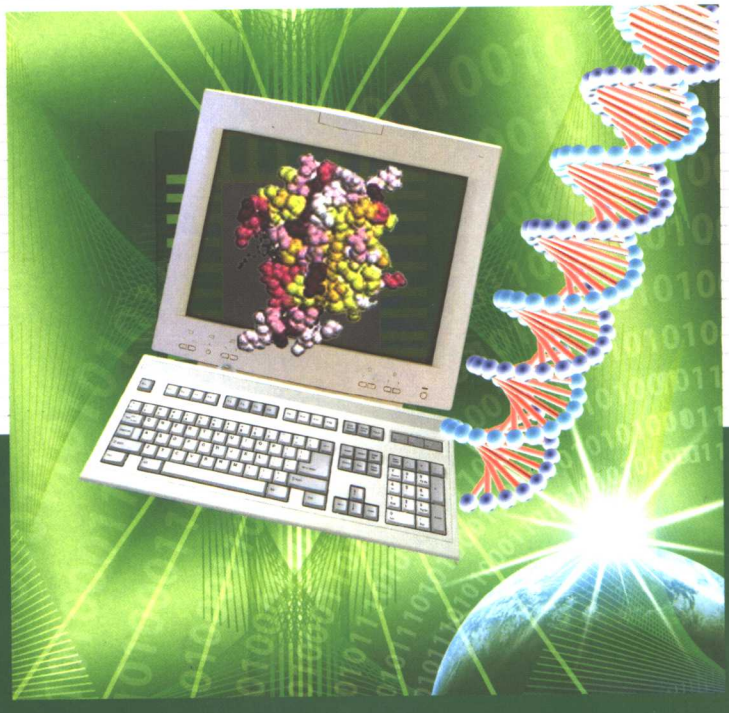
高等学校教材

生物工程 生物技术系列

生物信息学教程

蔡 禄 编著



化学工业出版社

高等学校教材

生物信息学教程

蔡禄 编著

化学工业出版社
·北京·

伴随着人类及其他生物基因组计划的实施，生物信息学迅速成为生命科学的前沿领域。生物信息学是生物学与物理学、化学、数学、信息科学及计算机科学交叉的学科。各种生物信息数据呈指数式增长趋势，运用计算机管理数据、控制误差、加速分析过程势在必行。在此基础上解释实验现象，认识导致实验现象发生的本质，在“整合”、“系统”等全新理念下探索生物学规律，进而了解和掌握生命的物质基础和生命的本质成为当今生物学发展的趋势。

本书从生物信息学研究对象及主要研究内容、生物信息学资源、序列分析、序列比对、系统发生分析、基因组生物信息学、生物芯片、后基因组生物信息学几个方面进行了详细的介绍。

本书内容新颖、简明扼要、内容深浅适度。可作为生物信息学、分子生物学、遗传学、细胞生物学、医学及其他相关专业高年级本科生、研究生的教材，也可供教师和研究人员学习、参阅使用。

图书在版编目（CIP）数据

生物信息学教程/蔡禄编著．—北京：化学工业出版社，2006.12
高等学校教材
ISBN 978-7-5025-9606-4

Ⅰ.生… Ⅱ.蔡… 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 161327 号

责任编辑：赵玉清　　文字编辑：俞方远　周　倜
责任校对：吴　静　　装帧设计：郑小红

出版发行：化学工业出版社（北京市东城区青年湖南街 13 号　邮政编码 100011）
印　　装：化学工业出版社印刷厂
720mm×1000mm　1/16　印张 19¾　字数 412 千字　　2007 年 3 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：30.00 元

序

生物信息学是生物学与物理学、化学、数学及计算机科学交叉的新生边缘学科。伴随着人类及其他生物基因组计划的实施，生物信息数据呈指数式增长趋势；面对着海量的数据，传统生物学的研究方法已经显得不能适应，生物信息学应运而生。其直接目的是解决生物信息的获取、处理、存储、联网、浏览等问题；更深层次的目的则是对大量数据的分析和解释，以及数据背后隐藏的生物学规律的探寻（数据挖掘）。

从伽利略和牛顿开始的近代科学规范的要点有两个：第一是实证性，第二是理性。实证性和理性的结合在近代物理学中已经获得了丰硕的成果，这种结合正在影响着自然科学的其他部门，其中特别重要的是理性化向生命科学渗透。传统的生物学都是实验的，而“正在建立的研究新模式是：由于全部基因将被知晓，储存在电子数据库中，生物学研究的出发点将是理论的。一个科学家将从理论假说出发，然后转向实验，去追随和检验这些假说。”（Gilbert 语）由此可见，把传统的实验方法和定量的逻辑的理性的方法相结合，从而揭示和了解生命规律，把它从一门实验科学提高到综合的理性的水平上来，这是科学历史发展的必然。事实上，从 20 世纪中叶分子生物学诞生以来，生物学理性化的努力就没有中断过。例如，60 年代的量子生物学，70 年代的耗散结构理论，80 年代以模拟个体发育中形态建成问题为中心的数学生物学，90 年代以研究分子序列为中心的计算生物学等。生物信息学是基因组海量数据出现的背景下生物学理性化的一个新阶段。从这个观点来理解和运用生物信息学，可以更好地把握它的关键和发展动向，同时也加深这个学科的理论深度，提高它的预测能力。

生物信息学的迅速发展迫切需要培养新生力量。尽管目前国内已经出版了不少的生物信息学专著或教材，但仍急需出版内容新颖、全面、系统、深浅适度的适合科研人员、研究生和高年级本科生学习使用的教材。本书是作者在多年从事生物信息学研究和教学工作的基础上，参考国内外优秀教材编写而成；内容新颖、全面、系统，既有必备的理论知识和基本原理，又有实用性很强的实践方法，还有生物信息学领域最新进展的介绍，适合科研工作者、研究生和高年级本科生学习。

期望本书的出版能对我国生物信息学及相关学科的发展起到推动作用。

罗云敏

2006 年 9 月于呼和浩特

前　言

如果说物理学是研究物质和能量的学科，那么生命科学就是研究生命物质基础上的信息的学科。随着人类基因组计划的实施，有关核酸、蛋白质的序列和结构数据呈指数增长，运用计算机管理数据、控制误差、加速分析过程势在必行。生物信息学的实质就是利用数理知识、信息和计算机科学及技术来研究生物学信息的组织、传递和表达规律等问题。从中获取基因编码、基因调控、序列-结构-功能关系等理性知识，阐明细胞、器官和个体的发生、发育、病变、衰亡的基本规律和时空联系，探索生命起源、生物进化、生命本质等重大理论问题，最终建立“生物学周期表”。在此基础上解释实验现象，认识导致实验现象发生的本质，在“整合”、“系统”等全新理念下探索生物学规律，进而了解和掌握生命的物质基础和生命的本质。

我国生物信息学研究从20世纪80年代开始起步，目前已经拥有一支数量颇具规模并具有相当水平的研究队伍。但总地来说与发达国家还有较大差距，并且在世界各国加大对生物信息学领域重视的背景下这一差距还有加大的趋势。生物信息学领域的专业工作者的培养相对其他领域较为困难。主要原因是这一交叉学科要求工作者具有诸多领域的知识，既要求有基础的数学、物理学、化学基本训练，全面的生物学知识，又必须具备新兴的信息科学和计算机科学的技能。所以培养生物信息学人才是一漫长而艰巨的任务，从大学学习到研究生培养大概需要7年的时间。发达国家和我国部分高校已在本科和研究生教育水平专门设置了生物信息学专业，大大促进了生物信息学人才的培养。尽管目前国内已经出版了不少生物信息学专著或教材，但由于生物信息学这一新兴学科发展非常迅速，亟待出版内容新颖、全面、系统、深浅适度的适合科研人员、研究生和高年级本科生学习使用的教材。本书是作者在内蒙古科技大学多年教授生物信息学课程讲义的基础上，历经一年多认真修改、补充完成的。本书在编写过程中参考了国内外优秀的专著或教材，主要有：孙啸，陆祖宏和谢建明（2005）的《生物信息学基础》；钟扬，张亮和赵琼（2001）的《简明生物信息学》；R. Durbin，S. Eddy，A. Krogh和G. Mitchison（1998）的Biological Sequence Analysis；T. K. Attwood和D. J. Parry-Smith（2002）的《生物信息学概论》（罗静初等译）；张成刚和贺福初（2002）的《生物信息学》；郝伯林和张淑誉（2000）的《生物信息学》；李巍（2004）的《生物信息学导论》；伍欣星，赵旻和罗晓忠（2005）的《生物信息学——基础与临床医学应用指南》。

本书共分9章，第1章介绍生物信息学的基本概念、研究对象、基本方法、发展历史和前沿技术；第2章概括总结了生物信息学学习必备的生物学知识；第3章详细介绍了生物信息学资源与数据挖掘工具；第4章介绍了DNA和蛋白质序列分

析的基本方法和软件；第5章介绍了生物信息学中常用的序列比对方法；第6章重点介绍生物信息学中最成熟的分子系统发生分析；第7章专门介绍基因组生物信息学内容；第8章介绍生物芯片这一最新技术；第9章以专题的形式介绍了生物信息学的最新领域——后基因组时代的生物信息学的基本概念、内容的发展动态。

感谢化学工业出版社的支持与帮助，感谢内蒙古科技大学教材建设基金的支持。在完成书稿之际，对父母及家人多年无私支持表示由衷的谢意。

生物信息学内容新、发展快、覆盖学科广，由于作者知识水平所限，难以对每一部分内容都有非常深刻的理解，加之编写时间仓促，书中难免有错误和疏漏之处，欢迎大家提出宝贵意见。

编　者

2006年9月于包头

目　录

第1章　生物信息学引论

本章提要：本章旨在介绍生物信息学的基本概念，指出生物信息学的研究目标和任务、研究意义、基本方法和前沿技术。简要回顾了生物信息学的产生和发展历史，基因组计划对生物信息学产生的巨大挑战，较为详细地介绍了目前阶段其主要研究内容，并列举了生物信息学的应用领域。

生命科学在20世纪得到了快速发展，在还原论思想的引导下，生理学、细胞生物学、分子遗传学、分子生物学等学科的发展使人们从器官、组织、细胞及生物大分子等各个层次认识了生命的物质基础。生物与其他物质有本质的区别，生物并非只是物质的简单堆积，生物体的生长发育是生命信息控制之下的复杂而有序的过程。如果说物理学是研究物质和能量的学科，那么生命科学就是研究生命物质基础上的信息的学科。21世纪是生命科学的时代，也是信息时代。随着人类基因组计划的实施，有关核酸、蛋白质的序列和结构数据呈指数式增长。面对巨大而复杂的数据，运用计算机管理数据、控制误差、加速分析过程势在必行。目前，对生命的奥秘还不甚了解，对生命信息的组织、传递和表达还知之甚少。既然这牵涉到信息的组织、传递和表达，就可以用信息科学的方法和技术来尝试认识和分析生命信息。在这样一个背景下，生物信息学作为一门学科应运而生并且得到了迅速发展。

1.1　引　　言

传统的生物学是一门实验科学，生物学研究依赖于对实验数据的处理和分析。生物学也是一门发现科学，通过实验发现新的现象、新的生物学规律，经过分析、归纳和总结，提炼出新的生物学知识。传统的还原论生物学研究方法在20世纪取得了重大成就，特别是分子生物学的出现。在21世纪的头几年生物学发生了重大的变化，传统的生物学研究模式受到了极大的挑战。随着基因组计划的迅速发展，生物数据的积累速度不断加快。因此，也就对生物数据的科学分析方法和实用分析工具提出了更新、更高的要求。在这个过程中，需要对实验数据进行处理并及时进行理论分析，在此基础上解释实验现象，认识导致实验现象发生的本质，在“整合”、“系统”等全新理念下探索固有的生物学规律，进而了解和掌握生命的物质基础和生命的本质。

1.1.1　生物信息学基本概念

无论从理论上来讲还是从实际情况来看，生物信息学的实质就是利用数理知识、信息和计算机科学及技术来研究生物学信息的组织、传递和表达规律等问题。生物信息学的诞生是由生物学对大量数据处理和分析的需求而引发的，是历史的必

然。作为一门交叉学科，生物信息学的发展依赖于计算机科学技术和生物技术的发展，而生物信息学的研究成果又促进了生物学特别是分子生物学的发展。

生物信息学（bioinformatics）有许多不同的定义。基于生物信息学与分子生物学的密切关系，狭义的生物信息学专指应用信息技术储存和分析基因组测序所产生的分子序列及其相关数据，也被称为分子生物信息学。

广义的生物信息学是指以核酸、蛋白质等生物大分子为主要研究对象，以信息、数理、计算机科学为主要研究手段，以计算机网络为主要研究环境，以计算机软件为主要研究工具，对序列数据进行储存、管理、注释、加工，对各种数据库进行查询、搜索、比较、分析，构建各种类型的专用数据库信息系统，研究开发面向生物学家的新一代计算机软件；并利用数理统计、模式识别、动态规划、密码解读、语意解析、信令传递、神经网络、遗传算法以及隐马氏模型等各种方法，对序列、结构数据进行定性和定量分析，从中获取基因编码、基因调控、序列-结构-功能关系等理性知识，阐明细胞、器官和个体的发生、发育、病变、衰亡的基本规律和时空联系，探索生命起源、生物进化、生命本质等重大理论问题，最终建立“生物学周期表”。

与生物信息学相关的概念还有计算分子生物学（computational molecular biology），它常被看作是生物信息学的同义词。两者确实十分相近，尤其是它们都将分子生物学数据分析作为主要研究内容。但一般认为，计算分子生物学主要研究分析方法，开发分析工具，促进生物分子数据的分析。计算生物学更侧重于发展理论模型和计算方法，应用领域则不如生物信息学覆盖面广。与生物信息学相近的另一个名词是生物计算，生物计算主要是用计算机技术分析和处理生物学数据。

总地来说，生物信息学中许多分支学科源于生物学的不同分支学科与信息科学的结合。例如，计算分子生物学和计算神经生物学（computational neurobiology）等，从名称上即可大致反映其内容。不同的研究单位和研究者一般依自己工作的重点来使用这些分支学科的名称，或采用类似的名称，如日本国立遗传研究所（NIG）中著名的“信息生物学中心（Center for Information Biology）”。此外，一批生物信息学的姊妹学科也已形成，如医学信息学（medical informatics）、化学信息学（chemical informatics）等。

1.1.2 生物信息学的研究目标和任务

揭示生物分子数据隐含的生物学信息是其长远目标和根本任务。生物分子数据之间存在着复杂的联系，这些数据中蕴涵着丰富的生物学知识和生物学规律。生物信息学的发展将揭示生物分子信息的本质，使人类彻底了解、掌握遗传信息的编码、传递及表达，从而加快人类了解自身的进程。

目前生物信息学的主要任务是研究生物分子数据的获取、存储和查询，发展数据分析方法。主要包括3个方面。

第一是收集和管理生物分子数据，使得生物学研究人员能够方便地使用这些数据，并为信息分析和数据挖掘打下基础。生物分子数据来自于生物学实验，应用信息学技术收集和管理这些数据，将各种数据以一定的表示形式存放在计算机中，建

立数据库系统，并提供数据查询、搜索和数据通信工具。

第二是进行数据处理和分析。通过数据分析，发现数据之间的关系，认识数据的本质，进而上升为生物学知识。并在此基础上，解释与生物分子信息复制、传递和表达有关的生物过程，解释在生物过程中出现的信息变化与疾病的关系，帮助发现新的药物作用目标，设计新的药物分子，为进一步的研究和应用打下基础。目前生物信息学的主要研究对象是 DNA 和蛋白质。在 DNA 分析方面，着重分析 DNA 序列中的基因信息及基因表达调控信息，分析基因表达数据，分析基因之间的相互作用关系，比较不同种属的基因组，研究基因组中非编码区域的生物学功能。在蛋白质分析方面，着重分析蛋白质序列与蛋白质结构及功能之间的关系，预测蛋白质的结构和功能，研究蛋白质的进化关系。

生物信息学研究的第三个方面是开发分析工具和实用软件，解决具体的问题，为具体的生物信息学应用服务。例如，开发生物分子序列比较工具、基因识别工具、生物分子结构预测工具、基因表达数据分析工具等。

生物分子数据类型的不断增多及数据量的不断膨胀促进了生物信息学的研究与应用。生物信息学的研究成果不断涌现，各种生物信息源如雨后春笋，层出不穷，而各种生物信息分析算法和工具也日益更新。

掌握互联网上各种生物信息学数据库以及相关软件的使用技术已成为生物学和医学研究人员的迫切需要。尤其是分子生物学的三大核心数据库 GenBank 核酸序列数据库、SWISS-PROT 蛋白质序列数据库和 PDB 生物大分子结构数据库，不仅是全世界分子生物学和医学研究人员获取生物分子序列、结构和其他信息的基本来源，而且是发表自己序列或结构测定结果的重要媒体。围绕这三大核心数据库还有众多面向各种特定应用的衍生数据库和分析软件，这些数据库分别从不同角度、以不同方式对各类生物信息学数据进行归纳、总结和注释，而各种分析软件为挖掘这些数据提供了有力的工具。

1.1.3　生物信息学的研究意义

生物信息学研究是从理论上认识生物本质的必要途径，通过生物信息学研究和探索，可以更为全面和深刻地认识生物科学中的本质问题，了解生物分子信息的组织和结构，破译基因组信息，阐明生物信息之间的关系。基因序列到蛋白质序列的三联密码关系是众所周知的，也是非常简单、非常确定的。然而，基因调控序列与基因表达之间的关系、蛋白质序列与蛋白质结构之间的关系则是未知的，也一定是非常复杂的。破译和阐明生物信息的本质将使得人类对生物界的认识跨越一个新台阶。

生物信息学的出现将改变生物学的研究方式。传统的生物学是一门实验科学，传统的分子生物学实验往往是集中精力研究一个基因、一条代谢路径，手工分析完全能够胜任。然而，随着分子生物学技术的发展，已经出现一些高通量的实验方法，如基因芯片，利用基因芯片一次可以获取上千个基因的表达数据。生物学已经从一次只分析一个生物分子的时代跳跃到同时分析成千上万个生物分子的时代。对于高通量的实验结果，必须利用计算机进行自动分析。因而，在高通量实验技术出

现的时代，生物信息学必然要介入生物学研究和实验。

再者，从生物分子数据本身来看，各种数据之间存在着密切的关系，如 DNA 序列与蛋白质序列、基因突变与疾病等，这些关系反映了生物学的规律。但是，这些关系可能是非常复杂的，是未知的，是简单的多元统计方法难以分析的。对于这些复杂的关系，必须运用现代信息学的方法去分析，去研究。因而，随着分子生物学研究的深入，必然需要生物信息学。

另外，现在全世界每天都会产生大量的核酸和蛋白质序列，不可能用实验的方法去详细研究每一条序列，必须首先进行信息处理和分析，去粗取精，去伪存真。通过预处理，发现有用的线索。在此基础上进行有针对性、有明确目的的分子生物学实验。因而，生物信息学在指导实验、精心设计实验方面将会发挥重要的作用。

生物信息学研究在医学上也有重要的意义。通过生物信息学分析，可以了解基因与疾病的关系，了解疾病产生的机理，为疾病的诊断和治疗提供依据。研究生物分子结构与功能的关系将是研制新药的基础，可以帮助确定新药作用的目标和作用的方式，从而为设计新药提供依据，揭示人类及重要动植物种类的基因的信息，继而开展生物大分子结构模拟和药物设计，人们甚至期望在单核苷酸多态性（SNP）研究的基础上开发针对个体或某一群体的药物。

1.2 生物信息学的产生与发展

1.2.1 生物信息学的发展历史

生物信息学的发展大致经历了 3 个阶段。

（1）前基因组时代（20 世纪 90 年代前） 早在 20 世纪 50 年代，生物信息学就已经开始孕育。1956 年在美国田纳西州的 Gatlinburg 召开了首次“生物学中的信息理论研讨会”。在 20 世纪 60 年代，一些计算生物学家开始进行相关研究，做了许多生物数据搜集和分析方面的工作。在这个时期，生物大分子携带信息成为分子生物学的重要理论，生物分子信息在概念上将计算生物学和计算机科学联系起来。大量的生物分子序列成为丰富的信息源，科学家们开始应用计算方法分析这些信息。相关或者同源蛋白质序列之间的相似性首先引起了人们的注意。1962 年，Zucherkandl 和 Pauling 研究序列变化与进化之间的关系，开创了一个新的领域——分子进化。随后，通过序列比较确定序列的功能及序列分类关系成为序列分析的主要工作。1964 年，蛋白质结构预测的研究由 Davies 的工作开始。氨基酸序列的收集是这个时期的一项重要工作，1967 年 Dayhoff 发表了蛋白质序列图集，该图集后来演变为著名的蛋白质信息源（PIR）。

20 世纪 60 年代是生物信息学形成雏形的阶段。一般认为，生物信息学的真正开端是 20 世纪 70 年代。从 70 年代到 80 年代初期，随着生物化学技术的发展，产生出许多生物分子序列数据，而在这个阶段数学统计方法和计算机技术都得到较快的发展，于是促使一部分计算机科学家应用计算机技术解决生物学问题，特别是与生物分子序列相关的问题。他们开始研究生物分子序列，研究如何根据序列推测结

构和功能。这时，生物信息学开始崭露头角。

从 20 世纪 70 年代初期到 80 年代初期，出现了一系列著名的序列比较方法。其中，Needleman 和 Wunsch 于 1970 年提出的序列比对算法是对生物信息学发展最重要的贡献。同年，Gibbs 和 McIntyre 发表的矩阵打点作图法也是进行序列比较的一个著名方法，该方法可用于寻找序列中的重复片段，从而推测其功能。Dayhoff 提出的基于点突变模型的 PAM 矩阵是第一个广泛使用的比较氨基酸相似性的得分矩阵，它大大地提高了序列比较算法的性能。1981 年，Smith 和 Waterman 提出了著名的公共子序列识别算法，同年 Doolittle 提出关于序列模体的概念。1983 年，Wilbur 和 Lipman 发表了数据库相似序列搜索算法。1985 年，出现了快速的蛋白质序列搜索算法 FASTP/FASTN。1988 年，Pearson 和 Lipman 发表了著名的序列比较算法 FASTA。1990 年，快速相似序列搜索算法 BLAST 问世。1997 年，BLAST 的改进版本 PSI-BLAST 投入实际应用。

在 20 世纪 70 年代，还不断涌现出许多生物信息分析方法。1972 年，Gatlin 将信息论引入序列分析。证实自然的生物分子序列是高度非随机的。1975 年，继第一批 RNA（tRNA）序列的发表之后，Pípas 和 McMahon 首先提出运用计算机技术预测 RNA 二级结构。1977 年，出现了将 DNA 序列翻译成蛋白质序列的算法。1978 年，核酸序列数据库出现，收录有发表的 5S 和 5.8S 核糖体 RNA 序列，Gingeras 等人研制出核酸序列中限制性酶切位点的识别软件。

20 世纪 80 年代以后，出现了一批生物信息服务机构和生物信息数据库。1982 年，核酸数据库 GenBank 第 3 版公开发行。1986 年，日本核酸序列数据库 DDBJ 诞生。1986 年，出现蛋白质数据库 SWISS-PROT。1988 年，美国国家卫生研究所和美国国家图书馆成立国家生物技术信息中心（NCBI）。同年，成立欧洲分子生物学网络（EMRnet），该网络专门发布各种生物数据库。

这一时期陆续出现了生物信息学相关的专著、刊物和关键性论文。1958 年，由 H. P. Yockey 编辑的《生物学中的信息理论讨论会》由纽约 Pergamon 出版社出版。1970 年，期刊 Computer Methods and Programs in Biomedicine 诞生。Science 期刊于 1980 年第 209 卷发表了 Gingeras 和 Roberts 关于计算分子生物学的综述：Steps towards a programmed analysis of nucleic acid sequences。1985 年，生物信息学专业期刊——Computer Application in the Biosciences 创刊。

（2）基因组时代（20 世纪 90 年代后至 2001 年）　生物信息学的真正发展则是在 20 世纪 90 年代，在人类基因组计划的推动下，生物信息学才得以迅猛发展。人类基因组计划产生的生物分子数据是生物信息学的源泉，而人类基因组计划所需要解决的问题则是生物信息学发展的动力。标志性工作包括基因寻找和识别，网络数据库系统的建立和交互界面的开发等。例如，建立与发展表达序列标签（expressed sequence tag，EST）数据库以及电子克隆（virtual cloning）技术等。

在 20 世纪 80 年代后，科学家们开始大规模的基因组研究。

1986 年，出现基因组学（genomics）概念，即研究基因组的作图、测序和分析。

1990 年，第一届国际电泳、超级计算和人类基因组会议在美国佛罗里达州会议中心举行，尽管会议的名称并没有出现生物信息学这一名词，实际上生物信息学却是会议的主要部分。国际人类基因组计划启动，被誉为生命科学的“阿波罗登月计划”。

1993 年，成立 Sanger 中心，该中心专门从事基因组研究。欧洲生物信息学研究所（EBI）获准成立。专业蛋白质分析系统网络服务器诞生。第一届 ISMB（Intelligent Systems for Molecular Biology）国际会议在 Bethesda 召开，会议一年一次，2006 年已是第 14 届。

1994 年，国际生物信息学系列会议由 Cambridge Healthtech 研究所接管，并走向商业化和联机化。澳大利亚 Macquarie 大学的 Marc Wilkins 和 Keith Williams 首先提出蛋白质组（proteome）的概念。第三届国际生物信息学和基因组研究会议在佛罗里达州会议中心举行。

1995 年，第一个细菌基因组被完全测序。

1996 年，酵母基因组被完全测序。

1996 年，Affymetrix 生产出第一块 DNA 芯片。

1997 年，Prusiner 因发现引发疯牛病的朊病毒而获得诺贝尔生理/医学奖。

1998 年，亚太生物信息学网络（APBioNet）成立。人类完成第一个多细胞生物——线虫的基因组全序列测定。生物信息学专业期刊——Comput. Appl. Biosci. 更名为 Bioinformatics。瑞士生物信息学研究所（SIB）成立。美国塞莱拉遗传公司成立，目标是到 2001 年绘制出完整的人体基因图谱，与国际人类基因组计划展开竞争。

1999 年，果蝇的基因组被完全测序。1999 年底，国际人类基因组计划联合研究小组宣布人类第一次获得对完整的人类染色体——第 22 号染色体的遗传序列。

2000 年 3 月 14 日，美国总统克林顿和英国首相布莱尔针对某些私营生物技术公司为商业利益而试图为自己的研究成果申请专利而发表联合声明，呼吁公开人类基因组研究成果。2000 年 5 月 8 日德国、日本等国科学家宣布，他们已基本完成人体第 21 对染色体的测序工作。2000 年 6 月 24 日，人类基因组计划协作组的 6 个国家研究机构在全球同一时间宣布已完成人类基因组的工作框架图。2000 年 12 月 14 日，美国、英国等国科学家宣布绘出拟南芥基因组的完整图谱，这是人类首次全部破译出一种植物的基因序列。

2001 年 2 月 12 日，中国、美国、日本、德国、法国、英国 6 国科学家和美国塞莱拉遗传公司联合公布人类基因组图谱及初步分析结果。

在此期间，生物信息学在人类基因组计划的促动之下迅速发展。

（3）后基因组时代（2001 年至今） 随着后基因组时代的到来，生物信息学研究的重点逐步转移到功能基因组信息研究，其研究的内容不仅包括基因的查询和同源性分析，而且进一步发展到基因和基因组的功能分析，即所谓的功能基因组学研究。其具体表现在：①将已知基因的序列与功能联系在一起进行研究；②从以常规克隆为基础的基因分离转向以序列分析和功能分析为基础的基因分离；③从单个基

因致病机理的研究转向多个基因致病机理的研究；④从组织与组织之间的比较来研究功能基因组和蛋白质组。组织与组织之间的比较主要表现在：正常与疾病组织之间的比较，正常与激活组织之间的比较，疾病与处理（或治疗）组织之间的比较，不同发育过程的比较等。标志是大规模基因组分析、蛋白质组分析以及各种数据的比较和整合。出现了蛋白质组学、药物基因组学、比较基因组学、功能基因组学、系统生物学、整合生物学等学科。研究思路也发生了本质的变化，从传统的还原论研究生命过程转到了综合论思想。综合论方法研究基因和各种生物大分子是怎样通过网络调控方式形成一个生物系统的。提出了层次抽提和相互作用网络等概念。继基因组概念之后，人们开始关注转录组（transcriptome）、蛋白质组（proteome）、相互作用组（interactome）、定位组（localizome）、折叠子组（foldome）、代谢组（metabolome）和表型组（phenome）等。

1.2.2　我国生物信息学发展现状

我国的生物信息学工作是逐步发展起来的。20 世纪 80 年代就有若干科研院所的生物、物理、信息、数学等学科的工作者从事生物信息学的研究工作。国内近年来开展生物信息学研究的单位主要有：北京大学、清华大学、中国科学院生物物理研究所、军事医学科学院、上海生命科学研究院、中国科学院生物化学研究所、中国科学院微生物研究所、中国科学院遗传所人类基因组中心、中国医学科学院、天津大学、内蒙古大学、复旦大学、南开大学、中国科技大学、东南大学等。开展工作如核酸序列统计分析、生物大分子二级结构预测、分子动力学等。我国虽然早在 1993 年就在中国人类基因组计划中列入了生物信息学的相关研究内容，但真正开始发展是在 1995～1996 年。目前我国生物信息数据源和分析软件多半来自于国外，依靠国外生物信息中心建立中国数据镜像中心，中国生物信息学的基础力量还比较薄弱。由于技术、人才、资金等多方面的原因，其研究水平与国际同行尚有较大差距。我国尚处于引进国外已有数据库，为国内研究人员提供服务的阶段。

近几年来，国内对生物信息学的研究和应用越来越重视。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。在一些著名院士和教授的带领下，我国的生物信息研究和应用在一些领域取得了一定成绩，有的在国际上还占有一席之地。我国在生物信息学领域取得了一些比较好的研究成果，特别是在基因预测算法、基因组信息分析、蛋白质分子设计、分子动力学等方面。从国内生物信息学研究与应用的整体情况来看，仍然与国际先进水平有较大的差距。

我国在基因组信息的收集与发布方面开展了一些工作，如北京大学生物信息学中心建立的生物信息学服务器和 EMBL 数据库的中国节点，已经成为国内最重要的生物信息学资源，为我国及世界各地的科学家提供生物信息查询、软件工具使用、文献查阅等多种服务。中国承担并顺利完成了人类基因组计划 1%的测序任务，测序技术取得了很大的进展，但是在生物信息分析、基因功能分析等方面的工

作还没有及时跟上。国内生物医学研究与开发对生物信息学的需求市场非常广阔，然而真正开展生物信息学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。与发达国家相比，在人力和财力投入上明显不足。目前，我国基因组和蛋白质组研究在国际上已经占据了重要的地位；在生物信息学研究和应用方面，相信经过科学家的努力，经过多学科专家的合作，完全有可能赶上甚至超过世界先进水平。

1.2.3 我国生物信息学研究的发展方向

从国内权威的政府科学研究基金“国家自然科学基金”的资助方向可大致了解我国生物信息学研究的主流发展方向。国家自然科学委员会数理科学部设立了一个“理论物理学及其交叉科学若干前沿问题”的重大项目，其科学目标是：围绕生物大分子理论及生物信息学中关键问题，在 DNA 链复杂性、基因组序列信息分析、编码区和非编码区的统计分析、基因组全信息的生物进化等方面提出新理论、建立新方法；开展多重时空尺度上的生物大分子和生物凝聚体的结构、相互作用、性质及其调控理论的创新研究。主要资助方向有以下两个方面。①生物信息学研究：基因识别（包括编码区和启动子区域识别）的新方法；分析多个基因组新方法并应用于分子进化；基因网络与系统生物学研究。②计算分子生物学与计算细胞生物学研究：单分子生物物理理论；蛋白质二级、三级结构预测新方法；生物大分子的自组装（如生物膜、肌纤、蛋白微管等）理论等。国家自然科学委员会数理科学部还设立了重点项目“基因功能预测的生物信息学”，项目强调发展物理与生物、化学、数学结合的新实验和理论方法来探索生物系统调控的基本规律。基于生物信息学的观点来揭示蛋白质序列，结构与功能的关系，蛋白质之间的相互作用及网络，基因表达调控网络的性质及关键环节等。生命科学部的“生物化学与分子生物学学科”、“遗传学与发育生物学学科”、“生物物理与生物医学工程学科”，信息科学部的“电子科学学科”均把“生物信息学”方向作为重要的前沿领域资助。

1997 年专门召开了香山会议，专题讨论我国生物信息学的发展。1999 年 4 月，国家自然科学委员会生命科学部、信息科学部、数理科学部、材料科学部在北京召开“生命科学中的信息科学问题”论坛。研讨主题主要集中在微观尺度上的基因组和蛋白质组信息学及宏观尺度上的信息生态学和信息农学。关于基因组和蛋白质组信息学当前的研究任务，会议一致认为应该建立国家生物医学数据库与服务系统，同时开展基因组及功能基因组信息分析工作，发现新基因和新单核苷酸多态性（single nucleotide polymorphisms，SNP）以及各种功能位点，发展大规模基因表达谱分析算法，研究基因表达调控网络，进行核酸、蛋白质空间结构的预测和模拟，研究蛋白质功能预测方法，开展遗传密码起源和生物进化的研究，建立生物信息学的新理论、新方法、新技术和新软件。目前我国生物信息学研究的主要方向如下。

（1）建立国家生物医学数据库与服务系统　已有专家建议在我国尽快建立国家级的“生物医学信息中心”，其首要任务是从国际上引进生物医学数据库和免费共

享软件，同时把我国在生物信息方面有特色的成果提供给国际科学界。需要开发适合我国用户的接口和界面系统，同时开展数据库管理、模型和算法等方面的研究以及教育培训等工作。

（2）人类基因组的信息结构分析　利用 EST 数据库（如 dbEST）并采用大规模并行计算，发现新的基因和单核苷酸多态性（SNP）以及各种功能位点；进行模式生物完整基因组的信息结构分析和比较研究。例如，对酵母（微生物）、线虫（动物）和拟南芥（植物）等模式生物进行比较基因组学研究。

（3）功能基因组相关信息分析　研究开发大规模基因表达谱分析相关的算法与软件，特别是研究基因表达调控网络；预测和模拟与基因组信息相关的核酸、蛋白质空间结构，进而预测蛋白质功能。

（4）遗传密码起源与生物进化（尤其是分子进化）的过程与机制　早期的工作主要是利用不同物种中同一种基因序列的异同来研究生物的进化，构建进化树。既可以用 DNA 序列也可以用其编码的氨基酸序列来做，甚至可通过相关蛋白质的结构比对来研究分子进化。以上研究已经积累了大量的工作。近年来由于较多模式生物基因组测序任务的完成，为从整个基因组的角度来研究分子进化提供了条件。可以设想，比较两个或多个完整基因组这一工作需要新的思路和方法，当然也渴望得到更丰硕的成果，这方面可做的工作是很多的。

（5）非编码区分析和 DNA 语言研究，是最重要的课题之一　研究占人类基因组 95%的非编码区的信息结构，建立理论模型以阐明非编码区的重要生物学功能；在人类基因组中，编码部分占总序列的 3%～5%，其他通常称为“垃圾”DNA。其实一点也不是垃圾，只是暂时还不知道其重要的功能。分析非编码区 DNA 序列需要大胆的想象及崭新的研究思路和方法。DNA 序列作为一种遗传语言，不仅体现在编码序列之中，而且隐含在非编码序列之中。

（6）基于结构的药物设计　人类基因组计划的目的之一在于阐明人的约 10 万种蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系，寻求各种治疗和预防方法，包括药物治疗。基于生物大分子结构的药物设计是生物信息学中的极为重要的研究领域。为了抑制某些酶或蛋白质的活性，在已知其三级结构的基础上，可以利用分子对接算法，在计算机上设计抑制剂分子，作为候选药物。这种发现新药物的方法有强大的生命力，也有着巨大的经济效益。

此外，结合重大科学问题的研究，发挥我国在理论生物学和信息科学领域的研究特色，发展生物信息学的新理论、新方法、新技术和新软件也是重要的发展方向之一。

在我国，有关生物信息学的研究已逐渐引起大家的重视，例如，在“HGP 1%的测序工作”、“中华民族基因组中若干位点基因结构的研究”和“重大疾病相关基因的定位、克隆、结构与功能研究”等项目中，生物信息学分析均发挥了重要作用。如何进一步根据我国在生物学方面的特点，建立高水平的理论与实验体系，加快培养优秀的青年人才，是发展我国生物信息学研究最为迫切的任务之一。