What are the frequencies of each amino acid? List them in decreasing order.

Total Chars: 8076

Amount for Each Amino Acid: {'L': 804, 'E': 682, 'G': 594, 'S': 562, 'A': 562, 'K': 510, 'R': 458, 'Q': 436, 'V': 414, 'P': 398, 'D': 398, 'T': 376, 'I': 356, 'N': 331, 'F': 295, 'Y': 259, 'M': 195, 'H': 175, 'C': 148, 'W': 123}

Frequencies of Each Amino Acid: {'L': 804/8076, 'E': 682/8076, 'G': 594/8076, 'S': 562/8076, 'A': 562/8076, 'K': 510/8076, 'R': 458/8076, 'Q': 436/8076, 'V': 414/8076, 'P': 398/8076, 'D': 398/8076, 'T': 376/8076, 'I': 356/8076, 'N': 331/8076, 'F': 295/8076, 'Y': 259/8076, 'M': 195/8076, 'H': 175/8076, 'C': 148/8076, 'W': 123/8076}

What are your 50 alignment scores?

The following is the current result corresponding to the random.txt:

alignment score result: [56, 64, 47, 57, 50, 37, 56, 44, 60, 39, 56, 50, 61, 43, 51, 52, 52, 52, 43, 52, 49, 59, 51, 50, 59, 45, 55, 62, 49, 44, 54, 57, 50, 54, 57, 53, 49, 47, 54, 56, 62, 51, 49, 49, 55, 56, 55, 64, 45, 50]
alignment score freq: {56: 5, 50: 5, 49: 5, 52: 4, 57: 3, 51: 3, 55: 3, 54: 3, 64: 2, 47: 2, 44: 2, 43: 2, 59: 2, 45: 2, 62: 2, 37: 1, 60: 1, 39: 1, 61: 1, 53: 1}

The following are multiple test result for different random sequence generating and correpsonding alignment scores to form a more robust conclusion:

- **Sample1:** [45, 65, 52, 52, 55, 78, 49, 51, 56, 50, 52, 65, 54, 47, 54, 56, 53, 58, 45, 62, 58, 56, 63, 59, 52, 60, 58, 56, 49, 51, 44, 55, 46, 43, 53, 49, 47, 65, 50, 47, 53, 47, 50, 50, 70, 59, 63, 56, 50, 51]

- **Sample 2:** [54, 52, 62, 55, 60, 56, 47, 51, 46, 70, 62, 58, 47, 46, 59, 51, 54, 49, 70, 47, 54, 54, 55, 51, 56, 52, 53, 51, 62, 53, 79, 53, 61, 59, 61, 61, 60, 48, 71, 62, 75, 64, 58, 65, 49, 53, 54, 43, 51, 47]
- **Sample 3:** [65, 47, 46, 54, 63, 52, 46, 46, 59, 51, 47, 53, 61, 47, 46, 51, 48, 56, 54, 59, 52, 51, 51, 49, 50, 53, 64, 68, 47, 49, 50, 69, 52, 49, 60, 54, 47, 55, 45, 55, 56, 48, 55, 49, 58, 50, 46, 58, 50, 56]
- **Sample 4:** [58, 47, 51, 41, 49, 53, 54, 56, 47, 61, 60, 49, 64, 53, 44, 67, 54, 48, 45, 46, 67, 51, 54, 62, 55, 56, 47, 45, 53, 46, 52, 48, 55, 50, 58, 48, 64, 61, 54, 54, 45, 51, 60, 47, 65, 51, 48, 65, 52, 53]
- **Sample 5:** [54, 51, 58, 48, 50, 49, 55, 51, 57, 53, 69, 49, 47, 46, 48, 45, 49, 48, 46, 53, 52, 60, 52, 41, 50, 60, 47, 54, 53, 61, 59, 44, 61, 43, 54, 62, 55, 56, 58, 58, 51, 50, 58, 50, 62, 53, 55, 51, 57, 49]

Do your data points appear to be clustered? Are there outliers?

With the above six samples in total, we can observe the frequency: {51: 23, 47: 21, 54: 21, 49: 20, 50: 19, 56: 18, 53: 18, 52: 17, 55: 15, 58: 13, 46: 12, 62: 10, 48: 10, 60: 9, 61: 9, 59: 9, 45: 9, 65: 7, 64: 6, 57: 5, 44: 5, 43: 5, 63: 3, 70: 3, 69: 2, 41: 2, 67: 2, 37: 1, 39: 1, 78: 1, 79: 1, 71: 1, 75: 1, 68: 1}

- The mean value is 53.616
- The standard deviation is 6.643

which means, yes, the data points seem to be clustered, several random generated sequence alignments are tested, and all of the results shows the scores are usually ranged from 40 to 60.

Outlier: When generating 100 random sequences multiple times to obtain a more convincing result, some but not many scores below 40 or higher than 70 are observed.

How might these statistics be relevant to the BLAST algorithm?

Consider the Scoring Matrix

Block: In the protein sequence database, the protein sequences are clustered into 500 groups based on the identity between the sequences (greater than a threshold), the sequences in each group are subjected to multiple sequence alignment, the blocks are defined by the conservative and ungapped regions.

Scoring Matrix: Thinking about the scoring matrix like PAM and BLOSUM, notice the LOD (log odds ratio) when talking about the Alternative Hypothesis, where $LOD(a\ pair\ of\ amino\ acid) = log\frac{observed\ frequency}{expected\ freqency\ in\ random\ model}$, with LOD we iterate the block to get the score.

Sequence Similarity:

1. With the above definition, we can notice the scoring matrix actually contains the information about sequence similarity. For example, BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
2. Now, notice the alignment score we get from the random model, it is clustered in the range from 40 to 60. That's to say, if the score is much higher than this range, it implies the alignment may have the statistical significance under the same scoring matrix system.
3. Be more specifically, here we use matrix PAM250. PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity. Then PAM250 is PAM1 times itself with 250 times, which is relatively equivalent to the BLOSUM45, which basically aligns with the random scores range we obtained.