# PB Assignment 1 Readme

## DnaA Box Motif Analysis in E. coli K12 Genome

by Sidharth Kumar, 2023526

## Overview

This script fetches the **FASTA sequence** of the *E. coli* K12 genome from NCBI, analyzes its **AT richness**, and then searches for specific **DnaA box motifs** within the genome. It also compares the **AT richness of the motifs** with the overall genome to study their potential functional significance.

### Requirements:

### 1. Downloading the Genome/Fetching the Genome Sequence

```
from Bio import Entrez, SeqIO
from Bio.Seq import Seq
import os
Entrez.email = "sidharth@random.com"
file_path = "EcoliFasta"
```

- Uses **NCBI Entrez** (from BioPython) to fetch the genome.
- Specifies a random email (required for Entrez access).
- Defines a file path to store the retrieved **FASTA file**.

```
def fetch_fasta(record_id, output_file):
    try:
        handle = Entrez.efetch(db="nucleotide", id=record_id,
rettype="fasta", retmode="text")
        fasta_data = handle.read()
        handle.close()
        with open(output_file, "w") as file:
            file.write(fasta_data)

        print(f"FASTA for {record_id} saved to {output_file}")

    except Exception as e:
        print(f"Error fetching {record_id}: {e}")

fetch_fasta("NC_000913.3", file_path)
```

- Fetches the **E. coli K12 RefSeq (NC_000913.3)** genome in **FASTA format** and saves it.
- Handles errors in case of connection issues.

## 2. Calculating the AT Content of the Genome

```
def analyze_fasta(file_path):
    for record in SeqIO.parse(file_path, "fasta"):
        sequence = record.seq.upper()
        seq_length = len(sequence)

        a_count = sequence.count("A")
        t_count = sequence.count("T")

        at_richness = (a_count + t_count) / seq_length * 100 if seq_length > 0 else 0

        print(f"Record ID: {record.id}")
        print(f"Description: {record.description}")
        print(f"Length: {seq_length}")
        print(f"AT Richness: {at_richness:.2f}%")
        print("-" * 40)
        return at_richness
```

```
genomeAT = analyze_fasta(file_path)
```

- **Reads** the saved EcoliFASTA file and **calculates the AT percentage**.
- **Returns the genome-wide AT richness** for comparison with motifs.

## 3. Calculating AT Richness of a the Motifs

```
def motifRichness(motif):
    length = len(motif)
    cost = 0
    for c in motif:
        if c == 'A' or c == 'T':
            cost += 1
    return (cost / length) * 100 if length > 0 else 0
```

- Counts **A and T nucleic bases** in the given **motif**.
- Returns its **AT richness percentage**.

## 4. Finding Motif Occurrences in the Genome

```
def motifCheck(motif, file_path, genomeAT):
    with open(file_path, "r") as file:
        for record in SeqIO.parse(file, "fasta"):
            print(f"Currently checking {motif} motif")
            motifAT = motifRichness(motif)
            print(f"AT richness of this motif is {motifAT:.2
f}%")
            print(f"It has {(motifAT - genomeAT):.2f}% higher
AT content compared to the EColi K12 Genome")
            occurences = 0
            genome_seq = str(record.seq)
            for i in range(len(genome_seq) - len(motif) + 1):
                if genome_seq[i:i+len(motif)] == motif:
                    print(f"Motif found at position {i+1}")
                    occurences += 1
    print(f"This motif {motif} had {occurences} occurrences i
n the entire Genome")
    print("-" * 40)
```

- Reads the genome sequence.
- Finds **all occurrences of the motif** and **compares its AT richness with the whole genome**.
- Prints **motif locations** within the genome.

### 5. Running the Motif Search

```
motiflist = ["TTATACACA", "TTATTCACA", "TTATGCACA", "TTATCCAC
A"]
for motif in motiflist:
    motifCheck(motif, file_path, genomeAT)

print("Successfully Executed")
```

- **The list of used DnaA box motifs**.
- **Checks occurrences of each motif** in the genome.
- Prints locations and richness and execution confirmation.

## Expected Output

```
FASTA for NC_000913.3 saved to EcoliFasta
Record ID: NC_000913.3
Description: NC_000913.3 Escherichia coli str. K-12 substr. M
G1655, complete genome
Length: 4641652
AT Richness: 49.21%
----------------------------------------
Currently checking TTATACACA motif
AT richness of this motif is 77.78%
```

```
It has 28.57% higher AT content compared to the EColi K12 Gen
ome
Motif found at position 997280
Motif found at position 3808286
Motif found at position 3925907
Motif found at position 4326803
This motif TTATACACA had 4 occurences in the entire Genome
----------------------------------------
Currently checking TTATTCACA motif
AT richness of this motif is 77.78%
It has 28.57% higher AT content compared to the EColi K12 Gen
ome
Motif found at position 336634
Motif found at position 1482914
Motif found at position 1848700
Motif found at position 1935716
Motif found at position 2339506
Motif found at position 3015365
Motif found at position 3361018
Motif found at position 3932292
Motif found at position 4392906
Motif found at position 4538113
This motif TTATTCACA had 10 occurences in the entire Genome
----------------------------------------
Currently checking TTATGCACA motif
AT richness of this motif is 66.67%
It has 17.46% higher AT content compared to the EColi K12 Gen
ome
Motif found at position 167570
Motif found at position 323427
Motif found at position 338821
Motif found at position 551861
Motif found at position 592793
Motif found at position 672664
Motif found at position 719397
Motif found at position 1538184
Motif found at position 1784091
Motif found at position 1865694
Motif found at position 2168242
Motif found at position 2624396
Motif found at position 2859365
Motif found at position 2863210
Motif found at position 2931164
Motif found at position 3020429
Motif found at position 3941623
Motif found at position 3974832
Motif found at position 4035346
Motif found at position 4046822
Motif found at position 4166474
Motif found at position 4259771
Motif found at position 4263933
This motif TTATGCACA had 23 occurences in the entire Genome
----------------------------------------
Currently checking TTATCCACA motif
AT richness of this motif is 66.67%
It has 17.46% higher AT content compared to the EColi K12 Genome
Motif found at position 984232
Motif found at position 1478112
Motif found at position 1526906
Motif found at position 1544607
Motif found at position 2304000
Motif found at position 2331359
Motif found at position 2344703
Motif found at position 2969346
Motif found at position 2995468
Motif found at position 3105553
Motif found at position 3181557
Motif found at position 3204645
Motif found at position 3333933
Motif found at position 3600890
Motif found at position 3664728
Motif found at position 3883939
Motif found at position 3925981
Motif found at position 4046834
Motif found at position 4313475
Motif found at position 4392744
Motif found at position 4462687
Motif found at position 4462764
This motif TTATCCACA had 22 occurences in the entire Genome
----------------------------------------
Successfully Executed
```

## Conclusion

This script **fetches, analyzes, and searches for DnaA box motifs** in *E. coli* K12 genome. By comparing AT content between motifs and the genome, we can infer why these motifs might play a role in DNA replication initiation. The results can be useful for studying **replication origins** and **bacterial genome organization**.