Jisoo Kim
BIMM 182
Professor Vineet Bafna
Jun 5, 2023

# Assignment 4

## Problem Set I [1]

Code available on **Question1.py**

Dinucleotide frequency table:

| | Expected | Observed | Diff |
|---|---|---|---|
| AA | 0.0572 | 0.0534 | -0.0038 |
| AT | 0.0554 | 0.0535 | -0.0019 |
| AC | 0.0616 | 0.0578 | -0.0038 |
| AG | 0.0611 | 0.0784 | 0.0173 |
| TA | 0.0554 | 0.0435 | -0.0119 |
| TT | 0.0537 | 0.0497 | -0.0040 |
| TC | 0.0596 | 0.0639 | 0.0043 |
| TG | 0.0592 | 0.0796 | 0.0204 |
| CA | 0.0616 | 0.0822 | 0.0206 |
| CT | 0.0596 | 0.0781 | 0.0185 |
| CC | 0.0662 | 0.0675 | 0.0013 |
| CG | 0.0657 | 0.0334 | -0.0323 |
| GA | 0.0611 | 0.0640 | 0.0029 |
| GT | 0.0592 | 0.0554 | -0.0038 |
| GC | 0.0657 | 0.0720 | 0.0063 |
| GG | 0.0652 | 0.0676 | 0.0024 |

Most of the differences in frequency between the observed and expected are less than 0.1-3%. Yet, we can notice that there are some Observed frequencies that show more than 1% or even 2% difference and 1-2% of a sequence with total length of 2,033,334 can have a significant impact on our interpretation.

## Problem Set I [2]

Code available on **Question2.py**

| CPG Markov Model | Non CPG Markov Model |
|---|---|
| AA 0.0509<br>AT 0.0510<br>AC 0.0574<br>AG 0.0778<br>TA 0.0414<br>TT 0.0476<br>TC 0.0634<br>TG 0.0789<br>CA 0.0811<br>CT 0.0775<br>CC 0.0696<br>CG 0.0383<br>GA 0.0638<br>GT 0.0551<br>GC 0.0761<br>GG 0.0700 | AA 0.0558<br>AT 0.0564<br>AC 0.0584<br>AG 0.0793<br>TA 0.0460<br>TT 0.0519<br>TC 0.0644<br>TG 0.0806<br>CA 0.0836<br>CT 0.0788<br>CC 0.0646<br>CG 0.0280<br>GA 0.0645<br>GT 0.0558<br>GC 0.0675<br>GG 0.0645 |

CpG region statistics are logged in CPG_stat.txt based on the Markov models above and calculating the CpG potential score. The first five CpG islands can be previewed below.

## Problem Set II [1]

Code available on Question3.py

Results are produced via **SLAMMER_isotopeProfile.csv**. Preview of P0 to P5 is shown below

SLAMMER_isotopeProfile

| | Peak Coefficient |
|---|---|
| 0 | 0.5795101428689200 |
| 1 | 0.2375219423431770 |
| 2 | 0.1201035816922640 |
| 3 | 0.035515000227235100 |
| 4 | 0.009713431734881780 |
| 5 | 0.002121215881559460 |

Pseudocode of Isotope Profile Calculation:

**Abundance** = Given list of abundance for each element in a dictionary
AND

**Amino** = list of each amino symbol molecular formula in the order of C,H,N,O,S in a dictionary,

peptide = 'SLAMMER'
**Peptide_formula** = []
for p in peptide:
    Find molecular of **p** from **Amino** and add to **Peptide_formula**

With the complete molecular formula of peptide:
Final_coeff = 1.0
for ith atom in **Peptide_formula**:
    **Amino_abundance** = Find abundance of that atom and set 0 if any abundance is less than 0.1

    Coeff = np.poly1d(**Amino_abundance**)****Peptide_formula[i]**
    **make sure highest number isotope have highest order
    **Final_coeff = Final_coeff*Coeff**

Return pd.DataFrame(Final_coeff, columns = ['Peak Coefficient']

Problem Set II [2]

C isotope abundance is adjust from C-12 (99.93) and C-13 (1.07) to C-12(49.465) and C-13(50.535)

Code available on **Question4.py**

Results are produced via **SLAMMER_isotopeProfile_0.5_C-12.csv**. Preview of P0 to P5 is shown below

SLAMMER_isotopeProfile_

| | Peak Coefficient |
|---|---|
| 0 | 6.74638597840582E-11 |
| 1 | 2.27803770187387E-09 |
| 2 | 3.73076181868566E-08 |
| 3 | 3.94746589055672E-07 |
| 4 | 3.03285068608089E-06 |
| 5 | 1.80290839723234E-05 |