

西安交通大学学术专著出版基金资助

医用生物信息学 概论

■ 楚雍烈 陈静宏 王治伦 主编

INTRODUCTION TO MEDICAL
BIOINFORMATICS



陕西科学技术出版社

西安交通大学学术专著出版基金资助

医用生物信息学概论

楚雍烈 陈静宏 王治伦 主编

陕西科学技术出版社

内 容 提 要

21 世纪是生物生命的世纪。生物信息学是一门新兴的交叉学科,已成为生命科学的重大前沿领域,它对生物医学的发展有着极大的推动作用。本书从生物医学的应用角度对生物信息学进行了概述,分别介绍了生物信息学的产生基础、基本概念、主要生物医学信息资源、生物信息学关键技术、生物大分子的分析方法和医学研究中的常用技术。为便于学习方便,每个章节后附有复习思考题,最后一章为实习指导,并在书后附有常用英语词汇、生物医学信息学相关网址和数据库。本书面向生物医学专业各个层次的读者,是一本实用性较强的教科书,有助于在生命医学领域中生物信息学的推广、普及、研究和应用。

本书由西安交通大学医学院教师和博士生在多年教学和科研实践的基础上编写而成,内容新颖、简明扼要、概念准确、条理清晰、注重实用,可作为生物医学各专业本科生、研究生的教材和教学参考书,同时也是从事生命科学、医学各专业科研工作者的必备工具书。

图书在版编目(CIP)数据

医用生物信息学概论/楚雍烈,陈静宏,王治伦主编. —西安:陕西科学技术出版社,2005.10

ISBN 7-5369-4007-6

I. 医... II. ①楚...②陈...③王... III. 医学—生物信息论—高等学校—教材 IV. R318.04

中国版本图书馆 CIP 数据核字(2005)第 114926 号

出 版 者 陕西科学技术出版社

西安北大街 131 号 邮编 710003

电话(029)87211894 传真 (029)87218236

<http://www.snstp.com>

发 行 者 陕西科学技术出版社

电话(029)87212206 87260001

印 刷 陕西新世纪印刷厂印刷

规 格 787mm × 1092mm 16 开本

印 张 16.25

字 数 358 千字

版 次 2005 年 10 月第 1 版

2005 年 10 月第 1 次印刷

定 价 21.00 元

版权所有 翻印必究

编 委 会

主 编 楚雍烈 陈静宏 王治伦

编写人员(以姓氏笔画为序)

卢 阳 代晓霞 刘文康

安 润 杨 娥 陈 萍

张 帆 答 嵘 楚 霆

前 言

现代科学以前所未有的速度向前发展,大量的生物医学信息需要收集、处理、存储、传递、分析、解释和应用,生物信息学(Bioinformatics)应运而生。生物信息学是一门新兴的交叉学科和覆盖面极广的综合性学科,它横跨了物理、数学、信息学、生物生命科学,特别是分子生物学和分子遗传学。其应用不但对促进生物医学科研水平的提高具有重要的理论意义,而且能产生极大的社会效益和经济效益。

生物信息学是一门实践性很强的学科,其中涉及许多计算机和互联网的具体知识,需要经常实践来熟练掌握关键技术手段。生物信息学已成为当今每个生物医学工作者的必修课和基本技术,成为彼此相互学习、相互交流必不可少的现代工具,成为科学研究中最能达到事半功倍的技术手段。作为一个现代生物医学工作者,若要赶超世界先进水平、作出创新和领先水平的成绩,就必须学习、掌握生物信息学的知识和技术。

应广大医学院校师生的迫切需要,我们在三年前为医学生开设了生物信息学选修课程,效果良好。为进一步满足广大青年教师,特别是医学研究生的要求,加快生物信息学在我国医务界的推广和应用,我们结合自己的教学和科研实践编写了《医用生物信息学概论》。

全书包括了生物信息学的基本理论和基本概念,介绍了生物信息学最新研究成果和观点,其相关技术方法,特别是分子生物学中常用的关键技术及常用软件。本书1~4章在对生物信息学、医学信息学的概念、产生、发展和研究内容作了介绍之后,简要地对生物信息学的分子生物学基础、计算机和互联网知识作了介绍;5~7章则主要介绍生物信息学的构成成分、搜索工具、浏览软件、著名的生物信息学中心和常用数据库;8~13章重点介绍生物信息学的应用内容,如生物信息资源的获取、核酸序列、基因组、蛋白质组分析及在基因芯片和现代制药中的应用等;考虑到生物信息学实践性很强的特点,结合生物医学实际,还设计了一些实验供学生练习操作,以巩固所学的知识和技术。应该强调,初学者一定要经常结合自己专业,在互联网上不断地实践、提高自己解决实际问题的能力。为了方便使用者学习和操作,书后附录列出了常用的有关生物信息学常用词汇和缩略语、生物信息学及医学信息学网址和数据库,供使用者选用和参考,希望能有利于大家学习和工作。

本书力求做到简明扼要、深入浅出、通俗易懂和新颖实用。作为概论,只能对经典生物信息学的基本内容作一般性介绍,未能对具体内容详尽地展开。由于本书的对象主要

针对医学专业教师、学生和医务工作者,重点是在医学教学、科研和医疗实践中如何使用,对其理论算法、数据处理、软件设计和数据库的建立未能涉及,有兴趣者可参阅有关专著。

在编写过程中我们参考了国外有关专著、文献和北京大学、清华大学、复旦大学及中科院的有关专著和教材,在此表示衷心的感谢!

生物信息学内容新颖、涵盖面广,信息数量巨大,正以前所未有的速度向前发展,需要我们随时学习和掌握,希望使用者掌握基本知识和技术后,不断地学习和补充新的知识。

由于我们的水平和能力有限,编写中难免有不足之处,欢迎批评指正。

楚雍烈

2005年5月于西安

目 录

第一章 绪论	1
第一节 生物信息学概述	1
第二节 信息学基础	10
第二章 医学信息学基础	15
第一节 医学信息学概述	15
第二节 医学信息学的研究与应用	22
第三节 医学信息学发展趋势	24
第三章 生物信息学的计算机和互联网基础	30
第一节 生物信息学的计算机基础	30
第二节 生物信息学的互联网络基础	32
第四章 生物信息学中的生物学、分子生物学	43
第一节 生物信息学中的生物学、分子生物学基础知识	43
第二节 基因工程和核酸研究技术简介	52
第五章 生物信息学信息资源检索的常用工具	56
第一节 概述	56
第二节 通用型检索工具	59
第三节 医学专业搜索引擎	66
第四节 生物医学信息检索工具的综合利用	71
第六章 生物信息学网上浏览及网络文件传输	73
第一节 浏览器及其应用技巧	73
第二节 文件压缩与解压	77
第三节 文件传输与 FTP	78
第四节 网络蚂蚁	80
第五节 离线浏览工具 WebZIP	82
第六节 文档阅读工具 Acrobat Reader	84
第七节 图像浏览软件 ACDSec	85
第七章 生物医学信息学的重要信息中心和主要数据库	88
第一节 重要生物信息中心简介	88
第二节 生物大分子序列核心数据库	92
第三节 生物信息学中常用的三大数据库	95
第四节 其他重要的生物医学网站介绍	104

第八章 生物医学网络资源及获取	107
第一节 网络生物医学资源	107
第二节 网络文献数据库检索	113
第三节 特殊网络医学资源	118
第九章 核酸序列分析	131
第一节 核酸序列的检索	131
第二节 核酸序列的基本分析	132
第三节 核酸序列比对分析	135
第十章 基因组信息分析	142
第一节 基因组生物信息概述	142
第二节 基因组信息分析	146
第三节 DNA 中的功能性位点搜索	150
第十一章 蛋白质序列分析和蛋白质功能预测	153
第一节 蛋白质和蛋白质组研究的意义	153
第二节 蛋白质序列检索	156
第三节 蛋白质序列基本性质分析	158
第四节 蛋白质功能预测	160
第五节 蛋白质结构预测与结构生物信息学	163
第六节 后基因组学、信息学和分子相互作用简介	169
第十二章 生物信息学与基因芯片	176
第一节 基因芯片概述	176
第二节 生物信息学和基因芯片的相互关系	182
第三节 生物信息学在基因芯片技术应用中的举例	185
第四节 基因芯片制备和应用中的原则及注意事项	191
第十三章 生物信息学与生物制药	195
第一节 概述	195
第二节 生物信息学在新药研发中的作用	196
第三节 生物信息学推动了药物基因组学的发展	203
第四节 未来药物研究方法展望	210
第十四章 医用生物信息学教学实习	216
实习一 生物医学网络原始文献的获取	216
实习二 核酸序列的搜索与分析	218
实习三 PCR 引物设计的原理和方法	221
实习四 基因组分析	225
实习五 蛋白质序列分析	227

附录	230
附录一 常用缩略语及词汇表	230
附录二 分子生物学与分子遗传学相关数据库表	238
附录三 生物医学有关数据库表	242
附录四 生物医学文献数据库表	247
参考文献	248

第一章 绪 论

第一节 生物信息学概述

一、生物信息学的重要地位

1. 对生物信息学的认识

生物信息学被誉为 21 世纪最具发展前景的学科。它是 21 世纪生命科学、信息科学的综合体现和结晶,是二者结合必然的产物。生物信息学又是未来自然科学的核心和前沿学科,是解决生物生命本质的基础。生物信息学把数学、物理、信息技术和生物科学紧密联系在一起,是科学研究中的一个令人心醉着迷的领域。许多重大的发现和奥秘不断地鼓舞人们涉足,并随之引发了一场涉及多个领域的深刻革命。现代生命学科早已从单纯观察进入到实验和观察的研究基础上,进而发展到观察、实验和理论阶段,并向更高级的观察、实验、理论和信息学阶段发展。生物信息学已使现代生命生物学科进入以观察、实验为基础,以理论和计算为推动力,用信息、计算机及网络为双翼的全方位腾飞的新时代。

2. 生物信息学的意义

(1) 生物信息学是生物学发展史上的里程碑。生物信息学属新学科、新技术、新理论、新方法,是生物学的崭新阶段。生物信息学是 21 世纪生物工作者最基本的手段和必修课,本世纪的生物医学生不懂生物信息学,就不可能成为一个好的生物学工作者,就要落伍或者被淘汰。

(2) 生物信息学是一个新型的交叉学科。生物信息的广泛渗透首先表现在生物学间的自身相互联系,如动物、植物和微生物之间的知识融合,农、林、畜、牧、渔、海洋和医学诸学科间广泛交织;生命信息学与数学、物理、军事等学科广泛的交叉。

(3) 生物信息学是人类共享的资源。生物信息学让人类真正做到资源共享、随意交流,共同受益。生物信息学真正做到了发挥国际“团队”精神,解决世界面临难题。它改变了人类的意识和思维方式,使人类的思维不可能再封闭,由线性思维方式向着二维、三维(空间)、四维(时间)等高级思维方式发展;由单纯的纵向思维方式朝着纵横结合、多向联系,到网络思维方式发展。

利用生物信息学能做出重大理论的突破,能产生巨大的经济及社会效益。如在结构与功能、分子运动、生命起源、物种关系、生物进化、人类保健和医疗以及药物开发等方面,都具有巨大的应用潜能。

二、生物信息学的产生背景

1. 生物信息的快速增长

生物信息从逐年有限量的增加,已发展到指数性爆炸性的增加,生物信息爆炸性的增长速度令人刮目相看。例如生物大分子核酸序列的碱基对(Base Pair, bp)数量,从早期只能手工测序,发展到自动化仪器测序(500bp/次·人)和基因芯片测序,进而发展到机器人测序,使得近年来 GenBank 中核酸序列的信息每半个月就翻一番。近几年来,每年有大于 2 亿的信息进入数据库,单是一个典型的基因测序中心,每年即可产生 10^{15} 字节的原始数据。当今,依据计算机和互联网技术的发展,充分利用其数据库、算法和应用软件,进行广泛的信息交流,达到资源共享和共同发展已成为生物医学工作者的重要使命。

当前,人类社会实践中能够产生数据量达 10^{15} 字节的有三个领域:一是高能物理实验,二是大脑活动成像,三是生物数据。生物数据量已超过前两者的($>10^{15}$ 字节)数据量。据估计,人类(包括已经去世的和仍然在世的)所说过的话的信息总量约为 5×10^{18} 字节。而如今生物学数据信息总量已接近甚至超过此数量级。生物信息学 GenBank 数据库中的碱基对数急剧增长,2000 年已大大超过医学论文数,人类已来不及消化这些信息并使之上升为理论知识。

2. 巨大的生物信息量

生物信息不但增长速度惊人,其信息的总量也达到让人类叹为观止的地步,表现在:

(1) 人类基因组计划已完成 3×10^9 bp 的测序工作,加上相关信息共约 500 亿信息的资料。该计划的实施和完成,促进了生物信息学的形成,而后基因组计划的提出,又对生物信息提出了新要求,丰富了其内容,促进了结构和功能基因组学的发展。

(2) 出现了大量的蛋白质信息。蛋白质的一级、二级、三级和四级结构及功能信息和其他不断增加的大分子信息使蛋白质信息快速增长。

(3) 12 种模式生物信息包括噬菌体、病毒(SV40, HIV 等)、大肠杆菌、酿酒酵母、线虫、果蝇、拟南芥(小草)、水稻、非洲爪蟾、斑马鱼、小鼠、家鼠、猪和狗的基因组计划已完成。

3. 高度复杂的生物信息

爆炸性增长和海量的生物信息又具有高度复杂性,其四维的特征和网络般的广泛联系,人类利用常规方法已难以胜任,生物信息已经令科学工作者面对错综复杂、眼花缭乱的复杂信息无能为力、一筹莫展,表现在以下几点:

(1) 生物大分子包括核酸、蛋白质、糖和脂四大种类,它们之间的组合在一级结构上就千变万化。

(2) 生物大分子三维立体结构的复杂性,特别是其空间构象更是千变万化。

(3) 生物信息的四维(空间、时间)观念,使分子运动变化和机体发育、组织的分化更为错综复杂。生物大分子之间的相互作用又构成了生物信息的“多维”概念。

(4) 各种信息间的联系:核酸、基因、基因组、蛋白与疾病关系的多样性,结构与功能的复杂性(生理与病理),使各种信息间联系错综复杂。

(5) 各种信息与机体整体关系,与生物进化、发育和各种生命现象的关系更为复

杂等。

上述五点已成为当今生物学科发展的瓶颈和重大障碍,到了山重水复地步,必须突破进入柳暗花明的持续发展新阶段。

4. 信息学的发展为生物学的进展提供了机遇

信息学的发展为生物学的突破性进展提供了机遇。计算机可以高速度(如数十倍光速收集)、高贮量(数百 G)、高效率(整理,加工整合)地处理信息,完全可以满足生物信息数据的采集、贮存、整理、加工和使用。

人类在科学研究中要求资源共享和充分交流,这加速了生物信息学的发展,互联网的诞生、互联网的高速传递(光纤,宽带和无线)、互联网广泛联系的网络系统和各种协议(IP 等),使人类生物信息得以广泛交流、人人使用,资源共享。

21 世纪是人类全面进入生命科学的世纪,是一个信息革命的时代。它对现代信息学提出了更高要求,一门新兴的交叉学科——生物信息学应运而生。

5. 生物信息学是生物学科发展的需要

人类面对客观世界遵循着这样一个规律:从了解和认识客观世界,逐步发展到利用和改造客观世界,最后要上升到创造客观条件为自己谋福利的阶段。科技工作者依据此理念,大大扩展了人类生命活动的范围,生物信息学的功能在满足人类需求上显得十分必要。因此,生物信息学迅速发展成实用性极强的现代科学工具,被推上令人瞩目的地位。

生物生命学科的研究首先从整体水平研究开始,继而逐步深入到人体的各个系统、器官、组织、细胞和亚细胞水平,最后深入到分子水平,实现了生物学的第一阶段大飞跃,取得了以完成人类基因组计划为指标的令人瞩目的成就。

通过分子水平的长期研究,人类经过不断反思,发现这种研究思路一直在还原论的指导下进行,虽然取得了很大进展,但获得的结果不能完全揭示生命现象的本质和规律。因为生命不是原子和分子,它必须在完整的、不断变化的活细胞中才能表现出来。此外,后成遗传学(表观遗传学)和系统生物学的诞生,使人类必须回到以功能、系统、完整和动态的研究四项原则上来,生命科学的研究必然过渡到从简单的分子联系到完整机体,从结构联系到功能,从静态联系到动态,从个体联系到群体,从单个细胞联系到复杂的组织分化和发育,从单一系统功能联系到多个系统功能的时代。这不是简单的轮回,而是生命科学研究第二发展阶段,是新层次上的创新和飞跃,产生了全新意义上的代谢组学和系统发育生物学。20 世纪是微观世界的黄金时代,是生物医学在分子水平上大发展的时代。21 世纪则是宏观与微观相结合大发展的新时代,生物医学则是分子水平与细胞、系统和整体相结合发展的新时代。

三、生物信息学的诞生与发展

1. 计算机分子生物学

20 世纪 70 年代、80 年代,由于 DNA 测序技术的问世,DNA 序列数据呈指数性增加,大量信息需要处理,计算机开始应用于生物学。此时一批数学和计算机专家加盟了分子生物学研究,进行资料收集、贮存、管理和处理数据。在 20 世纪 80 年代中叶,形成了计算

机分子生物学(计算生物学, Computational Biology),这是生物信息学的前身。在英语字典中原本没有 Bioinformatics 这个单词,它是一个新增加的学科名词。20 世纪 80 年代末期,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,为这一领域构思了这个合适的名称,他也因此赢得了“生物信息学之父”的美誉。

长期以来,生物信息学与计算生物学被混同使用。但计算生物学仅指将计算机系统和计算机运算法则应用于生物现象的模型研究。生物信息学中计算机技术仅作为研究手段和工具,同时采用数学、统计学的模型、模拟研究手法来解决生物科学的问题,因而成为生物学、统计学、数学、计算机学和工程学等多学科之间的交叉领域。生物信息学的发展一方面依赖于这些学科的发展,但它在自身发展的同时,又对相关学科提出新的要求和挑战,不断为其开辟新的研究领域并促进其发展。

2. 生物信息学的诞生

1990 年,人类基因组计划(HGP)正式启动,新技术发展迅速,比如,测序、克隆技术工作的速度较前快了 10 倍,出现了超级智能计算机等现代化的信息工具。这些问题使得用信息学观点和方法去研究生命信息成为一门科学,催生了早期的生物信息学。

3. 现代生物信息学的诞生与发展

(1) 基因组计划对生物信息学的发展起到了推动作用。

随着 HGP 广泛与深入地进行,其他生物基因组研究也相继开始启动,DNA 序列数据库(GenBank)的信息每半个月就翻一番。与此同时,蛋白质序列测量速度也大大加快,其三维空间研究资料飞速增加(蛋白氨基酸库中已有 2400 万氨基酸,6000 多种晶体结构)。HGP 的主要任务有两个:首先,“读出”HGP 全部 ATCG 语言,即对人基因组作图和测出全部核苷酸序列,主要工作属于结构基因组学的内容;再者就是“读懂”HGP 全部 ATCG 语言,了解 HGP 中全部基因的编码及功能,属功能基因组学内容。

传统生物信息学主要对结构基因组学信息进行研究,还不能很好对功能基因组学信息进行管理和研究。现在“后基因组计划”的提出,使生物学家不满足于只对序列的认识、基因结构及基因表达的注意,更想深入了解其调控、产物及其功能在生命活动中的作用。此时单纯的传统计算生物学已不能满足高通量和海量信息的需要,向生物信息学提出了更高要求,要对生物信息进行分析和解释,即必须对信息的结构(由低级到高级)和功能做出明确的说明。

(2) 生物医药工业推动了生物信息学的发展。

由于生物医药工业的发展需要生物信息学的支持来开发新药,如生物信息学海量的信息可提供大量原材料、药物作用的靶点、高通量分析和筛选的方法及基因药物的开发,许多公司投入了大量的资金,不但产生了经济效益,也极大刺激了生物信息学的发展。

(3) 信息学的自身发展促进了生物信息学的发展。

信息必须共享与交流。要完成 HGP 任务不是某一个国家和几个实验室能完成的工作,它需全世界范围合作和全人类共同努力来完成,需要包括生物、物理、数学、计算机信息和通讯诸多学科协作,生物医学、工业、农业等多部门的联合。目前已有 200 多个生物医学信息库及繁多的处理软件在全球互联网络上公开。由于现代生物医学进展在很大程度上将依赖计算机和信息技术,所以生物信息学的另一重大任务是把处理好的信息方

便快捷地传播,使人类共享,面对面地交流。例如,远程教育和远程会诊将越来越广泛的为人类健康服务。生物学与信息学更加紧密结合,相互促进共同发展,促使形成了旨在收集、检索、管理、处理、分析、利用和开发这些巨型生物信息的科学——生物信息学。

总之,生物信息学是一门全新的交叉学科,它不同于生物计算和生物统计,不同于计算机应用和计算机生物学,也不同于结构生物学和分子生物学。

四、生物信息学概念

1. 经典生物信息学

早期狭义的定义指用计算机的手段和方法来对生物遗传信息和数据进行管理和分析。后来则指利用计算机、应用数学和信息学理论、方法,对生物大分子(核酸和蛋白质层次)信息进行收集、贮存、分类、检索与综合分析,研究生命现象,以达到认识理解生物信息的生物学意义的科学。

2. 现代生物信息学

现代生物信息学是指利用包括信息学、计算机学等现代技术,在对生物大分子信息进行收集、贮存、分类、检索、交流以及在综合分析生命信息基础上,深入了解其三维结构和生物功能,并对其进行开发利用的一门新型科学。

两者相比,现代生物信息学有了质的飞跃:

(1)在获取生物信息手段上,不单是利用计算机,还包括分子生物学和生物物理学现代技术,如X射线晶体衍射、核磁共振、质谱、扫描隧道显微镜、DNA芯片、蛋白多肽芯片等技术。

(2)在获取生物信息深度上,它不但研究生物大分子一级结构,还要研究复杂的高级结构。

(3)在获取生物信息广度上,它把序列资料与三维结构和生物功能相联系。

(4)在获取生物信息功能上,扩展到对信息的解释(理解)、交流、综合分析和开发利用。

现代生物信息学是信息学、计算机学、应用数学(数理统计学)和生物生命学科的融合,是许多物理、信息和化学分析新技术以及分子生物学研究方法等生物技术的融合。

3. 广义生物信息学

广义生物信息学的定义泛指任何运用计算机及其网络和数据库进行大规模生物学数据的收集、组织、管理、分析和应用以及相关的所有生物学的工作,涉及生命科学全部研究领域的内容。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),它以崭新的理念和现代的先进技术吸引着科学家的注意。

生命现象是信息控制下的不同层次上的物质、能量与信息的活动。生命活动过程是指不同层次上活动的综合表现,如机体的核酸、蛋白质、细胞、器官、系统、整体等不同水平上的代谢活动和生命现象的综合。广义生物信息学主要包括如下几个方面的生物信息:

(1)遗传生物信息学:包括生物遗传信息的物质基础、结构、功能和传递。

(2)细胞生物信息学:包括细胞增殖、周期、分化、发育、凋亡、死亡的信息。

(3)整个有机体生物信息学:包括器官、系统、整体活动的调控节律、生物钟、生长、代谢,营养的吸收、传输、转化及对外界信号反应等遗传表型的呈现。

(4)生命活动的调控信息:包括基因的表达和调控,蛋白的结构、功能和生物活性,代谢信息。

(5)人体信号的生物信息学:包括生命活体在不同层次(基因、细胞、组织、整体等)的活动和不同属性(包括思维、精神)、生命活动时和外界环境(生命体周围至宇宙)相互作用时反映出来的各种信号信息。

(6)脑和神经系统与感觉系统的生物信息:包括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。视觉、听觉、嗅觉、味觉和感觉的生物信息。

(7)高通量医学生物工程设备和技术获得的信息:包括如基因芯片、蛋白质芯片和各种大型医疗设备等获得的种种信息。

需要强调的是,目前的生物信息学主要是基因组层次上的生物信息。生物信息学是把基因组 DNA 序列信息分析作为源头,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 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。总之,生物信息学的目的是研究和分析生物大分子信息结构和功能的科学,是以解释和认识生命的起源、进化、发育、遗传和变异的科学。

五、生物信息学研究内容

1. 生物信息学的研究对象

医学生物信息学的研究对象是人类所有生物大分子的序列、细胞、组织、器官的结构和功能,在机体生理、病理和药理过程中的生物信息。狭义的生物信息学研究对象为生物生命大分子序列、结构和功能。

2. 生物信息学的研究任务

生物信息学研究的任务具体包括:

(1)对各种生物数据库的建立和管理:收集、存贮、综合、分析。

(2)供人类共享、查询、交流、使用生物数据和构象模块,开发软件。

(3)从海量数据中,通过软件,分析计算提取新知识、新理论,开发新的产品。

(4)生物体活动聚群关系,基因表达的时空性,体内生化反应器的耦合关系。

生物信息学的根本任务是综合、分析、解释和利用生物学数据和构象;是让人类理解生物信息的本质和意义并能交流、应用、开发和创造。它把信息海洋中复杂和无序的信息数据转化为知识,能把单纯信息转化为可利用和可以再生的资源,为人类疾病的预测、诊断、预防、治疗和健康保健提供合理和有效的服务的学科。

3. 生物信息学的研究内容

生物信息学的研究内容包括:

- (1) 生物大分子(核酸、蛋白、脂肪、糖类等)结构、序列、构象和其特性有关资料的收集,存贮和分类(机体、器官、组织、细胞及亚细胞结构)。
- (2) 基因组结构和功能:用模式识别等软件分析其结构(转录谱和蛋白表达谱)、预测结构和功能;新基因的发现,基因组比较及多态性研究。
- (3) 生物技术运用:基因工程,蛋白质工程药物,大分子药物开发,改造和创造。
- (4) 大分子模拟动力学(如蛋白质折叠三级、四级结构)。
- (5) 生物统计、智能软件、新型光子计算机、量子计算机、生物计算机的开发。
- (6) 生物科学网络技术,生物信息传输、利用及使用。
- (7) 遗传密码的起源、生命起源与系统进化。

4. 生物医学信息学应用上的四个层次

- (1) 查阅生物医学文献资料数据库中的文献、书籍、资料和各种数据,更新知识。
- (2) 利用生物信息学平台,开展国内外信息交流,能分析生物大分子的结构和序列,做出预测。
- (3) 应用资料、信息和软件,对所需的大量生物数据进行生物医学、遗传学信息的分析和统计。通过补充新知识,解决科学研究中的实际问题。能自主设计实验,开展研究。
- (4) 具备上述三点能力,能够自我编程,建立数据库。可以改造、开发生物信息软件,进行新药设计,获得创造性结果和产品。

作为 21 世纪的生物医学工作者,必须掌握生物信息学的知识和技术,最起码要进入第一、第二层次;而当代的医学研究生应该进入到第三、第四层次,利用它做出创新性的科研成果。

5. 现代生物医学信息学研究的基本问题

现代生物信息学研究的基本问题已经从生物数据库的建立和 DNA 序列的分析阶段,发展到基因组生物信息学,目前又提升到后基因组时代的以多个组学为中心的新阶段。现代生物医学信息学研究的基本问题主要涉及:

- (1) 医学和生物学数据库的建立和搜索。

目前,除了已建立的 DNA、蛋白质和三维结构的三大系列数据库外,前后又建立了许多文献数据库和各种关系数据库。数据库不但是生物信息学存在和发展的基础,又是医学和生物工作者的知识源泉和材料宝库。今后,发展综合型数据库和建立功能强大的浏览器和搜索工具仍是生物信息学的重要任务。

- (2) DNA 和蛋白质序列分析,序列分析是生物信息学的基本任务和日常工作。
- (3) 预测基因组的基因结构,发展高智能化的运算法则和工具。
- (4) 预测蛋白质的结构和功能,主要依赖于结构生物学的大量数据。
- (5) 基因组数据的分析,提高基因组组装的效率和准确性,发展新的软件。
- (6) 比较基因组学和系统发生遗传学,了解生物多样性、个体差异性和多态性。
- (7) 分析功能基因组和蛋白质组学数据,预测大分子序列的结构和功能。
- (8) 了解代谢、信号传导和基因调节的途径、相互作用和相互调节的关系。

(9)分析脑和神经系统的功能和学习认知行为的本质。

(10)疾病发病的分子机制、疾病的基因诊断、治疗、预防和人类保健。

六、生物信息学研究的流程和构件

1. 生物信息学的流程(见图 1-1)

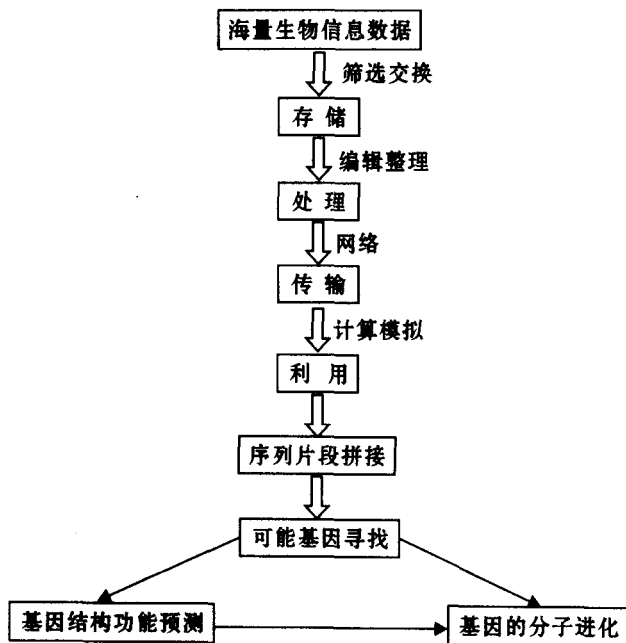


图 1-1 生物信息流程图

2. 生物信息学研究构件的三要素

生物信息学研究的支撑构件首先是数据库(实验室),其次是计算机及应用软件包,再者是互联网(计算机网络)。生物信息学除了收集所有生物信息外,离不开数据库建设、管理检索、模式识别、软件开发和网络技术。它是分子生物学、计算机学、数学、医药学等多学科的交叉学科。

世界一流的生物信息学中心包括:① 计算机中心,中继站、放大器、加工厂等;② 网络中心数据库;③ 分子生物学实验中心实验室;④ 结构研究中心(X线衍射,核磁共振,STE及质谱室等)。

七、生物信息学的作用

生物信息学以核酸、蛋白质等生物大分子数据库及其相关的图书、文献、资料为主要对象,以数学、信息学、计算机科学为主要手段,以计算机硬件、软件和计算机网络为主要工具,对浩如烟海的原始数据和原始资料进行存储、管理、注释、加工,使之成为具有明确生物学意义的生物信息。通过对生物信息的查询、搜索、比较和分析,从中获取基因的编码、调控、遗传、突变等知识,研究核酸和蛋白质等生物大分子的结构、功能及其相互关