DA Assignment 5

Shreeman Agrawal

October 2024

Problem Statement

This project involves aligning genome sequence reads to the reference sequence of the human X chromosome, with a focus on two specific genes (referred to as the "red" and "green" genes) associated with color vision. Given 3 million reads and the reference sequence of chromosome X, our objective is to count the reads that map to exons of these two genes. The steps are as follows:

1. **Read Alignment:** Using the Burrows-Wheeler Transform (BWT) along with pointers back to the reference, align each read to the chromosome X reference sequence. Each read will allow up to two mismatches during alignment, and any 'N' present in the reads will be interpreted as an 'A'.
2. **Exon Mapping and Counting:** For each mapped read, determine its location within the exons of the red and green genes. We will count each read as 1 if it maps unambiguously to one of the genes and as 0.5 per gene if it maps ambiguously to both.
3. **Configuration Probability and Inference:** For each hypothesized red-green gene configuration associated with color vision and color-blindness outcomes, calculate the probability of generating the observed read counts. Based on this, identify the most likely genetic configuration that leads to color-blindness.

Implementation Details

This implementation aligns DNA sequence reads from chromosome X to exons of the red and green opsin genes to analyze configurations associated with color vision. The process involves several key steps:

1. **Data Loading Functions:** Functions `loadLastCol`, `loadref`, `loadreads`, and `loadmap` read essential files, including the BWT last column, reference sequence, genome reads, and mappings from BWT indices to reference positions.

2. **Milestones Class:** The `Milestones` class enables efficient alignment through precomputed cumulative character ranks for the BWT last column. The functions `last_col_rank` and `col1_index_rank` facilitate locating read positions by calculating ranks and indices within the BWT.

3. **Read Alignment and Exon Counting:** Using `get_window`, potential alignments for each read and its reverse complement are identified, allowing up to two mismatches. The `identify_exon` function classifies reads that map to exons of red or green genes, contributing counts based on read specificity (1 for unambiguous, 0.5 for ambiguous).

4. **Configuration Probability Calculation:** The `ComputeProb` function evaluates the probability of observed exon counts across multiple gene configurations. Using combinatorial methods, it calculates the likelihood of each configuration, selecting the most probable outcome based on observed data.

The probability of observing exon counts $\{R_i, G_i\}_{i=1}^{n}$ for a given configuration is:

$$P(\{R_i, G_i\} \mid \text{configuration}) = \prod_{i=1}^{n} \binom{T_i}{R_i} (p_i^R)^{R_i} (p_i^G)^{G_i}$$

where:

- $T_i = R_i + G_i$ is the total read count for exon i,
- $\binom{T_i}{R_i}$ is the binomial coefficient,
- p_i^R and $p_i^G = 1 - p_i^R$ are the configuration probabilities for red and green genes.

The configuration with the highest $P(\{R_i, G_i\} \mid \text{configuration})$ is selected as the most likely.

Results

The exon read counts for the red and green genes are as follows:

$$[135, 74, 78, 145, 302, 358, 135, 186, 75, 123, 337, 358]$$

Each configuration specifies different probabilities for reads mapping to the red and green gene exons. We computed the likelihood of observing these exon counts for each configuration. The resulting probabilities for configurations 1 through 4 are:

$$[4.89 \times 10^{-33}, 0.0, 1.95 \times 10^{-17}, 1.99 \times 10^{-39}]$$

Configuration 3 yields the highest probability (1.95×10^{-17}) and is therefore the best match for the observed exon counts. This suggests that Configuration 3 most accurately represents the alignment patterns associated with color blindness.