

前言

電腦與全球資訊網正在快速而戲劇化地改變生物研究的面貌。當「典範轉移」(paradigm shift) 這個詞彙近來被用在描述新興商業趨勢、甚至最新的可樂口味時，生物科學卻正在經歷標準的轉移歷程：理論生物學與計算生物學發展至今已經有數十年之久，長久以來位處生物科學的邊陲地帶，而近幾年內基因體研究產生的新數據、與利用電腦分析基因體資料的研究方法，正如潮流般地衝擊著生命科學的每一個角落。以往的研究工作始於實驗室，如今先從電腦開始，因為科學家需要搜尋資料庫，尋找任何可能推導出新假說的重要資訊。

近二十年間，個人電腦與超級電腦變成所有領域的科學家都可以使用的工具。個人電腦是非常昂貴的時髦商品，僅僅只有少量運算能力，如今的運算能力已媲美十年前的超級電腦。電腦取代實驗室中用以蒐集數據與控制實驗的設備，就如同作家的打字機或是會計的帳簿被電腦取代了一樣，它更有可能完全取代實驗室儲存資料的筆記本與檔案櫃。資料庫若以數位方式存取資料，藉由電腦存取內容將非常容易，其程度遠超過傳統的紀錄方式。除了對實驗數據資料的儲存、分析與視覺化有所幫助之外，電腦更是一個強大工具可以來協助解析任何以數學方式所描述的系統，此應用過程更促成了計算生物學與新近生物資訊學的興起。

「生物資訊」是生物資料管理面向的資訊科技應用；它是一個演化進程相當快的學科。近二十年來，將生物資料儲存在公用資料庫中逐漸變得相當普遍，這些資料庫的資料量正以指數速率成長。同時，生物學文獻的數量也以指數速率在成長。即便是最積極勤奮的學者，也沒有辦法不依賴任何電腦工具，而能在其專業領域的知識上，保持於領先地位。全球資訊網（World Wide Web）介面如此簡易，使用者只需知道如何架構正確的基本環境，就能隨心所欲，使用任何一個網站上的程式與資料庫。

自始至終，生物資訊都將是生命科學的一部分。它多半試圖回答實際的問題，較少致力於發展出一種完美、精練的演算法。生物資訊學者是打造工具的人，他們必須了解生物學的問題，以及哪些是可行的電腦解決方案，才得以製造出有用的工具。生物資訊的演算法需引用科學假設，這些假設本身的複雜程度，會導致程式設計與資料模擬變得複雜而各有特色。

生物資訊與計算生物學中的研究可能會碰到各式各樣的問題，例如將生物系統特性化簡，變成數學與物理模型，或實作新的資料分析演算法，或發展資料庫並開發存取它的網路工具。一個生物學家要使用電腦從事研究，必須要對不同作業系統上的軟體工具使用自如。本書將介紹與解釋許多生物資訊研究領域中的熱門工具，內容亦包含許多額外的資訊與背景知識，希望幫助讀者如何善用這些工具，並了解它們的重要性。我們衷心期望，這些起步的引導，能夠幫助你在研究中運用電腦提高生產力。

本書的適用對象

大部分生命科學的學生與研究者，開始應用電腦來處理在文書處理、資料蒐集與繪圖之外的需求。許多人沒有電腦科學或計算理論（computational theory）的背景知識，對他們而言，計算生物學與生物資訊的領域看起來過於龐大與複雜、漫無頭緒。這本書寫作的動機，乃是來自與學生、同事們之間的互動。我們並不想要寫一本能解決所有問題的生物資訊聖經，而是經過深思熟慮，挑選一些在生物資訊領域中極為重要的主題來討論。書中重點包括搜尋生物序列、基因體與分子結構資料庫資訊的標準電腦技術，如何辨識基因與偵測基因家族特徵樣式，種屬親緣樹關係、分子結構與生化特性的模擬問題等。我們也將論及如何運用電腦組織資料、有系統地思考關於資料分析的過程，以及開始思考資料處理如何自動化的問題。

生物資訊是一個相當先進的研究主題，所以即便像是這本介紹性的書，都預設了某些基本程度的背景知識。倘若要了解本書大部分的內容，你應該具有分子生物學、化學與數學的課堂或實作經驗。如果你有修過一兩門大學部的電腦程式設計課程將會很有幫助。

本書的結構

我們對本書內容的安排，並不限定前後順序，讀者可以依據個人喜好或需求，從頭讀到尾、或者跳著讀、先吸收後面的部分再回來閱讀前面的章節都沒關係。全書分作四大部分：

第一部分，簡介

第一章：電腦時代的生物學。首先，在此為生物資訊學這個學科下定義，回顧一點相關的發展歷史，並簡介本書所涵括的主題與其收錄的理由。

第二章：應用電腦解決生物學的問題。介紹生物資訊與分子生物學的一些核心概念，以及造成生物學資料急速增加的一些科技與研究，另列出每個生物學家應具有的的基本電腦技能—這個清單總是在不斷擴增。

第二部分，生物資訊工作站

第三章：設定你的工作站。介紹 Unix，如何在 PC 上安裝 Linux，以及軟體的安裝執行。

第四章：Unix 中的檔案與目錄。介紹在 Unix 檔案系統中的基本操作，包括檔案階層、命名方式、常用的目錄命令，以及如何在多重使用者環境中工作。

第五章：在 Unix 系統中工作。說明許多常見的 Unix 命令，包括檢視、編輯檔案、從檔案中擷取資訊；正規表示式（regular expressions）；shell script；以及如何與其他電腦溝通。

第三部分，生物資訊的工具

第六章：全球資訊網中的生物研究資源。介紹是關於在網路上尋找生物資訊的技法。該章節介紹搜尋引擎、尋找科學文章與軟體、如何使用線上資訊，以及公用生物資料庫等資源。

第七章：序列分析、逐對排比與資料庫搜尋。從檢視分子演化開始，本章包括逐對序列分析技術的基礎介紹，例如預測基因位置、整體（global）與局部（local）排比，以及利用局部排比原理為基礎的 BLAST 與 FASTA 搜尋；最後介紹多重功能的序列分析工具。

第八章：多重序列排比、親緣樹與 profile。討論用以分析一群彼此具有關係的基因與蛋白質的研究方法與工具，包括多重序列排比的策略，使用諸如 ClustalW 與 Jalview 等工具；然後介紹親緣樹分析的工具，以及 profile 與 motif 的建立與應用。

第九章：蛋白質結構的視覺化與結構特性計算。論述有關於蛋白質的立體分析以及計算結構特性的工具。從蛋白質化學的分析開始，介紹主題包括網路上既有的蛋白質結構工具，結構的分類、比對與分析，溶劑可作用性與溶劑交互作用，以及計算蛋白質的生理化學特性。本章以結構最佳化，與蛋白質資源資料庫的導覽作為結尾。

第十章：從序列預測蛋白質結構與功能。本章討論蛋白質序列中的特徵偵測、二級結構預測與立體結構預測，包含許多應用序列資料來決定蛋白質結構的工具，最後以蛋白質模擬的研究範例作為總結。

第十一章：基因體學與蛋白質體學的工具。目前為止我們討論了分析單一序列或結構的工具與技術，以及對單一基因進行多重序列比較的方法，本章將全部方法作一整合，討論某些目前可以用來研究基因體中所有基因表現之整合功能，包含全基因體定序、在網路上存取基因體資訊、註解與分析全基因體序列的資料型態與工具，以及新科技和蛋白質體學。

第四部份，資料庫與視覺化

第十二章：用 Perl 來進行自動化的資料分析。以 Perl 這類的程式語言為範例，展現其從龐大資料來源中分析擷取所需資訊的能力。本章並不教導 Perl 的程式撰寫，而是提供其簡介，並且以一些實作範例，引導讀者自學程式寫作。

第十三章：建立生物資料庫。介紹資料庫的概念，包括生物研究中所使用的資料庫類型、資料庫軟體、資料庫語言（特別是 SQL 語言），以及開發 Web 式工具來與資料庫進行互動。

第十四章：視覺化與資料採擷。介紹一些協助研究者瞭解其實驗或分析結果的資訊工具與技術。本章的第一部分，介紹一些展現生物資訊研究結果的視覺化程式，從一般用途的繪圖與處理數值資料的統計套裝軟體，例如 Grace 與 gnuplot，到 T_EXshade 這類的軟體，會將序列與結構資訊轉換成可以明瞭的結果。所謂的資料採擷，就是在大量資料中尋找、解釋以及評選特定樣式；本章的第二部分，介紹在生物資訊研究中，所使用的資料採擷工具。

我們對生物資訊的研究取向

我們必須承認，身為結構生物學家（實際上是生物物理學家），很難僅僅關注於基因序列而不去想它們的蛋白質產物。在我們的觀點裡，DNA 序列不僅是序列，除了一些例外狀況，結構生物學家眼裡的基因代表著立體的結構、分子形狀以及構形改變、活化區域（active site）、化學反應以及分子間交互作用等等的諸多細節。本書著重的，是以結構生物學家與生物化學家所慣用的觀點來使用序列資訊，也就是研究生物學功能的化學基礎，也許對一些分子生物學家與遺傳學家非常關切的序列分析應用卻著墨不多，所以請讀者不吝指教。

本書中所參考的網址

我們在這本書以及額外生物資訊的參考資料中，所參照的網址的相關資訊請參閱記載在 批評與建議 中本書的網址。

本書體裁

本書使用底下的字型體裁：

斜體字

用來表示命令、檔案名稱、目錄名稱、變數、網址。

粗體字

當第一次提及某個專有名詞時的表示方法。

定寬字

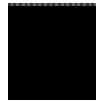
在程式碼範例中使用，顯示指令的輸出結果。

定寬斜體字

在「用法」中使用的短句，用來表示變數。



貓頭鷹的的圖案，標示一則與相鄰的內文中有關的重要註記。



火雞圖案，提醒讀者注意此處與前後文的內容之關連性。

批評與建議

我們已經竭盡所能地校閱並測試本書的內容，不過您也許會發現某些功能已經改變了（或者甚至發現書中的錯誤）。我們永遠樂意聽到讀者對於出版品的意見，包括如何讓本書更好的建議、指正本書的錯誤、或是往後改版時應該加進來的主題。底下是本公司的聯絡資料：

美商歐萊禮股份有限公司台灣分公司

電話：(02)2709-9669

傳真：(02)2703-8802

網址：<http://www.oreilly.com.tw>

電子郵件：

sales@oreilly.com.tw (業務部)

editors@oreilly.com.tw (編輯部)

bookquestion@oreilly.com.tw (疑難雜症)

請以電子郵件的方式與我們聯絡，這會比電話和傳統郵件方便。有興趣為本公司翻譯書籍的眾家高手，可與編輯部聯絡；如果您買到的書有印刷品質上的問題，可以寫信到業務部；若您對書籍內容有疑義，或是發現錯字，請寫信到bookquestion@oreilly.com.tw，謝謝您！

O'Reilly 的每一本書都有專屬網頁，你可以在此找到關於該本書籍的相關資訊，包括範例程式的下載、勘誤表與相關資源的連結。

<http://www.oreilly.com/catalog/devbioinfo/> (本書英文版的網頁)

致謝

Cynthia 的感言：我想要謝謝所有去年一年聽了我們說一千次「這本書快要完成了！」，然後還忍住沒有笑出來的朋友。謝謝家人與朋友們，在最後幾個月的時候忍受我完全渺無音訊；感謝 2000 年生物資訊課程秋季班的同學，是我第一年教授生物資訊課的學生，他們就像是實驗室白老鼠一樣，幫助我判斷哪些章節需要更為詳盡的解說；感謝維吉尼亞科技學院的同事，一年來一直討論著一些有趣的主题，像是「生物資訊代表著什麼意思？」、「生物資訊的學生應該知道些什麼？」；感謝我的朋友與同事 Jim Fenton，在本書構思過程早期的貢獻；以及我的論文指導教授 Shankar Subramaniam。另外要謝謝技術編輯，Sean Eddy、Peter Leopold、Andrew Odewahn、Clay Shirky，以及 Jim Tisdall，對本書提出許多實質的意見與非常棒的建議。最後，謝謝 O'Reilly 的工作人員，以及我們的編輯 Lorrie Leleune，在本書寫作期間付出無盡的耐心，以及在寫作過程中的支持。

Per 的感言：首先我要深深地感謝我的指導教授 Shankar Subramaniam。他在加州大學聖地牙哥校區的實驗室是很棒的工作環境，他更是我們不斷的靈感來源與支柱。我也要感謝我的兩位導師，UCSD 的 Charles Elkan 教授以及 Michael R. Brent 教授（現任教於華盛頓大學）。他們充滿智慧的導引方式，型塑了我對於電腦問題的了解。Sanna Herrgard 與 Markus Herrgard 先讀過本書的早期版本，並且提供了相當寶

貴的意見與精神支持。本書也獲益於 Ewan Birney、Phil Bourne、Jim Fenton、Mike Farnum、Brian Saunders，以及 Winny Tan 等人的回應與討論。謝謝 O'Reilly 的 Joe Johnston 提供第十二章 Perl 相關的建議與程式碼。我們的技術編輯提出重要的建議與貢獻，在此要特別感謝 Sean Eddy、Peter Leopold、Andrew Odewahn、Clay Shirky，以及 Jim Tisdall，謝謝他們在每個細節的謹慎與專注。與 O'Reilly 的同事們一起工作是相當快樂的事，特別是跟我們的編輯 Lorrie Lejeune，她很有耐心、並不斷地鼓舞著我們，以完成這個計畫。最後我想要謝謝我的家人，沒有他們的支持與鼓勵，我的部分是不可能完成的。