BITS, PILANI- DUBAI
DUBAI INTERNATIONAL ACADEMIC CITY
SECOND SEMESTER 2010-2011
COMPREHENSIVE EXAM

COURSE NO.: EA C414 **05.06.11** **MAXIMUM MARKS: 40**
COURSE TITLE: Introduction to Bioinformatics **DURATION: 3 Hours**
Answer Part A and Part B in separate answer sheets
Answer to the point; Answer all questions in the given sequence
Marks will be allotted for step-wise solution to problems

PART - A

Q1. (a) Define the major bonds involved in a nucleic acid molecule. [1]

(b) Classify proteins functionally with one example for each classification. [1]

(c) Why does polymerization take place only in the 5' – 3' direction? Explain schematically. [No theory required] [2]

(d) Depict the DNA Replication process pictorially. Ensure that each step is indicated clearly and also mark the relevant enzymes. [3]

(e) Define (i) NMR and (ii) DNA Microarray [2]

Q2. (a) Write a short note on X-Ray Crystallography. [2]

(b) State the significance of the 'Ramachandran Plot'. [1]

(c) Map the circular plasmid from the fragments shown below: [2]

DNA	**Size of Fragments (bp)**
EcoRI	4.36 kb
Sall	4.36 kb
Pstl	4.36 kb
EcoRI + Sall	3.98 kb, 0.38 kb
EcoRI + Pstl	3.61 kb, 0.75 kb
Pstl + Sall	3.23 kb, 1.13 kb

(d) Define the three possible models of DNA Replication. [1]

P.T.O

PART – B

Q1. a) Two tables describing the properties of amino acids are shown below.

amino_acid_tbl :

amino_acid	three_let_code	one_let_code	volume	surface_area	distal_grp

distal_group_tbl:

distal_group	h_bond_donor	h_bond_acceptor

Write the SQL statement for the following queries.

i) List the all amino acids, its three_let_code, and volume which has surface_area > 200 and distal_group = "Phenyl" [1]

ii) List all the amino acids, its three_let_code, and its h-bond properties, which belong to the distal_group = "Carboxyl" [2]

b) What are the advantages of storing biological data using XML format as compared to the relational database format? [2]

Q2. a) In the NW method of pairwise alignment, after the score in a square is calculated, a backward arrow to the one of the squares (left, diagonal or top) is placed. How is this determined and what is its significance. [3]

b) Explain what do you mean by an additive tree, with the help of a tree with 2 organisms and 1 common ancestor? [2]

Q3. (a) Given the 5 profile sequences.

TCAAGC, AGTAGC, TACTCG, TGTTCC, CGCTGG

And the query sequence

A T T T A G T A T C A A T G A T A A C A A T T C,

Find the alignment score of the profile pattern at the 10^{th} position of the query sequence, using PSSM method. [5]

(b) Explain the working of PSI-BLAST with suitable diagram and flow chart. [3]

Q4. a) The distance matrix for 5 species A, B, C, D, and E is given below. Obtain the phylogenetic tree using the UPGMA method. [5]

	A	B	C	D	E
A		19	22	30	23
B			28	33	17
C				30	40
D					43
E					

(b) Explain the maximum likelihood method of phylogenetic tree generation. [2]

BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai

IV Year (BIOTECH)
Second Semester, 2010-2011

Test 2 (Open Book)

Course No: EA C414 | **Course Title: Introduction to Bioinformatics**
Date: 17th Apr 2011 | **Weightage: 20%**
Duration: 50 minutes | **Max. Marks. 20**

1) For the following sequences find. **[4M]**
   a) Hamming Distance
      Seq 1: CATGA Seq 1: GGATTA
      Seq 2: CGATA Seq 2: CGGACT
   b) Levenshtein edit distance
      Seq 1: A_ _C T_ Seq 1: GT_GA
      Seq 2: AGTCAT Seq 2: _ TTCA

2) Using the dotmatrix method perform a pairwise alignment of the following two sequences. Assuming +1 for match and -1 for mismatch, find out the alignment score.
   Horizontal Seq: TAGCAGTCA
   Vertical Seq : TTAGGA **[5M]**

3) Using Needleman-Wunsch algorithm for global alignment, find out the optimal alignment for the pair of amino acid sequence shown below. Use BLOSUM62 (shown on pg 2) substitution matrix, with a gap penalty of -8. **[5M]**
   Horizontal Seq: HEAGAWE
   Vertical Seq: PAWE

4) Show how a dot plot diagram would look for the alignment of two sequences. **[6M]**
   a) Direct repeated residues in both sequences.
   b) Inverted repeats of residues in both sequences.
   c) A sequence is aligned with itself

P.T.O

BLOSUM62 MATRIX

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

BITS, PILANI- DUBAI
DUBAI INTERNATIONAL ACADEMIC CITY
SECOND SEMESTER 2010-2011
TEST – I (CLOSED BOOK)

COURSE NO.: EA C414 **27.02.11** **MAXIMUM MARKS: 25**
COURSE TITLE: Introduction to Bioinformatics **DURATION: 50 Minutes**
Answer to the point; Answer all questions in the given sequence

Q1. (a) DNA Coding Strand: 5'-**CCTGATGAGGAAAGGCTGACATTACAT** -3' [5]
(a) Write the mRNA formed after transcription
(b) Write the polypeptide formed after translation of the mRNA.
(c) Write the possible tRNA anti-codons for the mRNA sequence.
(d) If the sequence was 5'-**CCTGATGATGAAAGGCTGACATTACAT** -3', how would your answer change?
(e) Define termination sequences.

[P.T.O FOR AMINO ACID – NUCLEIC ACID DICTIONARY]

(b) Depict the translation process pictorially. [3]

(c) Name any two DNA families and mention one characteristic feature of each family. [2]

Q2. (a) Depict a dinucleotide and indicate the major bonds in the molecule. [1]

(b) You have a piece of DNA but you do not know the sequence of the DNA. Using Sanger method, you determine the sequence as: 3'-**CTTGAACGA** -5'
Using the steps of Sanger procedure, explain how this answer has been obtained. [4]

(c) What did M. Meselson and F. Stahl do? [3]

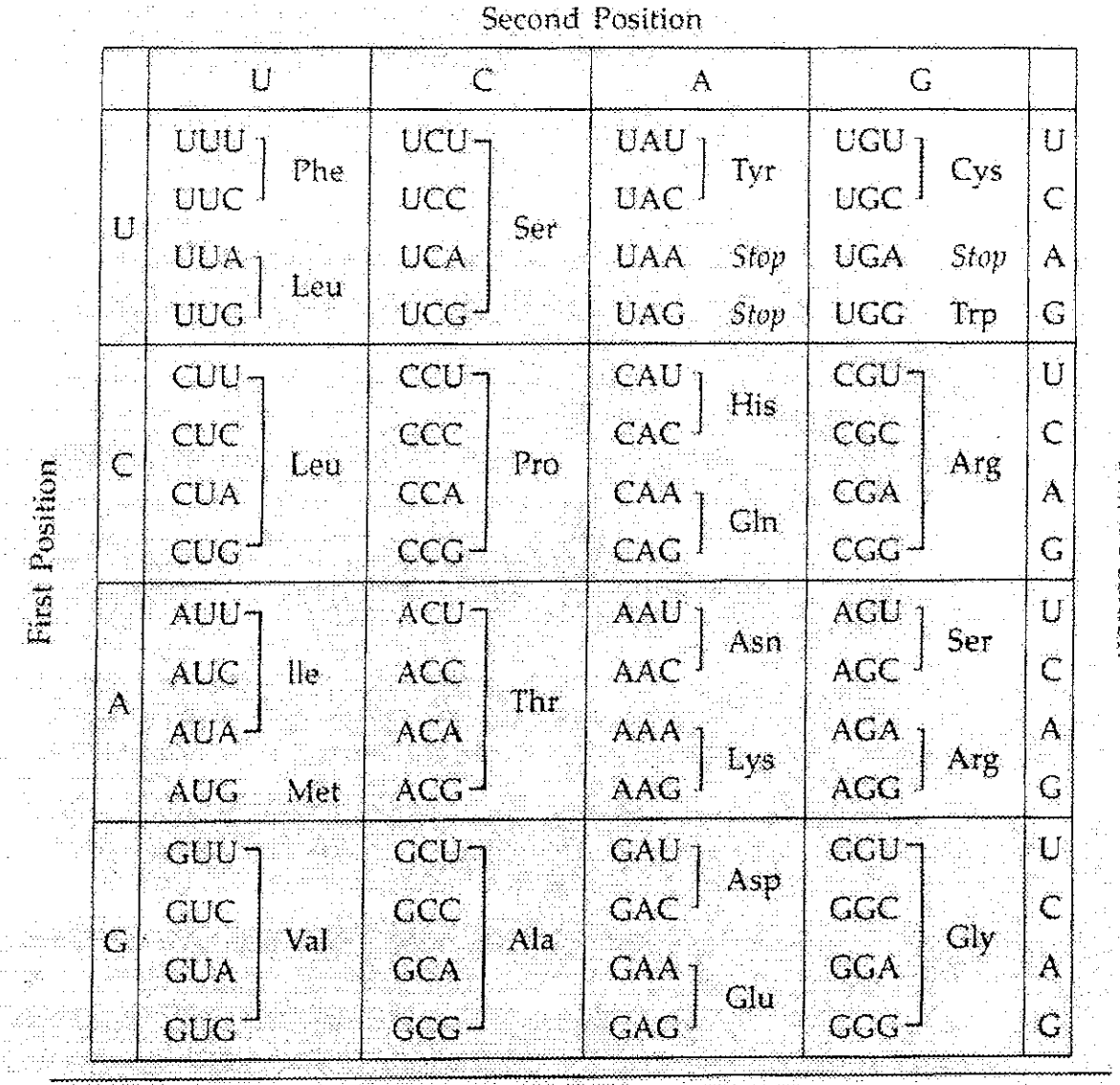
(d) Name the steps in lagging strand synthesis and also name the major enzyme involved in each step. [2]

Q3. (a) Name two common methods used for DNA Amplification. [1]

(b) Name two limitations in inferring the amino acid sequence from gene sequencing. [1]

(c) How are parallel and anti-parallel β pleated sheets different? [1]

(d) There is a strong electrostatic force of attraction between DNA and the histone proteins. Why is this bond so strong? [2]


Second Position
First Position
Third Position
U
C
A
G
UUU Phe
UUC
UUA Leu
UUG
UCU Ser
UCC
UCA
UCG
UAU Tyr
UAC
UAA Stop
UAG Stop
UGU Cys
UGC
UGA Stop
UGG Trp
CUU Leu
CUC
CUA
CUG
CCU Pro
CCC
CCA
CCG
CAU His
CAC
CAA Gln
CAG
CGU Arg
CGC
CGA
CGG
AUU Ile
AUC
AUA
AUG Met
ACU Thr
ACC
ACA
ACG
AAU Asn
AAC
AAA Lys
AAG
AGU Ser
AGC
AGA Arg
AGG
GUU Val
GUC
GUA
GUG
GCU Ala
GCC
GCA
GCG
GAU Asp
GAC
GAA Glu
GAG
GGU Gly
GGC
GGA
GGG

BITS PILANI, DUBAI CAMPUS

SECOND SEMESTER 2010 – 2011

Course Code: EA C414 **FOURH YEAR (BIOTECH)** Date: 02.05.11

Course Title: Introduction to Bioinformatics **QUIZ 2** Max Marks: 07

Duration : 20 minutes Weightage: 7%

Name: . . . . . ID No: . . . . . Discipline: . . . .

1. For the distance matrix shown for 6 organisms, A, B, C, D, E, and F using the Neighbour Joining algorithm,

	A	B	C	D	E	F
A		10	32	5	25	43
B			29	14	51	21
C				4	18	52
D					33	27
E						16
F						

For X=D, and Y=E,

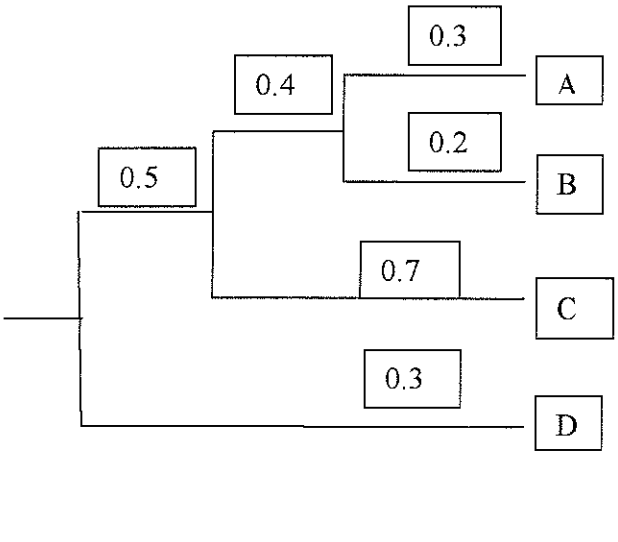
a) Draw the tree considering that D and E are closely matched organisms. **[1M]**

b) Show the expressions for $S_{XY,}$ x **[2M]**

c) Find the values for $S_{XY,}$ x **[2M]**

P.T.O

2. For the tree shown below, calculate the weighting factor for each of the organisms, A, B, C and D . **[2M]**



BITS, PILANI – DUBAI

SECOND SEMESTER 2010 – 2011

FOURH YEAR (BIOTECH)

QUIZ 1

Course Code: EA C414
Course Title: Introduction to Bioinformatics
Duration : 20 minutes

Date: 21.03.11
Max Marks: 08
Weightage: 8%

Name: . . . . . ID No: . . . . . Discipline: . . . . . .

1. Construct a restriction map of a circular DNA plasmid, using the following data. Your map should indicate the relative positions of the restriction sites along with distances between restriction sites: [3]

DNA	**Sizes of Fragments (bp)**
uncut DNA	7950
DNA cut with BglII	7950
DNA cut with EcoRI	7950
DNA cut with HpaI	7950
DNA cut with BglII + EcoRI	5416, 2534
DNA cut with BglII + HpaI	6632, 1318
DNA cut with EcoRI + HpaI	4098, 3852

P.T.O

2. What does the Bragg's law state? [0.5]

3. The resonance frequency of a particular substance is directly proportional to _____ [0.5]

4. The basic unit of data storage in relational data bases is called _____ [1]

5. The general format of an item in the FEATURE TABLE (FT) component in the EMBL Nucleotide Database file is [3]

BITS, PILANI- DUBAI
DUBAI INTERNATIONAL ACADEMIC CITY
SECOND SEMESTER 2010-2011
COMPREHENSIVE EXAM -- ANSWER KEY

COURSE NO.: EA C414 **05.06.11** **MAXIMUM MARKS: 40**
COURSE TITLE: Introduction to Bioinformatics **DURATION: 3 Hours**

PART - A

Q1. (a) Phosphodiester Bonds – Strong covalent bonds between a phosphate group and two carbon ring pentoses over two ester bonds
N Glycosidic Bond – Joins a carbohydrate molecule to another group, which may / may not be another carbohydrate. [1]

(b) Structural [Ex: Collagen], Carrier [Ex: Lipoproteins] and Regulator [Ex: Enzymes / Hormones] [1]

(c) Refer class notes. Show two figures – depicting polymerization in both directions and why one is advantageous in proof reading. [2]

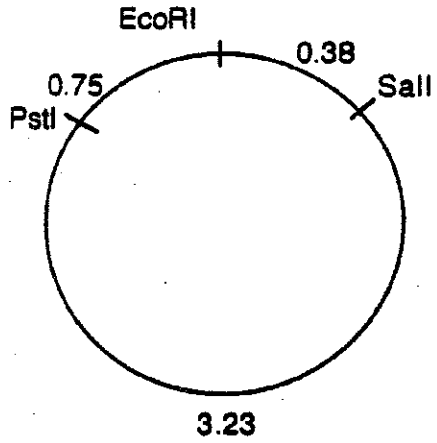
(d) Refer class notes or Pages 227 – 229 of Reference Book: "Principles of Genetics". [3]

(e) (i) NMR – Property that magnetic nuclei have in a magnetic field and applied electromagnetic pulse which cause the nuclei to absorb energy from EM pulse and radiate the energy back out, which is at a specific resonance frequency. This allows for the observation of specific quantum mechanical magnetic properties of an atomic nucleus.
(ii) DNA Microarray -- A technology which involves a collection of DNA spots [probes] attached to a solid surface which enables running of several genetic tests in parallel. [2]

Q2. (a) Refer class notes. [2]

(b) Rotation is permitted around the N-Cα and Cα-C single bonds in a protein structure – The angles φ and ψ around these bonds and the angle of rotation around the peptide bond define the conformation of a residue. The sequence of these angles in a protein defines the backbone conformation. The allowed ranges of φ and ψ for angle of rotation = 180° fall into defined regions in a graph commonly known as the Ramachandran plot. The allowed regions generate standard conformations. [1]

(c) Map the circular plasmid from the fragments shown below: [2]



(d) Conservative – Whole original double helix acts as the template for a new one, therefore one daughter molecule would consist of the original parental DNA and the other daughter would be totally new DNA
Semi-Conservative – Every daughter DNA molecule has an intact template strand and a newly replicated strand.
Dispersive – Some parts of the original double helix are conserved and some parts are not. [1]

PART – B

Q1. a)
i) List the all amino acids, its three_let_code, and volume which has surface_area > 200 and distal_group = "Phenyl" [1]

```
SELECT  amino_acid, three_let_code, volume
FROM  amino_acid_tbl
WHERE ((surface_area >200) AND (distal_group = "Phenyl"))
```

ii) List all the amino acids, its three_let_code, and its h-bond properties, which belong to the distal_group = "Carboxyl" [2]

```
SELECT amino_acid_tbl.amino_acid, amino_acid_tbl.three_let_code,
        distal_group_tbl.h_bond_donor, distal_group_tbl.h_bond_acceptor
FROM amino_acid_tbl, distal_group_tbl
WHERE ((amino_acid_tbl.distal_group=distal_group_tbl.distal_group) AND
        (amino_acid_tbl.distal_group=" Carboxyl"))
```

b) Advantages of XML [2]
semi-structured data (where the number of attributes varies and the type of attributes) can be easily represented.
Hierarchical information can be easily represented.

Q2. a) Score in a square is calculated as the best scores in left, diagonal or top + new alignment score/gap. A backward arrows from the current square to that square which yielded the best score. Significance is that it is guaranteed to give the best possible alignment and the backward arrow can be used to trace the path that produced the best alignment. [3]

b) The genetic distance between 2 organisms is the number of mismatches in the pair-wise alignment of the (DNA) sequences. By the property of additive tree, we can say that the genetic distance between the organisms can be expressed as the sum of the genetic distance between the organisms and their ancestor. [2]

Q3. (a)

Table 1: Given profile seq

Position → / Seq ↓	1	2	3	4	5	6
1	T	C	A	A	G	C
2	A	G	T	A	G	C
3	T	A	C	T	C	G
4	T	G	T	T	C	C
5	C	G	C	T	G	G

Table 2: Freq of occurrences of A C G and T

Position →	1	2	3	4	5	6
A	1	1	1	2	0	0
C	1	1	2	0	2	3
G	0	3	0	0	3	2
T	3	0	2	3	0	0

Table 3: Relative Freq of occurrences of A C G and T

Position →	1	2	3	4	5	6
A	0.2	0.2	0.2	0.4		0
C	0.2	0.2	0.4	0	0.4	0.6
G	0	1	0	0	0.6	0.4
T	0.6	0	0.4	0.6	0	0

Table 4: Alignment score

Query →	C	A	A	T	G	A
Position Seq	1	2	3	4	5	6
A		0,2	0,2			0
C	0.2					
G					0.6	
T				0.6		

Total alignment score = 0.2 + 0.2 + 0.2 + 0.6 + 0.6 + 0 = 1.8 **[1+1+1+1+1]**

(b) Explain the working of PSI-BLAST with suitable diagram and flow chart. [3]

A flowchart for PSI-BLAST (figure 5.7 in TB 1)

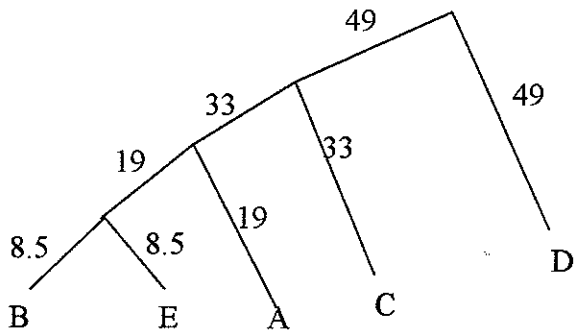
1. Probe each sequence in the chosen database independently for local regions of similarity to the query sequence, using a BLAST-type search but allowing gaps.
2. Collect significant hits. Construct a multiple sequence alignment table between the query sequence and the significant local matches.
3. Form a profile from the multiple sequence alignment.
4. Reprobe the database with the profile, still looking only for local matches.
5. Decide which hits are statistically significant and retain these only.
6. Go back to step 2, until a cycle produces no change. This accounts for the 'Iterated' in the program title.

Q4. a)

	A	BE	C	D
A		(19+23)/2 =21	22	30
BE			(28+40)/2 =34	(33+43)/2 =38
C				30
D				

	ABE	C	D
ABE		(22+34)/2 =28	(30+38)/2 =34
C			30
D			

	ABEC	D
ABEC		(34+30)/2 =32
D		



[5]

(b) Explain the maximum likelihood method of phylogenetic tree generation. [2]

- Assigns quantitative probabilities to mutations. Better that just counting them,
-reconstructs tree, (finds ancestor). Branch length – probalities.
- different trees are tried which maximizes the likelihood of observed seq.