Danny Diaz
eid: dd32387
2-9-19
BCH 394: HW 1

****

Question 1

****

Influenzae Observed Freq:
{'T': 30.832, 'A': 31.017, 'G': 18.985, 'C': 19.165}

Influenzae Observed Freq:
{'T': 30.831, 'A': 31.015, 'G': 18.984, 'C': 19.164, 'N': 0.003, 'K': 0.001, 'R': 0.001, 'S': 0.001,
'Y': 0.001, 'M': 0.001, 'W': 0.001}

****

Aquaticus Observed Freq:
{'G': 33.936, 'T': 15.867, 'C': 34.16, 'A': 16.037}

Aquaticus Observed Freq:
{'G': 33.936, 'T': 15.867, 'C': 34.16, 'A': 16.037}

****

Question 2, 3, and 4

****

Influenzae Expected Freq:
{'AA': 9.621, 'AC': 5.944, 'AG': 5.889, 'AT': 9.563, 'CA': 5.944, 'CC': 3.673, 'CG': 3.638, 'CT': 5.909,
'GA': 5.889, 'GC': 3.638, 'GG': 3.604, 'GT': 5.853, 'TA': 9.563, 'TC': 5.909, 'TG': 5.853, 'TT': 9.506}

Influenzae Observed Freq:
{'AA': 12.015, 'AC': 5.05, 'AG': 4.834, 'AT': 9.117, 'CA': 6.646, 'CC': 3.717, 'CG': 3.963, 'CT': 4.839,
'GA': 5.143, 'GC': 5.22, 'GG': 3.631, 'GT': 4.99, 'TA': 7.211, 'TC': 5.177, 'TG': 6.557, 'TT': 11.886}

****

Aquaticus Expected Freq:
{'AA': 2.572, 'AC': 5.478, 'AG': 5.442, 'AT': 2.545, 'CA': 5.478, 'CC': 11.669, 'CG': 11.593, 'CT': 5.42,
'GA': 5.442, 'GC': 11.593, 'GG': 11.517, 'GT': 5.385, 'TA': 2.545, 'TC': 5.42, 'TG': 5.385, 'TT': 2.518}

Aquaticus Observed Freq:
{'AA': 3.276, 'AC': 4.062, 'AG': 6.855, 'AT': 1.843, 'CA': 5.101, 'CC': 14.899, 'CG': 7.235, 'CT': 6.925,
'GA': 5.94, 'GC': 9.293, 'GG': 14.761, 'GT': 3.943, 'TA': 1.719, 'TC': 5.906, 'TG': 5.085, 'TT': 3.156}

****

The observed frequencies for both organisms do not match the expected frequencies. I believe this is due to the presence of genes in the genome. Genes consist of 3 nucleotide codons thatare translated into amino acids. The evolutionary pressure on these proteins to conserves their sequence is reflected in observed dinuecleotide frequencies observed.

****

Question 5

****

Mystery Gene 1 expected Freq:
{'AA': 9.887, 'AC': 4.682, 'AG': 6.708, 'AT': 10.167, 'CA': 4.682, 'CC': 2.217, 'CG': 3.176, 'CT': 4.814, 'GA': 6.708, 'GC': 3.176, 'GG': 4.551, 'GT': 6.898, 'TA': 10.167, 'TC': 4.814, 'TG': 6.898, 'TT': 10.454}

Mystery Gene 1 Observed Freq:
{'AA': 12.111, 'AC': 3.333, 'AG': 4.556, 'AT': 11.444, 'CA': 5.111, 'CC': 1.889, 'CG': 4.889, 'CT': 3.0, 'GA': 9.111, 'GC': 4.222, 'GG': 3.222, 'GT': 4.667, 'TA': 5.0, 'TC': 5.444, 'TG': 8.667, 'TT': 13.222}

****

Mystery Gene 2 expected Freq:
{'AA': 2.11, 'AC': 5.161, 'AG': 4.749, 'AT': 2.505, 'CA': 5.161, 'CC': 12.625, 'CG': 11.618, 'CT': 6.127, 'GA': 4.749, 'GC': 11.618, 'GG': 10.691, 'GT': 5.639, 'TA': 2.505, 'TC': 6.127, 'TG': 5.639, 'TT': 2.974}

Mystery Gene 2 Observed Freq:
{'AA': 2.488, 'AC': 2.662, 'AG': 7.407, 'AT': 1.968, 'CA': 4.572, 'CC': 16.377, 'CG': 7.35, 'CT': 7.234,
'GA': 5.498, 'GC': 9.086, 'GG': 13.252, 'GT': 4.861, 'TA': 1.968, 'TC': 7.35, 'TG': 4.688, 'TT': 3.183}

****

Mystery Gene 3 expected Freq:
{'AA': 10.931, 'AC': 5.645, 'AG': 7.302, 'AT': 9.184, 'CA': 5.645, 'CC': 2.915, 'CG': 3.771, 'CT': 4.743,
'GA': 7.302, 'GC': 3.771, 'GG': 4.878, 'GT': 6.135, 'TA': 9.184, 'TC': 4.743, 'TG': 6.135, 'TT': 7.716}

Mystery Gene 3 Observed Freq:
{'AA': 13.55, 'AC': 4.336, 'AG': 5.285, 'AT': 9.756, 'CA': 6.369, 'CC': 2.71, 'CG': 4.201, 'CT': 3.794,
'GA': 7.317, 'GC': 5.556, 'GG': 4.743, 'GT': 4.472, 'TA': 5.691, 'TC': 4.472, 'TG': 7.859, 'TT': 9.756}

****

p-value that mystery gene 1 is in the Influenza genome: 9.15e-01
p-value that mystery gene 2 is in the Influenza genome: 2.80e-15
p-value that mystery gene 3 is in the Influenza genome: 9.99e-01

****

p-value that mystery gene 1 is in the Aquaticus genome: 4.41e-23
p-value that mystery gene 2 is in the Aquaticus genome: 1.00e+00
p-value that mystery gene 3 is in the Aquaticus genome: 4.12e-17

****

The p-values provided are from a "Chi-Square Goodness of Fit" test conducted on the observeddinucleotide frequencies between each gene and each genome. The comparison of the Hinfluenzae genome with mystery gene 1 and 3 provided the respective p-values:.915 and .999. Thus, we cannot reject the null hypothesis that their is a statistically significant difference between their observed dinucleotide frequencies. Hence, mystery gene 1 and 3 are in the Hinfluenzae genome. The p-value for mystery gene 2 was statistically significant (< 0.05). Thus, we reject the null hypothesis that there is not a difference between the observed dinucleotide frequencies of mystery gene 2 and Hinfluenzae.

The comparison of the Taquaticus genome with the 3 mystery gene provided 2 p-values that are statistically significant (< 0.05): gene 1 and 3. Thus, we can reject the null hypothesis that that their is no difference between their observed dinucleotide frequencies. This makes perfect sense because we concluded that these genes were present in Hinfluenzae. The p-value for mystery gene 2 is 1.000; the null hypothesis cannot be rejected and mystery gene 2 observed dinucleotide frequencies show no difference from the the dinucleotide frequencies observed in Taquaticus. Hence, mystery gene 2 is in the Taquaticus genome.

****

Question 6

****

You can store your a copy of the entire human genome on your cell phone's 100 GB SD card if you assume

each nucleotide is 1 byte. The human genome is 3 giga base pairs and you have 100 giga bytes available.

To sequence the genome of every human diagnosed with cancer in the US each year (1,735,350 in 2018) will cost a little over $1.735 billion, assuming each genome is $1000 and we do not receive whole sale prices.

To store the genomes of every American diagnosed with cancer in 2018 you will require 5.21 Peta bytes of space on your local PC.

****

Question 7

****

The human genome is approximately 652 times larger than the E. Coli genome.

The density of genes in the human genome is approximately 133,333 bp per gene. This approximatationassumes that there is 22,500 genes in the human genome and that the genome is 3 giga base pairs.The density of genes in E. Coli is approximately 1,022 bp per gene. This is assuming that there are 4,500 genes and the genome length is 4.6 mega base pairs.

****

Question 8

****

Tryptophan is the amino acid that is least likely to be substituted by another. It has a negative substitution score with every amino acid besides phenylalanine and tyrosine, the other 2 aromatic amino acids. These scores are 1 and 2 respectively, however, with itself its blossum score is 15. Thus, it really does not like to be substituted by other aromatic amino acids when compared to self.

Other amino acids that show a similar profile to tryptophan are glycine and cysteine. For cysteine, its score with itself is 13 compared to tryptophan's 15. However, unlike tryptophan it does not have a positive blossum score for any other amino acid, but its score values are not as negative as tryptophan. Like cysteine, glycine does not have a positive blossum score for any substitution except for itself. However, the blossum score with itself is only an 8 and its most positive blossum score is 3 zero values with alanine, serine, and asparagine.

****

Question 9

****

Polar amino acids are most easily substituted by other polar amino acids. Specifically, asparagine and glutamine. Especially if you consider substitution scores of only non-negative numbers (including 0), these 2 amino acids scored the highest each with 9. In close second, lysine, histidine, and glutamic acid each scored 7 and serine scored 8.

****

Question 10

****

The most disfavored substitutions are: tryptophan to aspartic acid, tryptophan to cysteine, and phenylalanine to aspartic acid. The trend here is the substitution of an aromatic for a short acidic residue is highly disfavored.

****

Question 11

****

The average blossum score between DEH: 1/3

The average blossum score between VIL: 7/3

The average blossum score between the 2 groups: -11/3

Substitution between polar amino acids and greasy amino acids is favorable. Subsitution from a polar to a nonpolar or vice-versa is not favorable. This is due to their chemical properties and roles within a protein.