Megan Chu
A12814536
Assignment 4
5/29/18

Q1.

**Program Output:**

```
Running Q1(dinucelotide frequency)...
Total length of sequence:
        1989980
Observed frequency of each base:
        A       C       G       T
        0.2736  0.2278  0.2264  0.2722
Calculated Expected Frequency of each dinucleotide:
        Note that rows represent 1st base and columns represent 2nd base
                A       C       G       T
        A       0.0749  0.0623  0.062   0.0745
        C       0.0623  0.0519  0.0516  0.062
        G       0.062   0.0516  0.0513  0.0616
        T       0.0745  0.062   0.0616  0.0616
Observed Frequency of each dinucleotide:
        Note that rows represent 1st base and columns represent 2nd base
                A       C       G       T
        A       0.0834  0.0533  0.0711  0.0657
        C       0.078   0.0626  0.0161  0.0711
        G       0.0607  0.0511  0.0618  0.0528
        T       0.0514  0.0607  0.0775  0.0826
Observed/Expected frequencies for dinucleotides:
        Note that rows represent 1st base and columns represent 2nd base
                A       C       G       T
        A       1.1143  0.8559  1.1477  0.8828
        C       1.2521  1.2066  0.3117  1.1464
        G       0.9804  0.9905  1.2053  0.8568
        T       0.6905  0.9799  1.2567  1.3397
```

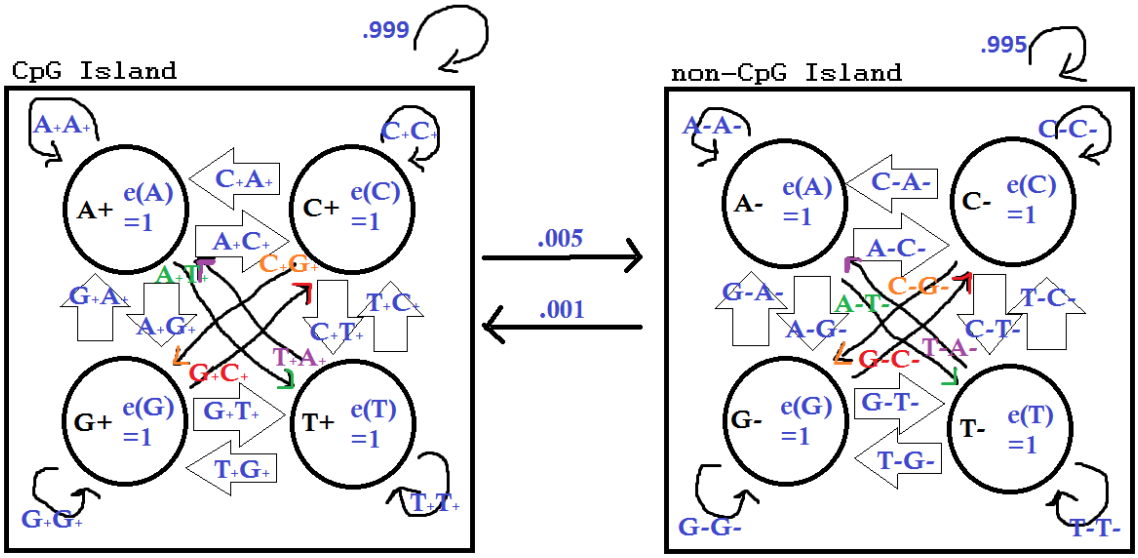
**Answer:**

Note that we see an observed/expected frequency that is the least (.3117) for the CG dinucleotide (this means the observed frequency is much less than the expected). We see an observed/expected frequency that is the highest (1.3397) for the TT dinucleotide (this means the observed frequency is much more than the expected). Ranking dinucleotides from a small difference between observed and expected frequencies (observed/expected close to 1) to a big difference between observed and expected frequencies (observed/expected not close to 1), we have GC(.9905), GA(.9804), TC(.9799), AA(1.1143), AT(.8828), GT(.8568), AC(.8559), CT(1.1464), AG(1.1477), GG(1.2053), CC(1.2066), CA(1.2521), TG(1.2567), TA(.6905), TT(1.3397), and CG(.3117).

Q2.

**Rough Sketch of HMM Designed(Note that transitions between + and – states are not shown in this sketch for simplicity, they will be defined at the end):**

T[j,k] is represented by jk on the diagram for simplicity. For example, the arrow labelled T+C+ on the diagram represents T[T+,C+]. Also note that within the CpG Island and non-CpG Island are 2 hidden states that encompass 4 smaller states within them.




Q = {A+, C+, G+, T+, A-, C-, G-, T-}

Note that all + states are CpG Island and all – states are non-CpG Island.

∑ = {A, C, G, T}

Π = 0 for all states in CpG island and 1 for non-CpG island

$e_j(S)$ table is shown below (Note that the rows represent S or the set of symbols ∑, and the columns represent j or the set of states Q).

| | A+ | C+ | G+ | T+ | A- | C- | G- | T- |
|---|---|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| G | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| T | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

T[j,k] is the probability of transitioning from state j to state k.

Within the CpG Island States:

We know P[CG] = .06 and all other dinucleotides equally likely thus P[all other dinucleotides] = (1-.06)/(16-1) = .94/15. Thus,

T[C+,G+] = Pr(CG)/[Pr(CA)+Pr(CC)+Pr(CG)+Pr(CT)+Pr(Escape starting from C+)] = .06/[.94/15*3 + .06 + .005/4] = .24072216649.

T[C+,Y+] = (.94/15)/[.94/15*3 + .06 + .005/4] = .25142092945

T[X+,Y+] = (.94/15)/[.94/15*4 + .005/4] = .24875951042

Table for T[j,k] is shown below (Note that the rows represent j and the columns represent k).

Within the non-CpG Island States:

We know P[CG] = .01 and all other dinucleotides equally likely thus P[all other dinucleotides = (1-.01)/(16-1) = .066. Thus,

T[C-,G-] = Pr(CG)/[Pr(CA)+Pr(CC)+Pr(CG)+Pr(CT)+Pr(Escape starting from C-)] = .01/[.066*3 + .01 + .001/4] = .04801920768.

T[C-,Y-] = .066/[.066*3 + .01 + .001/4] = .3169267707

T[X-,Y-] = .066/[.066*4 + .001/4] = .24976348155

From a CpG island state, j, to a non-CpG Island state, k:

We know transitioning from a CpG island state to non-CpG Island has probability .005. We also know that there are 16 possible such transitions((4 CpG states)*(4 non-CpG states)). Thus the transition probability of each is .005/16 = 0.0003125.

From a non-CpG island state, j, to a CpG Island state, k:

We know transitioning from a non-CpG island state to CpG Island has probability .001. We also know that there are 16 possible such transitions((4 non-CpG states)*(4 CpG states)). Thus the transition probability of each is .001/16 = 0.0000625.

Table for T[j,k] is shown below. Note that the rows represent the first state, j, and the columns represent the second state, k.

| | A+ | C+ | G+ | T+ | A- | C- | G- | T- |
|---|---|---|---|---|---|---|---|---|
| A+ | .2487595 1042 | .2487595 1042 | .2487595 1042 | .2487595 1042 | 0.000312 5 | 0.000312 5 | 0.000312 5 | 0.000312 5 |
| C+ | **.2514209 2945** | **.2514209 2945** | **.2407221 6649** | **.2514209 2945** | 0.000312 5 | 0.000312 5 | 0.000312 5 | 0.000312 5 |
| G+ | .2487595 1042 | .2487595 1042 | .2487595 1042 | .2487595 1042 | 0.000312 5 | 0.000312 5 | 0.000312 5 | 0.000312 5 |
| T+ | .2487595 1042 | .2487595 1042 | .2487595 1042 | .2487595 1042 | 0.000312 5 | 0.000312 5 | 0.000312 5 | 0.000312 5 |
| A- | 0.000062 5 | 0.000062 5 | 0.000062 5 | 0.000062 5 | .2497634 8155 | .2497634 8155 | .2497634 8155 | .2497634 8155 |
| C- | 0.000062 5 | 0.000062 5 | 0.000062 5 | 0.000062 5 | **.3169267 707** | **.3169267 707** | **.0480192 0768** | **.3169267 707** |
| G- | 0.000062 5 | 0.000062 5 | 0.000062 5 | 0.000062 5 | .2497634 8155 | .2497634 8155 | .2497634 8155 | .2497634 8155 |
| T- | 0.000062 5 | 0.000062 5 | 0.000062 5 | 0.000062 5 | .2497634 8155 | .2497634 8155 | .2497634 8155 | .2497634 8155 |

Q3.

**Program Output Snippet:**

*Note that actual output is too long to include, thus it is included with the code in a file named Q3_CpG.txt. Note that there is also a file named Q3_T.txt with the code that contains the matrix of Transition probabilities needed for the Viterbi algorithm in Q3.py.

```
Running Q3(identify CpG Islands)...
575 1954
35491 37580
116231 117208
161490 161714
181249 181533
187400 188553
203992 204400
219431 219702
…
Number of CpG Islands Identified: 99
```

Q4.

**Program Output:**

```
Running Q4(Calculate Transformation and Emission Probabilities)...

nII: 140104
nOO: 1859576
nIO: 159
nOI: 159
T[I,I]: nII/(nII+nIO) = 0.9988664152342386
T[O,O]: nOO/(nOO+nOI) = 0.9999145039481432
T[I,O]: nIO/(nIO+nII) = 0.0011335847657614625
T[O,I]: nOI/(nOI+nOO) = 8.549605185685057e-05
T[I,O]/16: 7.08490478600914e-05
T[O,I]/16: 5.343503241053161e-06
T[j,k] matrices, Note that rows are first state, j, columns are second
state, k.
```

| | A+ | C+ | G+ | T+ |
|---|---|---|---|---|
| A+ | 0.1425893033353069 | 0.3249457272547859 | 0.4333432011051904 | 0.09951647917900139 |
| C+ | 0.17015058651621545 | 0.36875431261923125 | 0.28100012176807243 | 0.18074440881600845 |
| G+ | 0.15171259489509237 | 0.34255882997589016 | 0.3744196704325821 | 0.13089046964353318 |
| T+ | 0.0672056102074608 | 0.3567741307100419 | 0.4269504236875426 | 0.14814454076166358 |

| | A- | C- | G- | T- |
|---|---|---|---|---|
| A- | 0.28267025343338525 | 0.21583883575954935 | 0.2931860013128596 | 0.2082856027355532 |
| C- | 0.3038218549670188 | 0.3078922824357573 | 0.10162141859356075 | 0.28659692006431653 |
| G- | 0.2369409667735553 | 0.248074725387187 | 0.3099649484887767 | 0.20506413317868596 |
| T- | 0.1759107420631134 | 0.24636912541689873 | 0.308029086939059 | 0.26973320699701986 |

Because there are not that many transitions between island and non-island, it may not be possible (also not very accurate) to calculate the transition probabilities from island to non-island and from non-island to island for every nucleotide. Thus, we will use T[I,O]/16 to estimate the probability of going from island to non-island and T[O,I]/16 to estimate the probability of going from non-island to island.

Here is a table of the compiled transition probabilities that are estimated from chrA.fasta. Note that rows represent the first state, j, and columns represent the second state, k.

| | A+ | C+ | G+ | T+ | A- | C- | G- | T- |
|---|---|---|---|---|---|---|---|---|
| A + | 0.142589303335306969 | 0.324945727254785959 | 0.433343201105190404 | 0.099516479179001390 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 |
| C + | 0.170150586516215450 | 0.368754312619231250 | 0.281000121768072430 | 0.180744408816008450 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 |
| G + | 0.151712594895092370 | 0.342558829975890160 | 0.374419670432582100 | 0.130890469643533180 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 |
| T + | 0.067205610207460800 | 0.356774130710041900 | 0.426950423687542600 | 0.148144540761663580 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 | 7.08490478600914e-05 |
| A - | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 0.282670253433385250 | 0.215838835759549350 | 0.293186001312859600 | 0.208285602735553200 |
| C - | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 0.303821854967018800 | 0.307892282435757300 | 0.101621418593560750 | 0.286596920064316530 |
| G - | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 0.236940966773555300 | 0.248074725387187000 | 0.309964948488776700 | 0.205064133178685960 |
| T - | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 5.34350324105316e-06 | 0.175910742063113400 | 0.246369125416898730 | 0.308029086939059000 | 0.269733206997019860 |

Since we designed our HMM such that emission probabilities were either 1 or 0, the emission probabilities for chrA are the same as what we had in Q3. It is listed in the table below for convenience.

$e_j(S)$ table is shown below (Note that the rows represent S or the set of symbols $\sum$, and the columns represent j or the set of states Q).

| | A+ | C+ | G+ | T+ | A- | C- | G- | T- |
|---|---|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| C | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| G | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| T | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |

Q5.

**Program Output Snippet:**
*Note that actual output is too long to include, thus it is included with the code in a file named Q5_CpG.txt. Note that there is also a file named Q5_T.txt generated by Q4.py submitted with the code that contains the matrix of Transition probabilities needed for the Viterbi algorithm in Q5.py.

```
Running Q5(identify CpG Islands with estimated parameters)...
635 1930
35340 36944
37064 37222
80796 81320
116228 117126
181241 181858
187572 188519
204026 204628
…
Number of CpG Islands Identified: 97
Number Shared CpG Islands: 50
Number in Q3 but not Q5: 49
Number in Q5 but not Q3: 47
```

As seen, there are quite a few shared CpG islands between Q3 and Q5. I counted any two islands that had even 1 nucleotide overlap as "shared". Just from visual checking, most of the "shared" islands had at least a couple hundred nucleotide overlaps (I wouldn't worry about counting two islands with only 1 nucleotide overlap since there seem to be none in the two output files). If one large island in Q5 was split into two smaller islands in Q3 I only counted that as 1 shared island. Although there many shared islands, there are almost as many different islands. We see the Q3 result has 49 islands that are have no overlap with islands in Q5, and the Q5 result has 47 islands that have no overlap with islands in Q3. Thus I would conclude that the Q5 results are very different compared to the labels obtained in Q3 (I would consider the results similar only if above 95% of the CpG islands were "shared" or had some overlap). I conclude that by slightly changing the transition probabilities one can get a large variance in CpG islands locations compared to before.