PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Thursday, October 21, 2021

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. (5 points) In the multiple alignment tutorial, we saw that pal2nal.pl creates a DNA multiple alignment. Pal2nal.pl requires 2 files as input: a multiple alignment of proteins and a set of unaligned DNA sequences. Why isn't it possible to create a DNA alignment directly from the aligned protein sequences, simply by reverse translating amino acids into the corresponding DNA?

2. (15 points) If you wanted to design an oligonucleotide as a hybridization probe, you want to ensure that the oligo sequence is unique within the genome ie. it is not likely to occur by random chance. To help in your calculations, a table is given with some relevant information.

a) How big would an oligo probe have to be for use with haploid yeast, *Saccharomyces cerevisiae*, (1N = 1.2 x 10^7 bp)? That is, how long does the oligo have to be to ensure that it is not likely to occur in the genome due to random chance?

b) Yeast also go through a diploid phase. If you were hybridizing to DNA extracted from diploid yeast, would you need to use a longer oligo? Explain.

n	4^n	2 x 4^n
10	1.05E+06	2.10E+06
11	4.19E+06	8.39E+06
12	1.68E+07	3.36E+07
13	6.71E+07	1.34E+08
14	2.68E+08	5.37E+08
15	1.07E+09	2.15E+09
16	4.29E+09	8.59E+09
17	1.72E+10	3.44E+10
18	6.87E+10	1.37E+11

c) Most eukaryotic genomes, especially for higher organisms, are largely composed of middle repetitive sequences such as the AluI family in mammals. How would this affect our estimates of the likelihood of finding a particular oligonucleotide in a eukaryotic genome?

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. (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 -v '^>' | grep -i -e GAATTC ASTRASTL2A.fsa
```

(-i tells grep to ignore case; -v '^>' tells grep to ignore name lines; -e GAATTC tells grep to search for the expression GAATTC.)

6. (5 points) The BLAST database services at NCBI must process over 100,000 BLAST searches per day. Researchers at NCBI realized that the most critical bottleneck in the process was the simple matter of reading in all the sequence data when comparing a query sequence with all sequences in a database. What solution was found to solve this problem?

7. (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

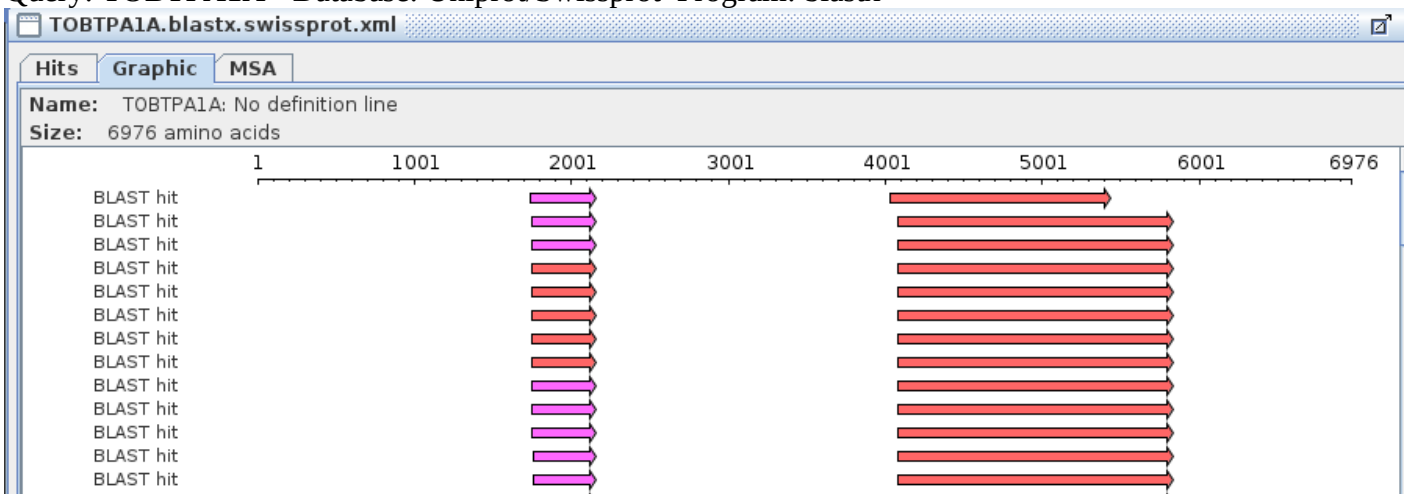
8. (10 points) A tobacco clone for the PAL gene (6976 bp) has the following features:

```
source          1 6976
gene            <1754 >5833
CDS             1754 5833
exon            <1754 2151
mRNA            <1754 5833
intron          2152 4083
exon            4084 >5833
```

The entire sequence was used as a query for a blastx search of swissprot, and for a tblastn search of refseq_rna. The top hits for a blastviewer graphic alignment from both searches is compared below. Hits represent sequences from many different species.

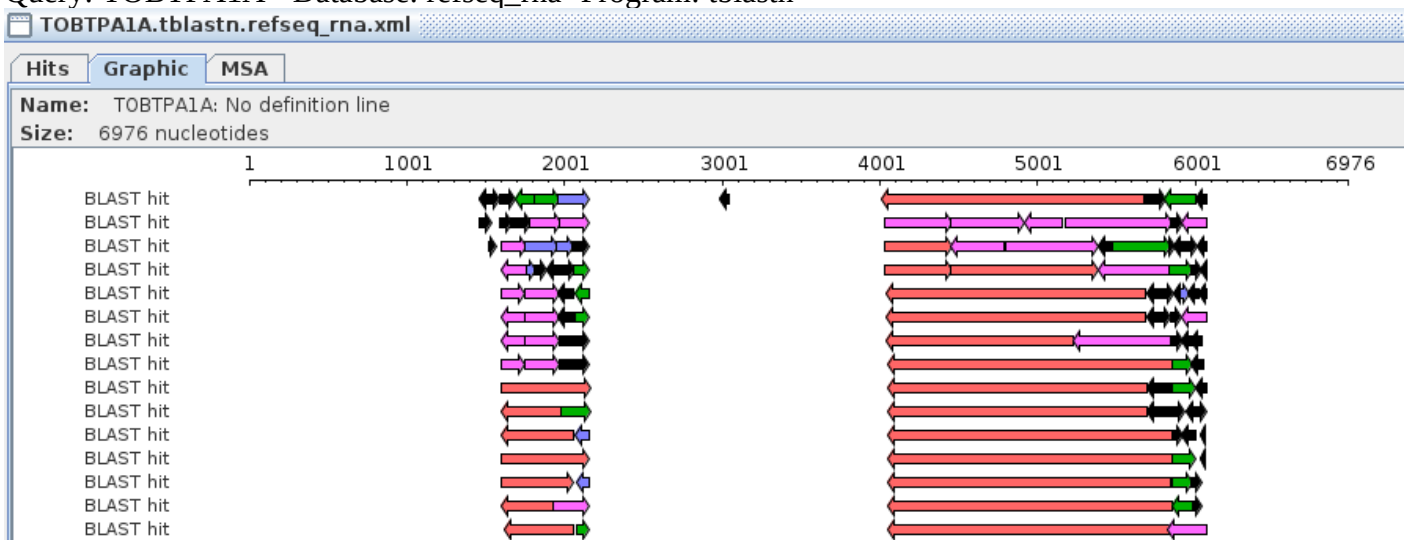
a) In the blastx/swissprot search, why are two solid arrows shown for the top hits?

Query: TOBTPA1A Database: Uniprot/Swissprot Program: blastx



b) In the tblastn/refseq_rna search, the arrows indicating hits are more complex. In general, what does this search tell you about the evolution of the PAL genes that was not seen in the UniprotSwissprot output?

Query: TOBTPA1A Database: refseq_rna Program: tblastn



9. (15 points) A user wants to find GenBank entries for the E. coli β-galactosidase gene. For each each NCBI keyword search in the nucleotide database, the query terms are shown, followed by the results. In each case, explain the results.

a) QUERY: E. coli [ALL] AND galactosidase [ALL] AND 1:500000[Sequence Length]

COUNT: 3552

b) QUERY: E. coli [Organism] AND galactosidase [ALL] AND 1:500000[Sequence Length]

COUNT: 2990

c) QUERY: E. coli [Organism] AND galactosidase [Protein name] AND 1:500000[Sequence Length]

COUNT: 135

5	#QKEY: 1			
6	#COUNT: 135			
7	#uid	Title	BioMol	Slen
8	NZ_WSPU01000610	Escherichia coli strain 8374wH5 NODE_612_length_576_cov_0.801782_ID_16276, ...	genomic	576
9	NZ_NWPN01000303	Escherichia coli strain MOD1-EC4310 MOD1-EC4310_653_length_394_cov_2.95584,...	genomic	394
10	RDTQ01000019	Escherichia coli strain EC45_ST57C scaffold_18, whole genome shotgun sequence	genomic	81272
11	VUEE01000019	Escherichia coli strain EcFF394 NODE_19_length_89234_cov_51.3448, whole gen...	genomic	89234
12	VUED01000054	Escherichia coli strain EcFF421 NODE_54_length_22559_cov_46.0373, whole gen...	genomic	22559
13	VUEM01000054	Escherichia coli strain EcFF211 NODE_54_length_22559_cov_44.1544, whole gen...	genomic	22559
14	VUEF01000019	Escherichia coli strain EcFF391 NODE_19_length_89233_cov_55.3069, whole gen...	genomic	89233
15	VRVV01000033	Escherichia coli strain CD64_7 NODE_28_length_73665_cov_34.551660, whole ge...	genomic	73665
16	SSUW01000071	Escherichia coli K-12 strain 70 GCID_CRE_0141_NODE_71, whole genome shotgun...	genomic	13077
17	QFAZ01000037	Escherichia coli strain E-4 NODE_37_length_15547_cov_22.1509, whole genome ...	genomic	15547
18	SRMZ01000007	Escherichia coli strain BX1S20 NODE_7_length_198876_cov_118.762, whole geno...	genomic	198876
19	QESC01000155	Escherichia coli strain 211_1 NODE_158_length_411_cov_0.809859_ID_5363, who...	genomic	411
20	SHKF01000081	Escherichia coli strain EC_03 NODE_81_length_12048_cov_19.433047, whole gen...	genomic	12048
21	SHJW01000062	Escherichia coli strain EC 103 NODE 62 length 17218 cov 20.372615, whole ge...	genomic	17218

(*Only a partial listing of hits is shown*.)

d) The gene for β-galactosidase is only about 2500 bp long. In c above, some of the hits are very large. Why is there such a range of sequence lengths for the hits?

e) What would be the problem with trying to find the sequence of this gene using a generic search engine such as Google?

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) 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 as shown above, write the annotation for the intron and exon features.

13. (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 $\frac{1}{p^k}$			
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	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