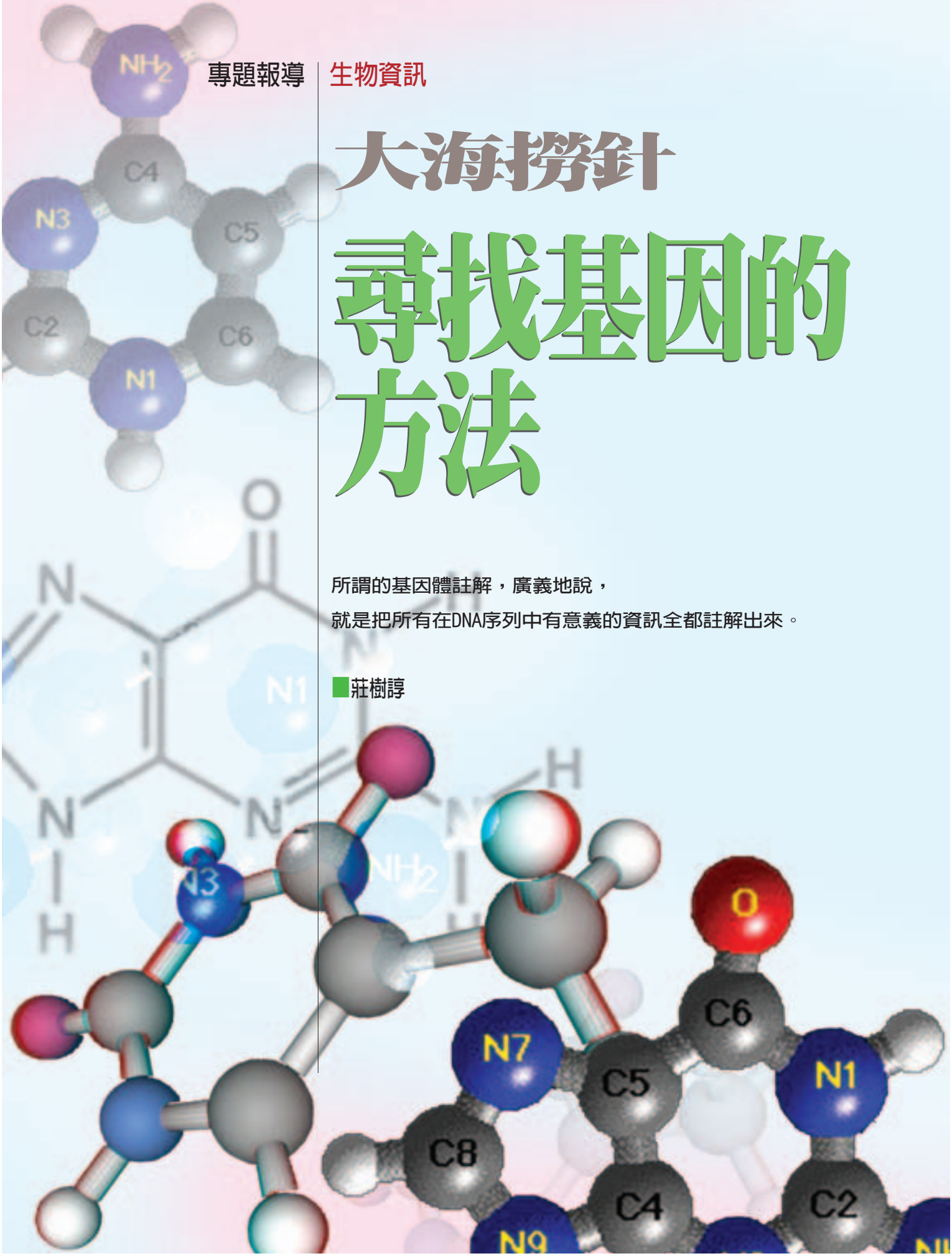
大海撈針 尋找基因的方法

所謂的基因體註解，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。

■莊樹諄



基因體註解

人類和其他生物眞的很不一樣，但爲什麼會不一樣呢？這是久遠以來人們一直都很感興趣的問題。隨著資訊科學在生物科技上的應用越來越普遍，電腦已成爲探索這些問題不可或缺的重要工具。其中一個應用相當廣泛的研究，就是基因體註解。目前，已經開發出相當多的應用軟體，在簡介這些資訊軟體之前，先來了解什麼是基因體註解。

基因體就是DNA序列，由4種字母A、C、G、T排列組合而成，分別代表4種去氧核醣核苷酸：腺嘌呤、胞嘧啶、鳥糞嘌呤、以及胸腺嘧啶。2001年喧騰一時已完成的人類基因體序列「初稿」(「初稿」表示尚未百分之百完成)，所指的就是已定出的DNA序列中的A、C、G、T排列組合。

人類的基因體序列估計約有30億個核苷酸（通常以鹼基對bp爲單位，也就是有3×10^9 bp)，目前定序的工作已差不多完成，僅剩下少數比較難定序的間隙。當基因體序列即將被完全定序完畢之際，我們非常急切知道的就是，這由4個字母編排出來的序列到底隱含了什麼樣的意義？

所謂的基因體註解，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，因爲基因會表現而產生功能。例如由DNA到蛋白質的過程，中間會經過DNA到RNA的轉錄，以及RNA到蛋白質的轉譯步驟。蛋白質是基因的產物，它會產生各種生命現象所需的功能，疾病便是可能由一個或數個基因出了差錯所造成的。

讓我們進一步解釋DNA到蛋白質的整個流程。以一條DNA序列爲例，假設編碼方向是由左向右，則左邊稱爲五端（5'end)，右邊稱爲三端（3'end)。如果白色長條代表一個基因所在，則這白色長條狀的區域會編碼產生pre-mRNA，它包括兩部分：表現子（exon）及介入子（intron)。

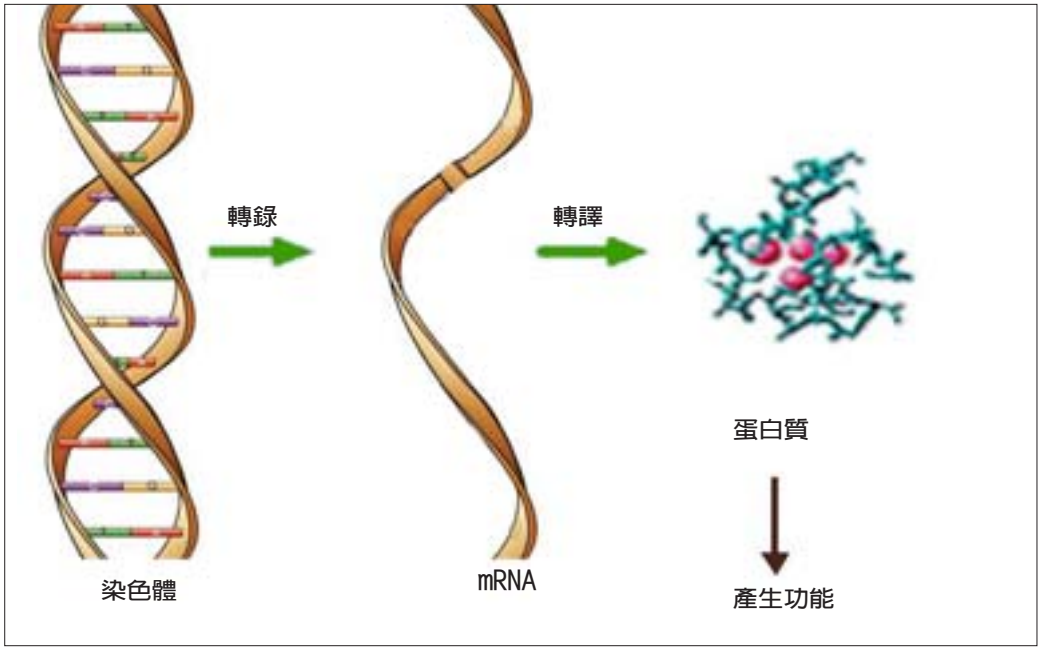
接下來在pre-mRNA上，會進一步把介入子切開去除，再把表現子連接起來。這條由表現子所連接而成的序列，就稱爲成熟的mRNA。mRNA和DNA不同的地方在於T會由U所取代，U就是尿嘧啶。爲了處理上的方便性與一致性，mRNA序列的U仍由T表示，放在資料庫裡，而這以T表示U的mRNA序列，便簡稱爲cDNA。

成熟的mRNA除了頭尾的五端與三端不轉錄區域外（不轉錄區域的

http://www.skcc.org/n_bioinformatics_core.html

人類和其他生物為什麼會不一樣呢？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。過去生物學家必須埋頭苦幹好幾年才能完成的實驗，如今靠著電腦的幫忙，可以在短短幾天之內就完成。

基因體就是DNA序列，由4種字母ACGT排列組合而成，分別代表4種去氧核醣核苷酸。2001年喧騰一時已完成的人類基因體序列初稿，所指的就是已定出的DNA序列中的ACGT排列組合。

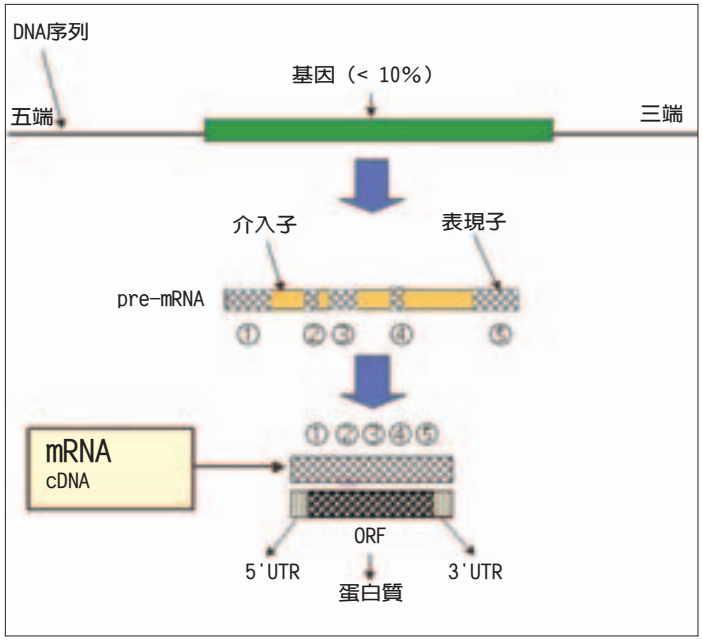


由DNA到蛋白質的大略流程　染色體DNA序列經轉錄動作產生mRNA，mRNA再經轉譯動作產生蛋白質，最後蛋白質產生生命所需的功能。

大小在各個基因上不一樣，有的甚至沒有），經轉譯的過程，會編碼產生蛋白質。而這段編碼產生蛋白質的序列，我們稱爲ORF（open reading frame）。

估計在人類的DNA序列中，屬於基因所在的範圍（包含表現子和介入子），大概僅占整個基因體30億個核苷酸的不到10%，而會編碼產生蛋白質的部分，即ORF，更是只有占整個基因體的2～3%。

因此，狹義的基因體註解就是，找出基因在DNA序列上的位置，並定義出表現子與介入子的界線。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。



基因的構造以及DNA序列到蛋白質的詳細流程　這pre-mRNA包含5個表現子（交叉線方塊部分）和4個介入子（橘色方塊部分）。

基因體註解工具

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，也就是找出基因的位置。因此，有的應用軟體乾脆直接就叫做基因認定工具。

在這些應用軟體中，大略可分成4大類。第一類是以統計預測爲基礎的演算法，它的特徵是不需要實驗上的資料作輔助，利用基因、蛋白質以及表現子與介入子結構在DNA序列上已知的一些特徵或訊號，在DNA序列上直接預測

基因的位置。

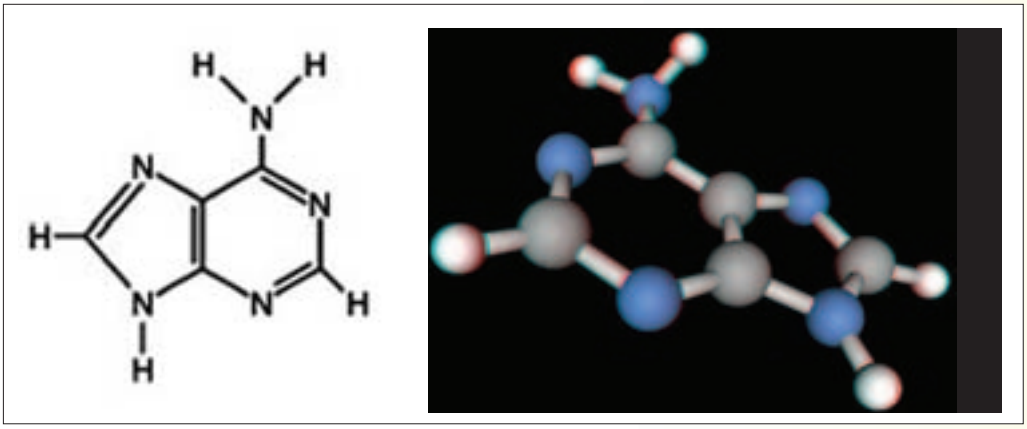
第二類是以資料比對爲基礎的演算法，它的特徵是需要實驗上的資料輔助，譬如說表現的序列片段（即mRNA的序列片段）、cDNA、蛋白質資料庫等實驗上的資料。利用這些實驗上的資料和DNA序列做比對，再篩選出可能的基因所在。

第三類是結合上列二類方式的演算法。第四類則是利用跨物種的基因體比對來尋找基因。由於老鼠、大鼠等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（也就是相似度很高的區域）很有可能是基因的位置，所以近年來這類方法的發展就變成基因體註解的一個新趨勢。

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，許多的應用軟體便應運而生。

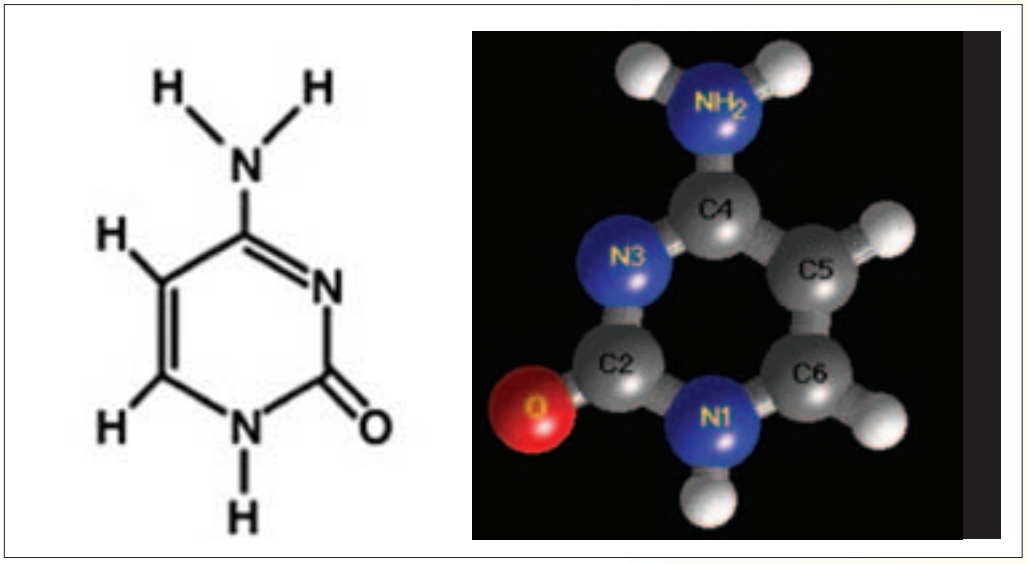
註解工具的優缺點

第一類方法的優點在於不需要實驗資料的輔助，當我們所要探討物種的實驗資料稀少時，這種方法是很好的選擇。另外，也因爲這類方法不需要大規模的資料庫比對，所以一般來說，它的註解速度比其他需要比對的方法來得快。這類方法的缺點就是容易高估基因的數目，也就是誤測值偏高。目前以這類方法預測的基因，高達百分之四、



http://coris.noaa.gov/glossary/adenine_186.jpg及
http://nautilus.fis.uc.pt/molecularium/stereo/molecules/c5h5n5.jpg

DNA序列中的腺嘌呤（adenine，A）



http://coris.noaa.gov/glossary/cytosine_186.jpg 及
http://resources.ed.gov.hk/biology/english/images/genetics/cytosine.jpg

胞嘧啶（cytosine，C）

五十都是高估的，當然這個結果會因所預測的DNA區域及物種不同而不一樣。

這類方法依其所使用的理論，又可細分成5小類：（1）以隱藏式馬可夫模式爲基礎的演算法，（2）以類神經網路爲基礎的演算法，（3）以決策樹爲基礎的演算法，（4）整合數種統計預測方法而成的演算法，（5）其他。在這些演算法中，最有名、使用最廣泛、且最有效率的，當推第1小類中的GENSCAN方法。

第二類方法是以資料比對爲基礎的演算法，它的缺點正是第一類方法的優點，因爲很多物種因其表現的序列片段（EST）資料庫缺乏，而難以使用這類方法來註解基因。所幸，以人類而言，表現的序列片段資料庫已經非常豐富，以這類方法來註解人類基因已經相當成熟。

一般來說，以表現的序列片段資料庫和DNA序列比對而篩選出的可能基因，其準確性遠比第一類方法所預測出來的高，這是因爲多了實驗的資料作輔證。不過，這類方法的困難度在於，表現的序列片段資料庫也很大，比對要花很多的計算時間及儲存空間。另外，表現序列片段資料庫的品質比DNA序列本身差，而且常常有人爲的錯誤，如實驗時的污染。因此，如何在龐大的比對結果中篩選、確認或修補可能的基因，便成爲這類方法最頭痛的問題。

這類方法依其使用的理論又可細分成兩小類：（1）區域性比對演算法，是以動態程式編製演算法爲基礎。這類方法大都是利用BLAST（basic local alignment sequence tool）或類似的演算法進行比對的工作。（2）以型樣爲基礎的比對演算法，即建立表現的序列片段資料庫或DNA序列的型樣資料庫，利用相同型樣直接做比對。在中研院計算中心的技術與生物計算平台的全力支援下，這一類方法中的CRASA法的線上服務已全面對外開放，網址是http://big.pcf.sinica.edu.tw/。

第三類是結合上述二類方式的演算法。這類方法的優點是，當用來比對的資料庫中已存有所要比對的蛋白質序列時，它的準確性是最高的。但這也是它的最大缺點，因爲用這類方法找到未知基因的可能性偏低。

另外，這類方法需要先到蛋白質資料庫中，比對找到適當的候選蛋白質序列，再使用第一類演算法預測基因結構。或者，先利用第一類演算法預測可能的基因落點，再根據這些落點到蛋白質資料庫作比對篩選。很明顯的，這類方法至少經過兩層的處理，錯誤率自然大幅降低。不過使用者在操作上，與前二類方法相比，便顯得十分不便。

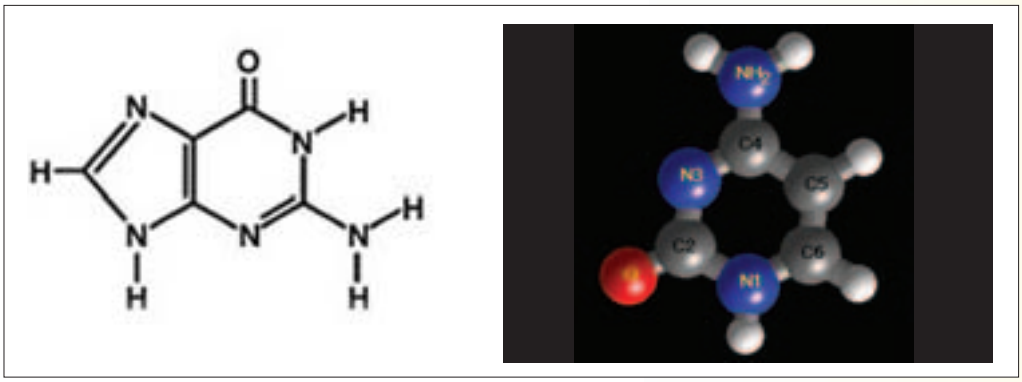
如何去選擇適當的候選蛋白質，及設定這個篩選的門檻，更是一件困難的事。不同的門檻影響精確度相當大，這是前二類方法所沒有的缺點。以目前的文獻記載來看，GenomeScan是這類方法中最精準的。不過，這類方法中最常使用以及常用來和其他方法比較的，則是GeneWise和Procrustes。

第四類是利用跨物種的基因體比對來尋找基因的方法，其中的PSEP（progressive signal extracting and patching—漸進式訊號擷取與補綴）是國人新近研發成功的基因體註解系統，在精確度的評估上都優於上列的其他方法，PSEP的查詢系統即將在中研院基因體研究中心公開。

這類方法提供了前述三類方法所缺乏的跨物

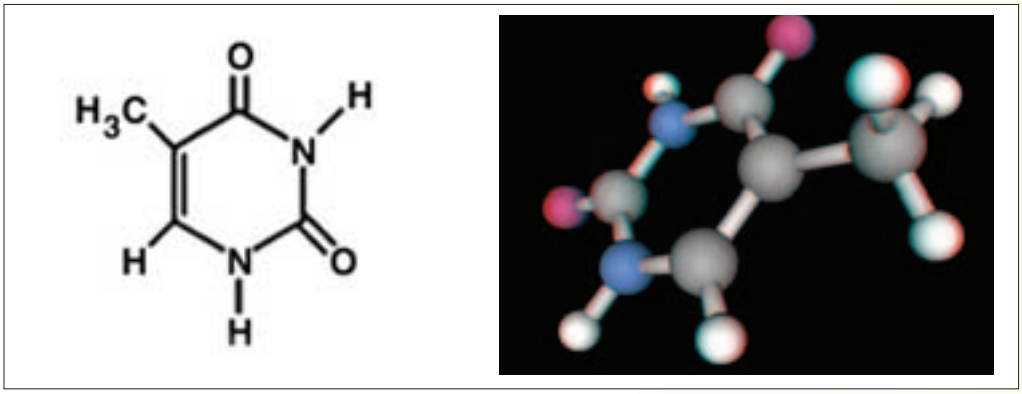
人類為何和其他生物不一樣呢？科學家正以電腦為工具，應用基因體註解方法努力地尋找答案。

圖片提供：張志玲



http://coris.noaa.gov/glossary/guanine_186.jpg
http://resources.ed.gov.hk/biology/english/images/genetics/guanine.jpg

DNA序列中的鳥糞嘌呤（guanine，G）



http://coris.noaa.gov/glossary/thymine_186.jpg 及
http://nautilus.fis.uc.pt/molecularium/stereo/moleculas/c5h6n2o2.jpg

DNA序列中的胸腺嘧啶（thymine，T）

種間保留區域的資訊，找到許多在物種演化上有意義的新線索。因此，除了尋找基因之外，這類方法還可應用到物種間演化探索的議題上。

這類方法的困難在於需要作比對，因此比對資料所遭遇到的困難，如第二與第三類方法般，一樣會在這類方法中出現。而且，這類方法需要物種間基因體對基因體的比對，可想而知，計算量與資料的儲存空間需求，是相當可觀的，電腦的工作量更甚於第二及第三類方法。再者，並非所有跨物種間高度保留區域都是屬於眞正基因的落點，因此如何作判斷篩選，值得進一步研究。

未來展望

以上所提的基因體註解方法，都是很成熟的註解方法，至少在人類的基因體註解上都有一定程度的效用。其中以第一類方法爲數最多，歷史也最悠久，那是因爲早期各種實驗資料都不完整的緣故。

不過，隨著人類或其他物種的基因體序列和表現的序列片段資料庫愈趨於完備，以資料比對爲基礎的演算法（即第二、三、四類方法）似乎愈形重要。尤其是第四類方法，利用跨物種間基因體比對的方法，更是在這兩、三年間如雨後春筍般地被研發發表出來，顯示這類方法已成爲基因體註解的主流。這也是國人積極發展 PSEP 系統的原因。

我們由衷希望，藉由這個完全由國人自行研發而成的演算法，能夠在國際基因體的研究上，也貢獻一份心力。目前國內相關學者的個人網頁上有部分比較數據，以及對人類第 20 號染色體基因註解結果，也有一些關於 PSEP 的簡介，有興趣者可上 http://www.sinica.edu.tw/~trees/PSEP/ 點閱。 □

莊樹諄
中央研究院基因體研究中心