Miriam Bern
Bio 131
May 11, 2017

Final Project

When you mentioned in class that the Dayhoff paper was readable and fairly simple for a layperson to follow, I knew that I wanted to read through the paper myself and see if I could understand the steps taken to derive the PAM 250 matrix. Therefore, I chose to work through the paper as my project so that I would get a chance to not only seriously read it, but also delve into the math and write my own code based on it. I was also curious as to why the toy DNA example you presented to us in class did not work, and wanted to see if I could fix it and make my toy example.

Because this project was based on the Dayhoff paper, I was able to pull all of my data from the paper itself. Some of the steps, however, I had to do further research on in order to be able to progress through the rest of the paper. I turned the relative mutabilites table (Table 21) and normalized frequencies (Table 22) into dictionaries. I also typed up the Accepted Point Mutations Matrix from the paper into a plain text file. Each row started with an amino acid and the last row contained the same order of the amino acids. I created a function "accepted_point_mutations" to turn this text file into a useable list of lists. I was able to continue this list of lists format until the creation of the PAM 250 matrix. Due to size limitation, I had to use the NumPy package to multiply the matrix rather than my own function (named "PAM_twoFifty_slow"). The NumPy package can create and multiply 2-D arrays and works much more quickly than my own function. However, because I had to convert

from a list of lists into the 2-D array format, the Mutation Probability Matrix for 1 PAM needed to be stripped of the amino acids in the front of each list and the final list containing another row of amino acids. This was a simple step, and only required a short chunk of code that deleted the first element off each list and removed the final list entirely.

The first step of my project, after turning the Accepted Point Mutations text file into a list of lists, was to calculate the proportionality constant (λ). The sum of the values in the Mutation Probability Matrix (Mij) multiplied by the frequency of amino acid i and then multiplied by 100 equals the number of conserved amino acids after being allowed to mutate over the evolutionary interval (PAM) represented by the matrix. This value is dependent on the choice of λ. For example, in the Mutations Probability Matrix for 1 PAM, the percent of conserved amino acids is 99%, and the choice of λ ensures that after 1 PAM evolutionary interval, a sequence will be 99% conserved. Unfortunately, the paper mentioned neither the constant used nor how to derive it. However, I was able to find this information on the point accepted mutation Wikipedia page (screenshot from Wikipedia):

$$0.99 = 1 - \lambda \sum_{j=1}^{20} f(j)m(j)$$

Using λ, which came out to be 0.000133, I was able to calculate the Mutation Probability Matrix for 1 PAM using the following equations taken from the paper:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}} \qquad M_{jj} = 1 - \lambda m_j$$

The equation on the left is for calculating the non-diagonal values of the matrix, where mj is the relative mutability of amino acid j, Aij is the number of mutations between amino acid i and amino acid j (taken from the Accepted Point Mutation Matrix). The equation on the right is for calculating the diagonal values of the matrix.

The next high level step was to multiply it by itself 250 times, and after that to divide it by the frequency of the amino acid in each row to get the Relatedness Odds Matrix:

$$R_{ij} = \frac{M_{ij}}{f_i}$$

After this, I needed to take the log of all the values in the Relatedness Odds Matrix and then rearrange the table according to the final Log Odds Matrix in the paper; the amino acids were rearranged to group chemically similar amino acids (ones most likely to mutate into each other) together.

After creating these matrices, I copied and modified my functions into a different file, toy_example.py and ran your DNA examples through my program. I also created a toy example in which I made up random values in the Accepted Point Mutations Matrix, frequencies, and relative mutabilities to see what the result would be.

Finally, I made up two DNA sequences and used the following example from the paper to pull the values of frequencies and relative mutabilites from my toy example:

Aligned sequences	A D A A D B		
Amino acids	A	B	D
Changes	1	1	0
Frequency of occurrence (total composition)	3	1	2
Relative mutability	.33	1	0

Figure 81. Sample computation of relative mutability. The two aligned sequences may be two experimentally observed sequences or an observed sequence and its inferred ancestor.

The results from my project were very pleasing; all of my matrices matched the matrices in the paper. There were a few values that were off by 1, but I believe this is most likely due to rounding errors. For my Log Odds Matrix, I wrote some code to calculate how many values in the table were asymmetric, checking each pair twice (once for the value table[i][j] and once for the value table[j][i].) The number of asymmetric values came out to be 34, which meant that there were at least 17 values in the table out of 400 total that did not match the Log Odds Matrix in the paper.

When looking through the DNA example presented in class, I noticed that the proportionality constant was not included in the creation of the Mutation Probability Matrix. At first I ran my code without the constant, simply substituting a value of 1 instead of the constant, and found that both my Relatedness Odds Matrix and Log Odds Matrix were different from the class examples. I then ran the DNA

examples with the proportionality constant with the intent of seeing if the matrix was symmetric after doing so. The matrix was more symmetric than the class example, but not perfectly symmetric. Looking at the values provided for the frequencies and relative mutabilites, I realized that those values were extremely unlikely, if not impossible, to pull from an actual dataset. I concluded that it was very unlikely for amino acids (or nucleotides) to have different probabilities of mutating but appear the exact same number of times. To investigate this further, I filled an Accepted Point Mutation Matrix with completely random values and then filled dictionaries of relative mutabilities and frequencies similarly with random values with no connection to each other. I ended up getting a Log Odds Matrix that was about as symmetric as the Log Odds Matrix that resulted from running the DNA class example with the proportionality constant.

Results:

My Mutation Probability Matrix for 1 PAM:

```
Mutation Probability Matrix for 1 PAM (Elements x10000):

['A', 9867, 2, 9, 10, 3, 8, 16, 21, 2, 6, 4, 2, 6, 2, 22, 35, 32, 0, 2, 18]
['R', 1, 9914, 1, 0, 1, 10, 0, 0, 10, 3, 1, 18, 4, 1, 4, 6, 1, 8, 0, 1]
['N', 4, 1, 9822, 36, 0, 4, 6, 6, 21, 3, 1, 12, 0, 1, 2, 20, 9, 1, 4, 1]
['D', 6, 0, 42, 9859, 0, 6, 51, 6, 4, 1, 0, 3, 0, 0, 1, 4, 3, 0, 0, 1]
['C', 1, 1, 0, 0, 9973, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 5, 1, 0, 3, 2]
['Q', 3, 9, 4, 5, 0, 9876, 26, 1, 23, 1, 3, 6, 4, 0, 6, 2, 2, 0, 0, 1]
['E', 10, 0, 7, 56, 0, 35, 9864, 4, 2, 3, 1, 7, 2, 0, 3, 4, 2, 0, 1, 2]
['G', 21, 1, 12, 11, 1, 2, 7, 9935, 1, 0, 1, 3, 2, 1, 3, 21, 3, 0, 0, 5]
['H', 1, 8, 18, 3, 1, 20, 1, 0, 9912, 0, 1, 1, 0, 2, 3, 1, 1, 1, 4, 1]
['I', 2, 2, 3, 1, 2, 1, 2, 0, 0, 9872, 9, 2, 12, 7, 0, 1, 7, 0, 1, 32]
['L', 3, 1, 3, 0, 0, 6, 1, 1, 4, 22, 9947, 1, 45, 13, 3, 1, 3, 4, 2, 15]
['K', 2, 37, 25, 6, 0, 12, 12, 2, 2, 4, 1, 9926, 19, 0, 3, 8, 11, 0, 1, 1]
['M', 1, 1, 0, 0, 0, 2, 0, 0, 0, 5, 8, 3, 9875, 1, 0, 1, 2, 0, 0, 4]
['F', 1, 1, 1, 0, 0, 0, 0, 1, 2, 8, 6, 0, 4, 9946, 0, 2, 1, 3, 28, 0]
['P', 13, 5, 2, 1, 1, 8, 2, 2, 5, 1, 2, 2, 1, 1, 9926, 12, 4, 0, 0, 2]
['S', 28, 11, 34, 7, 11, 4, 5, 16, 2, 2, 1, 6, 4, 3, 17, 9841, 38, 5, 2, 2]
['T', 22, 2, 13, 4, 1, 3, 2, 2, 1, 11, 2, 8, 6, 1, 5, 32, 9871, 0, 2, 9]
['W', 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 9976, 1, 0]
['Y', 1, 0, 3, 0, 3, 0, 1, 0, 4, 1, 1, 0, 0, 21, 0, 1, 1, 2, 9946, 1]
['V', 13, 2, 1, 1, 3, 2, 2, 3, 3, 57, 11, 1, 17, 1, 3, 2, 10, 0, 2, 9902]
['A', 'R', 'N', 'D', 'C', 'Q', 'E', 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V']
```

The published Mutation Probability Matrix for 1 PAM:

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

My Mutation Probability Matrix for 250 PAMs:

```
Mutation Probability Matrix for 250 PAMs (Elements x100):

A [13, 6, 9, 9, 5, 8, 9, 12, 6, 8, 6, 7, 7, 4, 11, 11, 11, 2, 4, 9]
R [3, 16, 4, 3, 2, 5, 3, 2, 6, 3, 2, 9, 4, 1, 4, 4, 3, 7, 2, 2]
N [4, 4, 6, 6, 2, 5, 6, 4, 6, 3, 2, 5, 3, 2, 4, 5, 4, 2, 3, 3]
D [5, 4, 7, 11, 1, 7, 10, 5, 5, 3, 2, 5, 3, 1, 4, 5, 5, 1, 2, 3]
C [2, 1, 1, 1, 52, 1, 1, 2, 2, 2, 1, 1, 1, 1, 2, 3, 2, 1, 4, 2]
Q [3, 5, 5, 5, 1, 10, 7, 3, 7, 2, 3, 5, 3, 1, 4, 3, 3, 1, 2, 2]
E [5, 4, 7, 11, 1, 9, 12, 5, 6, 3, 2, 6, 3, 1, 4, 5, 5, 1, 2, 3]
G [12, 5, 10, 10, 4, 7, 9, 26, 5, 5, 4, 6, 5, 3, 8, 11, 9, 2, 3, 7]
H [2, 5, 5, 4, 2, 7, 4, 2, 15, 2, 2, 3, 2, 2, 3, 3, 2, 2, 3, 2]
I [3, 2, 2, 2, 2, 2, 2, 2, 2, 10, 6, 2, 6, 5, 2, 3, 4, 1, 3, 9]
L [6, 4, 4, 3, 2, 6, 4, 3, 5, 15, 33, 4, 20, 13, 5, 4, 6, 6, 7, 13]
K [6, 18, 10, 9, 2, 10, 9, 6, 8, 5, 4, 24, 9, 2, 6, 8, 8, 4, 3, 5]
M [1, 1, 1, 1, 0, 1, 1, 1, 1, 2, 3, 2, 6, 2, 1, 1, 1, 1, 1, 2]
F [2, 1, 2, 1, 1, 1, 1, 1, 3, 5, 6, 1, 4, 32, 1, 2, 2, 4, 20, 3]
P [7, 5, 5, 4, 3, 5, 4, 5, 5, 3, 3, 4, 3, 2, 19, 6, 5, 1, 2, 4]
S [9, 6, 8, 7, 7, 6, 7, 9, 6, 5, 4, 7, 5, 3, 9, 10, 9, 4, 4, 6]
T [8, 5, 6, 6, 4, 5, 5, 6, 4, 6, 4, 6, 5, 3, 6, 8, 11, 2, 3, 6]
W [0, 2, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 55, 1, 0]
Y [1, 1, 2, 1, 3, 1, 1, 1, 3, 2, 2, 1, 2, 15, 1, 2, 2, 3, 31, 2]
V [7, 4, 4, 4, 4, 4, 4, 5, 4, 15, 10, 4, 10, 5, 5, 5, 7, 2, 4, 17]
```

The published Mutation Probability Matrix for 250 PAMs:

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

My Log Odds Matrix:

```
C [12]
S [0, 2]
T [-2, 1, 3]
P [-3, 1, 0, 6]
A [-2, 1, 1, 1, 2]
G [-3, 1, 0, -1, 1, 5]
N [-4, 1, 0, 0, 0, 0, 2]
D [-5, 0, 0, -1, 0, 1, 2, 4]
E [-5, 0, 0, 0, 0, 0, 2, 3, 4]
Q [-5, -1, -1, 0, 0, -1, 1, 2, 2, 4]
H [-4, -1, -1, 0, -1, -2, 2, 1, 1, 3, 6]
R [-4, 0, -1, 0, -2, -3, 0, -1, -1, 1, 2, 6]
K [-5, 0, 0, -1, -1, -1, 1, 0, 1, 1, 0, 3, 5]
M [-5, -2, -1, -2, -1, -3, -2, -3, -2, -1, -2, -1, 0, 6]
I [-2, -1, 0, -2, -1, -3, -2, -2, -2, -2, -2, -2, -2, 2, 4]
L [-6, -3, -2, -3, -2, -4, -3, -4, -3, -2, -2, -3, -3, 4, 2, 6]
V [-2, -1, 0, -1, 0, -1, -2, -2, -2, -2, -2, -3, -3, 2, 4, 2, 4]
F [-4, -3, -3, -5, -4, -5, -4, -6, -6, -5, -2, -4, -5, 0, 1, 2, -1, 9]
Y [0, -3, -3, -5, -3, -5, -2, -4, -4, -4, 0, -4, -5, -2, -1, -1, -2, 7, 10]
W [-8, -2, -5, -5, -6, -7, -4, -7, -7, -5, -2, 2, -3, -4, -5, -2, -6, 1, 0, 17]
```

The published Log Odds Matrix:

		C Cys	S Ser	T Thr	P Pro	A Ala	G Gly	N Asn	D Asp	E Glu	Q Gln	H His	R Arg	K Lys	M Met	I Ile	L Leu	V Val	F Phe	Y Tyr	W Trp
C	Cys	12																			
S	Ser	0	2																		
T	Thr	-2	1	3																	
P	Pro	-3	1	0	6																
A	Ala	-2	1	1	1	2															
G	Gly	-3	1	0	-1	1	5														
N	Asn	-4	1	0	-1	0	0	2													
D	Asp	-5	0	0	-1	0	1	2	4												
E	Glu	-5	0	0	-1	0	0	1	3	4											
Q	Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H	His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	Phe	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Figure 84. Log odds matrix for 250 PAMs. Elements are shown multiplied by 10. The neutral score is zero. A score of −10 means that the pair would be expected to occur only one-tenth as frequently in related sequences as random chance would predict, and a score of +2 means that the pair would be expected to occur 1.6 times as frequently. The order of the amino acids has been arranged to illustrate the patterns in the mutation data.

Toy Example

string1 = 'TACCGTAAAGG"
string2 = 'TTCCTTAACGA"

Length of string1 = 11
Length of string2 = 11
Total number of amino acids = **22**

String	A	C	G	T
1	4	2	3	2
2	3	3	1	4
Total	7	5	4	6

Frequency(j) = # of times j appears/total # of amino acids
Mutability(j) = # of changes between j and an amino acid i/# of times j appears

Nucleotide: A
string1 = 'TACCGTAAAGG"
string2 = 'TTCCTTAACGA"
Changes: 3

Nucleotide: C
string1 = 'TACCGTAAAGG"
string2 = 'TTCCTTAACGA"
Changes: 1

Nucleotide: G
string1 = 'TACCGTAAAGG"
string2 = 'TTCCTTAACGA"
Changes: 2

Nucleotide: T
string1 = 'TACCGTAAAGG"
string2 = 'TTCCTTAACGA"
Changes: 2

Nucleotide	Mutability	Frequency
A	3/7	7/22
C	1/5	5/22
G	2/4	4/22
T	2/6	6/22

Accepted Point Mutations Matrix

	A	C	G	T
A	0	1	1	1
C	1	0	0	0
G	1	0	0	1
T	1	0	1	0

Mutation Probability Matrix

	A	C	G	T
A	0.988	0.006	0.007	0.005
C	0.004	0.995	0.000	0.000
G	0.004	0.000	0.986	0.005
T	0.004	0.000	0.007	0.991

Mutation Probability Matrix (10 Units)

	A	C	G	T
A	0.881	0.056	0.068	0.047
C	0.040	0.942	0.001	0.001
G	0.039	0.001	0.862	0.046
T	0.040	0.001	0.069	0.906

Odds Matrix

	A	C	G	T
A	2.770	0.175	0.214	0.148
C	0.175	4.146	0.006	0.004
G	0.214	0.006	4.739	0.252
T	0.148	0.004	0.252	3.323

Log Odds Matrix

	A	C	G	T
A	0.442	-0.758	-0.669	-0.831
C	-0.758	0.618	-2.219	-2.386
G	-0.669	-2.219	0.676	-0.598
T	-0.831	-2.386	-0.598	0.521