Data Analytics - Assignment #6

Nishant Kumar (Sr.No: 21495)

November 2023

Color Blindness Detection

We have four different ways genes can be arranged to determine if someone is colour-blind. To figure out which arrangement is most likely to cause colour blindness, we need to compare the DNA sequences in a person's genes with a known standard. We'll count how many times the DNA sequences in the person's genes match the specific regions related to the red and green genes. Based on these counts, we can determine the most likely gene arrangement that leads to colour blindness.

Approach:

To solve the problem, we follow the same approach taught in class. We start by making four binary arrays for the characters A, G, C, and T from the BWT column. These binary arrays need 4n bits of storage.

Next, we use these binary arrays to keep track of the ranks in four separate integer arrays, using Δ milestones. This part needs 4n/Δ bytes of storage. When you want to find the rank of a character, it takes O(Δ) time. The Δ value is a way to balance how much time and memory we use in this problem.

After creating the rank arrays, we can use them to find where a specific part of a text matches a reference sequence. We do this by looking for a specific range of positions in the text's transformed version. To find this range, we use the rank query, which helps us identify the matching area. We also have a map that tells us how these positions in the transformed text correspond to the original reference sequence. This map helps us figure out where the matching parts are in the reference sequence.

We are going to see how many of these matched positions are inside the specific regions called "red exon" and "green exon" for each read. If a read-only matches with either the red or green exon (no confusion), we add 1 to the respective count. If it's not clear and matches with both, we add 0.5 to both the red and green exon counts.

When you're allowed to have up to two mistakes when comparing a read to a reference sequence, you break the read into three nearly equal parts. Then, you examine each of these segments to see if they match exactly with the reference sequence. If there's a match, you record the position (index) where it happens. Afterwards, you need to check the reference sequence at those positions to count how many mistakes (mismatches) are there, and you're allowed to have up to two of these mistakes.

How to find most probable configuration:

We have two sets of genes, one labelled "red" and the other "green." For each set, we count the number of matching exons, which we'll call Rc for the red genes and Gc for the green genes. These counts can be 1, 2, 3, 4, 5, 6 for both sets.

Now, we want to calculate the probability of getting these counts for each set, given a certain configuration, which we'll call "k." In this case, the configuration k has values for 2, 3, 4, 5.

We're using a binomial distribution to estimate this probability, which is a way of describing the likelihood of getting a certain number of successes (matching exons in this case) out of a certain number of trials (the total number of exons) for a given configuration k.

$$P(\{R_c, G_c\}|C_k) = \prod_{i=2}^{5} \binom{r_i + g_i}{r_i} p_{ki}^{r_i}(1 - p_{ki})^{c_i} \tag{1}$$

The value pkj represents the likelihood that the jth part of a genetic sequence is red for a specific arrangement (k). Conversely, (1 - pkj) is the probability that the j^{th} part is green.

Result:

The number of matches for red exons is **97, 237, 107.5, 167.5, 303.5, and 235**. For green exons, the matching numbers are **97, 156.5, 99.5, 87.5, 217.5, and 235**. It's worth noting that the first and last exons have the same number of matches for both red and green genes, as we would expect.

The table below shows the exact probability values and their logarithmic probabilities for four different setups, as presented in the slides. These probabilities were determined by rounding down the counts we previously measured.

Configuration	Probability	log(Probability)
I	5.6×10^{-98}	-223.92
II	0	$-\infty$
III	3.9×10^{-77}	-175.93
IV	4.3×10^{-112}	-256.41

Therefore, we can say that Configuration-III is the most likely one, and the individual with colour blindness has this specific red-green gene structure.

Select Query on Binary Array:

The task is to locate the i^{th} occurrence of the number 1 in a list of binary (0s and 1s) numbers with n elements. You are allowed to use extra memory space that is proportional to cn/Δ, and your goal is to make the value of c as small as possible. Additionally, you need to achieve this with a time complexity of O(Δ).