**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**II SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**COMPREHENSIVE EXAMINATION (CLOSED BOOK)**

**Duration: 3h** **Date: 5.6.2013**
**Weightage: 40%** **Max. Marks: 40**

*Note: a) answer all the questions, b) answer to the point and c) draw schematic diagrams if required.*

## PART-A

1. What are observables and data achives? Explain with suitable examples. [2.0]
2. What are the applications of SINES and LINES? Briefly outline. [2.0]
3. Why the secondary and tertiary protein structure analysis is important in bioinformatics? Explain with respect to the levels of organization and their role in protein stability. Mention the applications of protein structure prediction in medical biotechnology. [3.0]
4. Why model organisms are important in the analysis of genomes and proteomes? Explain with one example each for prokaryotic and eukaryotic systems in bioinformatics. [3.0]
5. What are the most common programming languages and tools used in bioinformatics? Explain with suitable examples. [2.0]
6. What is gene ontology? Mention how gene ontology it is organized and give any two applications. [2.0]
7. Write a short note on the following: [2.0]
   a. EcoCYC
   b. BioCYC
   c. KEGG
   d. PDB
8. How the protein stability and folding patterns are validated? Explain with respect to bond angles, secondary structure stability testing and validation methods and the role of side chains in protein stability. [3.0]
9. What is the role of hydrophobicity profiling in the determination protein structure prediction? Explain with schematic diagram for a cytosolic and membrane protein with seven transmembrane domains. [2.0]
10. What are the applications of mass spectrometry, DNA microarrays, NMR and X-ray crystallography in bioinformatics? Mention any two points for each. [2.0]

## PART-B

11) For the following sequences find. [3M]

a) Hamming Distance b) Edit Distance

Seq 1: ATATATAT
Seq 2: TATATATA

12) 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: ATTGCCCATG
Vertical Seq : ATGGCCATTG [3M]

13) 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. [4M]
Horizontal Seq: G A A T T C A G T T A
Vertical Seq: G G A T C G A

14) Build the Hidden Markov Model (HMM) and derive the probabilities for the given sequences: [3M]

A C A G
T C A T
A G A C
A G _ T
A C _ _

15) Generate a Position Specific Scoring Matrix from the profile [4M]

NTEGEWI
NITRGEW
NIAGECC

(i) Query Sequence:
NTEGWIHRACCAGGAGC

Generate a PSSM matrix from the profile and calculate the alignment score of the profile. Sequence starting at position1.

**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 | |

Ans key.

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**II SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**COMPREHENSIVE EXAMINATION (CLOSED BOOK)**

**Duration: 3h** **Date: 5.6.2013**
**Weightage: 40%** **Max. Marks: 40**

*Note: a) answer all the questions, b) answer to the point and c) draw schematic diagrams if required.*

## PART-A

1. What are observables and data achives? Explain with suitable examples. [2.0]
   Dna protein sequences, annotations, databanks, metabolic pathways, ncbi, pdb, swissprot.
2. What are the applications of SINES and LINES? Briefly outline. [2.0]
   Short interspersed nuclear elements, long interspersed nuclear elements, 70-500bp; 7000bp; phylogenetic analysis, evolutionary links.
3. Why the secondary and tertiary protein structure analysis is important in bioinformatics? Explain with respect to the levels of organization and their role in protein stability. Mention the applications of protein structure prediction in medical biotechnology. [3.0]
   Secondary structure, domains, tertiary structure, protein interaction, stabilizing forces, clinical-drug design application, parkinsons disease, alzhimers disease, drug targeting.
4. Why model organisms are important in the analysis of genomes and proteomes? Explain with one example each for prokaryotic and eukaryotic systems in bioinformatics. [3.0]
   E. coli, yeast, dna, protein sequences, comparison, data mining applications.
5. What are the most common programming languages and tools used in bioinformatics? Explain with suitable examples. [2.0]
   Perl, java, c, fortran, programming languages, ncbi server applications.
6. What is gene ontology? Mention how gene ontology it is organized and give any two applications. [2.0]
   To produce standardized scheme for describing function, molecular, biological, cellular components, metabolic pathways.
7. Write a short note on the following: [2.0]
   a. EcoCYC
      EcoCYC-E. coli.genes, genomics, proteins
   b. BioCYC
      Biological, genes, genomics, proteins
   c. KEGG
      KEGG multiple organisms, genes, genomics, proteins
   d. PDB
      Protein structure, databases
8. How the protein stability and folding patterns are validated? Explain with respect to bond angles, secondary structure stability testing and validation methods and the role of side chains in protein stability. [3.0]
   Secondary structure, plot, alpha, beta regions, side chains, size, charge, poloarty, shape and rigidity, hydrophobicity,
9. What is the role of hydrophobicity profiling in the determination protein structure prediction? Explain with schematic diagram for a cytosolic and membrane protein with seven transmembrane domains. [2.0]
   Aromatic amino acids, side chains, membrane proteins, graph.

10. What are the applications of mass spectrometry, DNA microarrays, NMR and X-ray crystallography in bioinformatics? Mention any two points for each. [2.0]
Protein identification, aminoacid sequence determination; identifying genes, transcriptomics, highthroughput analysis, structure determination, drug target and active site analysis.

## PART-B

11) For the following sequences find. [3M]
a) Hamming Distance b) Edit Distance
Seq 1: ATATATAT
Seq 2: TATATATA
Hamming, alignment, refer class notes

12) 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: ATTGCCCATG
Vertical Seq : ATGGCCATTG [3M]
Dot matrix, refer class notes

13) 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. [4M]
Horizontal Seq: G A A T T C A G T T A
Vertical Seq: G G A T C G A
refer class notes

14) Build the Hidden Markov Model (HMM) and derive the probabilities for the given sequences: [3M]

A C A G
T C A T
A G A C
A G _ T
A C _ _

refer class notes

15) Generate a Position Specific Scoring Matrix from the profile [4M]

NTEGEWI
NITRGEW
NIAGECC

(i) Query Sequence:
NTEGWIHRACCAGGAGC
Generate a PSSM matrix from the profile and calculate the alignment score of the profile. Sequence starting at position1.
refer class notes

## 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 | 6 | | | | | | | | | 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 CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**TEST-II (OPEN BOOK)**

**Duration: 50 min.** **Date: 5.5.2013** **Max. Marks: 20**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1. Why energy minimization is required in the ab initio protein structure prediction? Compare with the homology modeling and explain energy minimization. [3.0]

2. The protein stability is important parameter for drug discovery applications for finding suitable inhibitors or analogs for pharmaceutical applications, How will you approach to solve the problems of bioinformatics based protein structure prediction for protein stability and folding of α-helices, β-sheets and 3D conformation of proteins in terms biophysical data, bond angles, entropy and enthalpy? [4.0]

3. How will you identify the membrane proteins and soluble cytosolic proteins with respect to aminoacid composition and hydrophobicity profile? Explain with suitable examples. [3.0]

4. Generate a Position Specific Scoring Matrix from the profile

   (i)
   AGATGGATGG
   TGATTGATGT
   TGATGGATGG
   AGATTGATCG
   TGATGGATTG
   TGATGGATTG
   AGATGGATTG

   (ii) Query Sequence:
   ACTCAGCCCCAGCGGAGGCGAAGGACGTCCTTCCCCAGGAGC

   Generate a PSSM matrix from the profile and calculate the alignment score of the profile with 10 nucleotides of the query sequence starting at position 19. [5.0]

5. Write a PERL program to concatenate two DNA sequences without using dot operator and transcript it to RNA and print the output in uppercase. [5.0]
   Seq1: GAGTGAGGG
   Seq2: GAGCAGTTG

Ans key

**BITS PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**TEST-II (OPEN BOOK)**

**Duration: 50 min.** **Date: 5.5.2013** **Max. Marks: 20**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1. Why energy minimization is required in the ab initio protein structure prediction? Compare with the homology modeling and explain energy minimization. [3.0]
   It is called as *ab initio* method and is used when there is no useful template is available. *Ab initio* prediction relies on the thermodynamic hypothesis of protein folding. The thermodynamic hypothesis suggests that the native structure of a protein sequence corresponds to its global free energy minimum state. Accordingly, the *ab initio* prediction methods are generally formulated as optimizations. As such, they can be distinguished by the representation of a protein and its degrees of freedom, the function that defines the energy for each of the allowed conformations, and the optimization method that attempts to find the global minimum on a given energy surface.

2. The protein stability is important parameter for drug discovery applications for finding suitable inhibitors or analogs for pharmaceutical applications, How will you approach to solve the problems of bioinformatics based protein structure prediction for protein stability and folding of α-helices, β-sheets and 3D conformation of proteins in terms biophysical data, bond angles, entropy and enthalpy? [4.0]
   The mutations affect the protein function, mostly favorably during evolution; preliminary analysis of peptide stabilily, biophysical data on bond angles, amino acid composition, primary and secondary structure analysis. Clustering of residues, determination of bond angles, side chains,

3. How will you identify the membrane proteins and soluble cytosolic proteins with respect to aminoacid composition and hydrophobicity profile? Explain with suitable examples. [3.0]
   Hydrophobic aminoacids and transmembrane domains presence and in the case of soluble proteins higher percentage of charged aminoacids.

4. Generate a Position Specific Scoring Matrix from the profile

   (i) AGATGGATGG
   TGATTGATGT
   TGATGGATGG
   AGATTGATCG
   TGATGGATTG
   TGATGGATTG
   AGATGGATTG
   refer class notes on position specific scoring matrix, profile

   (ii) Query Sequence:
   ACTCAGCCCCAGCGGAGGCGAAGGACGTCCTTCCCCAGGAGC

Generate a PSSM matrix from the profile and calculate the alignment score of the profile with 10 nucleotides of the query sequence starting at position 19. [5.0]
Refer class notes on PSSM matrix, alignment scores.

5. Write a PERL program to concatenate two DNA sequences without using dot operator and transcript it to RNA and print the output in uppercase. [5.0]
   Seq1: GAGTGAGGG
   Seq2: GAGCAGTTG
   class notes on bioperl, sequence, program output formats.

**BITS PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**TEST-I (CLOSED BOOK)**

**Duration: 50 min.** **Date: 18.3.2013** **Max. Marks: 25**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1. Write a brief note on observables and data archives. How data archives help in bioinformatics? Explain with suitable examples. [2.0]
2. How advances in computing hardware and software help in bioinformatics? Explain with any three examples. [1.0]
3. What is FASTA format for protein and DNA sequence and why it is required? [1.0]
4. How phylogenetic relationships are analyzed using bioinformatics softwares? Why this analysis is required? [2.0]
5. Define SINES and LINES and mention their applications. [2.0]
6. What are the different levels of protein structure prediction? Briefly outline a method. [2.0]
7. What are model organisms and mention any one from prokaryotes and eukaryotes with their genome and protein properties? [2.0]
8. What are the applications of identification of genes associated with inherited diseases?
9. How evolutions help in the analysis of evolution of genomes? Briefly explain. [1.5]
10. What is metagenomics and mention their applications? [1.5]

11. For the following sequences find. [2.0]

a) Hamming Distance
Seq 1: AGCAA
Seq 2: ACATA
b) Levenshtein edit distance
Seq 1: AGCACAC_A
Seq 2: A_CACACTA

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. [3M]
Horizontal Seq: CGTTAGA
Vertical Seq : CGTAC

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: AAGT
Vertical Seq: AT

# 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 | |

Ans. key

**BITS PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**TEST-I (CLOSED BOOK)**

**Duration: 50 min.** **Date: 18.3.2013** **Max. Marks: 25**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1. Write a brief note on observables and data archives. How data archives help in bioinformatics? Explain with suitable examples. [2.0]
   Databases, protein, DNA, RNA, protein sequences, structure determination, alignment
2. How advances in computing hardware and software help in bioinformatics? Explain with any three examples. [1.0]
   Prediction software, data mining, cluster analysis
3. What is FASTA format for protein and DNA sequence and why it is required? [1.0]
   Systematic search algorithm for database, protein, DNA sequence
4. How phylogenetic relationships are analyzed using bioinformatics softwares? Why this analysis is required? [2.0]
   Evolutionary relationship analysis, phylip, NCBI
5. Define SINES and LINES and mention their applications. [2.0]
   Short and long signature sequences, genome analysis
6. What are the different levels of protein structure prediction? Briefly outline a method. [2.0]
   Homology, secondary, tertiary, energy minimization, protein folding
7. What are model organisms and mention any one from prokaryotes and eukaryotes with their genome and protein properties? [2.0]
   E. coli, yeast, model system for genome proteome analysis
8. What are the applications of identification of genes associated with inherited diseases?
9. How evolutions help in the analysis of evolution of genomes? Briefly explain. [1.5]
   Genetic analysis, DNA protein mutations, structure function
10. What is metagenomics and mention their applications? [1.5]
    Environmental sample analysis, genome, identification
11. For the following sequences find. [2.0]

a) Hamming Distance
   Seq 1: AGCAA
   Seq 2: ACATA
   refer. text book

b) Levenshtein edit distance
   Seq 1: AGCACAC_A
   Seq 2: A_CACACTA
   refer. text book

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. [3M]
   Horizontal Seq: CGTTAGA
   Vertical Seq : CGTAC
   refer. text book

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: AAGT

Vertical Seq: AT

refer. text book

**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 CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**QUIZ-I (CLOSED BOOK)**

**Duration: 20 min.** **Date: 3.4.2013** **Max. Marks: 8**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1) Consider four species characterized by homologous sequences ACTTC, AGTTC, CGTAC and CCTAC. Taking the number of differences as the measure of dissimilarity between each pair of species, use a simple clustering procedure to derive a phylogenetic tree. [2.0]

2) What are archives and give any two examples. [1.5]

3) What are the applications of NMR and X-ray crystallography in Bioinformatics? [1.5]

4) What is OMIM and mention the applications? [1.5]

5) What are protein information resources (PIR) give any two examples? [1.5]

Ans. key.

**BITS PILANI – DUBAI CAMPUS**
**DUBAI INTERNATIONAL ACADEMIC CITY**
**SECOND SEMESTER 2012 – 2013**
**EA C414 INTRODUCTION TO BIOINFORMATICS**
**QUIZ-I (CLOSED BOOK)**

**Duration: 20 min.** **Date: 3.4.2013** **Max. Marks: 8**

**Note:** *Answer all the questions*
*Answer to the point and draw suitable diagrams if required*

1) Consider four species characterized by homologous sequences ACTTC, AGTTC, CGTAC and CCTAC. Taking the number of differences as the measure of dissimilarity between each pair of species, use a simple clustering procedure to derive a phylogenetic tree. [2.0]
Clustering, dotmatrix, phylogenetic analysis, similarity analysis

2) What are archives and give any two examples. [1.5]
databases, sequence retreival, annotation, NCBI, PDB

3) What are the applications of NMR and X-ray crystallography in Bioinformatics? [1.5]
protein structure determination, homology modeling

4) What is OMIM and mention the applications? [1.5]
genetic disease database, genes and related information

5) What are protein information resources (PIR) give any two examples? [1.5]
Expassy, swissprot, pdb