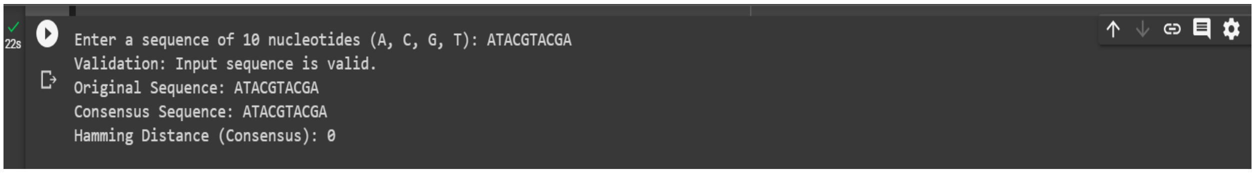
Abin assignment grp-6

Answer 1)

Well commented code for Q1 is submitted in the jupyter notebook.

1) write down the formula for the number of strings with 0, 1, 2 mutations-> (NCK*3^k) where N represents the string/sequence of length N. K represents the position in sequence or string.

The input sequence is: ATACGTACGA



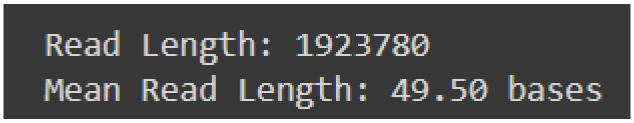
```
Enter a sequence of 10 nucleotides (A, C, G, T): ATACGTACGA
Validation: Input sequence is valid.
Original Sequence: ATACGTACGA
Consensus Sequence: ATACGTACGA
Hamming Distance (Consensus): 0
```

Difference between any 2 nucleotides is defined as 1 hamming distance

Answer 2)

1. Read length: The read length is 1923780

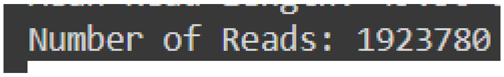
Output :



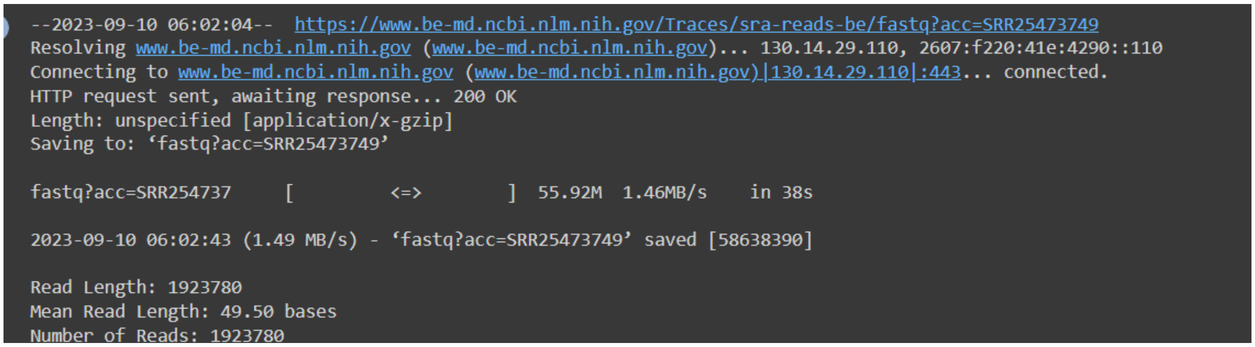
```
Read Length: 1923780
Mean Read Length: 49.50 bases
```

2. Number of reads: the total number of reads is same as the read length i.e. 1923780

output :



```
Number of Reads: 1923780
```



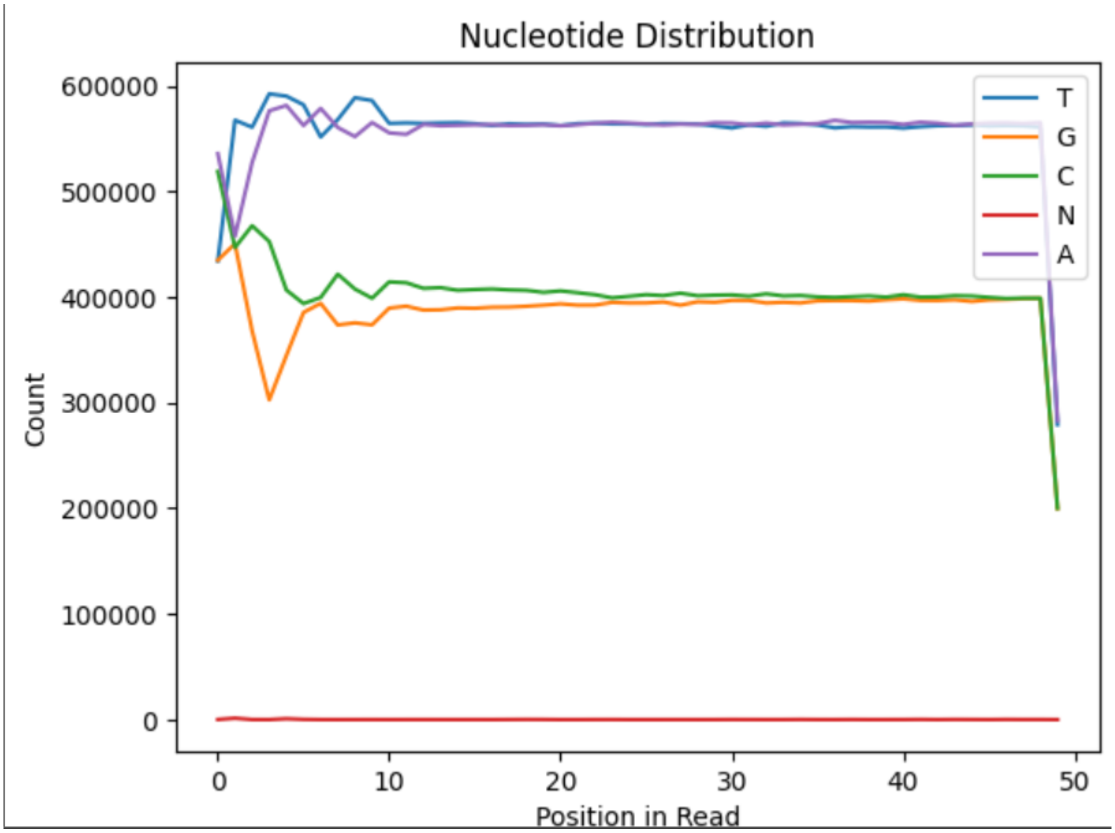
```
--2023-09-10 06:02:04--  https://www.be-md.ncbi.nlm.nih.gov/Traces/sra-reads-be/fastq?acc=SRR25473749
Resolving www.be-md.ncbi.nlm.nih.gov (www.be-md.ncbi.nlm.nih.gov)... 130.14.29.110, 2607:f220:41e:4290::110
Connecting to www.be-md.ncbi.nlm.nih.gov (www.be-md.ncbi.nlm.nih.gov)|130.14.29.110|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [application/x-gzip]
Saving to: ‘fastq?acc=SRR25473749’

fastq?acc=SRR254737     [         <=>        ]  55.92M  1.46MB/s    in 38s

2023-09-10 06:02:43 (1.49 MB/s) - ‘fastq?acc=SRR25473749’ saved [58638390]

Read Length: 1923780
Mean Read Length: 49.50 bases
Number of Reads: 1923780
```

3. The Plot nucleotide distribution across locations of the reads is shown in below:



The above graph shows the nucleotide distribution of the reads in the FASTQC file. The x-axis shows the position of the nucleotide in the read, and the y-axis shows the number of reads at that position. The most abundant nucleotide is A, followed by T, G, and C. There is a slight bias towards the beginning of the read, with more reads starting with A or T.

This bias could be due to a number of factors, such as the sequencing technology used or the biological sample being sequenced. It is important to investigate any biases in the data before analyzing it further.

i) The nucleotide distribution shows the number of reads at each position in the read.
ii) The most abundant nucleotide is A, followed by T, G, and C.
iii) There is a slight bias towards the beginning of the read, with more reads starting with A or T

Answer3)

(i) In the question we had to generate 100 random sequences of 1KB length each.
(ii) Then we had to plant a motif of length 10 and within that motif we can add 0,1,2 mutations

(iii) We Then apply gibbs sampling to identify motif locations or motifs which can be placed after (0/1/2) mutations and then we find the consensus motif which is basically the sequence which aligns with the dataset the most

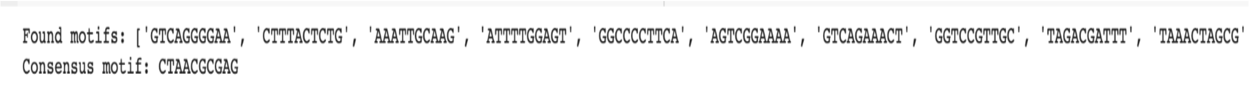
Gibbs sampler aims to iteratively adjust the motif positions within the sequences to find the set of motifs that maximizes their alignment within the dataset. This is achieved by probabilistically sampling new motif positions based on the current motifs and profiles. The algorithm continues until convergence or for a specified number of iterations, whichever comes first

In our code we apply 15 iterations which we felt is enough for meeting the convergence criteria

We basically go about the process of first initializing the motifs then we iterate for n number of times mentioned , then a random sequence from the dataset is selected and soon a profile matrix is created which is basically containing the score of the motifs

In the second part of the analysis we find the probabilities of the motifs using the profile matrix we have created , positions having a greater probability will be used in making the new motif which will be used for computation . We keep performing this process iteratively then until we aren't able to find any possible improvements in the code produced , basically meet the convergence criteria

These are the results which we get in one of the iterations



```
Found motifs: ['GTCAGGGGAA', 'CTTTACTCTG', 'AAATTGCAAG', 'ATTTTGGAGT', 'GGCCCCTTCA', 'AGTCGGAAAA', 'GTCAGAAACT', 'GGTCCGTTGC', 'TAGACGATTT', 'TAAACTAGCG'
Consensus motif: CTAACGCGAG
```