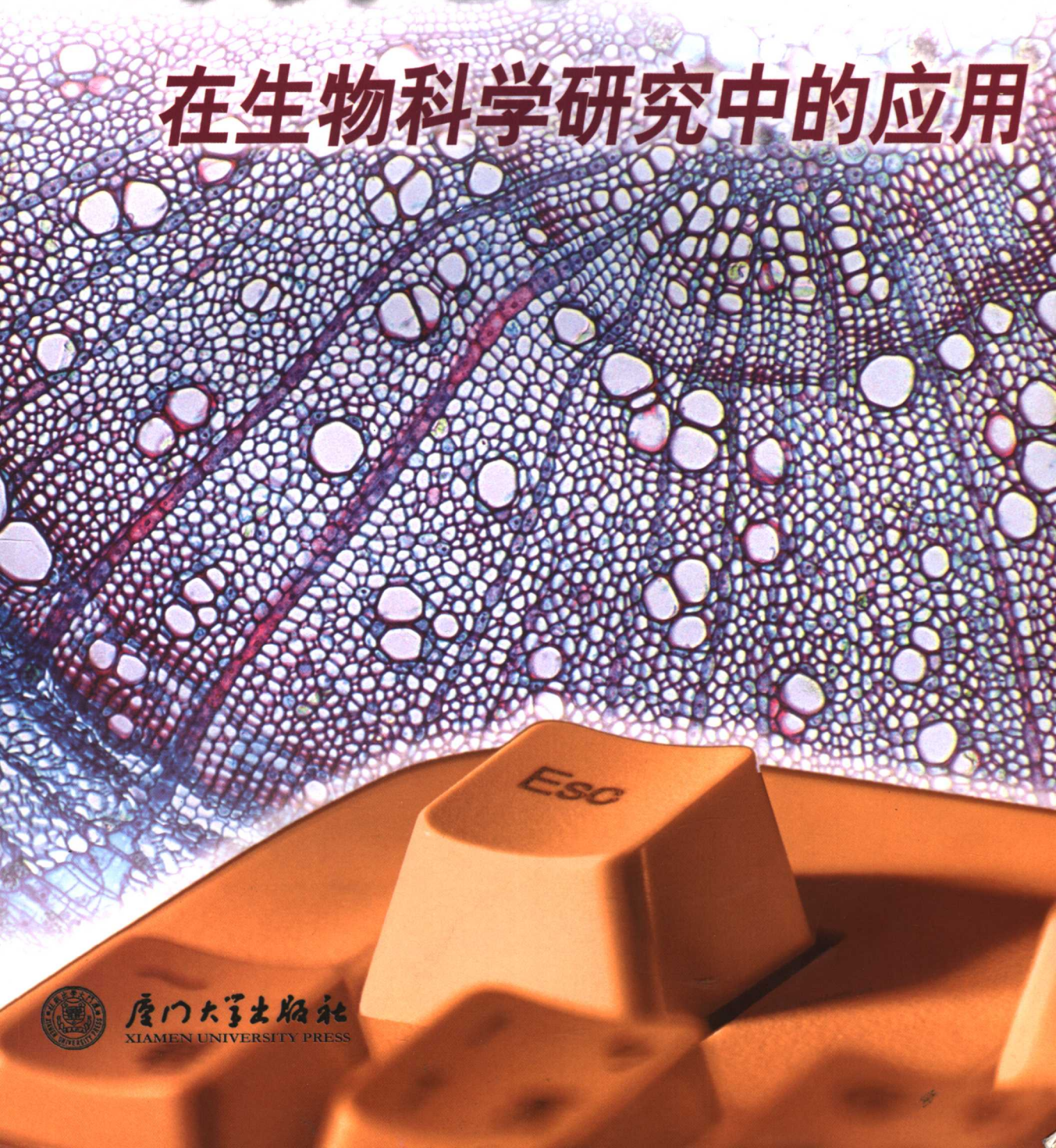
计算机在生物科学研究中的应用

宁正元 著

JISUANJI ZAI
SHENGWU KEXUE YANJIU ZHONG
DE YINGYONG



厦门大学出版社
XIAMEN UNIVERSITY PRESS

计算机在生物科学研究中的应用

宁正元　著

厦门大学出版社

图书在版编目(CIP)数据

计算机在生物科学研究中的应用/宁正元著. —厦门:厦门大学出版社,2006.11
ISBN 7-5615-2676-8

Ⅰ.计… Ⅱ.宁… Ⅲ.计算机应用-生物学 Ⅳ.Q-39

中国版本图书馆 CIP 数据核字(2006)第 134520 号

厦门大学出版社出版发行
(地址:厦门大学 邮编:361005)
http://www.xmupress.com
xmup @ public.xm.fj.cn
厦门昕嘉莹印刷有限公司印刷
2006 年 11 月第 1 版 2006 年 11 月第 1 次印刷
开本:787×1092 1/16 印张:17 字数:435 千字
定价:30.00 元

前 言

生物学是一门实验科学，它的研究方法是通过对实验的观察或实验结果的分析得出实验的结果，进而提出或验证一种生物学的理论。随着生物学中的生态学、遗传学和分子生物学等生物子学科的出现，在生物学研究中产生了海量的实验数据，处理这些实验数据需要具有较强的计算能力，而这种能力显然是手工计算所不具备的。正如伟大算法学家 D. R. Knuth 所说的："生物学中有一百年也解决不了的问题。"这就给生物学家造成了一个困惑：生物学研究中的高性能计算的出路在何方？

正是 20 世纪 40 年代末产生的电子计算机使生物学家走出了这个困境。1956 年，在美国田纳西州的凯特琳堡(Catlinburg)召开了首次"生物信息学中的信息理论讨论会"，第一次明确地提出了将计算机用于生物学研究的思想。20 世纪 80 年代，IBM 公司制造出第一台 PC 机以来，计算机迅速得到了普及，计算机与信息科学已经成为发展最为迅速的学科领域，也为生物学研究提供了更多的技术支持。正是在这个时期，利用计算机来进行生物学研究的学科——生物信息学产生了，它是当今生命科学和自然科学的重大前沿领域之一，也是 21 世纪自然科学的核心领域之一。

伴随着时间进入了 21 世纪的信息化社会，生物学研究进入了系统生物学的阶段，在系统生物学的研究中，揭示隐藏在生物自身中的丰富信息更加需要计算机技术的支持。数据库技术、人工智能技术、图形图像技术、多媒体技术等计算机技术与生物学的研究紧密结合，产生了不少举世瞩目的成果。例如：人类基因组计划就是在计算机技术的支持下完成的。

计算机在生物学研究中的应用是一个范畴较广的概念，包含了计算机在生物学的各个子学科研究中的应用。虽然近 10 年来，欧美各国和我国都出版了许多计算机在生物学研究的相关专著或教材，对计算机在生物学研究中的应用进行了很好的总结。但它们多侧重于介绍计算机在生物学的某一个领域或某一个子学科研究的应用，而系统介绍计算机在生物学研究中的应用专著还较为缺少。

本书作者阅读了大量的同类参考文献并结合自身的科研与教学实践，虽然也想尽力做到本书内容的"求全"，但是计算机在生物学研究中的应用是个范畴很大的概念，而本书限于篇幅，所以只能从计算机在几个大的生物子学科中的研究加以阐述。本书一共分成六章，分为：绪论、计算机在动物学研究中的应用、计算机在植物学研究中的应用、计算机在微生物学研究中的应用、计算机在分子生物学研究中的应用以及计算机在生物学其他分支的研究中的应用。此外，还在附录中提供了一些生物学研究的常用网址和一些常用的生物学软件。

本书可以作为生物信息学专业高年级的教学指导书，也可供相近专业的研究生以及进行

生物学研究的科研人员作为参考。

林文雄教授通读了本书全文并提出了许多宝贵的意见，在此表示衷心的感谢。

由于时间仓促和作者本身的水平有限，本书中一定存在不少的缺点，敬请广大读者指正。

作者

2006 年 9 月

目 录

第一章　计算机在生物学研究中的应用概述

什么是生物科学？在古时候，人们对生物学的认识是很有局限性的：对生物学的认识往往停留在观察上，如从中草药医生分辨草药的种类、原始的解剖学、博物学家独特的采集箱及科学航行到植物展览和动物展览等。到了19世纪，达尔文发表《物种起源》之后，生物学第一次总结出一个有重大哲学意义的普遍规律。此后，孟德尔发现了遗传学的规律，沃森和克里克发现的DNA双螺旋结构以及核酸是生命本质的一系列重大发现，为生物学发展奠定了坚实的基础，从而生物学正式摆脱了那种仅靠观察、比较的方法，发展成为一门实验科学。

传统的生物学是一门实验科学，生物学的研究主要依靠的是对实验所得的数据进行处理和分析。生物学还是一门发现科学，通过对在实验中发现的新现象、新生物规律进行分析、归纳和总结，提炼出新的生物学知识。在这个过程中，需要对实验数据进行处理和理论分析，并在此基础上解释实验的现象，认识导致实验现象发生的本质，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。随着生物科学技术的不断发展，生物数据的积累速度将不断加快，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。

进入到20世纪以来，人们渐渐意识到信息是十分重要的，随着信息科学和技术的发展，信息已经和物质、能量一起构成客观世界的三大要素。信息是客观实体运动状态和过程的抽象描述。人类已经进入了信息化的社会。作为信息社会中最为重要的工具，计算机在人们生活中发挥着日益重要的作用。随着网络技术和通信技术以及半导体技术的发展，计算机的功能越来越强大。计算机科学是对社会各个层面影响最大，渗透力最强的高新技术。

回顾20世纪人类所取得的科学成就，以计算机为代表的信息技术得到高速的发展和应用。在以计算机为代表的信息科学取得快速发展的同时，现代生物科学研究也取得了极大的成功。美国基因研究专家埃里克·兰德2005年在波士顿举行的“生物信息世界”会议上指出：“现代信息技术正在促使生命科学揭开新的一页。计算机科学家与生命科学家合作至关重要，基因学的发展不仅依赖于生物信息技术的进步，也离不开信息技术。”兰德目前担任美国麻省理工学院伍德海德研究中心负责人，他认为在过去的20至30年中，生物学已从一种以实验室为基础的科学转向以信息为基础的科学。

1.1 计算机在生物学研究中的应用回顾(1956—1980)

自 1946 年第一台计算机 EANIC 在美国出现之后,在一段时间内,计算机主要是用于军事用途的,在生物学上的应用几乎没有。随着计算机科学的发展和生物学实验对数据分析和数据计算的要求不断增加,计算机开始慢慢地进入生物学研究中。

计算机用于生物科学研究大致上是在 20 世纪 50 年代末。1956 年,在美国田纳西州的凯特琳堡(Catlinburg)召开了首次"生物信息学中的信息理论讨论会"。但此时计算机应用于生物学的研究,主要还是在计算机上利用数学模型、统计学方法对宏观生物学数据进行处理。也就是在这时候,出现了一些研究的成果,如:数量分类学和数量生态学以及数量生理学和生物力学等,经过了二十多年的发展,到了 20 世纪 70 年代时已经基本成熟(Sneath 和 Sokal,1973;Pielou,1976)。

在计算机应用于宏观生物学的研究取得一定进展后,计算机应用于生物的研究开始进入分子生物学等微观领域,其中,包括建立分子生物学数据库以及蛋白质结构的计算机辅助分析预测等;1962 年,Zuckerkandl 和 Pauling 提出了序列变异与其演化存在一定的关系,从而奠定了分子演化领域的研究基础;1964 年,Davies 开创了蛋白质结构预测的研究的新领域;1970 年,Needleman 和 Wunsch 提出了现在广泛应用于两序列比对的 N-W 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。

正是在这些科学工作者的不断努力下,在上述科学研究的领域中,人们已经逐步形成了理论基础和一批方法、模型与软件,这些领域有的到今天还在不断的发展着。

1.2 计算机在生物学研究中的应用现状(1980—)

自 20 世纪 80 年代,IBM 公司制造出第一台 PC 机以来,计算机迅速得到了普及。而且近二十年来,计算机与信息科学已经成为发展最为迅速的学科领域,也为生物学的研究提供了更多的技术支持。在这个时期,生物学与计算机科学相结合的学科——生物信息学产生了,它是当今生命科学和自然科学的重大前沿领域之一,也是 21 世纪自然科学的核心领域之一。从国外近几年的应用情况来看,生物信息学在理论上促进了生物学研究(特别是分子生物学)研究的发展,使人类对生命本质的认识更加深刻。生物信息学已经改变了传统生物学的研究方法,提高了生物学实验的科学性和研究的效率。

在这个阶段，计算机在生物学研究中的应用更为广泛与深远，这一时期在生物学研究中用到的计算机技术大体有以下几个方面：

1. 数据库技术、数据挖掘技术与海量存储技术

生物信息数据库具有数据结构和组织方式复杂、数据量增长十分迅速等特点。《核酸研究》(Nucleic Acids Research)杂志连续七年在其每年的第一期中详细介绍最新版本的各种生物学数据库。在2000年1月1日出版的28卷第一期中详细地介绍了115种通用和专用数据库，包括其详尽描述和访问网址。在DNA序列方面有GenBank、EMBL和DDBJ等。在蛋白质一级结构方面有SWISS－PROT、PIR和MIPS等。在蛋白质和其他生物大分子的结构方面有PDB等。在蛋白质结构分类方面有SCOP和CATH等。

很多数据库涉及非结构化的数据，例如：PDB中的蛋白质三级结构等[1]。利用传统的关系数据库对这些非结构化的数据进行管理就显得有些力不从心了，所以，必须要采用面向对象等数据库新技术来处理复杂结构的生物数据。生物信息数据库具有种类繁多的特点，目前各种生物信息数据库大约有600种，分布在全球各个数据库服务器中[2]。又因为这些数据库的结构各异且往往是分布在不同的位置的，采用分布式数据库技术对这些不同数据库中的数据进行整合。此外，生物数据库中的数据质量并不能保证完全可靠，所以常常采用ETL技术(Extraction，Transformation and Loading)对数据库进行清洗、转换和装载。

现代生物研究的数据量增长十分迅速，特别是随着人类基因组计划和人类脑计划等大型的科学工程的相继实施。如何处理如此大量的数据摆到了科研人员的面前，这里关键的问题是如何设计生物信息专用的海量存储技术。由于现有的存储技术发展与生物学数据量的增长相比相对滞后，而生物信息资源的有效使用率目前仍然较低，严重影响了生物信息的利用。信息存取已经成为生物学界一个具有挑战性的问题，同时也是对计算机科学的一个挑战。

所以，计算机技术在生物学的应用中，数据库技术是最基本的技术。生物学的实验数据的存储、管理、查询都是建立在数据库管理系统之上的。

随着数据库技术、计算机网络和人工智能等技术的发展，出现了一种新的信息管理技术，即数据仓库技术(data warehouse)。20世纪90年代，Immon第一次提出了数据仓库的概念之后，这种技术就迅速地发展起来。所谓的数据仓库是指从多个内容相关的、物理和逻辑上都相互独立的数据源中提取面向主题的数据集合，通过一定的技术将这些数据集成起来，并对这些数据按时间重新进行组织与集成，从而可以为用户提供决策支持分析。根据不同用户的不同需求可以完成多种数据的查询、分析与决策。

随着当代生物学实验的手段不断的进步，所产生的实验数据的信息量是十分庞大的。如何在如此浩渺的信息海洋中发现潜在的规律呢？而构建的数据仓库中的数据究竟蕴含着什么样的规律呢？这个问题让不少研究人员觉得头痛。而数据仓库技术中提供了一个解决方案，就是数据挖掘技术。

数据挖掘(data mining)就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘与聚类分析的方法在蛋白质的结构预测中也有广阔的应用空间。数据挖掘技术一般分成四个基本步骤：数据选择、数据转换、数据挖掘和结果分析。现在生物学的发展产生了大量的数据，这些数据中隐藏着许多目前未被我们发现的自然规律。由于，目前生物学主要还是以实验为

主要研究手段，如何从已有的海量生物学实验数据中挖掘出一些未知的生物学规律是数据挖掘技术令人心动的应用领域。

例如：数据挖掘可用于分析基因表达数据相似性度量，从中发现基因表达数据相似性和波动相似性类似，从而提出以波动相似性为依据的相似性度量函数。

2. 机器学习与模式识别技术

机器学习算法(machine learning methods)，抽象的统称，实质是一种统计学的方法，它自动地从一个样本的训练(training)过程中获得数据信息，这种方法适用于有大量数据但缺乏相应理论的情况。如 BRNNs(Bidirectional Recurrent Neural Networks，双向重复神经网络)算法即属于机器学习算法，它的训练过程即通过对样本进行有效编码，输入网络，训练网络各权值参数和阈值参数，使网络达到基本稳定。目前机器学习方法包括：神经网络法、决策树法、基于事例学习法、符号性知识优化法及基于逻辑的归纳学习法[4]。

数据是机器学习的基础，对于生物学实验数据也一样。在大多数情况下，生物学中的知识和数据可以用序列的模式或序列的特征来概括。

随着人工智能研究不断取得进展，人们逐渐发现研究人工智能的最好方法是向人类自身学习。因此引进了一些模拟进化的方法来解决复杂优化问题。其中较有代表性的是：进化主义思想和联接主义思想。近年来，许多科学家致力于这两种方法的研究。

模式识别是机器学习的一个主要任务。所谓模式，指的是对感兴趣客体定量的或者结构的描述，而模式识别就是利用计算机对客体进行鉴别，将相同或者相似的客体归入同种类别中。模式识别的关键是通过数据分析，提取分类对象的本质特征，建立分类特征模型。在此基础上设计模式分类规则和分类器，判别待识别模式的分类情况。分类特征模型描述各种目标对象的特征，以便于利用特征进行识别。模式识别主要有两种方法：一种是根据对象统计特征进行识别，另一种是根据对象的结构特征进行识别。

在机器学习中，数据分类与模式识别密切相关。数据分类就是根据数据的特性进行分类。与数据分类相关的另一种数据分析方法是数据聚类。数据分类是有监督的学习，这种方法是指在学习过程中接受外界输入的学习指导信号，数据聚类是无监督的学习，是完全靠自己的能力进行学习的。数据聚类分析方法在基因的表达数据分析中有着重要的应用。

利用机器学习的方法可以应用于蛋白质结构预测，但现在的问题是从蛋白质一级结构序列预测蛋白质二级结构和三级结构的准确率较低，还有许多问题需要解决。众所周知，蛋白质分子是由 20 种基本氨基酸通过肽键连接而形成的共价多肽链。天然蛋白质在生理特点大多是由其特定的空间结构确定的。遗传信息由 DNA 到 RNA 再到蛋白质的过程，一直是分子生物学研究的中心，通常称之为中心法则。由 DNA 到 RNA 再到多肽链合成的基本过程已经基本清楚，但是一定氨基酸序列连接形成的多肽链是如何形成特定的空间结构且具有生理功能的蛋白质分子仍然是个尚未解决的问题。蛋白质的一级结构决定高级结构是进行蛋白质结构预测的理论基础。蛋白质结构预测是生物信息学的核心问题，在这方面主要的研究在于如何通过已知的蛋白质一级结构序列和其对应的三级结构序列来挖掘知识，从而形成蛋白质一级结构序列与三级结构的对应关系的知识。

3. 人工心智和心脑科学在生物学中的应用

了解脑及其全部功能是21世纪重大挑战之一，人类脑计划开始于1993年，这项行动的主要目标：创立以web为基础的神经科学所有数据的数据库，并提供数据分析、整合、合成、建模与模拟的先进工具，有助于实现了解健康与有病神经系统功能的最终目标。脑是生物体内结构和功能最复杂的组织，人脑内有上千亿个神经细胞，神经突触超过1 014个，是生物体接受外界信号、产生感觉、形成意识、进行逻辑思维、发出指令产生行为的指挥部，但它的功能目前还不为人们所了解。

在人类脑科学计划提出后，产生了一门新的交叉学科——神经信息学。神经信息学产生的先进的信息学解决方案，将加速对脑的了解，并能将基础研究转化为诊断、监视、处理和预防脑疾病的更好手段。反过来，关于数据与信息的获得、存储、提取、分析、合成及可见的生物学机制的阐述，将更加清楚地解释信息学技术，以至随着时间的推移，计算机将能超过人脑的工作[5]。

人脑的结构和功能极其复杂，需要从不同的层次对其进行研究，包括：从DNA、RNA、蛋白、神经元、神经网络到全脑。其中对神经网络和全脑功能的研究近年来发展很快，成为神经信息学研究的重点。神经信息学主要从信息和信息处理的观点来研究人脑，研究神经系统信息的载体形式，神经信息的产生、传输与加工，以及神经信息的编码、存储与提取机理等，并从系统和信息的观点建立以生物学实际为基础的神经网络模型。以生物学实际为基础的神经网络模型的研究对仿脑计算的研究具有极大的促进作用。人工智能正在最新神经科学与心理科学成果的启发下朝着人工心智、情感计算与仿脑计算的方向发展。

4. 生物分子的计算机模拟技术

传统的生物分子研究主要是通过生物学实验来分析和表征生物分子，如利用测序技术确定DNA或RNA分子的序列，通过分子遗传学方法确定基因的多态性，通过X射线衍射技术来确定蛋白质等生物大分子的结构，通过生物化学实验来研究生物大分子之间的相互作用、药物分子和靶分子的结合等。

现代对生物分子的研究也可有采用计算机模拟生物分子的技术。所谓生物分子的计算机模拟就是从分子或者原子水平上的相互作用出发，建立分子体系的数学模型，利用计算机进行模拟实验，预测生物分子的结构和功能。可以模拟生物大分子与大分子之间的相互作用，模拟生物大分子与具有活性的小分子之间的相互作用，研究分子之间的识别与及分子间的特异性结合。生物分子的计算机模拟对从理论上解释实验现象、指导设计实验方案、发现新的现象以及产生新的科学假设具有重要的作用，美国哥伦比亚大学化学系的生物分子模拟中心就在NIH的支持下作了这方面的工作。

一般在生物分子模拟之前，首先为待模拟的分子体系建立模型，描述分子内和分子间的相互作用。常用的两种理论模型是量子力学模型和分子力学模型。利用这些理论模型进行分子系统的能量变化情况。然后用得到的模型进行模拟，模拟采用的方法一般是基于统计物理学的分子动力学方法、基于随机统计原理的蒙特卡罗方法、基于全局性的极小化方法的模拟退火方法。这些方法对于生物分子的计算机模拟，起到了积极的作用。

5. 网络技术

随着人类进入信息社会，网络已成为社会的基础设施，对人们的生活起着重要的影响。电子邮件和新闻组已经成为生物学科研中的最要交流工具。而且网络提供的各种服务，如：FTP服务，WEB服务等也为科研人员提供了重要的服务。

目前，Internet 上有着巨大的生物学资源和生物学的相关数据库与知识库。使用者可以通过网络查询或搜索所需要的生物学信息，使用各个网络站点提供的分析工具对生物进行分析。生物信息的研究者能够下载大量的数据，但如何集成这些数据不是一件容易的事。XML是一种元语言，可以用来定义和描述结构化数据，它是 Web Services 得以实现的语言基础。Web Services 的其他协议规范都是以 XML 形式来描述和表达的。Visual Genomics 开发了一种用于生物学信息处理的 XML 标记语言 BSML(Bioinformatic Sequence MarkupLanguage)[8]。它使基因数据能以动态的和可重用的方式传递给 BSML 浏览器。公共的和私有的数据库及应用能以 BSML 的格式传递。免费的 BSML 浏览器使研究人员能够以可视化的和交互式的方法来存取 BSML 数据和注释。Fenyo 开发了一种 Biopoly－met Markup Ltlllguage(BioML)用来对蛋白质和核酸的序列数据的复杂注释的表达。BioML 用 DTD 的方式获得各种信息源(核酸、蛋白质数据库)的数据的集成。欧洲生物信息研究所也开发了 XEMBL 来发布基于 XML 的 EMBL 数据库信息。

Web Services 技术由于使用标准的 Web 协议(http 、SMTP 等)和一系列标准协议(XML、SOAP、WSDL 等)，为生物信息集成提供了一种崭新的方法。当把 Web Services 应用到生物数据库中时，所有生物数据库系统都成了一个松散结构中的组件，系统接口、应用通信、数据转换和目录信息都是建立在开放的、被广为接受的标准之上，用户能迅速地访问到他们所需要的信息。Web Services 的最大特点是具有真正意义上的平台独立性和语言独立性。基于 Web Services 技术的生物信息集成方案，可以方便地实现各种已有生物数据库系统、新开发的 Web Services 应用等各种系统的集成，必将广泛地应用于生物信息的研究领域中[7]。

针对生物学有些研究的样品分布的范围广和分散的特点，基于移动网络技术的移动计算机在生物学研究中也起到了很重要的作用。特别是 Java 语言和 C# 语言对移动设备(手机，PDA 等移动终端)的支持较好，已经有一些可以应用的移动应用程序可以用于生物学的研究，由于 XML 技术的支持，它们往往还是可以跨越不同平台的。

6. 高速计算能力与网格计算技术

生物学研究需要对大量的样本进行分析计算或统计，这就为高性能计算提供了一个大的应用领域。生物学研究中的计算面临巨大的计算量与海量的数据，如利用分子动力学模拟一个蛋白质的折叠就需要一个巨型机几个星期的运算。这给高性能计算、并行计算和网格计算提出了挑战。

什么是并行？并行指的是若干个可以同时进行运算或操作的部分。并行性可以分为同时性和并发性两种。前者指的是两个以上事件在同一时刻内发生，后者指的是两个以上事件在同一时间间隔内发生。人们已经提出了多种的并行方案，它们大致上可以分成六大类：专用多功能单元、相联处理机、阵列处理机、数据流计算机、函数式编程语言处理机和多处理机系统。

被称作生物学界“坏小子”的美国基因学家克雷格·文特尔的基因测序实验室中就装备了300多台高性能的计算机。这些计算机可以看成是一张网上的结点，一个结点也许承受不了一条鱼的重量，但是，将这些结点组成一张网，它就能捞起一网鱼。现有的计算机大都是串行运算的，如何将它们协调起来，组成并行的系统，是一个研究的热点。

而且，在生物信息学研究中，专用高性能计算机已经成为一个重要的课题。IBM宣布将耗资1亿美元研制一套代号为“蓝色基因”(Blue Gene)的超级计算机，通过对各个蛋白质分子聚合到一起的多种力量加以测量，来研究人类蛋白质分子的折叠方式。IBM预计“蓝色基因”将拥有100万个处理器，计算能力达到1千万亿次浮点结果，比目前世界上最快的IBM计算机的12.3万亿次快了近百倍。据透露，“蓝色基因”将采用一种称为SMASH的全新体系结构，可以在简化指令的基础上实现800万个线程并行处理的能力，并能做到自稳定、自适应和自修复。整套系统由64个6英尺高的机柜互联而成，每个机柜配置8块主板，每块主板上有64个芯片，每个芯片上包含32个处理器。在国内，中科院计算技术研究所与华大测序中心合作，基于曙光3000超级计算机系统，开发了Balst，Phrap，Smith－Waterman的并行算法，并应用于华大测序中心的数据处理流程中。

7. 专家系统

专家系统(exepert system)是一种基于知识的智能系统，它将领域专家的知识用知识表现的方法表示出来，并放入知识库中，供推理机使用。专家系统利用知识和推理机解决那些需要特殊的、重要的人类专家知识才能解决的复杂问题。一般的专家系统是由六大部分：知识库、数据库、知识获取部分、推理机、解释机构和使用界面组成的。知识库中的知识也可以分成事实性知识和启发性知识两大类。生物学研究中已经有了不少的专家系统。

8. 计算机图形学

众所周知，DNA序列是两条碱基互补的脱氧核糖核酸形成的双螺旋结构。一般认为，它们可以用一条序列来进行表示。根据文献按照某种规则，人们可以把DNA序列转换为一条z型曲线，该z曲线与所表示的DNA序列的关系是一一对应的，即：一个特定的DNA序列有唯一的一条z型曲线与它对应；反之，对任意一条给定的z曲线，可找到唯一的一个DNA序列与之对应。也就是说，z曲线包含了DNA序列的全部信息。z曲线是与符号DNA等价的另一种表示形式。这样就可将复杂的DNA序列转换为一条空间中的曲线。对z曲线曲率和挠率的计算和分析，可用于识别DNA序列的不同的功能区等。DNA序列的几何学研究是建立在计算机图形学的基础上的，对DNA序列几何学的研究必将为计算机图形学的研究提出一些新的课题。

而蛋白质是由若干种基本氨基酸组成的，它的空间结构决定着它的功能。利用计算机图形学的研究成果，对于预测蛋白质的功能有着重要的作用。对蛋白质空间结构的比对、显示也将促进计算几何与计算机图形学的发展。

事实上，计算机科学和生物科学都是不断发展的学科，所以计算机在生物学中的应用也是不断发展的，一种新的计算机技术有可能为生物学的研究提供更好的支持；反之，一种新的生物学的需求，有可能带动相关的计算机技术的发展。

1.3 计算机在我国生物学研究中应用的情况与面临的挑战

我国的科研人员在20世纪60—70年代就开始利用计算机在生物学研究中进行数据的统计分析，但是应用的层次低，多用于教学和实验数据分析处理。我国的生物信息工作是逐步发展起来的，20世纪80年代初仅在个别单位开展了一些计算分子生物学的工作，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。虽然我国在1993年就在中国人类基因组计划中加入了生物信息学的相关研究内容，但是真正的开始是在1995年。目前，我国所用到的生物数据库和生物系列软件多半来自于国外，基础力量还比较薄弱。

1997年，香山会议专题讨论了我国生物信息学的发展。1999年，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了“生命科学中的信息科学问题”论坛，提出了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。2000年国家自然科学基金委员会主持召开的“生物信息学前沿方向”研讨会上，与会专家提出了我国生物信息学发展的方向是：建立国家生物医学数据库与服务系统、人类基因组信息结构分析、功能基因组相关信息分析和研究遗传密码起源与生物进化（尤其是分子进化）的过程与机制。

近几年来，我国对生物学中的计算机应用工作越来越重视，研究的层次也不断提高。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，计算机都起到了重要的作用。

北京大学于1997年3月成立了生物信息学中心，中科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。

2003年8月18日，“作为国内服务器品牌三甲之一”的曙光信息产业（北京）有限公司（以下简称曙光公司）与国内著名的基因组、生物信息研究中心华大基因联合推出国内第一款完全拥有自主知识产权的生物信息专用计算机，采用先进的基因数据库架构技术、数据定制可视化技术、数据密集技术、网格使能技术、在线扩展技术及机群系统等技术，为国内用户搭建了一套与国际生物信息研究主流趋势相接轨的系统平台。该系统是建立在华大基因和曙光公司在生物信息研究领域长期合作成果的基础之上，通过运用曙光公司每秒3万亿次浮点峰值运算能力的Linux超级服务器，以支持数据密集应用为主，为国内大量致力于基因组研究的科研工作者们提供方便、快捷的服务。“生物信息专用计算机”采用机群结构，系统中节点根据功能划分为计算节点、数据库节点、服务节点三种类型，为生物信息学研究提供了一个基于硬件、软件和数据库集成环境下的统一运行平台，为各个分析软件、子数据库模块提供一致的运行和管理环境。同时用户可以根据需要选择软件和数据库模块，无缝集成到平台上。平台提供ORACLE数据库和软件的集成接口和管理工具。生物信息专用计算机以模块化的方式提供大量基因组学、生物信息学研究的常用分析工具，并能实现分布式高性能计算。用户也可以根据需要定制分析软件，添加到该专用计算机应用平台中。

对于我国来说，生物信息学人才的培养是当务之急。生物信息学是一个交叉学科研究领

域，这对生物信息学研究人员在知识结构上提出了非常高的要求，特别是对于来自数学或计算机专业的研究人员，不仅要掌握生物学的基础知识，还要求深入了解生物学中的相关问题，这样的人才不是单一学科能够培养出来的，要求跨学科地培养生物学和信息科学的复合型人才。目前中国科学院和国内一些著名大学已经开始较大规模地培养生物信息学专业人才，这为我国今后生物信息学的发展奠定了良好的基础。可以相信，我国未来计算机在生物学中的应用一定会有着很大的进步与发展。

1.4　计算机在生物学研究中的应用展望

虽然计算机在生物学应用中取得了不小的成果，但还有许多的问题摆在人们面前。目前计算机在生物学研究中的应用面临着许多的挑战：

(1)需要建立交互性好的生物学应用软件，生物学数据库及相关的数据挖掘技术。现有的生物学软件种类繁多，功能也不尽相同，但是，大部份软件都要求用户有较强的计算机基础，甚至还有一些软件是基于 Linux 或 Windows 控制台的，起特殊的命令语法不是一般的科研人员所能掌握的。而且，有些软件的源代码不是公开的，特定用户就不能根据自己的需要对程序进行修改，进而适应自己研究的需求。寻求一种好的方法，来开发出交互性好，操作方便而功能强大的生物学研究软件是今后一个重要的目标。

在世界范围内，已经开发出了大量的结构不同的生物数据库，如何在它们之间实现数据集成与共享是有效利用生物信息资源的关键技术问题。在信息集成和数据挖掘的过程中，首先必须考虑生物在不同水平(层次)上信息之间的复杂联系。这需要在数据库技术和数据挖掘算法上再做进一步的研究。

(2)需要能提示大规模数据集合中不同组分之间关系的统计分析方法及优化算法。在生物学研究中，获取所得的实验数据往往可以根据其数据特征的不同分成若干组分，这些组分之间的关系是怎样的？如何在实验数据中确定分组的标准？如何用更快的算法更有效率的确定数据的分组标准等等都让科研人员十分困惑。

例如：不同物种间可能包含了同源或非同源的数据基因，而不同基因可能在 DNA 或蛋白质序列上具有较高的异质性。因而，在基因组水平上比较不同物种或不同基因之间的相似性，有助于揭示整个基因组进化与物种进化的规律。

(3)需要开发适合于微阵列和基因芯片等新技术的数据分析工具。微点阵杂交中涉及上万个寡核苷酸，并依杂交信号强弱、探针位置和序列确定靶 DNA 的表达及多态性等。目前，迫切需要提高检测的自动化程度和数据的并行处理能力。

生命科学也是不断发展的，它在发展过程中，也不断的对计算机技术提出新的要求。现在，人类基因组计划已经完成，我们已经进入了后基因组时代，研究的重点也从提示生命所有的遗传信息转移到在分子整体水平上对基因功能的研究，这种转变的一个标志就是产生了功能基因组学(functional genomics)，它主要是研究基因的转录调控信息、注释基因产物的功能、研究基因的表达机制和比较基因组学的研究。这里，如果没有计算机的参与，是不可能实现

的。

国际生物信息学产业和市场已经逐步形成。国外一个研究小组对生物信息学市场进行了调查，结果表明：许多IT公司正在逐步的进入这一领域。目前，这类的新兴公司大都在美国，并以出现了专业化的趋势，例如：有的公司专门提供生物学网络计算服务平台，有的提供数据处理、解释和可视化软件包，有的致力于建立生物数据和专家知识集成数据库系统和技术平台。像IBM、Motorola、HP、Compaq、SGI等著名的IT企业都已经介入生物信息学的领域，由此可以想到，日后，计算机科学将同生物科学融合得更紧密。

参考文献

[1] H. M. erman, J. Westbrook, Z. Feng, et al. The protein Data Bank[J]. *Nucleic Acids Research*, 2001, (28): 235～242

[2] D. R. Westhead, J. H. Prish, R. M. Twyman. Instant Notes in Bioinformatics[M]. United Kingdom: Bios Scientific Pub Ltd, 2002

[3] SCRATCH servers, http://hpdb. hbu. cn/thesis/2005/yht/principle/principle. asp[EB/OL]

[4] 卢美律. 蛋白质结构预测与机器学习[J]. 科学, 1996, 46(5): 22～27

[5] 沈均贤. 人类脑计划与神经信息学[J]. 生物物理学报, 2001, 12(17): 607～612

[6] Ligeng Ma, Jinming Li, LiJin qu, et al. Light control of Arabidopsis development entails coordinated regulation of genome expression and cellcular pathways[J]. *Plant Cell*, 2001, 139(12): 2589～2607

[7] 生物信息学对计算机科学发展的机遇与挑战[J]. 生物信息学, 2001, (3): 37～41

[8] BSML Organization. Bioinformatic Sequence Markup Language Version 3. 1[EB/OL]. http://www. bsml. org/resource/, 2003

[9] Fenyo. The biopolymer Markup Language[J], *Bioinformatics*, 1999, (15): 339～340

[10] Lichun wang. XEMBL: distributing EMBL, data in XML format[J]. *Bioinformatics*, 2002, (18): 1147～1148

[11] 郝柏林，刘寄星. 理论物理与生命科学[M]. 上海：上海科学技术出版社，1997

[12] Hang C T, Pickover C A, et al. Viusalizing Biological Informatin[M]. Singapore. World Science Pub co, 1993

[13] 钟扬，张亮等. 简明生物信息学[M]. 北京：高等教育出版社，2001

[14] 赵青，黄小兵. 生物信息研究的加速剂[J]. 互联网天地，2004，76～77