Question 1: How does one compute the amino acid frequencies from a sequence?

The frequency is equal the occurrences of the certain aa divided by the total numbers of aa in a seqence.

Question 2: What amino acid appears to be the most/least frequent for your choice

of data set?

I selected the set 1 hemn, and the p(W) is the least one and p(L) is the most frequent with 11.1%.

Question 3: There is a line in the graph, called “random”, at the frequency 5%.

Why 5%?

Question 4: How can one estimate P(α|β) from a sequence of amino acids?

P(α|β) = p(**β**/α)p(α)/p(β), p(**β**/α) could be big which turns out P(α|β) high.

Question 5: For your choice of amino acid to condition on, what amino acids

appear to have the highest/smallest probability? Do you have any biological explanation?

Question 6: What is the sum of all these probabilities, and why? (Hint: If unsure,

use the Matlab function sum on the returned variable plett, i.e. sum(plett).)

p = 0.999962962962966,

Question 7: Even if both P(α) and P(β) are relatively small, P(α|β) can be relatively

high. Explain why!

P(α|β) = p(**β**/α)p(α)/p(β), p(**β**/α) could be big which turns out P(α|β) high.

Question 8: Take one pair where P(α)P(β|α) > P(α)P(β) and in your own words

explain the meaning of this in terms of independent events.

EA: P(E)P(A) = 0.005739 P(E)P(A|E) = 0.010262,

GV: P(G)P(V) = 0.003132 P(G)P(V|G) = 0.006552,

Question 9: Suppose we were to produce random protein sequences (random but

using the known frequencies P(α)) and then count the number of αβ pairs in

these sequences. If you have the condition P(α)P(β|α) < P(α)P(β) in your real

sequences and also here count the number of αβ pairs. Would you expect to find

more or less αβ pairs in the random sequences?

PL: P(P)P(L) = 0.004618 P(P)P(L|P) = 0.002831,

MD: P(M)P(D) = 0.001163 P(M)P(D|M) = 0.001033

Question 10: Run the permutation test for αβ pairs from question 8. Can you

confirm the previous findings using the results found by running permtest? Give

an example



