Complex Systems in Bioinformatics

Freie Universität Berlin, Summer 2024
Martin Vingron · Ekin Deniz Aksu

Assignment 1
Due date: 25.06.2024 12:00 PM before the lecture

You should upload 2 files: a single PDF for theoretical exercises (can be LaTeX or handwritten), and a Python/Jupyter notebook file for the coding exercises.

Problem 1 (*30 Points; Gene Networks*)**.** Consider the following RPKM values from 5 RNA-seq experiments for the following genes (you can use the file rpkm.csv):

Gene	Exp1	Exp2	Exp3	Exp4	Exp5
A	7	9.2	14.6	20	35.1
B	19	14.2	6.6	14.6	18
C	8.6	7.0	6.5	7.3	8.7
D	6.8	7.9	5.5	2.3	2.9
E	0.9	1.8	3.9	4.8	6.2

Using the package NetworkX for Python 3, draw the gene network with the following criteria for edges:

(A) Draw an edge between genes X and Y if the Euclidean distance

$$d_E(X,Y) := \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} < 10$$

n is the number of samples (i.e. experiments) and $X, Y \in \{A, B, C, D, E\}$.

(B) Draw an edge between genes X and Y if the correlation coefficient $|r(X,Y)| > 0.75$. Color the edges with positive correlation red and the edges with negative correlation blue.

(C) Draw an edge between genes X and Y if the L_1-norm:

$$||X,Y||_{L_1} := \sum_{i=1}^{n} |x_i - y_i| < 20.$$

(D) Draw an edge between genes X and Y if the mutual information:

$$I(X,Y) := \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) > 0.65.$$

To calculate the mutual information, bin the RPKM values for each gene into 3 intervals. Show the MI values as edge weights.

Problem 2 (*35 Points; Probabilistic Distribution, Independence, Information Theory*)**.** Consider two random variables X and Y from which we drew the following samples:

$$\begin{aligned} x &= (0.3, 0.98, 0.54, 0.49, 0.39, 0.13, 0.03, 0.81, 0.65, 0.18) \\ y &= (0.74, 0.09, 0.48, 0.15, 0.71, 0.8, 0.53, 0.95, 0.61, 0.88) \end{aligned}$$

Therefore the first observation is ($x = 0.3$, $y = 0.74$) and so on (10 observations in total). First, bin the data by dividing the interval of $[0, 1]$ into 4 equally wide sub-intervals. Provide the following calculations *by hand*.

(A) Calculate the joint probability distribution $p_{X,Y}(x, y)$ of the binned data and write it in the following table:

$Y\|X$	x_1	x_2	x_3	x_4
y_1				
y_2				
y_3				
y_4				

(B) Calculate the marginal distributions $p_X(x)$ and $p_Y(y)$.

(C) Calculate the product of the two marginal distributions $p_X(x) \times p_Y(y)^T$ (matrix multiplication!) and compare it with the joint distribution $p_{X,Y}(x, y)$. Are the variables X and Y stochastically independent? Justify your answer.

(D) Calculate the conditional distribution $p_{X|Y}(x|y = y_3)$.

(E) Calculate the joint entropy $H(X, Y)$ and the marginal entropies $H(X)$ and $H(Y)$.

(F) Calculate the conditional entropies $H(X|Y)$ and $H(Y|X)$ using the chain rule.

(G) Calculate the mutual information $I(X, Y)$ using both, the definition and the relation to entropy. Are both results equal? Why?

Problem 3 (*15 Points; Bivariate Gaussian independence theorem*)**.** Prove the following theorem.

Theorem: Two random variables from a bivariate Gaussian distribution are independent if and only if the correlation coefficient between them is zero.

Problem 4 (*20 Points; Network topology*)**.** Install Cytoscape 3.8 and get familiar with its interface. For subsequent tasks use the package NetworkX for Python 3.

(A) Generate a random network G($n = 150$, $p = 0.08$) according to Erdős-Rényi model where n is the number of nodes and p is the probability of including an edge in the graph (remember to set a random seed). Analyze your network: calculate the number and size of connected components, plot histograms for the degree values, closeness and betweenness

centralities. Normalize the degree histogram and on top of it plot the theoretically appropriate Poisson probability mass function (you may use scipy.stats.poisson). Compare the two for different bin numbers of the histogram. Is the bin size important? Export the network as GraphML file and the node attributes (centralities, degree) as a csv file. Check what happens when you change the value of p (you do not need to hand in the results for other values of p).

(B) Import the network and its attributes from the previous task into Cytoscape. Visualize it so that: the size of the node reflects the closeness centrality value (the larger the closeness, the larger the node), betweenness centrality is represented by colour of the node (viridis colour palette - yellow for small values, dark blue for large values) and the degree value is set as the label of the node. Set layout to degree sorted circular and save the resulting visualization as an image. Zoom-in on the node with the largest degree and save the view.