GP_lab1

March 4, 2020

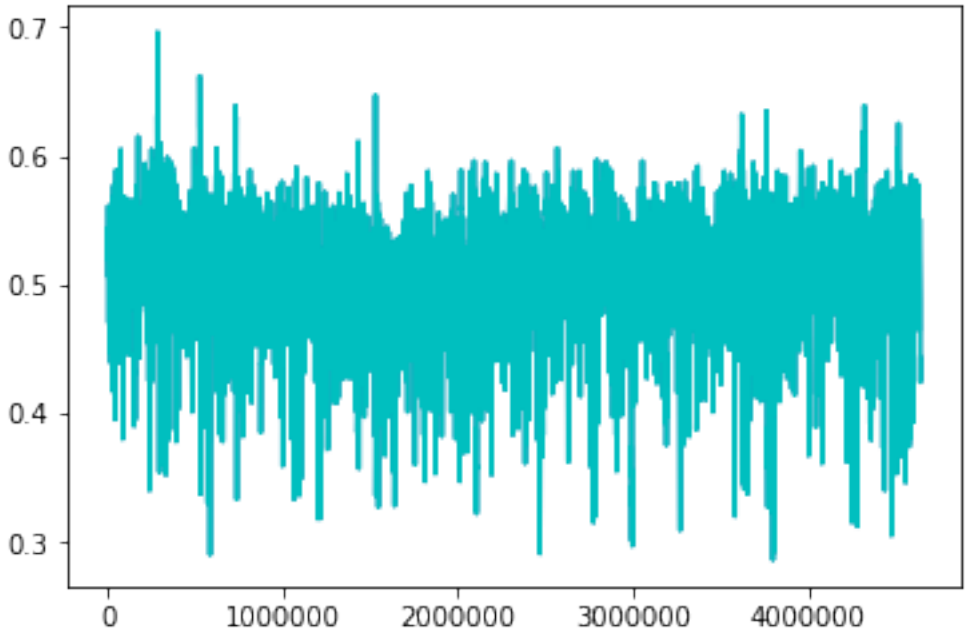
In [1]:
```
from collections import Counter
import matplotlib.pyplot as plt
```

In [2]:
```
class genome():
    def __init__(self, organism, sequence = None, content = None, iterations = None):
        self.sequence = sequence
        self.address = "E.coli K12\\" + organism + ".fna"
        self.content = content
        self.iterations = iterations

    #parser method parses given fasta file and puts it into
    #genome class attribute self.sequence.
    def parser(self):
        f = open(self.address)
        seq = ""

        for line in f:
            if line[0] != ">":
                seq += line.strip("\n")

        self.sequence = seq

        f.close()

    #gc_content_counter method computes GC content with regards
    #of window over which GC content is computed and step with which
    #window is moved forward along the given sequence.
    def gc_content_counter(self,window, step):
        seq = self.sequence
        iters = []
        content_lst = []

        for i in range(0, len(seq) - window, step):
            x = Counter(seq[i:i+window])
            content_lst.append((x['G'] + x['C'])/(x['A'] + x['C'] + x['T'] + x['G']))
            iters.append(i)

        self.content = content_lst
```

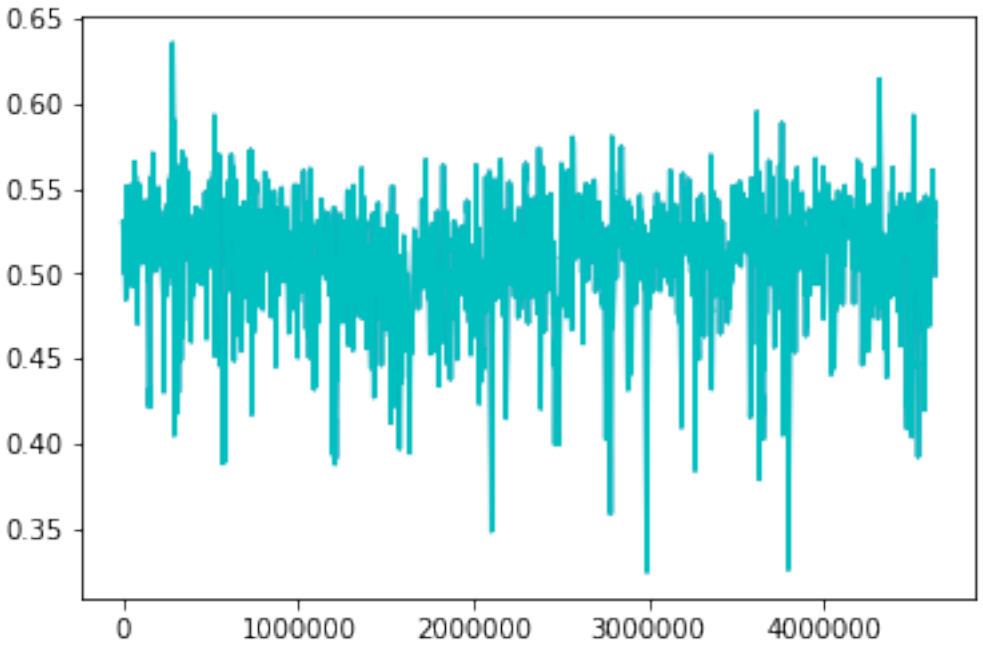
```
            self.iterations = iters

        #gc_plotter method plots GC content over the windows
        #over which GC content was computed.
        def gc_plotter(self):
            plt.plot(self.iterations, self.content)
            plt.plot(e_coli.iterations, e_coli.content, color = 'c')
            plt.show()
            plt.close()
```

In [3]:
```
e_coli = genome("E_coli_K12")
e_coli.parser()
```

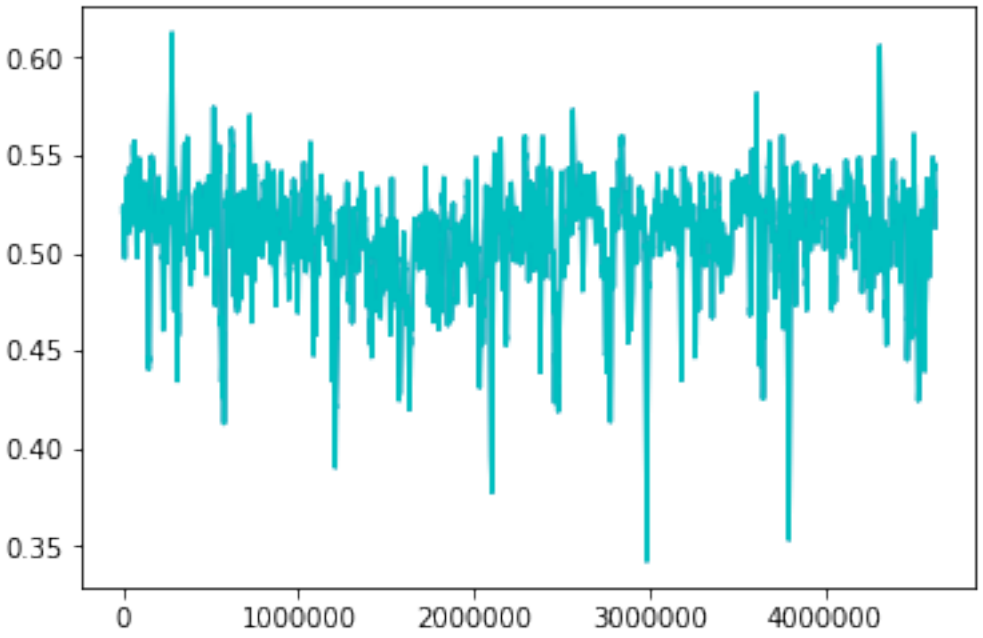
In [4]:
```
e_coli.gc_content_counter(1000,500)
e_coli.gc_plotter()
```



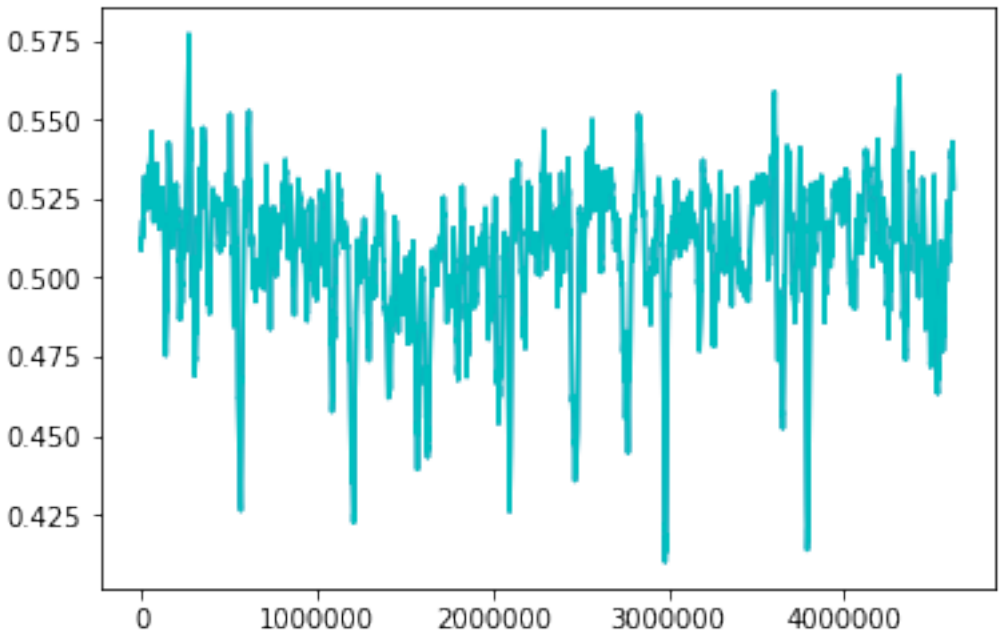
In [5]:
```
e_coli.gc_content_counter(5000,500)
e_coli.gc_plotter()
```



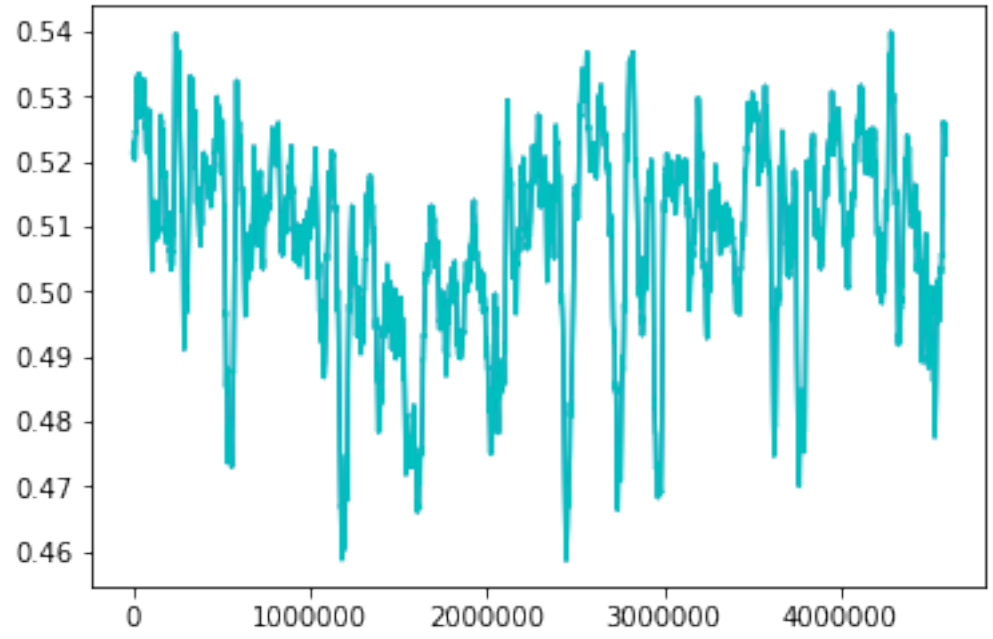
In [6]:
```
e_coli.gc_content_counter(10000,500)
e_coli.gc_plotter()
```



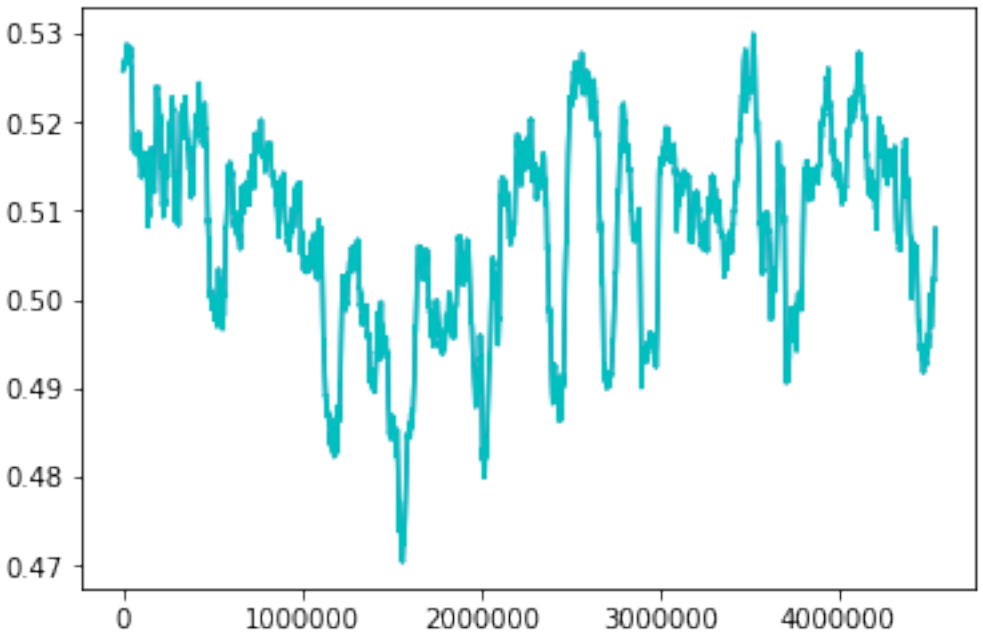
In [7]:
```
e_coli.gc_content_counter(20000,500)
e_coli.gc_plotter()
```



In [8]:
```
e_coli.gc_content_counter(50000,500)
e_coli.gc_plotter()
```



```
In [9]: e_coli.gc_content_counter(100000,500)
        e_coli.gc_plotter()
```



GC content is used to estimate average temperature of melting DNA strains in PCR reaction. The higher the GC content the higher the temperature.

GC content may also be used in taxonomy of prokaryota, e.g. Actinobacteria are classified as 'high GC-content bacteria'.