Data Analytics: Assignment - 4

Color Blindness

Author: Archit Agarwal

Agenda:

(a) reads from a genome – 3 million reads of the 150m we actually generated
(b) the reference sequence of chromosome X – 150m instead of 3b for the whole genome
(c) the BWT last column and the pointers back to the reference for chromosome X
(d) the locations of the exons of the red and green genes in chromosome X

Align the reads to the reference sequence with up to two mismatches, and then count reads mapping to exons of the red and green genes, counting 1 for each read that unambiguously maps to one of the two genes, and 1/2 for each gene for a read that maps ambiguously.

BTW:

The Burrows-Wheeler transform is a data transformation algorithm that restructures data in such a way that the transformed message is more compressible. Technically, This is obtained by taking circular shifts of the string and then arranging them in ascending order. After that, we throw away every column but the first and last, these two columns will be used to reconstruct the string back given the first index character of the string

Goals:

1. In the last column of the reference string, the reference string is the string of a healthy individual created over a long period of time. This is the reference gene we match our reads to
2. The reads are approx a 150-character long DNA sequence. These are the sequence at multiple positions of a DNA
3. Character Map, this is the first column but instead of the first column its the index

Algorithm:

1. Load the file and get the last column of the BWT array
2. Generate the binary array of the four characters ACGT
3. Generate the sum array with a delta value of 1000, this is the array that serves the cumulative sum of the array of every delta element
4. The rank function finds the rank with the help of a binary array and sum array
5. The match follows the BWT method to reduce the index range to finally find recursively the array

Results:

1. Observations:
    a. Characters in last column: 151100561
    b. Total count of A,C,G,T is 151100560
    c. count of A,C,G,T individually 45648952, 29813353, 29865831, 45772424
    d. character set observed {'A', '$', 'G', 'C', 'T'}

2. The result obtained was the following.
    a. Red exon match: 276
    b. Green exon match: 452
    c. The ratio is **~61%** hence color blindness

Known facts: For a perfectly normal result the exon ratio should be 50% for every red match there should be two green matches
Blue/Green mixup The other color blindness scenario is where the red-to-green percentage is less than 50%. That is where green exons mix with red exons and compromise blue receptors.
Red/Green mixup: The red-green mixup is a scenario where the red exons ratio to green is higher than 50% which is there in the above case, this compromises the ability to distinguish red/green color