王哲　主编

SHENGWU
XINXIXUE
GAILUN

生物信息学概论

第四军医大学出版社

生物信息学概论

主　编　王哲

编　者　王哲　王林　刘刚

主　审　杨守京

第四军医大学出版社

西安

图书在版编目(CIP)数据

生物信息学概论/王哲主编. —西安：第四军医大学出版社，2002.2
ISBN 7-81086-023-2

Ⅰ.生… Ⅱ.王… Ⅲ.生物信息论 Ⅳ.Q811.4

中国版本图书馆 CIP 数据核字(2002)第 002680 号

第四军医大学出版社
(西安市长乐西路 17 号 邮政编码:710032)
电话:029-3376765(发行部) 029-3376763(总编室)
传真:029-3376761 E—mail:fmmuP03@fmmu.edu.cn
第四军医大学印刷厂印刷
*
开本:850 ×1168 1/32 印张:7.35 字数:116 千字
2002 年 4 月第 1 版 2002 年 12 月第 2 次印刷
印数:1001～2000 册 定价:15.00 元
ISBN 7-81086-023-2/**Q**·1
(购买本社图书,凡有缺、损、倒、脱页者,本社负责调换)

序

生物信息学（Bioinformatics）是应用数理和信息科学的理论和方法研究生命现象，组织和分析日益剧增的生物信息数据库的一门新兴学科。它主要利用计算机、网络技术和不断发展的各种软件，研究遗传物质的载体 DNA 及其编码的功能大分子蛋白质，对逐日增多的序列和结构进行收集、整理、储存、发布、提取和加工，并从中分析和发现新的序列，从而不断揭示人体生理和病理过程的分子基础，为人类疾病的预防、诊断和治疗提供根本依据。实际上，生物信息学不仅已经成为生物医学、遗传学、农学等学科发展的强大动力，而且也为药物设计提供了有效途径。

随着人类基因组计划的不断发展，生物信息学的研究范围已从结构基因组学扩展到功能基因组学，随之又出现了进化基因组学。生物信息学的根本任务之一是发现新的基因、蛋白及其功能。生物信息学的特点是投资少，见效快，效益大，适合我国的现实条件。本书编著者是在生物信息学第一线工作的青年科学工作者，他们通过钻研与实践，已经基本掌握了如何从因特网上不断收集数据，并能进行分析、归类与重组，发现新线索、新现象和新规律，不仅发现并克隆了与肿瘤分化相关的新基因，并登录 GenBank，对有的新基因的功能也做了初步研究，并以此为基础获得了国家自然科学基金的资助。可贵的是，他们还把自己应用生物信息学的经验，在《生命科学》上介绍。为了加速我国生物信息学的不断快速发展，培养一批在数理、信息科学、计算机科学和分子生物学方面均有造诣的跨学科人才的任务十分迫切，愿本书能在这一方面发挥积极作用，吸引更多有志之士参与生物信息学研究，用不断发展的生物信息学推动我国生命科学的发展，发现更多具有我国自主知识产权的生物大分子，为我国科技创新做出贡献。

黄高昇

二〇〇二年三月于第四军医大学

序

苍宇时空无垠，科学前沿无涯。

近年来，随着分子生物学、人类基因组计划的快速发展，相应地产生了一门新兴的学科——生物信息学。它的出现是生命科学、计算机网络技术快速发展的必然结果，同时又对包括分子生物学、免疫学、神经科学在内的许多学科的发展起到了良好的促进作用。我们还欣喜地看到这门学科对科研思维、科学工作方法的扩展和改进都有助益。

但是，这毕竟是一门崭新的学科，有许多生物学研究者、临床工作者和青年学生对此不够了解。而在国内，系统地、深入浅出地介绍这方面知识的书籍很少见到。《生物信息学概论》一书的出版提供了极好的参考资料和学习读本，能够起到普及和提高的作用，使生命科学工作者受到这方面的训练和培养，使年轻学子易于掌握其基础知识和研究方法。

我校三位青年学者：王哲、王林、刘刚，近几年十分关注这一学科的发展。他们在完成各自研究课题的同时，悉心钻研，掌握了丰富的相关资讯。本书就是他们厚积薄发、大胆尝试之作。这是一本具有较高学术水平的参考书，它的出版无疑会对生物信息学的普及，以及生物学各领域的深入研究起到积极的推动作用。

我热忱地祝贺本书的出版，并向广大生命科学工作者，特别是青年学者推荐此书。

胡蕴玉

辛巳岁末

于第四军医大学

试读结束，需要全本PDF请购买 www.ertongbook.com

前 言

近十年，由于分子生物学在基因排列和蛋白质识别的研究上取得了可喜的进步，也由于对生物体功能和结构关系深入研究的必需，载录有数十亿数据信息的各类数据库需要有一个强有力的分析工具，用来描述数据与生物学意义之间的关联，用来收集、归纳、研究各类生物信息。这一工具就是生物信息学——一门传统生物学与计算生物学的交叉学科。

它的出现一方面是生命科学自身发展的需求；另一方面，信息科学、计算机及网络技术也为它的发展提供了理论支持和操作的平台。二者的结合使得对生物数据的演算、组织归纳和分析成为可能，并最终构架出具有生物学意义的本质内容。

如今，从事这一学科的研究开发、管理维护以及教学培训的专门人员已为数不少；应用这一工具为自己的科研服务的人就更多了。大致地，可以将他们分为 Doer 和 User 两类。前者是生物信息学的专业人员，包括各种研究机构（诸如：NCBI）的从业人员、大学里本专业的教研人员等等。他们中间有信息科学、分子生物学、结构生物学、计算机及网络技术、数学等方面的研究人员。而后者则是生物信息学的服务对象，包括生物学、医学、药学等学科的相关研究者。他们利用已建立好的各类数据库中的信息为自己的研究服务，同时也可能成为数据库的提交者和充实者。这本书就是为 User 提供基本知识的读物。

生物信息学的一个特点是发展速度很快。今天在网络上看到的东西已经与一年前有所不同了。形式上的不同仅是一方面，而更为重要的是内容上的变化。因此，写这方面的专著，常有跟不上变化的感觉；写成的东西也常常沦为“an Old Link”，而显得实用性不强。

在生物信息学的早期发展中，其变化固然多端；但在相对成

熟之后，其主要的形式和入口亦随之稳定下来。相关的方法学已详细地制定出来了，国际著名的一些数据库将会长期地发展下去。很多人发现，对生物信息学的基本内容有了相当的了解之后，追逐相关数据库的不断变化、进展，是一件令人着迷的事情。而这本书将就生物信息学的基础知识和最新进展做一系统的介绍。

这不是一本关于基因和蛋白质分析的实用手册，而是介绍基本概念、基本方法和生物学数据库最新资讯的专著。对于那些初入门的 User，这本书将是很有助益的。另外，本书的写作亦未过分简单化。其中的实用资料和解释，为研究者提供了有用的信息和帮助。

一年前，同学小聚。谈古论今之时，亦未敢遗忘正统学业。众人均对生物信息学有兴趣：言其发展神速，言其已使分子生物学进入了新境界，言其对研究方法、工作思维有深刻地影响，等等。深谈入巷，遂有著述之意。

其后的写作立即陷入了辛苦的套路之中，时常深感已入 harmless drudges 之境。然砥砺有成，今事随人愿。

但收获之余，有遗珠之恨；欣喜之际，有憾事不已。唯愿读者不吝赐教，以利我等不断地对此学问有新的领悟。

最后，感谢各位：

医学管理　医学硕士　王东光先生

病 理 学　医学博士　郭华章先生

放 射 学　医学学士　汤志华先生

病 理 学　医学博士　冯骥良先生

他们为本书的完成，提供了丰富的资讯服务和有益的信心支持。

作 者　谨　识

2001 年 10 月 2 日

于第四军医大学

目 录

试读结束，需要全本PDF请购买 www.ertongbook.com

第一章 概 论

当今世界，科学技术的发展日新月异。其中，生命科学的进展尤为引人注目。进入分子水平以来，人们发现在生物化学、分子生物学、免疫学以及遗传学领域的研究中有大量的数据资料需要处理。于是，随着计算机技术、网络通讯的飞速发展，产生了一门新兴的学科——生物信息学。它首先利用电子计算机技术，对在分子生物学等学科的研究中产生出来的大量原始数据进行收集、整理和管理；其次，对各种数据进行对比、分析、归纳并建立计算模型，以期更好地解释数据，并进行结构、功能的预测以及仿真，等等(图 1-1)。它的出现极大地推动了分子生物学的发展，在人类基因组计划的研究中发挥了重要的作用。这门学科在生物学、医学领域有着十分广泛的应用。其中的一些大型生物学数据库包含了众多的生物学信息资源，人们可以很方便地从国际互联网上寻找

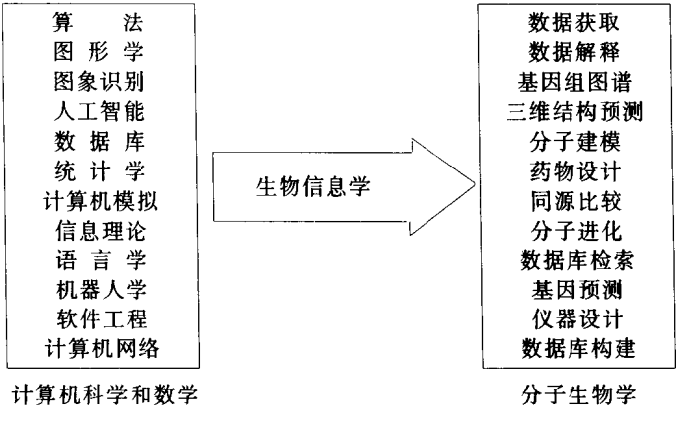

图 1-1 生物信息学是计算机科学、数学和分子生物学之间的桥梁

所需的资料和处理工具。这不仅方便了研究思想和资料的交流，减少了许多重复性的工作，而且也提供了一种崭新的思维方式和科研工作方法。近年来，互联网的高速发展为人们共享数据资源、合作研究提供了网络这一物质基础。越来越多的生物学、医学、药学工作者认识到生物信息学的重要性和实用性，其良好的发展前景业已显现。

第一节　生物信息学及其与生物学的关系

近十多年来，生命科学在分子水平上进行了广泛而深入地研究。随之而来的是大量的数据结果需要处理。特别是生物化学、分子生物学及遗传学的研究，各种各样的有关生物分子的原始实验数据，数量十分庞大。因此，利用计算机技术处理数据十分必要。另外，众多的学科，诸如结构生物学、酶学、细胞生物学、生理学、病理学、神经生物学等等，从不同角度的研究结果，可经过计算机的分类、组织和构建，形成具有生物学意义的新的研究结果。这些新的结果是对生命体的细胞结构和功能更为本质的反映。在这样的情形下，生物信息学应运而生。

生物信息学(Bioinformatics)的萌生可以追溯到1956年，那时还是计算机的初创期。在美国田纳西州的Gatlinburg，曾召开过首次“生物学中的信息理论讨论会”，这拉开了生物信息学的序幕。随着二十世纪八、九十年代计算机技术的迅猛发展，它才同时获得自身的快速成长。无论从理论上来讲，还是从现实情况来看，生物信息学都还是一门相当年轻的学科，它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。它的诞生和发展是应时所需，是历史的必然，并且已经悄然渗透到生命科学的每一个角落。以至于在整个科学界意识到它的存在之前，相关学科的研究者就已经离不开它了。

二十世纪末期，生命科学技术的迅猛发展，无论从数量上还是在质量上，都极大地丰富了生命科学的数据资源。数据资源的急剧膨胀首先迫使人们不得不考虑寻求一种强有力的工具，在有效地组织数据的同时，有利于对已知生物学知识的储存和进一步地加工利用。在大量多样化的生物学数据资源中，必然蕴含着许多重要的生物学规律。这些规律是我们解决许多生命之谜的关键所在。然而，继续沿用传统手段以人脑来分析如此庞杂的数据是不可能的。人们同样需要寻求一种强有力的工具去协助人脑完成这些分析工作。可以说，伴随着二十一世纪的到来，生命科学的重点和潜在的突破点已经由上个世纪的试验分析和数据积累，转移到数据分析及其指导下的实验验证上来。生命科学也正在经历着一个从分析还原思维到系统整合思维的转变。

那么，我们所寻求的那种强有力的数据处理分析工具，就成为未来生命科学的关键所在；伴随着生命科学这一需求的加剧，以数据处理分析为本质的计算机科学技术和网络技术获得了突飞猛进的发展，而自然地成为生命科学家的必然选择。计算机科学技术和网络技术正日益渗透到生命科学的方方面面，一门崭新的、拥有巨大发展潜力的生物信息学也就悄然而坚定地发展起来了。可以说，历史必然性地选择了生物信息学——生命科学与计算科学的融合体——作为新一代生物科学研究的重要工具。

生物信息学(Bioinformatics)这一名词的由来，还要从八十年代末期说起。美国佛罗里达州立大学超级计算机计算研究所的林华安博士认识到将计算机科学与生物学结合起来的重要意义，遂开始留意为这一新的领域构思一个合适的名称。考虑到与佛罗里达州立大学大型计算机计算研究所的关系，起初，他使用的是“CompBio”。当时，这一机构支持由他主办的一系列“生物信息学”的会议；之后，他又将其改为兼具法国风情的“bioinformatique”。因其拼写看起来似乎有些古怪，不久，他便进一步把它更改为“bio-informatics(或 bio/informatics)”。但由于当时的电子邮件系

统与今日不同，该名称中的‘—’或‘/’符号经常会引起许多系统问题。于是，林博士又将其去除。今天，我们所看到的“bioinformatics”就这样正式诞生了。林华安博士也因此赢得了“生物信息学之父”的美誉。

一、生物信息学的定义

生物信息学主要是由分子生物学与信息学、计算机技术、数学、物理学等学科交叉结合的产物。对于这样一门年轻的边缘科学，不同的学者对它的定义不尽相同(见附录二)。有不严格的定义称之为：分子生物学与计算生物学的交叉学科。国外学者一般认为，它是对现代分子生物学和生物化学技术带来的不断增加的复杂的资料进行分析、组织并使之系统化的一门科学。也有人认为，生物信息学应含有生物系统内信息链的内容，它主要指的是贮存于 DNA 或 RNA 中的信息，表现为核苷酸的序列并能通过翻译表达出重要的生命大分子——蛋白质。对这部分内容的研究，无疑是生物信息学在应用上的一个很重要的方面。我们认为生物信息学的含义是基于计算机和互联网的应用和信息科学的知识方法对生物信息进行收集、整理、分析研究、处理和应用的一门交叉学科。今天已经认识到：一项研究欲更加深刻地反映生物的本质规律，需要用到这门新兴的学科。例如，基因密码的含义与相对应的生物机体生理特点之间的关系、人脑的研究、基因与意识及心理行为的关系、系统遗传学家对各物种之间内在关系的研究等等，这类研究均需在计算机软件技术、各种不同类型的生物学数据库的辅助下完成。又如，结构生物信息学对靶蛋白质活性位点精细结构的描述可为新药的模拟设计提供良好的基础。总之，生物系统的复杂性需要生物学方法与计算技术的结合。所以，生物信息学是一门建立、管理并运用生物信息数据库研究生命现象，并最终模拟出生命有机体复杂性的科学。

一门学科的建立除了有应用上的需求外，还应当有相应的理论支持。信息学理论的发展是其重要的支柱之一。此外，计算机凭

借其强大的运算分析功能介入到生物学的研究中，使研究手段、工具方法迈上了新台阶。美国学者 H. Rashidi 和 L. K. Buehler 就认为生物信息学是建立在这样一个假设的基础上的：即基因结构、基因在基因组中的排列位置、蛋白质的功能以及在机体中引起能量代谢、繁殖和构成诸如身材、体型等蛋白质的相互作用之间存在着一个分等级的关系。而对其相互关联的研究，使人们意识到计算方法的介入为此提供了一个良好的平台。

二、生物学的发展与生物信息学

二十世纪初，人们用有机化学的方法研究三大物质的代谢途径，研究酶的组成及生理作用，等等。那时候，生物化学家没有分子生物学、基因的知识，并不知道核酸是生命的遗传单位。他们的研究是对各种实验现象的观测和记录。而时至今日，人们已经可以在电脑前完成基因测序、基因筛选、计算机识别蛋白质功能、计算机模拟蛋白质三维结构以及新药设计等工作，发展出计算和实验方法相结合的新的生物学研究模式。下面试举一例，来说明生物信息学在这一新模式中的用途。

基因是生命的遗传单位。在复制时，保持基因中分子信息的严密性和准确性是十分重要的。我们在研究某个基因突变与肿瘤发生的关系时，该基因的克隆是首先应完成的，因为这是获得核酸序列及寻找调节因子的第一步。首先，将我们需要的 DNA 片段从有关的基因组中分离出来，然后将这段基因插入到一个载体 DNA 中，从而制成重组 DNA。按生物进化的观点，所有生命体在遗传上是有密切的相关性的，所以人类基因在其他动物体或微生物体内操纵复制是完全有可能的。由此，人们将上述重组 DNA 置入细菌体内繁殖，从而达到基因克隆的目的并可复制出大量的基因拷贝。应用这种方法复制基因、扩增 DNA 简单而有效，而且避免了为纯化 DNA 或蛋白质而需获取大量的人体组织标本的过程。

基因克隆完成后，即可对基因测序。通过基因序列可预测其

试读结束，需要全本PDF请购买 www.ertongbook.com