PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 11:30 Thursday, October 22, 2020

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.

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) In this course, we have used numerous layers of software working in concert. Describe the distinction among the following. In other words, what do each of them do, and how do they differ, and how do they work together?

- Linux
- bash
- BIRCH
- BioLegato
- Thinlinc

3. (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?

4. (10 points) Two amino acid sequences were compared for similarity using SSEARCH. Next, the sequences were then randomized by local shuffling, and a second SSEARCH alignment was done on the two randomized sequences. The first five lines of the alignment are shown for each. Even if you didn't know which was the original, and which was done with randomized sequences, it should be easy to figure out. What differences do you see in the two outputs, that tells you which is which?

It is striking that the two alignments are about the same length 419 in the original vs. 435 aa in the alignment of randomized sequences. Normally we'd expect that the alignment with the original sequences would be much longer. What does it tell us that both alignments are of comparable length?

Original sequences

```
>>D_ 424 bp                                               (424 aa)
 s-w opt: 2259  Z-score: 2444.5  bits: 461.3 E(1): 2.4e-134
Smith-Waterman score: 2259; 80.4% identity (92.4% similar) in 419 aa overlap (3-417:7-423)

                   10         20          30        40        50
D_sali     MPSTSGASPFLPAAPA-LARRCSR--GPNGSSRRCSRAVPASSVSRSPTVAVQATL
             ..:..:: .  ::. .. : .:  . : . :  :: . :. ..   :: :. ::
D_     MAQRTATSSSSSPSIIYAPSPISNRSGRRAAANHGIRNGSRRA-AGRMGLCSTVQVNCTL
               10        20        30        40         50

             60        70        80        90       100       110
D_sali AMPSPD-SQRLRLQQQLQQQAQQQQAQQQLSGKDVEQAAMQACIRTATSVPPSSGVLDPS
       :::.:. .:..::::: ::: :::: :::::::.::. ::  ::.:: :::::.:.:.:
D_     AMPQPNHGQKMRLQQQQQQQLQQQQ-QQQLSGKQVEEQAMLQCIKTAQSVPPSTGLLNPR
      60        70        80         90       100       110

            120       130       140       150       160       170
D_sali GLRWRGGALEAAYERCGAVCKEYAKTFYLGTQLMTPVQARCIWAIYVWCRRTDELVDGPN
       ::::.:..::::::::::::.:::::::::::::::::::::::::::::::::::::::
D_     GLRWQGSSLEAAYERCGAVCSEYAKTFYLGTQLMTPVQARCIWAIYVWCRRTDELVDGPN
      120       130       140       150       160       170

            180       190       200       210       220       230
D_sali ASKITPQALDRWEERLNGVFQGRPYDVLDAALTDTISKFPLEVQPFRDMIEGMRMDLFKS
       ::::::::::::::::.:.:::.::::::::::::::::::::::::::::::::.::::
D_     ASKITPQALDRWEERLEGMFQGKPYDVLDAALTDTISKFPLEVQPFRDMIEGMRIDLFKS
      180       190       200       210       220       230
            240       250       260       270       280       290
D_sali RYQTFDELYEYCYRVAGTVGLMTVPVMGIDPNYKGPLDKVYRAALALGTANQLTNILRDV
       ::.::::::::::::::::::::.::::::::::::.::::.::::::::::::::::::
D_     RYHTFDELYEYCYRVAGTVGLMTMPVMGIDPNYKGPIDKVYKAALALGTANQLTNILRDV
      240       250       260       270       280       290
```

Randomized sequences

```
>>D_-rand 424 bp                                          (424 aa)
 s-w opt: 226  Z-score: 229.2  bits: 51.4 E(1): 5.9e-11
Smith-Waterman score: 226; 27.6% identity (49.0% similar) in 435 aa overlap (4-409:12-406)

                       10          20             30        40
D_sali         FGTAPSSSMPLAAPPAARR--LCSGSG-----RRSNPARPSVSCRSAARVVS
                  : :::.:    :. ::  .  .::     : ..     . :  .  ::.
D_-ran MTQSSASTRAIAYSSSIP----PSPRRAGISRNSGAINSHRNAGALAGRMRCGRTCSVVN
               10            20        30        40        50

          50        60        70         80        90       100
D_sali QTSPVMAAPPTSDSLQQLLQRQQLRQQQQQ-AQQQAAGEDQSLVQKQAIAMATTRCDGSS
        :. .      .   :::::.:: .::::: .: :   . . :  .:::   :.. . ..
D_-ran -TAQLPQQMHKNGMPQQLLQQQQRQQQQQQLSQGQQMQQKEVLEASQAICPPTVKGNLTG
          60        70        80        90       100       110

              110       120        130       140         150
D_sali VL-VP---SPGGWPALGSRRERCLAEG-YAAEFYAKYTCKVLMGTP--TQVQLYRAICWV
        : .:   . . :   ::.:  :  :  :: . :.  : ..    :  : ....:   :.
```

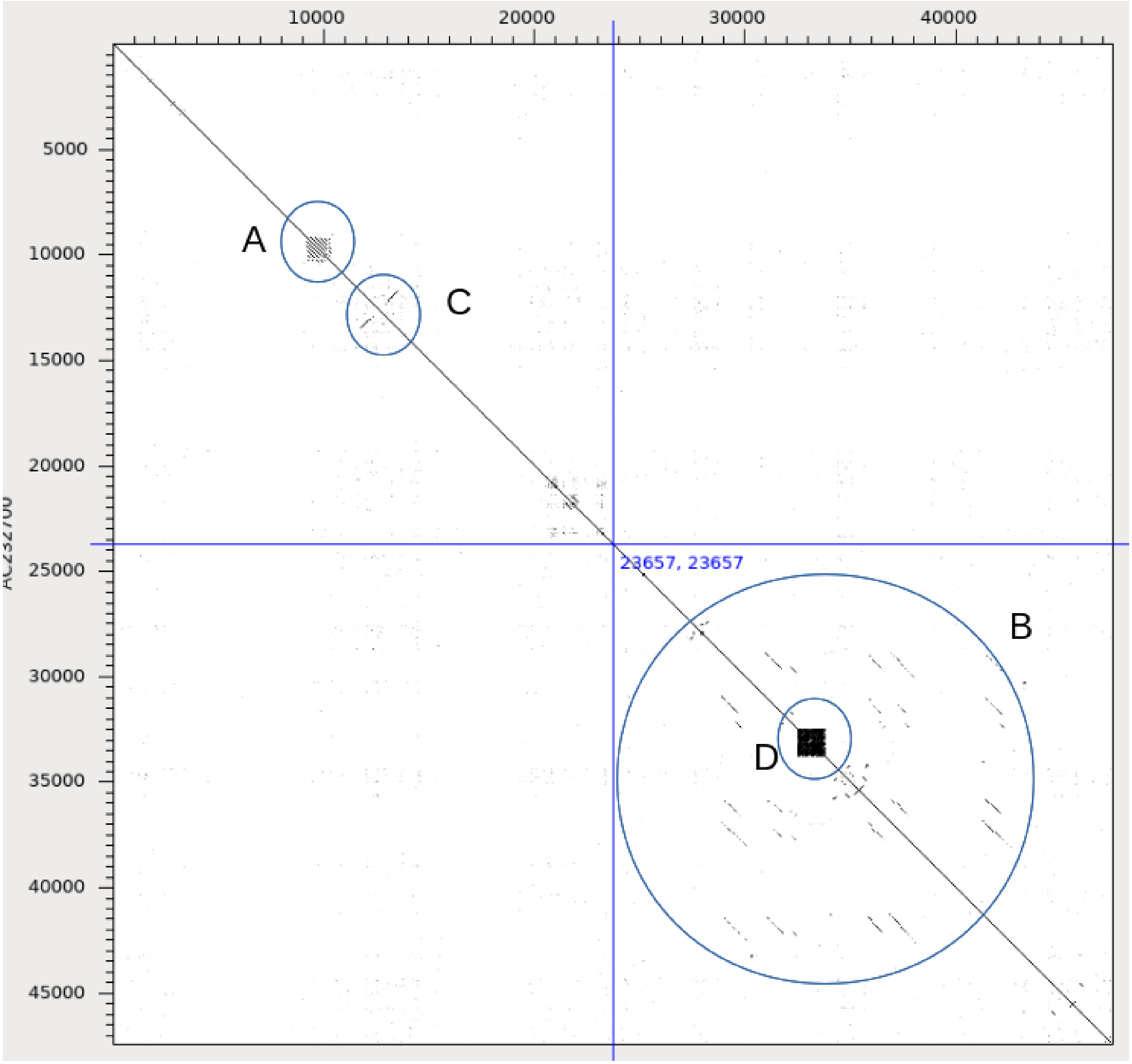
```
D_-ran SLRLPQREAASLWS--GSERVGCACEYLYAFKTYGTQTMQICALRPVCTYIRVWR---WA
         120         130       140       150       160          170

       160       170       180       190       200       210
D_sali IAWGRREDCVDTLKIAPPANQTSWGLRRLNDEELRGFDPQVYVIATSDDTKALQPDVFFR
          :.  . :.   : :    :.:      .:. :::  .. :.   : .  :.:  : .
D_-ran ---GEDPNDVSLADIRP---LTKWQ----AEEQGRGFMLEAKVLYDLDPATDLKP--FSE
                 180              190       200       210

        220       230          240           250          260
D_sali L-PERMDMEGMFILLRQS---YDTKEFVYYAETG----CYR---MLPGVMVGVTNPKYPI
       . :  .:.::.   ::.    .      .: : :     ::   . : ::.:  . :::
D_-ran ITPMIQDFEGVRDRLRKYSIMFTDLYHEFYCELGMGVAVYRTDNIPPTVMMG--GKKYPD
```

5. (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?

6. (10 points) The output below shows a pairwise comparison of a BAC clone from tomato with itself.

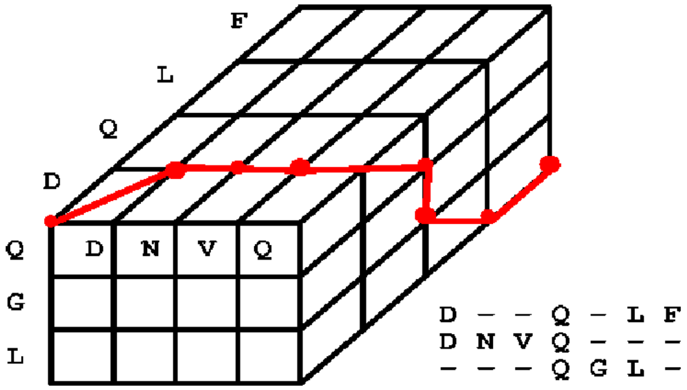


A (4 points) - Describe the two features labeled as A and D

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.)

7. (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 **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. **k^2, 2^k** or **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?

8. (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
```

9. (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
```

10. (5 points) The sizes of two protein databases in a 2020 release of GenBank were as follows:

DB name	size (Mb)
refseq_protein	170800
swissprot	695

Suppose that you have just cloned a new gene from an obscure species, and search both databases. It turns out that one match is present in both databases, giving identical alignments in both searches. However, the calculated E-values calculated for the refseq search is much larger than the E-value for Swissprot. Explain the different E-values. The equation for E is given:

$$E = Kmn\, e^{-\lambda S}$$

11. (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

12. (10 points) Answer the following questions about the table below:

a) By random chance alone, what is the probability that an amino acid chosen from one protein will match a given amino acid from another protein?

b) By random chance alone, what is the probability that a nucleotide from one DNA sequence will match a nucleotide from another DNA sequence?

c) When comparing two amino acid sequences for similarity, if you use a k value of 3, how much would you expect to speed up the search?

d) Typically, proteins are only a few hundred amino acids long. How might that affect the actual speedup of the algorithm, given a k value of 3?

e) When comparing two DNA sequences, what is the probability a 20 base segment from one sequence will match a given 20 base segment from another sequence? Express the answer as an exponential number ie. scientific notation.

Table 2.	Avg. dist. between k-matches 1 ------- $\mathbf{p^k}$			
Prob. of a match (**p**)	**k**= 2	3	4	5
0.050	400	8000		
0.075	178	2370		
0.100	100	1000		
0.150	44	296		
0.200	25	125		
0.250	16	64	256	1024
0.300	11	37	123	412
0.350	8	23	67	190
0.450	5	11	24	54
0.600	3	5	8	13
0.700	2	3	4	6
0.900	1	1	1	2

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	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