國立臺南大學資訊工程學系 106級畢業專題

多重資訊來源之基因表現群聚分析整合系統開發

A Platform for Cluster Analysis of Gene Expression Data with Multiple Information Sources

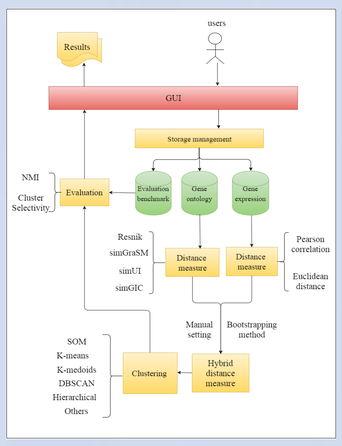
第三組　陳奕廷 臧芷藝

■ 摘要

由於生物科技不斷的進步，基因相關的資料量也逐漸龐大，以人力分析既費時又費工的，因此有許多的分群法以及各種基因相關的知識(例如：基因語意相似度、基因調控路徑…等)的產生，基因微陣列可以幫助我們觀察上千或是上百萬個基因的表現，而這些數量也會隨著時間不斷地變大，過去研究有結合基因知識的基因表現資料群集分析，例如藉由考慮基因表現資料相似度與語意相似度，希望有效提升群聚分析結果的生物意義相關性。若能有一個整合不同資訊來源與各種分析方法的基因表現群集分析平台，對於生醫領域使用者做基因表現分析時將很便利。因此，本專題設計開發了一個整合分析軟體平台系統(有雲端與單機兩種版本)，提供給生醫相關人員使用，除了實現使用拔靴法自動決定不同資訊來源的權重分配，並設計友善的圖形化使用者介面，讓使用者可以自己選擇想要做的分群法、距離測度法以及語意相似度計算法，並將分群結果與評估以圖形化方式呈現。並設計未來擴充其他基因資訊來源的整合介面。

■ 研究方法&系統架構

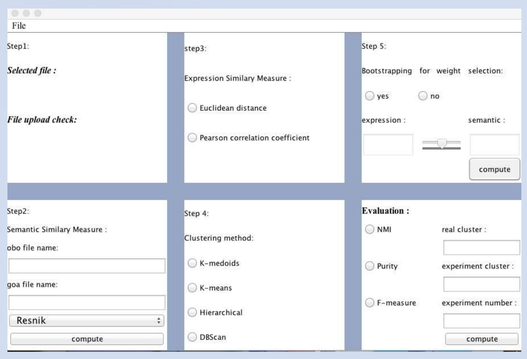
根據所讀的論文及相關資料，本專題設計之架構繪製於圖1。使用者首先經由我們所設計的系統介面，將所要運算的資料以特定格式儲存至資料庫中，這裡區分為三種資料，前兩者是基因表現資料與基因語意相似度資料；第三種evaluation benchmark儲存的是已知功能類別的基因資訊，之後會作為系統效能評估的標的資料。經由基因表現資料庫中所取出的資料，可讓使用者勾選一個或多個距離測度(像是皮爾森相關性及歐式距離)來得到基因表現相異度矩陣；而利用基因本體所提供的基因語意計算所得到的基因語意相似度資料，同樣也可讓使用者選擇一個或多個距離測度來得到基因語意相異度矩陣(方法有Resnik、simGraSM、simUI…等)。這兩種不同資訊來源所得到的矩陣再經過整合能得到混合式的距離矩陣(Hybrid distance measure)，當中決定這兩種distance measure 的權值(Weighting)方法有兩種，可以使用拔靴法(Bootstrapping method)，也可以讓使用者以手動方式設定權值(Manual setting)，例如基因表現相異度矩陣是0.2，則基因語意相異度矩陣是0.8，權值總和為1.0。



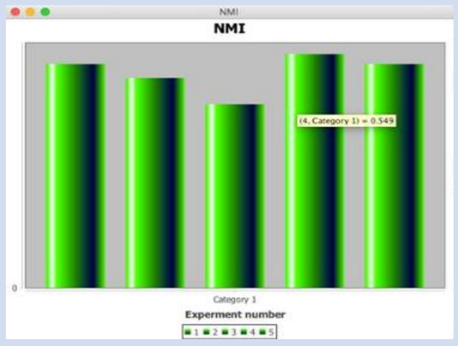
■ 圖1. 系統架構圖

得出混合式的距離矩陣之後，系統會提供多種分群法(Clustering)來讓使用者勾選想要的一個或多個分群法，做完分群法會進行分群效能評估(Evaluation)，這個部分我們會參考evaluation benchmark裡的資料來評估分群結果之生物意義關聯性成效，在這裡使用的方法是normalized mutual information (NMI) 以及F-measure，最後系統會將此次分群的評估結果輸出以提供使用者觀看與比較，並輸出圖形化結果顯示使用各種分群方法與使用不同距離測度的結果。

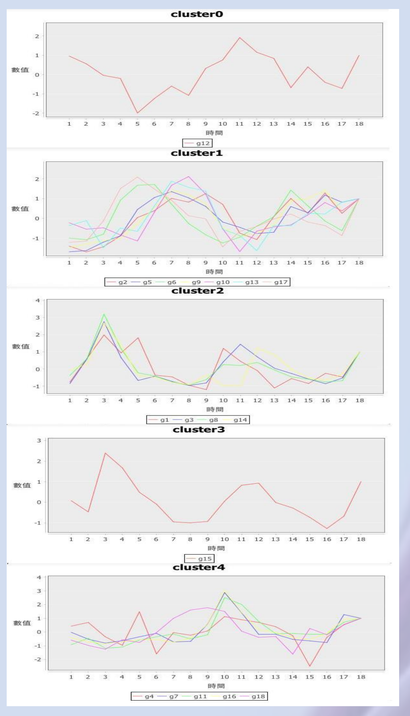
■ 單機版系統



■ 圖2. 單機版系統介面

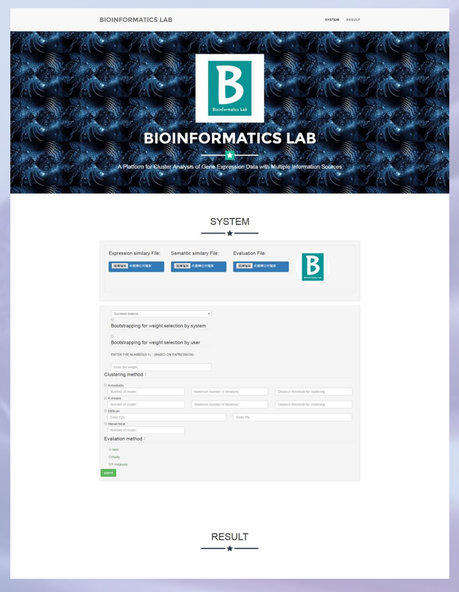


■ 圖4. 單機版系統評估結果

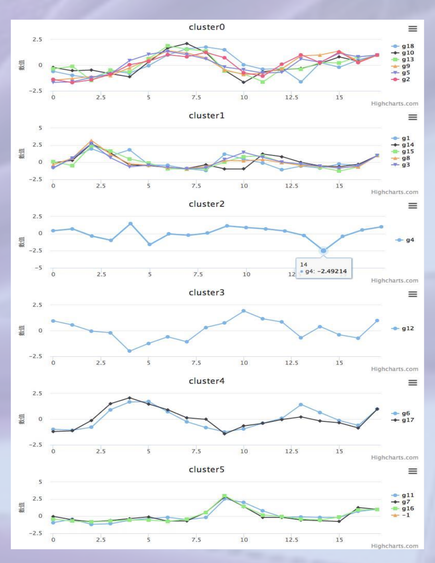


■ 圖3. 單機版系統分群結果

■ 雲端版系統(網頁)

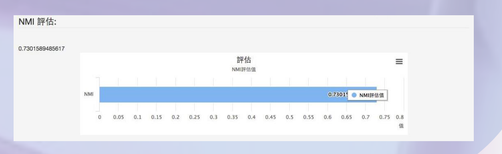


■ 圖5. 雲端版系統介面↑



■ 圖6. 雲端版系統分群結果

■ 圖7. 雲端版系統評估結果→



國立臺南大學-資訊工程學系
Department Of Computer Science And Information Engineering

Bioinformatics Lab

Bioinformatics Laboratory X NUTN CSIE

致謝：感謝科技部大專生研究計畫對本專題的部分經費補助，計畫編號105-2815-C-024-024-E。