量化基因體中的重複序列

Quantifying Repetitive Sequence in Genomes

指導教授：賀保羅

專題成員：李哲宇

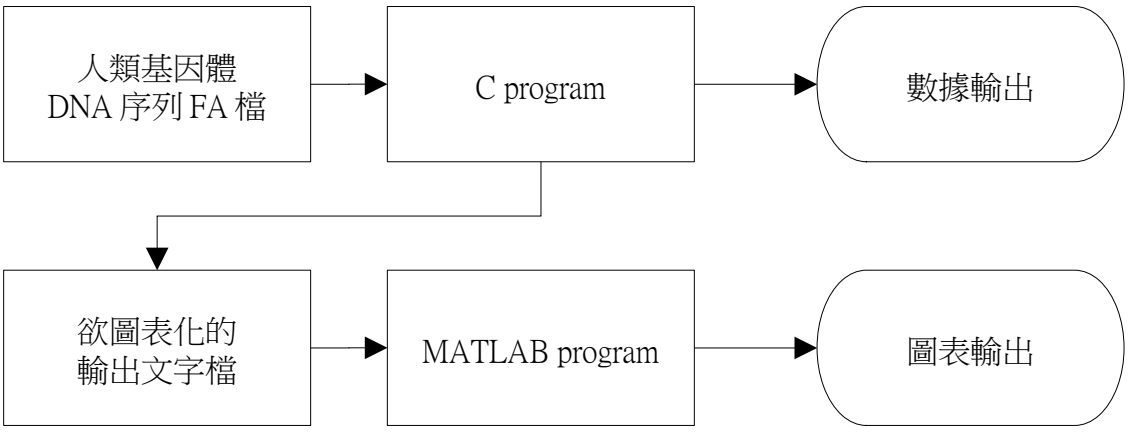
開發工具：GCC 8.1.0、Visual Studio Code、MATLAB R2022a

測試環境：Windows 10

一、簡介：

DNA 定序藉由分析 A、T、C、G 鹼基的排列方式，有效的推動分子生物學與現代醫學的發展。而其中研究基因體中高重複性的序列能對診斷基因疾病、血緣關係鑑定等等帶來極大的幫助。

本專題的目的是運用 C 語言實作 Markov Model、Autocorrelation 等等序列、訊號處理演算法來數據化，並運用 MATLAB 圖表化人類基因體中，重複出現的 DNA 序列片段。讓我們對人類基因體有更大的了解。



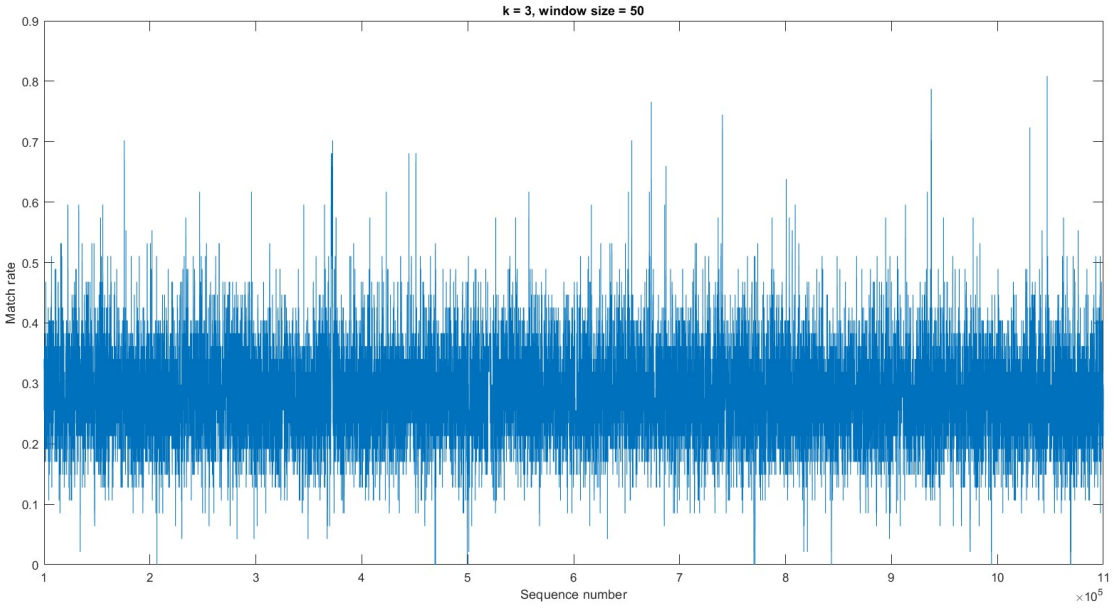
圖一：系統架構圖

二、測試結果：

將 NCBI 提供的人類基因體第六對染色體（NC_000006.12 Homo sapiens chromosome 6，GRCh38.p14）第 100001 至 1100000 個鹼基序列作為輸入，得到以下分別為 Markov Model 和 Autocorrelation 的輸出結果：

```
level 0 model log base 2 probability: -1988107.761833
level 1 model log base 2 probability: -1949355.207585
level 2 model log base 2 probability: -1932983.620669
```

圖二：Markov Model 的 log 以 2 為底之機率



圖三：k=3, window size=50 之 Autocorrelation 結果