High throughput

The second homework

Name: Chengzhi Dong

Student ID: 19331027

1 Code

1.1 Directory operations

1. Creat a directory named learning_python.

```
mkdir learning_python
```

2. Switch to learning_python directory

```
cd learning_python
```

3. Put the chrM.fa file in this directory.

```
wget http://222.200.186.169/highput/data/chrM.fa
```

4. Open vim editor with *vim count.py*

```
vim count.py
```

1.2 Python in biology

Edit your python code in vim editor, then save and exit. The program should include the following:

1. Calculate the percentage of each base of ATCG in the chrM chromosome.

```
#Question 1
def percent_Cal_1(DNA_seq):
    count_A = 0
    count_T = 0
    count_C = 0
    count_G = 0
    for i in DNA_seq:
        if i == "A":
            count_A += 1
        elif i == "T":
            count_T += 1
        elif i == "C":
            count_C += 1
        elif i == "G":
            count_G += 1
    all_count = len(DNA_seq)
    precent_A = count_A*100/all_count
    precent_T = count_T*100/all_count
    precent_C = count_C*100/all_count
```

```
    precent_G = count_G*100/all_count

    print("A | Frequency:{0} | Percentage:{1:.2f}%".format(count_A,precent_A))
    print("T | Frequency:{0} | Percentage:{1:.2f}%".format(count_T,precent_T))
    print("C | Frequency:{0} | Percentage:{1:.2f}%".format(count_C,precent_C))
    print("G | Frequency:{0} | Percentage:{1:.2f}%".format(count_G,precent_G))

def percent_Cal_2(DNA_seq):
    count_ls = set(DNA_seq)
    all_count = len(DNA_seq)/100
    for x in count_ls:
        frequency = DNA_seq.count(x)
        percent = frequency/all_count
        print("{0} | Frequency:{1} | Percentage:{2:.2f}%".format(x,frequency,percent))

def percent_Cal_3(DNA_seq):
    count_ls = dict()
    all_count = len(DNA_seq)/100
    for x in DNA_seq:
        if x in count_ls:
            count_ls[x] += 1
        else:
            count_ls[x] = 1
    for key in count_ls:
        count = count_ls[key]
        percent = count/all_count
        print("{0} | Frequency:{1} | Percentage:{2:.2f}%".format(key,count,percent))


def text_to_DNA(text_file_name):
    text_file = open(text_file_name)
    lines = text_file.readlines()
    text_file.close()
    line = [x.strip() for x in lines[1:]]
    lines = ''.join(line)
    DNA_seq = lines.upper()
    return DNA_seq

text_file_name = "chrM.fa"
DNA_seq = text_to_DNA(text_file_name)
print("Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by
    Funtion1.1:")
percent_Cal_1(DNA_seq)
print("Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by by
    Funtion1.2:")
percent_Cal_2(DNA_seq)
print("Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by by
    Funtion1.3:")
percent_Cal_3(DNA_seq)
```

2. Calculate insulin = "GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT" amino acid frequency in insulin sequence.

```
#Question 2
insulin = "GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT"
print("Question2: Calculate insulin = GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT amino
     acid frequency in insulin sequence by Funtion1.2:")
percent_Cal_2(insulin)
print("Question2: Calculate insulin = GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT amino
     acid frequency in insulin sequence by Funtion1.3:")
percent_Cal_3(insulin)
```

3. Run 5bp windows in the sequence "PRQTEINSEQWENCE" and count the number of occurrences of each window in the sequence.

```
#Question 3
def percent_Cal_3_1(DNA_seq):
    count_ls = dict()
    for i in range(len(DNA_seq)-4):
        seq_short = DNA_seq[i:i+5]
        if seq_short in count_ls:
            count_ls[seq_short] += 1
        else:
            count_ls[seq_short] = 1
    for key in count_ls:
        count = count_ls[key]
        print("{0} | Count:{1}".format(key,count))

print("Question3: Run 5bp windows in the sequence PRQTEINSEQWENCE and count the number of
    occurrences of each window in the sequence by Funtion3.1:")
sequence = "PRQTEINSEQWENCE"
percent_Cal_3_1(sequence)
```

4. Calculate GC percentage in chrM.fa

```
# Question4
print("Question4: Calculate GC percentage in chrM.fa:")
count_GC = (DNA_seq.count("G")+DNA_seq.count("C"))/len(DNA_seq)
print("GC percentage in chrM: {0:.2f}".format(count_GC))
```

5. Use *python3 count.py* to run your program from the command line

```
python3 count.py
```

2 Result

1. Question1

```
Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by Funtion1.1:
A | Frequency:5113 | Percentage:30.86%
T | Frequency:4086 | Percentage:24.66%
C | Frequency:5192 | Percentage:31.33%
G | Frequency:2180 | Percentage:13.16%
Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by by
Funtion1.2:
A | Frequency:5113 | Percentage:30.86%
T | Frequency:4086 | Percentage:24.66%
G | Frequency:2180 | Percentage:13.16%
C | Frequency:5192 | Percentage:31.33%
Question1: Calculate the percentage of each base of ATCG in the chrM chromosome by by
Funtion1.3:
G | Frequency:2180 | Percentage:13.16%
A | Frequency:5113 | Percentage:30.86%
T | Frequency:4086 | Percentage:24.66%
C | Frequency:5192 | Percentage:31.33%
```

图 1: The percentage of each base of ATCG in the chrM chromosome.

2. Question2

```
Question2: Calculate insulin = GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT amino acid
frequency in insulin sequence by Funtion1.3:
G | Frequency:4 | Percentage:7.84%
I | Frequency:2 | Percentage:3.92%
V | Frequency:4 | Percentage:7.84%
E | Frequency:4 | Percentage:7.84%
Q | Frequency:3 | Percentage:5.88%
C | Frequency:6 | Percentage:11.76%
T | Frequency:3 | Percentage:5.88%
S | Frequency:3 | Percentage:5.88%
L | Frequency:6 | Percentage:11.76%
Y | Frequency:4 | Percentage:7.84%
N | Frequency:3 | Percentage:5.88%
F | Frequency:3 | Percentage:5.88%
H | Frequency:2 | Percentage:3.92%
A | Frequency:1 | Percentage:1.96%
R | Frequency:1 | Percentage:1.96%
P | Frequency:1 | Percentage:1.96%
K | Frequency:1 | Percentage:1.96%
```

图 2: Insulin amino acid frequency in insulin sequence.

3. Question3

```
Question3: Calculate insulin = GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKT amino acid
frequency in insulin sequence by Funtion3.1:
PRQTE | Count:1
RQTEI | Count:1
QTEIN | Count:1
TEINS | Count:1
EINSE | Count:1
INSEQ | Count:1
NSEQW | Count:1
SEQWE | Count:1
EQWEN | Count:1
QWENC | Count:1
WENCE | Count:1
```

图 3: The number of occurrences of each window in the sequence.

4. Question4

```
Question4: Calculate GC percentage in chrM.fa:
GC percentage in chrM: 0.44
```

图 4: GC percentage in chrM.fa.