基因工程－－廿一世紀的新科學園地

李弘謙
中央大學物理系與生命科學系

一、人類基因工程

1988 年美國的華生(Watson)認識到時機已成熟，乃組織同儕，說服了有關政府研究機構—能源部(DOE) 與國家衛生院(NIH)，由他們資助並推動一項前所未有的、龐大的、全球性的研究計畫—人類因基工程。這個計畫的目標是將長約三十億個鹼基對的人類基因組儘快的排序出來。經過了一年多的討論、組織、試車及評選之後，人類因基工程(Human Genome Project)於 1990 年正式啓動，在政府及民間企業資助下，由數個經過嚴格挑選的跨國團隊執行，分段負責人類基因組的排序。這個工程也風起雲踴的帶動了許多其它基因組工程的啓動，包括老鼠、果蠅、酵母菌、線蟲，以及很多細菌的基因組工程。

人類以及其它物種的基因組工程在成立之初就有一個很特殊但影響深遠的共識。此共識的最高原則是: 經由工程所取得的智慧財產，屬於公眾，這個原則的實現就是工程中所取得的數據以及相關資訊，都必須在最短的時間內（通常是數日）上載於各執行團隊在互聯網的網站上，無約束供任何人下載。正因如此，所有對基因及相關課題有興趣的人，只要他可以上網，都能很方便的利用基因工程的原始成果，作下一步的應用或研究。另一方面，排序之後的下游工程，無論是研究或應用發展，也需要極爲龐大的人力與智慧。在這兩個因素的相互激盪之下，許多新的學門應運而生，其中最蓬勃也最具代表性的是生物訊息(Bioinformatics)。人們對這個熱門的名詞目前仍沒有完全的共識。最簡單的說法是它涵蓋了以基因工程載於互聯網上的資訊爲起點，以非試管及臨床實驗方法所取得的一切知識和取得與應用這些知識的軟體與方法。

人類基因工程最初的目標是在 2005 前完成人類基因組的排序。工程啓動初期進度緩慢，1995 第一個長約一百萬鹼基對的細菌基因組被排序完成之後，進度的腳步快速加快。1999 年第一個動物，三千七百萬鹼基對長的線蟲基因組被排序完成。在經驗的累積，電腦硬軟體質量的飛躍進步，無限商機的誘導等多重因素影響下，相關單位在 1999 年對基因工程作了新的評估，把工程完成日期提前至 2003 年。2000 年的成果更是令人興奮，迄五月，一億八千萬鹼基對長的果蠅基因組及人類第廿二與第廿一號染色體基因組相續出籠。一般咸認人類基因組將更提前於 2001 年被排序完成。

二、四個字母的生命百科全書

早在 1988 年之前卅五年，當時仍是生物學新鮮博士的華生，在劍橋大學與才智橫溢的年青物理學家克里克(Crick)在 1953 年共同發現了基因的結

構：它是兩條平行的鹼基鏈扭成的雙螺旋。這個被稱爲廿世紀最偉大發現之一的事件，啓動了遺傳學、基因學、分子生物學的快速成長，也直接促成了卅五之後人類基因工程的誕生。

基因組(genome)是一個生物的遺傳密碼載體。在多細胞生物中每一個細胞都複製了一套基因組。基因組也是一個生物所有染色體的總稱。基因組由一種名爲去氧核糖核酸的線型大化學分子，即大家所熟知的 DNA（deoxyribonucleic acid）。DNA 本身由兩條互補的鹼基鏈扭成雙螺旋組成，而每一條鹼基鏈又是由一個一個核苷酸(nucleotide)接成的鏈型的大分子。每一個核苷酸上附著一個鹼基或碱基（大陸用詞）。鹼基有四種：兩種爲嘌呤(purine)，即腺嘌呤(adenine)及鳥嘌呤(guanine)。另兩種爲密啶(pyrimidine)，即胸腺密啶(thymine)及鳥密啶(cytosine)。這四種鹼基一般都各以四個英文字母 A, G, T, C 代稱，因此一條鹼基鏈的組成一般也用一條以這四個英文字母組成的字串序列來代表。

之前我們稱雙螺旋形的兩條鹼基鏈互補，是因爲一條鏈的鹼基序列就決定了另一條鏈的序列。互補的規則非常簡單；A-T 或 T-A 永遠形成一種互補對，而 C-G 或 G-C 則永遠形成另一種互補對。例如一條鏈上有一段序列爲 GCCTAACT，則在另一條鏈上相對應的一段必爲 CGGATTGA。在結構上每個互補的鹼基對都由氫鍵聯結。如果我們把兩條鏈想像爲一張梯子的兩條縱桿,把聯結鹼基對想像爲聯結縱桿的橫向踏板，則 DNA 就是一張極長的扭成螺旋狀的梯子。正因爲兩條鏈是互補的，當我們知道一條鏈的鹼基序列之後，我們就完全知道另一條鏈的序列。克里克與華生發現雙螺旋構造的同時，也瞭解到爲何兩條所載資訊看似完全相同的鹼基鏈應該並存的原因：如此 DNA 就能簡便且不易發生錯誤的被複製。

每一個生命體的基因組，就是那個生命體自主製造的百科全書與運行手冊。人類的基因組與大英百科全書有許多可比之處，大英百科全書有冊、條、句、字、字母，基因組有染色體(chromosomes)，基因原(operons)，基因(genes)，寡核苷酸(oligonucleotides)，鹼基(bases)。有趣的是，大英百科全書有廿三大冊，而我們的基因組也恰巧有廿三個染色體。（人類共有廿四種染色體，但包括 x 與 y 染色體。但女人的基因組分有 x 或 y 染色體，而男人的只有 x 而無 y 染色體。）百科全書共約廿萬條名詞事物，而我們的基因組估計約有十萬個基因或基因原，百科全書用廿六個英文字母寫出，而基因組只用四個字母。百科全書長全約兩億字母，而人類基因組全長約卅億鹼基對。一般硬版大英百科全書體積約 80 公升，而人類基因組則是長約一公尺，直徑約百分之一毫米，體積約 10^{-13} 公升的一條細絲。由這個比較我們約略可以體會到生物包裝資訊的驚人能力：它的包裝密度高於一般書本約 10^{15} 倍！與最先進的電腦磁碟相比，它的包裝能力也高出十億倍。我們如果要真正體會一條人類基因組的大小，我們可以向全世界六十多億人每人要一條基因組，然後將這六十多億條基因組揉成一束繩子。這條繩子恰好與我們的一根頭髮一般粗細！

這條卅億字長四個字母編成的生命百科全書，還有一點與大英百科全書大大的不同。大英百科全書前有目錄，後有索引，字分大小，文有章、節、段、句，外加空白、標點符號的靈活運用，在在都是爲著要讓讀者能輕易的不費力的清楚瞭解它每一字每一句的涵意。

我們的生命百科全書則相反。這些出版商的玩意兒它全不用，它只用字形永遠相同的四個字母(A, G, T, C)，沒有空白沒有標點的寫出這卅億字長的天書。基因工程的成果基本上是把這本天書先印出來。至於能幫助我們閱讀這本天書的字典，專家們則認爲人類至少仍需努力一百年才能編出來。雖說如此，現階段我們對基因組仍已有不少瞭解。

三、基因組是建築籃圖

我們體內億萬個細胞裡每個都藏有一條完整的基因組，也有知道如何解讀基因組，從而製造蛋白質的小機器（這些小機器本身也是蛋白質或它的演化前身—核糖核酸（RNA））。基因組裡用四個字母寫的文章有兩大類，一種是編碼段，或基因段，一種是基際段(intergenic regions)。顧名思意，編碼段裡藏著基因的訊息，基際段則否。人數的基因組只有大約百分之三爲編碼段，其餘爲我們對它幾乎毫無所知的基際段。每一段編碼段裡編著一個基因，也就是製造一個蛋白質的密碼（被稱爲基因密碼，已被解）。蛋白質有兩大類，一類是建造我們身體各部的材料，另一類是讓我們身體運作的機器。基因組中的編碼段，絕大部份在每一個人身上都是極相似的，因而我們體內製造出來的蛋白質也極相似。如此就保證了我們每個人的體型、構造、生理機制等也基本相同。每個人編碼段中極微小的差異，就導致人與人間體型及體質的差異。編碼段的異型也是許多遺傳性疾病的根源。

四、基因組是使用手冊

雖然我們體內每一個細胞都包著一付完整的基因組，上面同樣寫著我們身上每一個蛋白質的編碼，然而每一個細胞卻依著它所在的時、地不同而作極不相同的事。也就是說細胞會因時、地不同而製造不同的蛋白質，並適時的放出蛋白質出來適應後者執行它特定的任務。正因如此，感光的神經細胞只長在網膜上，而不長在我們的指尖上。我們仍不知道基因組在何處或如何具有這種識時務的能力。也許將來我們有一天瞭解這個運作原理之後，可以改造我們的基因組，使我們能用指尖看電影，用耳朵聞香，用眼睛聽音樂。

五、基因組是演化歷史

物種的已知親緣關係與這些物種基因組之間的相似性有足夠的相干性使生物學者相信所有的物種有一個共同的祖先。根據這個（可信度極高的）假設，每一個現存的或曾經存在的基因組都是共同祖先基因組經過極漫長歲用的演化的結果。演化起沿於隨機性的突變或複製錯誤，再經過生存競爭的篩選而成。每一個基因組都記載著它演化的歷史。基因組很清楚的告訴我們人類與老鼠的親緣關係近於人類與果蠅的關係，而後者又近於人類與大腸桿菌的關係。我們也瞭解爲何人類、老鼠，雞和許多其他動物的早期胚胎爲何那麼相似。

經由統計理論，突變模型，我們可以推算真核（包括一切動、植物）與古菌（包括許多嗜熱菌）的祖先在大約二十三億年前分道演化，而它們的共同祖先則與真菌（大腸桿菌，根瘤菌）的祖先在大約三十五億年前分道。

六、基因組是終極奇妙的全自動機

基因組是一部天書，一部秘箋，是生命的籃圖，是運行的手冊。然而很奇妙的，它是一個自編、

自導、自演的全自動複製機。它在不運作時自動捲成一個構造複雜但有規則，半徑不及全長七萬分之一的小球，要製造蛋白質時會自動將相關的片段伸展開，要複製時更會自動將自己全部打開。最奇妙的，是基因組會自己幫著製造所有協助它執行這些任務的機器。基因組的終極目標，似是保證自己的永續生存。著名的古生物學家與作家多肯斯(Dawkins)就說：人的身體只不過是永續傳承人類基因組的暫時性的軀殼罷了。賦予基因組這些時空功能的機制，絕大部仍等著被人們瞭解。

七、互聯網是大眾的研究平台

互聯網上有許多資料庫，其中最主要的是美國的 GenBank，歐洲的 EMBL 及日本的 DDBJ。互聯網使這三個資料庫互通[1]。EMBL 及 GenBank 分別於 1982 年六月及十二月建立，DDBJ 建立於 1987 年七月。三個資料庫在 1992 初互聯。庫中所存鹼基對數，1983 年四月初次超過一百萬對，1992 年超過一億對，1997 年八月超過 10 億對。2002 年二月爲 57 億對，五月底已達 81 億對，目前以每六個月番一番的速率增加[2]。除了許多基因組片段之外，它存有近三十個細菌、酵母菌、線蟲、果蠅、人類等 22 及 21 個染色體的全基因序列。隨伴著這些基因序列，更有巨量的文字說明、圖片、軟體也一併存在資料庫中。而新的資料更以排山倒海之勢每日湧入。可喜的是，任何一個對這門新興知識感興趣而同時具有上網資源的人，都可以很容易的經由互聯網與這些知識接觸，或進而作不同層次的學習與研究（多種其他的學門在互聯網上也設了許多網站，提供讓人自由索取學習與研究的資料。這些學門網站的互聯性、涵蓋性、整合性與更新頻率遠遠不如生物訊息網站）。只要有一些基本的科學訓練，多一點耐心與堅持，把互聯網當作自己的研究平台不是一件難事。

八、生物訊息是整合數理工計學門的催生劑

人類基因工程經過互聯網呈現在大眾面前極巨量的數據與訊息，也同時提供了無數的問題。人們相信傳統的生物學者，即使有電腦專家的協助，也只有能力解決這些問題的極小部份。大部分的問題的解決需要具有物理、化學、數學、統計、工程等專業的研究者的參與。基因與蛋白質都是大分子，瞭解它們的性質須要化學專業知識，自不待言。牽涉到數學的重要課題包括基因與蛋白質序列的解碼，基因編碼段的辨認，物種分類，演化等。牽涉到物理的問題有基因組的彈性與折疊性質，蛋白質的結構與折疊，基因序列的演化機制等。牽涉到工程的一個明顯的例子是基因晶片的製造，而晶片上的探索碼的設計及晶片圖案的分析則是一個複雜的數理問題。

事實上，我們對一個生物體的真正認識，是始於該生物的基因組排序完成之後。初期最需要作的工作，也是對瞭解在分子層次的遺傳與醫藥學極爲重要的工作，是對所有基因編碼段的確認，對相對應於這些編碼段的蛋白質的結構與功能的瞭解，以及對各個基因相互配合呼應的形態與機制的瞭解。這個龐大的工作，只有經由跨領域的整合研究才能做好。因此，說生物訊息是整合數、理、化、生、計、工各學門的催生劑是絕不爲過的。讓我們共同面對，歡迎這廿一世紀的大挑戰吧！

———————

後記：本文校對期間，國科會公告了由生物處、自

然處、工程處共同推動的發展研究生物訊息學辦法，並開始徵求生、理、工整合性研究計畫[3]。

參考文獻：

1. GenBank, EMBL 與 DDBJ 之網址分爲：
www.ncbi.nlm.nib.gov,
www.ebi.ac.uk, www.ddbj.nig.ac.jp.
2. 參閱: www3.ebi.ac.uk/Services/DBStats/.
3. 參閱: www.nsc.gov.tw/bio/wwwbio35.html.