博雅書院期末報告–

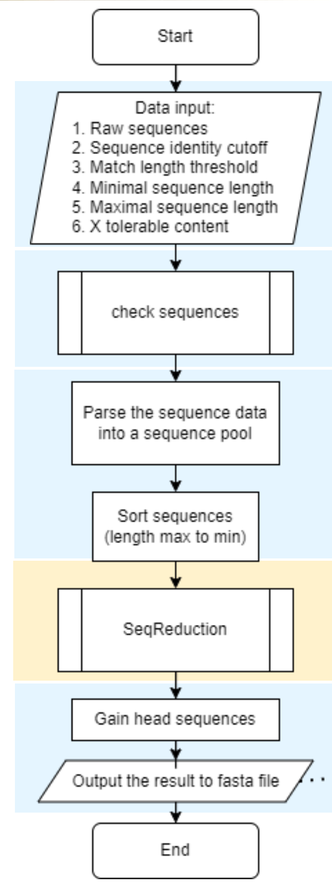
高效能生物資料去冗餘演算法

第十一組 高效能生物資料去冗餘演算法開發小組

想法理念

我們相信，臺灣作為資訊科技巨頭，有實力帶領全球開創數位生物科技新紀元。我們企圖拋磚引玉，以提升大數據分析效率為切入點，開發高效能生物序列資料去冗餘演算法，盼加速全球生物科技發展，並鼓舞更多資訊團　隊投入生物科技研發，提升臺灣的國際地位。大數據時代來臨，處理海量數據極耗資源與時間，但許多生物資料具高度相似性，全數分析並無必要。當前諸多生物資料去冗餘演算法仍有缺點，速度快者不夠精確，精確者運算速度卻難以跟上資料增加腳步。我們發現，幾乎所有序列去冗餘演算法都將去冗餘及分群共同封裝，但很多研究只需去冗餘之資料而非分群結果。因此，我們欲開發一演算法，把傳統同時進行去冗餘和分群的操作做兩階段拆分，並於去冗餘階段用 Golang 程式語言實現高度平行運算，來提升資料生產效率。期望在海量數據充斥的現今，能以臺灣強大資訊實力，提供全球一個快速獲取高密度原始資料並銳減資料總量之有效解決方案，推進各生物科技領域研發，共創新紀元。

演算流程



資料輸入

- 原始序列資料
- 初步篩選條件

資料篩選

- 利用輸入之初步篩選條件做序列資料前處理

序列排序

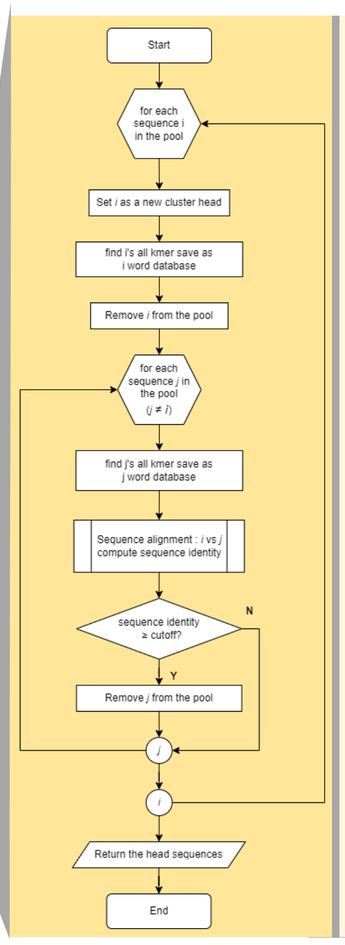
- 將序列整理存入記憶體 (序列池)，以便後續程式使用
- 將序列池中序列由長至短排序，以利後續序列比對之進行

序列去冗餘　*主程式*

- 進行序列比對並篩選出代表性序列

結果取得

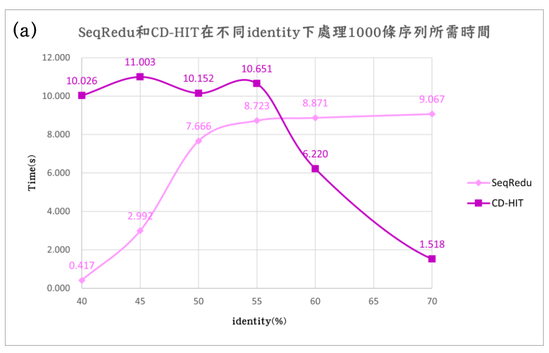
- 將取得之篩選結果輸出至檔案中
  - 檔案中含有序列entry及整條序列編碼

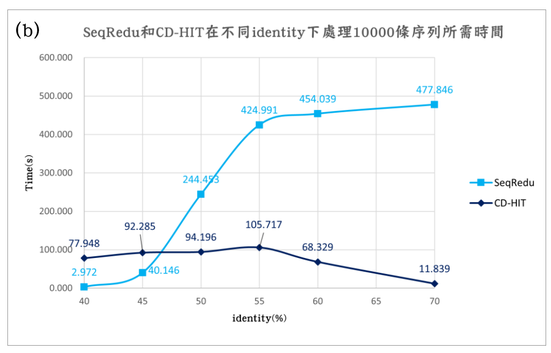


- 第一層迴圈：
  取出序列池中最開頭的序列直接當作代表性序列，計算並儲存其文字kmer(為某序列中某長度的序列短片段。如一序列：AGDEE，其長度為三的kmer有AGD、GDE、DEE)。
- 第二層迴圈：
  運用分散運算一次進行多執行緒計算，透過和其他序列進行序列比對後得出之相似度進行篩選，如相似度大於使用者設定值，此序列將被視為和代表性序列有一定相似度，而從序列池中被移除。
- 所有執行緒都完成計算：
  進入下回合迴圈，重複直至序列池中所有序列都已比對完成。過程中如遇到已計算過文字kmer的序列，將會從緩衝區中取出先前計算之結果，直接進行比對。
- Sequence alignment副程式：
  穿繞兩序列kmer集合資料，如有相同文字片段，則會透過已儲存在其中的文字片段位置，往下進行BLAST(Basic Local Alignment Search Tool)序列比對算法的計算，切除頭尾分數較低之配對後，再計算兩序列相似度

研究結果

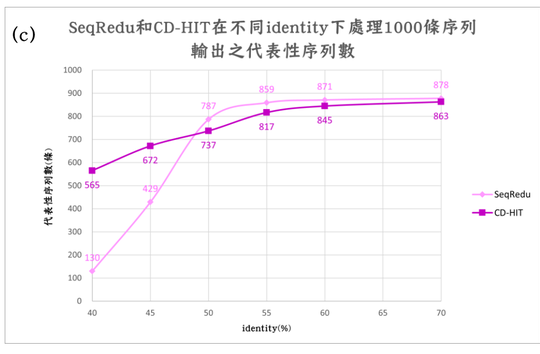
所需時間

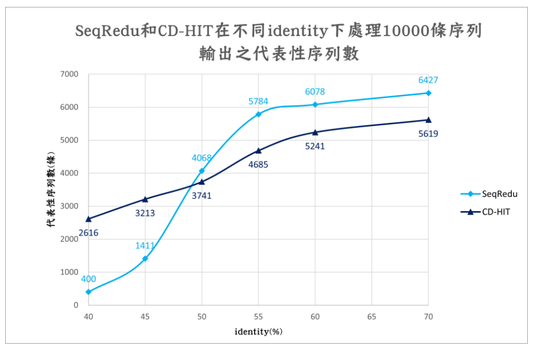




隨著處理序列數量上升(a至b圖)，在低identity時，SeqRedu的計算速度可以較CD-HIT快上許多，尤其當資料量越多時，耗費時間差距會越顯著；但在高identity時，SeqRedu所耗費之時間會急遽上升，尤其在處理序列數為*10000*條(b)時，CD-HIT計算70% identity只需約*11*秒，而SeqRedu所耗費的時間卻高達8分鐘左右。因此，如何優化SeqRedu，提升其在高identity時的計算效能，將是我們在未來開發過程中，需要彼此充分思考討論並解決的一大關卡。

序列數

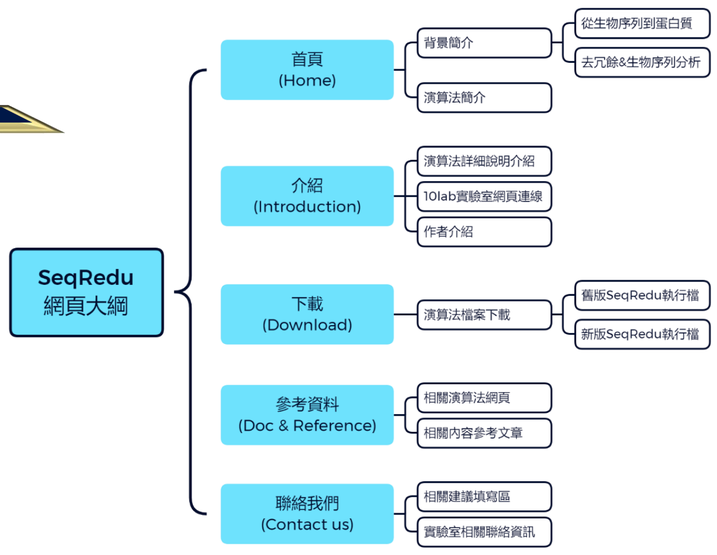




在高identity時，兩演算法輸出之代表性序列數會相近，SeqRedu輸出之序列數會稍多於CD-HIT，但當identity降低(由右至左)時，兩個演算法在不同資料量大小下，皆會在約50% identity處發生交叉，當identity低於50%，SeqRedu所篩選出之序列數相較CD-HIT會有明顯的降低。因為SeqRedu在計算過程中不會開gap(gapless alignment)，在程式停止後切除頭尾分數較低之配對，並直接計算兩序列相似度，最後再利用輸出之相似度進行序列去冗餘暨分群。因此，SeqRedu較不容易算出高identity的結果。

網站成果展示

網站QR code



Sequence Reduction
高效能生物資料序列去冗餘演算法

這是由鈞詒、千珊、詩玟、啓宏、暘典，在羅惟正教授及梁美智教授的指導下，做出的研究成果。

我們的目標是：開發新時代的生物資訊演算法，提升台灣在生物科技的國際地位

參與人員

Associate Professor
Wei-Cheng Lo (羅惟正)

WadeLo@nycu.edu.tw

Associate Professor
Mei-Chih Liang (梁美智)

mcliang@nycu.edu.tw

Undergraduate
Chun-Yi Yang (楊鈞詒)

necocat.bt09@nycu.edu.tw

Undergraduate
Chien-Shan Lin (林千珊)

lcs0909.bt09@nycu.edu.tw

Undergraduate
Shih-Wen Weng (翁詩玟)

zaki.bt09@nycu.edu.tw

Undergraduate
Yang-Tien Chou (周暘典)

csowo.cs09@nctu.edu.tw

Undergraduate
Qi-Hong Su (蘇啓宏)

megumi.bt09@nycu.edu.tw