隱馬可夫模型於 DNA 序列之分析

指導教授：賀保羅
專題成員：曾暐翔、謝承佑
開發工具：Visual Studio Code
測試環境：Microsoft Windows

一、專題簡介

本專題運用 **隱馬可夫模型**（Hidden Markov Model, HMM）對 DNA 序列進行統計建模與結構推測。透過觀察 DNA 中的鹼基（A、C、G、T）序列，推論其潛在的隱狀態（例如 GC 含量高的區段 High，以及含量低的區段 Low），並藉由機器學習方法自動學習其狀態轉移與發射機率。
我們以 **Viterbi 演算法**預測最可能的隱狀態序列，並利用 **Baum-Welch（EM）演算法**反覆更新模型參數，提升模型生成與解碼的準確度。最終模型可用於：

- 模擬產生具有統計結構的 DNA 序列，
- 評估實際測試序列的生成機率，
- 並與隨機模型 (1/4)n(1/4)^n(1/4)n 產生的機率作比較，以評估模型學習成效。

核心技術

- **隱馬可夫模型（HMM）建模**
- **Viterbi 演算法**：預測最可能的隱狀態序列
- **Baum-Welch EM 演算法**：最大期望參數訓練
- **前向／後向演算法**：計算後驗機率與 log-likelihood
- **C 語言實作**：高效訓練與推論大型 DNA 序列

系統功能模組

- **序列讀取**：從檔案載入 DNA 訓練與驗證資料。
- **初始參數估計**：結合 Viterbi 解碼與頻率統計進行快速初始化。
- **模型訓練**：使用 Baum-Welch 進行多輪 EM 訓練，修正模型參數。
- **狀態預測**：推測測試序列對應的隱狀態組成。
- **序列生成**：根據模型參數隨機產生具結構性的 DNA 序列。
- **生成機率驗證**：計算驗證序列的生成機率，並與隨機生成模型比較。

模型訓練與效能驗證

- 比較訓練後模型在驗證資料上的生成機率，與隨機模型 （1/4）^n 的生成機率差異。
- 分析兩者之間的機率比值（提升倍數），以證明模型學會了序列中的隱含統計結構。

專題貢獻與應用潛力

本研究展示了隱馬可夫模型應用於 DNA 分析的可行性，並具以下潛在應用：

- 基因區段辨識（如高/低 GC 區塊）
- 統計序列模擬與預測
- 基因異常區段初步偵測與篩選

二、測試結果