

Trabalho de simulação e conclusão da disciplina PPB

Dr. Célio Dias Santos Júnior

Instruções sobre forma, execução e entrega:

O trabalho deve ser feito de modo similar a um paper (usar o template da revista [Bioinformatics](#)), onde haverá um material e métodos explicando o que foi feito, bem como um parágrafo de data availability com um link para o respectivo GitHub repository criado para a disciplina.

Este repositório deve conter os scripts e uma pasta com os dados, da seguinte forma:

```
|_ README.txt  
|_ dados/  
|_ scripts/  
|_ analysis/
```

Os alunos devem produzir os scripts de forma organizada e com um arquivo readme no repositório explicando a função de cada script e a localização de cada dado gerado, bem como a estrutura de cada arquivo. [Será avaliado a adoção ao padrão [PEP8](#)]

O prazo para entrega é até 1 semana antes da conclusão do curso.

Procedimentos:

1. Gerar uma tabela de OTUs aleatória em formato TSV com 26 amostras (nas colunas) nomeadas com as letras do alfabeto, e 100 linhas (representando OTUs) que serão nomeadas como 'OTU_{n}', em que n representa o número da linha. Assim, a tabela gerada deve conter 27 colunas com a primeira coluna sendo para o nome das OTUs e nomeada como 'OTU'.
2. Cada amostra dessa tabela deve conter pelo menos 40% e no máximo 75% de células como zeros e somar até no máximo 100.000 contagens.

3. Criar uma função para rarefar a tabela até o mínimo de contagens por amostra da tabela original.

4. Rarefar a tabela de OTUs e salvar ela como um arquivo tabulado.

5. Criar uma função para normalização da tabela de OTUs, de modo que cada OTU dentro de amostra seja convertida a sua proporção em cada amostra, então multiplicando cada célula então por um número grande. O resultado deve novamente ser convertido a um número inteiro.

6. Normalizar a tabela original de OTUs.

6.5. Criar uma tabela em que os índices são as abordagens (normalizado, rarefeito, original) com 2 colunas (média de contagens por amostra e desvio padrão dessas médias). Plotar um boxplot das médias do tratamento rarefeito e normalizado.

7. Criar uma função para determinar a curva do coletor

8. Determinar a curva do coletor (e criar gráficos) para todas as amostras das 3 abordagens. -- *Idealmente serão 3 gráficos (normalizado, rarefeito e original)*

9. Criar uma função para determinar o índice de Shannon.

10. Calcular o índice de Shannon e determinar qual abordagem é a melhor.

-- *Alternativamente para pontos extras, o aluno pode criar ou buscar métodos para mostrar qual o melhor método de tratamento de tabelas de OTUs funciona melhor (e.g., normalização de contagens) ou criar suas teorias sobre o assunto, desde que suporte as suas conclusões com seus resultados e código.*

Pacotes recomendados:

- Numpy
- Pandas
- Scipy
- Seaborn
- Matplotlib.pyplot