Pristian Budi Dharmawan_2501983105_BB20_CB_Session 13_Quiz

1. EDA Analysis

   >protein_sequence
   MRKDLSPVLENYLLRVCVQGVAQKPVKALQKLQHSLTAAALYALQKKTAVVAPAAPLAAA

   a) Google Colab Link: Click Here or https://colab.research.google.com/drive/1nlIzVEmqDc9t_qOLs1IInRu-6byDu-Lb?usp=sharing

   b) Berikut adalah figure yang menunjukkan hasil frequency amino acid dari sequence diatas

```
# Amino Acid Frequency
from collections import Counter

# Reading our fasta file
quiz = SeqIO.read("sequence.fasta", "fasta")
# Freq
quiz_dna = quiz.seq
quiz_count = Counter(quiz_dna)
quiz_count
```

```
Counter({'M': 1,
         'R': 2,
         'K': 6,
         'D': 1,
         'L': 10,
         'S': 2,
         'P': 4,
         'V': 7,
         'E': 1,
         'N': 1,
         'Y': 2,
         'C': 1,
         'Q': 5,
         'G': 1,
         'A': 13,
         'H': 1,
         'T': 2})
```

```
# Data Frame (OPTIONAL BUT IMPORTANT FOR DATA SCIENCE)
import pandas as pd
df = pd.DataFrame({"amino_acids": protein_test_clean})

df['count'] = df['amino_acids'].str.len()

print(df.head())

print(df.nlargest(10, 'count'))

df.nlargest(10, 'count')
```

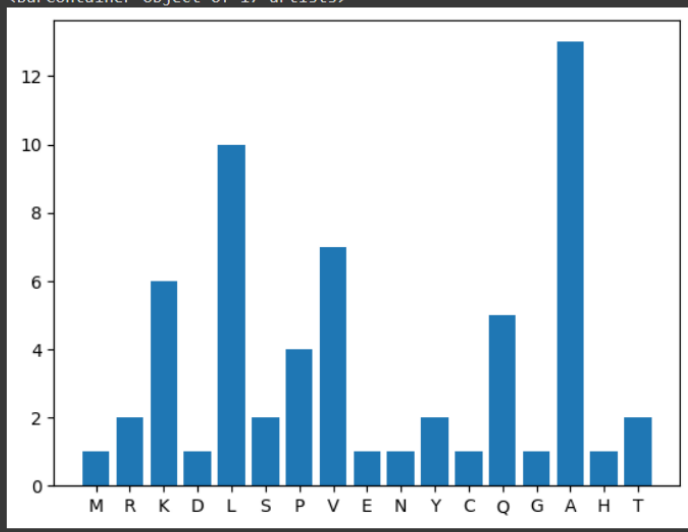
```
                                          amino_acids  count
0  MRKDLSPVLENYLLRVCVQGVAQKPVKALQKLQHSLTAAALYALQK...     60
                                          amino_acids  count
0  MRKDLSPVLENYLLRVCVQGVAQKPVKALQKLQHSLTAAALYALQK...     60
```

	amino_acids	count
0	MRKDLSPVLENYLLRVCVQGVAQKPVKALQKLQHSLTAAALYALQK...	60

   c) Analisis komposisi amino acid dari poin b)

```
# Plot the Amino Acid
import matplotlib.pyplot as plt
plt.bar(quiz_count.keys(), quiz_count.values())
```

```
<BarContainer object of 17 artists>
```



2. Interpretasi hasil analisis amino acid

Amino Acid	A	L	K	V	Q	R	C	S	T	Y	P	M	H
Freq	8	12	9	5	3	1	1	2	1	2	4	1	1

a) Amino acid dengan frequency terbanyak dan terendah
Amino acid dengan frequency terendah adalah R, C, T, M, H. Sedangkan amino acid dengan frequency tertinggi adalah L. Signifikansi ini memiliki arti bahwa amino acid L berperan besar terhadap protein yang sedang dianalisis dan dilanjutkan dengan K, A, dan seterusnya. Sedangkan amino acid dengan frequency rendah berarti tidak berperan signifikan terhadap protein yang dianalisis. Meskipun frequency sebuah amino acid rendah, hal ini bisa menjadikan sebuah patokan dalam menganalisa suatu penyakit tertentu.

b) Amino acid yang lebih tinggi atau lebih rendah dari rata-rata?
Rata-rata amino acid dari sequence diatas ialah 3,84615384615385. Sehingga amino acid dengan frequency yang lebih tinggi dari nilai tersebut ialah A, L, K, V, P. Sedangkan sisanya adalah amino acid yang lebih rendah. Tentu hal ini memiliki sebuah implikasi tersendiri terhadap struktur protein yang sedang dianalisis. Hal ini disebabkan apabila amino acid dengan frequency yang berubah maka hal ini juga akan berpengaruh terhadap fungsi protein yang dianalisi, bisa menjadi sebuah deficiency atau menjadi lebih efisien. Bahkan dari nilai ini kita bisa melihat apakah protein tersebut memiliki kecacatan atau tidak.

c) Kita bisa melakukan copy-paster amino acid tersebut ke dalam protein database seperti NCBI. Selain itu, kita juga bisa melakukan BLAST method untuk mencari kemiripan protein yang sedang dianalisis dengan protein yang sudah disimpan selama ini. Melalui metode ini, kita bisa mendapatkan informasi tambahan dari protein yang sedang dianalisis saat itu juga.