

Tarefa: Análise de Cadeias de Markov e Taxa de Entropia com Dados Reais de Sequências de DNA

Objetivo:

O objetivo desta tarefa é permitir que os alunos pratiquem a aplicação dos conceitos de processos estocásticos, taxa de entropia e cadeias de Markov usando o software GNU Octave/Matlab. A tarefa consistirá em duas partes principais: a análise de uma cadeia de Markov com dados reais de sequências de DNA e o cálculo da taxa de entropia de uma fonte de informação.

Parte 1: Análise de uma Cadeia de Markov com Dados de Sequências de DNA

1. Obtenha um conjunto de dados reais contendo sequências de DNA. Existem várias fontes onde esses dados podem ser encontrados, como bancos de dados públicos de sequenciamento genético.
2. Analise os dados e identifique os possíveis estados da cadeia de Markov. Nesse caso, os estados podem ser os nucleotídeos individuais (A, C, G, T) ou combinações de nucleotídeos.
3. Estime a matriz de transição da cadeia de Markov a partir dos dados disponíveis.
4. Implemente uma função em GNU Octave/Matlab que receba como entrada a matriz de transição e um estado inicial, e retorne uma sequência de estados gerada pela cadeia de Markov.
5. Utilizando a função implementada, gere uma sequência de estados com um tamanho mínimo de 1000.
6. Calcule a distribuição estacionária da cadeia de Markov estimada.
7. Verifique se a distribuição estacionária obtida é consistente com a matriz de transição estimada.

Parte 2: Cálculo da Taxa de Entropia

1. Implemente uma função em GNU Octave/Matlab que receba como entrada a matriz de transição de uma cadeia de Markov e calcule a taxa de entropia da fonte de informação.
2. Utilizando a função implementada, calcule a taxa de entropia para a cadeia de Markov na Parte 1.
3. Compare a taxa de entropia obtida com os resultados obtidos no trabalho anterior sobre propriedade da equipartição assintótica.
4. Discuta as implicações dos resultados obtidos em termos da quantidade de informação presente na sequência de DNA e sua relevância no contexto genômico.

Entrega:

Os alunos devem entregar um relatório contendo: 1. Uma descrição detalhada dos passos realizados para a implementação das funções. 2. Os resultados obtidos na Parte 1, incluindo a sequência de estados gerada e a distribuição estacionária encontrada. 3. O cálculo da taxa de entropia na Parte 2 e a análise comparativa com os resultados obtidos no trabalho prático anterior. 4. Uma discussão sobre as implicações dos resultados obtidos em termos da quantidade de informação presente na sequência de DNA e sua relevância no contexto genômico.

Observações:

- Os alunos podem trabalhar individualmente ou em grupos.
- O relatório deve conter o código fonte das funções implementadas.
- Explore diferentes conjuntos de dados de sequências de DNA e diferentes modelos de cadeias de Markov para análises comparativas.

Sugestão de leitura:

- Capítulo 17: The Genetic Code; Livro: Cryptography, Information Theory, and Error-Correction; Autores: Aiden A. Bruen, Mario A. Forcinito; DOI: <https://doi.org/10.1002/9781118033296.ch17>