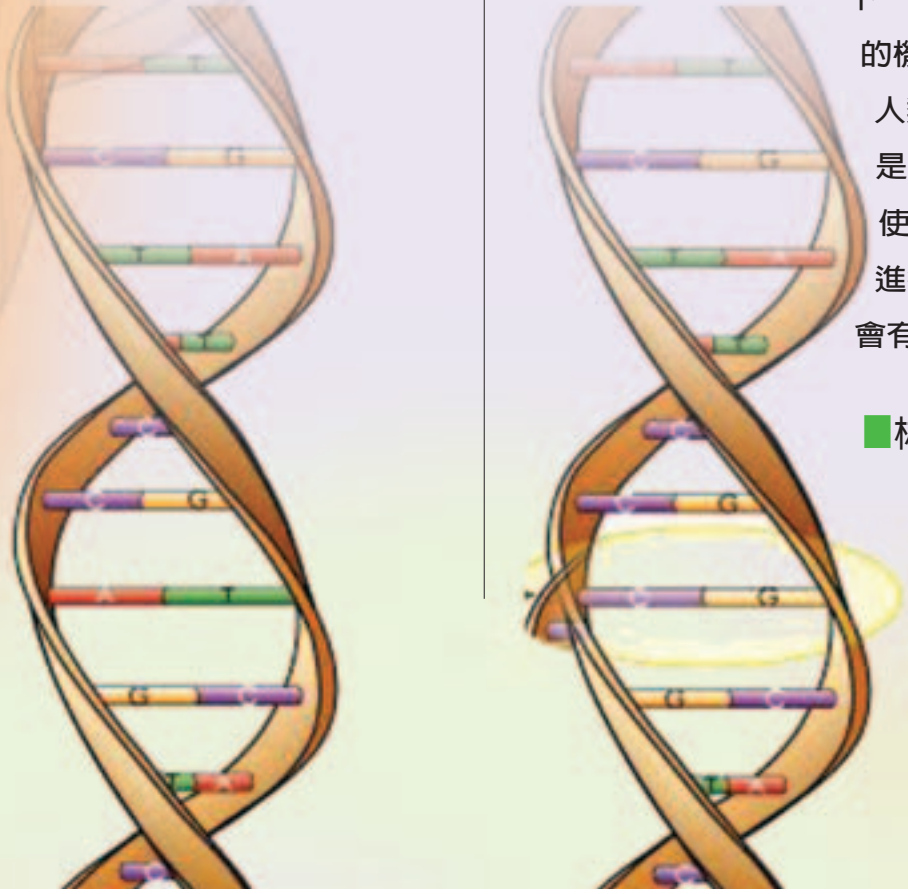
表現基因標記資料庫

生物資訊的舞台

人類基因體定序完成後，
生命之書的藍圖已出現，
下一步則是解開生命運作
的機制及協調的金鑰－基因。
人類表現基因資料庫
是隱藏基因祕密的大寶庫，
使用生物資訊技巧
進行基因資料採掘，
會有助於解開生命之書的奧祕。

林文昌



近年來由於自動定序技術的快速發展，核苷酸定序技術的成熟及成本下降，大規模的基因體計畫得以順利進展。約10年前，感冒嗜血桿菌物種首先被定序完成，接著是大腸桿菌等物種的基因體。初期的物種基因體工作，協助生物學家了解了基因體的基本資訊，如G-C的成分、重組序列、跳躍因子組成以及基因家族擴展資料等。這對於基因體上的所有基因資訊、總數目以及相互調控，有著不可或缺的重要性，可以說基因體是生命運作的重要藍圖，也是生物學家了解生命奧祕的踏腳石。

基因體是生命運作的重要藍圖，也是了解生命奧祕的踏腳石。因此，正確且迅速地辨識基因體內含的基因，就成為未來成功運用基因體資訊的重要基礎，而表現基因資料庫便是其中重要的關鍵。

在基因體研究後期，基因的功能與基因突變造成的遺傳性疾病，則成爲研究主要課題。因此，正確且迅速地辨識基因體內含的基因，就成爲未來成功運用基因體資訊的重要基礎，而表現基因資料庫便是其中重要的關鍵。

目前已有上百個物種的基因體被定序完成，而與我們息息相關的，自然是人類基因體計畫。80年代末期，以美國爲首的數十個國家，開始了人類基因體計畫的先期研究。首先是人類基因體的物理圖譜，以及遺傳基因圖譜的建立，以這爲藍圖，大規模的基因體定序工作便在全世界展開。由於計畫規模龐大，以及超高的研究經費，這項計畫也被比喻爲生物學界的登月計畫。

在2003年，也就是發現DNA雙螺旋結構的50周年，人類基因體中30億個鹼基對初步的定序宣布完成，這可說是生物學界的重大成就。但是眞正重要的功能基因體研究才正要開始！有了人類基因的完整資訊，以及生物功能全盤解析，研究人員才有可能了解細胞的運作以及病變的成因。因此發現及註解人類基因體上的所有基因，是當今最重要的課題。

爲何在完成所有人類基因體的定序後，仍然要花許多時間尋找人類基因？主要的原因是人類眞正的基因序列大約僅占基因體的百分之一，其餘百分之九十九的基因序列並不具有轉錄轉譯的功能，而且也不具備基因的基本要素。因此，基因辨識工作便成爲首要的難題。更複雜的是，人類基因並不是連續地存在於基因體上，而是在轉錄過程中由許多小片段（表現子，exon）組合而成的訊息片段。

舉例來說，假設在一個30GB的硬碟中，存有約4萬筆重要檔案，占有空間約30MB，但各個檔案分散在不同的磁區，且各個檔案約由10個磁區中的分散檔案組合而成。一般讀取檔案時，依據檔案目錄的索引，可把各個分散檔案聚合使用。但是如果硬碟檔案目錄毀損時，使用者雖仍擁有所有的資料，可是卻無法取出正確的資料，等於失去所有資料一般。解救方法是逐序掃描磁碟上面所有的磁區磁軌，再利用常見檔案特徵加以判斷組合。

http://www.oml.gov/sci/techresources/Human_Genome/graphics/slides/images/altcolhgplogo2.jpg

人類基因組解讀計畫的標章（logo），顯示這部「生命之書」除了生物學，化學、物理、工程和資訊科技之外，也涉及了倫理問題。

為何在完成所有人類基因體的定序後，仍然要花許多時間尋找人類基因？主要的原因是人類真正的基因序列大約僅占基因體的百分之一，其餘百分之九十九的基因序列並不具備基因體基本要素。

人類基因體有30億個鹼基對，分散在23對染色體上，生物資訊便是用來分析基因體資訊的工具。目前人類基因體計畫便有如掃描後的硬碟，生物學者正利用生物資訊工具，試著判斷、收取基因片段，並重新組合分析。

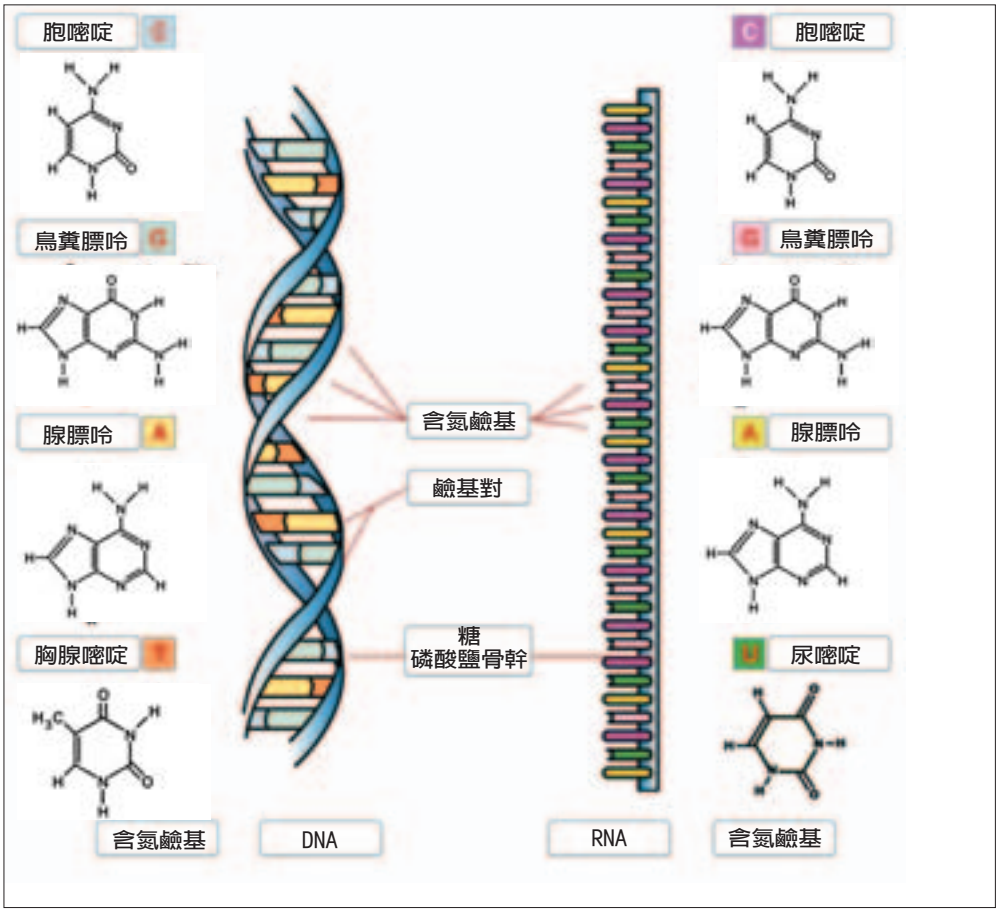
生物資訊是一門結合生命科學與資訊科學的新興學門，早期目的是爲了有效地處理基因體計畫產生的大量序列資料，但現在它的應用層面已延伸到所有生命科學領域，而生物資訊本身也已成爲另一項熱門的研究課題。

許多生物資訊工具及資料庫，因爲人類基因體計畫的推展而得到資源，使得資料庫快速擴充，如美國的GenBank資料庫。在GenBank下的子資料庫中，以表現基因標記資料庫成長最爲迅速。目前各個不同物種的表現基因標記資料庫，數目總和已超過2千萬筆，而人類表現基因標記資料庫就占有6百萬筆之多，因此善用人類表現基因標記資料庫，會有助於研究人員的人類基因解密及註解的工作。究竟表現基因標記資料庫是什麼？又是如何產生的呢？

人類基因體有30億個鹼基對，分散在23對染色體上，生物資訊便是用來分析基因體資訊的工具。目前生物學者正利用生物資訊工具，試著判斷、收取基因片段，並重新組合分析。

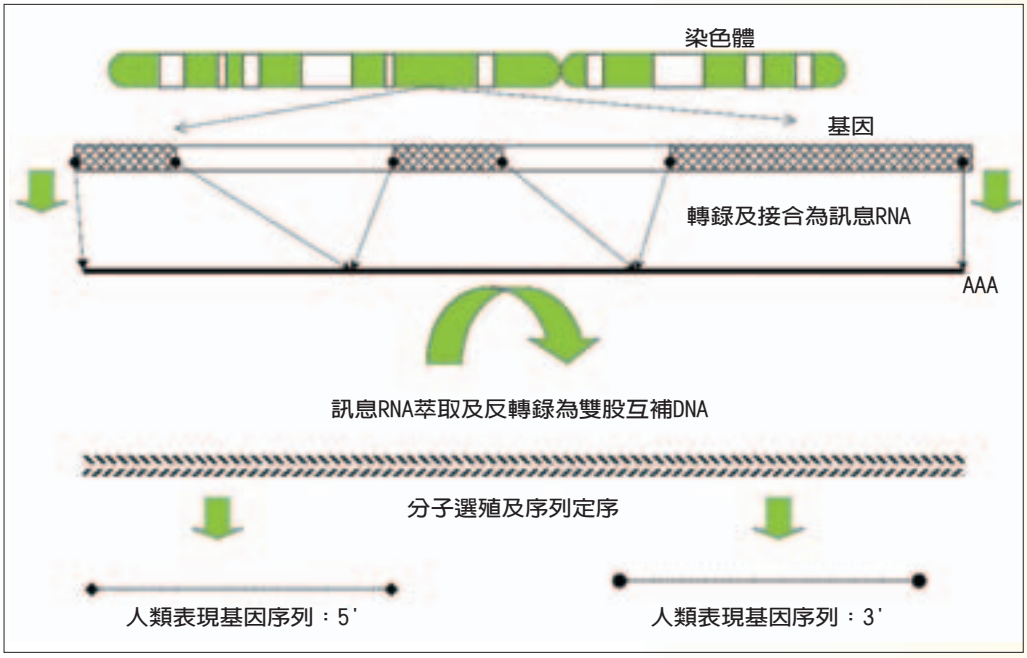
基因體中大約僅有百分之一是功能基因，而這些所謂基因的序列，便是細胞在適當的時機及地點，以它們雙股DNA的序列從事轉譯的作用，製造出單股的訊息RNA（mRNA），再經由所謂接合的動作，把許多片段的表現子，正確而完整地結合起來，以做爲合成生化蛋白質的模板。這些訊息RNA便是表現基因標記資料庫（expressed sequence tags，EST）的起源。

由於人類各部位細胞內都有著幾乎相同的基因體序列，但不同器官、組織的細胞是利用不同模組的訊息RNA從事不同的蛋白質轉譯，因此有著完全不同的生理



http://biotech.nstm.gov.tw/02/025.asp

DNA與RNA的結構差異 DNA是雙鏈結構，RNA則是單鏈結構，在五碳糖的第二個碳原子上，DNA連接的是氫原子，而RNA連接的是羥基。DNA所含的鹼基種類是ATCG，而RNA是AUCG 。



表現基因標記資料庫

功能。所以要了解各細胞在不同環境及時期的分子生物作用，唯有了解其表現的訊息RNA 組成，也就是表現基因總成，因此表現基因標記資料庫有著重要的生物應用意義。

由於基因是細胞執行功能的主要單元體，因此研究不同細胞之間基因的表現，有助於了解細胞眞正的生理生化機制。生物學家在取得某一種細胞的訊息RNA後，便利用反轉錄酶建立所謂的互補DNA（cDNA）圖庫，然後再利用自動化定序儀，大量地定序圖庫中各種基因的序列片段。但也由於使用了自動化序列定序儀，所以取得的序列長度便受到儀器的定序極限，通常是300至500 鹼基序列。因此在許多情況下，其實我們並未能取得一個基因的眞正全長，所以稱這種表現基因資料庫爲表現基因標記資料庫。

雖然我們取得的僅是數以百萬計的基因序列片段而已，而不是所有基因的完整資訊，但是這些數百萬個鹼基的序列片段，已足以告訴我們基因的部分接合資訊，更重要的是哪些基因表現在原本這個圖庫建構來源的細胞及組織中。我們也可以拼湊出細胞表現基因的總成，對於了解基因表現，表現基因標記資料庫有十分重要的貢獻。

隨著表現基因標記資料庫的推廣及成長，各種物種及不同組織的資料陸續加入，表現基因標記資料庫已成爲世界上最大的基因資料庫，有著2千萬筆以上的基因片段序列資料，有如世界基因寶庫，等待生物資訊研究人員進去尋寶。各種生物資訊工具，也針對表現基因標記資料庫，開發且建立了更有用的基因表現資訊。

基因體的座標　由於基因體的組成龐大，科學家需要利用不同的標記序列作爲基因體定序之用，而表現基因標記資料庫，提供了尋找及建立基因體標記的豐沛來源，且基因本身也是做爲基因遺傳圖譜的基礎。

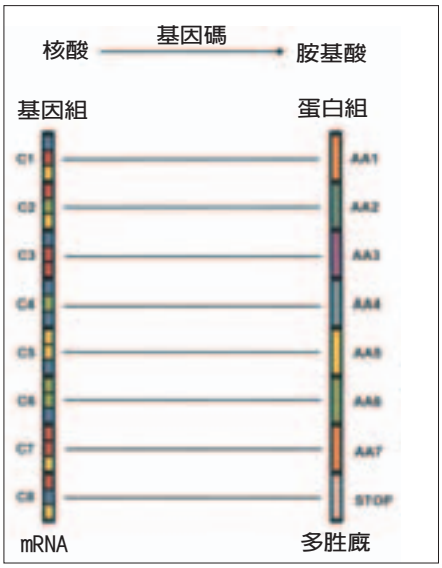
基因搜尋與辨識　表現基因標記資料庫雖只有片段基因序列資料，但是數量驚人，因此可以預期有許多序列資料重複出現在資料庫中。所以利用生物資訊比對序列工具，我們便有可能重組基因資料，進而建立起完整的基因序列資料。常見的是利用聚合方式，重組有相同片段DNA序列的標記基因。

人類基因體計畫已完成初稿定序，而後續的基因辨識工作也正積極進行中。完整且正確地辨識出所有人類基因，有助於未來功能基因體、結構基因體等的科學研究工作。

細胞或組織基因表現研究　由於各種細胞或組織表現的基因種類數量不同，因此建立數千種細胞或組織表現基因標記資料之餘，我們可以利用電腦程式及統計方法，比對不同組織基因表現標記數量的差異，而建立電子比對基因表現的生物資訊工具及資料庫。

辨識基因接合的圖譜分析　由於一半以上的人類基因有著不同的接合形式訊息，而探討表現基因標記資料庫可說是最佳的研究工作，不同表現子的接合狀態，可以藉由生物資訊工具進行表現基因標記資料庫的詳細分析。

一個胺基酸是由三個核苷酸轉譯而成的



http://codefun.com/Genetic_mapping.htm

人類單核苷酸多樣性資料庫建立的基礎　單核苷酸多樣性（single nucleotide polymorphism，SNP）是人類基因體計畫中最有醫學應用價值的資料。單核苷酸多樣性是指單一個核苷酸的自然變異，它也是人類基因體中數量最多的序列變異，預估在1千個鹼基上就有1個單核苷酸多樣性存在。因此了解每一個人體內的單核苷酸多樣性分布情形，就有可能了解個體差異現象，更能全盤解析單一個體的生化、生物反應的分子機制。

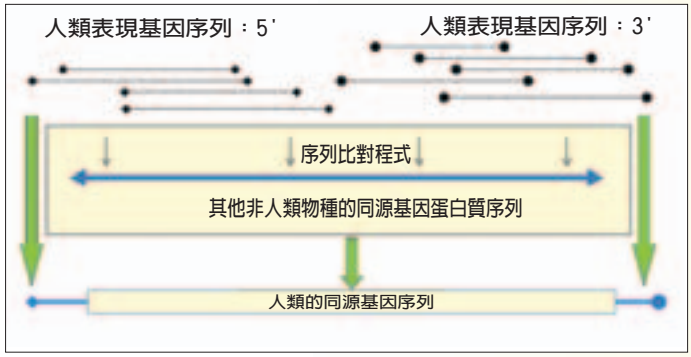
單核苷酸多樣性源於自然產生序列誤差的突變，再經由演化選擇及種族繁衍，存在於人類族群中高於百分之一的序列差異，才有資格稱爲多樣性。因爲單核苷酸多樣性的巨大數量，且高密度地存在於人類基因體上，預期未來單核苷酸多樣性在族群遺傳學、藥物開發及應用、刑事鑑定、以及人類疾病的研究及治療方面，會有重大的影響，這也是未來生物技術產業及基因型鑑定的發展基礎。

人類表現基因標記資料庫已收集由數千種組織cDNA來源的數百萬筆表現基因資料，可謂當前表現基因序列最豐富的資料庫。由於表現基因標記資料庫來自許多不同的組織，具有多樣性的特徵，可說是地球上不同人種之間最有代表性的基因序列資料庫，相當適合單核苷酸多變型的研究。

另一方面，利用基因體序列定序的方法，通常受限於樣本數目大小，僅能分析數個至數十個人的基因體，而造成由基因體序列定序方法發現的單核苷酸多變型代表性不足，無法應用在大規模分子流行病學研究。因此由表現基因標記資料庫所發現的單核苷酸多變型，相當具有臨床應用價值。

表現基因標記資料庫的另一項重要的特徵，是所有序列都是表現基因的片段，因此在資料庫裡發現的單核苷酸多樣性都是表現基因的一部分，極可能可以直接和基因變異及臨床病理特徵進行關聯性研究。由於基因分布在約百分之一的基因體範圍，一般基因體序列定序發現的單核苷酸多樣性常落於非基因區域，並不易直接應用在疾病基因分子生物機制的探討上。

人類基因體計畫已完成初稿定序，而後續的基因辨識工作也正積極進行中。完整且正確地辨識出所有人類基因，有助於未來功能基因體、結構基因體等的科學研究工作。在生物醫學研究方面，全盤了解人類基因組成及功能，是探討人類疾病起因及發展有效治療藥物不可或缺的基石。



比較性基因辨識法

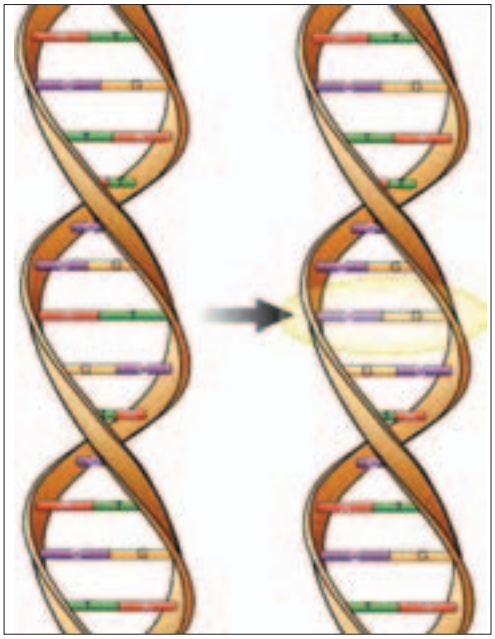
由於人類基因組成大約只占整個基因體的百分之一，判讀及正確地註解出所有人類基因，是目前生物資訊方面的重要課題。就現今的基因發現程式而言，雖在表現子的預測上有不錯的準確度，可是以1個人類基因平均10個左右表現子爲前提，要正確無誤地判讀出每個基因的所有表現子，尚有一段距離。況且過多的假性預測基因，也會造成後續分析上的負擔。因此，目前人類基因體計畫註解出的基因數目仍偏低，而部分無額外生物學證據的預測基因，也因無法得到採信而有遺珠之憾。

爲了眞正全面註解人類基因以及功能基因體研究，有必要建立一套有別於純粹DNA序列理論的預測方式。基於這樣的考量，有些研究人員開發了比較性基因辨識法的生物資訊程式。比較性基因辨識法是利用目前人類基因資訊最豐富的表現基因標記資料庫爲基礎，再加上其他已完成的定序物種蛋白體爲比對模板，用以辨識新的人類基因。

人類表現基因標記資料庫雖然是當前表現基因序列最豐富的資料庫，但由於資料庫中的序列定序錯誤及其他因素，造成在尋找及辨識基因時困難重重。

爲了更有效率地應用表現基因標記資料庫中的序列資料，研究人員便導入比較性基因辨識法，使用其他物種蛋白體胺基酸序列爲模板，以及BLAST（basic local alignment sequence tool）生物資訊序列比對程式，獲得演化中保存良好的人類直系基因資訊，並加入類神經網路資料採掘工具，協助判斷新人類基因。至今研究人員已找到150個以上人類全長完整基因，這項方法可應用在判讀及註解人類基因的重要工作上。

表現基因標記資料庫對於人類基因體計畫有顯著幫助，再加上生物資訊比較性基因辨識法，更可創造出一個新的資訊資料庫，採掘應用範例於實際基因註解及驗證。這表示利用舊資料及創新方法，可以使用在生物資訊方面，協助生物學者進行研究，並做爲未來的應用。□


http://www.bioteach.ubc.ca/BioinformaticsDeCODE/

單一核苷酸多型性—即使是DNA序列微小的改變，也可能對生命體外顯特徵或正常功能產生顯著的影響。

林文昌
中央研究院生物醫學研究所