生物信息学

蔡 禄 主编

Bioinformatics

科学出版社

生物信息学

主编　蔡　禄

参编　邢永强　崔向军　刘国庆　崔大超

科学出版社

内 容 简 介

本书首先介绍生物信息学的基本概念、产生与发展及主要研究内容，安排了生物学基础、统计学习与推断两章内容供读者选学；然后依次介绍生物信息学资源、序列分析与序列比对、分子系统发生分析等基本内容；接下来学习基因组信息学、生物芯片、转录组信息学、蛋白质组信息学等前沿内容；最后一章介绍系统生物学这一最新领域的有关内容。

本书内容新颖、简明扼要、深浅适度，可作为生物信息学、医学及其他生命科学相关专业本科高年级学生、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 / 蔡禄主编. —北京：科学出版社，2017.11
ISBN 978-7-03-055335-5

Ⅰ. ①生… Ⅱ. ①蔡… Ⅲ. ①生物信息论 Ⅳ. ① Q811.4

中国版本图书馆CIP数据核字（2017）第276483号

责任编辑：刘 丹 赵晓静 / 责任校对：郑金红
责任印制：吴兆东 / 封面设计：迷底书装

科 学 出 版 社出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 11 月第 一 版 开本：850×1168 1/16
2018 年 1 月第二次印刷 印张：19 1/4
字数：619 000

定价：59.00 元

（如有印装质量问题，我社负责调换）

序

信息是独立于物质和能量的第三范畴，生命始于信息。薛定谔（E. Schrodinger）在《什么是生命》中从德布吕克（M. Delbruck）噬菌体实验结果的推测——基因包含于微观体积（边长为 10 个原子距离的立方体）中——出发，建议将大分子作为一种非周期固体，可作为遗传信息的载体。薛定谔是第一个认识到信息对于生命的意义并提出遗传密码概念的人。20 世纪 50 年代，分子生物学诞生后，人们逐步确立了“分子序列-结构-功能”的生命逻辑。正是由于这条普适的铁的法则的存在，生物信息学才不仅是一个工作平台，还是一个独立的学科。

20 世纪 70 年代，戴霍夫（M. O. Dayhoff）对蛋白质序列的分析和卡巴特（E. A. Kabat）对抗体序列的分析开启了生物信息学研究的先河［后来人们称 Dayhoff 为生物信息学之父（father of bioinformatics）］。90 年代，伴随着人类基因组计划的开展，沃特曼（M. S. Waterman）提出计算生物学（computational biology）概念。进入 21 世纪，生物信息学 / 计算生物学已从理论生物学中分化出来，发展成为独立的分支学科。

当前，科学家已有很多机会可以在生物技术、大型制药行业中接触到生物信息学与大数据的工作。美国 *Science* 周刊上以“生物信息学，神秘的新职业”为题描述了这一盛况。“现在，更简单划算的工具促生了更多的数据，因此就更需要有专家能够以一种方式理清堆积成山的信息，让其对科学家和临床医生具有意义，并最终惠及客户和患者”“生物信息学的工作机会与以往相比有 100% 的增长”。

生物信息学已经获得十分广泛的应用，而且前途一片光明。有人估计，基因测序和分子诊断的全球市场价值每年都已达数十亿以至百亿美元以上，并且保持继续增长的势头。生物信息学在应用领域如此成功源于它有深厚的基础研究支持。关于基础科学与应用科学的关系，因发现调控血液和细胞内胆固醇代谢的受体而获得 1985 年的诺贝尔生理学或医学奖的戈尔斯坦（J. Goldstein）在论述中强调了基础研究的重要性，他列举了他的 9 位在美国国立卫生研究院学医的学生做基础研究从而改变了美国医疗制药史的过程。

“科学之父”泰勒斯因为对科学和哲学的全心追求，生活拮据，商人嘲笑他，后来他用天文学知识成功预测了橄榄丰收并大赚一笔。我们不是没有能力赚钱，只是我们有更有趣、更重要的事情要做而已。法拉第回答英国财政大臣的责问时说，“您会从今天看似无用的实验中大收其税的”。无数事实都说明，“今天的基础，明天的应用”；并且从基础研究到应用的转化时间愈来愈快，从 19～20 世纪的数十年至上百年到现在的几年至十年。

基础研究也是在不断发展的。从历史看，理论生物学曾是新兴学科生物信息学的孵化器。回顾过去，展望未来。当我们试图了解生物信息学基础研究的未来发展时，仍可从理论生物学的视角获得启发。例如，生命作为 DNA、RNA、蛋白质三者互作中形成的信息网络，基因表达调节三元网络比二元网络复杂得多，它具有哪些基本特征？信息传输必须具有精确性，传输系统必须具有容错性和耐攻击性。生命是如何适当安排冗余和选择网络的拓扑结构来实现这些目标的？表观遗传学提供了一条新的非 DNA 来源的信息传递途径，染色体中相关蛋白质是如何实现自我永生的？遗传的高度稳定性必须有一个物理原理来保证，核酸-蛋白质是一种特殊的生物凝聚态，它的量子特性可能是遗传稳定性的关键因子。

基因组作为一类信息系统，它的遗传和演化规律是怎样的？长期以来，进化的渐进论和间断平衡论争论不休，进化机制尚不清楚。目前，测序积累的海量数据和进化树的成功构建已经为发现这个系统的演化规律提供了资料和观念的准备。关键可能在于找出适当的动力学变量并建立其时间演化的动力学方程。我们希望通过了解基因组进化方程了解物种进化的方向性、信息的早期快速积累等实验事实，特别是了解新物种产生的突发性、随机性及其他相关规律。从初步研究结果看来，基因组的信息演化方程有可能建立，并且和物理学的物质能量演化方程类似，同样可以量子化。在 DNA 水平上新物种产生是否遵循超越经典轨道的量子规律？这个基本问题将有可能被解决。

上面只是抛砖引玉，列举了一些生物信息学的基础理论问题。2007 年图灵奖得主吉姆·格雷（Jim

Gray）的演讲“科学方法的革命”中提出将科学研究分为四类范式：实验归纳、模型推演、仿真模拟和数据密集型科学发现（data-intensive scientific discovery）。最后的“数据密集型”，就是后来人们所说的“科学大数据”。数据催生科学创新，19 世纪末 20 世纪初的原子光谱十万条谱线产生量子力学，就是一个典型例子。当时还没有计算机。今天的生物信息，地球上万千物种的生老病死，其数据量已远远超过当年的原子光谱。这海量密集数据，加上计算机方法的融合，是一个可以催生科学发现的大舞台。

蔡禄教授是国内最早进入生物信息学领域的研究者之一。近年来，他的团队进行了卓有成效的理论和实验相结合的生物信息学研究。生物信息学是高度综合性学科，需要多方面基础，包括理论的和实验的、生物学的、计算机科学的，以及数学的、物理学的，该书是蔡禄团队为大学生和研究生编写的入门书，也是为研究人员编写的参考手册，内容涵盖面广，材料组织精当，密切联系学科前沿，叙述讲解清楚，逻辑性和实用性并重，是不可多得的好书。希望该书能成为从事生物信息学应用研究的青年学生的引导者和助手，也希望它能启发和帮助他们去关注生物信息学的基础科学问题。

2017 年 9 月于内蒙古科技大学，鹿城

前　言

生物信息学的发展最早可追溯到20世纪50年代，20世纪下半叶，研究人员在数据库构建及检索、序列比对、基因识别、分子进化等领域做了大量工作。20世纪末，受人类基因组计划的直接推动，生物信息学进入了发展的快车道。进入21世纪后，各种高通量“组学”技术的发展加快了数据产生的速度和复杂度，各类算法大量涌现，生物信息学进入了全面开花、迅速发展的全新阶段。生物信息学分析已经成为实验生物学工作者的必备技能。

生物信息学实质是利用数理知识、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、生物进化等重大理论问题。在“整合”“系统”等全新理念下探索生物学规律，进而理解生命的本质。

我国生物信息学研究从20世纪80年代起步，目前已经拥有一支颇具规模，并具有相当水平的研究队伍。生物信息学领域专业工作者的培养相对其他领域较为艰难。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。培养生物信息学人才是一项漫长而艰巨的任务，从大学学习到研究生培养需要至少7年的周期。发达国家和我国部分高校已在本科生和研究生教育水平专门设置了生物信息学专业，极大地促进了生物信息学人才的培养。

尽管目前国内已经编写或翻译了为数不少的生物信息学专著或教材，但由于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，适合科研人员、研究生和高年级本科生学习使用的教材。我从1984年进入生物信息学研究领域，并且在内蒙古科技大学教授生物信息学课程也有17年了。期间一直在思考：生物信息学的内容到了稳定成型的时候了吗？生物信息学教学到底应该包含哪些基本或者核心内容？哪些最新的发展应该被及时收入进来？于是，在本书的编写过程中我们参考了国内外优秀的专著或教材，特别是孙啸等主编（2005）的《生物信息学基础》，陈铭主编（2012）的《生物信息学》，李霞主编（2013）的《生物信息学理论与医学实践》等，其他不能一一列举，在此一并谢过。

本书共分11章：第1章“生物信息学引论”，写这部分内容时尽管学习参考了许多书籍、综述和会议信息，还是感到战战兢兢；第2章“生物学基础”，供学习者选用；第3章“统计学习与推理”，便于读者查阅有关数学知识；第4章“生物信息学资源”，第5章“序列分析与序列比对”，第6章“分子系统发生分析”，这三章应该是生物信息学的最基本内容；第7章为“基因组信息学”，第8章介绍“生物芯片”这一新技术，第9章为“转录组信息学”，第10章为“蛋白质组信息学”，这四章内容应该是今后一个阶段生物信息学最热门的领域；最后一章“系统生物学”反映了生物信息学研究方法和理念上的革新，供读者参考。

本书具体编写分工如下：蔡禄编写了第1～8章和第11章，参编了第3、第6、第10章部分内容，并负责全书各章节的修改和校对工作。邢永强编写了第2章的第6节、第3章第3节及第9章，崔向军编写了第3章第1、第2节，刘国庆参与了第6章全部内容的修订，崔大超编写了第10章大部分内容。

感谢科学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，由衷地对父母和家人，以及本次参编同事表示诚挚的谢意。

生物信息学内容新、发展快、覆盖学科广，由于编者知识水平所限，难以对每一部分内容有非常深刻的理解，加之编写时间仓促，书中难免有不足之处，书中部分个人观点难免有失偏颇，欢迎大家提出宝贵意见。

蔡　禄

2017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。简要回顾了生物信息学的产生和发展历史，较为详细地介绍了目前阶段其主要研究内容。

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及生物分子相互作用等数据呈指数增长。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，我们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。既然这些都涉及信息的组织、传递和表达，我们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。在这样的背景下，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。随着基因组计划的不断延伸，各类“组学”技术迅速发展，生物数据的积累速度不断加快。因此，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开展理论分析，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）这个名词有许多不同的定义。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，对序列数据进行存储、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、隐马尔可夫模型及网络构建等各种方法，对序列、

结构、基因表达、分子相互作用等数据进行定性和定量分析，基于“综合论”思想和“系统生物学”框架，获取基因编码、基因调控、序列-结构-网络-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），计算分子生物学常常被看作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。计算分子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。

总的来说，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）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。

第一个方面是收集和管理生物分子数据，使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。

第二个方面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。早期生物信息学的主要研究对象是 DNA 和蛋白质，近些年研究人员逐渐关注生物分子相互作用及其网络表现的性质。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。在生物分子相互作用网络方面主要研究分子相互作用数据库构建，生物系统的网络建模，网络的动力学行为等。

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、基因或通路富集分析、各种组学数据分析软件、网络构建和生物仿真的数学描述等。

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、UniProt 蛋白质序列数据库和 PDB 生物大分子结构数据库，不但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，特别是在系统生物学框架下生物过程建模、仿真等也一定是非常复杂的。破译和阐明生物信息的本质将使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，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、各种组学技术等，利用基因芯片一次可以获取上千个基因的表达数据，转录组测序更是产生大量与基因表达和 RNA 加工有关的信息。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包括核酸序列、蛋白质序列和生物分子相互作用等数据，不可能用实验的方法去详细研究每一条数据，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。在此基础上有针对性地设计分子生物学实验。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。特别在系统生物学框架下，没有理论指导的实验几乎很难完成。

生物信息学研究在医学上也有重要意义。通过生物信息学分析，可以了解基因与疾病的关系，了解疾病产生的机制，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信息，继而开展生物大分子结构模拟和药物设计。在单核苷酸多态性（single nucleotide polymorphism，SNP）研究的基础上开发针对个体或某一群体的药物“个性化医疗”、加快生物学研究向临床应用的“转化医学”、在多种类型疾病相关数据基础上制订诊断、治疗和预后方案的“精准医学”被社会各界普遍寄予厚望。

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 年代以前）

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加特林堡首次召开了“生物学中的信息理论研讨会”。20 世纪 60 年代是生物信息学形成雏形的阶段，一些计算生物学家开始进行相关研究，做了许多生物数据收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。

1967 年，Dayhoff 发表了蛋白质序列图集，氨基酸序列的收集是这个时期的一项重要工作，该图集后来演变为著名的蛋白质信息资源（protein information resource，PIR）。

一般认为，生物信息学的真正开端是 20 世纪 70 年代。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，同时还不断涌现出许多生物信息数据库。

1970年，Needleman和Wunsch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，它极大地提高了序列比较算法的性能。

1971年，英国剑桥大学和美国布鲁克海文实验室宣告蛋白质结构数据库PDB开始运行。

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，Pipas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，继续涌现高效算法，出现了一批生物信息服务机构和生物信息数据库。

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。

1982年，核酸数据库GenBank第3版公开发行。

1983年，Wilbur和Lipman发表了数据库相似序列搜索算法。

1984年，蛋白质序列资源PIR正式上线。

1985年，出现了快速的蛋白质序列搜索算法FASTP/FASTN。

1986年，日本核酸序列数据库DDBJ诞生。同年，出现蛋白质数据库Swiss-Prot。

1988年，美国国立卫生研究院和美国国家图书馆成立国家生物技术信息中心（National Center for Biotechnology Information，NCBI）。同年，成立欧洲分子生物学网络（European Molecular Biology Network，EMBNet），该网络专门发布各种生物数据库。美国结构生物学研究协作组（Research Collaboratory for Structural Bioinformatics，RCSB）成立，负责PDB运行。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。

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*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年代至2003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（human genome project，HGP）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）数据库及电子克隆（virtual cloning）技术等。

20世纪80年代后，科学家开始进行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正式启动，被誉为生命科学的“阿波罗计划”。

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（The European Bioinformatics Institute，EBI）获准成立。专业蛋白质分析系统网络服务器诞生。第一届分子生物学智能系统（Intelligent Systems for Molecular Biology，ISMB）国际会议在Bethesda召开，会议一年一次。“系统生物学（system biology）”这一名词早在1968年就出现了，但真正得到深入研究是1993年Zieglgansberger和Tolle发表的研究神经系统疾病的工作。

1994年，国际生物信息学系列会议由Cambridge Healthtech Institute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。

1997 年，BLAST 的改进版本 PSI-BLAST 投入实际应用。

1996 年，酵母基因组被完全测序，Affymetrix 生产出第一块 DNA 芯片。

1998 年，亚太生物信息学网络成立。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—— *Computer Application in the Biosciences* 更名为 *Bioinformatics*。瑞士生物信息学研究所（Swiss Institute of Bioinformation，SIB）成立。美国塞莱拉基因技术公司成立，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。Prusiner 因发现引发疯牛病的朊病毒而获得诺贝尔生理学或医学奖。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完整的人类染色体——22 号染色体的遗传序列。中国于 1999 年 9 月积极参加到这项研究计划中的，承担其中 1% 的任务，即人类 3 号染色体短臂上约 3000 万个碱基对的测序任务。

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、日等国科学家宣布，他们已基本完成人体第 21 对染色体的测序工作。2000 年 6 月 26 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、英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、美、日、德、法、英 6 国科学家和美国塞莱拉基因技术公司联合公布人类基因组图谱及初步分析结果。

2003 年，人类基因组测序全部完成。国际蛋白质结构数据库联盟（world wide Protein Data Bank，wwPDB）成立。

2005 年，NCBI、EBI 和 DDBJ 成立国际核酸序列数据库联盟（International Nucleotide Sequence Database Collaboration，INSDC）。

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（2003 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，其研究的内容不但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从单个基因致病机制的研究转向多个基因致病机制的研究；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）组织之间的比较及不同发育过程的比较等。标志是大规模基因组分析、基因表达数据分析、蛋白质组分析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。开展工作如核酸序列统计分析、生物大分子二级结构预测、基因识别等。我国虽然早在 1993 年参与人类基因组计划时就列入了生物信息学的相关研究内容，但生物信息学真正开始发展是在 1995～1996 年。国内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，我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，我国生物信息学研究水平与国际同行尚有较大差距，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。

北京大学于 1997 年 3 月成立了生物信息中心，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。特别是在基因预测算法、基因组信息分析、蛋白质分子设计、非编码 RNA 研究等方面取得了一些比较好的研究成果。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中心建立的生物信息学服务器和EMBL数据库的中国节点，成为当时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，但是目前该中心这些功能已经基本丧失。中国科学院北京基因组研究所大数据中心等单位都对数据库建立充满热情，但国家级数据库的建立、维护及使用由于种种原因一直未能完成，使我国在该方面工作远远落后于发达国家。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，目前我国总体测序能力已处于世界领先水平。在生物信息分析、基因功能分析等方面迎头赶上，与发达国家之间的距离迅速缩小。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，相对而言开展生物信息学研究和服务的机构或公司较少。

国内近年来开展生物信息学研究的单位主要有：清华大学、北京大学、中国科学院生物物理研究所、军事医学科学院、中国科学院上海生命科学研究院、中国科学院微生物研究所、中国科学院北京基因组研究所、中国医学科学院、天津大学、内蒙古大学、复旦大学、中国科技大学、浙江大学、东南大学、哈尔滨医科大学、内蒙古科技大学等。近几年来，国内对生物信息学的研究和应用越来越重视，参与研究的人员迅速增加，“功能基因组信息学与系统生物学”会议的参会人数已达1000人/届以上，研究领域也愈加全面。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；转录组学、代谢组学研究迅猛发展。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”的资助方向和生物信息学相关的几次香山科学会议可大致了解我国生物信息学研究的主流发展方向。

早在2004年左右，国家自然科学基金委员会数学物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，主要资助方向包括以下两方面。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。数理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。

国家自然科学基金委员会2010年开始设立“医学科学部”，生命科学部五处首次将“遗传学学科”更改为“遗传学与生物信息学学科”。2017年，生命科学部遗传学与生物信息学学科强调生物信息学领域重点关注：发展新的算法和分析技术，用于研究基因组结构、功能与进化；整合组学数据与系统生物学分析；生物大数据的整合、标准化和可视化的方法研究；分子模块和网络的设计与合成；生物网络的研究等。鼓励生物信息学分析与生物实验验证相结合。信息科学部一处关注医学信息检测与处理、生物信息学中的信息处理与分析、生物大数据的信息分析方法、细胞和生物分子信息的检测与识别、生物系统信息网络与分析、生物系统功能建模与仿真等。2017年生命科学部设立“复杂性状的遗传解析、网络构建和调控机制”重点项目。

1997年12月召开第87次香山科学会议，专题讨论我国生物信息学的发展。会议围绕21世纪生物学和生物信息学现状及展望、蛋白质折叠规律、后基因组生物信息学研究、DNA序列分析及分子进化研究、互联网与生物数据库和信息论及其他数学理论在生物信息学研究中的应用5个议题开展了研讨。

1999年4月，国家自然科学基金委员会生命科学部、信息科学部、数学物理科学部、工程与材料科学部在北京召开“生命科学中的信息科学问题”论坛。研讨主题集中在微观尺度上的基因组和蛋白质组信息学和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SNP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。

2007 年 3 月，召开以“国际生物医学数据共享战略研讨”为主题的第 298 次香山科学会议，会议交流、分析了中美生物医学数据共享的基础和开展国际生物医学数据共享的重要意义，就如何开展国际生物医学数据共享进行了深入讨论，并达成了一些共识。

2012 年 4 月召开了以“系统生物医学中的生物信息学问题”为主题的香山科学会议第 420 次学术讨论会。会议中心议题包括：“适合人类健康与重大疾病预警的生物信息学新的技术方法探索”“系统生物医学与高性能科学计算”“信号通路调节的基因网络及其在癌变中的作用”等。会议建议重点开展以下 4 个方面的研究工作：①生物学医学资源的建立；不同层次组学的生物信息学方法；整合不同组学立体研究的生物信息学方法；生物信息学理论研究与临床研究相结合；复杂疾病的致病机制研究等。②重大疾病的风险因素（因子）的发现；相关疾病网络的建立及用于疾病诊断、风险评估，指导预防、诊断、治疗和预后；基于多源生物医学和临床数据，建立系统生物医学辅助诊断系统，进行诊断和个体化治疗；基于生物信息学发现新的疾病发生发展动态机制，通过临床实践验证其可靠性与价值。③运用系统生物医学的理论和方法来研究、识别新药靶点，从网络和多靶点的角度研究和开发新药；从多组学的角度来评价药物，实现个体化治疗。④建立复杂疾病的数据信息资源库，包括临床数据与生物医学组学（基因组、转录组、蛋白质组、代谢组和表观遗传组等）；面向临床应用，建立系统生物医学相关硬件平台；研究建立复杂疾病的生物信息学分析软件，用于疾病诊断和风险评估，指导预防、诊断、治疗和预后；系统生物医学相关人才的培养及平台推广。

2016 年 4 月 14～15 日在北京香山饭店召开以“生物大数据和精准医学时代的生物信息学核心理论问题与应用体系”为主题的第 557 次香山科学会议，会议围绕：①生命科学与医学大数据资源整合与平台建设中的生物信息学核心问题；②大数据与精准医学研究中的生物信息学新问题、新挑战，生物信息学核心理论与应用体系的未来发展方向；③面向精准医学临床实践的生物信息学应用体系等中心议题进行深入讨论。

目前较为主流的学会是挂靠在中国细胞生物学学会下的“功能基因组信息学与系统生物学”分会，2016 年在成都召开的年会包含以下 9 个议题：①大数据时代的测序技术发展与应用；②生物网络与系统的构建及功能分析；③序列信息挖掘与分子识别；④复杂生物过程与复杂疾病系统生物学及转化医学；⑤转录调控与表观遗传修饰，非编码 RNA 的辨识和功能分析；⑥蛋白质等生物大分子结构与功能研究；⑦合成生物学；⑧精准医学新理论新方法；⑨生物信息学的新概念新思想和发展趋势。从中可以看出我国生物信息学研究者主要研究或关注的热点领域。另外，中国生物工程学会、中国计算机学会等多个一级学会也设立了生物信息学分会。

1.3　生物信息学的主要研究内容

生物信息学作为一门新的交叉学科，其研究范畴是以基因组 DNA 序列的信息分析作为出发点，分析基因组结构，寻找或发现新基因，分析基因调控信息，并在此基础上研究基因的功能，研究基因的产物即蛋白质，模拟和预测蛋白质的空间结构，分析蛋白质的性质，其结果将为基于靶分子结构的药物分子设计和蛋白质分子改性设计提供依据。

1.3.1　生物大数据资源及信息学分析

1. 生物分子数据的收集与管理　核酸的序列测定是分子生物学的一大突破，并已经取得了非常大的进展，目前已测定的核酸序列的数量呈指数级增长。截至 2017 年 6 月，GenBank 存储了共计 2.017 亿条序列和 2350 亿对碱基；在蛋白质方面，UniProt 数据库中 Swiss-Prot 库包含 55.5 万条良好注释的记录，TrEMBL 含有 894 万多条记录，通过 X 射线衍射或核磁共振方法测定空间结构的蛋白质也有 133 759 个。

生物分子数据量巨大，特别是核酸序列的数据以千兆计。有组织地搜集和管理这些数据是各项工作的前提。为了便于其他研究人员共享这些数据，

及时得到最新的实验结果，也为保证数据的一致性、可靠性和完整性，国际上有专门的机构收集和管理这些数据。具体的工作包括构建数据库系统，建立网络服务器，开发数据查询和搜索工具，设计数据分析软件和数据可视化软件。对生物分子数据管理的一个特别要求是交叉索引，即数据库中的每一条数据应尽可能地与其他数据库中的相关数据链接起来。例如，从核酸数据库中的某段 DNA 序列到蛋白质序列数据库中对应蛋白质序列的链接，从蛋白质序列数据库到蛋白质结构数据库的链接。前者实际上说明了基因与其产物之间的联系，而后者则反映出蛋白质序列和结构之间的映射关系。

生物信息学发展很快，各种数据库不断涌现，并各有不同的特色。美国、日本、欧盟、加拿大等国家和地区都有基因组数据库，有的是国际性的，有的是本国的，有的公开，有的不公开。对于核酸序列，有 3 个权威组织在管理各自的数据库，一个是欧洲分子生物学实验室的 EMBL，一个是美国生物技术信息中心的 GenBank，另一个是日本遗传研究所的 DDBJ。3 个组织相互合作，各数据库中的最新数据完全一致，对于特定的查询，3 个数据库的返回结果基本一样。数据库中的数据来源于众多的研究机构和基因测序小组，或者来源于科学文献。比较著名的蛋白质序列数据库是美国生物医学基金会建立的 PIR 及瑞士生物信息学研究所和欧洲分子生物学实验室共同维护的 Swiss-Prot，2002 年整合为 UniProt。而比较著名的蛋白质结构数据库是美国 Brookhaven 实验室的大分子数据库 PDB。各种数据库可借助于 CD-ROM 发布，也可以通过互联网进行网络查询。

由于人类基因组等计划的顺利实施，生物分子数据量呈爆炸性增长，现有生物信息数据库中的数据量迅速膨胀，数据库的复杂程度也在不断增加，如核酸序列数据库、蛋白质序列数据库、大分子结构数据库、基因组信息数据库等。同时不断涌现出新的生物信息数据库，新的数据库反映了现代生物科学研究内容的拓宽和现代生物技术的发展，如基因表达数据库。有一些新数据库则是对原有数据库加工处理以后形成的二级数据库，这些数据库为特殊的应用服务，如蛋白质结构分类数据库。目前数据库中既包括主题数据、实验结果，还有大量的辅助资料，如作者、参考文献等。大多数数据库都配置了强大的查询和搜索工具，为用户使用提供最大的方便。网络环境下的数据库集成是目前生物信息数据库发展的重要特征。

数据库的内容十分丰富，除上述 DNA 序列、蛋白质序列和结构数据库之外，还有表达序列标记数据库、序列标记位点数据库、蛋白质序列功能位点数据库、基因图谱数据库、各种生物分子相互作用数据库、脊椎动物基因组数据库 Ensemble、生物通路数据库 KEGG 等一些具有特殊功能的数据库。

2. 数据库搜索及序列比对　对于许多新得到的生物分子序列，我们并不知道其相应的生物功能。生物学研究人员希望能够通过搜索序列数据库找到与新序列同源的已知序列，并根据同源性推测新序列的生物功能。搜索同源序列在一定程度上就是通过序列比较寻找相似序列。在分子生物学中 DNA 或蛋白质的相似性是多方面的，可能是核酸或氨基酸序列的相似，可能是结构的相似，也可能是功能的相似。一个普遍的规律是序列决定结构，结构决定功能。所以，当研究序列的相似性时，我们希望最终根据这个普遍规律推测出与新序列相应的结构或功能，也就是发现新的生物分子数据的内涵。这种方法在大多数情况下是成功的。当然，也有例外，也存在这样的情况，即两个序列几乎没有相似之处，但分子却折叠成相近的空间形状，并具有相似的生物功能。

对于 DNA 序列，同源搜索除有助于确定其功能之外，还有助于确定编码区域，确定基因。对于蛋白质，我们希望能够直接从蛋白质序列准确地预测蛋白质的结构和功能。通过序列的比较分析，特别是将一个未知结构、功能的蛋白质序列与已知结构、功能的蛋白质序列进行比较，可以得到一些关于蛋白质结构或功能的有用信息。通过比较不同种属的同源序列，还可以得到这些种属由它们共同祖先进化而来的信息。可以比较同类序列，也可以比较不同类型的序列，如比较 DNA 序列与蛋白质序列。当然，在比较之前，需要将不同类型的序列按照一定的规则转换成相同类型的序列，如将 DNA 序列按三联密码的关系转换为蛋白质序列。

随着各类数据库的不断涌现，数据库搜索不只局限于 DNA 和蛋白质序列，基因表达、基因组序列、生物大分子结构、生物分子相互作用、生物通路及网络等数据库搜索也迅速完善。

序列比较的基本操作就是比对（alignment），即将两个序列的各个字符（代表核苷酸或者氨基酸残基）按照对应等同或者置换关系进行对比排列，其结果是找出两个序列共有的排列顺序，这是序列相似程度的一种定性描述，它反映出在什么部位两个序列相似，在什么部位两个序列存在差别。最优比对反映了两个序列的最大相似程度，寻找最优比对

的基本算法就是动态规划算法。一个新序列与数据库中的某个序列的比较可以在很短的时间内完成，但由于序列数据库的数据量巨大，逐个与数据库中的每条序列进行比较需要很长的时间。因此，对于进行数据库搜索的序列比较算法要求具有较高的速度。目前在序列搜索方面有多种不同的实用程序，但较成功的两个程序是 BLAST 和 FASTA，它们能够根据所给定的目标序列，从 DNA 序列数据库或蛋白质序列数据库中快速地找出相似序列。它们采取专门的技术以加快搜索速度，如 BLAST 采用的是局部序列比对技术。现在，这两个程序已被广泛地应用于 DNA 或蛋白质序列分析。

与序列两两比对不一样，多重序列比对研究的是多个序列的共性。序列的多重比对可用来搜索基因组序列的功能区域，也可用于研究一组蛋白质之间的进化关系。在蛋白质研究方面，除序列数据库搜索之外，还有结构数据库搜索，而通过结构数据库的搜索，常常能发现蛋白质之间更深层的关系。例如，对于两个序列不相似的蛋白质，通过结构数据库搜索比较，有可能发现这两个蛋白质具有相似的空间结构，由此可以推测这两个蛋白质具有相似的生物学功能。

1.3.2　基因组信息学研究

1. 大规模测序与拼接　基因组计划的主要任务之一就是获得生物体全基因组序列，首先要做的就是大规模测序。不管是一代测序技术，还是二代测序技术，DNA 测序仪直接测序长度是有限的（小于 1000bp），一般要将长的基因组序列打断成小片段，短片段测序后的序列拼接都是必需的步骤。拼接是将打碎的已测序短片段按实际顺序还原出来的工作，在诸如表达序列标签（EST）数据库大量短片段中寻找 5′ 和 3′ 重叠部分，不断向两端延伸进行拼接需要高效的数学算法和高性能计算机的帮助才能实现。由华盛顿大学 Greent 和 Ewing 开发的 Phredt 和 Phrap 是其中代表性软件。

2. 基因的识别与定位　目前生物信息学仍有大量工作是针对基因组 DNA 序列的。DNA 序列是遗传信息的源泉，它对蛋白质的编码是我们所感兴趣的重要内涵。在 DNA 序列分析方面，识别蛋白质编码区域或识别基因是最关键的。如果发现一个新的基因，就可以通过生物学实验了解与其相关的生理功能或疾病的本质，为疾病防治和新药的开发提供依据。由于存在大量的 DNA 序列数据，发展识别编码区域和基因的算法也是最大限度利用生物分子数据所要求的。另外，从实验和计算的关系来看，在有些情况下，由于实验测定的编码区域并不一定完整，必须结合计算找到并证实所有的外显子（exon）。

从编码区域可以推导出基因的结构及其对应的蛋白质序列。就目前分子生物学技术的发展现状而言，实验测定 DNA 序列要比测定蛋白质序列容易得多，因此可以通过实验测定一段基因的序列，并由此推导蛋白质的氨基酸序列。实际上，许多蛋白质序列就是从为其编码的 DNA 序列直接推导而获得的。然而，要想由 DNA 序列直接得到蛋白质序列并非易事。一方面，由于许多蛋白质被编码在 DNA 序列的不同区域，对一段给定的 DNA 序列，生物学家必须猜测编码区域（即基因）从什么地方开始，到什么地方结束，在基因中间哪些地方会出现间隔。另一方面，由于人类基因组所拥有的 DNA 序列比编码蛋白质所需的多得多，给定的一段 DNA 序列也可能不为任何蛋白质编码。真核基因外显子不连续是基因识别中的一个困难，为解决这个问题，必须能够识别基因的可变剪切位点。

有许多线索可用于帮助寻找基因，如蛋白质编码区域的统计特征、基因结构中的一些特殊信号位点、基因转录调控区域的蛋白质结合位点等。在寻找基因的过程中，首先试图发现在 DNA 序列中哪一部分为蛋白质编码，如果在一段 DNA 区域含有许多终止密码子，则它成为编码区域的可能性极小。这虽然不能准确地说明蛋白质编码区域从什么地方开始，到什么地方结束，但却可以帮助猜测编码区域位于何处。编码区域统计特征、基因结构特征及基因调控信息组织特征，都可用以推测在 DNA 序列中编码蛋白质的区域处于什么地方。目前在编码区域识别或基因识别方面的算法大体可分为基于统计的方法、基于同源性的方法和基于机器学习（如人工神经网络）的方法。基于统计的方法和人工神经网络方法属于计算的方法，而基于同源性的方法属于分析比较的方法。神经网络具有非线性映射能力，能够发现数据的高阶相关性。在发现基因的过程中，利用现有与基因相关的数据可以提高基因识别的准确性。例如，使用基因表达标签 EST 序列数据或已知蛋白质序列数据可以证实基因预测的结果。使用 EST 序列信息寻找新基因是当前国际上基因争夺战的热点。另外，将理论识别方法与分子生物学实验结果结合起来，可以在一些特定的情况下较好地解决基因识别问题。生物信息学方法是发现新基因的重要手段。例如，啤酒酵母完整基因组大约 60% 的基因是通过信息分析得到的。