網絡分析技術於農業基因體資料之應用-以水稻為例

農試所作物組 吳東鴻 賴牧謙

一、前言

隨著資訊科技與農業生物技術不斷的發展，快速累積了大量作物的基因體資訊，如蛋白質體學、轉錄體學、基因表現與次世代定序等重要性狀遺傳調控數據，產出速度快且成本持續下降。其中生物資訊學可將數據有效地轉換成資訊，並轉化為知識。然而不同體學的在串聯與整合上有一定困難，研究者欲以宏觀角度通盤檢視巨量基因體資料實屬不易。據此，網絡分析的概念應運而生。

網絡分析 (network) 的概念從20世紀中期開始興起，基於網絡理論分析複合層次資料的資料分析方法，針對資料表單藉由不同的演算法梳理出串聯節點的關聯性，並以視覺化的網絡圖呈現重點脈絡趨勢。本篇將導論網絡分析技術以及在植物科學領域已建立的入口平台，並以水稻基因體資料為案例操作，分別在 RiceNet (v2) 與 STRING (v11.5) 等2個線上平台操作流程，導入視覺化工具剖析巨量複合基因體資料的重要樞紐因子與脈絡趨勢，有助於農業從業人員深化科研成果與探索跨域加值的可能性。

二、網絡分析原理介紹

相較於制式的單件表格資料，網絡分析圖利用圖形與線條串聯多層資料集，以視覺化展示整體資料的架構與彼此間的關聯性。網絡由節點與連線粗細所構成，其網絡區塊大小根據節點的數量與連線數目決定。節點依據研究目的，可以由基因、蛋白質或生理調控路徑所組成，連線粗細代表兩節點之間的關聯性強弱。以蛋白質間交互作用網絡為例，每一個蛋白質代表一個節點，而節點之間的連線代表兩個蛋白質之間具有交互作用。而網絡的拓樸性質 (topology) 可由中心性指標衡量，其探討在幾何圖形或空間中連續改變其形狀後還能保持不變的性質，著重物體間的相對位置

作　　者：吳東鴻副研究員
連絡電話：04-23317106

而不考慮其大小與形狀，常見的衡量指標為度數 (degree) 和群聚係數 (clustering coefficient)。度數的定義為該節點的與其他節點的總連線數；群聚係數則是用來描述網絡圖中節點與節點之間關係的緊密程度，其數值介於 0 到 1 之間。

網絡分布的結構特性亦可由度數與群聚係數判斷，若各節點的群聚係數趨近於1，則代表網絡的內部結構相當緊密，具有完全網絡 (complete network) 的趨勢，但節點若過於密集，會導致可閱讀性降低；反之，若各節點的群聚係數趨近於0，則表示網絡的內部結構相當鬆散，呈現無尺度網絡 (scale-free network) 的性質，但若節點與節點之間連線數太少，則能從中獲取的資訊含量較低。其中度數較高的節點稱為樞紐節點 (hub node)，通常是相當重要的基因、蛋白質或者是生理調控路徑。

三、高等植物網絡分析平台簡介

至今已有許多高等植物網絡分析平台可供使用，如阿拉伯芥的 AraNet、大豆的 SoyNet、玉米的MaizeNet 及水稻的 RiceNet 平台 (圖一)。由於阿拉伯芥是開花植物中研究完整度最高的模式植物，因此網絡架構也最早被建立，Lee 等人相繼於 2010 年與 2014 年分別建立 AraNet 和 AraNet(v2) 兩個阿拉伯芥網絡功能分析公開平台。另外平台版本的升級通常是更換新的核心演算法與納入新的同源基因資料集，在推演未知序列的功能性時，可參考兩條基因序列的相似性，若達80%以上就視為同源基因，因此藉由其他物種已知功能的基因連結阿拉伯芥的同源基因，進而推論該阿拉伯芥基因可能具備哪些功能，豐富阿拉伯芥的網絡功能資料庫更趨完整。第一代的 AraNet 平台，除了利用阿拉伯芥前人文獻之 DNA 微陣列晶片資料庫外，還納入了 24 個由人類、蚯蚓、酵母菌之直系同源基因所建構而成的關聯性資料庫，總計 19,647 個基因能涵蓋 73% 的阿拉伯芥基因體。相隔四年，進階版本的 AraNet(v2) 平台，除了採用機器學習為核心演算法之外，擴充納入斑馬魚、果蠅、線蟲之直系同源基因至網絡資料庫中，使阿拉伯芥總體基因體覆蓋率上升至 84%。

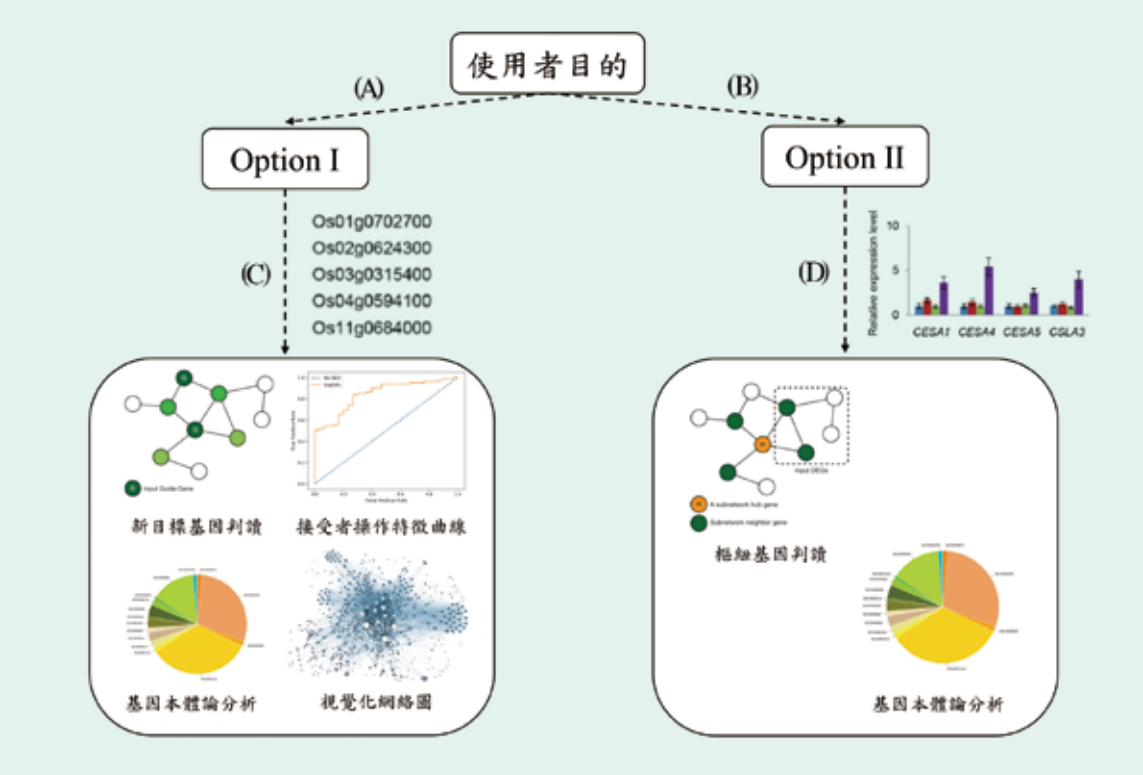
在水稻方面，同一個團隊的 Lee 等人也在 2011 與 2015 年分別建立了 RiceNet 和 RiceNet(v2) 兩個水稻網絡分

圖一、高等植物網絡分析平台。(A) 阿拉伯芥網絡分析平台；(B) 玉米網絡分析平台；(C) 大豆網絡分析平台；(D) 水稻網絡分析平台。

析平台。RiceNet 的建立利用水稻前人文獻之 DNA 微陣列晶片資料，並蒐集 24 個由人類、蚯蚓、蒼蠅、酵母菌之直系同源基因所建構而成的關聯性網絡資料庫，建構模式參照 AraNet，總計 41,203 個基因能約能涵蓋 50% 的水稻全基因組。進階版本的 RiceNet(v2) 平台不僅網絡基因涵蓋率達到 70.1%，亦新增了線蟲、果蠅和斑馬魚之直系同源基因所構成之蛋白質間交互作用網絡。網絡分析平台功能方面，則是新增了基因優化 (gene prioritization) 之新功能。基因優化的目的在於依據該基因與特定性狀的關聯性，透過演算法給予每一個候選基因分數，分數越高代表該基因與特定性狀之關聯程度愈高，反之愈低。

四、RiceNet(v2) 與 STRING(v11.5) 平台操作簡介

RiceNet(v2) 是一個免費公開的網路平台，主要功能在於利用網絡分析進行水稻基因清單優化，資料處理流程如圖二所示。首先，進入主頁面後點選Gene prioritization後有兩個選項，分別為Option Ⅰ:Gene prioritization based on network direct neighborhood 與 Option Ⅱ:Gene prioritization based on context associated hubs，研究者依據其需求選取相對應的分析工具。假若研



圖二、RiceNet(v2) 平台資料分析架構流程圖。(A) 方向1: 尋找新的目標基因；(B) 方向2: 找出參與其中的樞紐基因；(C) 候選基因匯入；(D) 差異表現基因集匯入。

究者的目的在於利用輸入的候選基因(candidate genes) 尋找新的目標基因，可以利用第一類型操作(Option Ⅰ)；若研究者的目的在於利用候選基因找出參與其中的樞紐基因 (hub genes)，則可以選擇第二類型操作(Option Ⅱ)進行分析。

第一類型操作(Option Ⅰ)方面，進入後須輸入欲進行分析之特定性狀的候選基因集。其中，基因名稱須符合編號共用原則，如粳稻 (*Oryza sativa* spp. *japonica*) 須以*LOC_Os01g01010*或*Os01g0100100*；秈稻 (*Oryza sativa* spp. *indica*) 則是*BGIOSIBCE000001*，其唯一的限制為輸入基因數大小不能超過500個。以平台所提供的逆境反應分子互動組 (stress response interactome) 範例基因為候選基因集，進行第一類型操作網絡分析之示範操作。第一項結果為接受者操作特徵曲線，用來判斷網絡模型的準確性，其統計檢定數值為曲面下面積 (area under curve, AUC)，和綠色對角線面積0.5相比，紅色曲線的面積通常大於0.7，數值越大代表網絡模型具有越好的鑑別能力。第二項結果為利用輸入基因集合找出的新基因。透過基因優化後所找出的新基因依據和輸入基因集合的總關聯性 (total connectivity) 得分由高至低排序，第一名的基因*Os03g0285800*的總關聯性得分為42.28。除此之外，表格中還會詳列出該基因所參與基因本體論 (gene ontology, GO) 之生理調控路徑，如*Os03g0285800*參與MAPK activity involved in osmosensory signaling pathway與response to hypoxia等訊息傳遞與生理調控路徑。

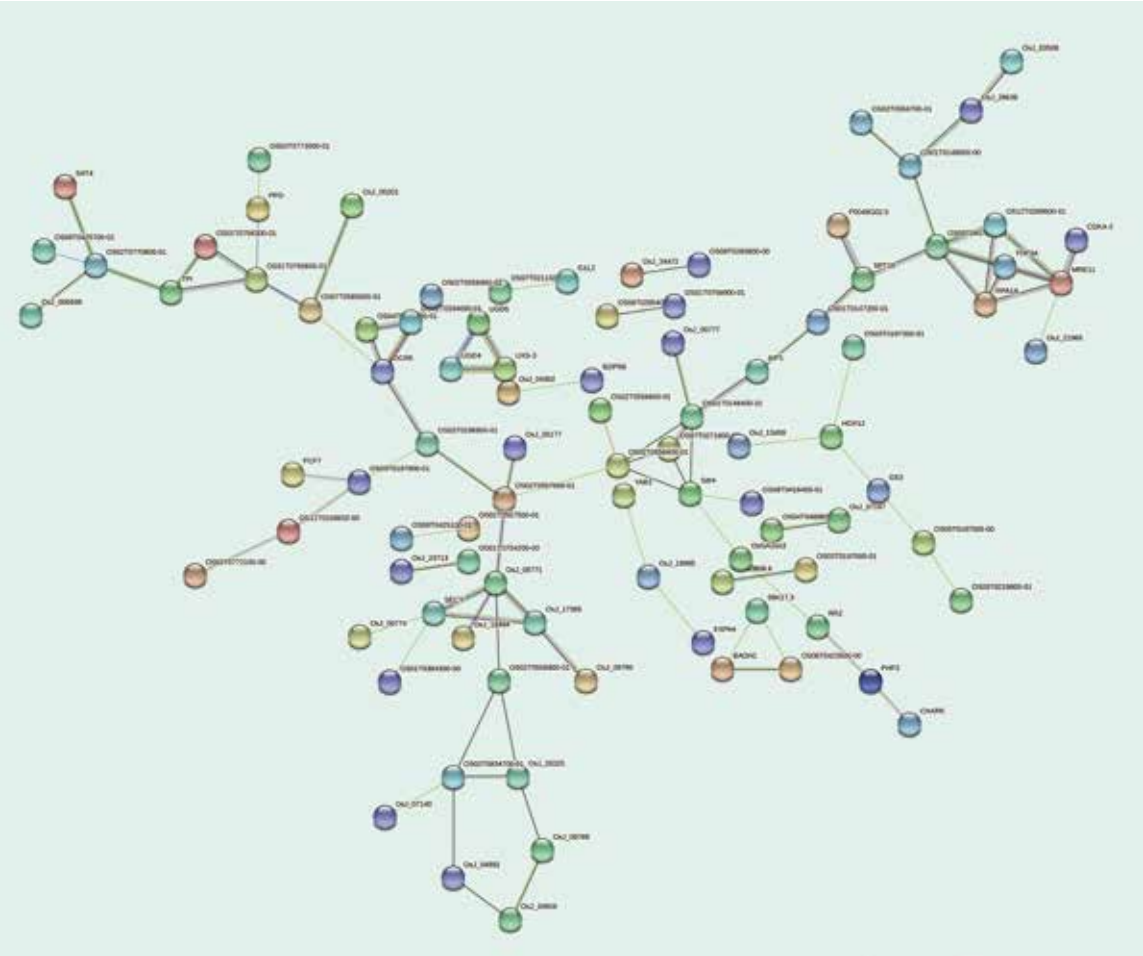
Option Ⅱ方面，輸入的基因集合主要由差異表現基因 (differentially expressed genes, DEGs) 所構成，如RNA定序資料和定量及時逆轉錄聚合酶連鎖反應資料。基因名稱和輸入基因集合大小規定與Option Ⅰ相同。以平台所提供的粳稻白葉枯病抗性基因*Xa21*所調控之範例DEGs集合為例進行分析。針對輸入基因對應的P值由小到大排列，P值越小代表該基因與其他基因的關聯性越高，在特定生理調控途徑的地位愈顯重要，愈有可能是樞紐基因。表格中亦提供該基因所參與之GO生理調控路徑，以利研究者進行分析。

在視覺化的網絡圖分析方面，由於 RiceNet(v2) 平台是利用Adobe flash player 產生網絡圖，但主流網頁瀏覽器(如Chrome、IE) 都已在2021年1月開始不支援Adobe flash player，因此 RiceNet(v2) 平台無法順利顯示網絡圖。據此，筆者推薦使用 STRING(v11.5) (https://string-db.org/) 平台產製網絡圖。STRING(v11.5) 為一個線上蛋白質-蛋白質交互作用網絡平台兼資料庫，最新的版本 version 11.5 於2021年8月更新，蒐集來自 14,094 個物種，67,592,464 個蛋白質共20,052,394,042 筆蛋白質交互作用資料。STRING(v11.5) 具有圖形化的操作介面，輸入資料的格式可以是基因名稱 (如*Os01g0146000*) 或者是蛋白質名稱 (如CDC15)，平台則

會一律轉換成蛋白質並計算關聯性。STRING(v11.5) 平台計算關聯性的來源從 KEGG (Kyoto Encyclopedia of Genes and Genomes) 等資料庫與實際上進行複合蛋白質純化後確認其真實狀況來進行整合與驗證，亦利用發表文獻進行整理。其中，不同粗細的連線表示蛋白質間關聯程度的強弱，不同顏色則代表來自不同資料庫的驗證 (圖三)。產生的結果可以依據研究者需求的格式進行輸出，如高解析度網絡圖片、蛋白質序列資料等。

五、結語

在這資訊爆炸的21世紀，基因體分析工具種類多樣、精確度高且成本大幅降低，根據美國國家衛生研究院統計，全基因體定序費用自2008年的100萬美金，相隔十年後的2018年只需要1千元美金即可完成。網絡分析技術面對龐大的基因體大數據抽絲剝繭，以條理分明的圖形來闡釋生物體複雜的生理調控機制，對農業研究人員來說不失為一項新興的研究利器。



圖三、蛋白質-蛋白質交互作用網絡圖。其中，節點代表蛋白質；連線的粗細代表關聯性的高低，越粗代表關聯性越高，反之越低。