File Copy

INSTITUTE OF BIOINFORMATICS AND APPLIED BIOTECHNOLOGY

TERM-END EXAMINATION | **MSc (2021-2023) Fourth Semester** | MONDAY 22 May 2023

BTBIH401: SYSTEMS BIOLOGY

Time: 4 hours *Maximum Marks: 70*

PART A Multiple choice questions (answer any 10 out of 12) (10 x 1 Mark)

A1. The time scale involved in the translation of a protein in a cell is of the order of

a) day **b)** minute **c)** millisecond **d)** microsecond

A2. The gram molecular weight of a small molecule used in a medicine is 100 grams. A tablet containing 5 mg of the molecule will approximately have ______ molecules.

a) 6×10^{23} **b)** 6×10^{21} **c)** 3×10^{21} **d)** 3×10^{19}

A3. The Genome-wide Association Study (GWAS) of immune biomarkers in serum from patients of a particular disease

a) establishes the correlation of the biomarker with the disease

b) identifies the biomarker as the causation of the disease

c) establishes both the correlation and the causation

d) none of the above

A4. If N(t) represents a population as a function of time and p and q are constants, identify the differential equation that describes the population dependent death

a) $\frac{dN}{dt} = pN$ **b)** $\frac{dN}{dt} = N + q$ **c)** $\frac{dN}{dt} = p - qN$ **d)** $\frac{dN}{dt} = (p + q)$

A5. In an enzyme catalyzed reaction, the plot of initial rate against the initial substrate concentration is a straight line with a slope of $0.17\ s$ and an intercept of $2.38\ mol^{-1}Ls$. The estimated value of Michelis Constant K_m in $mol\ L^{-1}$ is approximately

a) 14.0 **b)** 0.07 **c)** 6.30 **d)** 2.55

A6. Identify the phenomena that is not stochastic

a) daily weather **b)** coin toss **c)** planetary motion **d)** movement of gas molecules

A7. The fold change in a differential gene expression is given by the mean expression of a gene in treatment samples divided by the mean expression in control samples. If the log2(Fold Change) of a gene under this definition is -4, then the gene is

a) under expressed by a factor of 16

b) over expressed by a factor of 4

c) under expressed by a factor 4

d) over expressed by a factor of 16

A8. Identify the statistical distribution used in estimating the significance of a contingency table in the Gene Enrichment Analysis

a) binomial distribution

b) hypergeometric distribution

c) Poisson distribution

d) exponential distribution

A9. The *lac* operon is

a) an operon that lacks some short sequence in promotor region

b) an operon required for the transport and metabolism of lactose in certain bacteria

c) a set of genes that help in the metabolism of glucose in bacteria

d) an important enzyme in *E. coli*

A10. Hemoglobin is an example of

a) non-allosteric enzyme
b) allosteric protein that is an enzyme
c) non-allosteric protein
d) allosteric protein that is not an enzyme

A11. In network analysis, the eigenvector centrality score of a node reflects

a) the location of the node closer to the center of the network
b) the number of connections to the node
c) the connection of the node to high scoring nodes of the network
d) the connection of the node to the peripheral regions of the network

A12. The algorithm used for optimizing an Objective function in Flux balance analysis is called

a) Linear programming
b) Non-linear programming
c) Logistic Regression
d) Random forest flux optimization

PART B Short answer questions (answer any 10 out of 12) (10 x 2 Marks)

B1. Briefly explain the *correlative* and *explanatory* models in systems biology.

B2. Write down the differential equation for the constant rate of growth of a quantity **M** as a function of time ***t***. Plot the shape of the solution.

B3. What is the purpose of *Stability analysis* and *Sensitivity analysis* encountered in the steady state analysis of dynamical systems?

B4. Explain the assumptions of *spatial homogeneity* and *continuous hypothesis* made in the study of dynamical behavior of reaction networks.

B5. What is the essential difference between *p-value correction* and *False discovery rate* employed in multiple testing of hypothesis?

B6. In terms of statistical methods employed, how does the differential gene expression analysis of data from microarrays differs from that of RNA sequencing?

B7. Define the Carrying *capacity* of a biological species in a environment. Is it different from *equilibrium* value?

B8. Give one example each for a *directed network* and *undirected network* that can be contructed using data from biology.

B9. Write few sentences on your understanding of *Flux Balance Analysis*.

B10. Compute the average degrees of a directed and an undirected network, both having 50 nodes and 3000 links.

B11. Plot the shape of the curve $N(t) = \frac{200}{2050-t}$

B12. Give an example study in life sciences in which *K-means* clustering is employed.

PART C Long-answer/problem type questions (answer any 8 out of 10) (8 x 5 Marks).

C1. Write down the logistic equation for the density dependent growth and explain various terms. Write the expression for the solution of the equations and sketch the solution.

C2. A system of differential equations for two variables is given by,

$$\frac{dX}{dt} = 2Y - 3X^4 \quad \text{and} \quad \frac{dY}{dt} = 0.5e^{X}Y - 2Y^2$$

(a) Linearize the above equations around the operating point $(0.8, 0.4)$

(b) Write the pair of linear equations in Matrix form and identify the Jacobian Matrix.

C3. The aim of a clinical study was to know whether a particular allele A is associated with the Gender of the person with a disease. Of the 25 patients signed up for the trial, 13 were Male and 12 were Female. Among the 13 Men, 4 were detected with the Allele and the remaining did not have it. Among the 12 Women, 8 had the Allele and the remaining did not have it.

(a) Set up a contingency table for this data

(b) Perform a Fisher's exact test to know whether the allele is associated with the gender among the people with this disease. Use a significance level of 0.05 for this test. Write your null and alternate hypothesis clearly.

C4. An epidemic broke out in a goat farm. The data on the observed number of goats infected with the disease as a function of time(days) is given below

Time = {3,7,8,12,15,16,20,26,29,33,38,43,48}

Infected proportion = {0.013,0.06,0.23,0.39,0.56,0.41,0.28,0.2,0.1,0.02,0.05, 0.017, 0.02}

Fit an SIR model to this data in R with the following parameter values and plot S, I and R as a function of time on the same graph. Beta = 0.079 per day, gamma = 1/8.2 $days^{-1}$
The initial values are as follows:

Initially 3 goats out of 1000 were infected. $I = 3/1000,\ S = 1\ \ R = 0$

C5. An open reaction network representing a chemical reaction $A + B \rightarrow B + D$ is described by the following system of differential equations:

$$\frac{dA}{dt} = K_1 - K_2A - K_3AB$$

$$\frac{dB}{dt} = K_2A - K_3AB$$

$$\frac{dC}{dt} = K_3AB - K_4C$$

$\frac{dD}{dt} = K_3AB - K_5D$ where the values of the constants are given by

$K_1 = 3.1\ mM/s,\ K_2 = 2.2/s, K_3 = 2.5/nM/s,\ K_4 = 3/s,\ K_5 = 4.5/s$
Solve the above system of equations in R. Plot the functions A(t), B(t), C(t) and D(t) on the same graph to study their temporal behavior.

C6. The attached file "filtered_gene_list.csv" contains a list of 238 genes of the species *Mus musculus* differentially expressed under certain conditions in a microarray experiment. Use this gene list to perform gene enrichment analysis using the DAVID web-tool. Sort the enriched KEGG pathways by adjusted ***p*** value (Benjamini) and identify the top six pathways. Visualize the number of genes in these pathways as a bar plot.

C7. The attached file "protein_interactions.csv" contains the interaction between certain number of proteins. The names of the proteins can be accessed from the file "protein_list.csv".
Using the "igraph" and "igraphdata" packages in **R**,

(a) Construct a directed graph describing these protein-protein interactions and plot the graph.
(b) Plot the histogram of the degrees of the nodes.
(c) Compute the closeness centrality of the nodes and print them.

C8. The p-values of 20 differentially expressed genes obtained from an experiment are given below:
pvalue = {0.049, 0.065, 0.0528, 0.001, 0.126, 0.177, 0.008, 0.0192, 0.0856, 0.0295, 0.0122, 0.0939, 0.0722, 0.267, 0.458, 0.0322, 0.0013, 0.025, 0.089, 0.04}
Using the p.adjust() function in **R**, apply Benjamini Hochburg algorithm to control the False Discovery Rate of the above p-values. Print the original and adjusted p-values. How many genes have p-values less than 0.05 after the FDR control?

C9. The given file expression table.csv contains the microarray expression levels of 56 human genes from 9 disease samples and 9 normal(control) samples. In the data table, the rows are genes and the columns are control-treatment samples. Write an R script that performs a hierarchical clustering of this data and creates a heat map using the "heatmap" function of **R** or the "pheatmap package" of Bioconductor library.

Comment on the separation of control and treatment samples in the heatmap.

C10. Consider a population in which a predator and its prey co-exist. For prey, the growth rate per capita = 3.5, death rate per capita = 0.5 and the rate at which a prey is eaten by predator is 1. The growth and death rate per capita of the predator is 1 and 3.2. The equations for this simple predator and prey model are as follows:

$$\frac{d[prey]}{dt} = (growth\ rate\ for\ prey)(prey) - (death\ rate\ for\ prey)(prey)$$

$$- (rate\ at\ which\ prey\ eaten\ by\ predator)(prey)(predator)$$

$$\frac{d[predator]}{dt} = (growth\ rate\ for\ predator)(predator)(prey)$$

$$- (death\ rate\ of\ predator)(predator)$$

Let the initial number of preys is 15 and predators is 7. Solve this system of equations in **R** and sketch the number of predator and prey as function of time.