COMP90016 – Assignment 2

Haonan Li

May 4, 2018

1 Introduction

This assignment consists of three tasks.

In the first task, we discuss the application of HMMs to CNV detection in theory. The second task is an implementation of a CNV detection algorithm using segmentation. In the final task, we discuss the biology of a particular cancer sample.

2 Task1

The lectures introduced a HMM designed to detect CNVs in a haploid organism. It consisted of three states: One for the normal copy number (CP1) and two alternate copy number states (Cp0, CP2).

Question 1 How would you adapt this approach to a diploid organism, such as human (with states representing different integer copy numbers)?

Answer: For a diploid organism. The normal copy number of chromosome is 2. Take account of copy number changes. four alternate copy number states should be used (CP0, CP1, CP3, CP4). All of them maybe exist in the real world. However, it should be adjust according to the datasets. If there are some bins with 5 copy numbers. We might add CP5 to the state set.

Question 2: Explain the trade-off between the sensitivity of such a HMM and the computational complexity to solve the Viterbi algorithm for it.

Answer: The complexicity of HMM model is $O(nm^2)$. Where n represents the length of the sequence and m is the number of states. If m is large, of course the experiment can contain more possible copy numbers and the result will more precisely. But it will also take more time It will take more than twice as much time for m change from 4 to 6. So, if there are almost no bin with 5 or 6 copy numbers. It is not worth to extend more states.

Question 3: Consider the data shown in Figure 1. The plot shows read depth of bins normalized by the average in a diploid organism. How would you use this data to parametrize the emission probabilities of your HMM? Explain what about Figure 1 is general and what is specific to the data that this plot was derived from. How does this affect the HMM in terms of its application to different data sets? Also, how could this data be utilized to derive the transition probabilities for a CNV detection HMM?

Answer: First, to build a HMM model, we should know each bin's copy number and average depth. We can calculate the emission probabilites by statistic of these informations. For a specified copy number CPX (X = 0,1,2,...). We count how many bins with CPX, assumed N (N¿=0). And for each average read depth D (maybe a range), we count the number of bins with the read depth, assumed M. So the emission probability of CPX to read depth D is M/N.

In Figure 1. There is a peak whose ratio is 1 which is general for diploid organism. But another peak with ratio=0.5 is spicific, which means this sample have some bins with deletions. This will increase the predictions of deletion if apply this to other data sets.

Computing transation probabilities is similar with computing emission probalabilitits. Count the changes copy number pairs of adjacent bins. And compute the proportion of the end state from a specified start state.

Question 4: Describe why a HMM, such as discussed in this task is not very useful for non-clonal data (such as shown in Figure 2 below). Could this shortcoming be alleviated by introducing non-integer CN states?
Answer:

3 Task 2

The implementation of cirular binary segmentation is not complex.

First, inputing. Read data from file and build two lists, one is to store the tuple of start and end positions, the other stores the read depths of all bins.

Second, init the settings. We use numoy to speed up computing. Transfer the list of read depths to numpy array. Then take the median from the first third of the input read depths as 'normal' bin m. Tranform read depth each bin b into log ratios $log2(b/m)$. Then remove extreme values, specifically, set any log2-ratio that are larger then 2 or smaller than -5 to 0. After this init a segment mark array with 0 for later recersive use.

Third, a revursive cbs function, this is the most important part of the algorithm. The function have five input parameters: log-ratio array X, segment mark array I, start position a, end position b and z-threshold t. For each call of the function. First cut X from a to b, which is the current interval we analyze, name it X'.

Then compute cumulative sum of bins $S_i = X_1 + X_2 + ... + X_i$

After computing cumulative sum. We build a 2d array Z and compute Z by:

$$Z_{ij} = (\frac{1}{j-i} + \frac{1}{n-j+i})^{-\frac{1}{2}} \times (\frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i})$$

Then find the maximum of Z, if the maximum is larger than threshold t, its coordinates are new segment points, suppose x, y. We mark them in segment mark array I (set corresponding positions to 1) and call three new cbs functions with new interval (a,x),(x,y),(y,b).

Finally, the segment mark array should have marked with 1 in some positions. Find the corresponding interval between two adjacent marks. Compute the average log-ratio and output.

As for the theoretical complexity of the algorithm. For a file with n bins. we compute a $n \times n$ matrix and seperate it to 3 segment and compute recursively. So the complexity is:

$$Complexity = O(n^2 + 3^1(\frac{n}{3^1})^2 + 3^2(\frac{n}{3^2})^2 + ... + 3^t(\frac{n}{3^t})^2) = O(n^2)$$

References