PLNT4610 BIOINFORMATICS

MID-TERM EXAMINATION

08:30 - 9:45 Tuesday, October 25, 2022

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 13 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 (CDS) sequences. Why isn't it possible to create a DNA alignment using only 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. (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?

5. (5 points)

In the Basic shell scripting tutorial, we created a script called testprot.sh, that reads a FASTA format file and tests whether the file contains DNA or protein.

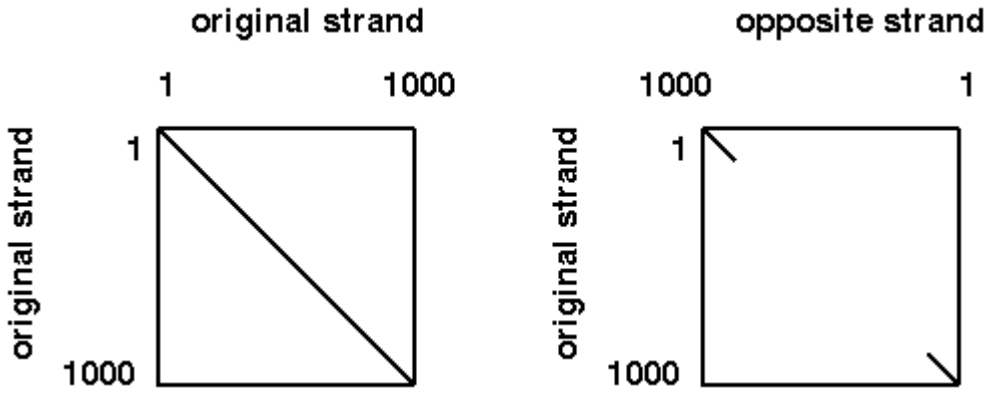
If this script were used with a GenBank DNA file, would it correctly indicate whether the file contained DNA or protein?

Explain your answer.

```
#!/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
```

6. (5 points) The longest chromosome sequenced so far is the *T. aestivum* (wheat) chromosome 3B at 851,934,019 bp. Since wheat is one of the largest known genomes, it is unlikely that chromosomes will be found in other species that are much larger than that. DNA sequences are normally represented as a string of bytes, one byte per nucleotide. Based on these numbers, could a typical desktop computer hold the entire sequence in memory at once? Should we be worried that someday nature will give us a chromosomal sequence too big to read into a computer's memory (RAM)?

7. (10 points) A sequence was compared with its opposite strand, showing short diagonals at each end. Explain this observation.



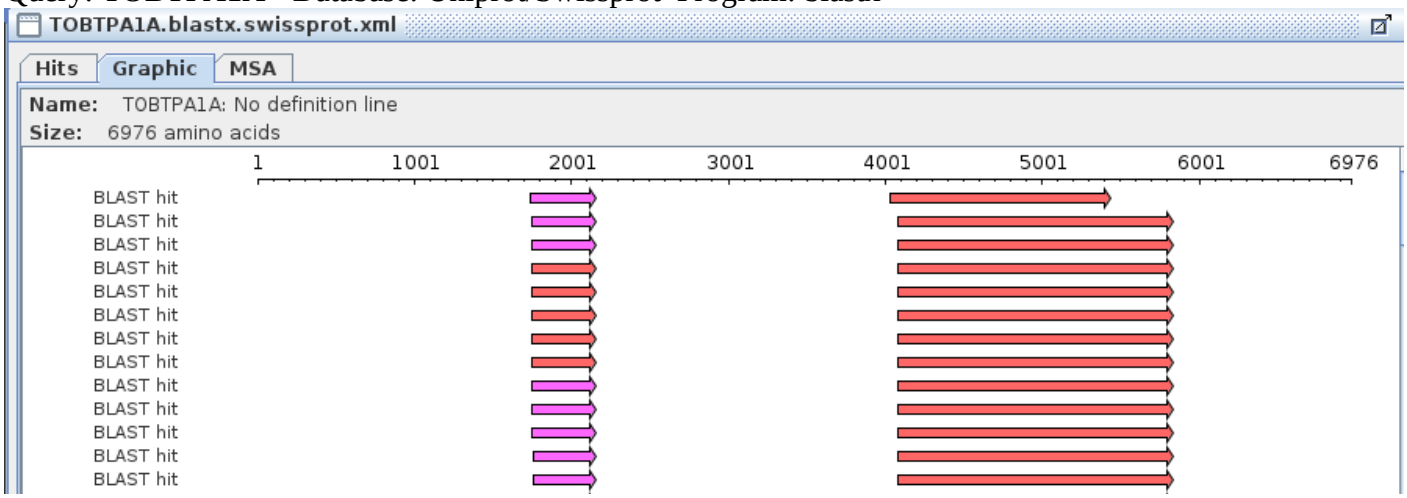
8. (10 points) A tobacco clone for the PAL gene (6976 bp) has the following features:

```
source          1..6976
gene            <1754..>5833
CDS             1754..5833
exon            <1754.. 2151
mRNA            join(<1754..2151,4084..>5833)
intron          2152..4083
exon            4084..>5833
```

The entire sequence was used as a query for a blastx search of swissprot, and for a tblastx 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.

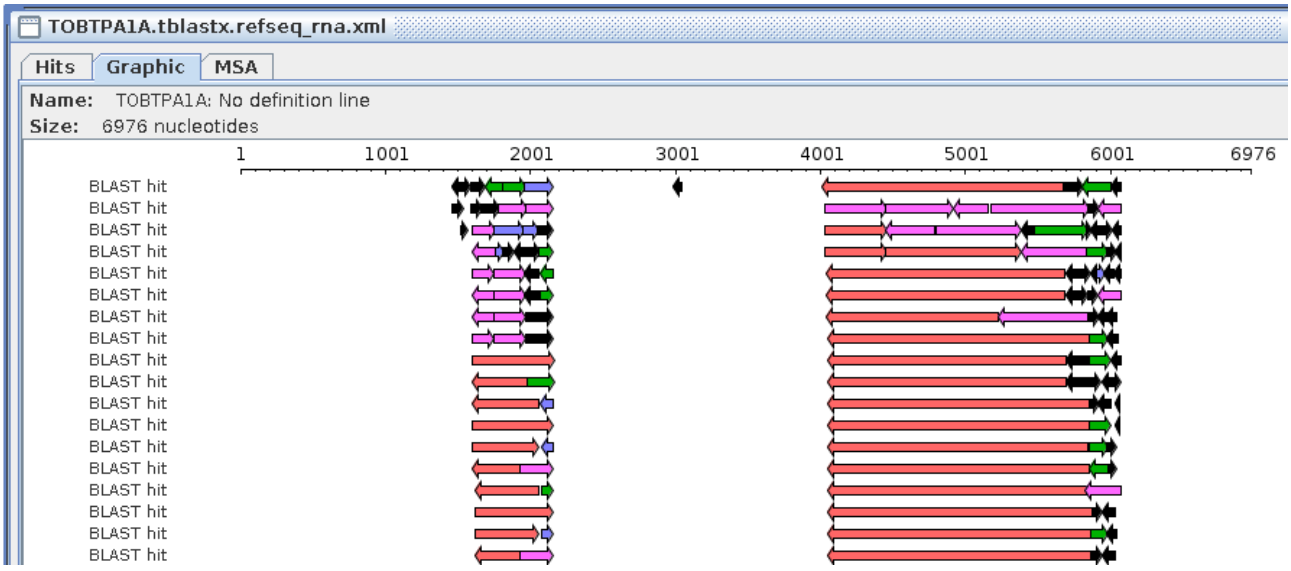
a) In the blastx/swissprot search, why are two sets of solid arrows shown for the top hits?

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: tblastx



9. (15 points) A researcher 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*.)

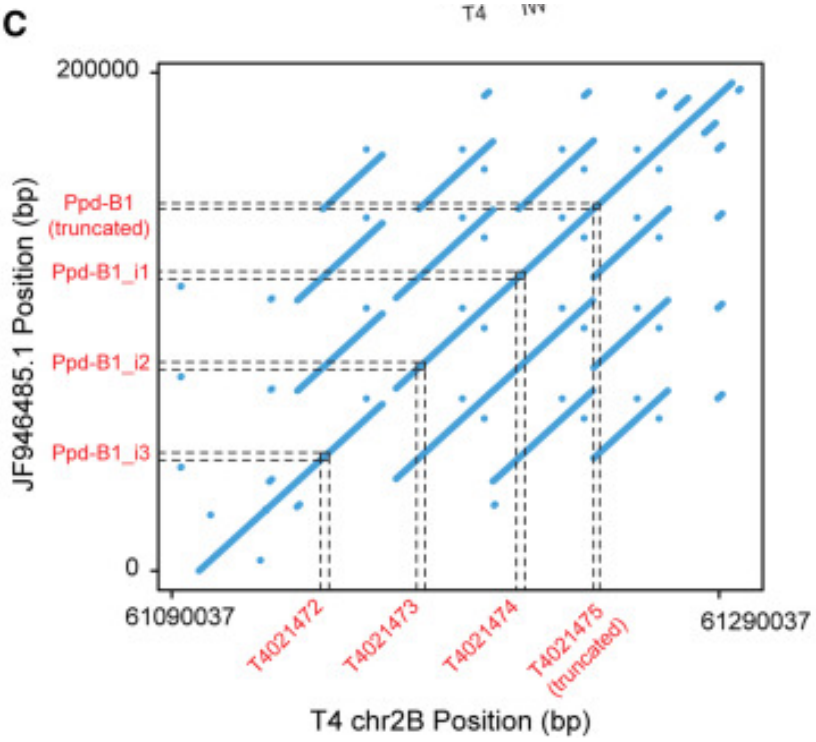
d) The gene for β-galactosidase is only about 2500 bp long. In c above, some of the hits are very large. Why is there such a range of sequence lengths for the hits?

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

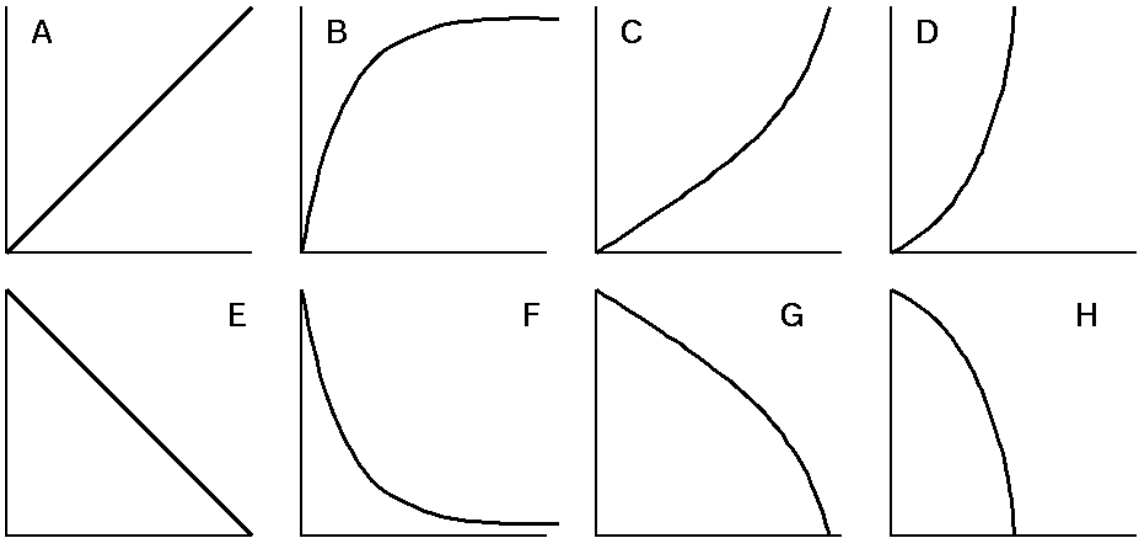
10. (10 points)

The dotplot at right shows a comparison of a 200,000 bp region of a chromosome from two different wheat varieties. The dashed lines show the locations of copies genes encoding a pseudo response regulator protein (PRR).

What do these results tell you about the recent evolution of this chromsomal region in wheat?



11 (10 points) For each choice I - V, indicate which graph most closely approximates the sentence. You can interpret "vs." to mean "as a function of". Graphs may be used twice. Some graphs will not be used at all.



I - E-value vs. size of database in a sequence database search.

II Time required to do a pairwise alignment of 2 sequences.

III. Alignment score vs. the number of iterations of MAFFT algorithm FFT-NSi.

IV. Time required to build a lookup table vs. sequence length.

V. Time required to do a multiple sequence alignment using a k-dimensional Needleman-Wunsch algorithm.

12. (10 points) Five algorithms are stated below. Using a table for A - E, fill in the roman numeral for the phrase describing each algorithm. Not all phrases match an algorithm.

A	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

B	pre-calculate all pairwise alignments between each pair of sequences to create a library of aligned sequence pairs for each aligned sequence pair in the library calculate all possible alignemts with sequence c choose the highest-scoring three-way alignment

C	**input**: Sequences: s of length m, t of length n
	const: MINPER // minimum percentage match
	output: matrix a[1..m,1..n]
	Maketable(TAB(x,y,z),t) // make lookup table using t
	for i = 1..m // for each nucleotide in s
	  set x,y,z to central triplet in window
	  for *each position t listed in TAB(x,y,z)*
	  if *MINPER/l bases match* **then**
	  a[i,j] **=** CharCode(MINPER/l)
	  //CharCode returns character to print
	  // for a given percent identity

D	**input**: Sequences: s of length m, t of length n
	output: matrix a[1..m,1..n]
	
	for i = 1..m // for each nucleotide in s
	  for j = 1..n // for each nucleotide in t
	  a[i,j] **= max(a[i,j-1-2, a[i-1,j-1]+p(i,j), a[i-1,j]-2)**

E	**input**: Sequences: s of length m, t of length n
	output: matrix a[1..m,1..n]
	
	for i = 1..m // for each nucleotide in s
	  for j = 1..n // for each nucleotide in t
	  if s[i] = t[j] **then**
	  a[i,j] **= 'A'**

I - Dot-matrix similarity plot of two sequences, **l = 1**
II - Dot-matrix similarity plot of two sequences, **l >= k**
III - Multiple sequence alignment (eg. Clustal)
IV - Multiple sequence alignment (eg. TCOFFEE,MAAFT)with WSP and consistency scores
V - Needleman-Wunsch algorithms
VI - BLAST algorithm

A	
B	
C	
D	
E	

13. (10 points) Two antifreeze proteins were aligned using both GGLSEARCH and GLSEARCH.

a) Which of the two alignments is deemed to be more statistically significant? Give a reason.

b) Why does the GGSEARCH alignment have a long gap, followed by a phenylalanine (F) at the end of ISP2_H? How does that gap contribute to the difference in Needleman-Wunsch (n-w) scores?

GGSEARCH

```
Algorithm: Global/Global affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL62 matrix (11:-4), open/ext: -11/-1

>>ISP2_OSMMO 175 bp                                          (175 aa)
 n-w opt: 315  Z-score: 295.7  bits: 61.1 E(1): 1.3e-133
global/global (N-W) score: 315; 39.1% identity (65.4% similar) in 179 aa overlap (1-163:1-175)

               10         20         30         40         50
ISP2_H MLTVSLLVCAMMALTQA-NDDKILKGTATEAGPVSQRAPPNCPAGWQPLGDRCIYYETTA
       ::. .:::::::.:::.: : :   ... : .   : .   .::. :. .. ::. ..
ISP2_O MLA-ALLVCAMVALTRAANGDTGKEAVMTGS---SGKNLTECPTDWKMFNGRCFLFNPLQ
                10         20         30           40         50

      60         70         80         90          100        110
ISP2_H MTWALAETNCMKLGGHLASIHSQEEHSFIQTLN-AGVV--WIGGSACLQAGAWTWSDGTP
       . :: :. .::: :..:::::: ::..:.. :. ::..  ::::: :  .  : : :.:
ISP2_O LHWAHAQISCMKDGANLASIHSLEEYAFVKELTTAGLIPAWIGGSDCHVSTYWFWMDSTS
         60         70         80         90         100        110

        120        130        140        150        160
ISP2_H MNFRSWCSTKPDDVLAACCMQMTAAADQCWDDLPCPASHKSVCAMT-----------F
       :.: .::...:: .:. ::.:..... .::.: ::   : ::::             .
ISP2_O MDFTDWCAAQPDFTLTECCIQINVGVGKCWNDTPCTHLHASVCAKPATVIPEVTPPSIM
        120        130        140        150        160        170
```

GLSEARCH

```
Algorithm: Global/Local affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL62 matrix (11:-4), open/ext: -11/-1

>>ISP2_OSMMO 175 bp                                          (175 aa)
 n-w opt: 336  Z-score: 328.6  bits: 67.2 E(1): 4e-171
global/local score: 336; 41.9% identity (69.5% similar) in 167 aa overlap (1-163:1-163)

               10         20         30         40         50
ISP2_H MLTVSLLVCAMMALTQA-NDDKILKGTATEAGPVSQRAPPNCPAGWQPLGDRCIYYETTA
       ::. .:::::::.:::.: : :   ... : .   : .   .::. :. .. ::. ..
ISP2_O MLA-ALLVCAMVALTRAANGDTGKEAVMTGS---SGKNLTECPTDWKMFNGRCFLFNPLQ
                10         20         30           40         50

      60         70         80         90          100        110
ISP2_H MTWALAETNCMKLGGHLASIHSQEEHSFIQTLN-AGVV--WIGGSACLQAGAWTWSDGTP
       . :: :. .::: :..:::::: ::..:.. :. ::..  ::::: :  .  : : :.:
ISP2_O LHWAHAQISCMKDGANLASIHSLEEYAFVKELTTAGLIPAWIGGSDCHVSTYWFWMDSTS
         60         70         80         90         100        110

        120        130        140        150        160
ISP2_H MNFRSWCSTKPDDVLAACCMQMTAAADQCWDDLPCPASHKSVCAMTF
       :.: .::...:: .:. ::.:..... .::.: ::   : ::::
ISP2_O MDFTDWCAAQPDFTLTECCIQINVGVGKCWNDTPCTHLHASVCAKPATVIPEVTPPSIM
    120    130    140    150    160    170
```

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