運用馬可夫程序理論於 B 型肝炎病毒基因序列突變率之研究

Apply Markov Process Theory on Mutation Rate of Hepatitis B Virus Sequence

孫光天[a] 李? 毅[a] 張定宗[b] 劉紋君[b]

[a]國立臺南大學資訊教育研究所 [b]國立成功大學基礎醫學研究所

在 B 型肝炎病毒(HBV)中，某些位置相對容易發生突變，而某些位置則相對不容易發生突變，動機為想找出這些位置與突變率的行為特性，這些行為特徵可以轉換成馬可夫模型理論中的元素-轉置矩陣，藉此特性，可以預測未來的 DNA 基因序列與分析突變行為。至於突變率公式，則採用公認最精準的突變率 Substitution rate:

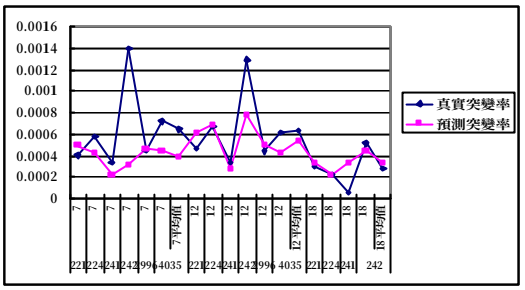
$$\frac{-(3/4)\ln[1-(4/3)p]}{2T} \quad (1)$$

其中 T 為時間而 P 為突變個數除以序列全長。本研究所使用的馬可夫程序理論也屬於一種隨機程序(Stochastic process)，與回歸模式不同之處在於無法用一條曲線表示，一條曲線僅能表示一次實驗(experiment)，而每次實驗所繪出之曲線都可能不同。總結來說，隨機程序是隨機變數的家族，描述某些程序隨著時間而演化的過程。 所謂的馬可夫特性若以數學式表示如下:

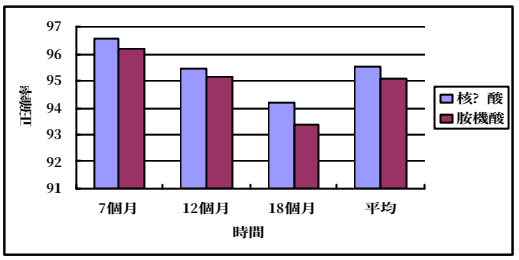
$$P(X_{n+1}=s_{n+1} \mid X_n=s_n, X_{n-1}=s_{n-1}.....,X_0=s_0)$$
$$=P(X_{n+1}=s_{n+1} \mid X_n=s_n), where 0=t_0<t_1<...<t_n<t_{n+1}, \forall s_i \in S$$

也就是說 X_{n+1} 的 CDF 與上一個觀察值 X_n 相關，屬於一種短期相依 (short-rang dependence).

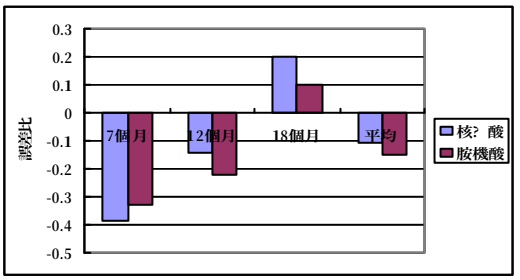
本研究之研究對象為由成大醫學院張定宗研究團隊所提供，共計 6 位病患各 14 筆抽血之 B 型肝炎病毒(HBV)基因序列，並將連續一年之資料當作訓練集，用來計算轉置機率矩陣，而與第一筆資料相距 18 個月之序列為測試集。研究結果如下:



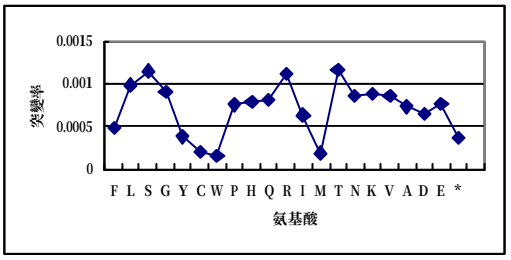
核? 酸突變率折線圖



核? 酸與胺基酸特異點序列預測正確率比較圖



核? 酸與胺基酸突變率預測誤差比比較圖



胺基酸突變的機率分配圖