PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Tuesday, October 25, 2011

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 equal 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) Below are several fragments from a VERY small chromosome. Your job is to indicate the order of the fragments, based on where the sequences overlap. (For simplicity, all data shown are from the same strand.)

a) If you were to assemble a composite sequence from these fragments by looking for overlaps between fragments, what would the order be?

b) By itself, is the data given adequate to determine whether this is a linear or circular chromosome?

a) 5' GTTTCACCCTTACCATGCCTAGGAATCGGGATCTT 3'

b) 5' AGGAATCGGGATCTTGACATGCACACCACACACACACACACAACA 3'

c) 5' GCCGCCGCTAACAAATCCTAGCGGGGTTTCA 3'

d) 5' AGCGGGGTTTCACCCTTACCATGCCTAGGAATCGGGATCTTGACAT 3'

e) 5' CACACACACACACACACACACACACAACAGTACGCCGCCG 3'

2. (10 points) Which of the following are NOT assumptions of multiple sequence alignment?

i) All sequences are homologous
ii) No duplicate sequences are present
iii) In each column, amino acid residues are homologous
iv) The alignment is optimal with minimal gaps
v) No back mutation has occurred (some methods take this into account)
vi) All sequences are the same length

3. (25 points) This question relates to restriction recognition sequences. To refresh your memory, a few examples of restriction sequences are listed below.

EcoRI	G^AATTC	5' protruding ends(*Escherichia coli)*
HindIII	A^AGCTT	" " " (*Haemophilus influenza*)
SmaI	CCC^GGG	blunt ends (*Serratia marcescens)*
XmaI	C^CCGGG	5' protruding ends (*Xanthomonas malvacaerum)* an isoschizomer [1]of Sma1
PstI	CTGCA^G	3' protruding ends (*Providencia stuarti*)
HinfI	G^ANTC	5' protruding ends (*H. influenza*). Degenerate recognition site. (GAATC,GAGTC,GACTC,GATTC)
HaeII	RGCGC^Y	3' protruding (*H. aegyptius*) 2^2=4 possible cuting sites: AGCGCC, AGCGCT, GGCGCC, GGCGCT
BglI	5'GCCN NNN^NGGC3' 3'CGGN^NNN NCCG 5'	assymetric, 3'protuding, (*Bacillus globigii*)
BbvI	5'GCAGC(N) $_8$3' 3'CGTCG(N) $_{12}$3'	assymetric, 3'recessed
[1]**isoschizomer** - restriction endonucleases that recognize the same sequences R = purine; Y = pyrimidine; N = {A,G,C or T}		

a) (5 points) Could you use a program like FASTA or BLAST to search for restriction sites in a DNA sequence? (Hint: P(k-mer) = $1/p^k$)

b) (10 points) If you wanted to write an efficient program to search for restriction sites in a DNA sequence, would it be possible to use a lookup table of k-mers to speed up the search, as is used in DXHOM, FASTA, or BLAST?

c) (5 points) Does a restriction site search program need to search both strands of a DNA sequence? Explain.

d) (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
```

4. (5 points) What does the algorithm below do?

```
Calculate distances between all possible pairs of sequences
Construct a Neighbor-Joining tree from pairwise distances
while not (all nodes on the tree have been visited)
     align each pair of sequences or profiles at the terminal nodes
     replace aligned sequences with a profile representing the alignment
        of all sequences in below that node
```

5. (5 points) What is an insertion/deletion event?

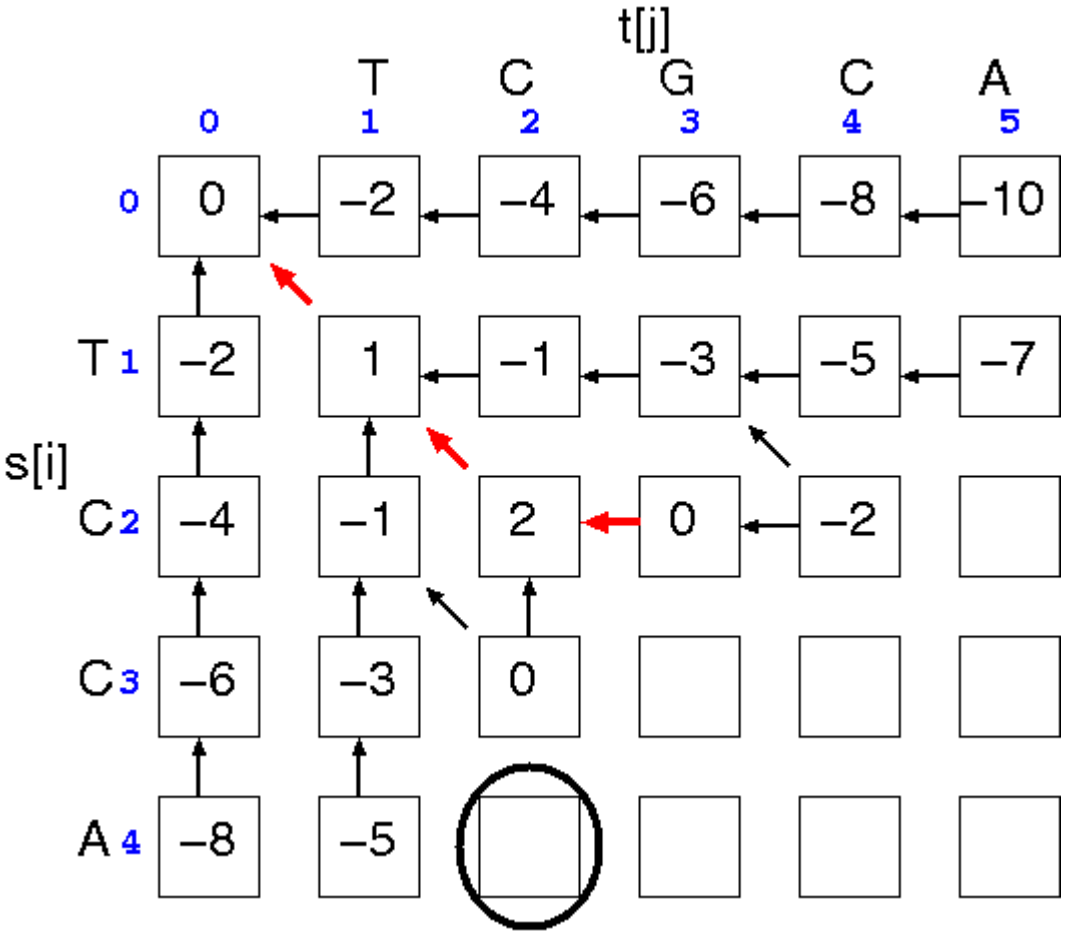
6. (5 points) What is the distinction between a global pairwise alignment and a local pairwise alignment?

7. (10 points) Affine gap penalties assign different scores for insertion of a gap and extension of a gap.

a) Why are there two distinct gap penalties?

b) What is the unrealistic assumption behind affine gap penalties?

8. (10 points)



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

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

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

```
AB017061_16  V  M  S  K  V  R  E  M  P  V
AC027034_19  A  L  S  E  V  R  E  M  P  I
```

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. (10 points) The CDS taken from GenBank entry X76860 (a genomic sequence) was translated, and the protein used as a query. The FASTA search shows an alignment between the query and the CDS for X76860 as found in the GenPept database. This is not surprising, since the sequence would be expected to find itself in the database. In the second search, the query protein is compared with the GenBank Plant division, using TFASTX, which translates each DNA sequence in the database into protein.

a) The TFASTX alignment has exactly the same amino acids as the FASTA alignment, but two gaps were inserted. Why does the insertion of two gaps lead to a higher E value in the TFASTX search, versus the E value in the FASTA search?

b) What is a simple explanation for the presence of two gaps in the TFASTX alignment?

FASTA - Query: protein; Database: GenPept (translation of CDS features)

```
>>X76860_1 X76860 1052551 type V Thionin A.squarrosa Ath  (131 aa)
 initn: 911 init1: 911 opt: 911  Z-score: 1246.6  bits: 236.3 E(): 1.2e-60
Smith-Waterman score: 911; 100.0% identity (100.0% similar) in 131 aa overlap (1-131:1-131)

               10        20        30        40        50        60
X76860 MGGGQKGLESAIVCLLVLGLVLEQVQVEGVDCGANPFKVACFNSCLLGPSTVFQCADFCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
X76860 MGGGQKGLESAIVCLLVLGLVLEQVQVEGVDCGANPFKVACFNSCLLGPSTVFQCADFCA
               10        20        30        40        50        60

               70        80        90       100       110       120
X76860 CRLPAGLASVRSSDEPNAIEYCSLGCRSSVCDNMINRADNSTEEMKLYVKRCGVACDSFC
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
X76860 CRLPAGLASVRSSDEPNAIEYCSLGCRSSVCDNMINRADNSTEEMKLYVKRCGVACDSFC
               70        80        90       100       110       120

              130
X76860 KGDTLLASLDD*
       ::::::::::::
X76860 KGDTLLASLDD
              130
```

TFASTX - Query: protein; Database: GenBank Plant division, translated

```
>>X76860 - A.squarrosa AthV1 gene.                          (781 aa)
 initn: 883 init1: 504 opt: 506  Z-score: 714.1  bits: 140.4 E(): 1e-31
trans. Smith-Waterman score: 662; 55.0% identity (55.0% similar) in 240 aa overlap (1-132:39-757)

               10        20        30        40        50        60
X76860 MGGGQKGLESAIVCLLVLGLVLEQVQVEGVDCGANPFKVACFNSCLLGPSTVFQCADFCA
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
X76860 MGGGQKGLESAIVCLLVLGLVLEQVQVEGVDCGANPFKVACFNSCLLGPSTVFQCADFCA
             60        90       120       150       180       210

               70
X76860 CRLPAGLASVRSS-----------------------------------------------
       :::::::::::::
X76860 CRLPAGLASVRSSGKDRQLATSNLYTILNLLADICMVRFDMNNVSIILKSFYQKIDNKE*
            240       270       300       330       360       390

                                      80        90
X76860 ----------------------------DEPNAIEYCSLGCRSSVCDNMINR-------
                                   ::::::::::::::::::::::::
X76860 MIEAMPIPKTSTYCNEGIIYVGSISFWQ/DEPNAIEYCSLGCRSSVCDNMINRGK*NPSE
            420       450       480        510       540       570
```

```
                                100       110       120       130
X76860 --------------------------ADNSTEEMKLYVKRCGVACDSFCKGDTLLASLDD
                                  :::::::::::::::::::::::::::::::::
X76860 YICISFCTCKLIGNLVLDAYYPILAPADNSTEEMKLYVKRCGVACDSFCKGDTLLASLDD
            600       630       660       690       720       750
```

11. (15 points) The graph below illustrates scores from a FASTA search against a protein database.
a) What is the distinction between the continuous curve and the diamonds?
b) This is not a normal, or Gaussian, Distribution, but rather, an Extreme Value Distribution (EVD). Based on the shape of the curve, what can you say about the difference between normally-distributed data, and data that falls on an EVD?
c) Draw a similar graph, showing what the curve would look like if many statistically significant "hits" were found in the database.