**N.Krishnateja-G01449836**

**Objective:**

To find the CpG islands or CG adjacencies with much higher frequencies of a DNA code taken from a human chromosome of a Parkinson diseased patient. The DNA code has been filed as *Human\_chromosome-1\_PARK7.fasta*.

**Analysis:**

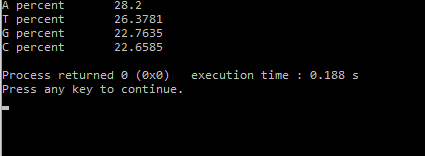
The analysis has been done in two levels. First, a preliminary level to observe the frequency of occurrence of all the four nucleotides in the DNA strand. Next, a secondary level designed and developed to find the CpG islands in the DNA code.

**Preliminary level:**

The DNA code file *Human\_chromosome-1\_PARK7.fasta* has been read and the nucleotide frequency for all the four nucleotides as follows:-

* Adenine: 28.2%age
* Thiamine: 26.378%age
* Guanine: 22.7635%age
* Cytosine: 22.65%age

The screenshot for the console output developed through C++ has been provided for the reference.



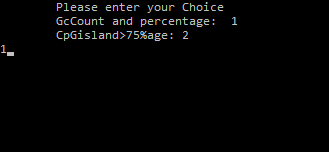
**Secondary Level:**

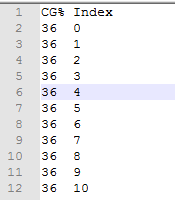
**Design:** The main idea of the design is to generate a text file that contains the Observed to Expected ratio of CG adjacencies and corresponding Window frame indices. In mammals, a Cpg island can be theoretically explained as, minimal base pair length of 200bp containing more than 50% of GC nucleotide frequency and average Observed/Expected ratio(O/E). O/E can be mathematically defined as

**O/E: (# CpG occurrence/(#C occurrence\*#G occrence))\*size of sequence**

**Programming:** The file *Human\_chromosome-1\_PARK7.fasta* contains the DNA code of person diseased with Parkinson disease has been read into C++ platform to proceed further. The file uniquely contains the DNA sequence with various number of lines being split. Hence, the primary motto is to make a single object of DNA code on appending all the lines in the file. Two hundred base pairs windows are assumed to find the CpG islands. Therefore, substrings for the DNA code string are formed and are considered as the window frame and looped throughout the DNA code string. Program has been customized to find the CpG ratio greater that 0.75.The program also codes for another text file to generate the text file for CG percent along with its window frame.

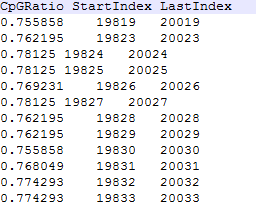
**Results:** When the application has been run, it asks for choices to select the type of text file.



When the selected choice has been 1,a text file with CpGpercent, FirstIndex has been generated.

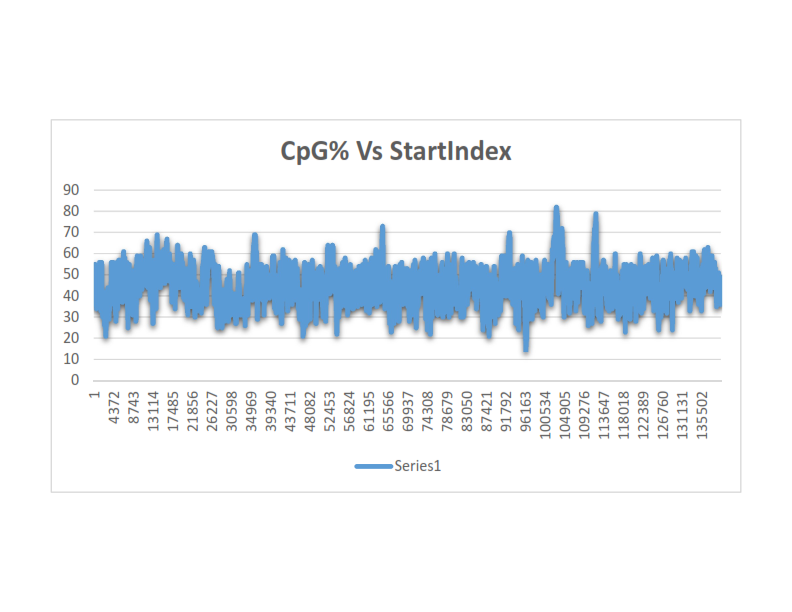
**Figure1:** sample of text file showing CGpercent and First Index

The output has been a generated text file with the tabulated values of CpgRatio, Starting index of the window frame, last index of the window frame when the slected choice is 2. Sample output has been shown below.



**Figure2** : showing the tabulated values from the generated text file.

**Discussion:** The file generated to observe the CG percent values has been made a pattern through excel and graph has been generated. The highest CG percent has been observed around 82.5% at 104905 index frame.



From the generated text, it has been observed that there are 1712 window frames with more than 0.75 as O/E of CpG ratio and GC percentage as more than 50%. However, the data collected has been graphically represented to observe the trends of the CpG islands.

