亞洲大學|資電學院
Asia University| College of Information and Electrical Engineering
資訊工程學系
Computer Science & Information Engineering




人類與病毒共同DNA序列查詢系統

亞洲大學 資訊工程學系　學生：楊永聖、王家恩、林建宏
指導教授：王經篤副教授

摘要

本專題將建立一個人類與病毒共有基因體DNA序列的網際網路介面的查詢系統。這個系統包括三個的子系統(1)基因體序列之可擴充性儲存系統（2）DNA序列線上比對系統 （3)病毒分類查詢系統。「基因體序列之可擴充性儲存系統」是將非關聯式資料庫HBase，建置在實驗室自行架設的Hadoop Cluster系統(圖5)，來儲存大量的共有基因體DNA序列與這些序列在不同類別(人類與各種病毒)出現次數分布。「病毒分類查詢系統」則是使用傳統的關聯式MySQL資料庫來查詢病毒的分類資訊（科屬種）。「DNA序列線上比對系統」則是利用網路機器人， 將查詢者輸入的DNA序列，呼叫NCBI Blast API完成線上即時比對，然後篩選回傳比對結果。本專題所開發的查詢系統，可以提供生物或病毒學家做全面性地「人類與病毒共有基因體DNA序列」研究參考。

根據系統概念圖(如圖1)，本專題分為四個部分:(1)資料來源 (2)基因體序列之可擴充性儲存系統 (3) DNA序列線上比對系統 (4)病毒分類查詢系統。

1.資料來源: 從 NCBI (National Center for Biotechnology Information) [1] 取得人類24條染色體以及7,538種病毒的序列， 並透過maximal repeat方法[2]在大量序列中取得最大重複序列與這些序列在人類與病毒出現次數的分布[3] (圖1左上實心黑框)

2.基因體序列之可擴充性儲存系統:人類與病毒共有基因體DNA序列可擴充性儲存系統存入 **Apache HBase** [4](圖1綠色虛線框)並且透過網頁讀取這個系統的資料(圖2)。

圖2 HBase資料格式（key, value）

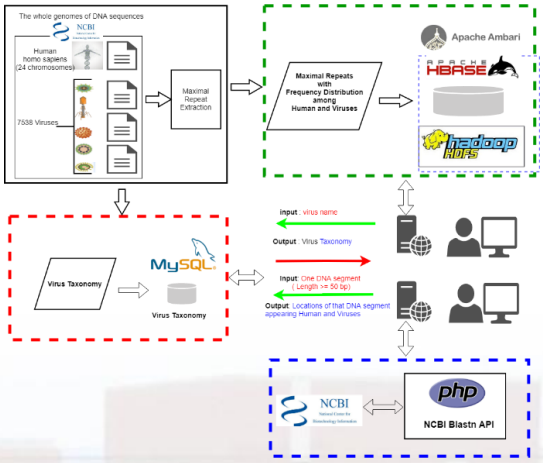

圖1 系統概念圖

3.病毒分類查詢系統:使用MySql資料庫存放病毒科屬種的詳細資料(圖1紅色虛線框)，讓使用者搭配基因體序列之可擴充性儲存系統了解病毒的詳細資料(圖3)。

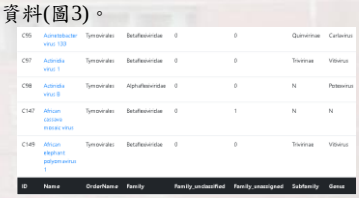

圖3 病毒分類資訊（NCBI擷取）

4.DNA序列線上比對系統(圖1藍色虛線框):透過另外兩個系統查詢完使用者感興趣的病毒資訊後(圖4)，即可透過「病毒分類查詢系統」中的連結透過NCBI Blast API到NCBI資料庫中與大量的DNA序列進行比對。

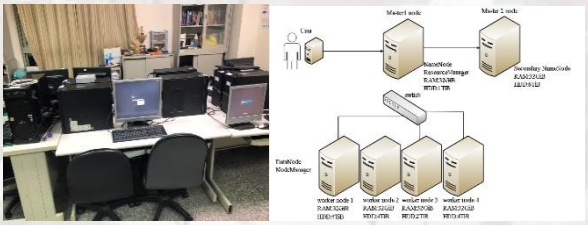

圖5 Hadoop Cluster系統概念圖及實際環境(Lab H204)

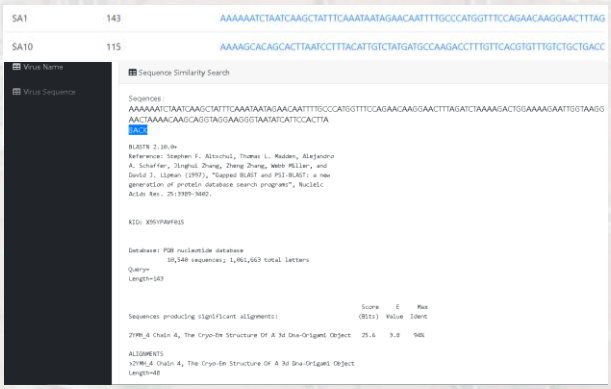

圖4: NCBI Blastn API 查詢結果

參考文獻
[1] NCBI FTP Site, “ftp://ftp.ncbi.nih.gov/genomes/”
[2] Wang, Ching-Tu. Method for Extracting Maximal Repeat Patterns and Computing Frequency Distribution Tables. U.S. Patent No. 10,409,844 (September 10, 2019)
[3] Jing-Doo Wang, Yi-Chun Wang, Rouh-Mei Hu and Jeffrey Tsai, Extracting the Co-occurrences of DNA Maximal Repeats in both Human and Viruses The 17th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2017) pages 106-111, 2017. Washington DC, U.S.A.
[4] Apache HBase™, “https://hbase.apache.org/”