知識天地

基因表現的精巧控制：利用機器學習探勘轉錄因子的結合偏好

蔡懷寬研究員、黃佳欣博士後研究（資訊科學研究所）

摘要

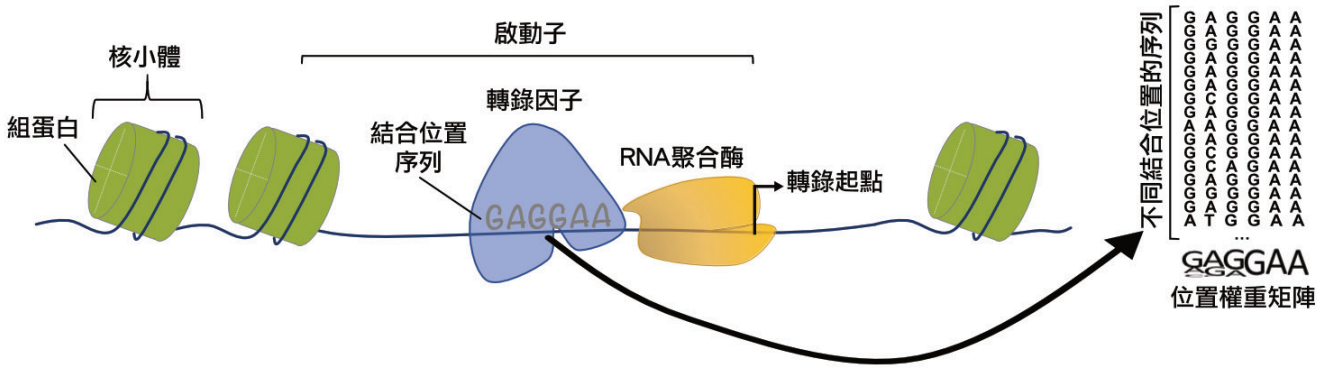
基因表現的第一步是轉錄，轉錄過程由許多蛋白質共同調控，其中一類為轉錄因子，會結合在特定DNA序列並調控轉錄過程。然而，傳統上利用特定DNA序列作為預測轉錄因子的結合位置並不完全正確。近年的研究顯示DNA的結構與染色質狀態都會影響DNA對轉錄因子的親和性。本研究利用機器學習技術整合各類DNA特徵包含序列、結構與染色質狀態探勘各類特徵對於預測結合位置的重要性，並歸納出三個重要的特徵可以精準地預測轉錄因子的結合位置。

基因轉錄表現的調控概觀

人類基因體的DNA總長度大約有三十億個鹼基對，雖然人體全部的細胞都使用同一組基因體，但是在不同的器官跟組織中只會有特定的基因被轉錄表現，目前認為，決定這些基因的轉錄與否主要由一群會結合在基因上游DNA的蛋白質來調控，而這些蛋白質被稱作為轉錄因子。轉錄因子跟DNA的結合是細胞控制基因表現的關鍵，例如人類大約有一千個不同的轉錄因子，透過搭配不同轉錄因子跟DNA的動態結合使得細胞能夠精確地讓生物體得以成長發育，並且對周遭環境完成適當反應跟調控。辨識轉錄因子之結合位置可以用來量化細胞內的基因調控情形，因此一直是生命科學研究中的重要課題之一。隨著生物科技的進步，諸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。這些實驗方法提供生物體內轉錄因子與DNA交互作用的證據，或是在生物體外檢測轉錄因子與各種合成DNA序列的親和程度。然而，由於這些技術仍有高成本與高耗時的問題，再則，轉錄因子偏好的結合位置通常是很短的DNA序列（約5-20個鹼基），其結合位置上的鹼基序列也不會完全相同。如何在整組基因體序列中快速大量且精準的尋找轉錄因子結合位置，至今仍是一項艱難的研究課題。而生物資訊正好提供解決這個棘手問題的可能方向。

預測轉錄因子的結合位置

轉錄因子經常會結合在目標基因的上游啟動子（promoter）區域，以往計算生物學（computational biology）將研究轉錄因子的結合位置轉換並簡化為尋找特殊序列的問題，也就是試圖從給定某轉錄因子的一群目標基因的啟動子序列中尋找重複出現的特殊DNA序列片段，或是統整經由生物實驗方法所收集到跟某轉錄因子結合的DNA序列片段（圖一）。由於這些結合位置之特定DNA序列的鹼基並不總是完全相同，經常以位置權重矩陣（position weight matrix, PWM）來表示，位置權重矩陣是利用已知的結合位置的序列產生一矩陣，該矩陣的每個位置是以量化數值來表達各個鹼基（A, T, C, G）出現的頻率高低，即在同一位置上某鹼基在所有已知結合序列出現的次數越多則數值越高。除了特定位置上的序列可能不同，生物實驗方法所收集到的結合位置之特殊序列也非固定的長度。而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。列舉法的主要概念為分析所有鹼基序列組合字串出現的頻率，然後找出經常出現的鹼基序列字串作為基礎來產生一組固定長度的位置權重矩陣。而概率法之原理則是先將給定的生物序列的鹼基組合字串建立多程序列比對（multiple sequence alignment, MSA），然後利用機器學習方法最佳化多程序列比對與位置權重矩陣的參數。將轉錄因子偏好的特殊結合序列轉換成位置權重矩陣可以方便應用在全基因體上的結合位置之預測。



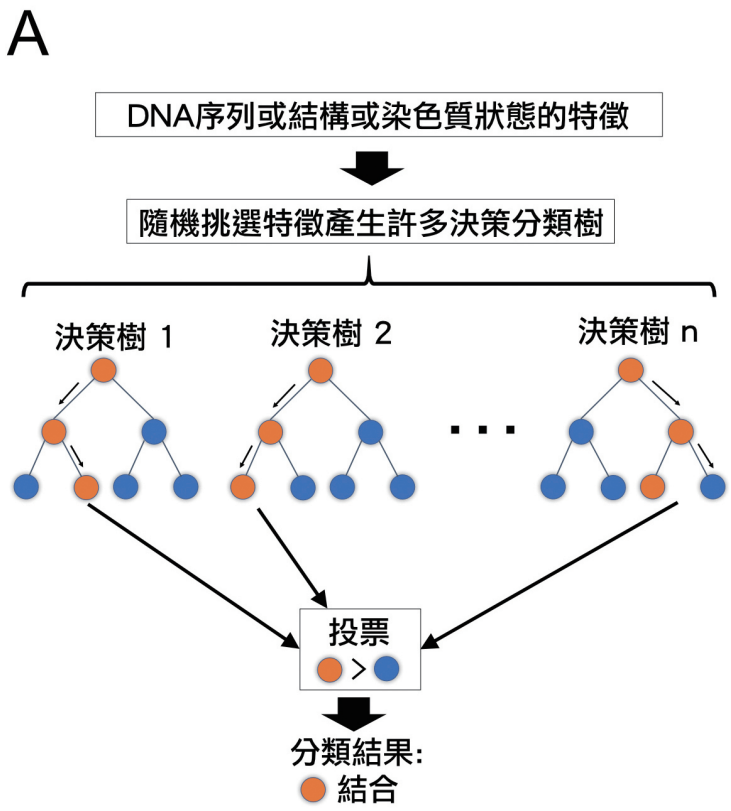
圖一、真核生物基因轉錄調控的基本架構，計算生物學經常以位置權重矩陣（position weight matrix, PWM）來表示轉錄因子結合位置序列，字母大小表示該鹼基在所有結合位置上出現頻率的高低。

近日有賴於高通量實驗技術如次世代定序（next generation sequence, NGS）的蓬勃發展，以及大量全基因體的資料累積，越來越多的證據顯示DNA序列並非決定轉錄因子結合的唯一因素，許多研究顯示即使是符合位置加權矩陣的特殊序列也不一定真的會發生轉錄因子結合的現象。這代表我們對於轉錄因子跟DNA的結合還有其他的因素需要列入考量。

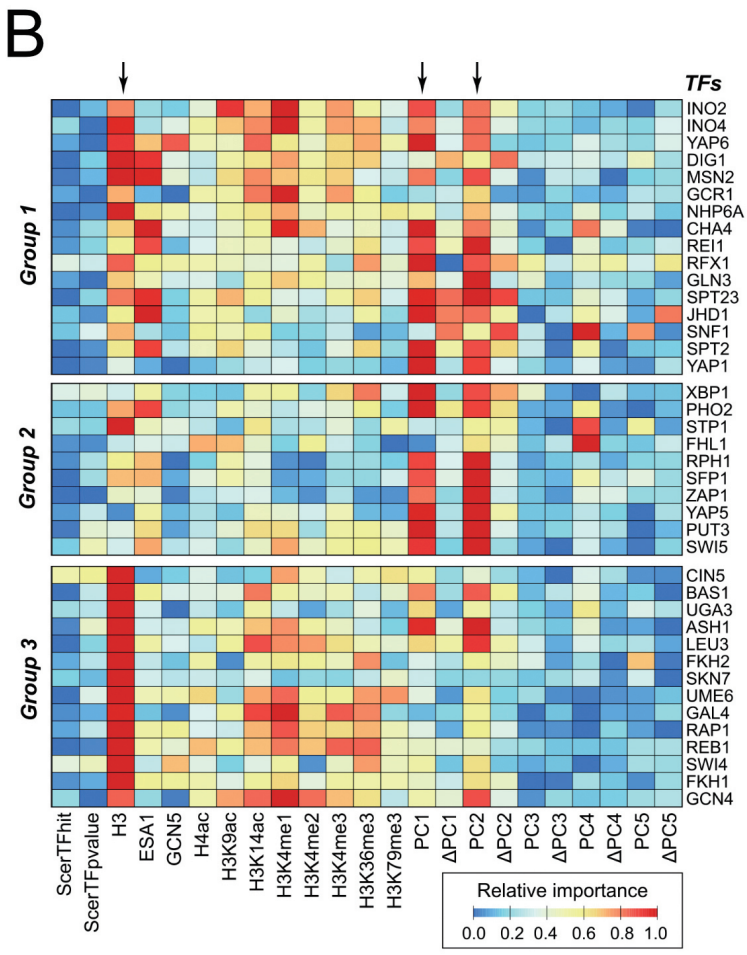
真核生物的染色質並非只是一條長長的DNA分子，實際上每一小段DNA會纏繞在八個單位一組的組蛋白（histone）形成核小體（nucleosome）的結構。當轉錄因子要與染色質（即DNA）結合的時候，必須解開核小體的結構使DNA分子得以鬆綁。另外，鬆散的DNA分子也會自己纏繞而形成特殊的物理性結構，類似繩結的構造。從近年的分子生物學研究，已知核小體的結構會造成染色質缺乏可親和性（accessibility）以及DNA的纏繞結構都會減低跟轉錄因子的結合，因此即使基因體的DNA有一段符合位置加權矩陣的特殊序列，也不一定會發生真正的結合與調控基因表現。最近已經有其他研究證實染色質的可親和性或DNA纏繞結構特性跟控制轉錄因子結合與否具有良好的相關性。雖然日前已有研究改採用這兩種染色質特性來預測轉錄因子結合位置，但目前比較各個影響染色質的特徵對於預測的效果付之闕如，且並沒有任何預測方法能成功整合以上所有特徵包括DNA序列，染色質可親和性狀態，與DNA的纏繞結構來預測轉錄因子的結合位置。因此我們以模式生物酵母菌為實驗對象，應用機器學習技術（這裡我們用了隨機森林）探討各特徵對轉錄因子結合之影響力。

隨機森林演算法探勘DNA特徵跟轉錄因子的結合偏好

隨機森林（Random Forest）機器學習演算法是以隨機方式建立許多決策樹形成一個森林，每一個決策樹包含上述提到DNA序列或結構或染色質狀態的部分特徵來將某段基因體上的區域分類成可能或不能被轉錄因子結合，最後統合整個森林的投票結果來給定最後的預測答案（圖二A）。經由我們的研究發現，在預測轉錄因子結合狀態時，使用染色質狀態或DNA結構特性建立的分類器會比傳統上單用DNA位置權重矩陣的序列要好。此外，同時考慮染色質狀態與DNA結構特性會使分類效果顯著地提升，充分顯示這兩種特徵具有互補效應。進一步觀察則發現，不同特徵對於預測不同轉錄因子的結合位置時大相逕庭，顯示不同的轉錄因子有著相異的結合機制。我們最後歸納出三個直接自DNA序列轉換而得的特徵（nucleosome occupancy, major groove geometry, and dinucleotide free energy）對轉錄因子結合有最顯著的影響，而整合該三項特性的資訊方法亦能得到良好的轉錄因子結合預測結果（圖二B）。這項研究的成果可以廣泛地應用於任何已定序之物種，能單就其DNA序列來預測轉錄因子結合的可能區域。該研究之相關成果發表於PLoS Computational Biology, 11(8), e1004418。



圖二、(A) 應用隨機森林分類器探討轉錄因子結合機制。



圖二、(B) DNA 序列、染色體狀態、與 DNA 結構特性三大類特徵各自對不同轉錄因子結合位置預測的重要性。圖中箭頭所指為對於所有轉錄因子都相對重要的三項特徵。