*Name:*_____
ID : _____

ECS 129: Structural Bioinformatics

March 19, 2018

Notes:

1) The final exam is open book, open notes.
2) The final is divided into 3 parts, and graded over 100 points, with 8 possible extra credit points (part III)
3) You can answer directly on these sheets (preferred), or on loose paper.
4) Please write your name at least on the front page!
5) Please, check your work! If possible, show your work when multiple steps are involved.

Part I (15 questions, each 4 points; total 60 points)

1) How many possible alignments of length M, with no gaps, can you form when you compare two sequences of length N and M, with N > M?

A) 1
B) N-M
C) N-M+1
D) M
E) N

There are N-M+1 positions in sequence 1for the first letter of the sequence of length M that leads to an alignment of length M.

2) In the dynamic programming matrix below, what is the score in the cell identified with an interrogation mark (?). Assume that the score for a perfect match is set to 10, the score of a mismatch is set to 0, and gap penalties are set to -2, independent of length

	G	Y	W	W	C	A
W	0	-2	8	8	-2	-2
W	-2	0	8	18	8	6
C	-2	-2	0	**8**	28	16

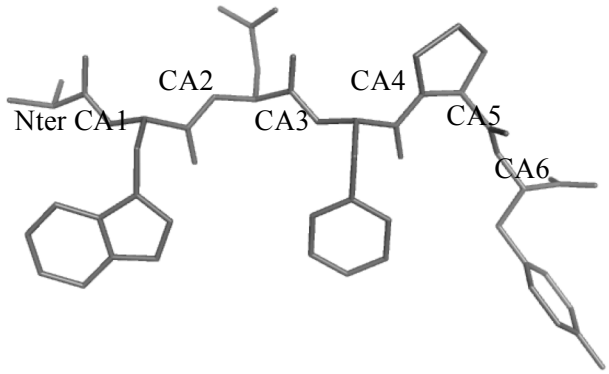
A) 10
B) 18
C) 8
D) 6
E) 0

3) The Ramachandran plot:

A) Compares the conformation of the side-chains of a protein.
B) Shows the accessibility of all amino acids in a protein
C) Shows the relationship between the torsion angles ϕ and ψ, for each amino acid in the protein
D) Shows the torsion angle around the peptide bond, for each amino acid in the protein
E) Shows the number of hydrogen bonds that stabilize a protein

*Name:*_____
ID : _____

4) The figure below shows a small peptide of six amino acids; give its sequence: (hint: there is one charged amino acid at physiological pH – from pH 5.5 to pH 8.0; hydrogens are not shown)



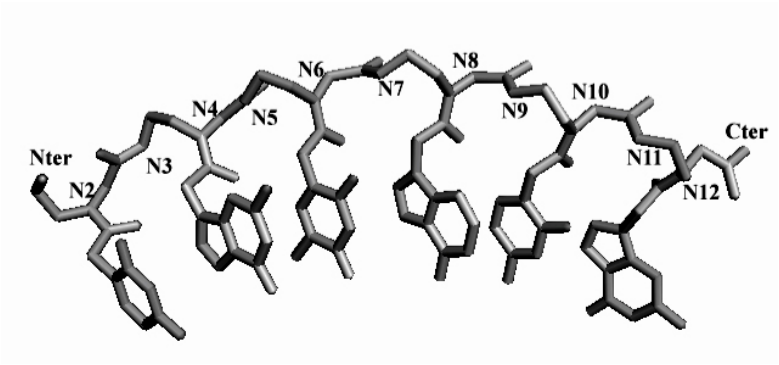
A) AWEFGF

B) AWDFPY

C) AYDYPW

D) AWNFPY

5) Peptide Nucleic Acids, or PNAs, are synthetic oligomers with a protein backbone on which bases (purines and pyrimidines) are linked every second N. Unlike DNA, PNAs do not contain sugars or phosphate groups. PNAs are represented as proteins, from Nter to Cter. Find the "sequence" from Nter to Cter of the PNA shown below:



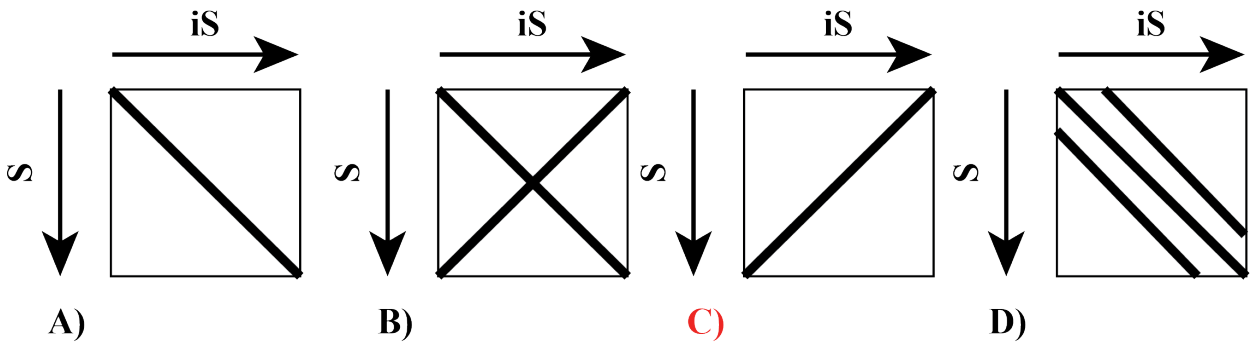
A) Nter-TACGTA-Cter

B) Nter-CGTACG-Cter

C) Nter-CATGCA-Cter

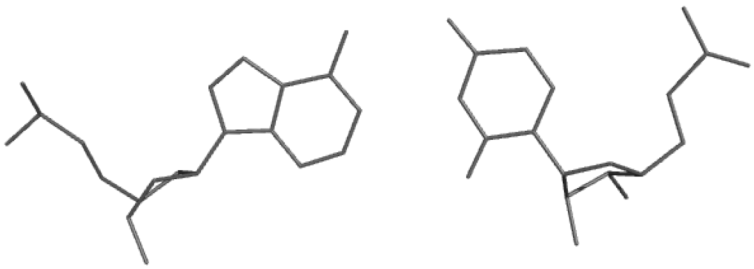
D) Nter-TGCATG-Cter

6) Given two DNA sequences that are each other's inverse (for example 5'-GATCAT-3' and 5'-TACTAG-3'), what does their dotplot look like?



*Name:*_____
ID : _____

7) The figure below shows a nucleotide base pair; identify it (note that dX indicates a deoxyribonucleotide, as contained in a DNA molecule, while rX refers to a ribonucleotide, as found in an RNA molecule). Hydrogen atoms are omitted.



A) dG-dC
B) dG-rC
C) dA-dT
D) dA-rU
E) dG-rU

8) Cytochrome P450 enzymes form a super-family of haem-containing oxygenases that are found in all kingdoms of life. These proteins have very similar structures but show extraordinary diversity in their reaction chemistry. Let us consider these three examples: (A) the human CYP46A1, an enzyme that controls cholesterol turnover in the brain, (B), a human prostacyclin synthase (prostacyclin is a small lipid that inhibits platelet aggregation), and (C), Xpla, a cytochrome P450 from rhodococcus (aerobic bacterium) that has been found to break down explosive pollutants. (A), (B) and (C) are homologous proteins; what else can you say?

A) (A), (B) and (C) are orthologous
B) (A), (B) and (C) are paralogous
C) (A) and (B) are paralogous, while (A) and (C) are orthologous
D) (A) and (B) are orthologous, while (A) and (C) are paralogous
E) (B) and (C) are paralogous, while (A) and (B) are orthologous

9) We want to find the best alignment(s) between the protein sequences WWYCTY and WCYTY. The scoring scheme S is defined as follows: S(i,i) = 10, S(i,j) = 5 if i and j are both aromatic amino acids, and S(i,j) = 0 otherwise. There is a constant gap penalty of 5 (gaps at the beginning are considered, see below). The score Sbest and the number N of optimal alignments are (show your final dynamic programming matrix for full credit):

	W	W	Y	C	T	Y
W	10	5	0	-5	-5	0
C	-5	10	5	15	5	5
Y	0	10	20	5	15	20
T	-5	5	10	20	25	15
Y	0	10	15	15	20	35

A) Sbest = 40, N = 1
B) Sbest = 35, N = 2
C) Sbest = 35, N = 1
D) Sbest = 40, N = 2
E) Sbest = 30, N = 1

*Name:*_____
ID : _____

There is only one optimal alignment (traceback shown in bold):

```
WWYCTY
WCY-TY
```

10) How many DNA coding sequences (where a coding sequence includes the START and STOP codon, but no introns) could lead to the following protein sequence:
Met- Lys-Ser-Trp-Leu-Phe-Trp-Ala, assuming the standard genetic code?

A) 1
B) 576
C) 1152
D) 1728
E) 4096

Count number of codons for each amino acid (3 for stop codons):
1 (Met) x 2 (Lys) x 6 (Ser) x 1 (Trp) x 6 (Leu) x 2 (Phe) x 1 (Trp) x 4 (Ala) x 3 (STOP)= 1728

11) Which combination of program / substitution matrix will most likely give you the best alignment between two sequences that are highly similar?

A) BLAST / Blosum45
B) Dynamic programming / Blosum45
C) BLAST / Blosum90
D) Dynamic programming / Blosum90
E) BLAST / Blosum10

Dynamic programming gives a better alignment than BLAST. As the two sequences are very similar, it is best to use Blosum90.

12) A BLAST search is most useful when you want to do the following:

A) Find inverted repeats within a protein sequence
B) Generate the best possible alignment between the target and template sequences to be used as input for homology modeling
C) Find a rat paralog to a human gene
D) Find a rice ortholog to a yeast gene
E) Predict the secondary structures of a protein

BLAST finds similarity between sequences and does not provide the best alignment. It cannot detect inverted repeats as it is biased toward reading from left to right. Paralog are genes from the same specie. BLAST is not used for secondary structure prediction.

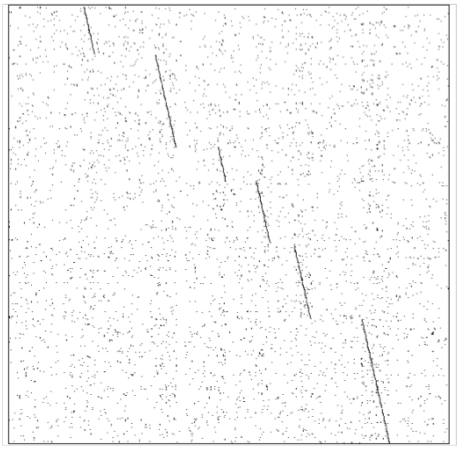
13) We want to find the best alignment(s) between the protein sequences FAFWC and FWFC. The scoring scheme S is defined as follows: S(i,i) = P, and S(i,j) = M otherwise. There is a constant gap penalty of G (gaps at the beginning are considered). The dynamic programming matrix is shown below. What were the values of P, M, and G:

*Name:*_____
ID : _____

	F	A	F	W	C
F	5	-4	3	-4	-4
W	-4	3	1	8	1
F	3	1	8	-1	6
C	-4	1	-1	6	11

A) P=5, M=0, G=0
B) P=5, M=-1, G=-1
C) P=5, M=-2, G=-2
D) P=5, M=-3, G=-1
E) P=5, M=-1, G=-3

14) The dotplot shown below compares the DNA sequence of the actin muscle gene from Pisaster ochraceus (horizontal) with the RNA corresponding to the same gene (vertical). The six regions of high similarity that shows as black lines correspond to:



A) Introns
B) Repeats
C) Inverted repeats
D) Exons
E) All of the above

Conserved regions between RNA and DNA corresponds to coding regions, hence exons.

15) When we compare the sequences of proteins that belong to the same family, we observe that some regions in the sequences are more conserved than others (see for example Figure 1 below).

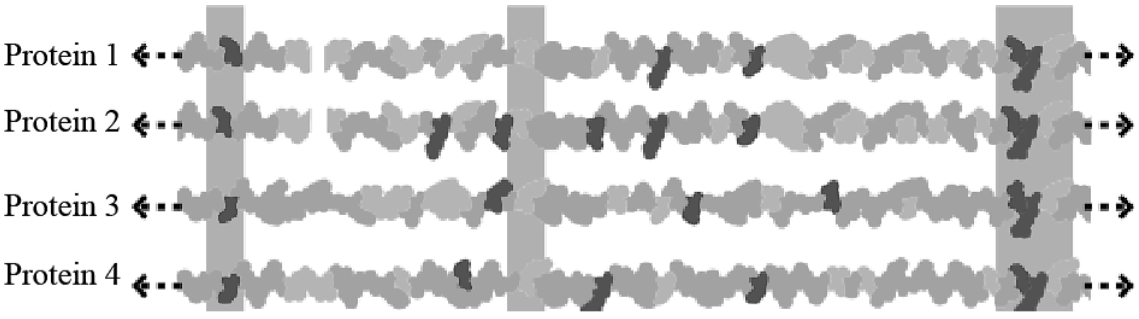


Figure: Residues conserved among various G protein coupled receptors are highlighted in horizontal gray bars (from http://en.wikipedia.org/wiki/Conserved_sequence).

The presence of such conserved regions is (choose the most likely answer):

A) irrelevant: conserved residues are only found at the beginning and end of protein sequences and do not play a role in function.

B) relevant: conserved residues are usually essential for the structure and the function of the protein.

C) irrelevant: the function of a protein is not defined by its sequence (function is only defined by structure)..

D) irrelevant: residue conservation is merely an indirect consequence of the degeneracy and non uniformity of the genetic code, i.e. some residue types are associated with more codons and are therefore less sensitive to mutations at the DNA level.

Part II (2 problems; total 40 points)

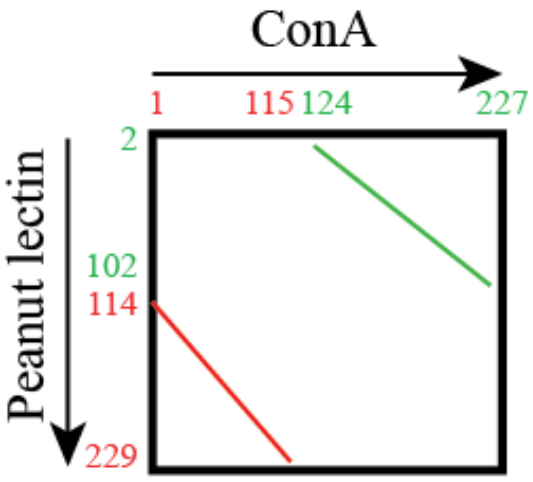
Problem 1 (4 questions, each 5 points: total 20)

Concanavalin A (ConA) is a lectin (carbohydrate-binding protein) originally extracted from the jack bean, *Canavalia ensiformis*. It binds specifically to certain sugars, glycoproteins, and glycolipids. The structure of concanavilin has been determined by X-ray crystallography, and is stored in the PDB. You are interested to know how similar this lectin is from the other lectins that are known, in particular to the lectin from peanut, whose structure is also known. First, you run BLAST, starting from the sequence of ConA. BLAST does find a match with the peanut lectin:

a) BLAST found two alignments between subsets of the sequences of ConA and the peanut lectin. Are these two alignments significant? Justify your answer

The two local alignments have E-values of $3x10^{-22}$ and $1x10^{-16}$, respectively: as such, they are highly significant (the corresponding P-values, i.e. the probabilities that these alignments are random, are $3x10^{-22}$ and $1x10^{-16,}$, respectively).

b) Based on these results from BLAST, draw schematically the dotplot between ConA and the peanut lectin. Only show the major correspondences between the two sequences



*Name:*_____
ID : _____

c) The two local alignments found by BLAST are 116 residues long and 106 residues long, respectively. Based on the specificity of these two alignments and the schematic dotplot you have drawn (from question b), explain why BLAST could not have found a single alignment of length at least 222.

The two alignments are not sequential along the ConA sequence from Nter to Cter: in fact, ConA contains two domains that have been swapped in peanut lectin (a circular permutation). As BLAST (just like most sequence comparison techniques) works with the sequence given from Nter to Cter, it is not designed to detect domain swap.

d) From these results, do you expect the structures of ConA and peanut lectin to be similar? Justify your answer.

As the two alignments are highly significant, it is expected that the two domains of ConA (from residue 1 to 115 and from residue 124 to 227, respectively) be structurally conserved in peanut lectin. We do not know however if the 3D arrangement of these two domains (i.e. how these two domains are packed in the full protein) is conserved.

Problem 2 (4 questions, each 5 points: total 20)

Below is the double-stranded DNA sequence of part of a hypothetical yeast genome, which happens to contain a very small gene.

```
5′ — CTATAAAGAGCCATGCATGAACTAGATAAAAGGCTCTGAGAATTTATCTCTAG— 3′
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||
3′ — GATATTTCTCGGTACGTACTTGATCTATTTTCCGAGACTCTTAAATAGAGATC— 5′
```

a) Which strand of DNA shown, the top or the bottom, is the template strand? Justify your answer

The longest ORF is found on the top strand; this strand is then the coding strand, and the bottom strand is then the template strand.

b) What is the mRNA sequence corresponding to the ORF for the gene?

The longest ORF, found on the top strand is:

5’ ATG CAT GAA CTA GAT AAA AGG CTC TGA-3’

The corresponding RNA sequence is:

5’ AUG CAU GAA CUA GAU AAA AGG CUC UGA-3’

c) What is the sequence of the protein produced from the mRNA in (b)? Label the N and C termini.
The protein sequence is obtained directly using the genetic code:
Nter – Met His Glu Leu Asp Lys Arg Leu – Cter

d) Predict the secondary structure of this protein using the Chou and Fasman method and the table provided in the Appendix. Justify your answer

To predict the secondary structure of this peptide, we use the Chou and Fassman propensities:

	M	H	E	L	D	K	R	L
P(a)	1.47	1.22	1.44	1.30	1.04	1.23	096	1.30
P(b)	0.97	1.08	0.75	1.02	0.72	0.77	0.99	1.02

Helix:

- nucleation sequence: MHELDK
- extension: add R (1.3+1.04+1.23+0.96 =4.53 > 4) and L (1.04+1.23+0.96+1.30=4.53 > 4) on Cter side
- Compute average over 8 residues: 1.245 > 1.0

All 8 residues predicted to be helical.

Strand:

- no nucleation site

The prediction is therefore: HHHHHHHH

Part III (Extra credit; one problem, 8 points)

Below is the double-stranded DNA sequence of part of a hypothetical bacterial genome, which happens to contain a very small gene.

```
5′ — TATAAATTATGTCTGCTATAAAATAACCCGGT— 3′
     ||||||||||||||||||||||||||||||||
3′ — ATATTTAATACAGACGATATTTTATTGGGCCA— 5′
```

a) What is the sequence of the longest protein that can be produced by this DNA sequence? Label the N and C termini.

The top strand sequence S contains one ATG (start codon) with one TAA (stop codon), in phase with the ATG. Consequently, the longest ORF is:

5’ ATG TCT GCT ATA AAA TAA -3’

*Name:*_____
ID : _____

The corresponding RNA sequence is:

5' AUG UCU GCU AUA AAA UAA -3'

The protein sequence is obtained directly using the genetic code:
Nter – Met Ser Ala Ile Lys – Cter

b) Propose a single base pair deletion that will lead to the mutated sequence still coding for a protein, albeit smaller, with the same START codon. Note that you still need a STOP codon in phase with the START codon. Give the sequence of the shorter protein. Label the N and C termini.

One option is to remove the A before the TAAAA in the top sequence:

```
5′ – TATAAATTATGTCTGCTATAAAATAACCCGGT– 3′
     |||||||||||||||||||||||||||||||||
3′ – ATATTTAATACAGACGATATTTTATTGGGCCA– 5′
```

This leads to the new sequence:

```
5′ – TATAAATTATGTCTGCTTAAAATAACCCGGT– 3′
     |||||||||||||||||||||||||||||||
3′ – ATATTTAATACAGACGAATTTTATTGGGCCA– 5′
```

This sequence has a new TAA, in phase with the original START codon ATG. This leads to a new longest ORF is:

5' ATG TCT GCT TAA -3'

The corresponding RNA sequence is:

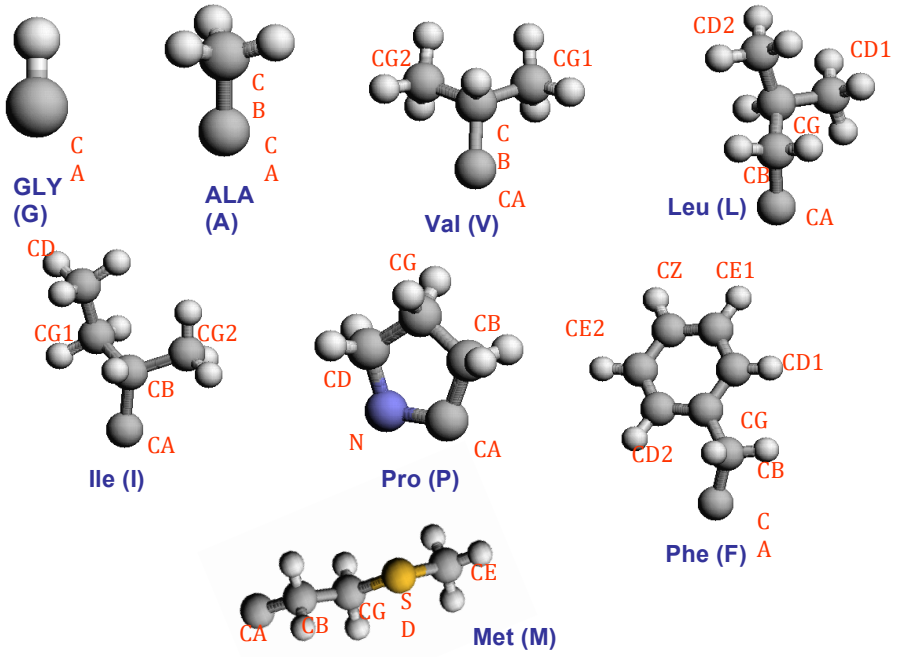
5' AUG UCU GCU UAA -3'

The protein sequence is obtained directly using the genetic code:
Nter – Met Ser Ala– Cter

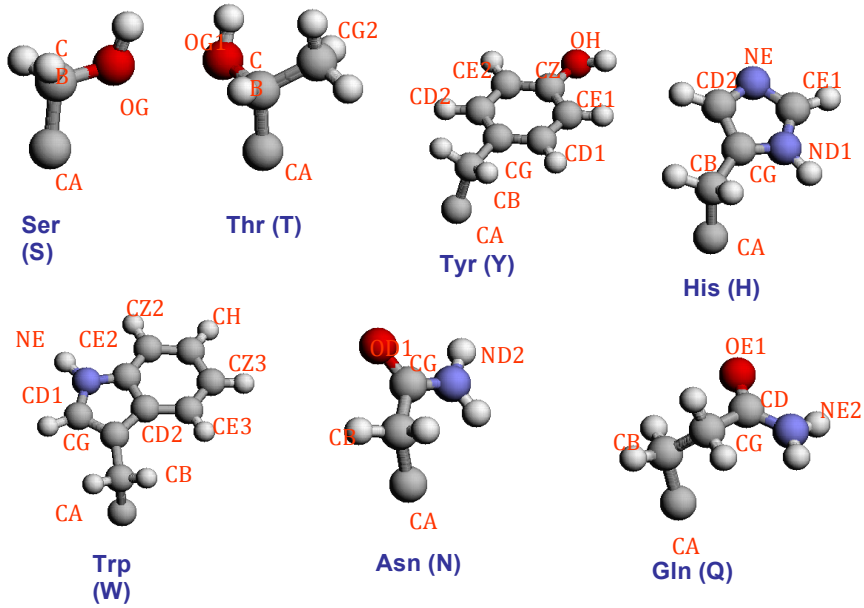
*Name:*_____
ID : _____

Appendix A: Amino Acids

Hydrophobic Amino Acids

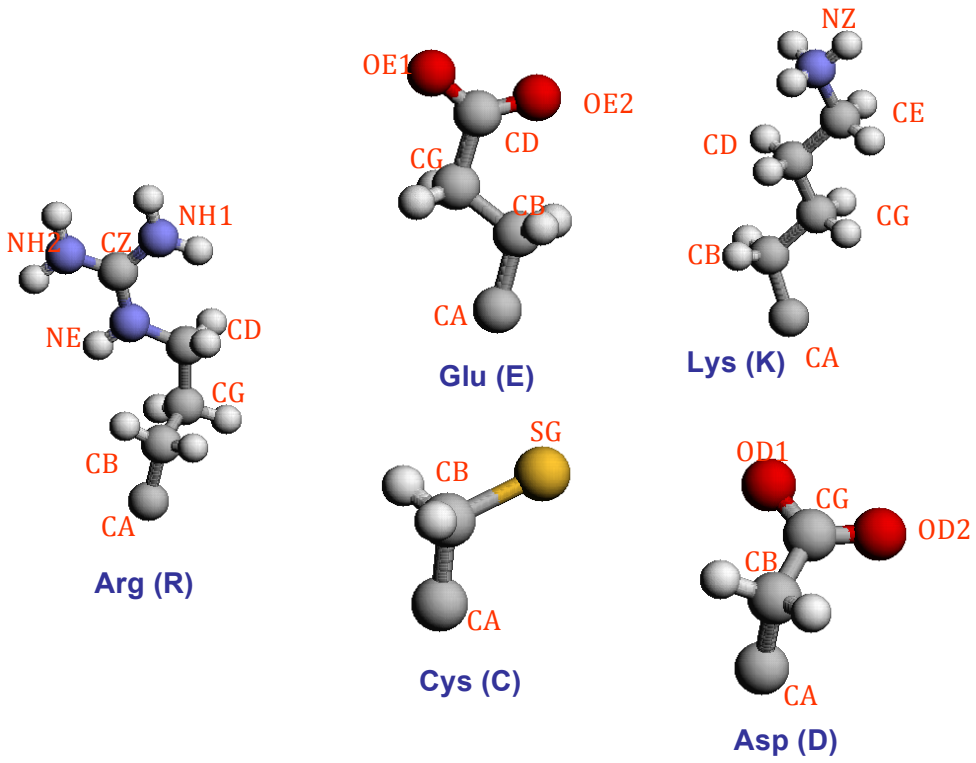


Polar Amino Acid

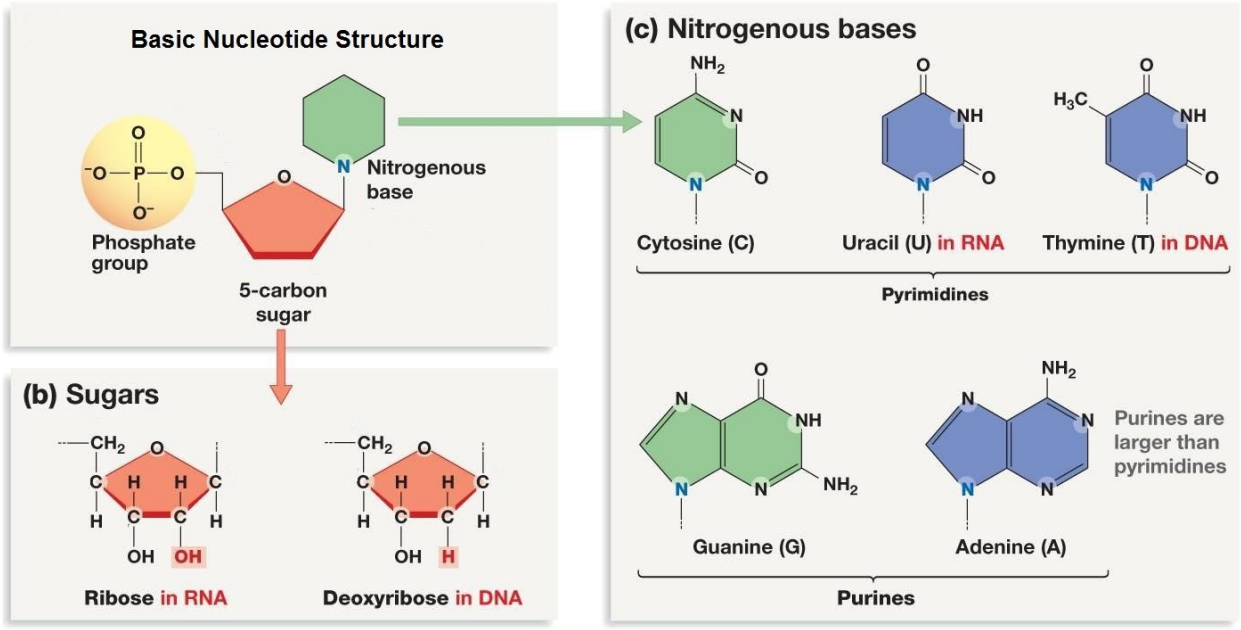


*Name:*_____
ID : _____

Polar Amino Acids



Appendix B: Nucleotides



*Name:*_____
ID : _____

Appendix C: Genetic Code

	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/Start	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Appendix D: Chou and Fassman Propensities

Amino Acid	**Helix**	**Strand**	**Turn**
Ala	1.29	0.90	0.78
Cys	1.11	0.74	0.80
Leu	1.30	1.02	0.59
Met	1.47	0.97	0.39
Glu	1.44	0.75	1.00
Gln	1.27	0.80	0.97
His	1.22	1.08	0.69
Lys	1.23	0.77	0.96
Val	0.91	1.49	0.47
Ile	0.97	1.45	0.51
Phe	1.07	1.32	0.58
Tyr	0.72	1.25	1.05
Trp	0.99	1.14	0.75
Thr	0.82	1.21	1.03
Gly	0.56	0.92	1.64
Ser	0.82	0.95	1.33
Asp	1.04	0.72	1.41
Asn	0.90	0.76	1.23
Pro	0.52	0.64	1.91
Arg	0.96	0.99	0.88