T-61.5110 Modeling biological networks

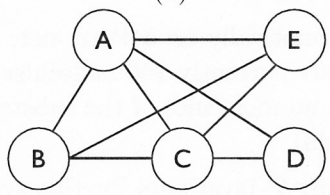
Exam, February 16, 2016

You are NOT allowed to use calculators or any other additional equipments/material in the exam. Please write your answers in English. Please write carefully.

1. a) Compute the average node degree statistic for the network shown in Fig. 1 (a). (3 points)

   b) The (approximate) distribution of the average node degree for the 5-node Bernoulli random network model (without self-loops) with parameter $p = 0.5$ is shown in Fig. 1 (b). Given a significance level of $\alpha = 0.05$, is the network in Fig. 1 (a) statistically significantly different from this Bernoulli random network model? (The exact p-value is not important because you are not allowed to use calculators, but explain the way you make your conclusion.) (3 points)

(a)



(b)

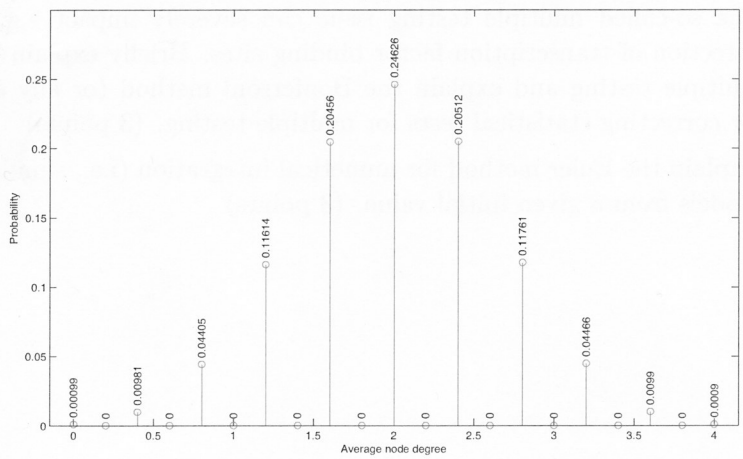


Figure 1: (a) An undirected biological network. (b) The distribution of the average node degree for 5-node Bernoulli network model with $p = 0.5$.

2. Assume you are given a position specific frequency matrix (PSFM) model θ for a DNA binding protein P. Assume the width of the PSFM (and thereby the width of the binding site/motif) is n nucleotides.

   a) You are given two DNA sequences, a wild-type $X = (x_1 x_2 x_3 \ldots x_n)$ and a mutant $X^* = (x_1 x_2^* x_3 \ldots x_n)$, where $x_2 \neq x_2^*$. Describe how you can compu-

tationally decide which one of the sequences, X or X^*, is more likely to bind protein A. (3 points)

b) You are given a DNA sequence $Z = (z_1 z_2 \ldots z_\ell)$, where $\ell > n$. Explain how you can assess the statistical significance that Z contains a binding site for protein P. (3 points)

3. A standard (state-of-the-art) approach to identify protein-DNA interactions for a selected protein is to carry out chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). Describe the MACS method for identifying protein-DNA binding sites from ChIP-seq data, assuming a control input-DNA sequencing data is also available from the same biological sample. (6 points)

4. Consider the simple model of Michaelis-Menten enzyme kinetics

$$\begin{aligned} S + E &\rightarrow SE \\ SE &\rightarrow S + E \\ SE &\rightarrow P + E \end{aligned}$$

a) Represent it mathematically as a Petri net, $N = (P, T, Pre, Post, M)$, assuming that there are currently 100 molecules of substrate S, 20 molecules of the enzyme E, and no molecules of the substrate-enzyme complex SE or the product P. (3 points)

b) Can you identify any T-invariants for this system? If you can, identify one T-invariant. (3 points)

5. a) The so-called multiple testing issue can severely impact e.g. genome-wide detection of transcription factor binding sites. Briefly explain the concept of multiple testing and explain the Bonferroni method (or any other method) for correcting statistical tests for multiple testing. (3 points)

b) Explain the Euler method for numerical integration (i.e., simulation) of ODE models from a given initial value. (3 points)