is the compromise made to take that advantage?

4. (10 points) Suppose that you wanted to do exhaustive pairwise similarity comparisons between very large sequences using the Smith-Waterman algorithm ie. global sequence alignment by dynamic programming. Consider the fact that the entire similarity matrix has to be stored in random access memory (RAM). If a typical PC has about 10 Gb (gigabytes) of RAM, what would be the maximum length of sequences that could be aligned?

Assumptions:

- Each bp in a sequence can be represented in a single charcter (A,G,C,T), which is a single byte.
- the memory taken up by software, the operating system etc. is negligible
- both sequences are the same length

5. (10 points) Suppose you wanted to create a dataset that would accurately sample sequences among different major taxonomic groups. Based on the data in the table below, what are some of the problems with creating such a dataset? Can you think of a strategy that would help you overcome these problems?

taxon	estimated number of species	percentage of species	number of sequences in NCBI UniGene	percentage of sequences
insects	830000	69.2	239944	12.7
molluscs	110000	9.2	40311	2.1
other animals	100000	8.3	216337	11.4
arachnids	60000	5.0	26582	1.4
crustaceans	50000	4.2	95901	5.1
vertebrates	50000	4.2	1275236	67.3
total	1200000		1894311	

estimates from Stoeckle et al. Barcoding Life Illustrated.

http://barcoding.si.edu/PDF/BLIllustrated26jan04v1-3.pdf

6. (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

N G K E H G

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 18-mer oligos?

7. (10 points) TFASTA and TBLASTN use protein query sequences to search against DNA databases. How do these programs translate the sequences in the DNA databases into proteins? Suppose that you were searching a DNA database consisting of 100 billion nucleotides. How many amino acids would that correspond to?

8. (10 points) For the following pairwise alignment, calculate the similarity score, using the BLOSUM45 scoring matrix provided.

```
SOSODCY:CDS1_1        D V R H A G D L G L
PEACUZNSD:CDS1_1      E T R H A G D L G N
```

Blosum 45 Amino Acid Similarity Matrix

	G	P	D	E	N	H	Q	K	R	S	T	A	M	V	I	L	F	Y	W	C
G	7																			
P	-2	9																		
D	-1	-1	7																	
E	-2	0	2	6																
N	0	-2	2	0	6															
H	-2	-2	0	0	1	10														
Q	-2	-1	0	2	0	1	6													
K	-2	-1	0	1	0	-1	1	5												
R	-2	-2	-1	0	0	0	1	3	7											
S	0	-1	0	0	1	-1	0	-1	-1	4										
T	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
A	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
M	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
V	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
I	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
L	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
F	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Y	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
W	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
C	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

9. (5 points) It might seem trivial to generate the opposite strand of a sequence, so simple, in fact that you might be able to do it by a simple search and replace:

original sequence	AATCGTTTGCCCCCCTA
Step 1: replace A with 1	11TCGTTTGCCCCCCT1
Step 2: replace G with 2	11TC2TTT2CCCCCCT1
Step 3: replace T with A	11AC2AAA2CCCCCCA1
Step 4: replace C with G	11AG2AAA2GGGGGGA1
Step 5: replace 1 with T	TTAG2AAA2GGGGGGAT
Step 6: replace 2 with C	TTAGCAAACGGGGGGAT

What is the problem with this approach?

10) (5 points) Below is an example of a FASTA file called ASTRASTL2A.fsa.

```
>ASTRASTL2A - Avana sativa thaumatin-like pathogenesis-related p
cccatagcaagctcggcacacagcaacactagcaaagcttgctagagcttgtagcgatggcgacctcctccgcgg
tgctgtttttcctcctcgccgtcttcgccgccggtgccagcgcggccaccttccgcatcaccaacaactgcggct
tcacggtgtggccggcgggcatcccggtgggcggaggcttccagctcaactcgaagcagtcgtccaacatcaacg
tgcccgcgggcaccagcgccggcaggatatggggccgcaccggctgctccttcaacaacgggagagggagctgcg
cgaccggagactgcgccggcgcgctgtcctgcaccctctccgggcagccggcgacgctggccgagtacaccatcg
gcggctcccaggacttctacgacatctcggtgatcgacggctacaacctcgccatggacttctcctgcagcaccg
gcgtcgcgctcaagtgcagggatgccaactgccccgacgcctatcaccaccccaacgacgtcgccacgcacgctt
gcaacggcaacagcaactaccagatcaccttctgcccatgaagaccctatgccgcgccgccaataaccggcgtac
atatacgaccgtataaatagtgtaaactgtgtaatgcttacatcgcggtatcatatatctgtattccagccgttg
tagtagttgacaaacggccaaataaagttcaataaagacggtgcacacatgtgtgcatgtcgacgttatctattt
aaaa
```

Explain whether or not it be appropriate to search for restriction sites using the grep command? For example, to search for EcoRI sites you might try the command

```
grep GAATTC ASTRASTL2A.fsa
```

11. (20 points) (Answer letters b through e. Letter a is an example.) An excerpt from a tblastn search at NCBI is shown below. Given the following statements about NCBI BLAST tell which aspect of the output illustrates one of these statements:

a. The alignment score is expressed as a deviation from randomness, according to information theory.
b. In scoring matrices such as PAM and BLOSUM, perfect identities between two sequences give the highest score, conservative replacements give intermediate scores, and uncommonly observed replacements give the lowest scores.
c. The BLAST programs filter out low complexity sequences in the query sequences
d. The length of a hit contributes to its score.

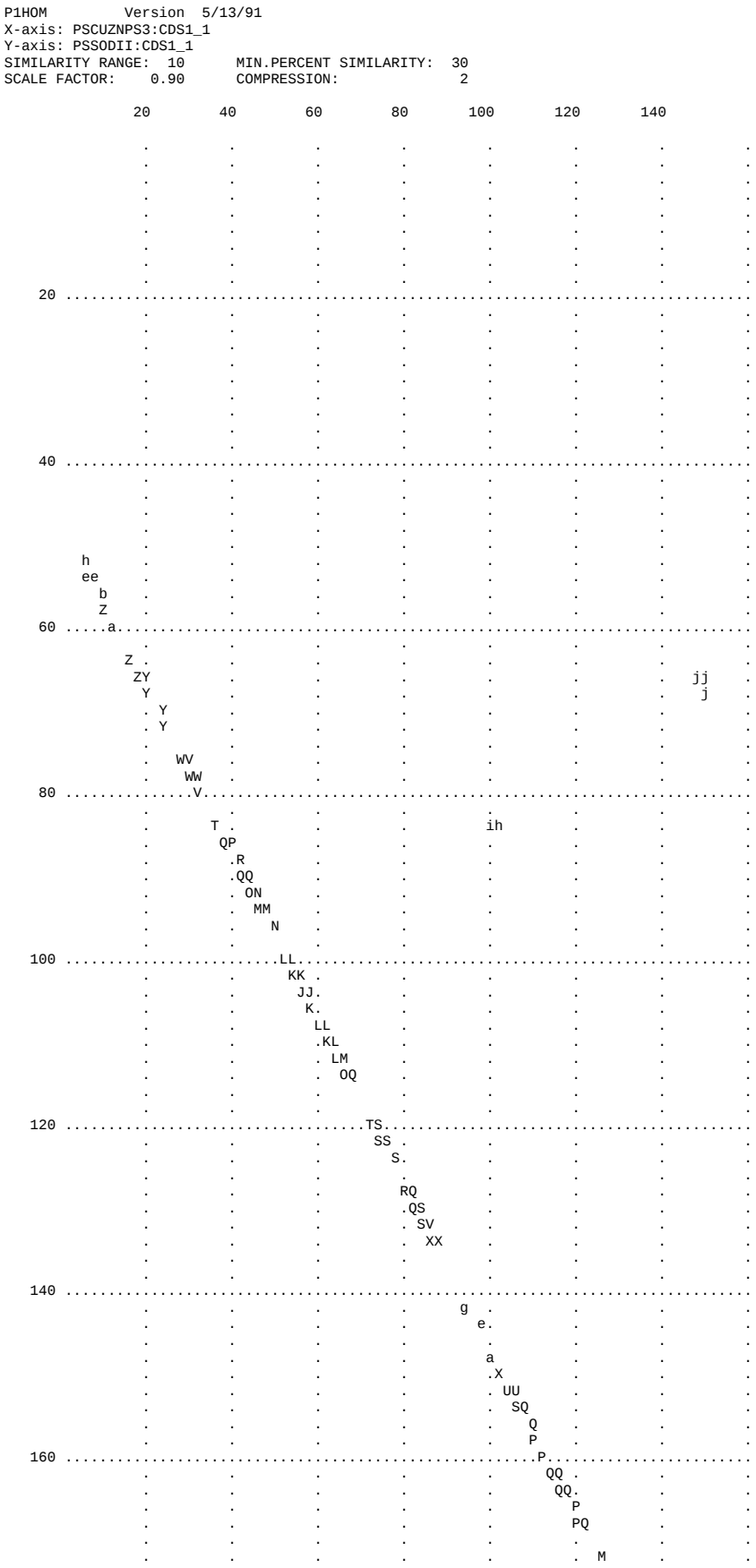
Example: For **a** above, your answer might be something like: The alignment score is shown in the output both as bits of information, and as the actual score in parenthese, calculated from the scoring matrix.

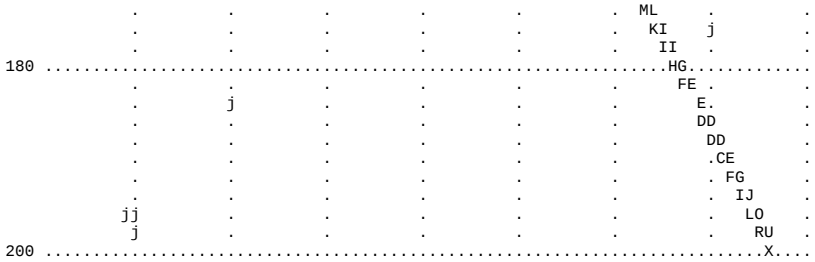
e. How much more statistically significant is the hit with AK120826, than the hit with XM002468536 eg. 2 times, 5 times 1000 times better? Give a number, and explain your reason.

```
>gi|37990449|dbj|AK120826.1| Oryza sativa Japonica Group cDNA clone:J023019E10, full insert
sequence
Length=540

 Score = 88.2 bits (217),  Expect = 3e-19, Method: Compositional matrix adjust.
 Identities = 57/83 (69%), Positives = 61/83 (73%), Gaps = 0/83 (0%)
 Frame = +2

Query  7    QSSMEAPRKlvsaalllvlllaaTGEMGGPVAVAEARKCESLSHRFAGLCLRGHNCANVC  66
            +  MEA RK+ SA LL+VLLLAATGEMGGPV VAEAR CES SHRF G C R  NCA+VC
Sbjct  35   KEEMEASRKVFSAMLLMVLLLAATGEMGGPVMVAEARTCESQSHRFKGPCARKANCASVC  214

Query  67   RTEGFPGGKCRGASRRCFCTTHC  89
             TEGFP G C G  RRC CT  C
Sbjct  215  NTEGFPDGYCHGVRRRCMCTKPC  283


>gi|242042372|ref|XM_002468536.1| Sorghum bicolor hypothetical protein, mRNA
Length=612

 Score = 87.0 bits (214),  Expect = 1e-18, Method: Compositional matrix adjust.
 Identities = 43/56 (77%), Positives = 45/56 (80%), Gaps = 0/56 (0%)
 Frame = +2

Query  35   GPVAVAEARKCESLSHRFAGLCLRGHNCANVCRTEGFPGGKCRGASRRCFCTTHCR  90
            G VAVAEAR C+S SHRF G C+R  NCANVCRTEGFP GKCRG  RRCFC THCR
```

Sbjct 92 GAVAVAEARTCQSQSHRFRGPCVRRENCANVCRTEGFPDGKCRGFRRRCFCLTHCR 259

12. (10 points) The dot-matrix plot below shows a comparison of two superoxide dismutase proteins. What are the most important observations you can make based on this data?




ML
KI
II
j
HG
FE
E.
DD
DD
.CE
. FG
. IJ
. LO
RU
180
200
X

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		