CSC2417
Assignment 3

Jared Simpson
Joseph C. Somody

Instructions

This assignment is due on **Thursday 8 December 2016**, by **23:59**. To turn in your assignment, make a `tar.gz` file containing your source code and a README file, and e-mail it to `jared.simpson+csc2417-a3@gmail.com`. The README file should contain:

- your name, your student number, and your e-mail address;
- instructions on how to compile your code, if necessary;
- instructions on how to run your code to generate the answers to the programming problems; and
- answers to the written questions.

If you prefer to write in LaTeX rather than in plain text, you can instead turn in a PDF containing the same information. For the programming problems, you can use Python (preferred), C/C++, Java, or Perl. If you want to use a different language, please contact me first to ask.

The assignments are intended to be solved individually—you can discuss the problems with your classmates, but please do not give the answers away. For questions requiring a written answer, please provide a complete description of both how the algorithm works and why it works; however, formal proofs are not required. If the instructions or problem descriptions are unclear, please ask via Google Groups.

Late submissions will be penalised by a deduction of 5% of the maximum grade per day (or partial day).

1 Modelling Whole-Genome Shotgun Sequencing

In whole-genome shotgun sequencing (WGS), we fragment many copies of a genome and sequence fragments uniformly at random (that is, each fragment has equal chance of being selected for sequencing). The uniformity assumption allows us to calculate how much of the genome will be covered by reads as a function of our sequencing parameters (read length, number of reads, and genome size). This is known as the Lander–Waterman model. Let N be the number of reads we have sequenced, G be the genome size, and L be the read length (in this simple model, all reads have the same length). The *coverage* of the genome is $a = NL/G$. Answer the following questions, referring to these slides describing the Lander–Waterman model:
http://www.math.ucsd.edu/~gptesler/186/slides/shotgun_15-handout.pdf

(a) Slide 11 gives the probability that there are no reads that start in an interval of length L for both a Binomial and Poisson model of sequencing. Starting from the Binomial distribution, show how the Binomial result was obtained. Hint: what is the number of trials, the number of successes, and the probability of success of each trial?

(b) Using a Poisson model of sequencing, the expected percentage of the genome that is not covered by a sequencing read is e^{-a}. What coverage do we need such that we expect 99.99% of the genome to be covered by at least one read?

(c) Verify the previous result by simulating a genome-sequencing process. Use the coverage you calculated in (b) to determine how many reads of length L from a genome of length G to simulate (you pick L and G, but make sure $G >> L$). You don't need to simulate full read sequences and map them to a reference genome—it is sufficient to create an array of length G to store the coverage at each base of the genome and directly accumulate values in this array during your simulation. Discuss how well your results match the calculation in (b).

(d) The expected number of *contigs* (contiguous intervals of the genome that are covered by at least one read) depends on coverage, genome size, and read length. Using the formula on Slide 13, make a plot of the expected number of contigs when sequencing a human genome ($G = 3\,000\,000\,000$) with reads of length 100, 1000 and 10 000 as a function of coverage in the range 0 to 10X.

2 Hidden Markov Models

Hidden Markov Models (HMMs) have many applications in genome sequence analysis. In class, we saw how HMMs can be used to find CpG islands and genes. In this exercise, we will explore some properties and applications of HMMs.

(a) On Slide 49 of the Week 9 lecture, we described the *decoding* problem as calculating:

$$p^* = \arg\max_p P(p|x) = \arg\max_p P(p, x)$$

Explain why finding a path of states that maximizes the joint probability $P(p, x)$ gives the same path of states that maximizes the conditional probability $P(p|x)$.

(b) One major type of genetic variation is *copy-number variation*, where a long segment of a genome (> 10 kbp) has either been duplicated or deleted. One method of detecting copy-number variation is to map the reads from a sequenced sample to the reference genome and look for regions where the number of mapped reads is either much higher or much lower than expected from random sampling. For example, we could count the number of reads that map to contiguous 1-kbp windows spanning the genome and assign each window a copy number of 0 (deletion), 1 copy (no change), or 2 copies (duplication). In this exercise, you will design an HMM for this task, but you don't have to implement it. You may assume reads are sampled uniformly at random as in the previous question.

i Describe the states, emission distributions, and state-to-state transitions of your HMM.

ii How would you train the parameters of your model given labelled training data?

iii How would you train the parameters of your model if you only have a vector of windowed read counts but no labels?

iv Do you expect any reads to map to the regions with a deletion (copy number 0)? Why or why not?

v We assumed that each sequencing read was selected from the genome uniformly. Real DNA sequencing data often deviates from this model due to sequence composition biases (for example, reads with high GC content are over-represented in Illumina sequencing data). How might you modify your model to account for sequencing biases?