



KENYATTA UNIVERSITY
UNIVERSITY EXAMINATIONS 2011/2012

**SECOND SEMESTER EXAMINATION FOR THE DEGREE OF BACHELOR
OF SCIENCE IN BIOCHEMISTRY AND BIOTECHNOLOGY**

SBC 354 : INTRODUCTION TO BIOINFORMATICS

DATE: THURSDAY 29TH MARCH 2012

TIME: 11.00 A.M. – 1.00 P.M.

INSTRUCTIONS

- **Answer ALL Questions in Section A and B and Any TWO questions in Section C**
- **SECTION A: Compulsory (20 marks) (Each question is one mark)**

1. The field of bioinformatics...
 - a) Uses biomimetic algorithms to develop more efficient software.
 - b) Integrates concepts and techniques from information technology and molecular biology.
 - c) Requires complete genome sequences to be useful.
 - d) Has only developed in the last 5 years.
2. Human genome contains about
 - a) 2 billion base pairs
 - b) 3 billion base pairs
 - c) 4 billion base pairs
 - d) 5 billion base pairs
3. Analysing or comparing entire genome of species
 - a) Bioinformatics
 - b) Genomics
 - c) Proteomics
 - d) Pharmacogenomics
4. Characterizing molecular component is
 - a) Genomics
 - b) Cheminformatics
 - c) Proteomics
 - d) Bioinformatics
5. What is the principle application of the BLAST family of algorithms?
 - e) Identifying sequences that are similar to a protein or nucleotide sequence in a biological sequence database.
 - f) Aligning two nucleotide sequences from end to end.
 - g) Identifying the best possible alignment of two short protein sequences.
 - h) Finding the minimum-energy configuration of a polypeptide sequence.

6. Which of the following is NOT true of a BLOSUM80 matrix?
- It gives the log-odds of substitution of any given pair of amino acids.
 - It is calculated based on the Gibbs free energy (ΔG) of amino acid substitutions.
 - It is based on the BLOCKS database of sequence motifs.
 - It is best used for analysis of very distantly related proteins.
7. At the NCBI web site, the default scoring matrix for protein-protein BLAST analyses (BLASTP) is the BLOSUM62 matrix. Which of the following is a likely reason that this matrix was selected as the default?
- It represents a compromise between information content for each residue and the amount of information that contributes to the matrix.
 - Only this matrix has been shown to be appropriate for all protein comparisons.
 - It incorporates 62 position-specific scoring patterns.
 - Unlike PAM matrices, this matrix is based on explicit phylogenetic information.
8. An example of Homology and similarity tool?
- PROSPECT
 - EMBOSS
 - RASMOL
 - BLAST
9. The first bioinformatics database was created by
- Richard Durbin
 - Dayhoff
 - Michael Dunn
 - Pearson
10. In blast database, Blastn is used to search for
- Protein sequences
 - Aminoacid sequences
 - Nucleotide sequences
 - All of above
11. Multiple sequence alignment is used to
- Predict nucleotide similarity
 - Predict bacterial identity
 - Predict protein sequence
 - None of the above
12. Labeling a stretch of DNA according to its function is called
- Recombinant DNA technology.
 - Functional analysis.
 - Annotation.

- d) Screening.
13. If a single organism contains two loci encoding homologous genes as the result of a gene duplication, the sequences encoded at these loci are said to be:
- a) Orthologous
 - b) Metalogous
 - c) Paralogous
 - d) Xenologous
14. What characteristics of protein-coding regions are (potentially) useful for detecting these sequences among genomic sequence?
- a) Patterns of base-compositional bias.
 - b) Lower frequency of stop codons in one reading frame than would be expected under a random model.
 - c) Fragments of sequence that closely resembles those of known protein-coding genes.
 - d) All of the above
15. The translated genes of genomes that encode proteins are referred to as
- a) The open reading frame.
 - b) Introns.
 - c) Codons.
 - d) Pseudogenes.
 - e) A gene microarray.

For Questions 16-20: Fill in the Blank. Provide the single most appropriate word or phrase. (1/2 mark each).

16. In parsimony analysis, the _____ that requires the smallest number of character-state transitions is preferred.
17. Maximum likelihood, minimum evolution, and maximum parsimony are _____ criteria.
18. GenBank is the comprehensive database in the USA for published _____ sequences.
19. The GenBank _____ is a unique identifier for a sequence that is permanently associated with that sequence.
20. The ----- is used as a convenient way to create a significance threshold for reporting results.

Section B: Compulsory - attempt all questions (30 Marks)

1. a) Diagrammatically describe the “Central dogma” of biology [3 Marks]

b) What is the format of the sequence given below? [2 Marks]

>MTH748 MTH748 upstream sequence, from -200 to -1, size 200

agtgacacaggccacgtttagccttcacagagggcttgatttcagtgaagaaagg

tgggtgacttaaaagtactttctgagtggcgaaggcctggtgcagaattgaaggtac

aggaaccgtatacatgcagagccggagcattgatagcttcgttgatggtaacacccat

gtccccctcaaggagttaga

2. For a phylogenetic tree it's (always) better to use the amino acid sequence?
[5 Marks]

3. What is the main purpose of “bootstrap analysis”? [5 Marks]

4. What can we do or not do with sequence analysis? [5 Marks]

5. How do we detect homology? [5 Marks].

6. Briefly describe the most important indicators of alignment quality that are returned with the results [5 Marks].

Section C: Answer any two questions (10 Marks each)

1. Use the Genetic code table below to determine the messenger RNA of the query sequence MVAETN.

Genetic Code					
	T	C	A	G	
T	TTT Phe [F]	TCT Ser [S]	TAT Tyr [Y]	TGT Cys [C]	T
	TTC Phe [F]	TCC Ser [S]	TAC Tyr [Y]	TGC Cys [C]	C
	TTA Leu [L]	TCA Ser [S]	TAA Ter [end]	TGA Ter [end]	A
	TTG Leu [L]	TCG Ser [S]	TAG Ter [end]	TGG Trp [W]	G
C	CTT Leu [L]	CCT Pro [P]	CAT His [H]	CGT Arg [R]	T
	CTC Leu [L]	CCC Pro [P]	CAC His [H]	CGC Arg [R]	C
	CTA Leu [L]	CCA Pro [P]	CAA Gln [Q]	CGA Arg [R]	A
	CTG Leu [L]	CCG Pro [P]	CAG Gln [Q]	CGG Arg [R]	G
A	ATT Ile [I]	ACT Thr [T]	AAT Asn [N]	AGT Ser [S]	T
	ATC Ile [I]	ACC Thr [T]	AAC Asn [N]	AGC Ser [S]	C
	ATA Ile [I]	ACA Thr [T]	AAA Lys [K]	AGA Arg [R]	A
	ATG Met [M]	ACG Thr [T]	AAG Lys [K]	AGG Arg [R]	G

G	GTT Val [V]	GCT Ala [A]	GAT Asp [D]	GGT Gly [G]	T
	GTC Val [V]	GCC Ala [A]	GAC Asp [D]	GGC Gly [G]	C
	GTA Val [V]	GCA Ala [A]	GAA Glu [E]	GGA Gly [G]	A
	GTG Val [V]	GCG Ala [A]	GAG Glu [E]	GGG Gly [G]	G

2. Describe in details the three main types of methods that are used for phylogenetic reconstruction and list the main factors affecting the outcome of these analyses.
3. Describe the functionality of the different variants of the BLAST algorithm.