PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

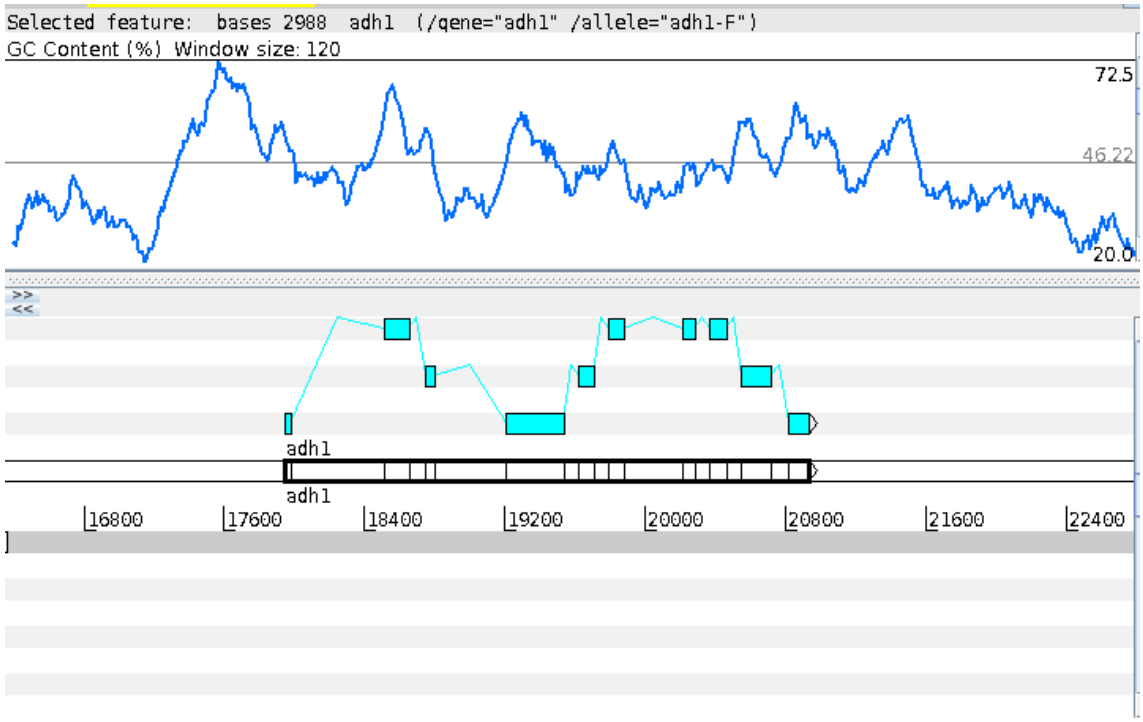
08:30 - 9:45 Tuesday, October 20, 2015

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. (10 points) A map of one of the loci encoding alcohol dehyrogenase 1 (adh1-f) in Maize is shown. What can you say about the structure of this gene?



2. (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

F T Q D G D

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? Show your work.

3. (10 points) The following features are annotated in a mouse sequence found in GenBank.

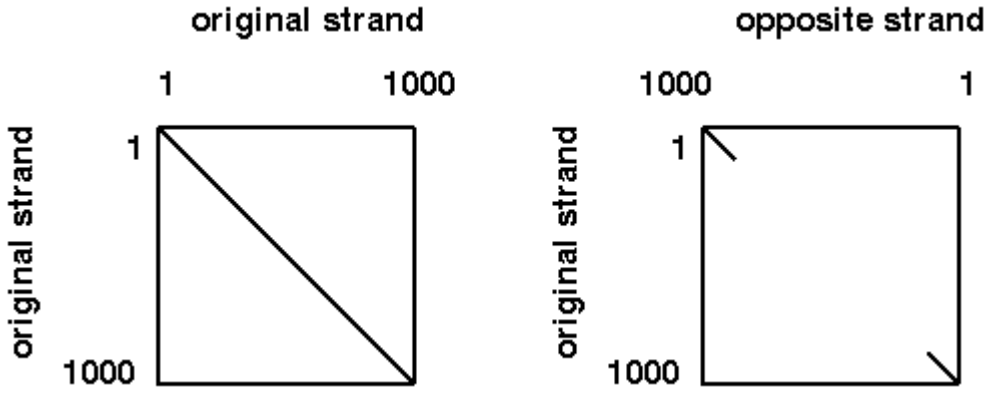
```
Key             Location/Qualifiers

source          1..1509
                /organism="Mus musculus"
                /strain="CD1"
                /mol_type="genomic DNA"
promoter        <1..9
                /gene="ubc42"
mRNA            join(10..567,789..1320)
                /gene="ubc42"
CDS             join(54..567,789..1254)
                /gene="ubc42"
                /product="ubiquitin conjugating enzyme"
                /function="cell division control"
```

a) The annotation for the promoter is expressed as "<1..9". Explain what is meant by this annotation.

b) The mRNA and CDS features imply the existence of intron and exon features. In the format of the Features Table, write the annotation for these features.

4. (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?

5. (10 points) A sequence was compared with its opposite strand, showing short diagonals at each end. Explain this observation.



6. (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?

7. (15 points) Fill in the blanks. In your exam booklet, just write a term for a - e. You don't need to rewrite the entire text. Note that for e, two terms should be given.

Protein database searches compared to DNA database searches

Speed

- A protein coding DNA sequence contains 3 times as many characters as the corresponding amino acid sequence
- Protein databases are much smaller than DNA databases because
  - _____a_____ are not present
  - where several DNA sequences encode identical proteins, only 1 protein database entry is usually created
- Speedup using lookup tables is more efficient with proteins. For proteins, k=2 yeilds a ___b___-fold speedup, whereas in DNA, k=4 yeilds only a 256-fold speedup.

Sensitivity

- The degeneracy of the genetic code makes it possible for DNA sequence to _____ c_____ more rapidly than amino acid sequence
- The greater complexity of _____ d _____ allows matches to be detected at very low levels of similarity
- The small alphabet size of DNA (4) compared to proteins (20) makes protein alignments more robust. That is, it is often far more obvious which _____ e-1 _____ should be aligned, compared to which _____e-2_____.
- DNA alignments tend to have more gaps, compared to protein alignments.

8. (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?

9. (10 points) For the following pairwise alignment, calculate the similarity score, using the BLOSUM45 scoring matrix provided. Show your work.

```
RICSOD    G  S  V  S  G  L  K  P  G  L
                .  .  :  :     :  :  :
PETSOD    V  R  I  T  G  L  A  P  G  L
```

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

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:

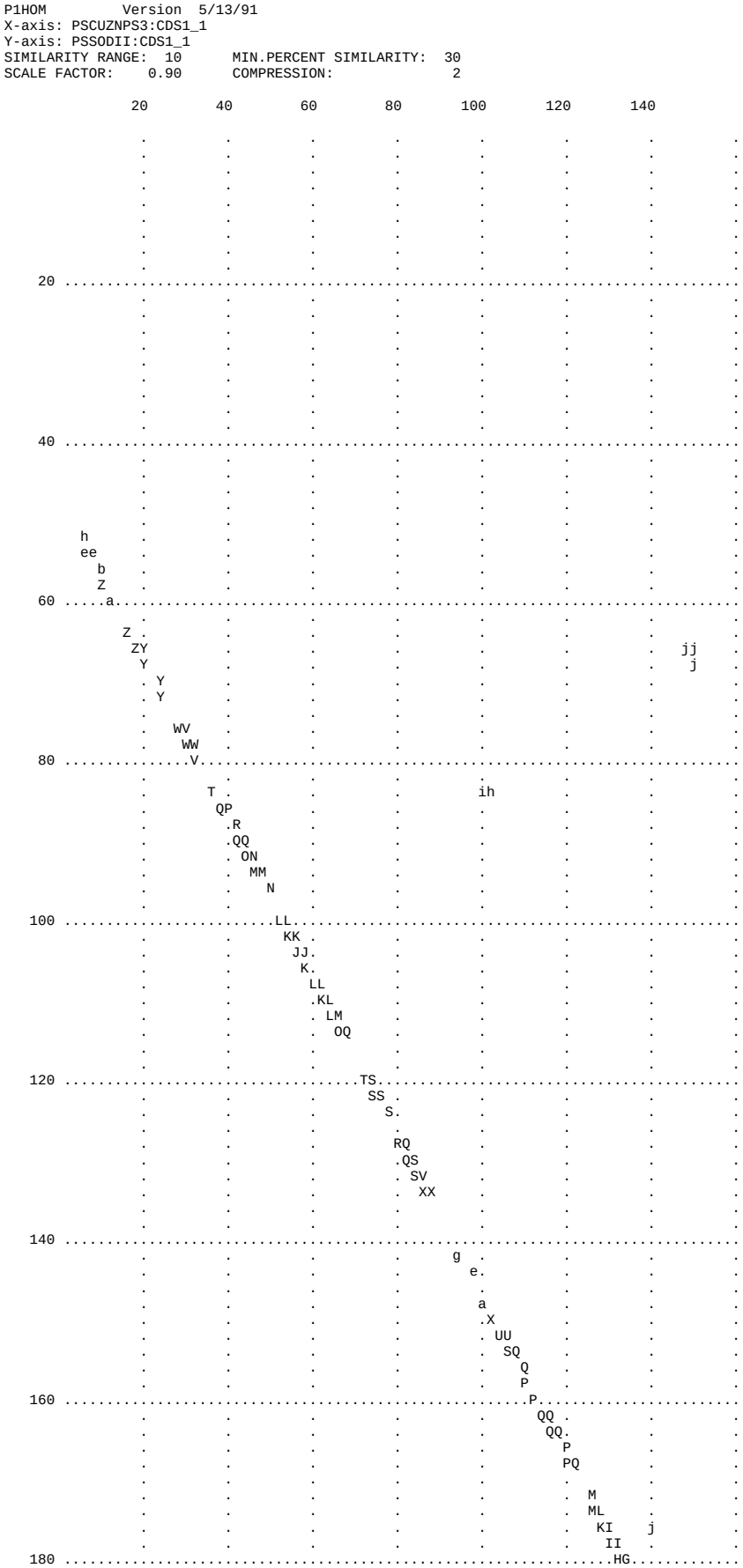
a. The alignment score is expressed as a deviation from randomness, according to information theory.
b. In scoring matrices such as PAM and BLOSUM, perfect identities between two sequences give the highest score, conservative replacements give intermediate scores, and uncommonly observed replacements 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.

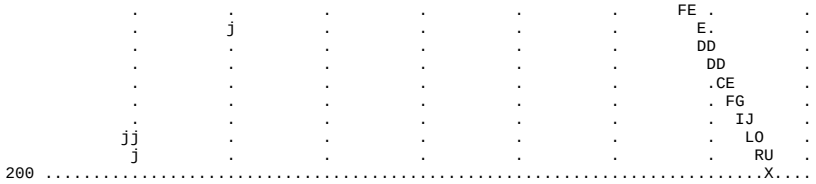
Example: For **a** above, your answer might be something like: The alignment score is shown in the output both as bits of information, and as the actual score in parenthese, calculated from the scoring matrix.

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?




FE
E.
DD
DD
.CE
. FG
. IJ
. LO
. RU
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		