Department of Computer Science Fall 2014 New Mexico State University

Ph.D. Qualifying Exam: Bioinformatics

This is a closed book exam. The total score is 100 points. Please answer all questions.

1. The negative binomial (NB) distribution has been used to characterize randomness in next-generation sequencing (NGS) data.

(15 points) (a) Let k be the number of successes until the r-th failure is observed in a sequence of Bernoulli trials. Let p be a given success rate in the Bernoulli trial. The random variable k follows the NB distribution. Please write down the probability mass function of the NB distribution.

(10 points) (b) NB has certain properties more desirable than the Poisson distribution. What is the main difference between Poisson and negative binomial distributions?

(5 points) (c) Why is such a difference favorable to NB in NGS data analysis?

2. Design an efficient algorithm to detect whether a short-gun RNA read is formed by splicing two exons of a same gene in the genome. The input includes both the RNA read sequence and the gene sequence. We assume that the exons and introns are not annotated for the gene sequence.

(20 points) (a) Please give the pseudocode for the algorithm.

(10 points) (b) Please describe how to set any parameters that are important to detect exon splicing.

3. In a genome sequencing project, one obtained 50 million DNA reads of length 100 from a genome of 1 billion base pairs.

(15 points) (a) What is the coverage of the sequencing project? Show how you derive the answer.

(15 points) (b) What is the probability that a specific location in the genome is not covered by any read? Justify your answer.

(10 points) 4. Describe an open-problem in biology that needs heavy use of computation. Please describe the goal of the biology problem and the input and output of the corresponding computation problem.