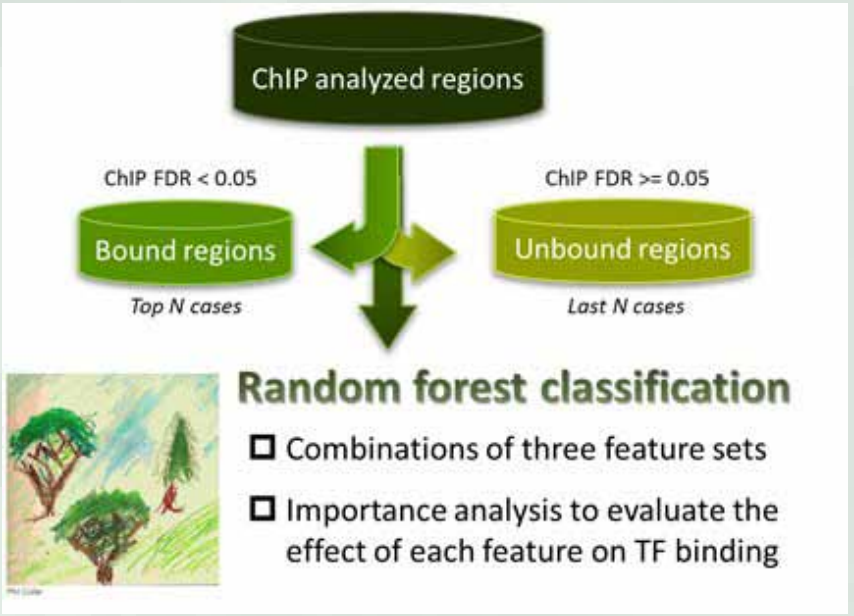
運用生物資訊方法
研究生物體內轉錄調控機制

蔡懷寬　研究員



應用隨機森林分類器探討轉錄因子結合機制。將分子生物問題與實驗數據轉化為資訊科學擅長解決的分類問題，並從解答過程中更深入的探討分子機制（圖一）。

生物體內的轉錄調控機制對生命體的運作極為重要，藉由轉錄因子與其結合位置的動態結合，基因得以精確地對生物過程與周遭環境完成適當反應，因此辨識轉錄因子之結合位置，從而量化細胞內的基因調控情形，一直是生命科學中重要的課題。然而，精確辨識轉錄因子之結合位置是項大難題，因為大多數的結合位置不僅很短（大約 5-20 個鹼基對，但人類的基因體長度超過三十億個鹼基對），而且具有高度退化的特性（簡單來說，轉錄因子可以與類似序列的 DNA 結合，不須完全相同）。隨著生物科技的進步，包括 EMSA、SELEX、PBM 以及 ChIP 等許多生物實驗技術已被發展來偵測辨識轉錄因子之結合位置。這些方法提供生物體內轉錄因子與 DNA 交互作用的證據，以及在生物體外量測轉錄因子與 DNA 結合的親和力。然而到目前為止，由於這些技術仍有高成本與高耗時的問題，該如何在全基因體中大量快速且精準的預測（尋找）轉錄因子之結合位置，仍舊是一項棘手的課題。而生物資訊，正提供了一個解決的可能方向。

目前已有許多計算生物學研究預測轉錄因子的結合位置，這課題可以簡化為特殊序列的尋找問題，也就是從給定的一組生物資料中（例如某轉錄因子目標基因的啟動子序列）尋找特殊的 DNA 序列片段。這些片段通常利用位置權重矩陣 (PWM) 來表示，位置權重矩陣是利用已知的結合位置產生一矩陣，可利用該矩陣來代表結合序列的強度。目前用來尋找轉錄因子結合位置之特殊序列的演算法主要可分為兩種，分別為列舉法以及概率法。列舉法主要精神為分析所有字串出現的頻率，然後找出經常出現的字串作為基礎來生成位置權重矩陣。而概率法之原理則是先將給定的生物字串建立多程序列比對 (MSA)，然後利用機器學習方法（例如 EM 演算法或是 Gibbs