Homework 3

CS/HG/BIOINFO 124/224

Due: May 13, 2014

- All assignments must be submitted by a hard copy to TA's office by 4 pm on the day they are due.

- Show all your work.

1 General Re-sequencing problems

Discuss the following problems in a few sentences in your own words.

1. What is re-sequencing, and how would you re-sequence your genome (what do you need to have to re-sequence your genome)?

2. What are some of its problems?

3. Is it better to have high coverage or low coverage to have low-error rate of re-sequencing? Why?

Solve the following problems.

4. The trivial mapping algorithm slides the read along the genome and counts the total mismatches between the read and genome. If the mismatches are below a threshold, it is a match. Let N be the length of genome, M be the total number of short reads, L be the read length, t be the time to perform each comparison of match vs. mismatch, and D be the mismatches allowed. What is the time complexity of this trivial mapping algorithm? Give an answer in Big-O notation.

5. Let D be 2, N be 3,000,000,000, and M be 300,000,000. We discussed the indexing algorithm that generates hashtable (or index) for the genome and finds a perfect matching read substring. If L=24, what is the length of each entry (key) in the hashtable? How many positions does each key have on average? What if L=36?

6. What is the time complexity of indexing algorithm? Give an answer in Big-O notation.

7. To solve sequencing errors problem, we collect abundant reads (coverage) and take consensus among them. Let's say error rate of reads is 1%, and we are going to predict the consensus sequence. What is the total error rate if the coverage is 5?

2 Coverage of Re-sequencing

- Assume that the coverage of re-sequencing follows Poisson distribution for the following problems.
- Also, please use R to solve the problems, and show a transcript of your code and results.

1. Let's assume that the overall coverage is 15 (15x coverage). What percentages of genome have coverage less than 1 (< 1)?

2. Again, let's say the overall coverage is 15. What percentages of genome have coverage exactly equal to 1?

3. If we want at least 10 coverage for 90% of genome, what is the minimum overall coverage do we need?

4. What if we want at least 15 coverage for 90% of genome?

Following problems are related to diplod case. Remember that humans have 2 chromosomes. Show how to compute the following probabilities using math equations (you can use R functions such as dbinom, ppois, etc). Also, show R code to compute the equations and its output.

5. Let's assume that for specific positions of genome, we have 30 short reads. What is the probability of having at least 10 coverage for each chromosome? (Hint: Use binomial distribution)

6. Let's assume that the overall coverage is 30. What is the probability of having at least 10 coverage for each chromosome over the whole genome?