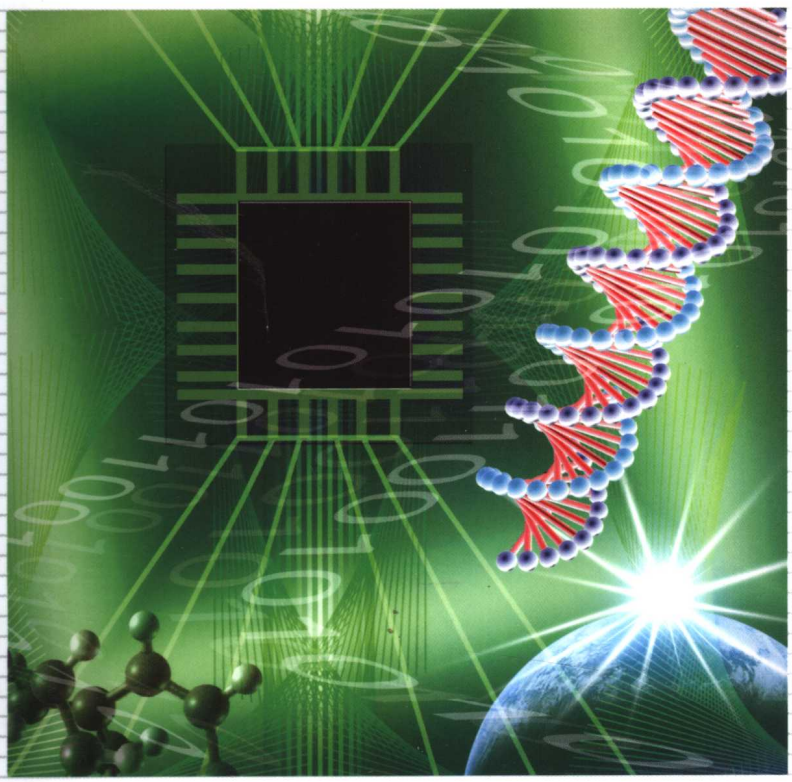
高等学校教材

生物工程 生物技术系列

简明生物信息学教程

张革新 主编



化学工业出版社
教材出版中心

高等学校教材

简明生物信息学教程

张革新　主编

化学工业出版社
教材出版中心
·北京·

图书在版编目（CIP）数据

简明生物信息学教程/张革新主编. —北京：化学工业出版社，2006.3

高等学校教材

ISBN 7-5025-8404-8

Ⅰ.简… Ⅱ.张… Ⅲ.生物信息论-高等学校-教材 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字（2006）第 022603 号

高等学校教材

简明生物信息学教程

张革新 主编

责任编辑：赵玉清

文字编辑：焦欣渝

责任校对：陈 静 宋 夏

封面设计：郑小红

*

化学工业出版社
教材出版中心 出版发行

（北京市朝阳区惠新里 3 号 邮政编码 100029）

购书咨询：(010)64982530

(010)64918013

购书传真：(010)64982630

http://www.cip.com.cn

*

新华书店北京发行所经销

化学工业出版社印刷厂印装

开本 787mm×1092mm 1/16 印张 11¾ 字数 271 千字

2006 年 6 月第 1 版 2006 年 6 月北京第 1 次印刷

ISBN 7-5025-8404-8

定 价：25.00 元

前　言

生物信息学（bioinformatics）是一门新兴的交叉学科。由于有各种各样公开的互联网资源可用，那些对生物信息学感兴趣的人要开始对它的研究并不困难。生物信息学不仅是一门新兴的学科，随着基因组研究的发展，它又是一门覆盖面极广的综合性学科。显然，在这样一本教材中想要囊括生物信息学的理论以及在相关学科中应用是不切实际的。因此，本教材以介绍方法为主，并选择一些有代表性的实例来使读者加深对方法的理解。

全书除第1章外共分为四部分：①计算机技术；②网络资源及其利用；③生物信息学数据挖掘方法；④生物信息学研究实例。教材内容紧紧围绕应用的目的，尽可能做到深入浅出，同时也有适量的理论推导，使读者能在理解的基础上掌握各种方法的适用条件、应用范围、优缺点等。为使读者巩固概念，掌握方法，第1～9章都安排了一定量的复习思考题供读者练习。生物信息学研究实例是编者研究工作的一部分。由于本书论及一个典型的交叉学科研究领域，涉及基础知识比较多，这些内容不可能在本书中都作介绍，有需要的读者可阅读有关书籍。

本书是由3位编者共同完成。汪明磊编写第1，7，10章和第2章的2.1～2.3节；宋江宁编写第3，4，5，11章；张革新编写第6，8，9，12章和第2章的2.4节，以及复习思考题。张革新负责组织和全书的审校工作。

本书的编者们都是在科研第一线从事生物信息学或与生物信息学相关研究的人员，化学工业出版社编辑为此书的出版付出了辛勤劳动，对于他们在百忙中完成这一写作任务及促成本书出版表示深切的感谢！在本书的编写过程中，参考了大量的互联网资源及有关参考书，在此谨对为本书提供帮助的网站和作者表示衷心的感谢。由于时间限制，书中存在疏漏在所难免，希望得到读者的批评和指正。3位编者的Email分别如下：汪明磊 wml_yh@yahoo.com.cn；宋江宁 jiangning_song@yahoo.com.cn；张革新 zgexin@pub.wx.jsinfo.net。

目前生物信息学的研究正在飞速发展，我们相信，今后生物信息学发展还会不断加速，因此本书所涉及的内容将会不断更新。随着时间的推移，书中内容恐有落后于研究进展之处，尚希读者予以谅解。

张革新

2005年12月于江南大学

目　录

第 1 章　什么是生物信息学

1985 年美国科学家率先提出了人类基因组计划（human genome project，简称 HGP），其目的在于阐明人类基因组核苷酸序列，破译人类全部遗传信息。HGP 于 1990 年正式启动。随着 HGP 产生的数据爆炸，一门新兴学科——生物信息学应运而生。

1.1　生物信息学的起源和特点

传统的生物学研究是一种基于观察，而不是基于推理的科学。但是在过去的十余年里，由于测序技术的飞速发展使分子生物学经历了信息革命时代。随着各项大规模的测序计划的完成，尤其是随着人类基因组计划的完成、多种生物的基因组全序列相继公布、蛋白质组学的迅速发展，使生物大分子序列数据获得了海量积累。只有使用计算机技术，才有可能应付日益增长的生物信息数据。在 21 世纪初，生物科学的重点正在从 20 世纪的试验分析和数据积累，向数据分析及其指导下的试验验证方向转移，生物科学正在经历着从分析还原思维到系统整合思维的转变。

早在 1956 年，在美国召开的首次“生物学中的信息理论研讨会”上便产生了生物信息学的概念。1987 年林华安（Hwa A. Lim）博士正式为这一领域定下“生物信息学”（bioinformatics）这一称谓。此后，计算机在生物学中的广泛应用孕育了生物信息学这一新兴学科。目前，生物信息学是生命科学中最活跃的领域之一，是一门由生物学、计算机科学、数学、物理学、化学等学科相互结合而产生的学科。目前，一般认为，生物信息学主要是一门研究生物学系统和生物学过程中信息流的综合系统科学，通过其独特的桥梁作用和整合作用，使人们能够从各生物学科众多分散的观测资料中，获得对生物学系统和生物学过程运作机制的理解，最终达到自由应用于实践的目的。生物信息学的实质就是利用计算机科学和网络技术来解决生物学问题。

1.2　生物信息学内涵

1.2.1　生物信息学的科学基础

生物信息学从事对生物信息的获取、加工、储存、分配、分析和释读，并综合运用数学、计算机科学和生物学工具，以达到理解数据中的生物学含义的目的。与此相应，生物信息学具有三方面的科学基础：首先，它需要发达的、复杂的、可相互交流的数据库系统；其次，生物信息学需要强有力的创新算法和软件；最后，也是十分重要的一个方面，是自动化的大规模高通量的生物学研究方法与平台技术。

1.2.2　生物信息学的意义

从宏观上说，生物信息学的发展必然会推动与之相关的前述各个学科的进一步发展，

并以此为基础萌生出一系列分支学科，如 DNA 计算。这一切，一方面给各个领域的发展带来了无限的机会；另一方面，伴随着生物信息学的发展，人类必将揭示更多的生命活动本质规律，其中当然会有很多是与人类自身健康、疾病、衰老、死亡相关的生物信息，而它们的发现必然导致新药物的设计与研发周期大幅度变短以及基因治疗的最终实现，从而彻底地改变人类自身的命运，这无疑是人类文明的又一次飞跃。

从微观上说，生物信息学的产生和发展对于生命科学的研究具有划时代的意义。它第一次大量地在生物学中引入了数学模型，它标志着生物学已经从实验学科向理论学科转变。对于生物学本身而言，这就是一次从量变到质变的飞跃。在生物信息学形成以前，一切生物学理论的发展都是通过大量实验证据所得到的经典理论，然而生物信息学加入之后，我们终于看到了一种希望，生物学理论的研究用于指导、验证实验生物学。这将会使得实验生物学的目的更加明确，并且大大缩短实验周期。当然，目前的生物信息学还难以实现这样一种宏伟目标，但是随着研究的深入，理论的进一步成熟，生物学家必将建立一个属于理论生物学家的“生物元素周期表”，生物学将从此走上理论学科的光辉大道，并且生物信息学最终将推动应用生物学的蓬勃发展。

1.2.3 生物信息学的研究内容

（1）生物信息的收集、存储和管理　现在已经建立了种类繁多的生物信息学数据库，在核酸方面，GenBank、EMBL、DDBJ 三个一级骨干数据库的建立，收集了所有已测序的核酸序列；蛋白质方面则有 SWISS-PROT、PIR 等一系列一级序列库；同时，PDB 等结构数据库提供了生物大分子详细的结构信息；在此基础之上，又创建了种类繁多的二级数据库。

（2）基因组序列信息的提取和分析　面对数量巨大且发展迅猛的数据，如何检出所需要的特定信息以及如何从大量的数据中发现可能存在的规律，探讨生命活动的本质，已成为当务之急。同时，在一个数据库搜索结果中可能会包含大量的信息，这往往会掩盖有意义的信息。因此，快速、功能强大的信息检索工具以及搜索后信息的自动化处理技术成为迫切的需要。

基因识别的基本问题是给定基因组序列后，正确识别基因的范围和在基因组序列中的精确位置。非编码区由内含子（intron）组成，一般在形成蛋白质后被丢弃，但从实验中，如果去除非编码区，又不能完成基因的复制。显然，DNA 序列作为一种遗传语言，既包含在编码区，又隐含在非编码序列中。分析非编码区 DNA 序列目前没有一般性的指导方法。在人类基因组中，并非所有的序列均被编码，已完成编码部分仅占人类基因总序列的 3%～5%，显然，手工搜索如此大的基因序列是难以想象的。侦测密码区的方法包括测量密码区密码子（codon）的频率、一阶和二阶马尔可夫链、开放阅读框（open reading frames，ORF）、启动子（promoter）识别等。

（3）功能基因组相关信息分析　功能基因组学是后基因组研究的核心内容，它强调用发展和整体的实验方法分析基因组序列信息来阐明基因功能，特点是采用高通量的实验方法结合大规模数据统计进行研究。

在后基因时代，生物信息学家面对的不只是序列和基因，而是越来越多的完整基因组。人们想知道，独立生活的最小生物至少需要多少基因？这些基因是如何使它们活起来的？不同物种之间基因组大小相似，可差异为何如此之大？由完整基因组研究引出的比较

基因组学必将为后基因组研究开辟新的领域。在分子水平对基因表达调控进行研究，如调节蛋白是如何作用于它的顺式调控元件（如启动子、增强子等）的？调节蛋白又是被什么调整的？依次下去就成了一个网络。功能基因组研究开展后，将使我们的认识上升到一个新阶段：基因是如何动作来产生结构、产生信息的？这样生物就“活”起来了。

（4）生物大分子结构模拟和药物模拟　人类基因组计划的目的在于阐明人体 3 万～4 万种蛋白质的结构、功能、相互作用以及人类各种疾病之间的关系和各种预防的方法，还包括药物治疗在内的治疗方法。

蛋白质结构比对和预测基本问题是比较两个或两个以上蛋白质分子空间结构的相似性或不相似性。蛋白质的结构与功能是密切相关的，一般认为，具有相似功能的蛋白质结构一般相似。蛋白质是由氨基酸组成的长链，长度从 50AA（amino acids）到 1000～3000AA，蛋白质具有多种功能，如酶、物质的存储和运输、信号传递、抗体等。氨基酸的序列决定了蛋白质的三维结构。一般认为，蛋白质有四级不同的结构。研究蛋白质结构和预测的理由是：医药上可以理解生物的功能，寻找对接药物（docking drugs）的目标；农业上获得更好的农作物的基因工程；工业上有利用酶的合成。直接对蛋白质结构进行比对的原因是由于蛋白质的三维结构比其一级结构在进化中更稳定地保留，同时也包含了较氨基酸序列更多的信息。蛋白质三维结构研究的前提假设是内在的氨基酸序列与三维结构一一对应（不一定全真），物理上可用最小能量来解释。从观察和总结已知结构的蛋白质结构规律出发来预测未知蛋白质的结构。同源建模（homology modeling）和指认（threading）方法属于这一范畴。同源建模用于寻找具有高度相似性的蛋白质结构（超过 30%氨基酸相同），后者则用于比较进化族中不同的蛋白质结构。然而，蛋白结构预测研究现状还远远不能满足实际需要。

经过 30 余年的努力，蛋白质结构预测研究现状仍远远不能满足实际需要。预测蛋白质空间结构，进而实现针对性的药物设计，已经是迫在眉睫的任务。

（5）生物信息分析的技术与方法研究　如发展有效的能支持大尺度作图与测序需要的软件、数据库以及数据库分析工具，改进现有的理论分析方法，创建适用于基因组信息分析的新方法、新技术等。

生物信息的规模之大给数据挖掘提出了新课题和挑战，需要新的思想的加入。常规的计算机算法虽仍可以应用于生物数据分析中，但越来越不适用于序列分析问题。究其原因，是由于生物系统本质上的模型复杂性及缺乏在分子层上建立的完备的生命组织理论。Simon 曾给出学习的定义：学习是系统的变化，这种变化可使系统做相同工作时更有效。机器学习的目的是期望能从数据中自动地获得相应的理论，通过采用如推理、模型拟合及从样本中学习，尤其适用于缺乏一般性的理论、“噪声”模式、大规模数据集。因此，机器学习形成了与常规方法互补的可行的方法。机器学习使得利用计算机从海量的生物信息中提取有用知识、发现知识成为可能。机器学习方法在大样本、多向量的数据分析工作中发挥着日益重要的作用，而目前大量的基因数据库处理需要计算机能自动识别、标注，以避免既耗时又花费巨大的人工处理方法。早期的科学方法（观测和假设）面对快速的数据获取率和客观分析的要求，已经不能仅依赖于人的感知来处理了。因而，生物信息学与机器学习相结合也就成了必然。机器学习中最基本的理论框架是建立在概率基础上的，从某种意义来说，是统计模型拟合的延续，其目的均为提取有用信息。机器学习与模式识别和

统计推理密切相关。学习方法包括数据聚类、神经网络分类器和非线性回归等。隐马尔可夫模型也广泛用于预测 DNA 的基因结构。目前研究重心包括：

① 观测和探索有趣的现象　目前机器学习方法（ML）研究的焦点是如何可视化和探索高维向量数据。一般的方法是将其约简至低维空间，如常规的主成分分析（PCA）、核主成分分析（KPCA）、独立成分分析（independent component analysis）、局部线性嵌套（locally linear embedding）。

② 生成假设和形式化模型来解释现象　大多数聚类方法可看成是拟合向量数据至某种简单分布的混合。在生物信息学中聚类方法已经用于 microarray 数据分析、癌症类型分类及其他方向中。机器学习也用于从基因数据库中获得相应的现象解释。机器学习加速了生物信息学的进展，也带来相应的问题。机器学习方法大多假定数据符合某种相对固定的模型，而一般数据结构通常是可变的，在生物信息学中尤其如此。因此，有必要建立一套不依赖于假定数据结构的一般性方法来寻找数据集的内在结构。其次，机器学习方法中常采用“黑箱”操作，如神经网络和隐马尔可夫模型，对于获得特定解的内在机理仍不清楚。

1.3 生物信息学的研究现状和趋势

国际上一直非常重视生物信息学的发展，各种专业研究机构和公司大量涌现出来，生物科技公司和制药工业内部的生物信息学部门的数量也与日俱增。对生物信息学的需求如此迅猛，以至于像美国这样的发达国家也面临着供不应求、人才匮乏的局面。

尽管在许多大学和研究机构已经成立了自己的生物信息学部门或中心，1999 年 6 月 3 日，美国国家卫生研究院（NIH）的专家委员会还是建议，迅速在大学和研究机构中建立 20 个生物计算中心，给予每个中心每年 800 万美元的支持，从事有关研究和人才培养。

近来，英国鉴于国内对生物信息学专业人才日益迫切的需求，所有主要的研究资助机构医学研究委员会（Medical Reasearch Council，MRC）、生物技术和生物科学研究委员会、工程学和物理科学研究委员会（Engineering and Physical Sciences Research Council，EPSRC）、粒子物理和天文学研究委员会（Particle Physics and Astronomy Research Council，PPARC）等不仅已经达成共识，认为应该高度优先地满足对生物信息学技术的需求，而且已经实现了对生物信息学人才培养的大力资助。

事实上，欧美等发达国家在生物信息方面已有较长时间的积累。

从数据库的角度来讲，早在 20 世纪 60 年代，美国就建立了手工搜集数据的蛋白质数据库。美国洛斯阿拉莫斯国家实验室 1979 年就已经建立起 GenBank 数据库；欧洲分子生物学实验室 1982 年就已经提供核酸序列数据库 EMBL 的服务；日本也于 1984 年着手建立国家级的核酸序列数据库 DDBJ 并于 1987 年开始提供服务。

从专业机构的角度来讲，美国于 1988 年在国会的支持下成立了国家生物技术信息中心（NCBI），其目的是进行计算分子生物学的基础研究，构建和散布分子生物学数据库；欧洲于 1993 年 3 月就着手建立欧洲生物信息学研究所（EBI）；日本也于 1995 年 4 月组建了自己的信息生物学中心（CIB）。

从数据分析技术的角度来讲，早在 1962 年，Zuckerkandl 和 Pauling 就将序列变异分

析与其演化关系联系起来，从而开辟了分子演化的崭新研究领域；1964 年，Davies 开创了蛋白质结构预测的研究；1970 年，Needleman 和 Wunsch 发表了广受重视的两序列比较算法；1974 年，Ratner 首先运用理论方法对分子遗传调控系统进行处理分析；1975 年，Pipas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构；随着 1976 年之后大量生物学数据分析技术的涌现，Science 于 1980 年第 209 卷就已经发表了关于计算分子生物学的综述。正如我们现在所看到的那样，在 20 世纪 80～90 年代，生物学数据分析技术在国外更是获得了突飞猛进的发展。

从专业出版业来看，由于没有专业领域专门的期刊，起初的专业文献都散落在各种其他领域的期刊中。到了 1970 年，出现了《Computer Methods and Programs in Biomedicine》这一相关期刊；到 1985 年 4 月，就有了第一种生物信息学专业期刊——《Computer Application in the Biosciences》；现在，我们可以看到的专业期刊已经很多了，包括书面期刊和网上期刊两种，如《Bioinformatics》(Formerly Computer Applications in the Biosciences)、《Acta Biotheoretica》、《Bioinformatics Technology & Systems》、《Bioinform Newsletter》、《Briefings in Bioinformatics》和《Journal of Computational Biology》等。

从网络资源来看，国外互联网上的生物信息学网点非常繁多，大到代表国家级研究机构的、小到代表专业实验室的都有。大型机构的网点一般提供相关新闻、数据库服务和软件在线服务，小型科研机构一般是介绍自己的研究成果，有的还提供自己设计的算法的在线服务。总体而言，基本都是面向生物信息学专业人士，各种分析方法虽然很全面，但却分散在不同的网点，分析结果也需专业人士来解读。

目前，绝大部分的核酸和蛋白质数据库由美国、欧洲和日本的三家数据库系统产生；他们共同组成了 DDBJ/EMBL/GenBank 国际核酸序列数据库，每天交换数据，同步更新。其他一些国家，如德国、法国、意大利等，在分享网络共享资源的同时，也分别建有自己的生物信息学机构、二级或更高级的具有各自特色的专业数据库以及自己的分析技术，服务于本国生物（医学）研究和开发，有些服务也开放于全世界。

国内对生物信息学领域也越来越重视，在很多领域取得了一定成绩，在国际上还占有一席之地，如北京大学的罗静初和顾孝诚教授（在生物信息学网站建设方面）、中国科学院生物物理所的陈润生研究员（在 EST 序列拼接方面以及在基因组演化方面）、天津大学的张春霆院士（在 DNA 序列的几何学分析方面）、中国科学院理论物理所郝柏林院士、清华大学的李衍达院士和孙之荣教授、内蒙古大学的罗辽复教授、上海的丁达夫教授等。北京大学于 1997 年 3 月成立了生物信息学中心，中国科学院上海生命科学研究院也于 2000 年 3 月成立了生物信息学中心，分别维护着国内两个专业水平相对较高的生物信息学网站。但从全国总体上来看与国际水平差距很大，一方面，国内生物（医药）科学研究与开发对生物信息学研究和服务的需求市场非常广阔；另一方面，真正开展生物信息学具体研究和服务的机构或公司却相对较少，仅有的几家科研机构主要开展生物信息学理论研究，声称提供生物信息学服务的公司所提供的服务也仅局限于简单的计算机辅助分子生物学实验设计，而且服务体系并不完善。目前国内互联网上已经有了几家生物信息学网站，但大部分偏于所有生物（医）学领域的新闻报道，生物信息学专业技术服务的含量太少，这就与国外有了较大差距。

1.4 生物信息学的学习和实践

生物信息学是一门崭新的学科，完全体现了在过去十几年内，由于基因组学研究的迅速发展而引起的对海量生物学数据的管理和解释的需要，覆盖了基因组学、生物技术和信息技术，包含了对数据的分析和解释，对生物现象的建模，以及对算法和统计学的发展。正因为这是一门交叉性极强的学科，使得生物信息学的教学变得相对困难和具有极大的挑战性。

生物信息学最初由于 Margaret O. Dayhoff、Walter M. Fitch、Russell F. Doolittle 等的开创性工作而起始于 20 世纪 60 年代，至今已成为一门全面发展的成熟的学科。对于包括笔者在内的很多研究人员来说，生物信息学笼统地包含有分子进化、生物建模、生物物理和系统生物学；对于其他更多的研究人员来说，它就是一门将计算机技术应用到生物学领域的学科。但不可否认的是，生物信息学是一门在科技的前沿蓬勃发展的学科，人类正越来越注重于从这个“鲜活”世界得到数据、传播数据和使用这些数据。生物信息学正是处于这一切活动的中心地带，成为非常引人注目的研究领域，而这一领域的学生也将在产业和学术方面的需求中大受其益。

1.4.1 对生物信息学学科的定位

随着计算机技术的发展，人类已经成为“数据的收集者”，管理着生活的方方面面。任何事物都能够，也将变为数据，任何活动都是一种对数据的收集，而这些数据是可存储的、可传播的、可分析的。我们从数据中正获取巨额的“利润”：对事物的预期控制（从地震、疾病到经济和社会的稳定），对化学、生物学、天文学……的深刻理解。但不幸的是，数据同样也带来了混乱和噪声，对于数据的认识和解释的速度远远赶不上数据的积累速度。这大量的数据最重要的问题就是它的多维性，以及如何最大可能地解释它内含的信息（通常用非线性的方法）；另一个主要问题是，我们用这些数据能做什么。科学有时总是由谬误和想象所推动的，而不是纯粹的由观察到理解的逻辑过程。如果使用了错误的指导思想，数据将不能产生知识。

对生物学数据的收集、传播、建模和分析是科学探索中一个相对年轻的领域，通常被称为“生物信息学”、“计算生物学”、“生物分子信息学”或者“计算分子生物学”。其中有些名称比其他更为严格，某些也只是专指将生物大分子作为计算的对象（如：计算分子生物学）。笔者更倾向于将那些由数据推动的领域视为生物信息学。生物信息学正是借由计算机的普及而带来的人类计算能力的快速增长和基因组学的发展而发展的。基因组学的研究使得从广泛的来源获取核酸序列和结构信息变为可能，并且将这些信息变得可以被进一步分析和研究。比如，序列被用于与那些已知结构的球蛋白比对，以及被用于高通量的分析方法（比如 DNA 的微阵列）来研究基因表达的模式。由序列和生化数据推导出来的信息被用于构建代谢网络。这些研究产生了海量的数据，正在用计算机、统计学和机器学习的方法进行分析。由这些探索活动而得到的信息会给人以这样的假象：想象和假说在生物学数据的获取中不会起任何作用。在这个复杂多变的现实世界中，生物信息学成为了唯一一门由假说推动研究的科学。生物信息学能够提供强有力的方法来理解生命的复杂性，而这个理念需要在教学过程中被清楚地表达。

和生物信息学相关的学科有很多方面，最常见的是与生物分子相关的知识，因此，既

要求对生物化学、分子生物学、分子进化、热动力学、生物物理、分子工程和统计方法都有了解，同时又要求对计算机科学、数学和统计学原理的掌握，实验通常也是生物信息学研究的一部分。一个典型的例子就是 RNA 在空间的折叠。RNA 的折叠服从于一个相对简单的基本的生物物理学模型，这就是表现型反映了基因型，并且表达了一个丰富的统计结构，它包括了神经网络和在遗传的可变性和分子的可塑性方面的合适的变化。这个生物物理模型和真实世界相联系，成为了 RNA 实际三维结构的几何学、热动力学的框架。有研究显示，在线虫的体外实验中，RNA 序列中存在中性路径，这一令人震撼的研究结果表明，在实验科学和理论科学之间有着共同的领域，因为它关系到大分子世界的特征。生物信息学就是处于实验科学和理论科学的交叉路口，它不仅与数据建模和挖掘有关，也是从进化和机理的角度对分子世界的理解。如同生物技术和基因组学一样，生物信息学正在从应用转变为基础科学，从开发工具转变为发展理论与假说。

在生物信息学中有两大主题：进化与复杂性。前者往生物学中加入了动态的因素，后者则是从全局的观点来研究生物系统的相互作用，这两大主题是互补的。

1.4.2　对教学需求的定位

生物信息学代表了多项专门技术的融合，因此生物信息学的教学就成为一项极富挑战意义的工作。由于不同学科的学生教学需要和教学目的变化很大，这一学科的教学变得尤为困难。为了满足计算机研究人员的需要，必须提供基本的生物化学方面的知识，这不仅包括对生化过程的介绍，而且包括对分子生物学、基因组学和生物技术相关的问题和思想的介绍。而为了满足生物学研究人员的需要，又必须讲授计算机科学与信息技术的基本技术和原理，包括程序的编写技术。因为很多方面必须被表达和理解得足够清楚（比如生物化学和统计学原理），所以学生的背景要求相对广泛，要对所涉及的领域都有基本的了解。生物物理、数学、统计学方面的学生都可以加入进来。但与此同时，对专业背景、研究兴趣都有很大差异的学生授课极大地增加了教学的复杂程度，必须使课程变得有效且有用。比如，对于农学院的学生，能够从生物信息学课程得到很大的益处，特别是他们以后如果是从事与此相关的研发工作（如农业生物技术、农业化学、生物材料以及营养学）。课程需要以较快的速度讲授，覆盖的面要力求宽泛，但不能对深度有过高的要求。相反，对于计算机专业的学生，需要要求他们更好地去适应生物信息学的研究方法，如何掌握生物信息学研究中统计学和机器学习的应用，这时就并不需要一个快节奏的教学，也不必过度强调生物学因素的重要性。那么，我们怎样以有效的方式完成这些呢？怎样合理地组织生物信息学中的各门课程呢？

一个主体必然需要一个客体。幸运的是，生物信息学是一门新兴的学科，许多不同的领域正在加速融入这个新体系当中，但是这门学科的研究者具有不同的广泛的背景知识以及不同的研究目的。如前文所述，大量不同专业的学生正在这一领域内寻找不同的挑战。对于一部分人来说，生物信息学就是开发更好的计算机程序，开发更全面更复杂的数据库，开发最精妙的算法，或者是寻找能够覆盖更广泛的模型而接近真实的世界；而对另一部分人来说，生物信息学就是一整套工具，它们能帮助人们研究不同的生物学领域，帮助研究不同的基因和生物学过程，帮助人们取得更广泛的观点；甚至对一部分人来说，生物信息学代表了他们的职业中更具竞争性的方面。因此，不同的利益将造成不同水平的研究动机和研究目的。很多致力于生物信息学以外的学科的研究人员，希望有更好的教学来更深地覆盖他们的研究项目所涉及的内容，相反，另一部分将生物信息学视作自己学科的补

充的人则只是希望更广泛的而不追求细节的覆盖。因此，当需要的时候，他们将只是探索生物信息学中对他们有用的方面。

1.4.3 教学策略的定位

目前，在很多高校与科研机构，都开设有与生物信息学相关的课程，这都是由于后基因组世代的到来而出现的科学发展的一种趋势。但教学方法却不相同，生物信息学若要作为一门专业被开设，就必须要有足够数量的教师，但这是很困难的，因为这个发展中的领域的专家自身也正在发展。如果只是希望注重于生物学中的“计算生物学”，那么使用已有的师资，将是非常有效的，但是这种情况下，教学需要作适当的调整。同样，个别的生物信息学课程也可以很有效地加入到已经很完善的其他生物学课程中间去。后两种策略是服务于不同目的的，但它们的共同之处在于必须加强个别课程的教学。

一个完全不同的策略是建立起一个能够覆盖生物信息学所有方面的，并能提供相应基本知识的一整套课程体系的框架。这一整套体系必须能联系到其他的不同专业。这样的方法有其自身的优点。从开始就具有的广泛的知识背景将能为学生提供更好的理解本专业各个侧重点的能力。对进化感兴趣的学生，将会注重基因组学（如功能基因组学、结构基因组学、比较基因组学等）和进化生物学（如系统发育、群体生物学）的学习；而对复杂性感兴趣的学生则会更注重生物化学、生物物理学（如生物分子工程、纳米生物技术、代谢网络、系统生物学）的学习。对于那些从计算机科学或统计学领域来的学生将能接受到从事生物信息学研究所必需的生物学基本知识的教育；对原来就是生物学领域的学生将能拓展他们的知识面，并能准备好在统计学、生物物理学、计算机科学或生物分子工程方面的学习。笔者认为，这是一个有效的、能节约大量教学资源的策略。课程的教材也不应当只用一本书，而应当使用多本书与讲义结合。

1.4.4 生物信息学的教材

目前由于国内尚没有统一编写的教材，各个学校使用的教材各不相同，甚至是教师根据自己的科研经历及学校的实际情况编写的讲义，虽然各具特色，并且已经对生物信息学人才的培养起到了很大的推动作用，但没有统一的教材作为参考，显然是不利于该门学科的长远发展与规划的，因此编写教材是一项当务之急的工作。在目前情况下，除了教师自己编写的讲义之外，还应推荐学生（尤其是研究生）阅读相关的外文期刊，如《Bioinformatics》、《Nuclide Acid Research》数据库专刊等。前者主要刊登生物信息学领域目前最新的研究进展和高水平的研究论文，对学生开阔视野、拓展科研思路、把握该领域世界最前沿的研究动态有重要作用；后者并非专门的生物信息学期刊，但它每年的第一期专门用来刊登生物学领域内数据库的建立发展的近况，而生物数据库是生物信息学研究的重要基础，有着直接的制约作用。

1.4.5 生物信息学的教学方法

由于目前国内的生物信息学本科专业尚不多见，在多数学校，生物信息学作为研究生的研究方向，通常是挂靠在其他专业之下，如分子生物学专业、自动化专业、发酵工程专业等，这样就使得学习生物信息学的学生的水平各有偏重，有些偏向于生物，有些则偏向于计算机，有些偏向于理论，有些偏向于工程等。实际上，生物信息学的教师多数也是从不同的研究领域转过来的，而生物信息学本身又要求将这些原本是不同领域的知识很好地统一起来，从而使这门学科的教学较为困难。对此，笔者提出了以下方法：

对不同专业来源的学生因材施教，充分发挥其已经具有的长处，不断弥补其不足之处。在教学初期，应尽快了解学生的现有水平，对计算机知识不足的学生，要帮助其在短时间内熟悉计算机操作、一些生物信息学研究常用的工具软件的用法，并至少掌握一门常用的编程语言（如 Perl、Fortran、VC^{++}、VB 等）；而对生物学知识不足的学生应推荐其阅读分子生物学、结构生物学的入门书籍，熟悉生物领域的常用名词与基本知识。

在生物信息学教学中，尤其是对研究生的教学过程中，应避免单纯的课堂讲解，在经过初期的“突击”后，可以要求学生通过亲手操作具体案例，来提高学生的水平。例如，生物信息学研究涉及大量的文本操作，可以让学生自己动手编程实现文本的搜索、提取。互联网在生物信息学教学中扮演着重要的角色，在教学中应充分利用，使学生能够确实掌握网上主要的核酸与蛋白质数据库的组织结构与使用方法，熟悉生物序列的查询方法。

生物信息学的研究方法涉及到很多的数学与统计知识，对此无论是对于来自计算机领域，还是来自生物领域的学生，短期内提高均有一定的困难。因此，笔者主张在讲解某些研究方法或算法时，应避免大量的数学推导，而要侧重于方法的应用。例如，在讲解到蛋白质结构预测的方法时，都要讲到神经网络、支持向量机、隐马尔可夫模型等，应该把侧重点放在应用上，对于这些技术不能一味地要求学生自己编程实现，毕竟，这不仅要求对数学原理的深刻理解，而且需要较高的编程技术，这对于刚涉及该领域的学生来说，是不适合的。可以通过第三方软件来使用这些技术，如著名的 Matlab 软件就包含有神经网络工具箱，互联网上也有支持向量机的 Matlab 工具箱可供免费下载，这样可以在较短时间内充分将这些技术应用于生物信息学研究。

但是，对于某些基本的、重要的算法与方法则要求全面掌握，不能囫囵吞枣、知其然而不知其所以然。例如数据库搜索，它的基础是序列的相似性比对，序列比对的基本思想是找出检测序列和目标序列的相似性，它的最终实现，必须依赖于某个数学模型。虽然已经有了专门的序列比对程序（Blast、Clastal W 等），但应该仍然要求学生掌握其原理，并能够编写简单的序列比对的程序。

在进行研究生教学时，应多采用 seminar（课堂讨论会）的形式，而不应是教师的“一言堂”。对于生物信息学这样由多种学科相互交叉融合而产生的新学科，尤其如此。师生在课堂上的讨论有时甚至比教师一人的讲解更有效果，因此要求学生各抒己见，发表对某一个问题的见解，提出自己的问题供大家讨论，是生物信息学教学的一个重要方法。

在生物信息学的教学过程中，也可以多使用一些多媒体课件，特别是对生物学基础较为薄弱的学生。互联网上很多分子生物学方面的课件有助于学生理解诸如 DNA 复制、转录、翻译等基本生物学概念，对教学有很大帮助。

思考与练习

1. 生物信息学的科学基础是什么？
2. 生物信息学的主要研究内容是什么？

参 考 文 献

1　赵国屏，陈润生，罗静初等. 生物信息学. 北京：科学出版社，2002
2　蒋彦，王小行，曹毅等. 基础生物信息学及应用. 北京：清华大学出版社，2003

3 黄韧，薛成，任瑞文等. 生物信息学网络资源与应用. 广州：中山大学出版社，2003
4 Simon H A. 人类的认知：思维的信息加工理论. 北京：科学出版社，1986
5 Baldi P. Brunank S. BIOINFORMATICS: The Machine Learning Approach (生物信息学——机器学习方法). The MIT press, Cambridge, Massachusetts, London, England, 1998
6 Mjolsness E, DeCoste D. Machine Learning for Science: State of the Art and Future Prospects. Science, computer and science, 2001, 9, 293 (14): 2051～2055