第一部分

簡介

第一章　電腦時代的生物學

第二章　應用電腦解決生物學的問題

1

資訊時代的生物學

生物學是一門研究生物的學問，範圍從物種與族群的互動，到單一個體中組織與細胞的功能。在這門學問中，生物學家收集並詮釋資料。在二十一世紀開始的現在，我們擁有精密的高科技實驗室，讓我們能夠更快速地收集與詮釋資料。我們的電腦中儲存了大量的 DNA 序列資料，但如何才知道 DNA 的哪個部分，控制了不同生命的化學反應呢？我們已經了解了某些蛋白質的結構與功能，但要如何辨別新的蛋白質呢？還有，要如何依據序列來預測蛋白質的外觀？我們了解將 DNA 轉譯成蛋白質的簡單編碼，但該如何在密碼中找到有意義的新「字」，然後把它們加到 DNA-蛋白質的「字典」中呢？

生物資訊（bioinformatics）就是以資訊來了解生物學的科學，它是能夠解答上述問題與類似問題的工具。然而，由於描繪基因圖譜被過度渲染，生物資訊成了流行的行話，「生物資訊」這個詞擁有不同的定義，端看是誰使用這個詞。嚴格說來，生物資訊是更大範圍的計算生物學（computational biology）的子集合，是模擬生物系統時，量化分析的應用技術。在本書中，我們將從生物資訊談到計算生物學，然後再回到生物資訊。兩者之間的差異並不是本書的重點，我們所著重的，是要了解這些對分子生物學家相當重要的工具與技術，以及如何應用現有的基本電腦工具。

生物資訊這個領域極度依賴統計學的方法，以及樣式辨識（pattern recognition）。生物資訊研究者來自許多不同領域，包括了數學、資訊科學、以及語言學。不幸的是，生物資訊不僅專精，也相當廣泛。生物資訊對在沒有充分了解生物資料從何而來，以及其意義的情況下，想要進行樣式辨識或預測的人來說，充滿重重的困難。生物學提供了演算法、資料庫、使用者介面以及統計工具等，讓研究者可以作一些令人振奮的事情，例如排比 DNA 序列，或演算出可能相當重要的結果。「可能相當重要」大概是最重要的詞。這些新工具同時也讓你有詮釋以往從未被發現的資料的機會。我們不能誇大了解這些工具的重要性，但只要你一開始接觸生物資訊，並成為生物資訊研究方法的知識消費者，你的研究進度速度將會相當驚人。

電腦如何改變生物學？

生物個體的遺傳與功能資訊儲存在 DNA、RNA 與蛋白質中，它們全部都是以更小的分子組成的直鏈。這些大分子是由已經充分了解的化學物質所組成，它們由特定的字母所代表：DNA 是由四種去氧核糖核酸組成（腺嘌呤 adenine、胸腺嘧啶 thymine、胞嘧啶 cytosine、與鳥糞嘌呤 guanine），RNA 是由四種核糖核酸組成（腺嘌呤 adenine、尿嘧啶 uracil、胞嘧啶 cytosine、鳥糞嘌呤 guanine），蛋白質是由二十種胺基酸組成。因為這些大分子是定義元件的直鏈，它們能以序列符號來表示。在這些序列可以透過比較，找出功能或形態類似的分子之間的關聯性。

序列排比可能是分子生物學家最有用的電腦工具。WWW 讓世界各地的使用者能夠透過一致的介面，使用公用基因序列資料庫。透過常用的程式 fsBLAST，分子生物學家能夠將一個未知的 DNA 序列，與公用 DNA 序列資料庫進行排比。下一節中，將舉例說明如何利用 BLAST 進行序列的排比，協助你更深入瞭解疾病的真相。

果蠅的眼睛

果蠅（Drosophila melanogaster）是常拿來進行生物發育研究的生物體。果蠅有一種「無眼」（eyeless）基因，如果少了這種基因（用分子生物學的方法將它移除），果蠅就會沒有眼睛；「無眼」基因對眼睛發展的影響是非常明顯的。

研究者發現有一種人類基因，與虹膜缺損（aniridia）有關。如果人類缺乏這種基因（或是基因發生突變，讓蛋白質無法正常運作），眼睛就不會發育出虹膜。

如果將虹膜缺損的基因，以實驗手法插入被剔除無眼基因的果蠅基因體中，便會生出眼睛正常的果蠅。這是非常有趣的巧合，即使果蠅和人類是截然不同的生物，但無眼基因和虹膜缺損基因功能之間是否有相似處呢？也許有吧。我們可以透過比對其基因序列來了解，無眼和虹膜缺損這兩個基因的作用機制。然而重要的是，基因會互相影響，要獲得確切的答案必須經過非常謹慎的實驗。

大約十五年前，要比對無眼與虹膜缺損基因兩者之間的 DNA 序列，其難度有如大海撈針。大部分的科學家運用文字處理器手工排比不同基因序列，試圖一個字母一個字母找出配對。這實在是太浪費時間了，更甭提眼睛會受不了。

到了 80 年代，能夠快速比對序列的電腦程式將分子生物學帶入全新的階段。生物序列的逐對（pairwise）排比是所有生物資訊技術中的基礎，從多重排比（multiple alignment）、親緣樹分析（phylogenetic analysis）、motif 的辨識（motif identification）、同源模擬軟體（homology-modeling software），到網路資料庫搜尋服務等許多技術，都將逐對排比的演算法視為其功能核心。

現在，生物學家能用序列排比程式（如 BLAST 或 FASTA）在幾秒內找到數十筆符合條件的序列。這些程式如此常用，所以運用這些生物資訊工具或生物資料庫時，最先遇到的可能就是國家生物科技資訊中心（NCBI）的 BLAST Web 介面。圖 1-1 顯示提交資料到 NCBI 作 BLAST 搜尋比對的標準表單。

基因序列的標記

在你急著去用 BLAST 比對無眼與虹膜缺損的基因序列之前，先讓我們來談一下序列排比如何運作。

首先，我們要知道生物的序列（DNA 或蛋白質）雖然具有化學功能，但如果將其化約成單一字母的編碼，它也會以如同條碼般的獨特標記來呈現、發揮作用。從資訊科技的觀點來看，序列資訊相當重要：序列標記可以應用在基因、其產物與功能，以及在細胞代謝中的角色等方面。想要找出特定基因相關資訊的使用者，可以用逐對序列排比的方式，來找到與序列標記有關的任何資訊。

這些序列標記最重要的地方，在於它們不僅標示出一個特定基因，同時也包含了具有生物學意義的樣式（pattern），能讓使用者比對不同標記、從而連接資訊，並作出推斷。所以標記不只能夠把一個基因的所有資訊連結起來，還能幫助使用者將細微差異，或序列上完全不相似的基因之間的資訊建立關聯。

圖 1-1：在 NCBI 網站中透過表單於核酸資料庫進行 BLAST 搜尋比對

如果簡單的標記就可以代表複雜的生物意義，那麼只要把一個代碼（例如 GenBANK 的 ID）插入任何 DNA 序列就夠了。但生物序列與演化息息相關，所以在兩個序列標記中的局部樣式比對（partial pattern match）就變得非常重要。BLAST 與簡單的關鍵字搜尋不同，它能夠在整個蛋白質序列中偵測到部分符合的資訊。

利用 BLAST 比較無眼基因與虹膜缺損基因

當你用 BLAST 來比對兩個序列，會發現無眼基因與虹膜缺損基因有部分符合。下面的文字是 BLAST 搜尋所產生的原始資料：

```
pir||A41644 homeotic protein aniridia - human
       Length = 447

 Score = 256 bits (647), Expect = 5e-67
 Identities = 128/146 (87%), Positives = 134/146 (91%), Gaps = 1/146 (0%)

Query: 24 IERLPSLEDMAHKGHSGVNQLGGVFVGGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 83
          I R P+  M + HSGVNQLGGVFV GRPLPDSTRQKIVELAHSGARPCDISRILQVSN
Sbjct: 17 IPRPPARASMQNS-HSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCDISRILQVSN 75

Query: 84 GCVSKILGRYYETGSIRPRAIGGSKPRVATAEVVSKISQYKRECPSIFAWEIRDRLLQEN 143
          GCVSKILGRYYETGSIRPRAIGGSKPRVAT EVVSKI+QYKRECPSIFAWEIRDRLL E
Sbjct: 76 GCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYKRECPSIFAWEIRDRLLSEG 135

Query: 144 VCTNDNIPSVSSINRVLRNLAAQKEQ 169
           VCTNDNIPSVSSINRVLRNLA++K+Q
Sbjct: 136 VCTNDNIPSVSSINRVLRNLASEKQQ 161


 Score = 142 bits (354), Expect = 1e-32
 Identities = 68/80 (85%), Positives = 74/80 (92%)

Query: 398 TEDDQARLILKRKLQRNRTSFTNDQIDSLEKEFERTHYPDVFARERLAGKIGLPEARIQV 457
           +++ Q RL LKRKLQRNRTSFT +QI++LEKEFERTHYPDVFARERLA KI LPEARIQV
Sbjct: 222 SDEAQMRLQLKRKLQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQV 281

Query: 458 WFSNRRAKWRREEKLRNQRR 477
           WFSNRRAKWRREEKLRNQRR
Sbjct: 282 WFSNRRAKWRREEKLRNQRR 301
```

這些資料顯示無眼與虹膜缺損基因中，蛋白質序列高度相似的局部排比（local alignment）結果。每三行中，查詢序列位在第一行（輸入 BLAST 的無眼序列），虹膜缺損序列則在最後一行。中間那一行是兩者相同的部分。如果中間一行有一個字母，代表序列完全相同。如果中間一行有加號，表示兩個序列於該位置上的胺基酸不同，但仍有一些化學特性相似，例如 D（天門冬胺酸；aspartic aicd）和 E（麩胺酸；glutamic acid）。如果中間那一行空白，表示在這個位置上，兩個序列完全沒有共通性。

這個例子中可以看到，如果以整條的無眼基因序列尋找完全相同的配對（如一般的關鍵字查詢），你沒辦法找到任何東西。局部序列範圍只能找到完整蛋白質其中的一部分：無眼基因第 24~169 個胺基酸，與人類的虹膜缺損基因第 17~161 個胺基酸相

吻合，而無眼基因第 398~477 個胺基酸，與虹膜缺損基因的第 222~301 個胺基酸相吻合。但其他的序列卻完全不能配對！即使上述兩個範圍中，吻合的兩段序列也並非如關鍵字查詢的結果一樣 100% 相同。

然而，這些相同的片段卻非常重要。即使我們對虹膜缺損基因所知有限，這個結果讓我們瞭解人類的虹膜缺損基因，與果蠅的無眼基因的確有所關聯。而且，我們對無眼基因了解頗多，從它基因的結構與功能（它是 DNA 的結合蛋白，能夠促進其他基因的活動），到如何影響表現型（phenotype）—成熟果蠅的型態。

縱使在序列並非完全相符的情況下，BLAST 能夠幫我們找出相吻合的序列。BLAST 的比較不只是單一字母的排比，而是透過動態規劃，限定排比分數高於某個預先設定的閾值，反覆嘗試各種方法前後位移進行排比。於是，BLAST 能夠在序列到序列間，找到相吻合卻不完全一致的樣式，因而建構出遙遠的關係；這些關係並非全然精確，但可能具有生物學上的意義。

依照這兩個序列的排比結果，我們可以將一個序列上的生物意義，透過標記轉移到另一條序列上。從兩個完整的序列所得到的高品質序列配對，可以讓研究者得出兩者功能相似的假說，但這種假說必須要經過嚴謹的實驗才能更加確認。在無眼基因與虹膜缺損基因的例子中，科學家希望透過果蠅的無眼基因能夠讓我們更了解，虹膜缺損基因如是何影響人類的眼睛發育。

生物資訊只是建立資料庫而已嗎？

我們目前談到的大部分的生物資訊—包括序列排比、序列資料庫搜尋、序列分析—都比單純設計建立資料庫更為複雜。生物資訊學家（或計算生物學家）的工作範圍相當廣泛，不只是擷取、整理與呈現資料，或是從全然不同的領域中得到靈感（包括統計、物理、資訊科學、與工程學科）。圖 1-2 即顯示不同領域的科學與生物學相關聯的程度；從序列資料與蛋白質結構的分析，到代謝模型、族群與生態的大規模資料分析。

生物資訊是生物科學中第一，也是最重要的組成要素。生物資訊的主要目標不是發展出精緻的演算法，或是完成神秘難解的分析；而是找出生命運作的秘密。正如同大大地拓展了生物學家研究能力的分子生物學方法，生物資訊也是一種工具，但並不止於是工具而已。生物資訊學家是製造工具的人，他們必須了解生物學問題與資訊的解決方案，才能製造出有用的工具。

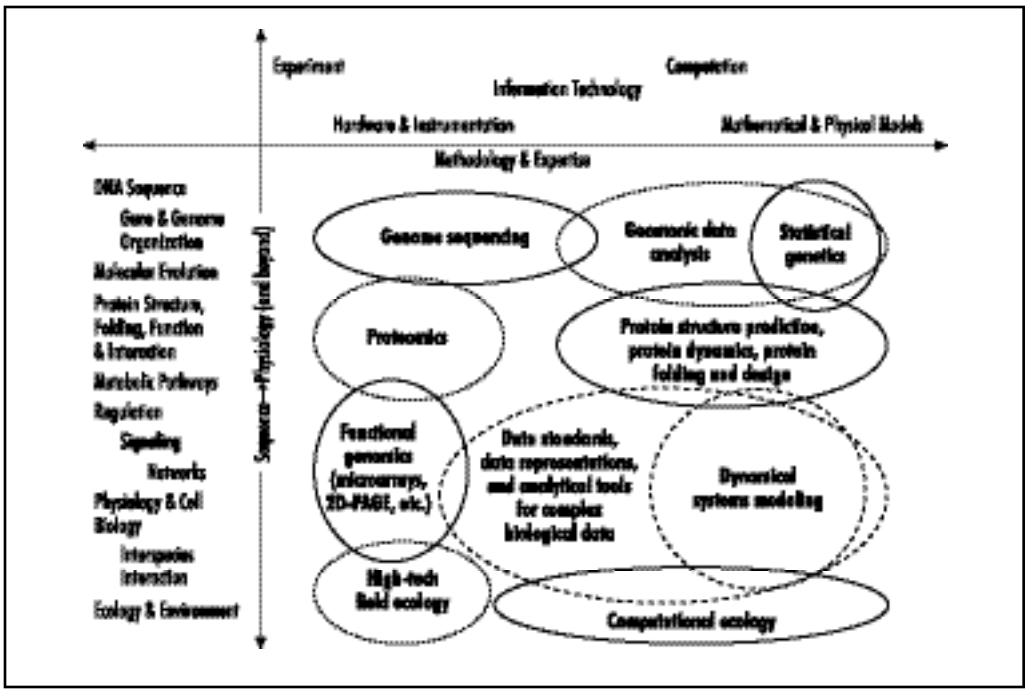

圖 1-2：不同領域的科學與生物學的關聯

生物資訊和計算生物學的研究可以將生物系統的特性變成生物與數學模型，然後再變成新的資料分析演算法，最終發展成可用的資料庫與 Web 工具。

生物學的第一個資訊時代

在生物學的領域中，生物學家必須要記住很多細節，卻也不能忽略大原則。生物學家從十七世紀開始就試圖解決資訊管理的問題。

演化概念的起源，來自於早期生物學家對生物物種所進行的分類與比較的工作。近三個世紀來，生物學家所關切的問題就是物種的分類，從動物與植物開始，一直到顯微鏡發明後所觀察到的微生物。時至今日我們仍持續地發掘新的生物型態、與未知的絕種生物化石。

這些對動植物的分類，在當時來說是非常龐大的資料。十六世紀中期，Otto Brunfels 出版了第一本描述植物物種的現代書籍《Herbarium vitae eicones》。隨著歐洲人足跡踏遍世界，已分類的物種數量也大量增加，而且植物園與植物標本室也逐漸興起。在亞里斯多德的學生 Theophrastus 的時代，植物種類有 500 種；西元 1623 年，Casper Bauhin 觀察到 6,000 種的植物。不久之後，John Ray 提出一種區分動植物物種的觀念，並以解剖學的特徵發展出能夠確定物種的指導方針。西元 1730 年代，Carolus Linnaeus 分類出 18,000 種植物與超過 4,000 種動物，並建立以界門屬種來作為分類基礎的現代分類命名系統。十八世紀末，Baron Cuvier 已經列出了超過 50,000 種的植物。

生物學家在同一時期專注在動植物的探尋與分類上，並非純屬偶然；在這段期間中的探索與分類，將生物有系統地納入分類學中。植物學文件可能包含了不同而大量的資料，每一種生物都有詳細的插圖與解說。生物學家面臨了如何組織、存取、有系統地增加資料的問題。有些業餘觀察者會發現，一些生物與另一種生物非常相似。大鼠（rat）和小鼠（mouse）的相似性就遠大於小鼠（mouse）和狗的相似性。但生物學家如何不用背著一冊冊的素描，就能知道大鼠與小鼠的相似處（rat 不只是 mouse 的別名）？於是，我們極需發明一種能夠獨特地區別不同的生物，並且理出與其他物種可能關係的命名法則。

解決方法相當簡單，但在那個時代卻是偉大的革新。生物以一串表現物種特性的單字來命名。首先區分是動物還是植物，也就是生物體所屬的界（kingdom）。然後隨著增加的各種特性，將之區分為綱（class）、屬（genera）、種（species）。這種圖表式的物種分類方法可以在圖 1-3 中看到，也就是今日所說的「生命之樹」（Tree of Life）。

縱使是對最積極的生物學家來說，地球上數百萬物種的分類資料太過複雜，以致於很難記住。還好，現在電腦可以幫忙分類這些大量的生物資料。亞利桑那大學的生命之樹計畫（Tree of Life project）以及國家生物科技資訊中心（NCBI）的分類學資料庫，就是線上分類學計畫的兩個例子。

分類學是現代生物學中的第一個資訊問題。現在生物學家已經到達了收集與分類基因資訊超載的臨界點。如何組織這些大量的資訊，並與科學社群分享基因層次的知識，並非發展一個命名法則所能處理的。這個問題在一開始就需要電腦與資料庫才能解決。

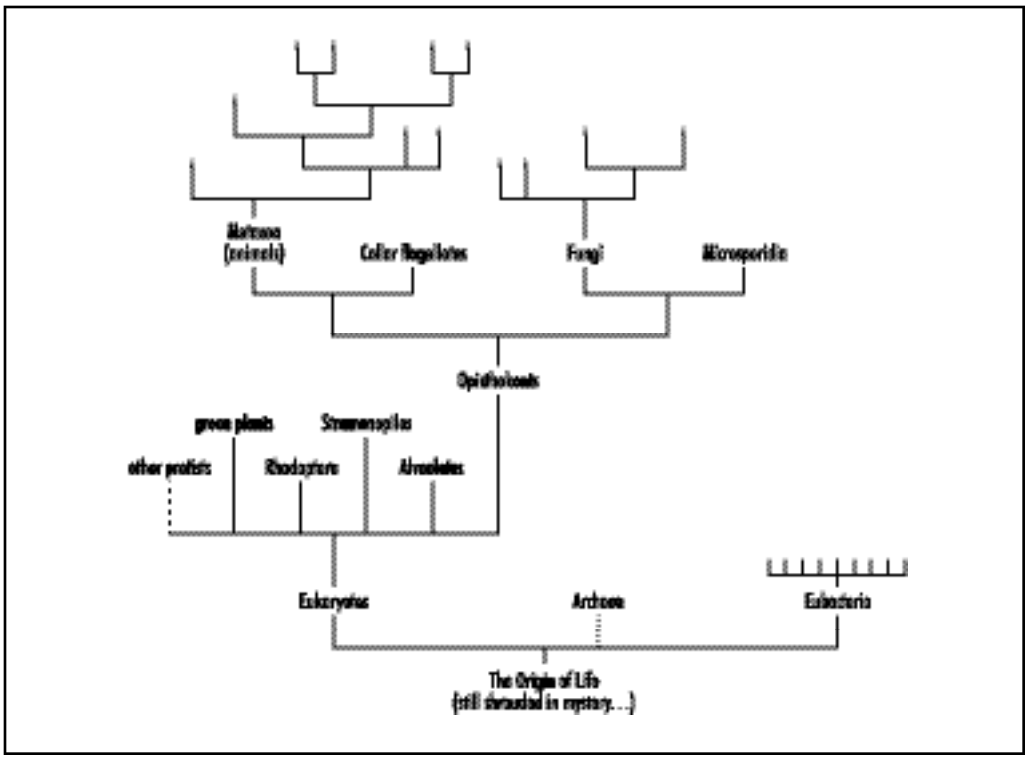

圖 1-3：「生命之樹」描述物種分類的學名命名系統

半個世紀以來，電腦的演進與物理科學偶然的平行發展，使得我們能夠逐漸地看清生物系統極細微的部分。圖 1-4 描繪了過去二十年中生物學知識驚人地蓬勃發展的成長速率。

現在，在茫茫的資訊大海中撈到正確的針，早已是一門獨立的研究課題了。在 80 年代末期，在序列資料庫中找到一個正確的配對，大概就可以能讓你寫成一份五頁的論文，然而目前這個程序已經算是例行公事了。但是，在我們擁有搜尋序列與建構資料庫的能力之後，許多問題也伴隨而來；這些問題便是生物資訊領域的前進動力。

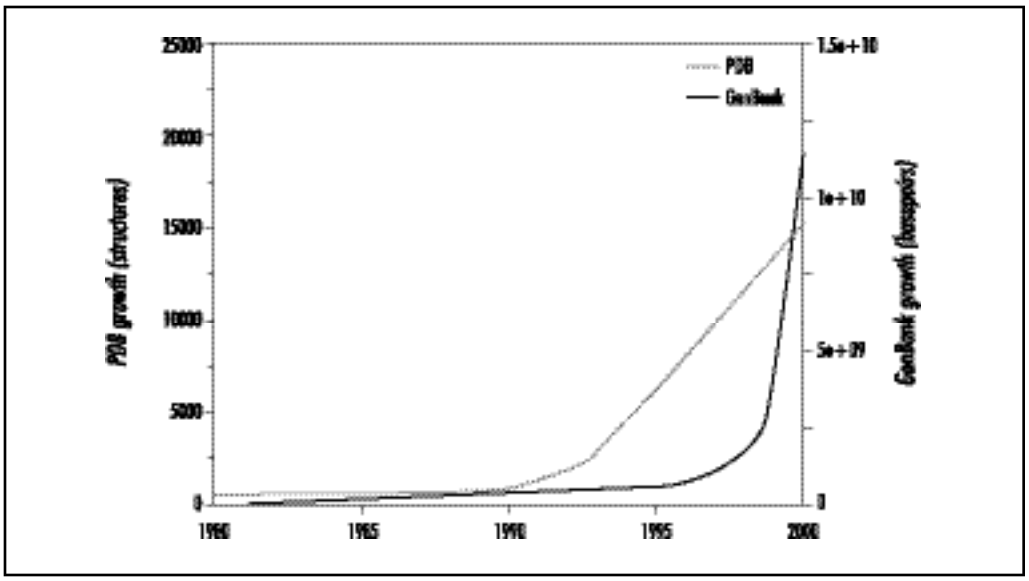

圖 1-4：GenBank 與 Protein Data Bank（PDB）的成長已相當龐大

資訊學對生物學家的意義何在？

資訊學是一門與呈現、組織、控制、散佈、維護、以及應用資訊（主要是數位形式）的科學。生物資訊與生物和資訊兩者都有相關，因此有很多不同的詮釋；你可能找到一份關於生物資訊的工作，工作內容卻與你想像的完全不同。

首先從功能面來看，生物資訊的重點在於如何呈現、儲存、與散播資料。生物資訊的基礎建設包括了設計聰明的資料格式、創造出方便查詢資料庫的工具、以及能夠讓使用者詢問複雜問題的使用者介面設計。

生物資訊第二個也較為科學的面向，是發展在資料中探詢與發現知識的分析工具。我們在不同層次使用生物學的資訊：例如，根據新發現的基因排比序列，進而發展出功能方面的假說；或者將已知的蛋白質立體結構拆解，以尋找足以預測蛋白質如何摺疊的序列樣式；又或者模擬蛋白質與細胞中的代謝物，如何協同工作讓細胞得以正常運作。分析取向生物資訊學家的終極目標，在於發展出能夠讓科學家只需根據生物的基因體序列，便能模擬其功能與表現型的預測方法。這是一個宏大的目標，必須由許多的科學家共同努力，一步一步地慢慢實現。

生物學帶給資訊科學家什麼樣的挑戰？

在基因體計畫的時代中，生物學的目標是加強對基因編碼如何組成生命體的認知。

基因體解碼的工作相當複雜。即便是在最簡單的層次，用電腦分析基因序列時，我們仍然難以辨識未知的基因；我們也仍無法預測或模擬，一條胺基酸鏈如何摺疊成特定結構的蛋白質。

在單一分子的層次之上，挑戰更為巨大。GenBank 中大量的資料目前是以指數的速率成長；當 DNA、RNA、以及蛋白質序列之外的資料類型與數量也如此龐大時，簡單的管理、存取、與清楚的呈現，都將是一個個不同的難題。人機介面專家需要與生物學研究者與臨床研究者緊密合作，來解決大量資料所帶來的問題。

生物學的資料非常複雜，且盤根錯節。例如，DNA 微陣列上的一個點不只是與其強度的即時資訊有關，還包括關於基因位置、DNA 序列、結構、功能等等不同層次的資訊。對電腦科學家而言，創造出能讓生物學家清楚掌握這些關連的資訊系統，並且不會迷失在茫茫的資訊大海中，也是一個很大的挑戰。

最後，每一個基因在基因體中並非獨立運作。多重的基因交互作用，形成了生化的反應路徑，其結果又參與了其他的反應路徑。生物化學被外在環境、病原體或其他刺激所影響。將基因體資訊與生物化學資料同時放入量化的生化與生理預測模型中，將是這個世代的計算生物學家的任務。電腦科學家、數學家、統計學家也將在這個過程中扮演著重要的角色。

生物資訊學家應該具備什麼技術？

如果你對生物資訊有興趣，可能要廣泛接觸不同的主題，而且一定沒辦法每一項都完全掌握。然而，我們與 Celera Genomics 與 Eli Lilly 這類公司工作的科學家訪談後，整理出生物資訊學家的「必備條件」：

- 你必須在分子生物學領域有一個專業背景，它可以是生物化學、分子生物學、分子生物物理學、甚至是分子模擬。如果你沒有分子生物學的專業知識，可能會如同人們所說的「經常碰壁」。

- 你必須完全了解分子生物學的「中心法則」（central dogma）。了解 DNA 序列如何轉錄成 RNA，然後轉譯成蛋白質是很重要的（在第二章中，我們會為這個中心法則下定義，並且解釋「轉錄」與「轉譯」）。
- 你必須具備至少一到兩個分子生物學軟體的實際操作經驗，不管是為了分析序列，還是為了分子模擬。這些軟體學習經驗能夠讓你更快學會使用其他的軟體。
- 你必須適應命令列的電腦環境。使用 Linux 或 Unix 可以提供這樣的經驗。
- 你必須擁有寫 C/C++、Perl、或 Python 等電腦程式語言的經驗。

還有很多進階技術能夠幫助你增強這個背景：例如，分子演化與系統體學（systematics）、物理化學—動力學、熱力學與統計力學、統計學與機率方法、資料庫設計與實作、演算法開發、分子生物學實驗方法，以及其他。

為何生物學家需要電腦？

電腦是了解任何以數學方法描述的系統，最強大的工具。隨著我們對生命的過程了解更多並且更為深入，古典生物學、數學、與電腦資訊逐漸發展成計算生物學與新近的生物資訊，這也就不令人意外了。

資料蒐集的新方法

生物化學常常是敘述性的科學（anecdotal science）。如果你注意到一種疾病或發現了感興趣的遺傳性狀（trait），了解這些問題的過程，會將研究進展帶往新的方向。基於對特定生化反應的興趣，生化學家已經能夠一次確定單一基因的序列或結構、或單一基因產物的表現特徵。通常這會帶領我們了解生化反應路徑，甚至是深入瞭解蛋白質。反應路徑或蛋白質如何與其他生物構成要素交互作用，可能是一個難解之謎；原因可能是人手不夠，甚至也可能是因為進行類似實驗的科學家之間，彼此沒有良好的溝通與互動。

網路改變了科學家分享資料的方法，並且使得整個研究社群可以共用一個資料的中央倉儲。更重要的是，實驗科技快速的發展，使得研究者能夠系統化的在一個中央「工廠」（factory）中蒐集到特定種類的全部資料，並散播給不同的研究者來解釋。

90 年代，生物學社群開始了一個史無前例的計畫：定序人類基因體中的所有 DNA。即使人類基因體序列的初稿已經完成，但自動定序器仍然日以繼夜地運作，並從各種生物研究中常用的生物型態，定序出基因體的序列資料。並且，我們也仍在整理過去十年中所蒐集到的人類基因體序列。在大量成串的資料中，我們只了解極少部分的重要基因及其位置，而資料仍在不斷地快速產出。透過運用影像處理技術，現在能夠更迅速地呈現整個基因圖譜，超越以往所使用的化學結構對應技術。但即使如此，完整而詳盡地對應出整個基因圖譜，也得花上幾年的時間才能完成。

近來，X 光結晶學（X-ray crystallography）的技術，已經能在幾分鐘之內得到一組完整的蛋白質結晶繞射資料（reflection），這在過去可能得花上幾個小時或幾天。從前可能要花幾個月確定結構，現在自動分析軟體能在幾天或幾個星期內便可完成。這項用在人類基因體計畫中的定序方法，讓我們突然發現，可以用這種高產能的方式來進行結構判定。即使蛋白質結晶化的步驟仍有限制，在未來的五到十年中，現有足供研究的蛋白質結構資料的數量仍可能會倍數地成長。

平行運算是一個發展很久的概念。將問題分解成數個電腦可運算與控制的部分，運用多個處理器同步解決這些部分，而非分次逐一解決。這種平行的方法運用 DNA 微陣列的技術時，已經在實驗分子生物學中獲得成功。微陣列技術讓研究者在非常小的晶片中，同時處理成千上萬的基因表現的實驗。微型的平行實驗（miniaturized parallel experiment）絕對需要電腦來蒐集與分析資料。他們也需要數位化的資料，因為大量資料集（large dataset）中的資訊，對收集資料者來說也許毫無用處，但可能對其他人有極大的吸引力。利用資料庫搜尋資訊，著實可為科學家省下在實驗室工作檯上幾年的工作時間。

由於網際網路與全球資訊網的發展，促成了溝通與資料傳輸的進步，也使得這些高通量實驗方法的成果能夠被研究者分享。

實驗分子生物學的自動化與資訊科技在生物學上的應用，從根本改變了生物學研究的方法。除了敘述研究（anecdotal research）—每次標示一個基因的位置，並仔細研究—我們現在將所有現有的資料分門別類，製作完整的圖譜，好讓我們日後可以在感興趣的地方加以註記。在序列和結構的領域當中也是如此，並也影響了其他型態資料的處理方法。趨勢朝向著將所有類型的原始生物資料儲存在公用資料庫，開放給研究社群使用。科學家可以先查詢資料庫，省下在實驗室作前置研究的時間與資源。

如何設定 PC 來進行生物資訊研究？

到目前為止，你可能已經在 Windows 或 Mac 這類有著友善介面的作業系統下，使用過文字處理軟體與其他套裝軟體。為了要充分掌握生物資訊學，你必須學習 Unix，這是作為伺服器或工作站，經典而強大的作業系統。大部分的科學軟體都是在 Unix 環境中研發，而嚴謹的研究者會想要用只在 Unix 上運作的程式。Unix 有幾種不同選擇，最受好評的是 BSD 和 SunOs。然而最近第三個選擇進入了市場，那就是 Linux。Linux 是 Unix 系統的開放原始碼版本。在第三、四、五章中，我們會討論如何在 Linux 上架設生物資訊工作站。我們將介紹作業系統，並說明如何下列的工作如何進行：檔案的整理，程式的運作，還有更重要的是，如何輸入正確指令讓電腦作你想做的事。

為什麼要用 Unix 或是 Linux？

在電腦上安裝 Linux 作業系統，能讓你使用專門為 Unix 系統開發的重要科學研究工具。隨著 Linux 市場佔有率的提高，Linux 保留了 Unix 系統中開發程式、編譯、執行程式、網路連線與多重使用者登入的能力，並且也提供了為桌上型電腦量身訂做的軟體版本，包括了文字處理、繪圖軟體，甚至影像編修工具。本書假設你將會學習如何使用 Unix 系統，並會在安裝了 Linux 或是其他 Unix 系統的電腦上工作。在許多特定的生物資訊工具中，Unix 是最實際的選擇。

另一方面來說，Unix 系統在 Mac 或 PC 已經佔有優勢地位的辦公室裡，就不一定很實際了。Linux 的文字處裡、桌面排版軟體、和週邊設備隨著這個作業系統的普及而有進步；然而，它還無法與這個領域中與消費者的作業系統競爭。Linux 並不會比一般 PC 作業系統更難維護，但是你所需的技術與將面臨的故障排除都會是全新的挑戰。

當我在寫這本書的時候，我的桌上型電腦已經安裝運作 Linux 將近五個月了，之間只有幾天因為硬體的關係停擺。軟體沒有當掉，也沒有當機和惱人的時候，沒有 **.dll* 檔案遺失或是出現神秘的錯誤訊息。第一次安裝 Linux 需要一些技術協助，與大約兩天的時間。第二次大概只需要一個小時（安裝在筆記型電腦上也差不多）。事實上，我遇到的主要問題是如何打開從 Mac 使用者寄來的附件檔案。

幸好，有些公司販賣的套裝 Linux 大致將安裝程序自動化，並且提供 90 天的電話與網路技術安裝支援。如 Red Hat 與 SuSE 等公司，與 Debian 這類機構，他們提供了 PC 的 Linux 發行套件；Yellow Dog（以及其他的公司）則提供了支援 Mac 系統的 Linux 發行套件。

以下是幾個讓你逐漸熟悉 Linux 的方法。當然，如果你有一台以上的電腦，你可以實驗一下將一台機器轉換成 Linux 系統，把另一台電腦保留以往習慣的作業系統。另一個選擇是使用多重開機安裝設定（dual boot installation）。在多重開機安裝設定中，你在硬碟上建立兩個區域（稱為「分割」，partition），把 Linux 安裝在其中一個，舊的作業系統則裝在另外一個區域。然後當你啟動電腦，你可以選擇要啟動 Linux 系統還是另一個作業系統。你可以將所有的舊檔案和程式保留在原處，在 Linux 的分割中開始新的工作。新版本的 Linux，如 PowerPC 使用的 Yellow Dog Linux，可以讓使用者在 Linux 上模擬 Mac 作業系統環境，並可同時使用兩種不同平台的軟體與檔案。

有哪些可用的資訊和軟體？

第六章〈全球資訊網中的生物研究資源〉中，我們將討論電腦資訊的運用。僅僅幾年之前，生物學家必須了解如何在適當的科技期刊，其出版索引中檢索所需要的參考文獻。現代生物學家在 Web 介面的資料庫中搜尋同樣的資料，同時也能找到其他類型的資訊。無論是不是數位的資料，了解如何搜尋資料對生物學家來說都是非常重要的技能。

再來將介紹如何使用資料庫、電腦程式和其他網路資源等基本工具，讓你能把這些資源傳送到你的電腦上，並且讓你一取得資料就能夠使用。第七章到第十一章，我們將討論特定類型的科學問題，以及解決這些問題所需的工具（例如胺基酸及核酸序列排比專用的 BLAST 與 FASTA）。在其他領域，當解決問題的方法仍是開放的研究議題時，可能會有幾個功能不相上下的工具，或者並沒有任何工具能完全解決問題。

為什麼我需要從網路上安裝程式？

處理大量的複雜資料需要一個有系統而自動的方法。如果你想在資料庫裡用一筆查詢搜尋資料，Web 介面就可以作到。但如果要用 10,000 筆查詢來搜尋符合的資料，然後從取得的資料中分類，並找到其中的交互關係呢？你一定不想在 Web 介面上輸入 10,000 筆搜尋，而且大概也不想讓每次查詢的結果，漂漂亮亮地以網頁格式呈現。共享的公用 Web 伺服器通常很慢，用它們來作資料批次處裡也很不實際。第十二章〈用 Perl 自動分析資料〉舉出許多例子，包含如何以 Perl 作為工具，用自己的電腦和程式處理大量的資料。

我可以不上課就學會一種程式設計語言嗎？

任何具備設計實驗與進行實驗能力的人，都有撰寫電腦程式的基本能力。實驗室中的每個實驗都是由一個問題開始，然後推演成一個可驗證的假說：也就是可透過實驗結果來檢證其真實性的敘述。為了驗證假說而發展出來的程序，可以等同於電腦程式。實驗的本質是：如果你對系統 X 做某事，會產生什麼結果？每個實驗必須經過設計，才能產生足以解釋問題的實驗結果。電腦程式也必須要被謹慎地設計，這樣當「值」從某一部份傳遞到下一個部分時，才能被清楚地解釋。程式設計師（人）必須給電腦清楚的指示設定，必須思考不同型態的結果代表著什麼意思，並且思考電腦應該怎麼處理它們。電腦程式設計實作的主要訓練，在於審慎思考的能力，能夠設計解答問題的程序，並且瞭解清楚地回答這些問題所需的資訊。

即使你有了這些能力，學習一種電腦語言仍然不是一項簡單的工作，但是最近幾年隨著 Perl 語言的發展，學習電腦語言變得越來越容易了。Perl 原作者認為 Perl 是「網際網路與所有東西的萬能膠帶」，從一開始的系統語言（scripting language），演化成為最佳化的資料處理工具。當 Perl 持續演化成為一種全功能的程式設計語言，而且開發任何類型的電腦程式或應用軟體原型的工作時，Perl 也相當實用。Perl 是一個相當有彈性的語言；你可以只學一些基礎便足以解決單個問題，然後當你實作一兩次之後，便有足以繼續發展的核心知識。學習 Perl 的關鍵在於使用它，並且是馬上使用。就像讀一大堆教科書不會讓你流暢地說出西班牙文，反覆閱讀 O'Reilly 的《Perl 學習手冊》而不動手操作，也不會讓你學會 Perl。在第十二章，我們提供了解析常用的生物學資料型態、處理其他語言開發程式輸出的結果，甚至包括解決常見計算生物學問題、實作的 Perl 範例程式碼。我們希望這些範例能夠提供一些啟發，讓你能開始嘗試寫一些自己的小小程式。

我該如何使用網路的資訊？

第六章也介紹了網際網路的公用資料庫，其中生物學資料被整理儲存成檔案，供全世界的研究者分享。

當你可以藉由填寫網頁的表單搜尋公用的資料庫，快速地找到一個蛋白質結構檔案，或 DNA 序列檔案時，接下來很可能會想要一次處理多筆資料。你會開始蒐集、儲存自己的資料檔案；也可能想要製作一種新型態的資料，提供給研究社群來運用。為了有效地進行這些事情，你需要在自己的電腦上儲存資料，並且把這些資料結構化。瞭解結構化與非結構化資料的差異，並且設計適合你資料儲存與使用的資料格式，是增加資料可用性的關鍵。

有許多方法可以組織資料。大部分生物資料仍然用很陽春的檔案資料庫方式儲存，當儲存的資料量越來越龐大，這種資料庫便變得相當沒有效率。第十三章〈建立生物資料庫〉中，將說明資料庫的基本概念，與資料庫專家討論時，或建立自己的資料庫時可能需要這些概念。我們將討論文字檔與關連式資料庫的差異，介紹最棒的公用資料庫管理工具，並且告訴你如何使用它們來儲存與取用你的資料。

我如何瞭解序列排比資料？

如果沒有視覺化的工具，就很難理解你手中的資料。想要理解生物資料，需要擷取複雜而有著多重變異的資料，進行跨段落與不同子集合的資料處理。第十三章將會討論到，如何在結構化的資料庫中儲存資料，以建立複雜資料分析的基礎建設。

一旦以易用、有彈性的格式儲存了資料，下一步就是擷取對你來說重要的資訊，並且將其視覺化。無論你想用長條圖，或顯示立體的分子結構並即時檢視它的變動，視覺化的工具可以都解決你的需求。第十四章〈視覺化與資料採擷〉中，則介紹包括資料分析與資料視覺化的工具，從一般繪圖軟體，到可用在生物序列排比標示程式、顯示分子結構、建立親緣樹（phylogenetic tree），以及其他目的的特殊用途軟體。

我如何撰寫程式來比對兩個生物序列？

任何電腦科學訓練中一個重要的部分，是判斷什麼時候需要自己動手寫程式，什麼時候你可以利用別人已經開發完成的程式碼。有效率的程式設計師是一個懶惰的專家；她從不浪費時間來開發一個別人已經完成的完美程式。如果你正在進行生物資訊的例行公事，例如比對兩個蛋白質序列，請相信一定有人已經寫好你所需要的程式，透過搜尋，甚至可能找到一些原始程式碼來看看。類似的狀況，許多數學與統計的問題，可以用函式庫中既存的標準程式碼來解決；這些函式庫都是免費的。Perl 的程式設計師用模組化的方式撰寫程式碼，藉此來讓標準運作得以簡化。有許多免費的模組可以管理 Web 相關的程序，事實上，學術界也有正在執行的研究計畫，以建立可以處理生物序列資料的標準模組。

我如何從序列資料預測蛋白質結構？

有些問題的答案我們目前還無法提供；這即是其中之一。事實上，這是計算生物學中最熱門的研究領域與難題之一。我們能夠提供你的是尋找資料的工具，以及其他正在發展中的工具，甚至有靈感的話，你自己也能找到解決的方法。生物資訊如同其他科學領域，並不總是能夠針對難題提供迅速而簡單的解答。

生物資訊能夠回答什麼樣的問題？

人類在應用生物學領域中近百年來所關切的諸多重要問題，正是驅動生物資訊研究的主要動力。我們如何治療疾病？如何預防感染？在人口爆炸的狀況下，如何生產足夠多、足以餵飽所有人類的食物？研發新藥、農藥、混種植物、塑膠與其他石化產品，以及開發環境整治的生物技術等，這類的公司企業也增設了生物資訊部門，期待生物資訊能提供新的目標並幫助減少稀有天然資源的消耗。

基因體計畫的存在，意味著人們想要運用所產生序列資料的企圖。現代分子生物學的目標，簡單地說，是瞭解生命體的完整基因體資料，進而辨識每一個基因，比對每個基因與其所組成的蛋白質，以及辨識每個蛋白質的結構與功能。我們期待根據基因序列的細部知識、蛋白質結構與功能，和基因的表現類型，可深入了解生命運作的過程。隱含於其中的是，能夠精確並正確地操控生命體的能力。