蛋白質序列分析

Computational analysis of protein sequences

指導教授：賀保羅
專題成員：邱鈺証
開發工具：Python 3.11 VS CodeTensorFlow2.x
測試環境：Windows 11 Anaconda

一、簡介：

蛋白質的功能與其三維結構密切相關，而三維結構中的 二級結構（Secondary Structure） 是一個重要的中介資訊，反映了胺基酸序列在局部範圍內如何形成如 α 螺旋（Helix, H）、β 摺板（Strand, E）與無規則線圈（Coil, C）等結構單元。常見的簡化表示為 Q3 格式，將 DSSP 標註的 8 種原始結構類型簡化為三類（H/E/C），方便進行分類與預測。

本研究的資料來自蛋白質的 .dssp 檔案，這些檔案是由 PDB 結構檔透過 DSSP 工具轉換而來，記錄了每個胺基酸殘基的結構類型。經前處理後，轉換為 .txt 檔格式，包含每條序列的胺基酸序列與對應的 Q3 結構標註。

本研究使用 隱馬可夫模型（Hidden Markov Model, HMM） 對這些序列進行結構預測。我們先實作了基本的 一階 HMM，僅考慮目前狀態與前一個狀態的轉移關係；接著推廣為 二階 HMM，利用前兩個狀態提供更豐富的上下文資訊。進一步，我們設計了一種 中階上下文感知 HMM（Mid-Order Contextual HMM），將每個位置的左右胺基酸作為條件資訊納入發射機率，模擬胺基酸物化性質對結構的影響；最後，我們導入機率平滑與 log 機率機制，避免低頻問題並提升模型穩定性。

資料經過 80% / 20% 的訓練與測試切分後，四種模型分別進行學習與預測，並比較其準確率與泛化能力，驗證所提出方法之有效性。

二、測試結果：

總筆數：23391 條蛋白質序列
平均序列長度：超過 100 個胺基酸
預處理方式：
將字元轉為整數編碼
補齊（padding）序列長度以對齊矩陣
切分為 80% 訓練集 / 20% 測試集

共讀取資料筆數： 23391
範例資料：
('MSSHEGGKKKALKQPKKQAKEMDEEEKAFKQKQKEEQKKLEVLKAKVVG
KGPLATGGIKKSGKK','CCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHCCHHHCCCCCCCCCC')

模型名稱　說明
一階 HMM :基本隱馬可夫模型，考慮前一個結構狀態與當前發射機率
二階 HMM :考慮前兩個結構狀態之轉移，參數量較大但表現不穩
中階 HMM :發射機率依賴左右胺基酸上下文