Mohit Kumar
SR No. 04-01-03-10-51-21-1-19825
MTech Artificial Intelligence

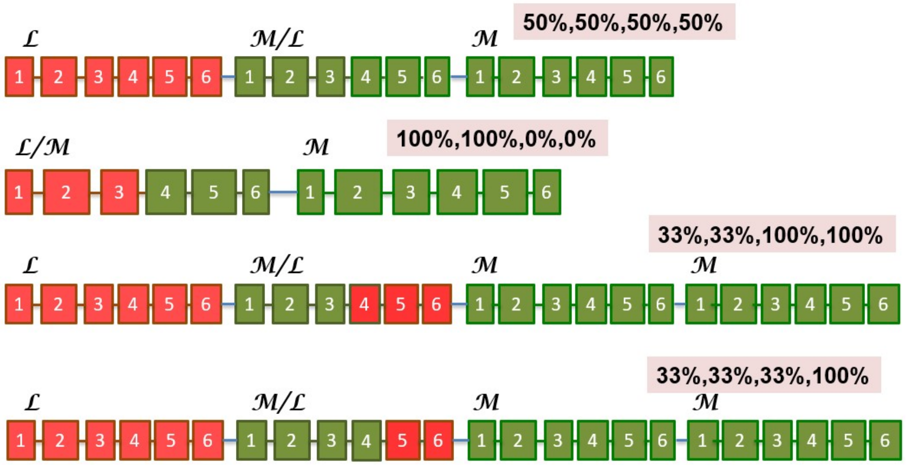
Assignment 4: Color Blindness

---

python packages used: numpy

Implementation details and Observations

- I have used Delta = 1 to optimize the code with respect to time. I have stored the rank of every element.
- To allow for a maximum of two mismatches, I divided each read into three equal parts. Even if a single part gets exactly matched to the reference sequence then I check the mismatches between the complete read and the portion of the reference sequence calculated with the help of the exact match of the subpart. If the number of these mismatches is less than equal to 2 then I count the read as a match.
- We 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.
- To calculate the conditional probabilities of observing the above number of counts given the various configurations, I assumed each read that got mapped to the two genes as a Bernoulli random variable with p depending upon the given configuration.
- Since Probability(counts —configuration 3) is the greatest among the four possibilities, Configuration 3 has the maximum likelihood.
- The code is implemented according to the template and is well commented for easy understanding of the reader.


50%,50%,50%,50%
$\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}/\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6
100%,100%,0%,0%
$\mathcal{L}/\mathcal{M}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6
33%,33%,100%,100%
$\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}/\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6
33%,33%,33%,100%
$\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}/\mathcal{L}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6
$\mathcal{M}$ 1 2 3 4 5 6

- We don't need to match all the reads with the reference sequence. The reads are given in increasing order, therefore the initial reads in the reads file are matched to the initial part of the reference sequence and the later reads match to the later part. The reads corresponding to the red and the green genes are found between $(2936000, 2948000)$ in the reads file.

Count of reads belonging to Red Exons

$R1$	$R2$	$R3$	$R4$	$R5$	$R6$
97	143	85.5	160.5	279.5	235

Count of reads belonging to Green Exons

$G1$	$G2$	$G3$	$G4$	$G5$	$G6$
119	207	149.5	146.5	352.5	259

P(counts|config0)	P(counts|config1)	P(counts|config2)	P(counts|config3)
2.49e-25	0	1.61e-20	5.62e-43

Therefore we see that since the probability of observing the above read counts is maximum for configuration 2. Configuration 2 is the most probable gene configuration.

Program execution time : 2579.22 s when we only match reads between $(2936000, 2948000)$. The code was run on a 8 GB RAM machine with intelcore-i5-11th gen processor(64 bit Windows11 OS).