生物資訊的研究與應用

文．圖／陳倩瑜

生物資訊（Bioinformatics）和計算生物學（Computational Biology）是兩個經常被交替使用的名詞，泛指利用計算方法研究細胞中各種巨分子之間的交互作用，進而瞭解生物分子運作機制的學科。生物資訊研究方法涵蓋演算法、統計、機器學習（Machine learning）以及近日很熱門的深度學習（Deep learning），除了資料分析，也包含各種序列和結構資料庫的建置，將各式資料分享給全世界的研究人員。而其研究素材除了DNA外，還有RNA和蛋白質，因為DNA的差異會影響RNA 的數量，也會影響蛋白質功能，各物種間 DNA 的差異是物種鑑別的基礎，而個體間的 DNA差異則是形成不同性狀的主因。生物個體中 DNA和性狀之間的關聯性就是生物資訊要探討的，如：DNA如何影響面貌？如何影響身高？如何影響罹病風險？為什麼有些昆蟲不怕農藥？為什麼有些植物不怕蟲害？帝雉為什麼適合生長在高海拔山區？沉香的香味和基因體有關嗎？

讓農民頭痛的東方果實蠅

昆蟲抗藥性一直是控制農業害蟲的關鍵議題。基因體定序（Genome sequencing）和轉錄體（Transcriptome）的定序與定量可能可以提供解決這個困難問題的解方。東方果實蠅（*Bactrocera dorsalis*）是世界上最具破壞性的農業害蟲之一，最近已被應用於研究昆蟲抗藥性相關的遺傳機制。然而，在和臺大昆蟲系許如君老師合作研究之前，與東方果實蠅的相關研究大多侷限於透過鄰近模式生物 ——果蠅（*D. melanogaster*）——的基因序列。為了提供昆蟲抗藥性更廣泛的研究素材，我們使用次世代定序（Next-generation sequencing，NGS）產生的短序列進行東方果實蠅全轉錄體分析，並使用所組裝的序列識別大量可能與昆蟲抗藥性相關的基因家族[1]。這個研究一共註解了90個P450、42個GST和37個COE相關基因，這三個酵素家族是主要影響昆蟲抗藥性中代謝和抗藥能力的基因。此外，我們發現36個基因序列含有與四類抗藥性基因相關的標靶點位，同時也分析了胺基酸序列中的序列特徵，並將它們與特定的基因和蛋白質功能關聯起來。

不怕豆象的野生綠豆

綠豆（Mungbean）是南亞和東南亞具有高營養價值的重要豆科作物，這種作物植株易受到豆象（Bruchids）這類害蟲侵害造成儲存上的經濟損失。有些野生和栽培的綠豆

品系對豆象顯示出抗性，我們與中央研究院植物暨微生物學研究所陳榮芳研究員合作研究，進行抗豆象綠豆和易感綠豆的基因體和轉錄體比較，可能揭示基因體中與豆象抗性相關的基因，並提供對綠豆抗豆象性的瞭解[2]。在本研究所測試的綠豆品種中，抗豆象綠豆的基因體大小變化了61Mb（百萬鹼基對），我們使用次世代定序對一個抗豆象品系（RIL59）的基因體進行組裝，釋出超過42,000個基因。針對抗豆象綠豆和易感綠豆的轉錄體比較則找到91個差異表達基因（Differentially expressed genes，DEG），註解為17個主要和74個次要的豆象抗性相關基因。我們在68個DEG的啟動子範圍內找到408個核苷酸變異。此外，在148個蛋白質上找到282個變異。這些基因大部分被定位到染色體5上的一個區域，該區域對於豆象抗性基因型具有很高的診斷性，未來可以應用於開發豆象抗性的綠豆品系。

比黃金還貴的沉香

沉香是珍貴的藥用植物，然對於其基因體的知識仍極為有限。

沉香木是來自沉香屬的樹木，其在國際市場交易已經受到《瀕危野生動植物種國際貿易公約》附錄II的嚴格管制。許多沉香木的次級代謝物被認為對人類具有藥用價值，包括一些已經被證明具有鎮定效果和抗癌性質的化合物。然而，關於沉香木的基因體、轉錄體以及負責產生這些次級代謝物的生物合成途徑的知識非常有限。在與中央研究院植物暨微生物學研究所陳榮芳研究員的研究合作中，我們釋出了一個來自沉香木（*Aquilaria agallocha*）的基因體草圖[3]，並提出產生葫蘆素E和I（已知具有藥用價值的次級代謝物）之可能生化途徑。DNA和RNA數據被用來註釋基因體草圖中的許多基因和蛋白質功能，葫蘆素E和I的表達變化與已知的*A. agallocha*對生物壓力的反應一致。這項研究是第一次嘗試從實驗室栽培的沉香木中識別出葫蘆素E和I，並為沉香屬中的首個基因體草圖，所提供的分子數據將有助於未來研究沉香屬和其他非模式藥用植物的次級代謝物途徑。

帝雉——高海拔山區的美麗身影

帝雉（*Syrmaticus Mikado*）是一種幾乎瀕臨絕種的物種，原生於臺灣的高海拔地區。帝雉提供研究學者一個機會來研究地理隔離後的演化過程，目前其遺傳背景和適應性演化仍不清楚。我們和生物技術研究中心合作，組裝了帝雉的基因體草圖[4]，它由1.04 Gb的DNA和15,972個蛋白質編碼基因組成。帝雉的基因體內容顯示其與適應性演化

相關的特性，如能量代謝、氧氣運輸、血紅蛋白結合、輻射反應、免疫反應和DNA修復。帝雉的主要組織相容性複合體（MHC），與原雞的相同基因之序列相比，顯示出高度的同源性和幾個快速演化的基因。本研究也對完整的粒線體基因組進行了定序組裝，並與4種長尾雉進行比較，分子時鐘分析的結果表明，帝雉的祖先大約在347萬年前從北方遷移到臺灣。這項研究為帝雉提供了寶貴的基因資源，對其適應高海拔的洞察，以及對Syrmaticus屬的演化史，對未來研究分子演化，基因體學，生態學和免疫遺傳學提供基礎。

應用生物資訊學研究臺灣瀕危的特有種——帝雉。（攝影／謝郁震）

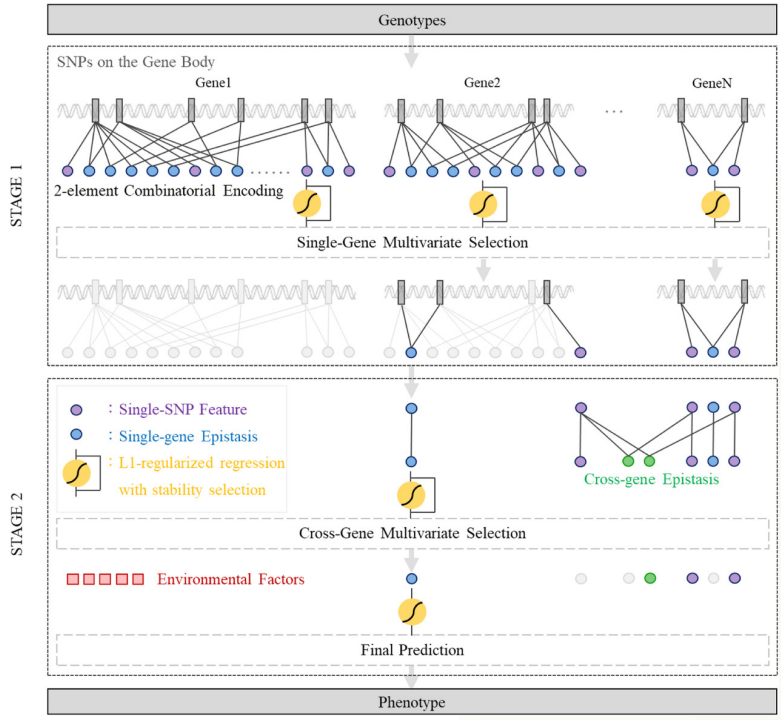
尋找阿茲海默症的基因交互作用

全基因組關聯研究（Genome-wide association study，GWAS）提供了一種有效的方式來偵測遺傳變異與人類疾病或動植物性狀之間的關聯。然而，目前GWAS應用於尋找遺傳變異之間交互作用的能力仍然有其侷限性，為此，一個名為GenEpi的計算方法[5]，透過所提出的機器學習方法來尋找與複雜疾病相關的基因交互作用，這對於阿茲海默症（AD）等複雜疾病尤其重要。GenEpi通過兩階段流程來識別基因內和跨基因的交互作用， GenEpi在產生特徵時採用兩元組合編碼，並通過具有穩定選擇的L1正則化迴歸來構建預測模型。模擬數據顯示，GenEpi在檢測表現性方面優於其他廣泛使用的方法，在真實數據上，本研究以AD為例，揭示了GenEpi在找到與疾病相關的變異以及變異交互作用方面的能力。模擬數據和AD的結果表明，GenEpi具有高效檢測與表型相關的能力，所釋出的程式碼可廣泛地促進許多複雜疾病的研究。

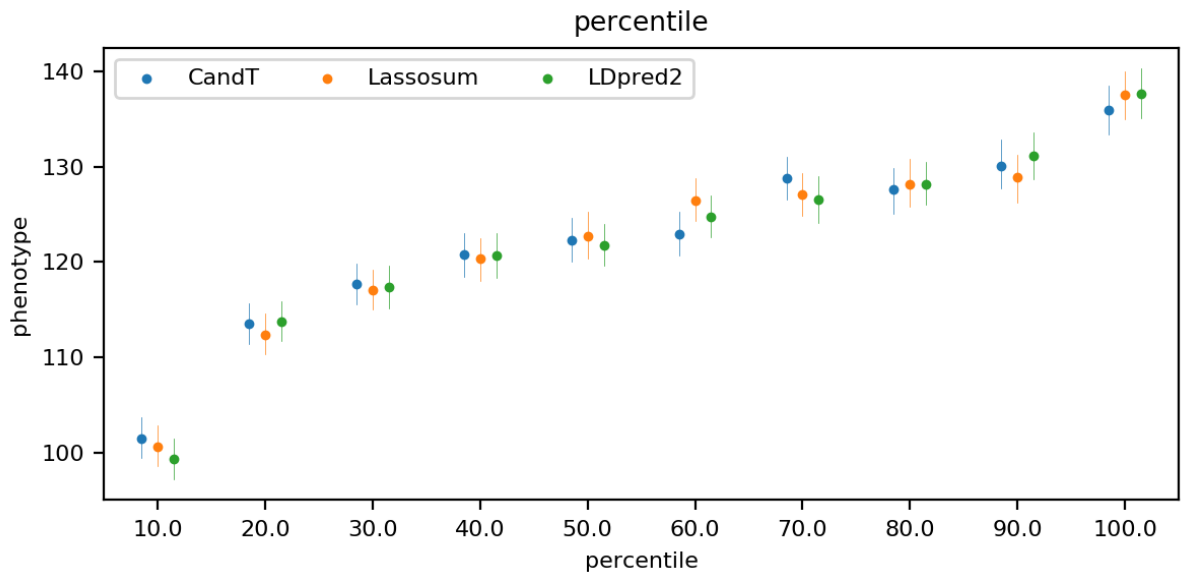
個體差異與疾病風險

多基因風險評分（Polygenic risk score，PRS）在這幾年突然變成一個熱門名詞，但其實它的概念並不新穎，簡單來說，就是把一個人的基因型（genotype）當成特徵，每一個SNP都是一個變數，使用成千上萬的特徵建立一個線性或非線性的分類或迴歸模型，每一個模型分別用來評估一個人在一生中會得某種疾病的風險，或是預測一些可量化的屬性，例如：身高或體重。在與臺灣人工智慧實驗室的基因 AI 團隊的合作中，我們用低密度脂

蛋白（low-density lipoprotein，LDL）為範例說明 PRS 模型的效用。這個模型使用臺大研究團隊向臺灣人體生物資料庫申請的6萬人 SNP 基因晶片資料作為示範，其中 68,080 人當作訓練資料，6,898人當作測試資料，下圖是三種 PRS 模型在 6,898 筆測試資料的評估結果。由圖可以看出，不論是哪一個方法（C+T、Lassosum 或 LDpred2），在將測試資料的6千人依單一模型預測的分數排序後，模型認為高風險的人（percentile 90~100%），這群人的 LDL 數值平均的確都比較高。我們知道 LDL 通常被稱為“壞的”膽固醇，因為過高的LDL可以導致一些健康問題。美國 CDC 告訴民眾：「預知你得到某種疾病的風險程度，將能幫助你採取行動去預防它，或是在疾病相對容易處理的階段就早期發現它」，這突顯各種 PRS 模型在未來的臨床醫學可能作為預防性醫學的有效工具。



GenEpi通過兩階段流程來識別基因內和跨基因的交互作用。



應用PRS模型分析臺灣人SNP 基因晶片資料。

在AI加持下的未來

尋找基因體中可能和性狀相關的區域只是生物資訊的起點，更多生物資訊研究所開發的演算法旨於從高通量的生物技術所產生的大量數據，建立可用於瞭解生物系統或進而預測系統行為的模型，這類研究方法之於人類疾病致病機轉的瞭解，或是設計藥物治癒疾病將扮演非常重要的角色。機器學習在生物資訊研究的發展中一直是不可或缺的工具，近幾年更因為引入深度學習幫助突破解決許多重要的生物問題，例如：蛋白質結構預測和藥物設計。生物資訊學的研究對醫學有深遠影響，同時，它在物種保護、作物分子育種及病蟲害管理等領域也展現出顯著的效益。熱門的基因編輯技術對於醫療、農業與生態都有廣泛的應用前景，我們可以預見，在不久的將來，人類將透過掌握基因資訊來實現許多重大的變革。（本期專題策畫／生農學院李達源副院長&公衛學院郭柏秀副院長）

參考文獻：

[1] Hsu, J.-C., et al., *Discovery of genes related to insecticide resistance in Bactrocera dorsalis by functional genomic analysis of a de novo assembled transcriptome*. 2012.

[2] Liu, M.-S., et al., Genomic and transcriptomic comparison of nucleotide variations for insights into bruchid resistance of mungbean （Vigna radiata [L.] R. Wilczek）. *BMC Plant Biology*, 2016. 16（1）: p. 1-16.

[3] Chen, C.-H., et al., Identification of cucurbitacins and assembly of a draft genome for Aquilaria agallocha. *BMC genomics*, 2014. 15: p. 1-11.

[4] Lee, C.-Y., et al., Whole-genome de novo sequencing reveals unique genes that contributed to the adaptive evolution of the Mikado pheasant. *GigaScience,* 2018. 7（5）: p. giy044.

[5] Chang, Y.-C., et al., GenEpi: gene-based epistasis discovery using machine learning. *BMC bioinformatics*, 2020. 21: p. 1-13.

陳倩瑜小檔案

陳倩瑜博士是臺灣大學生物機電工程學系的教授。她於 2003 年從臺大獲得資訊工程學的博士學位。在此之前，她在臺大（1996）與史丹福大學（1998）分別獲得電機工程學的學士和碩士學位。她的研究專長包括生物資訊學、機器學習和基因調控。她的實驗室開發計算方法來解決使用基因體、轉錄體或表觀基因體數據的生物問題。她的主要研究興趣是透過建立機器學習和深度學習模型來預測變異的致病性，標註人類基因體的功能區域（例如增強子、轉錄因子結合位點、eQTL 等）及其效應。最近，她的研究合作團隊建立了一個名為 TaiwanGenomes 的變異數據庫，並發布了一些用於研究臺灣人的遺傳數據的多基因風險評分（PRS）模型。