Assignment 4: Color Blindness

Adarsh T Shah, MTech AI, 19473

# Implementation Summary:

* In the beginning, all the data is initially extracted from files.
* The Milestone is computed from burrows wheeler transform with delta=100. Here delta signifies the number of characters between two milestones. Each line of file contains 100 characters. *Milestone* is a dictionary which maintains rank of A, C, G and T from beginning till the milestone.
* *Extract* is a utility function to extract strings of given length from reference string starting from given index.
* The *ReverseComplement* function generates reverse complement of read and it is searched along with read.
* *Count\_miss* is a utility function to check if two strings have less than equal to two mismatches.
* The function *Search* performs the search using BWT and Bands as explained in the slides. The code is well commented, and its implementation is much clearer there.
* The *Search* function returns on facing the first mismatch. It returns the position till the suffix matches and the corresponding band.
* Based on bands and the indices obtained from the *chrX\_max.txt,* the string is extracted from reference using *extract* function, compared with the read and using the range of positions given in the Readme file for exons belonging to R and G, it’s decided if read belongs to R or G or both genes.

# Result:

Count = 1540

This metric is obtained by Count+=0.5 if read is present in both genes else Count+=1 if read is present in exactly one gene.

Red Gene Matches:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 181 | 89 | 94 | 178 | 332 | 444 |

Green Gene Matches:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 181 | 239 | 148 | 159 | 398 | 444 |

Approximate Runtime duration = 1.5 hrs

# Problem 2:

1. 50%,50%,50%,50%
   1. This configuration clearly demarcates green gene from red gene. Hence, color blindness is not possible.
2. 100%,100%,0%,0%
   1. This configuration is most likely to cause color blindness as exons 4 and 5 of red and green genes are identical
3. 33%,33%,100%,100%
   1. This configuration is most likely to cause color blindness as exons 4 and 5 of red and green genes are identical
   2. The counts obtained from the data support this condition to be the cause of color blindness.
4. 33%,33%,33%,100%
   1. This configuration has some probability to cause color blindness because exon 5 of red and green genes is identical.