Сравнение распределения длин белков *Escherichia coli* и *Saccharomyces cerevisiae*

Введение

Escherichia coli (*E. coli*) является широко изученным прокариотическим организмом, в то время как *Saccharomyces cerevisiae* (дрожжи) служат модельным организмом для эукариот. Оба они играют важную роль в биологических исследованиях и биотехнологии. Цель данного анализа — сравнить распределения длин белков этих двух организмов и выявить различия.

Данные для этого анализа были получены из UniProt и загружены в виде несжатых TSV по следующим ссылкам:

- *E. coli* data[1]
- Yeast data[2]

Датасеты содержат следующую информацию: entry name, protein names, gene names, organism, length.

Использованные библиотеки

В этом анализе использовались следующие библиотеки Python:

- **pandas**: Для обработки и очистки данных.
- **matplotlib**: Для создания визуализаций, включая гистограммы, диаграммы ящиков и ECDF.
- **seaborn**: Для улучшенного визуального оформления диаграмм ящиков.
- **numpy**: Для числовых операций и предварительной обработки данных.

[1]Также используемые данные для *E. coli*, organism id 83333 можно посмотреть тут
[2]Также используемые данные для Yeast, organism id 559292 можно посмотреть тут

Методы

Анализ включал следующие методы:

- **Гистограмма**: используется для визуализации распределения частот длин белков.
- **Box Plot**: отображает центральную тенденцию (*медиану*[3]), изменчивость (*межквартильный размах*[4]), усы[5] и выбросы[6] в длинах белков.
- **ECDF (эмпирическая кумулятивная функция распределения)**: иллюстрирует *кумулятивные вероятности*[7] для сравнения распределений.

Результаты

Гистограммы

Гистограмма отображает распределение частот длин белков для *E. coli* и дрожжей.

Используя следующий код:

```
import pandas as pd
import matplotlib.pyplot as plt

# Load data
ec_data = pd.read_csv('ecoli_data.tsv', sep='\t')
yeast_data = pd.read_csv('yeast_data.tsv', sep='\t')

# Extract lengths
ec_lengths = ec_data['Length']
yeast_lengths = yeast_data['Length']

# Plot histogram
plt.hist(
    ec_lengths, bins=50, alpha=0.7, label='E. coli', color='blue', density
        =True
)
plt.hist(
    yeast_lengths, bins=50, alpha=0.7, label='Yeast', color='orange',
        density=True
```

[3]Медиана — значение, делящее набор данных на две равные части.

[4]Межквартильный размах — разница между третьим (Q3) и первым (Q1) квартилями, показывающая диапазон значений в центральных 50% данных.

[5]Усы — линии, выходящие за границы межквартильного размаха и простирающиеся до минимального и максимального значений без выбросов.

[6]Выбросы — значения, значительно отличающиеся от остальных и находящиеся за пределами усов.

[7]Кумулятивные вероятности — это вероятности, показывающие долю значений, меньших или равных заданному значению.

```
)
plt.xlabel('Protein Length (Amino Acids)')
plt.ylabel('Frequency')
plt.title('Joint Histogram of Protein Lengths')
plt.legend()
plt.show()
```

Код 1: Гистограмма

Получим:

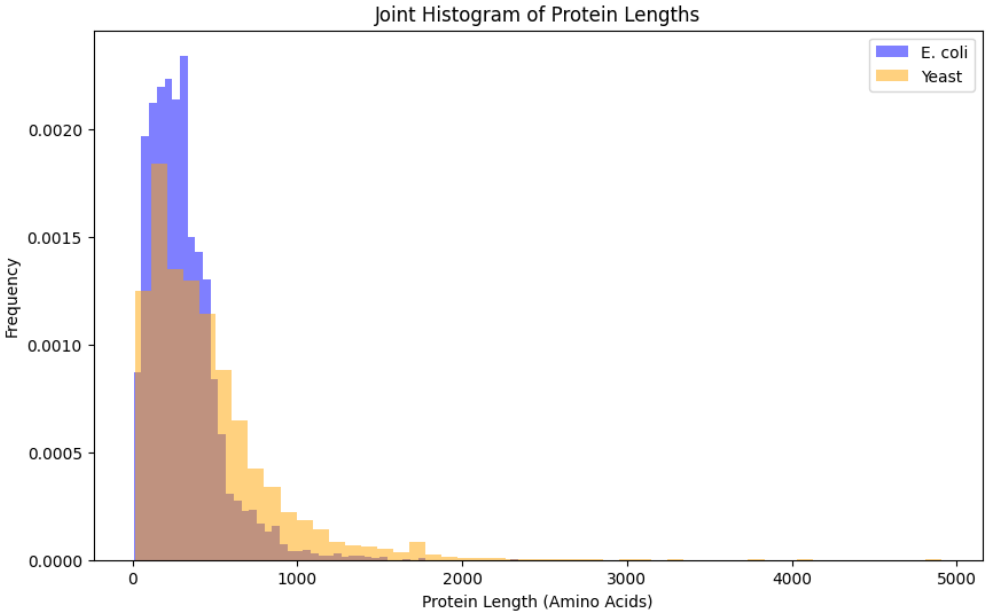

Рис 1: Гистограмма длин белков для *E. coli* и дрожжей.

Box Plot

Box plot показывает медиану (центральную линию), межквартильный размах (IQR, представленный ящиком), усы (представляющие данные в пределах 1,5 IQR) и выбросы.

Используя следующий код:

```
import seaborn as sns

# Combine data for plotting
data = pd.DataFrame({
    'Organism': ['E. coli'] * len(ec_lengths) + ['Yeast'] * len(
        yeast_lengths),
    'Length': pd.concat([ec_lengths, yeast_lengths])
```

```
})

# Plot box plot
sns.boxplot(
    x='Organism', y='Length', data=data, palette=['blue', 'orange']
)
plt.xlabel('Organism')
plt.ylabel('Protein Length (Amino Acids)')
plt.title('Box Plots of Protein Lengths')
plt.show()
```

Код 2: Box Plot

Получим:

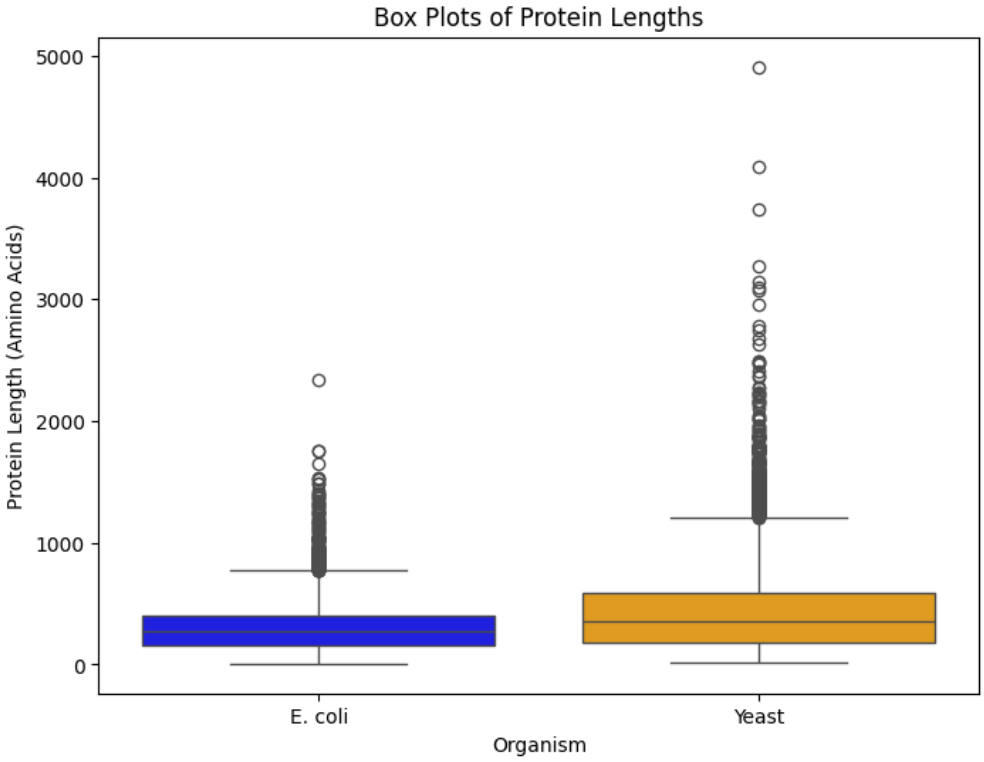

Рис 2: Диаграммы длин белков для *E. coli* и дрожжей.

ECDF

ECDF иллюстрирует кумулятивные вероятности, помогая выявлять различия в распределениях.

Используя следующий код:

```
def ecdf(data):
    x = np.sort(data)
    y = np.arange(1, len(x) + 1) / len(x)
    return x, y

# Compute ECDF for both datasets
ec_x, ec_y = ecdf(ec_lengths)
yeast_x, yeast_y = ecdf(yeast_lengths)

# Plot ECDF
plt.plot(ec_x, ec_y, label='E. coli', color='blue')
plt.plot(yeast_x, yeast_y, label='Yeast', color='orange')
plt.xlabel('Protein Length (Amino Acids)')
plt.ylabel('Cumulative Probability')
plt.title('Empirical Cumulative Distribution Function (ECDF)')
plt.legend()
plt.show()
```

Код 3: ECDF

Получим:

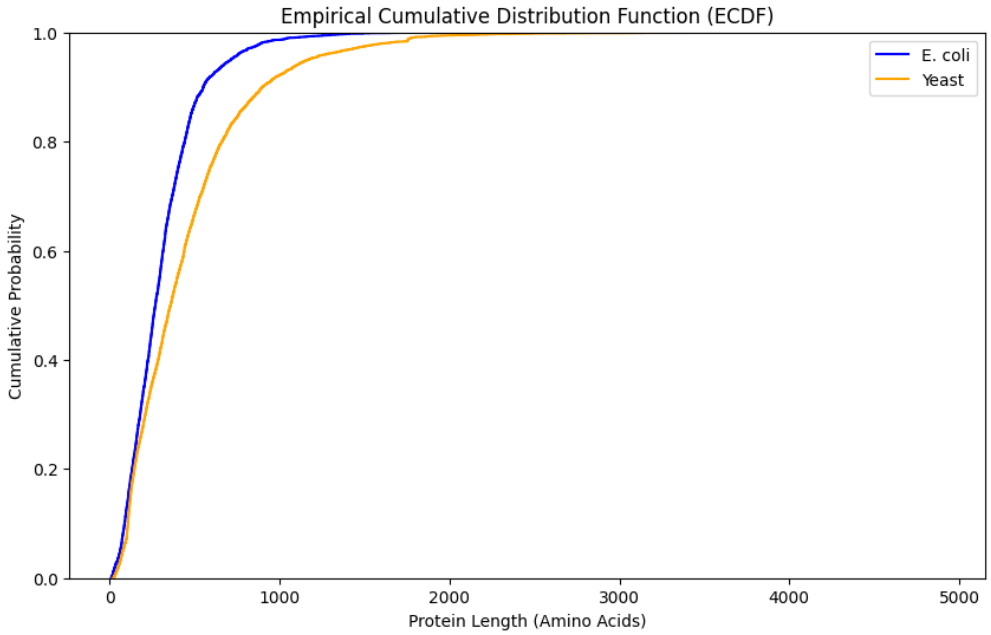

Рис 3: ECDF длин белков для *E. coli* и дрожжей.

Также, для удобства, можно рассмотреть логарифмическую шкалу длин для ECDF.

Используя следующий код:

```
plt.figure(figsize=(10, 6))
sns.ecdfplot(ecoli_lengths, label='E. coli', color='blue')
sns.ecdfplot(yeast_lengths, label='Yeast', color='orange')

# Set logarithmic scale for x-axis
plt.xscale('log')

plt.title('Empirical Cumulative Distribution Function (ECDF)')
plt.xlabel('Protein Length (Amino Acids, Log Scale)')
plt.ylabel('Cumulative Probability')
plt.legend()
plt.grid(True)
plt.show()
```

Код 4: ECDF log-scale

Получим:

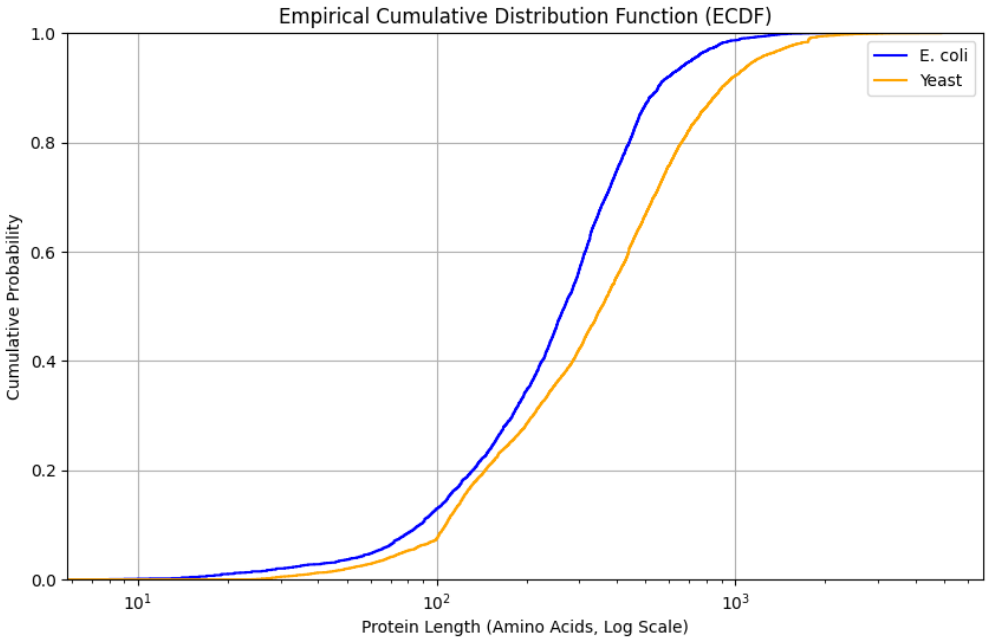

Рис 4: ECDF логарифмов длин белков для *E. coli* и дрожжей.

Результаты

Результаты показывают, что белки *E. coli* в целом короче и имеют более узкий диапазон длин по сравнению с белками дрожжей. Гистограмма (Рис. 1) демонстрирует, что белки дрожжей имеют более широкое распределение, а диаграмма размаха (Рис. 2) подчёркивает наличие большего количества выбросов у дрожжей, что указывает на большую изменчивость. ECDF (Рис. 3) подтверждает, что большая доля белков *E. coli* попадает в диапазоны меньшей длины по сравнению с белками дрожжей.

Эти различия, вероятно, связаны с большей сложностью эукариотических организмов, требующих более длинных и универсальных белков для выполнения клеточных функций.

Заключение

Данный анализ выявил значительные различия в распределении длин белков между *E. coli* и дрожжами, что отражает их различные биологические характеристики. Белки *E. coli* короче и менее вариативны, тогда как белки дрожжей характеризуются большей длиной и разнообразием, что свидетельствует о сложности эукариот.