Institute of Bioinformatics and Applied Biotechnology
Term-end Examination
MSc10 Fourth Semester
Tuesday 19 July 2022

BTBIH-401: Systems Biology

Toal Marks : 70 **Duration : 3 hours**

PART A Answer any 10 questions (1 mark each)

1. An order of magnitude estimate for the volume of a mammalian cell could in the range
   (a) $1 - 10\ \mu m^3$
   (b) $1000 - 10000\ \mu m^3$
   (c) $1 - 10\ nm^3$
   (d) $1 - 10\ mm^3$

2. The time taken by a cell to make proteins could be in the range
   (a) 20 seconds to several minutes
   (b) 1 to 500 milliseconds
   (c) 4 to 8 hours
   (d) 1 to 100 microseconds

3. What is an operon?
   (a) a part of the promoter region of a gene
   (b) a part of RNA polymerase
   (c) a unit made up of linked genes that code for proteins needed to do a specific task
   (d) A protein needed for DNA replication

4. If $\frac{dy}{dt} = 3t$, then the funtion y(t) is
   (a) an exponential curve
   (b) a stright line with a slope of 3
   (c) a parabola
   (d) none of the above

5. When the velocity of enzyme activity is plotted against substrate concentration, which of the following is obtained?
   (a) Parabola
   (b) Straight line with positive slope
   (c) Straight line with negative slope
   (d) Hyperbolic curve

6. In a gene expression analysis, 500 genes were selected from a data set at a significance level of 0.05. What is the total number of genes in the data set?
   (a) 1000

(b) 5000
(c) 10000
(d) The given information is not sufficient to compute the total number of genes in the data set.

7. Gene set enrichment analysis (GSEA)
(a) estimates the significance level of gene expression between disease and normal conditions
(b) creates a network of genes under a disease condition.
(c) is used to discover new pathways from experimental data.
(d) associates a disease phenotype to a group of genes/proteins

8. In network theory, the term "size of the network" refers to
(a) mean number of edges per node
(b) number of edges in the network
(c) the number of nodes in the netwoek
(d) number of isolated nodes in the network

9. The eigenvalues of a system of Ordinary Differential Equations is used to determine
(a) whether an equilibrium point is stable or unstable
(b) the equilibrium points of a system
(c) whether the system has non-trivial solutions
(d) the trivial solution of the system

10. Which one of the following mathematical methods is used for analyzing the flow of metabolites through a metabolic network?
(a) Hidden Markov model
(b) Network analysis
(c) Pathway enrichment analysis
(d) Flux Balance Analysis

11. Something performed on computer or computer simulation is
(a) insilico
(b) silico
(c) invitro
(d) none of these

12. Which of the following is a benefit of the use of mathematical models?
(a) They accurately describe a biological system.
(b) Field and lab observations are no longer necessary.
(c) They can describe observed biological systems in a manner that is quantifiable.
(d) There are more mathematicians than there are biologists.

PART B Answer any 10 questions (2 marks each)

1. Explain the law of mass action.

2. What is the advantage of using a logarithmic scale for Fold Change in gene expression analysis? Explain with an example.

3. Briefly explain the False Discovery Rate (FDR) in statistics.

4. What is the major difference between the gene expression analysis of microarray and RNA-seq data in terms of statistical methods employed?

5. Write down the three differential equations of Susceptibility-Infectious-Recivered (SIR) population model and describe the terms in it.

6. In what ways a steady state (ie., dynamic equilibrium) of a biological system is different from its chemical equilibrium?

7. An undirected network consists of 6 nodes that lie along a circle, with each node connected to its previous node and the next node. Draw the network and the histogram of its degree distribution.

8. What do you understand by the term "objective function" in the Flux Balance Analysis? Explain it.

9. Write the general form of Hill function and and explain its usage in enzyme kinetics.

10. What are the three major Gene Ontology (GO) categories?

11. Describe the "basic reproduction number" and its role in a disease model.

12. With any given gene list as input, mention at least four functionalities provided by the functional annotation tool DAVID.

PART C Answer any 8 questions (5 marks each)

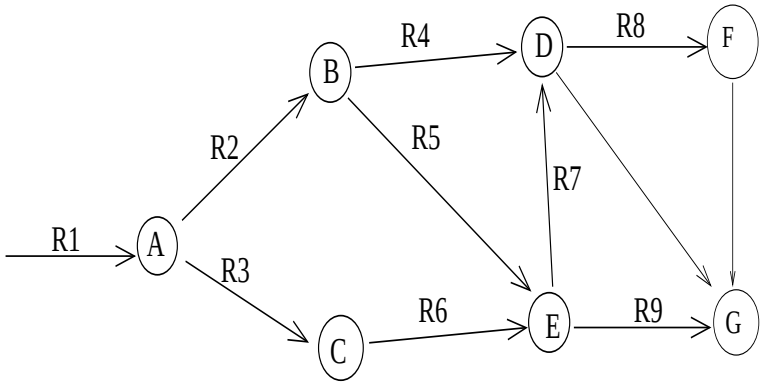
1. Consider the system : $\frac{dx}{dt} = x + 4y + xe^x, \qquad \frac{dy}{dt} = -y - ye^x$

(a) Check that (0, 0) is an equilibrium point of the system
(b) Find the general expression for the Jacobian of this system
(c) Find the Jacobian at the point (0, 0)

2. In a pathway enrichment analysis, the following results were obtained: Of the 15 genes that were differentially expressed, 2 were in the pathway, and the remaining were not found in the pathway. Similarly, of the 16 genes not expressed, 13 were in the pathway and the remaining were not in the pathway.

Construct a contingency table for this data. Perform Fishers exact test for this contingency table to compute a p-value under the null hypothesis that the presence or absence of genes in the pathway independent of their expression levels.

3. Write down the Stochastic Matrix for the following metabolic network:



4. Consider the logistic model for the density dependent growth:

$$\frac{dN}{dt} = rN(1 - N/K)$$

where, N(t) is population as a function of time, r and K are parameters of the model.

The exact solution to the above logistic equation is given by,

$$N(t) = \frac{N(0)e^{rt}}{1 + \frac{N(0)}{K}(e^{rt} - 1)}$$

The data on the United states population(in millions) from census taken every 10 years from 1790 to 1930 is given below:

year	t	population(millions)
1790	0	3.9
1800	10	5.3
1810	20	7.2
1820	30	9.6
1830	50	12.9
1840	60	17.8
1850	70	23.2
1860	80	31.4
1870	90	38.5
1880	100	50.2
1890	110	62.9
1900	120	76.2
1910	130	92.2
1920	140	106.2
1930	150	123.2
1940	160	132.2
1950	170	151.3
1960	180	179.3
1970	190	203.3
1980	200	226.5
1990	210	248.7
2000-	220	281.42
2010	230	308.7

(a) Plot time versus population in a graph

(b) For the values $r = 0.0312 \quad and \quad K = 198.6$, compute N(t) as a function of t using above solution formula. Plot this curve on the same plot along with data points. Upto which year, the fit is good? Is the logistic euation is good model for this population growth data?. Comment.

5. The data table in the given file "GSE56960_expression_data.csv" contains the gene expressions of two types of samples from microarray data along with fold changes in log2 scale and p-values of statistical test for finding differential gene expression. Perform the following analysis in R :

(a) Create a Volcano plot of fold change versus p-value in appropriate scale.

(b) Filter the genes whose absolute fold changes for over/under expression are more than 1.5 and the p-value less than 0.05 and write their names into a file.

6. Solve the following set of coupled ODE's in R.

$\frac{dx_1}{dt} = x_1 - x_2$

$\frac{dx_2}{dt} = -x_2$

$\frac{dx_3}{dt} = 10x_1 - x_3^2$

with the initial condition $x_1 = x_2 = x_3 = 0$ when $t = 0$

Plot the solutions x_1, x_1 and x_1 as a function of t on the same graph.

7. Use the gene list from the given file DAVID_input_entrez.txt to perform gene enrichment analysis using the DAVID web-tool.

(a) Sort the enriched KEGG pathways by adjusted p value (Benjamini) and identify the top ten pathways. Visualize the number of genes in these pathways as a bar plot.

(b) Sort the enriched INTERPRO protein domains by adjusted p value (Benjamini) and identify the top ten domains. Visualize the number of genes in these pathways as a bar plot.

8. Using the igraph package in R, create and plot an unweighted graph that connects the following pairs of nodes:

$N1 \rightarrow N2, N1 \rightarrow N3, N3 \rightarrow N1, N3 \rightarrow N4, N3 \rightarrow N5, N5 \rightarrow N6, N6 \rightarrow N4, N2 \rightarrow N5, N5 \rightarrow N7$

For the same graph, add the edge weights (2.0, 1.5, 3.0, 4.0, 3.0, 5.0, 2.5, 3.5, 2.0) and plot the graph with weights.

9. 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 prey is 15 and predator is 7.

Solve this system of equations in R to sketch the number of predator and prey as function of time.

10. . In a Michaelis -Menten process, the reaction rates 'V' were measured as a function of substrate concentration 'S'. The data is presented below:

S : 0.5, 1.0, 1.5, 2.0, 4, 6, 8, 10, 15, 20, 25, 30, 40, 50, 75, 100
V : 0.16, 0.30, 0.42, 0.52, 0.85, 1.06, 1.22, 1.34, 1.54, 1.67, 1.75, 1.81, 1.90, 1.95, 2.03, 2.07

in R, plot 'S' versus 'V' on a linear scale and get read the approximate values of V_{max} and K_m from the graph. Print them.