第一章 什麼是生物資訊學

1. 生物資訊學(或稱為生物信息學)是英文 *Bioinformatics* 的直接翻譯。

2. 目前，各種生物數據庫急速的發展，如何組織數據，並提取生物學新知識，需要以電腦(或計算機，*computer*)和網際網路(*Internet*)為重要憑藉，來迅速達成目標。

3. 生物資訊學突飛猛進的發展，正在引發生物學研究的一場革命。這些發展必將影響到廿一世紀的農業、林業、醫藥和許多人類的活動。

4.

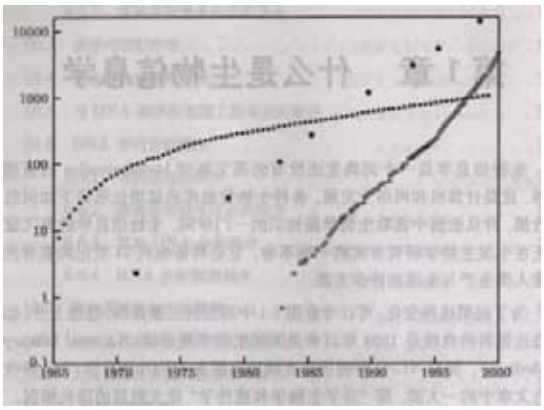

Figure 1

為了說明這種變化，可以看 Figure 1 中的三條曲線。

a. 第一條線以小圓黑點為記號，緩慢上升，似乎趨近飽和的曲線(實心圓點，高度需再乘以 10^3)是 1966 年以來美國國家醫學圖書館(*National Library of Medicine*，簡稱 NLM)所提供的線上檢索服務 MEDLINE [R1-1]所收錄之文章的一大類，即“分子生物學和遺傳學”論文數目的增長情況。MEDLINE 的選用範圍超出醫學，而幾乎囊括全部重要的生物學期刊，這條線大致反映了人類消化理解實驗事實和數據，使之提升為科學知識的過程。

b. 第二條線以空心圓點為記號，是一條迅速增長的曲線(空心圓點，高度需再乘以 10^6)。表達 20 世紀 80 年代初，美國核酸序列數據庫 GenBank[R1-2]中核酸序列數目的增長情況。這一條曲線清楚地說明，當數據增長越來越快時，傳統的研究方式已經來不及迅速消化新數據。

c. 第三條線(大實心圓點，高度需再乘以 10^3)由八個數據點構成，它反映出

大規模集體電路單個 CPU 晶片上三極管數目的增長速率。這一個進步的技術提供了解決問題的關鍵方法。當前一個典型的基因測序中心，每年可以產生 10^{14} 字節即 100 000GB 的原始數據[R1-3]。數據的產生、搜集和分析，都必須依靠電腦和網路，也必須發展數據庫、演算法和程式。這是生物資訊學的使命。

5. 廿世紀後半葉，因著分子生物學的長足進展，將生命活動的物質基礎追溯到核酸和蛋白質兩大類生物大分子的序列，它們構成了生物數據的主要部分。關於這些生物大分子的結構、相互作用和生物功能的研究，也產生著大量數據。直到不久之前，人類科學研究產生數據量最大的領域，還是高能物理實驗和腦神經活動成像，兩者都達到每年 10^{15} 個字節。目前生物數據的產生率已經達到同樣的水平，而且很快地就要超出前兩者。

6. 生物資訊學與計算生物學或生物計算有密切關係，但又不儘相同。目前歸入生物資訊學領域的大致有以下幾個方面：

   a. 各種生物數據庫的建立和管理。

      這是一切生物資訊學工作的基礎，通常需要有電腦科學背景的專業人員與生物學者密切合作。

   b. 數據庫接口和檢索工具的研製。

      數據庫的內容來自許多生物學者日積月累的研究成果，但一般生物學研究者並不具備專業的電腦和網路訓練，因此必須發展查詢數據庫和由數據庫提供數據的方便接口，以便利各種研究的進行。這些工作必須由具備專業電腦和網路知識與經驗的人員來完成。

   c. 人類基因組計畫的實行，配合大規模的 DNA 自動測序，對資訊的採集處理提出了空前的要求。

      從各種圖譜的分析，大量序列片段的配聯、電腦克隆和預測結構與功能，到數據和研究成果的視覺化等，無不需要高效率的演算和程式。研究新演算法、發展方便適用的程式，是生物資訊學的重要任務。就基因組分析這一角度來看，生物資訊學主要是指核酸和蛋白質序列數據的電腦處理和分析。

   d. 生物資訊學最重要的任務，是從極大量數據中擷取新知識。

      這首先要從 DNA 序列中識別編碼蛋白質的基因，以及調控基因表達的各種訊號。其次，從基因組編碼序列翻譯出來的蛋白質序列數目急劇增加，根本不可能用實驗方法一一確定它們的結構和功能。我們只能從已經累積的數據和知識出發，來預測蛋白質的三維結構和功能。這個已經成為常規的研究課題。

   e. DNA 晶片(DNA *microchip*)和微陣列(*micro array*)的發展，將一定組織或生物體內萬千基因時空表達的研究提上日程。

研究基因表達過程中的群聚關係，從其中擷取調控網路和代謝途徑的知識，進而從整體上掌握細胞內全部互相耦合的生化反應，這一切都需要新的演算法和程式。

7. 生物資訊學與生物實驗的關係。

生物資訊學的發展，將造就一批不直接做實驗而每天坐再電腦終端機前的生命科學研究人員。“生物學是一門實驗科學”這一種的敘述，在目前數據大量產生的現實環境中已經不再完全符合當前的情況。我們在正式進入這個領域之前，必須正確地領會這其中的關係與內涵。

a. 作為生物資訊學基礎和出發點的核酸與蛋白質序列都是來自貨真價實的實驗數據。即使是高產量的自動測序機，也都是基於以往的實驗成就。這種情況的出現最主要是因為測量實驗技術已經發展成為現代化的生產線。就好像商業產品源源不絕地從工廠裡頭的生產線上組裝而出。
b. 現在我們所面對的情況是要在全球每天所產生的數以千萬計之鹼基對核酸序列中，能夠翻譯出成百之可能的蛋白質序列。使用傳統的實驗方法，逐一地確定它們的結構和功能，是不可能做到的事。我們只有根據以往累積的數據與經驗，對大量的新序列進行分析篩選，才能有效率地挑選出必須由實驗去決斷的問題。這種決策過程需要借助電腦來完成。
c. 越來越多物種的基因組將被完全地測定。那種傾畢生精力研究單個基因、單條代謝途徑、單種生理週期的時代已經一去不返。目前研究人員正試著闡明細胞內全部互相耦合的調控網路和代謝網路，細胞間全部訊號的傳導過程，以及從授精卵發展到成體的全部生理和病理之基因表達的變化，等等。這一切都超出了手工分析的可能性。
d. 回顧物理學的發展，19 世紀曾是實驗科學，20 世紀上半葉發展成為理論和實驗密切配合的科學，而 20 世紀下半葉則發展成為以實驗、理論和電腦三種方法密切配合的成熟科學。生物學發展的過程應該也是如此。

8. 蛋白質序列測定(*Protein Sequencing*)

序列測定(*sequencing*)已有 60 多年的歷史，但開始時進展十分緩慢。最初，人們致力於建立蛋白質(*protein*)和多肽的分離技術，並確定其中胺基酸(*amino acid*)的種類及含量。1945 年以前，沒有任何蛋白質序列定量測定的方法。以後十年中，由於色譜技術和標記方法的快速進展，第一個完整的多肽激素(胰島素)的全序列測定於 1955 年完成(Ryle 等，)。1960 年第一個酵素(或稱為酶，*enzyme*；核糖核酸酶，*ribonuclease*)序列測定完成[R1-4]。1965 年，約有 20 個長度為 100 多個殘基的蛋白質序列被測定。截至 1999 年，這個數字已達 30 萬個，這在 60 年前是難以想像的。

9. 核酸列測定(*Nucleic Acid Sequencing*)

20 世紀 60~70 年代，科學家們只能測定轉移核糖核酸(*transfer ribonucleic acid*, 即 tRNA)，這些分子長度很短，通常只有 70~90 個核苷酸(*nucleotide*)，而且也比較容易將純化的單個分子分離出來。要測定脫氧核糖核酸(*deoxyribonucleic acid*, 即 DNA)則情況複雜許多。就人類的染色體而言，每個染色體大約含有 0.55~2.5 億個鹼基對(*basepair*，bp)，其長度遠遠地大於 RNA 分子。測定一個染色體 DNA 分子的核苷酸序列，是一向艱鉅的任務。通常一次實驗中可以測定的 DNA 片段約為 500 bp。基因克隆(*gene cloning*)和多聚酶鏈式反應(*polymerase chain reaction*，簡稱 PCR)，提供從染色體中分離特定 DNA 片段的快速方法。另外，1977 年基於鏈終止(*chain termination*)和化學降解(*chemical degradation*)測序方法研究成功，成為 80 年代和 90 年代序列測定革命的基礎，生物資訊學也應運而生。

10. 序列與結構(*Sequence and Structure*)

序列和結構這兩大類不同性質的生物資訊數據在數量上有著天壤之別。根據 1998 年的資料，國際上公開且非重複的資料庫中，儲存了超過 30 萬個蛋白質序列。而已公佈的序列片段和表現序列標籤(*expressed sequence tag*，EST)數據庫的數目也已達百萬個，但是蛋白質三維結構數據庫中獨立的原子座標依然不足 1500 套。這是因為結構數據的測定、儲存與處理，遠比序列數據複雜。

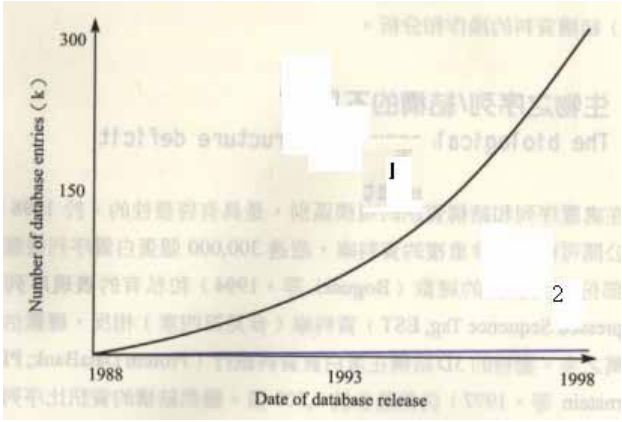

Figure 2

這兩種既不相同卻又密切相關之數據間信息量的巨大差異可以見 Figure 2。其中曲線 1 為序列數目的成長，而曲線 2 為結構數據的成長。

11.

12.

13.

14.

參考資料

R1-1 EDLINE 是美國國家醫學圖書館的文獻摘要庫，反應美國及其他國家 4500 多種醫學和生物期刊的論文摘要和引用情形。

R1-2 GenBank 是 NCBI (*National Center for Biotechnology Information*)所維護的 DNA 序列總數據庫。
其網址為 *http*://www.ncbi.nlm.nih.gov/Web/GenBank/。

R1-3 *Science* **284** (1999) 1742。

R1-4 Hirs, C.H.W.,Moore, S. and Stein, W.H. (1960) *Journal of Biological Chemistry,* **235**, 633-647。