医学生物信息学

YIXUE SHENGWU XINXIXUE

主 编◆赵雨杰

人民军医出版社

医学生物信息学

YIXUE SHENGWU XINXIXUE

主　编　赵雨杰
主　审　孙开来　何钦成
编　者　孙开来　何钦成　赵雨杰　何　群
马汝海　刘树春　陆祖宏　孙　啸
周伟强　马佳明

人民军医出版社
People's Military Medical Publisher
北　京

图书在版编目(CIP)数据

医学生物信息学/赵雨杰主编．—北京：人民军医出版社，2002.11
ISBN 7-80157-646-2

Ⅰ．医… Ⅱ．赵… Ⅲ．医学：生物学：信息学 Ⅳ．R318

中国版本图书馆 CIP 数据核字(2002)第 065227 号

人民军医出版社出版
（北京市复兴路 22 号甲 3 号）
（邮政编码：100842　电话：68222916）
人民军医出版社激光照排中心排版
北京天宇星印刷厂印刷
桃园装订厂装订
新华书店总店北京发行所发行
*
开本：787×1092mm 1/16 · 印张：18.75 · 字数：452 千字
2002 年 11 月第 1 版　（北京）第 1 次印刷

印数：0001～4000　定价：40.00 元

前　言

医学生物信息学是一门崭新的学科，它利用现代计算机技术对基因组测序、蛋白质序列测定、结构解析等实验获得的有关生物分子的原始数据进行收集、整理、管理；对数据进行比对、分析，建立计算模型，进行仿真、预测与验证，根据生物分子在基因表达调控中的作用研究生物结构、功能的相关信息，描述人类疾病的诊断、治疗的内在规律。人类基因组计划完成后，科学家们研究的重点将从基因组测序转向对基因组表达的分析，转向对蛋白质结构与功能的预测。除了单一基因及其所表达的蛋白质的功能之外，多个基因、多种蛋白质相互作用更是人们感兴趣的课题。在生物学界，尤其是在医学领域，要想对基因及其所翻译的蛋白质进行分析与预测，需要物理学、计算科学、系统科学、控制科学、信息科学与生物学的多学科的科学家共同完成。

20 世纪末生物科学技术的迅猛发展，生物科学的数据资源在质量上发生了许多突破性的进展，并且在数量上也得到了极大丰富，数据资源爆炸式的发展，迫切要求一种强有力的工具去组织它们，以利于对已知生物学知识的储存和进一步的加工利用。大量多样化的生物学数据资源中蕴含着大量重要的生物学规律，这些规律是我们解决许多生命之谜的关键。然而，继续沿用传统手段，以人脑来分析如此庞杂的数据是很难完成的，尤其在生物医学领域，各学科的研究已经深入到了分子生物学水平，科学家们迫切需要利用生物信息学的方法分析基因的组织结构、复制、转录和翻译以及基因表达的调控规律。生物信息学的发展已经对我们了解生命、了解人类自身，对于医药、保健、农业等起到了极大的作用。

在我国，生物信息学正在兴起。目前，国内已有大学开始招收生物信息学的本科生、硕士研究生和博士研究生。除此之外，目前从事分子生物学研究的研究人员，急需了解人类基因组计划的进展及意义，分析基因的奥秘，探索基因和蛋白的功能，解析基因调控网络的规律。但目前国内尚缺少适合医学工作者的生物信息学教材。因此，作者于 2001 年初开始编写本书，在编写过程中参考了国内外多种版本的相关专著。本书共 11 章，尤其在第 8 章详尽介绍了生物信息学中与生物医学有关的多种应用软件的操作方法。我们在考虑基础性和系统性的同时，更加重视可操作性，介绍了已有典型的工具和数据库，包括应用软件、Internet 资源、向数据库提交 DNA 序列以及进行序列分析和利用核酸序列与蛋白质序列进行预测的方法，使读者借助这本“生物信息学工具书”可以步入这一崭新的、有待开发的、颇具科学魅力的生物信息学殿堂。

赵雨杰

2002 年 7 月于沈阳

目　录

绪　论

生物学与信息科学是当今世界上发展最迅速、影响最大的两门科学。而这两门科学的交叉融合形成了广义的生物信息学(Bioinformatics),该学科正以崭新的理念吸引着科学家的注意。生命现象是在信息控制下不同层次上的物质、能量与信息的交换。传递过程是指不同层次核酸、蛋白质、细胞、器官、系统、整体等。广义生物信息学主要包括如下几个方面:

- 生物的遗传信息:DNA-RNA-蛋白质,遗传信息-转录-翻译,遗传信息生物信息学。
- 生命活动的调控:基因的功能、表达和调控蛋白的结构、功能及细胞活动(分化、发育、衰老、死亡)的调控;器官、系统、整体活动的调控节律、生物钟、分蘖、生长、开花、结果、营养的吸收、传输、转化、对外界信号的反应等。
- 生物电磁学与电磁生物学:生物电磁:生命活体在不同层次(电子、离子、原子、基因、细胞、组织、整体等)的活动和不同属性(包括思维、精神)活动时以及和外界环境(生命体周围直至宇宙)相互作用时反映出来的各种电磁信息。人体的电磁辐射(包括发光):频率、强度、频谱。人体信号的调制方式:调幅、调频、编码。电磁生物学:电磁辐射对生物体的影响,电磁场导致 DNA 突变,体内细胞电离、极化状态变化而导致疾病。
- 视觉系统与光信息处理:视网膜神经元回路与信息处理、彩色视觉及彩色图像的编码、变换机制、眼动成像机制及宽视场、消色差动态成像系统、视觉认知机制及其图像信息的智能模式识别、不同状态立体视觉机制和静态、动态立体视锐度。
- 脑和神经系统与信息:脑的感知觉信息处理原理及其应用,学习、记忆、思维,逻辑思维和形象思维,思维模型与信息处理系统新原理的研究,新的计算模型、新型计算机,如:神经计算机。
- 生物体结构与微光机电系统:DNA 驱动的微细机器人,生物大分子到细胞基本结构体系的自组装、自组织,创造新物质的分子工程学研究,分子聚集体的化学。
- 基因芯片、蛋白质芯片等。

而目前一般意义的生物信息学是基因层次的。

基因层次的生物信息学是近年来发展并完善起来的交叉学科。这门学科是综合运用生物学、数学、物理学、信息科学以及计算机科学等诸多学科的理论方法的崭新交叉学科。近年来随着快速序列测定、基因重组、多维核磁共振、同步辐射、机器人等技术的应用,生物学实验数据呈爆炸趋势增长,同时计算机和国际互联网络的发展使对大规模数据的贮存、处理和传输成为可能。现在某一实验室的基因研究成果一旦进入生物信息网络便为全球科学家所共享,生物信息学是内涵非常丰富的学科,生物信息学是把基因组 DNA 序列信息分析作为源头,在获得了蛋白质编码区的信息之后进行蛋白质空间结构模拟和预测,然后依据特定蛋白质的功能进行分析处理。其核心是基因组信息学,包括基因组信息的获取、处理、存储、分配和解释。基因组信息学的关键是"读懂"基因组的核苷酸顺序,即全部基因在染色体上的确切位置以及各 DNA 片段的功能;同时在发现了新基因信息之后进行蛋白质空间结构模拟和预测,然后依据特定蛋白质的功能进行药物设计。了解基因表达的调控机制也是生物信息学的重要内容,根

据生物分子在基因调控中的作用，描述人类疾病诊断、治疗的内在规律。它的研究目标是揭示“基因组信息结构的复杂性及遗传信息的根本规律”，解释生命的遗传规律。生物信息学已成为整个生命科学发展的重要组成部分，成为生命科学研究的前沿。当前，基因组信息、蛋白质的结构模拟以及药物设计有机地联系在一起，它们是生物信息学的3个重要组成部分，生物信息学目前已在理论生物学领域占有了核心地位，它广泛地应用在生物、医药、农业、环境等学科。

近来的研究表明，基因组不仅是基因的简单排列，它有其特有的组织结构和信息结构，这种结构是在长期的演化过程中产生的，也是基因发挥其功能所必需的。弄清楚生物体基因组特有的组织结构和信息结构，是了解生命遗传规律的关键。

一、生物信息学的研究内容

（一）生物信息的收集、存储、管理与提供

包括建立国际基本生物信息库和生物信息传输的国际联网系统；建立生物信息数据质量的评估与检测系统；生物信息的在线服务；生物信息可视化和专家系统。建立数据库是生物信息学的重要内容，当前在Internet上可找到的各种数据库几乎覆盖了生命科学的各个领域，核酸序列数据库有GenBank、EMBL、DDBJ等，蛋白质序列数据库有SWISS-PROT、PIR、OWL、NRL3D、TrEMBL等，三维结构数据库有PDB、NDB、BioMagResBank、CCSD等，与蛋白质结构相关的数据库还有SCOP、CATH、FSSP、3D-ALI、DSSP等，与基因组有关的数据库有ES-Tdb、OMIM、GDB、GSDB等，文献数据库有Medline、Uncover等，此外还有其他数据库数百种。另外一些公司还开发了商业数据库如MDL等。数据库内容呈爆炸性增长，除了在数量上的增长，数据库的复杂程度也在不断提高，它包括了大量注释、参考文献及软件，并通过指针将相关内容连接到其他数据库。数据库结构层次的加深客观上要求管理的进步，对数据库管理方法正在逐步取代旧的模式。在基因组相关数据库的发展中，建立基因组信息的评估与检测系统，实现数据标准化，基因组信息的可视化和专家系统的研究，发展次级与专业数据库，用户与数据库间迅速、有效地传递信息是基因组信息的收集、管理与使用一个要素，目前与基因组信息相关的数据库都有了自己的Internet地址和主页，同时在网上还出现了很多相关的在线服务器。

生物信息学各个领域中的软件数目庞大。并行算法、遗传算法、面向对象算法、并行虚拟机技术等已被应用到最新的程序中，生物信息学网络为用户提供更多的数据库服务。

生物信息学数据库覆盖面广，分布分散且格式不统一，因此，一些生物计算中心将多个数据库整合在一起提供综合服务，提供了数据库的一体化和集成环境，生物信息网格中的数据库服务广泛采用服务器-客户式结构，这些服务器包括为数众多的数据库搜索和序列对比服务器以及各专业领域的服务器，甚至有服务器将各搜索算法硬件化，实行并行计算和先进的内存管理，令搜索速度大幅度提高。

我国在基因组信息的收集与提供方面做了一定的工作：北京大学物理化学研究所建立了PDB数据库中国节点；北京大学生命科学院建立了EMBL数据库的中国节点；中国科学院生物物理所与日本JIPID合作，收集了我国科学家测定的DNA和蛋白质序列，并与国际相应数据库进行交流；此外，还有中国医学科学院肿瘤研究所建立的NEE-HOW服务器等。相信这一领域在我国会迅速发展。

(二)基因组序列信息的提取和分析

基因组信息学的根本任务是破译人类的遗传密码。迄今为止在人类基因组中真正掌握信息存储与表达规律的,只有 DNA 上编码蛋白质的区域,也就是基因。这部分只占人类基因组的 1%～3%,其余 97%的基因组序列人们尚不知其功能(所谓的"Junk"DNA)。"Junk"DNA 是许多对生命过程富有活力的不同类型的 DNA 的复合体,它们至少包含如下类型的 DNA 成分或由其表达的 RNA 成分:内含子、卫星 DNA、小卫星 DNA、微卫星 DNA、非均一核 RNA、SINEs 元件、LINEs 元件、假基因等。除此之外,顺式调控元件,如启动子、增强子等也属于非编码序列。目前普遍认为非编码区与基因在四维时空的表达调控有关。因此寻找这些区域的编码特征、信息调节与表达规律是未来相当长时间内的热点课题。因此基因组序列信息的提取和分析不仅要发现与确定新的基因,更重要的是发现存在于 95% "Junk" DNA 中的信息表达与调控规律,使用基因组信息学的方法是发现与鉴定新基因的重要手段。发现和研究新基因的生理功能或疾病的本质,可以为新药的开发、设计奠定基础。

可以利用 EST 数据库(dbEST)发现新基因,EST 表达序列标签(Expressed Sequence Tags)是从基因表达的短 cDNA 序列,它们携带着完整基因某些片段的信息。由于 EST 中包括了大量未发现的人类基因的信息,如何利用这些信息发现新基因成了近几年的重要研究课题。

近年来应用高维分布的统计方法、神经网络方法、分形方法等,将密码学方法用于识别编码区,从基因组 DNA 测序数据中确定编码区也取得了较好的效果。

人们熟知的碱基三联体数($4^3=64$)是大于 20(氨基酸的种类数)且最接近 20 的碱基组合,按这样推理由于结构单元是一一对应的,单考虑到 DNA 与 RNA 分子空间结构的复杂性,科学家们认为可能会存在其他的非三联体的编码方式,部分生物信息学家正在寻找其他编码方式。

Jacob 和 Monod 的乳糖操纵子模型给出了基因表达调控的最基本模式,但很多实验证据表明编码区和非编码区中信息调节规律更为复杂。

人类基因组及相关的模式生物基因组提供的大量信息,对于遗传密码起源的研究、基因组结构的形式与演化、生物进化等提供了有力的帮助。使生物信息学的研究拓展到更广泛的领域。

在基因组信息分析的方法研究方面,科学家们努力发展有效的能支持大尺度作图与测序需要的软件和数据库以及若干数据库工具,改进现有的理论分析方法,创建一切适用于基因组信息分析的新方法、新技术,建立严格的多序列比较方法。

(三)功能基因组相关信息分析

包括与大规模基因表达谱分析相关的算法、软件研究,基因表达调控网络的研究;与基因组信息相关的核酸、蛋白质空间结构的预测和模拟,以及蛋白质功能预测的研究。

在具有生命的细胞中基因表达的数量和种类是随着时间、环境的不同而变化的,所表达的基因相互之间有着必然的联系,基因之间的控制和被控制的关系构成复杂的网络关系。在分子水平对基因表达调控进行研究,研究调节蛋白是如何作用于它的顺式调控元件,像启动子、增强子等的;调节蛋白又是被什么调整的? 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。

蛋白质分子是由 20 种不同的氨基酸通过共价键连接而成的线性多肽链,每一种蛋白质在

天然条件下都有自己特定的空间结构。但以一定氨基酸顺序排列的多肽链是如何形成有一定空间结构的蛋白质分子的，也就是蛋白质结构的预测，仍是没有完全解决的问题。

分子模拟技术是利用计算机建立原子水平的分子模型来模拟分子的结构与行为，进而模拟分子体系的各种物理与化学性质。利用分子模拟技术结合计算机图形技术可以更形象、更直观地研究蛋白质等生物大分子的结构。蛋白质空间结构更清晰的表述和研究对揭示蛋白质的结构和功能的关系、总结蛋白质结构的规律、预测蛋白质肽链折叠和蛋白质结构是非常重要的。

蛋白质分子模拟技术主要借助于先进的计算机图形工作站，通过友好的图形环境，使用者可利用鼠标极为方便地建立多肽、蛋白分子的初始模型。同时，也可以对已经被测定的蛋白质分子的三维结构进行显示，并对这些结构进行灵活方便的平移、旋转、放大及缩小等操作，分子模型的建立为下一步进行的分子模拟以及了解结构与功能的关系打下了基础。

蛋白质结构预测的目的是利用已知的一级序列来构建出蛋白质的立体结构模型，对蛋白质进行结构预测需要具体问题具体分析，在不同的已知条件下对于不同的蛋白质采取不同的策略，目前预测蛋白质空间结构的方法可以分为两大类：

1. 分子动力学方法

采用分子力学、分子动力学的方法，根据物理化学的基本原理，从理论上计算蛋白质分子的空间结构，这类理论计算方法依据一个基本热力学假定：一个蛋白质分子在溶液中的天然构象是相应于热力学上最稳定、自由能最低的构象，但这一方法目前存在着3个主要问题，首先，用以描述蛋白质—溶剂系统工程力场和能量函数还处于半定量阶段；其次，数学上还没有有效方法解决能量极小化问题；第三，目前并没有证据证明蛋白质的天然构象就是全局自由能最小的构象。

2. 基于知识的预测方法

通过对已知空间结构的蛋白质进行研究和分析，找出蛋白质一级结构和空间之间的联系，总结出一定的规律并建立一些经验规则。这类方法已经被成功地应用于同源蛋白质空间结构预测的研究。然而对于同源性低的和非同源蛋白质分子来说，由于受二级结构预测精度的限制，这种方法的应用具有明显的局限性。

通过对大量已知空间结构的蛋白质分子的研究和分析，发现一条多肽链可能采取的构象的数目是相当大的，但在蛋白质分子中二级结构预测是解决从蛋白质的一级结构预测其空间结构这一问题的关键步骤，现有的预测方法都假定蛋白质的二级结构主要由邻近残基的短程相互作用所决定的，然后通过对一些已知空间结构的蛋白质分子进行分析、归纳，制定出一套预测规则，并根据这些规则对其他已知或未知结构的蛋白质分子的二级结构进行预测。目前常用的方法有：基于单残基统计的 Chou-Fasman 方法，基于信息论和统计的 Garnier 方法，Lim 方法，人工神经网络方法等。据一些检验结果，上述几种方法的预测率分别为 50%、56%、59%和 64%，而现在一般认为二级结构的预测准确率如果达到 80%的话，我们就可以基本准确地预测一个蛋白质分子的三维空间结构，因此进一步提高蛋白质二级结构预测的精度是非常必要的。

随着分子模拟技术的飞速发展，逐步形成了一些商品化的软件。应用于生物大分子领域的商品化分子模拟软件主要有美国 MSI 公司的 Insight II 软件和 Quanta 软件，以及 Tripos 公司的 Sybyl 软件；在国内，北京大学物理化学研究所也开发了一套“北京大学蛋白质分子设

计系统”。这些商品化软件在不断的变化和发展中，有些软件模块，每年都更新版本，不断完善这些软件的功能。

生物大分子结构模拟和药物设计包括：RNA（核糖核酸）的结构模拟和反义 RNA 的分子设计，蛋白质空间结构模拟和分子设计，具有不同功能域的复合蛋白质以及连接肽的设计，生物活性分子的电子结构计算和设计，纳米生物材料的模拟与设计，基于酶和功能蛋白质结构、细胞表面受体结构的药物设计，基于 DNA 结构的药物设计等。

分子图形学在药物研究领域被广泛应用，传统的药物研制主要是从大量的天然产物，如动物、植物、微生物和合成有机、无机化合物以及矿物中进行筛选。由于生物信息学的发展，相当数量的蛋白质以及一些核酸、糖类三维结构已被人们精确测定，使得基于蛋白质和核酸结构的药物设计成为可能。一般认为，任何有功能的蛋白质都可以作为蛋白质工程的改造对象，但实际上选择目标时往往要考虑以下几个方面：改性对象有没有测出空间结构；结构和生物功能的联系是否明确；所选对象的重要性；是否易于进行分子设计和最后的基因工程生产。当前的分子设计主要以能显示图形图像的计算机为工具，在了解了需要改造的蛋白质的性能及其相应的结构基础后，采用基于物理学原理的各种模拟方法及基于蛋白质分子结构知识的模型构建方法，提出蛋白质改性的设计方案。通过计算预测改性后的蛋白质的氨基酸顺序与天然蛋白质不同、新序列的空间结构和电子结构的变化，推论出新蛋白质生物学特性。

要了解蛋白质的功能，找到其致病的分子基础，只有氨基酸顺序是不够的，必须知道它们的三维结构。要设计药物治疗这些疾患也需要了解蛋白质的三维结构，目前的 X 射线晶体学技术、多维核磁共振波谱学技术等测定蛋白质空间结构的方法还不能很好的满足研究需要。因此，生物信息学中的理论模拟与结构预测就显示了重要性。模拟的结果对于在分子、亚分子和电子结构层次上了解生命现象的基本过程具有重要意义，为天然生物大分子的改性和基于受体结构的药物分子设计提供了依据。

另外，药物的治疗作用主要是通过药物与受体的相互作用而实现的。在生物体中受体多半是生物大分子，像蛋白质和核酸，而以蛋白质居多。如果人们了解了受体蛋白的结构，就可以根据其结构来研究药物是怎样改变它的构象、进而产生治疗作用的。目前已有很多生物大分子作为药物设计的受体模型，例如：基于酶结构的药物设计，基于抗体结构的药物设计，基于致癌、抑癌基因表达产物的药物设计，基于细胞表面受体结构的药物设计，基于转录因子结构的药物设计。随着人类基因组计划的进一步进行、化合物合成技术的进步和一些先进技术的使用已使受体药物筛选发展成为高通量筛选。利用生物信息学技术所建立的化合物库是筛选化合物的重要来源。

二、生物信息分析的技术与方法研究

随着分子生物学的发展，高通量、快速获取生物大分子生物信息的实验方法得到了飞速发展。DNA 芯片（DNA 微阵列）技术具有高集成度、高并行处理、可自动化分析的能力，因此它可对不同组织来源，不同细胞类型，不同生理状态的基因表达进行监测，获得基因表达的功能谱。与此同时，DNA 芯片还可用来进行 DNA 的快速测序，DNA 突变检测，药物筛选等。另外，生物功能的主要体现者是蛋白质，而蛋白质有其自身特有的活动规律，例如蛋白质的修饰加工、转运定位、结构变化、相互作用等活动，仅仅从基因的角度来研究是不够的，需要通过二维凝胶电泳、测序质谱技术、蛋白芯片、蛋白芯片-飞行质谱等新技术来研究。无论是生物芯片

还是蛋白质组技术的发展都更强烈地依赖于生物信息学的理论与工具。

生物信息应用与发展研究：汇集与疾病相关的人类基因信息，发展患者样品序列信息检测技术和基于序列信息选择表达载体、引物的技术，建立与动植物良种繁育相关的数据库以及与大分子设计和药物设计相关的数据库。

人们已经意识到，生物信息学的发展将对我们了解生命、了解人类自身，对于医药、保健、农业等都将起极大作用，因而生物信息学已引起世界各国的高度重视，纷纷加大投入，发展十分迅速。

近年来随着结构生物学的发展，相当数量的蛋白质以及一些核酸、多糖的三维结构获得精确测定，基于生物大分子结构知识的药物设计成为当前的热点。生物信息学的研究不仅可提供生物大分子空间结构的信息，还能提供电子结构的信息，如能级、表面电荷分布、分子轨道相互作用等以及动力学行为的信息，如生物化学反应中的能量变化、电荷转移、构象变化等。理论模拟还可研究包括生物分子及其周围环境的复杂体系和生物分子的量子效应。

但生物信息学的任务远不止于此。在以上工作的基础上，最重要的是如何运用数理理论成果对生物体进行完整系统的数理模型描述，使得人类能够从一个更加明确的角度和一个更加易于操作的途径来认识和控制自身以及所有其他的生命体。

参 考 文 献

1 陈润生．生物信息学．生物物理学报，1999；15：5

2 李维忠，王任小，林大威，等．国内外生物信息学数据库服务新进展．生物化学与生物物理进展，1999；26：22

3 王志新．蛋白质结构预测的现状与展望．生命的化学，1999；6：192

4 来鲁华．蛋白质的结构预测与分子设计．北京：北京大学出版社，1993

5 阎隆飞，孙之荣．蛋白质分子结构．北京：清华大学出版社，1999

6 吕秋军，高 月．受体药物筛选研究进展．中国药学杂志，1999；1：6

7 张亮仁．以结构为基础的药物设计与分子模拟．药物学研究与展望．北京：科学出版社，1999

8 朱 杰，张万年，周有骏，等．CoMFA 网格点的生成及选取对结果影响的研究．中国药学杂志，1999；34：417

9 赵雨杰，孙 啸，何志跃，等．基因芯片在医学中的应用．临床检验杂志，2000；18(6)：373

10 赵雨杰，陆祖宏，程 璐，等．基因多态性研究中基因芯片的应用．中华医学杂志，2000；80(9)：661

11 Hall-Alan-H. Computer modeling and computational toxicology in new chemical and pharmaceutical product development. Toxicology Letters Shannon, 1998; 102-103:623

12 Stoesser-Guenter. The EMBL nucleotide sequence database. Nucleic Acids Research, 1999; 27(1):18

13 Kraemer-Eileen-T. Molecules to maps: Tools for visualization and interaction in support of computational biology. Bioinformatics Oxford, 1998; 14(9):764

14 Vincens-Pierre, Buffat L, Andre C, *et al*. A strategy for finding regions of similarity in complete genome sequences. Bioinformatics Oxford, 1998; 14(8):715

15 Khalak-Hanif. Analysing biological molecules: Onward to function: (Computational Genomics II) (Virginia, USA; October 3-November 3, 1998). Trends in Biotechnology, 1999; 17(7):262

16 Jordan-Bertrand-R. 'Genomics': Buzzword or reality? Journal of Biomedical Science, 1999; 6(3):145

17 Scriver CR, Nowacki PM. Bioinformatics-rapid searching of sequence databases. Drug Discovery Today,

1999;4(10):482

18 Blaxter M, Bettomley S. Bioinformatics guide for evaluating bioinformatic software. Drug Discovery Today, 1999;4(5):240

第一章　Internet基础与生物医学文献检索

生物信息学是一门新兴的并正在迅速发展的交叉学科，是生命科学和信息科学尤其是分子生物学与计算机信息处理技术相结合而形成的交叉学科，是采用数学、统计学和计算机方法对生物学数据信息进行采集、存储、传播、分析、归类、解释的科学。该学科涉及生物学、计算机科学、数学、统计学、医学、化学等众多学科，其共同点在于任何类别的数据皆依赖于计算机处理、存储、分析。Internet网络是信息传输、检索、获取、交流的重要手段。因此，Internet的应用是生物信息学研究的关键。

第一节　Internet基础

Internet不是单一的计算机网络，而是建立在计算机网络之上的网络，是通过同一组计算机网络通信协议——TCP/IP(Transfer Control Protocol/Internet Protocol)将许多计算机局域网连接起来形成的巨型计算机网络。国内对Internet的称谓有很多，如：因特网、国际互联网、交互网等，Internet从20世纪60年代末诞生到80年代网络基本形成，目前已经发展为连接世界上170多个国家的国际性网络。作为一种新的信息传播媒体，因特网以其信息量大、数据更新迅速、文件传输速度快、使用方法简便等诸多特点，已经成为科研人员获取最新科技信息的主要途径。

一、IP地址与域名

IP地址(IP Address)是因特网上惟一标识计算机主机的数字地址。在因特网上，计算机相连接需要对每一台计算机进行标识，给每一个体计算机一个具体地址，从而使文件传输与交换得以实现。因此，所有与因特网直接相连的计算机都必须有一个IP地址。在同一时刻，一个IP地址只能标识一台计算机。IP地址是由4个8位二进制组以小数点相隔的数字串组成，每个8位组用十进制数0～255表示。例如，美国国立卫生研究院(NIH)生物技术信息中心(NCBI)的主服务器的IP地址为130.14.25.20。该数字组的意义从左到右依次为：主域(130.14表示NIH)、子网(.25表示NIH的国家药物实验室)及个体计算机(.20表示该计算机在国家药物实验室的编号)。

IP地址特指性强，但不便记忆。因此，网络设计人员设计了Internet域名系统(DNS)，DNS使得Internet用户可以用域名(Domain Name)取代IP地址代表主机。域名是以文字形式标识因特网组织结构的某个层次。IP地址通常都有相对应的正式域名，由域名服务器在后台将域名动态翻译成IP地址。相对于枯燥的数字来说，用户更喜欢利用域名在网上访问某一