T-61.5110 Modeling biological networks

Exam, April 10, 2015

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. Assume you are given a position specific frequency matrix (PSFM) model θ for a DNA binding protein A. 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 $X = (x_1 x_2 x_3 \dots x_n)$ and a mutant $X^* = (x_1 x_2^* x_3 \dots x_n)$, where $x_2 \neq x_2^*$. Describe how you can computationally 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 $Z = (z_1 z_2 \dots z_\ell)$, where $\ell > n$. Explain how you can assess the statistical significance that Z contains a binding site for protein A. (3 points)

2. 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)

3. a) Explain the Gillespie algorithm for simulating coupled chemical reactions. You can assume general stochastic rate constants $c_1, \dots, c_v$ and hazard functions $h_1(\mathbf{x}, c_1), \dots, h_v(\mathbf{x}, c_v)$ for all reactions. (3 points)
   b) In addition to the Gillespie simulation algorithm itself, explain the relationship between the expected value of the stochastic kinetic model (i.e., coupled chemical reactions) and the continuous deterministic formulation (i.e., ODE system). (3 points)

4. a) Explain the Euler method for numerical integration (i.e., simulation) of ODE models from a given initial value. (3 points)
   b) Consider mass-action stochastic kinetics and a second-order reaction R_i : $X_k + X_k \rightarrow X_l$. Explain how the reaction hazard $h_i(\mathbf{x}, c_i)$ for the ith reaction R_i is computed. In addition, explain how the reaction hazard is computed for a third-order reaction R_j : $X_k + X_l + X_m \rightarrow X_n$. (3 points)

5. Explain the principles of Bayesian model selection (including the marginal likelihood), and also explain how the Bayesian approach can be applied for choosing an optimal biological network model. You can use any of the modeling frameworks (ODE models, ODE models with the first order approximation (=regression), ODE models with the linear approximation (=regression), Bayesian networks, etc.) as an example. (Note: exact technical details/derivations of e.g. the marginal likelihood are NOT required.) (6 points)