药物生物信息学

郑 珩 王 非 编著

化学工业出版社
现代生物技术与医药科技出版中心
·北 京·

需要全本请访问 www.ertongbook.com

(京)新登字 039 号

图书在版编目(CIP)数据

药物生物信息学/郑珩,王非编著. —北京:化学工业出版社,2004.3
ISBN 7-5025-5318-5

Ⅰ.药… Ⅱ.①郑…②王… Ⅲ.药物-生物信息论 Ⅳ.R915

中国版本图书馆 CIP 数据核字(2004)第 013983 号

药物生物信息学

郑 珩 王 非 编著

责任编辑:余晓捷 杨燕玲 孙小芳

文字编辑:李 丽

责任校对:李 林

封面设计:关 飞

*

化 学 工 业 出 版 社
现代生物技术与医药科技出版中心 出版发行

(北京市朝阳区惠新里 3 号 邮政编码 100029)

发行电话:(010)64982530

http://www.cip.com.cn

*

新华书店北京发行所经销

聚鑫印印有限责任公司印刷

三河市延风装订厂装订

开本 787 毫米×1092 毫米 1/16 印张 19¼ 字数 447 千字

2004 年 4 月第 1 版 2004 年 4 月北京第 1 次印刷

ISBN 7-5025-5318-5/R·202

定 价:45.00 元

版权所有 违者必究

该书如有缺页、倒页、脱页者,本社发行部负责退换

序

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与完成，有关核酸、蛋白质的序列和结构数据呈指数增长。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。从 20 世纪 80 年代末开始，生物信息学逐渐兴起并蓬勃发展，业已成为一门工具型学科，广泛渗透到医学、药学、农学等相关领域，从而衍生出一系列交叉性前沿学科。

基因组学和生物信息学研究将极大地改变新药开发的思路与模式。对基因组计划产生的核酸序列分析一方面可从本质上认识疾病尤其是遗传性疾病的发生原因，为这些疾病的诊断、预防和治疗奠定基础；另一方面，这些核酸序列自身就是丰富的药物资源，其中蕴含大量目前尚不明了的基因，这些基因产物有的是调节机体机能的蛋白质类药物，有的可作为药物作用潜在的靶点而应用于新的药物开发。更重要的是，从药物作用与人类基因型的关系开拓了药物基因组学这一新的研究领域，为减小药物副作用、实现针对基因型的个性化治疗展示了宏伟蓝图。

对生物大分子信息处理分析为基于结构生物学的合理药物设计奠定了基础。通过偶然发现和随机生理活性筛选的方法已发现许多药物，但其盲目性较大，目前用这些方法发现药物已越来越难。随着分子生物学、分子药理学、计算机科学等的发展，根据药物与其作用靶点相互关系来设计新药成为当前国内外新药研究的重要手段，并已成功应用于酶抑制剂、受体抑制剂的研究开发；在蛋白质和多肽类药物分子设计方面，对新型蛋白质或多肽的分析及计算机辅助结构预测更不可缺少；生物信息的积累同时使计算机辅助抗原性分析、表位预测等成为可能，为降低生物大分子药物免疫原性，或提高疫苗活性等的分子设计研究提供了理论分析手段。

鉴于国内尚缺乏系统性、针对性介绍药物研发中的生物信息学工具参考书，作者编著了适用于我国高等教育的《药物生物信息学》。该书覆盖了核酸和蛋白质分析、计算机辅助药物设计、疫苗设计、生物芯片技术、药物基因组学等，较全面地介绍了生物信息学基本原理和技术，及其在药物研究开发中的应用，使学生通过本门课程的学习，可以较好地掌握药物生物信息学的基本理论、基本知识和基本技能。对于从事药学、生物技术、生命科学及生物信息学研究的科技工作者，也具有重要的参考价值。

吴梧桐

2004 年 1 月

前　言

药物生物信息学是一门崭新的综合性交叉学科。随着人类基因组计划的实施和深入，核酸、蛋白质序列及相关的生物信息数据呈爆炸性增长，这些生物信息为药学研究提供了丰富的资源，如从人类基因组序列中可预测和发现新基因，它们的产物可能是新的蛋白质类药物或作为药物开发的靶标；生物体尤其是病原微生物特异性基因序列分析，可为感染性疾病诊断芯片的设计奠定基础；蛋白质结构相关信息，可为虚拟药物筛选和合理药物设计提供帮助。另一方面，这些生物信息资源的处理分析和开发利用也对药学工作者提出了新的挑战，对这些错综繁杂的海量生物信息的提取和分析，离不开计算机和各种分析模型、算法软件的支持。药学、生物学、物理学、数学和计算机科学等学科的相互交叉渗透，促进了药物生物信息学这一新的研究领域的诞生。简单地说，药物生物信息学就是综合应用生命科学、数学、计算机科学等多学科的理论和方法，对伴随基因组计划产生的海量生物信息进行整理和分析，然后应用于药物的设计和开发，以达到合理药物设计目的的一门交叉性学科。

本书较系统地介绍了药物生物信息学这一新兴学科的概况，全书共分 13 章，前六章主要是生物信息学的基本原理，包括生物信息学概述、相关计算机基础、核酸和蛋白质数据库、常用生物信息和医药信息资源及检索方法、序列比对和数据库搜索原理等。后面七章则偏重于生物信息学原理和技术在药物开发上的应用，如核酸分析、蛋白质分析、新药开发相关生物信息学软件、计算机辅助药物设计、计算机辅助疫苗设计、生物芯片和药物基因组学等。药物生物信息学是一门实践性很强的学科，本书选择了一些应用性实例，如 SARS 病毒基因序列检索及分析、计算机辅助 HIV 蛋白酶抑制剂设计等，希望能给读者一些启发，促进我国生物信息学的普及和应用，提高药物研发和创新能力。本书可作为药学、生化药学、生物信息学、生物技术制药等专业的高年级本科生、研究生教材，也可供从事药物研发工作的技术人员参考。

本书编写工作自 2001 年底始，于 2002 年 9 月完成初稿。在初稿完成后，我们印制了试用教材在中国药科大学生物工程和生物技术专业高年级本科生以及生化药学专业研究生所修“药物生物信息学”课程中试用，获得了良好的教学效果及宝贵的反馈信息。与此同时，鉴于药物生物信息学及其相关学

科发展迅速，新理论和新方法层出不穷，我们数易其稿，以期本书不但能够适用于课程教学，还能够反映该学科的最新进展。

本书编写过程中，李菁博士在蛋白质结构预测部分给予了很大的帮助并编写了蛋白质分析一章中的部分内容，杨欣博士在生物信息软件部分给予了有益的建议并提供了部分参考材料。与远在美国的吴滔博士、胡容博士在生物信息学内容方面的讨论，以及张丽君博士、宋鹏飞博士赠送的《Computational Molecular Biology》一书，都使作者受益匪浅，在此对他们的支持和帮助表示由衷的感谢。同时特别感谢化学工业出版社有关编辑，他们在本书出版过程中所付出的大量而卓有成效的工作保证了本书的顺利出版。

药物生物信息学发展迅速，新理论、新技术、新发现和新成果层出不穷，书中取舍不当、疏漏和错误之处，恳请各位同行和广大读者批评指正。

郑珩　王非

2004 年 1 月

需要全本请访问 www.ertongbook.com

目　　录

第一章　生物信息学概述 …… 1
第一节　生物信息学及其发展 …… 1
一、什么是生物信息学 …… 1
二、生物信息学的诞生和发展 …… 1
三、生物信息学研究的主要推动力 …… 3
第二节　生物信息学的研究领域 …… 5
一、生物信息学在生物学领域的研究内容 …… 5
二、生物信息学在计算机领域的研究内容 …… 13
第三节　生物信息学在药学领域的应用 …… 14
一、初始阶段 …… 14
二、药物设计阶段 …… 15
三、药物开发阶段 …… 17
参考文献 …… 17
第二章　计算机网络基础 …… 19
第一节　Internet 简介 …… 19
一、Internet 的建立和发展 …… 19
二、TCP/IP 协议 …… 20
三、IP 地址与子网掩码 …… 20
四、域名系统 …… 22
第二节　Internet 提供的服务 …… 23
一、World Wide Web …… 23
二、电子邮件 …… 23
三、文件传输（FTP） …… 24
第三节　搜索引擎及其使用 …… 24
一、搜索引擎基本类型 …… 24
二、常用搜索引擎简介 …… 25
三、搜索引擎的使用方法 …… 28
参考文献 …… 30
第三章　核酸序列及基因组数据库 …… 31
第一节　核酸序列数据库 …… 31
一、概述 …… 31
二、GenBank 数据库 …… 33
三、EMBL 数据库 …… 38

第二节　基因组与功能基因组数据库 …… 44
一、GDB …… 44
二、UniGene …… 45
三、SGD …… 45
四、TDB …… 46
五、AceDB …… 46
六、dbEST …… 46
七、TRANSFAC …… 46
八、TRRD …… 47
九、COMPEL …… 47
十、CUTG …… 48
参考文献 …… 48
第四章　蛋白质序列及结构数据库 …… 50
第一节　蛋白质序列数据库 …… 50
一、概述 …… 50
二、SwissProt 蛋白质序列数据库 …… 50
三、PIR 蛋白质序列数据库 …… 55
四、PROSITE …… 56
五、PRINTS …… 58
六、TrEMBL …… 59
七、NRL-3D …… 59
第二节　蛋白质序列复合数据库 …… 59
一、NRDB …… 60
二、OWL …… 60
三、MIPSX …… 60
四、SWISS-PROT＋TrEMBL …… 61
第三节　蛋白质结构数据库 …… 61
一、蛋白质结构数据库 PDB …… 61
二、结构信息显示 …… 65
三、结构浏览器 …… 68
第四节　蛋白质结构分类数据库 …… 70
一、蛋白质结构域的折叠模式 …… 70
二、蛋白质结构分类数据库 …… 71
参考文献 …… 74
第五章　生物医药学信息检索 …… 76
第一节　集成式数据库检索系统 …… 76
一、Entrez 系统 …… 76
二、SRS 系统 …… 85
第二节　生物医药文献资源 …… 90

一、Medline …… 90
二、医药搜索引擎 …… 91
三、专利全文检索系统 …… 92
四、中文生物医药文献检索 …… 93
参考文献 …… 97
第六章　序列比对方法和相似性搜索 …… 98
第一节　概述 …… 98
第二节　双重序列比对 …… 99
一、序列比对的基本概念 …… 99
二、动态规划算法与全局比对 …… 100
三、局部比对 …… 103
四、打分矩阵和空位罚分 …… 104
第三节　多序列比对 …… 109
第四节　点阵作图法 …… 112
第五节　序列比对的统计学显著性 …… 115
第六节　数据库中的相似性搜索 …… 116
一、概述 …… 116
二、FASTA 搜索 …… 117
三、BLAST 搜索 …… 120
参考文献 …… 127
第七章　核酸序列分析 …… 129
第一节　基因结构分析 …… 129
一、生物体基因结构 …… 129
二、基因识别及预测 …… 132
三、其他位点分析 …… 137
四、基因结构分析的一般步骤 …… 141
第二节　核酸序列组装和基因定位分析 …… 142
一、核酸序列组装及电子克隆 …… 142
二、基因定位分析 …… 144
第三节　新基因的发现及功能预测 …… 145
一、利用 EST 数据库（dbEST）发现新基因 …… 145
二、从基因组 DNA 序列中预测新 ORF …… 146
三、通过核酸序列同源性比较进行功能预测 …… 146
参考文献 …… 146
第八章　蛋白质序列分析 …… 147
第一节　蛋白质组成及基本性质 …… 147
一、分子量及等电点 …… 147
二、蛋白质辨识 …… 147
三、酶切及断裂位点 …… 149

四、疏水性…… 150
第二节　蛋白质二级结构预测…… 152
一、二级结构预测方法…… 152
二、二级结构预测软件…… 155
三、二级结构预测的准确度…… 157
第三节　特殊结构或结构特征分析…… 159
一、跨膜区…… 159
二、信号肽和蛋白质定位…… 163
三、卷曲螺旋分析…… 163
第四节　蛋白质三级结构预测…… 164
一、同源模建…… 164
二、折叠模式识别…… 165
三、从头预测…… 167
参考文献…… 168
第九章　新药开发中的生物信息学软件 …… 170
第一节　概述…… 170
一、核酸和蛋白质序列分析…… 170
二、蛋白质高级结构分析…… 171
三、蛋白质组研究图谱的处理…… 171
四、蛋白质、核酸三维结构显示…… 171
五、格式转换…… 172
第二节　PCR 引物设计及其软件介绍 …… 172
一、引物设计的原则…… 172
二、WONDERFUL 生物信息学系统的引物设计功能介绍 …… 173
第三节　DNA 序列分析软件 …… 177
一、酶切位点分析…… 178
二、核酸基序分析…… 178
三、环状序列分析…… 179
四、开放阅读框 ORF 搜索 …… 179
五、CpG 岛搜索 …… 179
六、编码测试法预测编码区…… 180
七、序列转换…… 181
八、序列统计…… 181
第四节　蛋白质序列分析软件…… 182
一、蛋白酶切位点分析…… 182
二、蛋白质基序分析…… 183
三、蛋白质特性分析…… 183
四、蛋白质二级结构预测…… 184
第五节　序列比对软件…… 184

一、双序列比对…… 185
二、多序列比对…… 187
第六节 RNA 结构预测及其软件介绍 …… 188
一、RNA 二级结构预测的理论和方法 …… 188
二、RNA 结构预测软件 …… 191
第七节 生物信息软件的工业标准 GCG 简介 …… 194
第八节 生物分子与建模软件 Insight Ⅱ …… 199
一、功能及模块…… 199
二、应用领域…… 203
参考文献…… 206
第十章 计算机辅助药物设计…… 207
第一节 生物信息学与药物设计…… 207
一、新药研究与开发过程…… 207
二、合理药物设计…… 208
三、生物信息学与新药设计…… 209
第二节 先导化合物的计算机辅助设计…… 212
一、计算机辅助先导化合物设计的基本原理和方法…… 212
二、基于分子可视化和模建技术设计先导化合物…… 213
三、三维结构搜寻…… 214
四、全新药物设计…… 218
五、间接药物设计…… 221
第三节 先导化合物的计算机辅助优化…… 222
一、定量构效关系中的 Hansch 方法 …… 222
二、定量构效关系研究中的 Free-Wilson 方法 …… 223
三、分子连接性法…… 223
四、三维定量构效关系（3D-QSAR） …… 224
第四节 新型多肽及蛋白质分子设计…… 227
一、定点突变…… 228
二、盒式突变…… 228
三、蛋白质融合…… 229
四、蛋白质的从头设计…… 229
参考文献…… 231
第十一章 计算机辅助疫苗设计 …… 233
第一节 免疫学基本概念…… 233
一、基本概念…… 233
二、主要组织相容性系统…… 233
三、人工免疫与疫苗…… 235
四、新疫苗的研制和展望…… 236
第二节 免疫应答…… 238

一、概述…… 238
二、免疫应答的细胞学基础…… 238
三、T 细胞介导的免疫应答…… 239
四、B 细胞介导的免疫应答 …… 240
第三节　抗原表位预测方法…… 240
一、表位测定方法…… 240
二、基于基序的表位预测方法…… 241
三、基于人工神经网络的表位预测…… 243
四、基于分子模拟的表位预测方法…… 244
第四节　计算机辅助疫苗设计相关软件及应用…… 244
一、SYFPEITHI …… 245
二、TEPITOPE …… 246
三、MHCPred …… 246
四、其他软件及数据库…… 247
参考文献…… 248
第十二章　生物芯片技术 …… 249
第一节　基因芯片…… 249
一、基因芯片的制备…… 250
二、基因芯片的应用…… 251
第二节　蛋白质芯片…… 255
一、蛋白质芯片的类型与制备…… 255
二、蛋白质芯片的应用…… 256
三、存在问题和发展前景…… 258
第三节　生物芯片技术在新药研究中的应用…… 258
一、基因芯片与药物靶标…… 258
二、药理作用及药物毒性的评价…… 259
三、药物筛选…… 260
四、中药标准化鉴定…… 261
五、诊断芯片…… 261
第四节　生物芯片与生物信息学…… 262
一、确定芯片检测目标…… 262
二、芯片设计…… 263
三、生物芯片检测数据分析…… 269
参考文献…… 270
第十三章　药物基因组学 …… 272
第一节　药物基因组学的发展…… 272
一、概述…… 272
二、药物基因组学的产生和发展…… 273
第二节　药物基因组学的研究内容和方法…… 274
需要上中请访问：www.ertongbook.com

一、药物基因组学的研究内容…… 274
二、药物基因组学的研究方法和技术…… 276
第三节　药物基因组学的应用前景…… 277
一、为药物开发开辟一个全新领域…… 277
二、合理用药与个体化治疗…… 279
三、为医药公司带来无限商机…… 279
四、药物的临床和临床前研究…… 280
五、在中药现代化研究中的应用…… 280
参考文献…… 281
附 1　生物信息学大事记 …… 282
附 2　FASTA 和 BLAST 搜索参考 …… 284
附 3　生物信息学常用数据库 …… 287

第一章

生物信息学概述

第一节 生物信息学及其发展

一、什么是生物信息学

随着人类基因组计划的全面展开和深入，相关信息出现了爆炸性增长。自 1995 年科学家破译全长为 180 万核苷酸的流感嗜血杆菌基因组以来，至 2004 年 2 月，已有 170 多个模式生物如酵母、线虫、果蝇、拟南芥等的基因组被测定。根据国际数据库的统计，1990 年核酸碱基数目为 4 千万，1995 年核酸碱基数目为 2.5 亿，2000 年核酸碱基数目为 58 亿，到 2004 年 2 月这一数目已超过 200 亿。要对这些海量数据进行处理分析就必须发展新的分析理论、方法、技术和工具，也必须依赖计算机的信息处理。

生物信息学是在此背景下发展起来的综合运用生物学、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，它研究的材料是生物学的数据，而它所采用的方法，则是从各种计算技术衍生出来的。广义地说，生物信息学从事生物信息的获取、加工、储存、分配、分析和释读，并综合运用数学、计算机科学和生物学工具，以达到理解数据中的生物学含义的目标。生物信息学是内涵非常丰富的学科，其核心内容至少包括基因组信息学、蛋白质组信息学和代谢调控信息学三大部分。基因组信息学指基因组信息的获取、处理、存储、分配和解释，目的是确定全部基因在染色体上的确切位置以及各 DNA 片段的功能；蛋白质组信息学即对细胞、组织中蛋白质结构、表达、定位以及各蛋白质之间的相互作用信息进行处理和解释，最终确定各蛋白质的功能及相互作用关系，蛋白质组学研究不仅能为解释生命活动规律提供基础，还能为众多种疾病机理的阐明及攻克提供理论根据和解决途径；代谢调控信息学的目的是阐明基因表达的调控机理，它通过研究生物分子在基因调控中的作用，来了解人类疾病诊断、治疗的内在规律。生物信息学的研究目标是揭示“基因组信息结构的复杂性及遗传语言的根本规律”，解释生命的遗传语言，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。

二、生物信息学的诞生和发展

生物信息学就其萌生而言，是一门有“较长历史的学科”，早在计算机初创期的 1956 年就已在美国田纳西州的 Gatlinburg 召开过首次“生物学中的信息理论讨论会”。而就其发展而言，却是一门相当年轻的学科，因为继 20 多年的沉默之后，伴随着 20 世纪八九十年代计算机技术的迅猛发展，它才得以巨大发展。

关于生物信息学（bioinformatics）一词的来源，据说 20 世纪 80 年代末期，一位叫林

华安的博士认识到将计算机与生物学结合起来的重要意义，开始留意要为这一领域构建一个合适的名称。起初，考虑到与将要支持他主办一系列生物信息学会议的佛罗里达州立大学超型计算机研究所的关系，他使用的是“CompBio”；之后，又将其更改为兼具法国风情的“bioinformatique”，看起来似乎有些古怪。因此不久，他便进一步把它更名为“bioinformatics（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/符号常会引起许多系统问题，于是林博士将其去除，就是今天我们所看见的“bioinformatics”。

事实上，生物信息学的工作在此名词诞生之前就早已开始了。从数据库角度看，早在20世纪60年代，美国就建立了手工收集数据的蛋白质数据库。美国洛斯阿拉莫斯国家实验室1979年就已经建立起GenBank数据库。欧洲分子生物学实验室1982年就已经提供核酸序列数据库EMBL的服务。日本也于1984年着手建立国家级的核酸序列数据库DDBJ，并于1987年开始提供服务。目前绝大部分核酸和蛋白质数据由美国、欧洲和日本3家产生，以上三家共同组成了DDBJ/EMBL/GenBank国际核酸序列数据库，每天交换数据，同步更新。其他国家如德国、法国、意大利、澳大利亚、瑞士、瑞典、丹麦、加拿大、以色列、南非等，在分享网络资源的同时，还纷纷建立自己的生物信息中心，为本国服务。

从专业机构角度看，美国于1988年在国会的支持下成立了国家生物技术信息中心（NCBI），其目的是进行计算分子生物学的基础研究，构建和发布分子生物学数据库。欧洲于1993年3月就着手建立欧洲生物信息学研究所（EBI）。日本也于1995年4月组建了自己的信息生物学中心（CIB）。

从数据分析技术角度看，早在1962年，Zuckerkandl和Pauling就将序列变异分析与其演化关系联系起来，从而开辟了分子演化的崭新研究领域；1964年，Davies开创了蛋白质结构预测的研究；1970年，Needleran和Wunsch发表了广受重视的两序列比较算法；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。正如我们现在所看到的那样，在二十世纪八九十年代，生物学数据分析技术在国外更是获得了突飞猛进的发展。

从专业出版物来看，由于没有专业领域专门的期刊，起初的专业文献散落在各种其他领域的期刊中，到了1970年，出现了Computer Methods and Programs in Biomedicine这本相关期刊；到了1985年4月，第一种生物信息学专业期刊——Computer Application in the Biosciences正式出版；现在，相关专业期刊已经较多了，包括书面期刊和网上期刊两种，如Bioinformatics（即原先的Computer Application in the Biosciences）、Acta Biotheoretica、BioInformatics Technology &System、Bioinform Newsletter、Briefings in Bioinformatics和Journal of Computational Biology等。

从网络资源看，国际互联网上的生物信息学网点非常繁多，大到代表国家级研究机构的，小到代表专业实验室的都有。大型机构的网点一般提供相关新闻、数据库服务和软件在线服务，小型科研机构一般是介绍自己的研究成果。有的还提供自己设计的算法的在线服务。总体而言，基本都是面向生物信息学专业人士，各种分析方法虽然很全面，但却分

散在不同的网点，分析结果也需要专业人士来解读。

1996 年，北京大学加入欧洲分子生物学网络组织，成为该组织的中国国家节点，并成立生物信息中心，为生物、医学、制药、农业、环境等领域的研究开发提供生物信息资源服务，并开展二次数据库构建、软件集成、基因组分析等研究。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。近年来开展生物信息学研究的单位主要有：北京大学、清华大学、中国科学院生物物理所、军事医学科学院、上海生命科学院、中国科学院生物化学研究所、中国科学院微生物研究所、中国科学院遗传研究所人类基因组中心、中国医学科学院等。在一些著名院士和教授的带领下，部分领域取得了一定成绩，在国际上占有一席之地，如北京大学在生物信息学网站建设方面、中国科学院生物物理所在 EST 序列拼接方面以及在基因组演化方面、天津大学在 DNA 序列的几何学分析方面等，但从全国总体来看与国际的水平差距很大。

三、生物信息学研究的主要推动力

生物信息学的产生归因于 20 世纪生物学、分子生物学及计算机科学的迅猛发展。现代分子生物学的发展，特别是人类基因组计划的实施，使生物学家面临的数据不再是实验记录本上或文献上的简单数字，而是公共数据库中数以千兆计的记录。生物信息学就如同一个向导，帮助生物学家从这个公共数据库中寻找他们所需要的生物学知识。推动生物信息学研究的动力主要有现代分子生物学与计算机技术、人类基因组计划以及生物医药工业。

（一）现代分子生物学与计算机技术

20 世纪 50 年代沃森（James Dewey Waston）和克里克（Francis Harry Compton Crick）建立了 DNA 双螺旋结构的模式，标志着分子生物学时代的来临；60 年代人们发现了限制性内切酶，促使 DNA 重组技术诞生和发展；70 年代建立的 DNA 序列测定技术，推动生物学、遗传学等学科向分子微观层次的迅速扩展，生命科学研究亦步入前所未有的高速发展阶段。伴随生命科学迅速发展的是生物数据如生命科学相关文献及生物序列、结构信息的大量积累。分子生物学和遗传学的文献积累从 20 世纪 60 年代的近 10 万篇增长到 80 年代中期的 30 万篇。此后，至 90 年代中期，文献数已增长至 40 多万篇，平均每年增长 1 万篇。而至 2000 年，则达到近 50 万篇，平均每年 2 万篇。同时，以 DNA 序列为核心的序列数据也在迅速增长，从 20 世纪 80 年代初的百余条序列、几十万碱基迅速上升至 20 世纪 90 年代末的数百万条序列、数十亿碱基。而至 2000 年底，国际数据库记录的接近 1000 万条 DNA 序列的碱基数已超过 100 亿！如果用传统的纸张来书写，以每个核苷酸作为一个字符，则需要印制 1 万本每本 1000 页每页 1000 字的书！另外，二维凝胶电泳技术、测序质谱技术以及生物芯片技术的高速发展和广泛应用，也使得大量的数据信息已经无法用传统的文献形式发表，而更多的需以数据库形式，通过文字、图像、超链接等多种方式来记录。

生物学相关信息量的革命性的爆炸，产生了对海量生物信息进行处理的需求。幸运的是，在过去的二十多年里，电子计算机芯片对于数字处理的能力的增长基本符合 Moore 定律（指数增长）。每个 CPU 所含晶体管数从 20 世纪 70 年代初的几千个迅速而稳定地增长到 80 年代末的上百万个，即平均每 2 年翻一番。此后，至 20 世纪 90 年代末，又上升至上亿个，即平均每 2.5 年翻一番，如今的大型计算机的数据处理能力，已经发展到每秒

数千亿次乃至数万亿次计算的水平。有了这一技术支持条件，连同计算方法的创新和发展，基因组研究和其他生物学研究所产生的海量数据，才能够得以有效地管理和运行，于是，生物信息学便在综合计算生物学的研究和生物学信息的计算机处理的基础上迅速而成功地发展起来。

（二）人类基因组计划

1977 年，第一个病毒（$\phi x174$）基因被测序，它含有约 5000 个核苷酸，11 个基因；1995 年，第一个细菌基因序列被全部测定，即流感嗜血杆菌（*H. influenzae*），其基因组包含了 180 万个核苷酸，约有 1700 个基因。在此之后，掀起了一股全基因组测序的浪潮，并促进科学家们有计划地合作测定一系列生物体的全基因组。美国自 1990 年开始实施人类基因组计划（human genome project，HGP），其目标是测出人类和一些模式生物的基因组 DNA 全序列，制作完成全序列的物理图谱，找出全部人类基因，以及发展出一系列实现上述目标所需的技术。2001 年 3 月在 Nature 和 Science 杂志上同时发表了由人类基因组协作组和 Celera 公司分别完成的人类基因组序列及其初步的分析，已经被测定的人类基因组含有约 29 亿碱基，其物理图谱覆盖率达 96%，序列覆盖率为 94%。至 2004 年 2 月，已有 172 个模式生物的基因组被测定，包括 20 个真核生物（eukaryota）、18 个古细菌（archaea）、134 个细菌（bacteria）。

另一方面，各种生物的表达序列标签（expressed sequence tag，EST）序列数据也在迅速扩增。EST 是从已建好的 cDNA 库中随机取一个克隆，从 5′端或 3′端对插入的cDNA 片段进行一轮自动测序，测得的约 500bp 的一段序列就称为一个 EST。每一个 EST 代表一个表达基因的部分转录片段，通过对 EST 分析，有助于发现新基因、分析基因在不同组织中的表达特异性和建立物理图谱。至 2004 年 1 月 30 日，dbEST 数据库已收录超过 1991 万条序列，其中人类基因 EST 序列超过 547 万（http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html）。对如此巨量的数据进行存储、分类、分析等工作，没有计算机进行处理是不可想像的。

由于人类基因组计划的启动，生物信息学已成为分子生物学的一个组成部分。生物信息学的首要功能就是要有效地组织和管理数据，并用新的信息技术为实验计划提供帮助。同时，有关信息资料的基础设施已经建立起来，在互联网上提供了各种各样广泛的数据库和计算资源。更重要的是，生物信息学面对的下一个主要课题是从大量的数据信息中挖掘出有用的生物学信息，这些数据信息不仅包括高通量的 DNA 测序技术带来的序列数据，而且还包括新的 DNA 芯片技术将带来的表达和突变数据。

（三）生物医药工业

生物医药工业也是推动生物信息学发展的重要动力。近年来生物医药公司纷纷投资生物信息学研究，主要原因在于生物信息学可大大加速新药开发过程。HGP 所推动的大规模 DNA 测序也为生物医药工业提供了大量可用于新药开发的原材料，有些基因产物可以直接作为药物，而有些基因则可以成为药物作用的靶标。欧美许多医药公司纷纷投资 HGP 特别是 EST 的研究工作，另一个不可忽视的原因是基因和基因序列可以申请专利保护。至今已有数千项基因专利注册成功。因为生物信息学能够大大加快传统基因发现和研究，因而成为各盈利性研究机构和医药公司争夺基因专利的重要工具，同时这一竞争又反过来极大地刺激了生物信息学的发展。

第二节 生物信息学的研究领域

一、生物信息学在生物学领域的研究内容

生物信息学作为一门新兴学科，已经渗入生物学研究的各个领域。与生物学研究的多层次性相应，笔者认为其研究也至少包括分子层次、细胞层次和生物种群等三大部分。在分子层次的研究主要针对核酸、蛋白质和多糖三大类生物大分子相关信息的收集和处理；细胞层次的研究则涉及基因组学、蛋白组学、代谢调控网络等；而生物种群层次主要是基于分子信息的进化研究等。就目前的研究来看，由于测定多糖序列及空间结构的困难性和复杂性，其数据量远远少于核酸和蛋白质相关信息，少数多糖序列信息往往附加在其他数据库中。为此，以下将生物信息学在生物学领域的研究内容分为核酸及基因组信息、蛋白质及蛋白质组信息、分子相互作用及代谢调控网络、生物进化等四大部分进行简要阐述。

（一）核酸及基因组信息

生物体的发育、分化、生长和代谢的过程，始终是遗传信息从储存到表达、加工及传递的过程。实质上，其中的一大部分是从基因组到基因再到 mRNA（cDNA）的信息传递过程。而生物体的遗传、变异和进化问题，则主要体现在遗传信息的复制、重组、变异和选择。要从根本上揭示生命过程，首要前提是认识基因组本身，也就不可回避大规模 DNA 序列测定及相关数据获取、储存、处理和利用。

1. 大规模核酸测序及拼接

基因组研究最基本的任务之一就是获得生物体全基因序列，在此过程中离不开基因的大规模测序。目前广泛应用的两个测序策略是：全基因散弹法和逐步克隆法。对小基因组（1～5Mb）用散弹法非常成功，但由于人类基因中高比例重复序列的存在，克隆文库不可避免的存在间隙和基因多态性，这些原因导致散弹法测序的组装非常困难，需要更为复杂的信息分析和处理技术。为此，Celera 公司设计了一系列的信息分析软件和相关的实验方法。逐步克隆法的限制是需要制作物理图谱。然而不管用什么策略，其序列测定和拼接的基本流程相似，包括测序仪的光密度采样与分析、碱基读取、载体标识与去除、拼接、填补序列间隙、重复序列标识、可读框预测和基因标注等，在此过程中每一步都与基因组信息学的软件和数据库分不开。

大规模基因组测序所产生的信息量之大是难以形容的。一般来说，一个测序中心至少有 1/3 的人力来管理和处理这些数据。在测序信息分析中，序列拼接和填补序列间隙是最为关键的首要难题，这一过程尤其需要发展适当的算法及相应的软件，把实验设计和信息分析时刻联系在一起。在人类基因组全序列测定的过程中，拼接和间隙填补的困难，不仅来自它的巨大的数据量，而且在于它含有高度重复的序列。目前常用的软件有 Phil Green 实验室发展的 Phred-Phrap 软件包和加利福尼亚大学发展的 GihAssembler 软件包等。

2. 基因识别与定位

经过全世界科学家的通力合作，人类基因组序列工作草图已于 2000 年 6 月绘制完成，进一步的序列拼接和注释工作也在加紧进行中。我国科学家也参加了这一伟大的合作，负责全基因组 1%的测序、拼接以及注释工作。现在，遍布全世界的测序仪每天都在产生着

需要全本请购买 www.ertongbook.com