知識天地

生物資訊在基因調控及基因預測上的研究

蔡懷寬（基因體研究中心博士後研究員）、莊樹諄（基因體研究中心助研究員）

一個生物如何調控它的基因？人類有哪些基因？這是長久以來人們一直都很感興趣的兩個問題。隨著資訊科學在生物科技上的應用越來越普遍，電腦已成為探索這些問題不可或缺的重要工具。在此，我們介紹兩個應用相當廣泛的研究方向，基因調控預測以及基因體註解（或基因落點預測）。

基因調控預測

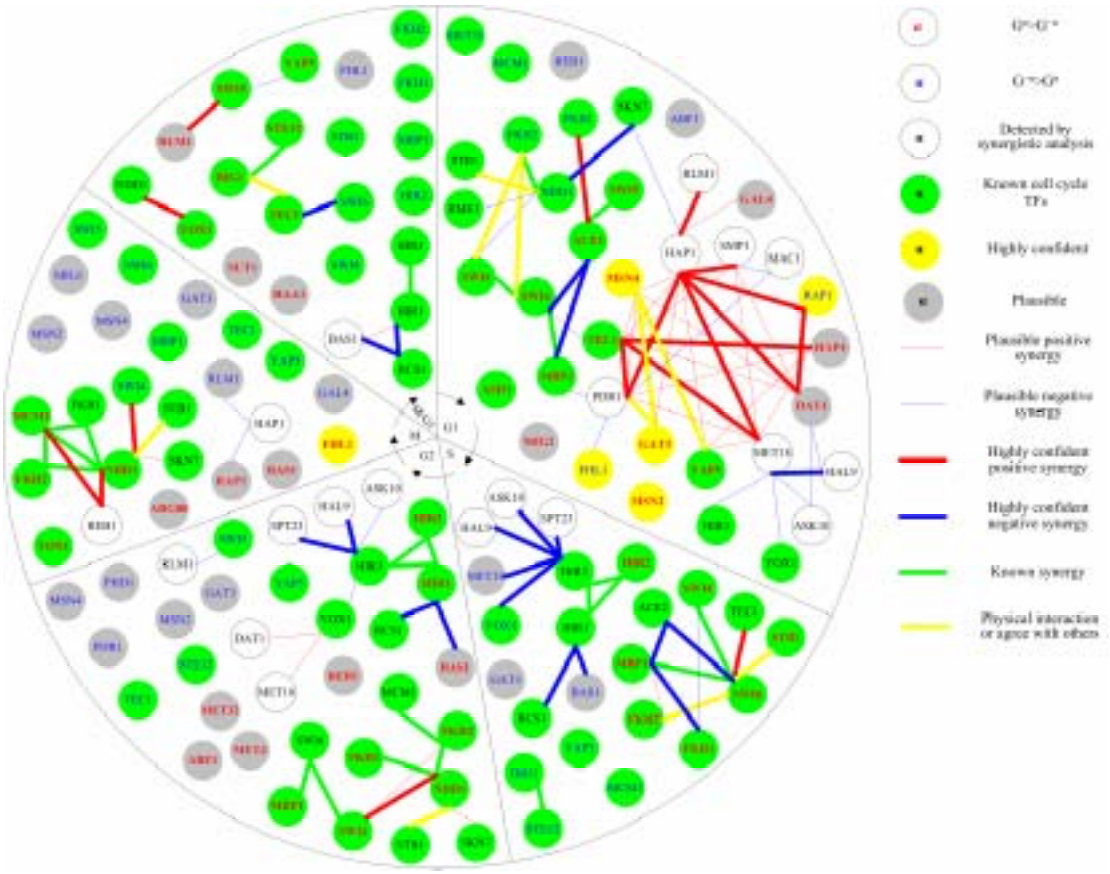
真核生物的細胞規律地進行 DNA 複製，染色體分離以及有絲分裂，形成有趣的細胞週期現象（cell cycle）。要瞭解這些參與在細胞週期中的基因如何被調控，分析哪些轉錄因子（transcription factors）調控細胞週期是非常重要的。轉錄因子為基因表現的調控因子，會辨認基因的序列前頭有段被稱為「啟動子」（promoter）的特殊序列，以便在適當的時機開啟該基因的表現，也藉此控制基因表現的程度。轉錄因子與 DNA 之間的交互作用控制許多重要的生理作用；例如面對發育及環境變化的處理機制。這些轉錄因子的缺陷可能引發疾病。

通常我們都利用一些模型生物（model organisms）來研究細胞週期。酵母菌不僅是生活中不可或缺的「飲食父母」（我們的美酒跟麵包可是得靠它呢），更是現今最佳的模型生物。因為它的基因體，遺傳學及生理學在真核生物中是最簡單，且瞭解最清楚的。更重要的是，它的實驗操控的技術最為成熟。因此我們利用基因表現晶片（microarray）與染色質免疫沈澱晶片（chromatin immunoprecipitation, ChIP）資料，並結合統計學與資訊學來找出酵母菌細胞週期的基因調控轉錄因子。

我們最主要的觀念是：若某一轉錄因子為酵母菌細胞週期的調控因子，則其所調控的基因在某些階段（phases）的基因表現量將會與沒有被其所調控的基因表現有明顯的差異。基於這樣的想法，我們先利用染色質免疫沈澱晶片技術以及文獻，對每一個轉錄因子找出兩組基因群，分別為可能調控基因以及最不可能被其調控的基因。接著利用 Kolmogorov-Smirnov (KS) 檢定法偵測此兩組基因群在細胞週期某一（些）階段的基因表現是否顯著不同。若在某階段顯著不同，則可推測此轉錄因子為細胞週期調控因子，並在此階段進行活化或抑制的功能。我們並延伸這樣的想法，利用多變數分析（ANOVA）來找出共同調控基因表現的調控模組。我們成功的預測了 50 個轉錄因子之間相互作用形成的網絡以及其細胞週期範圍與功能（見圖一）。

本篇研究最大的突破在於，我們的方法除了可以偵測單一或具協調作用轉錄因子作用的時間，並同時可以指出它的功能。此外，因為方法的特性，所需要的基因表現晶片的數量可以大幅減少，甚至於單一晶片即可。最重要的是，除了細胞週期的分析之外，此方法可以很容易的應用到其他生化反應，例如半乳糖反應、出芽、新陳代謝以及因應外在環境改變的基因調控等等。

本篇文章發表在美國國家科學期刊： Proc Natl Acad Sci, 2005, Vol. 102, No. 38. pp.13532-13537。



圖一：參與細胞週期的轉錄因子及其協調作用之預測圖。

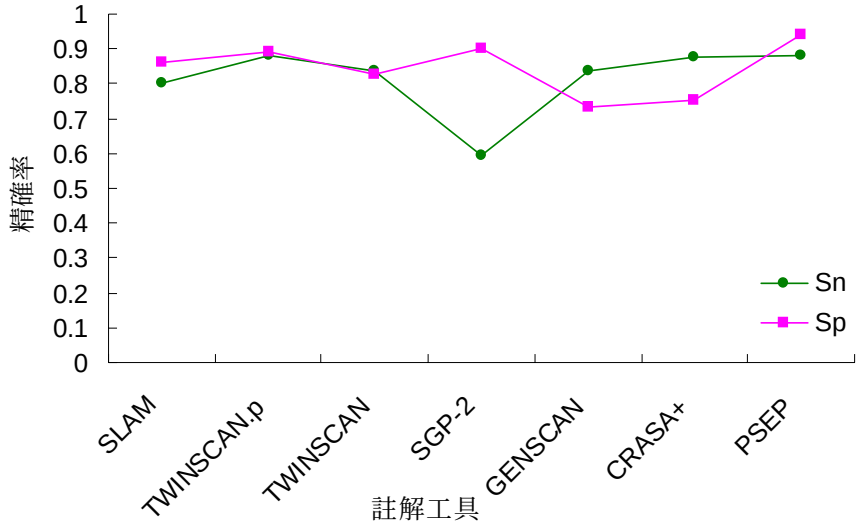
基因體註解

所謂的基因體註解 (genome annotation)，廣義地說，就是把所有在 DNA 序列中有意義的資訊全都註解出來。在這些有意義的資訊中，最重要的莫過於基因的位置，因為基因會表現（expression）而產生功能（function）。估計人類的 DNA 序列中，屬於基因所在的範圍的大小（包含 intron (介入子) 和 exon (表現子)），大概僅佔整個基因體三十億個核苷酸的不到 10％，而會編碼產生蛋白質的部分 (即 ORF)，更是只有約 2％而已。所以，狹義的基因體註解就是，找出基因在 DNA 序列（基因體）上的位置，並定義出介入子與表現子的界線。也就是說，以狹義的基因體註解而言，我們的工作像是大海撈針，在茫茫的基因體大海中，尋找不到 10％的基因的下落。由於基因體相當龐大，越是高等的生物可能越複雜，目前尚未發現一種萬無一失的通則來定義基因的位置。因此，基因註解工作挑戰性很高，許許多多的應用軟體便應運而生。通常基因體註解工具所要註解的，大都是指狹義的註解而言，也就是找出基因的位置。所以，有的應用軟體乾脆直接就被叫做基因認定（gene identification）工具。由於小鼠（mouse）、大鼠（rat）等哺乳類基因體的初稿陸續被定序完成公開，而且研究顯示，不同物種間序列保留的區域（conserved region，亦即相似度很高的區域） 很有可能是基因的位置，所以近年來利用跨物種比較的方式來尋找基因便成為基因體註解的一個主流趨勢。

基於跨物種的比對結果，我們成功研發完成一個可以預測基因以及分析多樣性切割（Alternative Splicing）的工具—漸進式訊號擷取與補綴（PSEP, Progressive signal extracting and patching）演算法。這個演算法基於表現的序列片段（expressed sequence tags；ESTs）對基因體以及基因體對基因體的比對的結果，來預測基因與多樣性切割。因此，整個系統包含兩個主要步驟：序列比對以及比對結果的後處理。後處理包含一連串的漸進式訊號擷取與補綴動作，其在基因預測方面，成功地濾掉高達 88%可能的雜訊。在整體的精確度比較上，以三個公認的標準測試資料組：ELN 基因區域、HoxA 叢集、ROSETTA 基因組來測試，我們的方法優於或者至少相當於現有知名的跨物種基因預測程式，如 ROSETA 程式、TWINSCAN、SGP-1/-2、以及 SLAM 等（見圖二）。由於 PSEP 具備跨物種的序列保留分析以及多

樣性切割分析兩個功能，PSEP 非常適合應用於多樣性切割式樣在演化上的研究以及尋找未定義的基因表現特徵。此外，我們也設計一個 Web 介面（ESTviewer），將 PSEP 在人類基因體上的註解結果視覺化呈現出來。在此介面中，也同時顯現 UCSC 及 Vega 兩個國際知名的註解工具的基因與多樣性切割註解結果，以茲比較。此外，此介面最重要的特色是，顯示出六個物種包括人、小鼠、大鼠、牛、豬、及雞的表現序列（EST）中和人的基因區段高度保留的片段。這些跨物種的保留區域具備很重要的演化及功能上的意義，因為這些保留區域顯現出在演化過程中在基因區域和轉錄區受演化壓力的程度。特別地，ESTviewer 提供很方便的方式來比較高度保留的非人的表現片段與人的多樣性切割間的差異，使用者可利用人基因序列的 ID 或是人基因體序列的絕對座標來查詢 PSEP、UCSC、及 Vega 所註解的基因架構及多樣性切割變異。此介面的網址是: http://gate.sinica.edu.tw/ ~trees/ESTviewer/ESTviewer.htm。

以上成果分別發表在生物資訊期刊：Bioinformatics, 2004, Vol. 20, No. 17, pp.3064-3079 以及 Bioinformatics, 2005, Vol. 21, No. 10, pp.2510-2513。



圖二、PSEP 和其他國際著名的註解工具在 ELN 基因區域上的註解結果的靈敏度（Sn）與精準度（Sp）比較。Sn 和 Sp 值都是越高越好越精確，值等於 1 時表示百分之百預測正確。