**SOURCE CODE**

```
'''
Author: Nam Than
Date: 02/07/2021
UT Austin
Bioinformatics

Problems 1-5
'''
import numpy as np
import matplotlib.pyplot as plt

known_nuc = {'A', 'G', 'T', 'C'}  # Define a non-duplicated set of nuc


def nuc_freq_cal(name):
    nuc_dict = {}
    file = open(name, 'r')
    for line in file:
        line = line.strip('\r\n')
        for let in line:
            if let not in known_nuc:
                pass
            else:
                if let not in nuc_dict:
                    nuc_dict[let] = 1
                else:
                    nuc_dict[let] += 1
    file.close()
    seq_len = sum(nuc_dict.values())  # Sum of all counts in the dict
    nuc_freq = {}
    for nuc in known_nuc:
        nuc_freq[nuc] = nuc_dict[nuc]/seq_len
    file = open('Nuc_freq_'+name, 'w')
    for nuc in nuc_freq:
        file.write('The nucleotide {0} appears {1} times or {2} %.\n'.
                   format(nuc, nuc_dict[nuc], 100.0 * nuc_freq[nuc]))
    file.close()
    return nuc_dict, nuc_freq

def dinuc_freq_cal(name):
    dinuc_dict = {}
    long_line = ''
    file = open(name, 'r')
    # Make a long line of all nuc
    for line in file:
        long_line += line.strip('\r\n')
    # Remove unwanted nuc
    all_nuc = set(long_line)
    unwanted_nuc = all_nuc.difference(known_nuc)  # Find the wrong nuc
    for i in unwanted_nuc:
        long_line = long_line.replace(i,'')
    # Check for dinucleotides
    for i in range(len(long_line)-1):
        dinuc = long_line[i] + long_line[i+1]
        if dinuc not in dinuc_dict:
            dinuc_dict[dinuc] = 1
        else:
            dinuc_dict[dinuc] += 1
    file.close()
    seq_len = sum(dinuc_dict.values())
    dinuc_freq = {}
    for nuc in dinuc_dict:
        dinuc_freq[nuc] = dinuc_dict[nuc]/seq_len
    file = open('Dinuc_freq_'+name, 'w')
    for dinuc in dinuc_freq:
```

```
        file.write('The dinucleotide {0} appears {1} times or {2} %.\n'.
                   format(dinuc, dinuc_dict[dinuc], 100.0 * dinuc_freq[dinuc]))
    file.close()
    return dinuc_dict, dinuc_freq


def expected_freq_cal(nuc_freq: dict):
    expected_freq = {}
    for first_nuc in nuc_freq.keys():
        for second_nuc in nuc_freq:
            dinuc = first_nuc + second_nuc
            probability = nuc_freq[first_nuc] * nuc_freq[second_nuc]
            expected_freq[dinuc] = probability
    return expected_freq


def plot_freq(freq_1: dict, freq_2: dict, label_1: str, label_2: str):
    # Rearrange so keys line up
    sorted_key_list = sorted(freq_1.keys())
    sorted_freq_1 = {}
    sorted_freq_2 = {}
    for key in sorted_key_list:
        sorted_freq_1[key] = freq_1[key]
        sorted_freq_2[key] = freq_2[key]
    # Plot
    x_axis = np.arange(len(sorted_key_list))
    fig = plt.subplot(111)
    fig.bar(x_axis, sorted_freq_1.values(), width=0.2, color='g',
            align='center')
    fig.bar(x_axis+0.2, sorted_freq_2.values(), width=0.2, color='b',
            align='center')
    fig.legend((label_1, label_2))  # arg is a tuple
    plt.title('Frequency comparison', fontsize=15)
    plt.xticks(x_axis, sorted_key_list)
    plt.savefig('{0}_{1}_freq_barplot'.format(label_1, label_2))
    plt.close()

def main():
    file_dir = ['Hinfluenzae.txt',
                'Taquaticus.txt',
                'MysteryGene1.txt',
                'MysteryGene2.txt',
                'MysteryGene3.txt']
    # Problem 1:
    print('Problem 1')
    HF_nuc_freq = nuc_freq_cal(file_dir[0])[1]
    TA_nuc_freq = nuc_freq_cal(file_dir[1])[1]

    # Problem 2:
    print('Problem 2')
    HF_dinuc_freq = dinuc_freq_cal(file_dir[0])[1]

    # Problem 3:
    print('Problem 3')
    TA_dinuc_freq = dinuc_freq_cal(file_dir[1])[1]

    # Problem 4:
    print('Problem 4')
    expected_dinuc_freq = expected_freq_cal(HF_nuc_freq)
    plot_freq(HF_dinuc_freq, expected_dinuc_freq, 'observed', 'expected')

    # Problem 5
    print('Problem 5')
    myst_1 = dinuc_freq_cal(file_dir[2])[1]
    plot_freq(myst_1, HF_dinuc_freq, 'myst_1', 'HF')
    plot_freq(myst_1, TA_dinuc_freq, 'myst_1', 'TA')
```

```
    myst_2 = dinuc_freq_cal(file_dir[3])[1]
    plot_freq(myst_2, HF_dinuc_freq, 'myst_2', 'HF')
    plot_freq(myst_2, TA_dinuc_freq, 'myst_2', 'TA')

    myst_3 = dinuc_freq_cal(file_dir[4])[1]
    plot_freq(myst_3, HF_dinuc_freq, 'myst_3', 'HF')
    plot_freq(myst_3, TA_dinuc_freq, 'myst_3', 'TA')

if __name__ == '__main__':
    main()
```

**ANSWERS**

**Problem 1:**

- H influenzae

The nucleotide C appears 350723 times or 19.164950385869467 %.
The nucleotide A appears 567623 times or 31.01726043880323 %.
The nucleotide G appears 347436 times or 18.98533515698983 %.
The nucleotide T appears 564241 times or 30.832454018337472 %.

- T aquaticus

The nucleotide C appears 737506 times or 34.160196353527134 %.
The nucleotide A appears 346223 times or 16.036541617433926 %.
The nucleotide G appears 732676 times or 33.936477836813324 %.
The nucleotide T appears 342558 times or 15.86678419222562 %.

**Problem 2:**

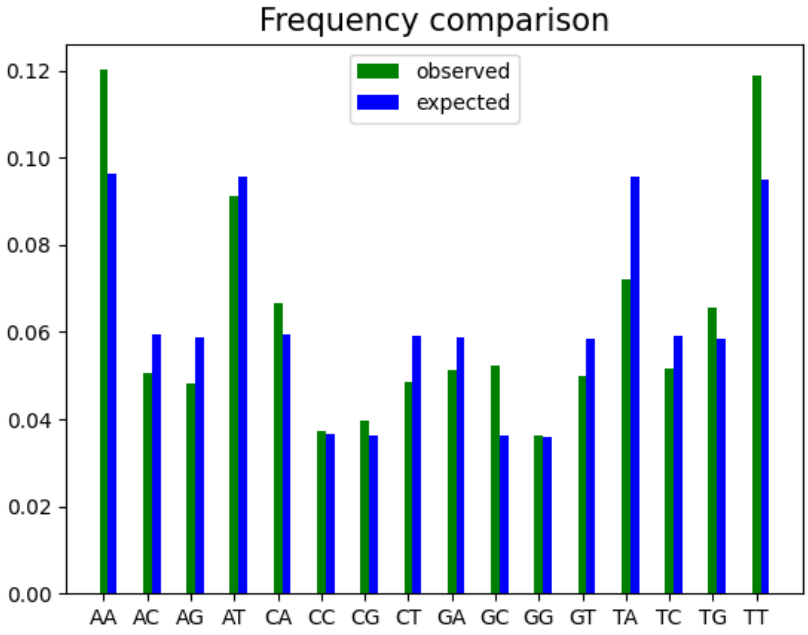
The dinucleotide TA appears 131964 times or 7.211060850634582 %.
The dinucleotide AT appears 166845 times or 9.117103510231024 %.
The dinucleotide TG appears 120001 times or 6.557352862424605 %.
The dinucleotide GG appears 66449 times or 3.6310492442167366 %.
The dinucleotide GC appears 95531 times or 5.2202104674151455 %.
The dinucleotide CA appears 121629 times or 6.6463135415858385 %.
The dinucleotide AA appears 219894 times or 12.015921120073966 %.
The dinucleotide TT appears 217522 times or 11.88630519195944 %.
The dinucleotide GT appears 91320 times or 4.9901039441055905 %.
The dinucleotide TC appears 94753 times or 5.1776973173000105 %.
The dinucleotide CG appears 72525 times or 3.9630671106686153 %.
The dinucleotide CC appears 68016 times or 3.7166766301170147 %.
The dinucleotide AG appears 88461 times or 4.8338763140552405 %.
The dinucleotide AC appears 92423 times or 5.0503764435618805 %.
The dinucleotide GA appears 94136 times or 5.143981875627725 %.
The dinucleotide CT appears 88553 times or 4.838903576022583 %.

**Problem 3:**

The dinucleotide GT appears 85124 times or 3.9428206703036 %.
The dinucleotide TG appears 109792 times or 5.085406783445007 %.
The dinucleotide GG appears 318676 times or 14.76061181252843 %.
The dinucleotide GC appears 200632 times or 9.292984313758186 %.
The dinucleotide CC appears 321674 times or 14.899474840224144 %.
The dinucleotide CT appears 149501 times or 6.924670281366693 %.
The dinucleotide TT appears 68144 times or 3.1563316075039767 %.
The dinucleotide GA appears 128243 times or 5.940030440554303 %.
The dinucleotide AC appears 87691 times or 4.061720400822247 %.
The dinucleotide CG appears 156200 times or 7.23495828087757 %.
The dinucleotide CA appears 110131 times or 5.101108773568039 %.
The dinucleotide AG appears 148007 times or 6.855470360293511 %.
The dinucleotide TC appears 127509 times or 5.90603262123187 %.
The dinucleotide AT appears 39789 times or 1.842968982316502 %.

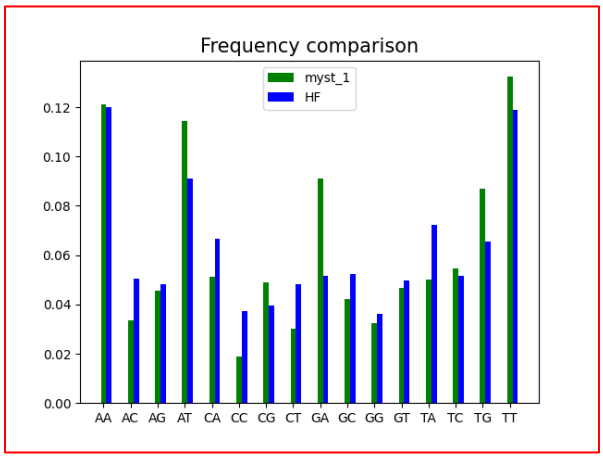
The dinucleotide AA appears 70736 times or 3.2763893018960037 %.
The dinucleotide TA appears 37113 times or 1.7190205293099183 %.

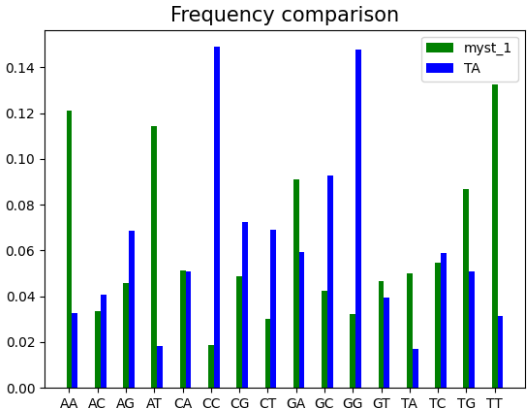
**Problem 4:**

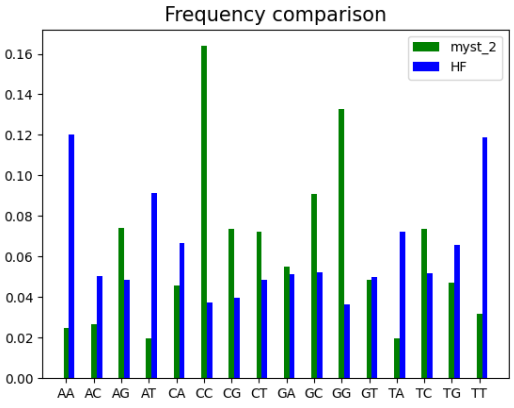


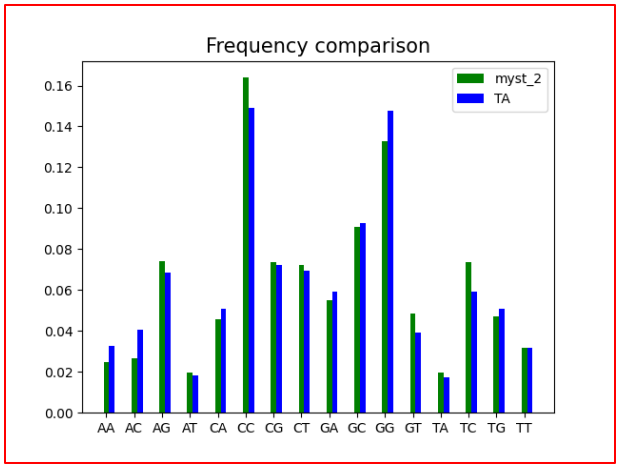

Some mismatches are visible in some dinucleotides, such as AA, GC, TA, and TT, but overall, the expected and the observed frequencies are quite nicely matched up.

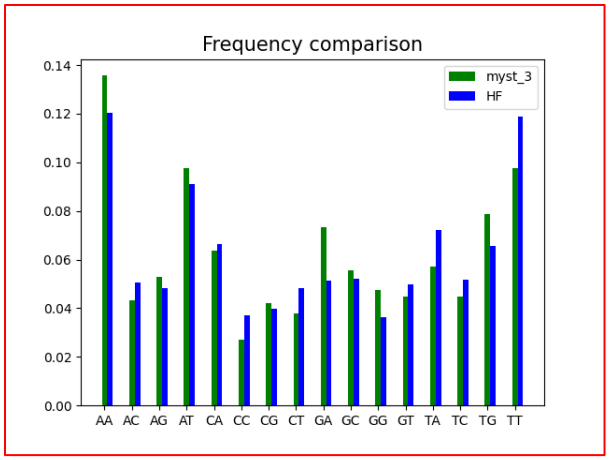
**Problem 5:**
Below are six figures depicting the frequency of each of Mystery Genes against either HF or TA. It is obvious that Mystery 1 is HF, Mystery 2 is TA, and Mystery 3 is HF.

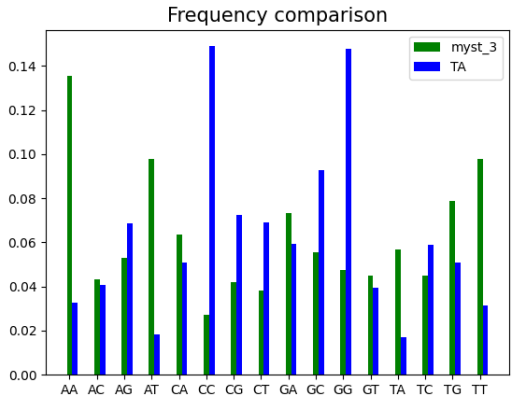





- Mystery gene 1:

The dinucleotide AT appears 103 times or 11.457174638487208 %.
The dinucleotide TG appears 78 times or 8.676307007786429 %.
The dinucleotide GA appears 82 times or 9.121245828698555 %.
The dinucleotide AA appears 109 times or 12.124582869855395 %.
The dinucleotide AC appears 30 times or 3.337041156840934 %.
The dinucleotide CG appears 44 times or 4.894327030033371 %.
The dinucleotide CA appears 46 times or 5.116796440489433 %.
The dinucleotide AG appears 41 times or 4.560622914349278 %.
The dinucleotide GC appears 38 times or 4.226918798665183 %.
The dinucleotide TT appears 119 times or 13.236929922135706 %.
The dinucleotide TC appears 49 times or 5.450500556173527 %.
The dinucleotide CC appears 17 times or 1.8909899888765296 %.
The dinucleotide CT appears 27 times or 3.0033370411568407 %.
The dinucleotide TA appears 45 times or 5.005561735261402 %.
The dinucleotide GG appears 29 times or 3.225806451612903 %.
The dinucleotide GT appears 42 times or 4.671857619577309 %.

- Mystery gene 2:

The dinucleotide CT appears 125 times or 7.2379849449913145 %.
The dinucleotide TA appears 34 times or 1.9687319050376375 %.
The dinucleotide AG appears 128 times or 7.411696583671105 %.
The dinucleotide GG appears 229 times or 13.25998841922409 %.
The dinucleotide GC appears 157 times or 9.090909090909092 %.
The dinucleotide CG appears 127 times or 7.353792704111176 %.

The dinucleotide CC appears 283 times or 16.386797915460335 %.
The dinucleotide TC appears 127 times or 7.353792704111176 %.
The dinucleotide TT appears 55 times or 3.1847133757961785 %.
The dinucleotide TG appears 81 times or 4.6902142443543715 %.
The dinucleotide GA appears 95 times or 5.500868558193399 %.
The dinucleotide CA appears 79 times or 4.57440648523451 %.
The dinucleotide GT appears 84 times or 4.863925883034163 %.
The dinucleotide AA appears 43 times or 2.4898668210770123 %.
The dinucleotide AC appears 46 times or 2.6635784597568035 %.
The dinucleotide AT appears 34 times or 1.9687319050376375 %.

- Mystery gene 3:

The dinucleotide AT appears 72 times or 9.76933514246947 %.
The dinucleotide TG appears 58 times or 7.869742198100408 %.
The dinucleotide GC appears 41 times or 5.563093622795115 %.
The dinucleotide CG appears 31 times or 4.2062415196743554 %.
The dinucleotide CA appears 47 times or 6.377204884667571 %.
The dinucleotide TA appears 42 times or 5.698778833107191 %.
The dinucleotide AC appears 32 times or 4.341926729986431 %.
The dinucleotide CC appears 20 times or 2.7137042062415198 %.
The dinucleotide AA appears 100 times or 13.568521031207597 %.
The dinucleotide AG appears 39 times or 5.291723202170964 %.
The dinucleotide GA appears 54 times or 7.327001356852103 %.
The dinucleotide TT appears 72 times or 9.76933514246947 %.
The dinucleotide TC appears 33 times or 4.477611940298507 %.
The dinucleotide CT appears 28 times or 3.7991858887381276 %.
The dinucleotide GT appears 33 times or 4.477611940298507 %.
The dinucleotide GG appears 35 times or 4.74898236092266 %.

**Problem 6:**
A human genome size is about 3 billion base pairs long, or 3 GB. Yes, you can store it on your 100 GB SD card, along with 32 other people.

Using the 2013-2017 rate, there are 442.4 new cases of cancer per 100000 people. The US population was 328 million in 2019. So that is about 1.45 million new cases yearly. For $1000, it would cost 1.45 billion to sequence all new cancer patients every year. It would take 3 GB * 1.45E6 ~= 4.35 Petabytes, or 2.7% of the TACC's supercomputer's memory.

**Problem 7:**
*E. coli* genome contains 4.2 million base pairs. The human genome is 714 times larger than that of *E. coli*.
*E. coli* gene density = 4500/4.2 Mb = 1071 genes/Mb
Human gene density = 25000/3000 Mb = 8.33 genes/Mb

**Problem 8:**
For each amino acid, all the non-self scores are summed and sorted:

| | Other | Self |
|---|---|---|
| W | -50 | 15 |
| C | -49 | 13 |
| G | -45 | 8 |
| P | -44 | 10 |
| D | -38 | 8 |
| F | -37 | 8 |

| | | |
|---|---|---|
| I | -35 | 5 |
| L | -34 | 5 |
| R | -32 | 7 |
| V | -32 | 5 |
| E | -25 | 6 |
| H | -25 | 10 |
| N | -24 | 7 |
| A | -23 | 5 |
| K | -23 | 6 |
| Y | -21 | 8 |
| M | -20 | 7 |
| S | -20 | 5 |
| T | -19 | 5 |
| Q | -16 | 7 |

W, or tryptophan, is least likely to be substituted by one of the rest because it has the lowest total substitution score against others and the highest substitution score against itself.

**Problem 9:**
Q, or glutamine, is the most likely to be substituted by one of the rest because it has the highest substitution score among the “Other” group. Its self substitution score is also among the lower tier.

**Problem 10:**
Below is the BLOSUM50 table as a heatmap to reveal the lowest individual substitution. The lowest score, which is -5, is held by F-D, W-D, and W-C substitutions.
F = Phenylalanine, W = Tryptophan, D = Aspartic acid, C = Cysteine.

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 5 | -2 | -1 | -2 | -1 | -1 | -1 | 0 | -2 | -1 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 |
| R | -2 | 7 | -1 | -2 | -4 | 1 | 0 | -3 | 0 | -4 | -3 | 3 | -2 | -3 | -3 | -1 | -1 | -3 | -1 | -3 |
| N | -1 | -1 | 7 | 2 | -2 | 0 | 0 | 0 | 1 | -3 | -4 | 0 | -2 | -4 | -2 | 1 | 0 | -4 | -2 | -3 |
| D | -2 | -2 | 2 | 8 | -4 | 0 | 2 | -1 | -1 | -4 | -4 | -1 | -4 | -5 | -1 | 0 | -1 | -5 | -3 | -4 |
| C | -1 | -4 | -2 | -4 | 13 | -3 | -3 | -3 | -3 | -2 | -2 | -3 | -2 | -2 | -4 | -1 | -1 | -5 | -3 | -1 |
| Q | -1 | 1 | 0 | 0 | -3 | 7 | 2 | -2 | 1 | -3 | -2 | 2 | 0 | -4 | -1 | 0 | -1 | -1 | -1 | -3 |
| E | -1 | 0 | 0 | 2 | -3 | 2 | 6 | -3 | 0 | -4 | -3 | 1 | -2 | -3 | -1 | -1 | -1 | -3 | -2 | -3 |
| G | 0 | -3 | 0 | -1 | -3 | -2 | -3 | 8 | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 0 | -2 | -3 | -3 | -4 |
| H | -2 | 0 | 1 | -1 | -3 | 1 | 0 | -2 | 10 | -4 | -3 | 0 | -1 | -1 | -2 | -1 | -2 | -3 | 2 | -4 |
| I | -1 | -4 | -3 | -4 | -2 | -3 | -4 | -4 | -4 | 5 | 2 | -3 | 2 | 0 | -3 | -3 | -1 | -3 | -1 | 4 |
| L | -2 | -3 | -4 | -4 | -2 | -2 | -3 | -4 | -3 | 2 | 5 | -3 | 3 | 1 | -4 | -3 | -1 | -2 | -1 | 1 |
| K | -1 | 3 | 0 | -1 | -3 | 2 | 1 | -2 | 0 | -3 | -3 | 6 | -2 | -4 | -1 | 0 | -1 | -3 | -2 | -3 |
| M | -1 | -2 | -2 | -4 | -2 | 0 | -2 | -3 | -1 | 2 | 3 | -2 | 7 | 0 | -3 | -2 | -1 | -1 | 0 | 1 |
| F | -3 | -3 | -4 | -5 | -2 | -4 | -3 | -4 | -1 | 0 | 1 | -4 | 0 | 8 | -4 | -3 | -2 | 1 | 4 | -1 |
| P | -1 | -3 | -2 | -1 | -4 | -1 | -1 | -2 | -2 | -3 | -4 | -1 | -3 | -4 | 10 | -1 | -1 | -4 | -3 | -3 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | -1 | 0 | -1 | -3 | -3 | 0 | -2 | -3 | -1 | 5 | 2 | -4 | -2 | -2 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 2 | 5 | -3 | -2 | 0 |
| W | -3 | -3 | -4 | -5 | -5 | -1 | -3 | -3 | -3 | -3 | -2 | -3 | -1 | 1 | -4 | -4 | -3 | 15 | 2 | -3 |
| Y | -2 | -1 | -2 | -3 | -3 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | 0 | 4 | -3 | -2 | -2 | 2 | 8 | -1 |
| V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 4 | 1 | -3 | 1 | -1 | -3 | -2 | 0 | -3 | -1 | 5 |

**Problem 11:**

| | D | E | H |
|---|---|---|---|
| D | 8 | 2 | -1 |
| E | 2 | 6 | 0 |
| H | -1 | 0 | 10 |

Ave 2.888889

| | I | L | V |
|---|---|---|---|
| I | 5 | 2 | 4 |
| L | 2 | 5 | 1 |
| V | 4 | 1 | 5 |

Ave 3.222222

Within each group, the score is positive on average because they have the same properties (charged/hydrophobic), making them easy to be replaced by one another.

| | D | E | H |
|---|---|---|---|
| I | -4 | -4 | -4 |
| L | -4 | -3 | -3 |
| V | -4 | -3 | -4 |

ave -3.66667

The average is negative for between groups. This is probably because hydrophobic amino acids tend to be buried inside protein folds to avoid water, while charged amino acids tend to be exposed to the outside toward polarly charged water molecules. Thus, amino acid substitution between these groups is unfavorable.