PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Thursday, October 17, 2019

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. (10 points) In a very real sense, the cell has to work with information in ways that are analogous to how we work with data in bioinformatics. For example, one might think of the eukaryotic nucleus as being analogous to the hard drive of a computer, where the data is stored as a DNA sequence. In this analogy, the fact that chromatin domains are uncoiled in the nucleus to allow transcription factors to find any gene, would be analogous to the fact that a disk drive is a random-access device, on which any file can be found by rotating the disk, and moving the read/write head in or out on the disk.

Describe another cellular process that has an analogy in bioinformatics or computer science. How does your analogy fit the process, and in what ways does the analogy break down? Feel free to use diagrams to make your point.

2. (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
```

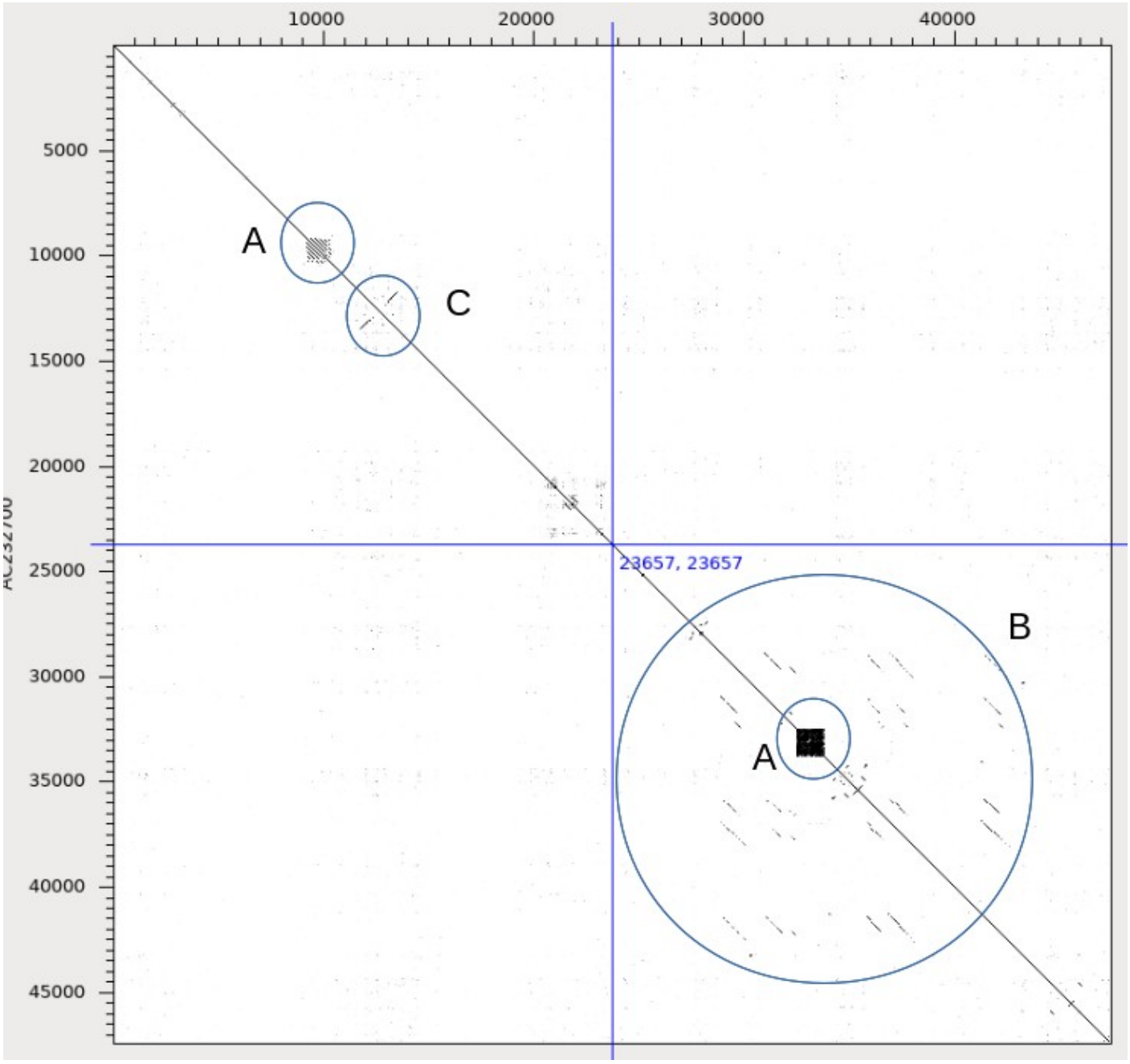
3. (10 points) Describe how thin clients such as Thinlinc divide tasks between the user's desktop computer and the remote server. How does this division of tasks make it possible for "Any user can do any task from anywhere".

4. (10 points) The following is an excerpt from a genomic sequence for a chlorophyll a/b binding protein from cotton (Accession number X54090).

```
mRNA            join(<454..599,690..>1341)
                /gene="cab"
gene            454..1341
                /gene="cab"
exon            <454..597
                /gene="cab"
                /number=1
CDS             join(454..599,690..1341)
                /gene="cab"
                /codon_start=1
                /product="chlorophyll ab binding protein"
                /protein_id="CAA38025.1"
                /db_xref="GI:452314"
                /db_xref="SWISS-PROT:P27518"
                /translation="MATSAIQQSAFAGQTALKQSNELVCKIGAVGGGRVSMRRTVKSA
                PTSIWYGPDRPKYLGPFSDQIPSYLTGEFPGDYGWDTAGLSADPETFAKNRELEVIHC
                RWAMLGALGCVFPEILSKNGVKFGEAVWFKAGSQIFSEGGLDYLGNPNLIHAQSILAI
                WACQVVLMGFVEGYRVGGGPLGEGLDPIYPGGAFDPLGLADDPDAFAELKVKEIKNGR
                LAMFSMFGFFVQAIVTGKGPIENLFDHLADPVANNAWAYATNFVPGK"
intron          600..689
                /gene="cab"
                /number=1
exon            691..>1341
                /gene="cab"
                /number=2
```

What is the difference between the join statements for the mRNA and CDS features, and what does that difference signifiy?

5. (10 points) The output below shows a pairwise comparison of a BAC clone from tomato with itself.



A (4 points) - Describe the two features labeled as A.

B (4 points) - Describe the reason for the parallel diagonals in region B.

C (2 points) - Describe the region labeled as C. (Note: This output is from the Dotter program, which shows similarity between the two forward strands as diagonals running from upper left to lower right, and similarity between the forward and reverse strand as diagonals running from lower left to upper right.)

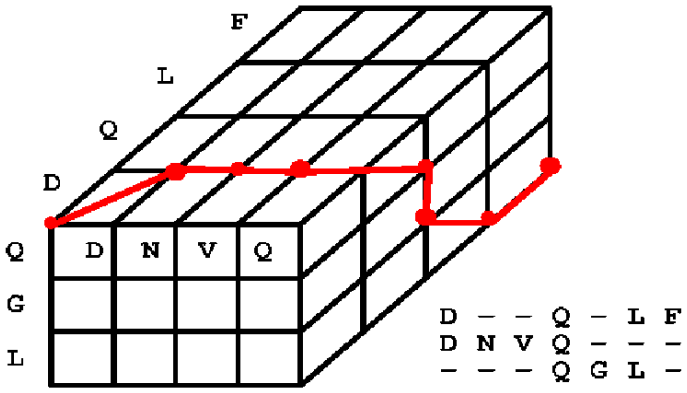
6. (10 points) You wish to design an oligonucleotide probe that would identify genes encoding the Superoxide dismutase protein. Given the following amino acid sequence from the SOD protein

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.

7. (5 points) Why is it important to eliminate duplicate sequences before doing a multiple protein alignment?

8. (10 points) We have discussed the problem of multiple sequence alignment by extending the Needleman-Wunsch (Smith-Waterman) pairwise alignment algorithm to k sequences. This is illustrated for k = 3 sequences below:



The time required for multiple alignment by this algorithm is $\mathbf{O(k^2 2^k n^k)}$, where

k is the number of sequences
n is the length of the alignment (assume all sequences are the same length)

Match each of the following phrases to one of the three terms in the expression above (ie. $\mathbf{k^2}$, $\mathbf{2^k}$ or $\mathbf{n^k}$)

a) the number of calculations that must be done to fill any given cell in the matrix
b) total number cells in the k-dimensional matrix
c) the number of pairwise comparisons between sequences at any given position in the alignment

d) Which of these three terms is the most important reason that exhaustive multiple alignment becomes impractical beyond a small number of sequences? That is, which term increases most rapidly as the number of sequences increases?

e) Aside from computational time, the memory (RAM) required to store the k-dimensional matrix also becomes a limitation. If you want to align 100 sequences, each of 200 amino acids in length, how many units of memory is needed to store the entire matrix?

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
```

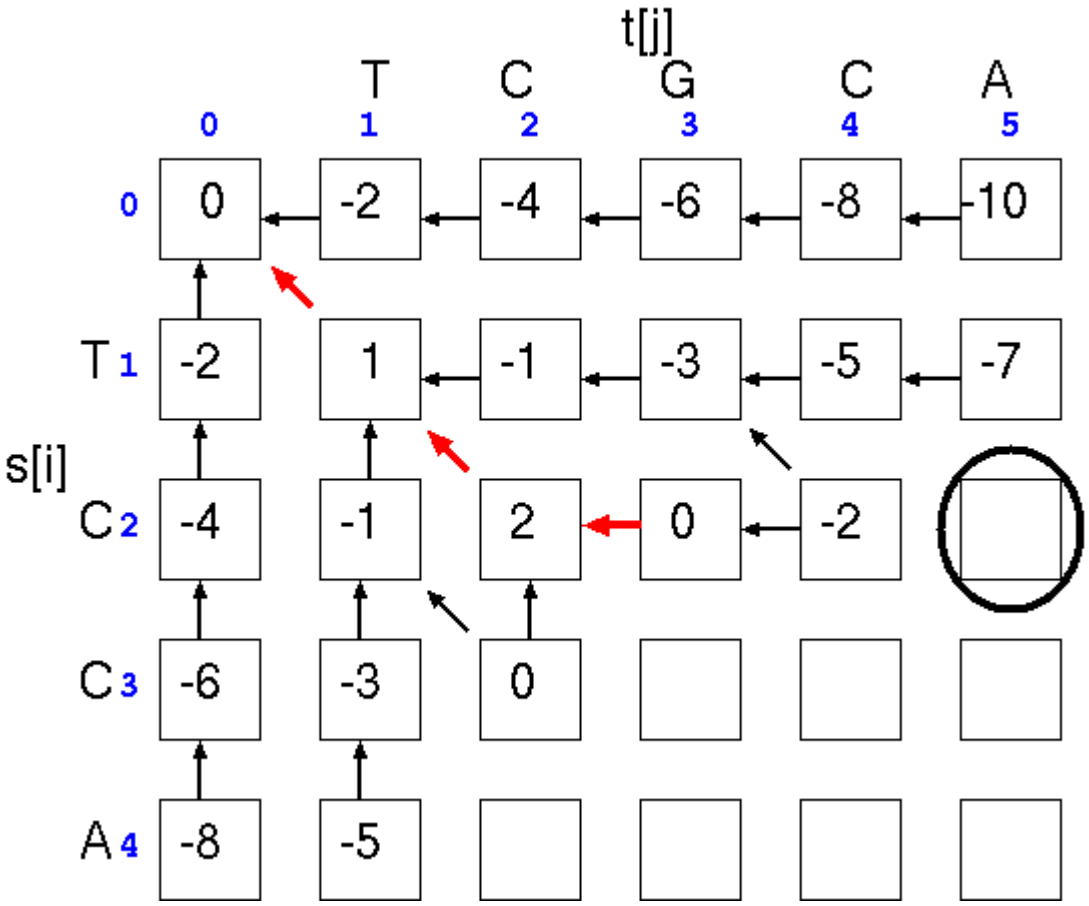
would result in the output

```
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 (similar to tblastn) 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
```

12. (20 points) Tblastn compares a protein sequence against sequences from a nucleotide database. As each database sequence is read, it is translated into protein in all 6 reading frames, and the proteins compared to the query sequence. On the next page, tblastn results are shown in which the query sequence was a 418 amino acid sequence for the human alpha-1-antitrypsin precursor (NP_001121174). The best hit from the RefSeq Gene database was a 20946 bp gene for serpin, a trypsin inhibitor (NG_008290). Some of the feature annotation from the serpin gene is shown on page 9.

a) (15 points) Keeping in mind that the query sequence is 418 amino acids long, explain why four shorter alignments were found. Use information from the annotation to support your explanation.

b) (5 points) In the tblastn output, the matches are almost perfect, with two exceptions. The last four positions in the first alignment show two mismatches, and the beginning of the third alignment has a region of very poor match, while the rest of the alignment matches perfectly. These sections of poor similarity are an artifact of how tblastn works. Explain the reason that these poor matches are shown in the alignment.

TBLASTN RESULTS

>NG_008290.1 Homo sapiens serpin family A member 1 (SERPINA1), RefSeqGene
on chromosome 14
Length=20946

```
 Score = 449 bits (1154),  Expect = 4e-141, Method: Compositional matrix adjust.
 Identities = 218/221 (99%), Positives = 219/221 (99%), Gaps = 0/221 (0%)
 Frame = +3

Query  1      MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFS  60
              MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFS
Sbjct  12456  MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDHPTFNKITPNLAEFAFS  12635

Query  61     LYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIHEGF  120
              LYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIHEGF
Sbjct  12636  LYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFNLTEIPEAQIHEGF  12815

Query  121    QELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQ  180
              QELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQ
Sbjct  12816  QELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSEAFTVNFGDTEEAKKQ  12995

Query  181    INDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKWERP  221
              INDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGK  +P
Sbjct  12996  INDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKVAQP  13118


 Score = 195 bits (495),  Expect = 3e-53, Method: Compositional matrix adjust.
 Identities = 91/91 (100%), Positives = 91/91 (100%), Gaps = 0/91 (0%)
 Frame = +1

Query  216    GKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAI  275
              GKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAI
Sbjct  14551  GKWERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAI  14730

Query  276    FFLPDEGKLQHLENELTHDIITKFLENEDRR  306
              FFLPDEGKLQHLENELTHDIITKFLENEDRR
Sbjct  14731  FFLPDEGKLQHLENELTHDIITKFLENEDRR  14823


 Score = 130 bits (328),  Expect = 3e-31, Method: Compositional matrix adjust.
 Identities = 67/80 (84%), Positives = 70/80 (88%), Gaps = 3/80 (4%)
 Frame = +1

Query  339    GADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVFLM  398
              G  L+     +PL+   AVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVFLM
Sbjct  17011  GISLTTCLCFSPLQ---AVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVFLM  17181

Query  399    IEQNTKSPLFMGKVVNPTQK  418
              IEQNTKSPLFMGKVVNPTQK
Sbjct  17182  IEQNTKSPLFMGKVVNPTQK  17241


 Score = 100 bits (248),  Expect = 4e-21, Method: Compositional matrix adjust.
 Identities = 50/50 (100%), Positives = 50/50 (100%), Gaps = 0/50 (0%)
 Frame = +3

Query  306    RSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSK  355
              RSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSK
Sbjct  16080  RSASLHLPKLSITGTYDLKSVLGQLGITKVFSNGADLSGVTEEAPLKLSK  16229
```

FEATURE ANNOTATION

```
   gene            7091..18946
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /note="serpin family A member 1"
                   /db_xref="GeneID:5265"
     mRNA          join(7091..7133,12452..13101,14552..14822,16082..16229,
                   17053..18946)
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /product="serpin family A member 1, transcript variant 1"
                   /transcript_id="NM_000295.5"
                   /db_xref="GeneID:5265"
     exon          7091..7133
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /inference="alignment:Splign:2.1.0"
                   /number=1
     exon          12452..13101
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /inference="alignment:Splign:2.1.0"
                   /number=2
     CDS           join(12456..13101,14552..14822,16082..16229,17053..17244)
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /note="protease inhibitor 1 (anti-elastase),
                   alpha-1-antitrypsin; serpin peptidase inhibitor, clade A
                   (alpha-1 antiproteinase, antitrypsin), member 1; alpha-1
                   antitrypsin; serine (or cysteine) proteinase inhibitor,
                   clade A, member 1; alpha-1-antitrypsin null; serpin A1;
                   epididymis secretory sperm binding protein;
                   alpha-1-antiproteinase; alpha-1 protease inhibitor; serpin
                   peptidase inhibitor clade A member 1; alpha-1-antitrypsin
                   short transcript variant 1C4; serpin peptidase inhibitor
                   clade A (alpha-1antiproteinase, antitrypsin) member 1;
                   alpha-1-antitrypsin short transcript variant 1C5"
                   /codon_start=1
                   /product="alpha-1-antitrypsin precursor"
                   /protein_id="NP_000286.3"
                   /db_xref="CCDS:CCDS9925.1"
                   /db_xref="GeneID:5265"
                   /translation="MPSSVSWGILLLAGLCCLVPVSLAEDPQGDAAQKTDTSHHDQDH
                   PTFNKITPNLAEFAFSLYRQLAHQSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILE
                   GLNFNLTEIPEAQIHEGFQELLRTLNQPDSQLQLTTGNGLFLSEGLKLVDKFLEDVKK
                   LYHSEAFTVNFGDTEEAKKQINDYVEKGTQGKIVDLVKELDRDTVFALVNYIFFKGKW
                   ERPFEVKDTEEEDFHVDQVTTVKVPMMKRLGMFNIQHCKKLSSWVLLMKYLGNATAIF
                   FLPDEGKLQHLENELTHDIITKFLENEDRRSASLHLPKLSITGTYDLKSVLGQLGITK
                   VFSNGADLSGVTEEAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNK
                   PFVFLMIEQNTKSPLFMGKVVNPTQK"
     exon          14552..14822
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /inference="alignment:Splign:2.1.0"
                   /number=3
     exon          16082..16229
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /inference="alignment:Splign:2.1.0"
                   /number=4
     exon          17053..18946
                   /gene="SERPINA1"
                   /gene_synonym="A1A; A1AT; AAT; alpha1AT; nNIF; PI; PI1;
                   PRO2275"
                   /inference="alignment:Splign:2.1.0"
                   /number=5
```

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		

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