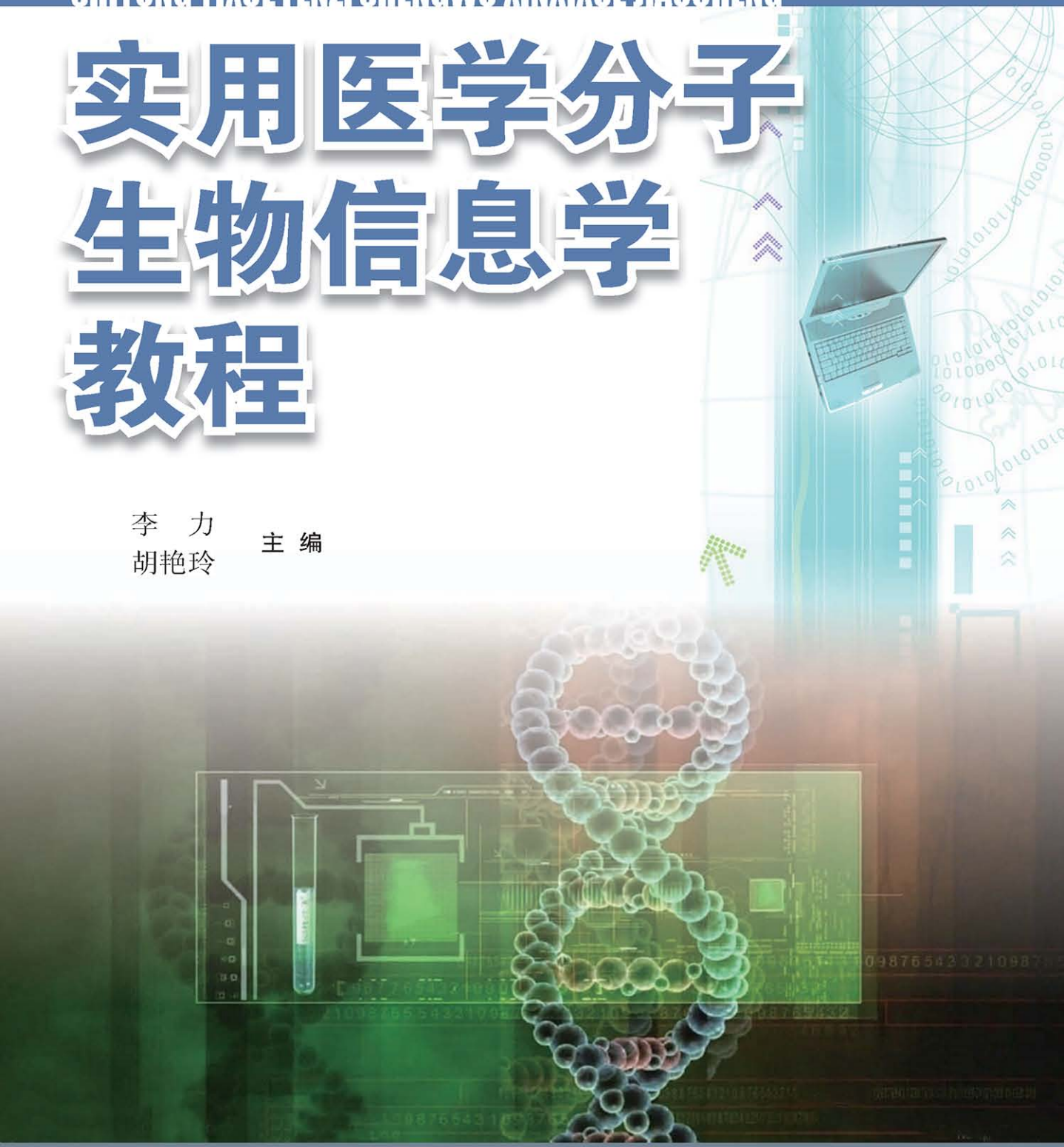
SHIYONG YIXUE FENZI SHENGWU XINXIXUE JIAOCHENG

实用医学分子生物信息学教程

李　力
胡艳玲　主编



广西科学技术出版社

图书在版编目（CIP）数据

实用医学分子生物信息学教程 / 李力，胡艳玲主编.
—南宁：广西科学技术出版社，2013.12
ISBN 978-7-5551-0002-7

Ⅰ.①实… Ⅱ.①李… ②胡… Ⅲ.①医学—分子生物学—信息学—教材 Ⅳ.①Q7-05

中国版本图书馆CIP数据核字（2013）第279133号

实用医学分子生物信息学教程

主　编：李　力　胡艳玲
副主编：王　琪　尹富强　杨小丽
　　钟艳平　谢　莹　廖　明

责任编辑：林　坚　陈　婧　　封面设计：韦娇林
责任校对：梁　斌　　责任印制：韦文印

出 版 人：韦鸿学　　出版发行：广西科学技术出版社
社　　址：广西南宁市东葛路66号　　邮政编码：530022
网　　址：http://www.gxkjs.com

经　　销：全国各地新华书店
印　　刷：广西大华印刷有限公司
地　　址：广西南宁市高新区科园路62号　　邮政编码：530007
开　　本：787 mm×1092 mm　1/16
字　　数：322千字　　印　　张：17.25
版　　次：2013年12月第1版　　印　　次：2013年12月第1次印刷
书　　号：ISBN 978-7-5551-0002-7
定　　价：38.00元

前 言

生物信息学是随着分子生物学、计算机科学、数学等多学科发展而形成的，集生物学、数学、统筹学、计算机科学、信息学、化学与物理学等为一体的交叉学科，包括生物信息的获取、处理、存储、分发、分析和解释等方面的内容。医学生物信息学，顾名思义即生物信息学在医学领域的应用，它的目标是以核酸、蛋白质代谢等生物分子数据库为主要研究对象，运用数学、信息学、计算机科学、统筹学为主要手段，以计算机硬件、软件和计算机网络为主要工具，对庞大复杂的原始数据进行存储、管理、注释、加工，使之成为具有明确生物意义的生物信息；通过对生物信息的查询、搜索、比较、分析，从中获取基因编码、基因调控、核酸、蛋白质和代谢物质结构功能及其相互关系等理性知识；在大量信息和知识的基础上，探索生命起源、生物进化，以及细胞、器官和个体的生长、发育、病变、衰亡等生命科学中的重大问题，搞清它们的基本规律和时空联系，建立生物学周期。随着对疾病的认识从大体解剖细胞水平再到分子水平，医学生物信息学已成为医学临床和基础研究的重要工具。

广西医科大学科学实验中心自 2007 年起承担学校研究生和本科生的医学生物信息学的教学工作，在多年的教学实践中认识到医学生特别是临床医学生由于过去接受的相关的基础理论知识与理工科学生有所不同，数学、统筹学、计算机科学、信息学、化学与物理学是他们所应具备的基础知识中的弱项。为了让医学生更好地掌握医学生物信息学并能综合运用到医学临床和基础研究工作中，有必要在医学生现有的知识水平的基础上，编写一本理论与实践相结合，便于医学生使用的相关教材。《实用医学分子生物信息学教程》的编者都是一线从事医学生物信息学教学的中青年教师，他们本身具备较好的生物学、数学、统筹学、计算机科学、信息学等多学科知识和掌握生物信息学的各种工具，同时在多年从事医学生物信息学教学实践过程中也深知医学生学习这门课的难点所在。因此在这本 《实用医学分子生物信息学教程》 中，在各章节开始便注明本章节重点掌握的知识点，而且在阐明医学分子生物信息学基础理论时尽可能言简意赅，并较为详尽地介绍了生物信息学在医学科研和临床应用中的最新信息及资料；另外每章节后都附有如何将本章节知识点用于医学临床和基础研究的实例分析，尽可能便于学生学以致

用，理论联系实际。《实用医学分子生物信息学教程》既可作为生物信息学课程的教材，也是一本实用性很强的生物信息学参考书。

由于编写人员理论水平和实践经验有限，书中难免存在许多不足之处，还望教师、学生及读者提出批评和改进意见，以便今后进一步修订提高。

编者

2013 年 01 月 14 日

目 录

第一章 生物信息学概论 …… 1
第一节 人类基因组计划与生物信息学的产生 …… 1
1 人类基因组计划及延续 …… 1
2 生物信息学的产生 …… 2
第二节 生物信息学的研究目标和内容 …… 3
1 生物信息学的生物学内涵 …… 3
2 生物信息学的研究目标和内容 …… 4
第三节 生物信息学的发展 …… 7
1 人类基因组计划时代的基因组信息学 …… 7
2 后基因组时代的生物信息学 …… 8
第四节 生物信息学的应用和展望 …… 9
1 生物信息学的应用 …… 9
2 生物信息学的展望 …… 10
习 题 …… 11
第二章 生物信息学数据库 …… 12
第一节 序列数据库 …… 13
1 核酸序列数据库 …… 13
2 蛋白质序列数据库 …… 22
第二节 基因组数据库 …… 29
1 基因组数据库 GDB …… 29
2 线虫基因组 AceDB 数据库 …… 30
3 酵母基因组 SGD 数据库 …… 30
4 美国基因组研究所的 TDB 数据库 …… 31
5 UniGene 数据库 …… 31
6 HapMap 数据库 …… 31

第三节　其他数据库…… 32
1　综合数据库 …… 33
2　序列与结构数据库 …… 35
3　疾病遗传信息数据库 …… 36
4　蛋白质综合数据库 …… 40
5　专科疾病信息数据库 …… 41
6　基因与蛋白质表达及调控数据库 …… 42
7　与疾病相关的低等生物数据库 …… 43
8　生物网络及分子代谢通路数据库 …… 44
9　RNA 序列和核糖体数据库 …… 45
10　基因图谱及生物基因组数据库 …… 45
11　蛋白质互作、蛋白质-RNA 和蛋白质-DNA 结合数据库 …… 47
12　其他类型数据库 …… 49
习　题…… 51
第三章　核酸序列及信息分析 …… 52
第一节　核酸序列的获取及序列比对…… 52
1　核酸序列的获取与递交 …… 52
2　序列比对和数据库搜索 …… 56
第二节　核酸结构和功能的预测分析…… 62
1　针对核酸序列的预测方法 …… 62
2　重复序列的识别与分析 …… 63
3　编码区统计特性分析 …… 63
4　基因结构预测 …… 63
5　tRNA 基因识别 …… 67
6　基于核酸序列的电子基因定位 …… 68
7　基于序列同源性分析的蛋白质功能预测 …… 68
第三节　实例分析…… 69
1　卵巢癌组织中 CCL18 的 cDNA 分析 …… 69
2　CDS 序列的染色体电子定位 …… 71
3　进行序列表达谱分析（GEO，UniGene） …… 72
4　编码序列分析 …… 74
习　题…… 75

第四章 蛋白质结构与功能预测 …… 77
第一节 蛋白质一级结构分析 …… 77
1 常用的蛋白质数据库 …… 78
2 蛋白质理化性质分析 …… 79
3 一级结构与功能的关系 …… 84
第二节 蛋白质二级结构分析 …… 86
1 模体（motif） …… 87
2 PredictProtein（蛋白质序列和结构预测服务网站） …… 87
3 PSIPRED（蛋白质二级结构预测网站） …… 89
第三节 蛋白质三级结构 …… 90
1 结构域（Domain） …… 90
2 蛋白质结构域与功能分析 …… 91
第四节 蛋白质三维结构分析 …… 97
1 同源建模 …… 98
2 线串法 …… 101
3 从头预测 …… 101
4 蛋白质三维结构观察 …… 101
5 蛋白质空间构象对其功能的影响 …… 102
第五节 实例分析：对 GFAP 人胶质纤维酸性蛋白（glial fibrillary acidic protein）进行结构与功能预测 …… 103
1 GFAP 概况 …… 103
2 GFAP 一级结构分析 …… 106
3 蛋白质二级结构分析 …… 112
4 蛋白质结构域与功能分析 …… 113
5 蛋白质三维结构分析 …… 115
习 题 …… 116
第五章 基因组遗传多态性及信息分析 …… 117
第一节 微卫星及信息分析 …… 117
1 微卫星标记简介 …… 117
2 微卫星 DNA 的特点 …… 118
3 微卫星在人类疾病研究中的应用 …… 118

4 微卫星相关的生物信息数据库 …… 119
第二节 单核苷酸多态性及功能注释 …… 124
1 单核苷酸多态性简介 …… 124
2 单核苷酸多态性在人类疾病研究中的应用 …… 125
3 SNPs 的信息分析 …… 126
第三节 连锁不平衡及单体型分析 …… 133
1 连锁不平衡 …… 133
2 单体型 …… 135
3 单体域和标签 SNP …… 137
4 连锁不平衡及单体型分析的主要软件及数据库 …… 138
第四节 拷贝数变异（CNV）及信息分析 …… 143
1 拷贝数变异简介 …… 143
2 CNV 在人类疾病研究中的应用 …… 144
3 CNV 相关的生物信息数据库 …… 144
第五节 全基因组关联分析 …… 146
1 全基因组关联研究简介 …… 146
2 GWAS 研究设计类型 …… 147
3 GWAS 研究设计表型选择 …… 148
4 GWAS 研究的统计分析 …… 148
5 GWAS 研究分析的软件及数据库 …… 150
第六节 实例分析：以 COMT 基因为例，选择该基因上与功能有潜在关系的 SNPs 位点 …… 154
习 题 …… 158
第六章 表观遗传与代谢组生物信息学分析 …… 159
第一节 microRNA 的生物信息学分析 …… 159
1 microRNA 简介 …… 159
2 microRNA 常用数据库 …… 160
3 microRNA 进化的生物信息学知识 …… 163
4 microRNA 与细胞网络的相互作用 …… 165
5 实例分析 …… 168
第二节 DNA 甲基化的生物信息学检测 …… 170

1 DNA 甲基化 …… 171
2 DNA 甲基化相关数据库 …… 173
3 甲基化预测、分析相关工具 …… 174
4 实例分析 …… 177
第三节 代谢组学的生物信息学分析 …… 180
1 代谢组学简介 …… 180
2 代谢组学的数据特点和标准 …… 181
3 代谢组学相关的生物信息学分析数据库 …… 182
4 代谢组学的生物信息学分析举例 …… 186
习 题 …… 189
第七章 后基因组时代的生物信息 …… 190
第一节 基因表达谱芯片的数据分析 …… 191
1 表达数据的获取和标准化 …… 191
2 基因表达矩阵的构建 …… 192
3 差异表达基因的筛选 …… 192
4 基因表达聚类分析 …… 193
5 基于表达谱的调控网络构建 …… 194
6 表达芯片数据的元分析 …… 198
第二节 生物医学信息的文本挖掘 …… 199
1 文本挖掘的概念 …… 200
2 文本挖掘的主要研究方向 …… 200
3 文本挖掘技术在生物医学的应用 …… 201
4 常用的生物医学文献及文本挖掘数据库 …… 202
第三节 分子进化和系统发生分析 …… 203
1 分子系统发生分析 …… 203
2 系统发生树 …… 205
3 距离和特征 …… 206
4 分子系统发生分析过程 …… 207
5 进化及系统发生数据库及软件 …… 209
第四节 实例分析：以肝癌为例，从肝癌基因表达芯片筛选功能相关基因
…… 210
1 GEO 数据的查询与下载 …… 210

2 差异基因的选择 …… 212
3 探针号与基因名的转换 …… 213
4 基因表达数据集的聚类分析 …… 215
5 通路富集分析 …… 216
6 基因信息的文本挖掘 …… 217
习 题 …… 218
第八章 生物信息学软件使用 …… 219
第一节 序列分析软件 DNAMAN 的使用方法 …… 219
1 将待分析序列装入 Channel …… 220
2 以不同形式显示序列 …… 220
3 DNA 序列的限制性酶切位点分析 …… 220
4 DNA 序列比对分析 …… 223
5 序列同源性分析 …… 224
6 PCR 引物设计 …… 228
7 画质粒模式图 …… 231
第二节 引物设计原理及软件 …… 234
1 设计原理 …… 234
2 引物设计与筛选 …… 235
3 实例分析: Primer Premier 5.0 引物设计 …… 238
第三节 DAVID: 系统性和综合性的大型基因分析数据库 …… 241
1 功能注释 …… 241
2 基因功能聚类 …… 245
3 基因 ID 转换 …… 246
4 相关基因批量显示 …… 247
第四节 进化分析软件 …… 248
1 ClustalX 和 Phylip 软件相结合建进化树 …… 248
2 MEGA 4.0 软件 …… 252
习 题 …… 256
参考文献 …… 258

第一章　生物信息学概论

第一节　人类基因组计划与生物信息学的产生

重点提示:

1. 2005 年人类基因组计划完成了人类基因组大约 30 亿个碱基对的全序列测定。

2. 根据基因组图谱，人类的基因数实际仅有 2 万 ~2.5 万个。

3. 生物信息学是随着基因组的发展而发展起来的，人类基因组计划的完成带动了生物信息学的迅速发展。

1　人类基因组计划及延续

1986 年 3 月，诺贝尔生理学或医学奖得主 R. Dulbecco 在 《Science》 上发表文章 “A turning point in cancer research: Sequencing the genome”，认为要彻底阐明癌症的发生、演进、侵袭和转移的机制，必须对人体细胞的基因组进行全测序。1990 年 10 月，美国政府正式启动为期 15 年的人类基因组计划（Human Genome Project，HGP），预期到 2005 年完成，此计划完成后将完成人类基因组大约 30 亿个碱基对的全序列测定。人类基因组计划是一项国际性的研究计划，它的主要任务是人类基因组以及一些模式生物体（细菌、酵母、线虫、果蝇、白鼠等）基因组的遗传图谱和物理图谱的测序和基因识别。HGP 的目的是解码生命，了解生命的起源，了解生命体生长发育的规律，探究种属之间和个体之间存在差异的起因，认识疾病产生的机制以及长寿与衰老等生命现象，为疾病的诊治提供科学依据。1999 年 9 月，中国积极参加到人类基因组计划中，承担其中 1% 的任务，即人类 3 号染色体上约3 000万个碱基对的测序任务。中国是参加这项研究计划的唯一的发展中国家。2000 年 6 月 26 日，人类基因组工作草图完成；2001 年 2 月，已被鉴定的人类基因组图谱在 《Nature》 杂志上公布。让人意想不到的是，根据基因组图谱，人类的基因数实际仅有 2 万 ~2.5 万个，比国际人类基因组计划 2001 年公布的人类拥有 3 万 ~4 万个基因要低。如此少的人类基因数，却能产生如此复杂的结构和功能，研究人员对此作了如下解释：首先，每个基因平均可能合成几种蛋白质，因为目前发现

的人类基因有 3 万 ~4 万个，而蛋白质却有 25 万种之多。其次，人体蛋白质在合成后进行多种修饰，即分别黏附上了不同的糖和其他物质，从而产生不同功能的蛋白质。最后，目前对基因数目的计算可能存在失误。人类教育科学主席 W. Haseltine 博士至今还坚持 10 万 ~12 万个基因数的估计。他认为，两个小组的研究人员可能采用错误的计算方法，寻找基因的方法也有所欠缺。

2001 年 2 月 12 日，参与人类基因组计划的美、日、法、中等六国科学家和美国塞莱拉公司在《Science》和《Nature》杂志上公布了人类基因组精细图谱及其初步分析结果，标志生命科学进入后基因组时代（post-genome era）。后基因组时代的研究重心转向了揭示基因组及其包含的全部基因的功能，以及对基因产物——蛋白质结构和功能的研究和预测（蛋白质组学，proteomics），即从整个基因组及其全套蛋白质产物的结构、功能和机制的深度去了解生命活动的全貌，并系统地整合有关生命科学的全部知识，阐明遗传、发育、进化、功能调控等基本生物学问题，以及与人类健康和疾病相关的生物医学问题。

2 生物信息学的产生

生物信息学（Bioinformatics）是近 20 年来迅速发展起来的一门新兴学科，它是集生物学、数学、统筹学、计算机科学、信息学、化学与物理学等为一体的交叉学科。生物信息学的产生与基因组的发展有很大的关系，最初被称为基因组信息学，因此生物信息学的研究内容与基因组研究的发展密切相关。

人类基因组计划的直接结果是获得了难以计数的不连续的数据，由于计算机数据库等技术的迅速发展，20 世纪 80 年代初开始建立了美国 GenBank、欧洲分子生物学实验室数据库（EMBL）和日本 DNA 数据库（DDBJ），用户可通过各种存储媒体以及互联网利用这些数据库的资源。生物信息最基本的表达形式是一维的分子排列顺序，即序列，包括核酸序列和氨基酸序列。从 1990 年 1 月至 2010 年 12 月，仅登录在美国 GenBank 数据库中的核酸序列总量已达4 398 085 644条，基于 cDNA 序列测定所建立起来的 EST 数据库记录也已达上百万条。另外，相继构建的蛋白质一级结构即氨基酸序列数据库数据也与日俱增，比较著名的有 PIR、SWISS-PORT 和 PRIDE 等。随着生物信息的爆炸式增长，目前各种类型的数据库陆续产生。截至 2011 年 3 月，已有的生物医学类数据库已达上千个，包括基因组学，蛋白组学，临床疾病检测，细菌、病毒与宿主关系，表观遗传信息数据库等，这些数据库的文献已近万篇，这一切构成了一个生物信息学数据的海洋。这种巨大积累规模的科学数据，在人类的科学研究历史上是空前的。这些品种繁多、信息各异的数据库，通过网络连接，构成了极其复杂、规模巨大的生物信息资源网

络。

数据并不等于信息和知识，但却是信息和知识的源泉。如何收集、存储、分析这些数据，尤其是如何从这些不相连的数据中获取有用的生物学信息是问题的关键所在。生物数据量的迅猛增长，最终是要把这些生物学问题转化成为有用的资料。为此，就必须发展新的分析理论、方法、技术和工具，依赖多学科如计算机科学、数学、统筹学等交叉学科提取、整理和转化成有用的知识。于是，伴随着美国国立卫生研究院（NIH）的人类基因组计划，生物信息学应运而生。

第二节　生物信息学的研究目标和内容

重点提示:

1. 生物信息学是多个领域交叉的新兴边缘学科，需要一定的计算能力，需要强有力的创新算法和软件，以及实验科学来证明。

2. 生物信息学的研究目标：通过认识生命的起源、进化、遗传和发育的本质，破译隐藏在DNA序列中的遗传信息，揭示人体生理和病理过程的分子基础。

3. 生物信息学的研究内容包括对基因组序列分析和解释、药物分子设计、基因多态性分析、基因表达调控、疾病相关基因功能验证等。

1　生物信息学的生物学内涵

生物信息学是在生命科学、计算机科学和数学基础上逐步发展而形成的一门新兴的边缘学科，它以核酸和蛋白质为主要研究对象，以数学、计算机科学为主要研究手段，对生物学实验数据进行获取、加工、存储、检索和分析，从而达到揭示数据所蕴含的生物学意义的目的。生物信息学是伴随基因组研究而产生的，其研究内容紧随着基因组研究发展，其核心是基因组信息学。生物信息学是把基因组DNA序列信息分析作为源头，找到基因组序列中代表蛋白质和RNA基因的编码区；同时阐明基因组中大量存在的非编码区的信息实质，破译隐藏在DNA序列中的遗传信息规律；据此归纳、整理跟基因组遗传信息和其调控相关的转录谱和蛋白质谱的数据，从而认识代谢、发育、分化、进化的规律。另外，生物信息学可利用基因组中编码区的信息进行蛋白质空间结构的模拟和蛋白质功能的预测，并将此类信息与生物体和生命过程的生理生化信息相结合，阐明其分子机制，最终进行蛋白质和核酸的分子设计、药物设计和个体化的医疗保健设计。

基因组信息学作为一个学科领域，关键是弄清楚全部基因在染色体上的确切位置及

各 DNA 片段的功能。其具体内涵包括：第一，要发展有效的能处理通量数据的软件和数据库；第二，须产生包括电子网络等远程通信工具的若干数据库工具，能方便地处理日益增长的物理图、遗传图、染色体图和序列信息，并在这些数据资料中进行比较；第三，要研究算法和分析技术，用于解释基因组的信息，例如预测序列的功能区域等。生物信息学的另一个重要任务是进行蛋白质、RNA 等结构模拟和分子设计，以及随之而来的药物设计。

总之，要把生物学问题转化成对数据符号的处理问题，从事生物信息学研究应具备多方面的科学基础。第一，它需要一定的计算能力，包括相应的软、硬件设备。第二，要有各种数据库或者能与国际、国内的数据库进行系统有效的交流，要有发达、稳定的互联网络系统。第三，生物信息学需要强有力的创新算法和软件。没有算法创新，生物信息学就无法获得持续的发展。第四，它要与实验科学，特别是与自动化的大规模高通量的生物学研究方法与技术平台建立广泛、紧密的联系。这些技术，既是产生生物信息数据的主要方法，又是验证生物信息学研究结果的关键手段。

2 生物信息学的研究目标和内容

20 世纪 90 年代以来，伴随着各种基因组测序计划的展开和分子结构测定技术的突破以及互联网的普及，上千个生物医学数据库迅速出现和成长，也对生物信息学工作者提出了严峻的挑战：数以亿计的 ACGT 序列中包含着什么信息？基因组中的冗余序列是否真的冗余？与疾病相关的基因表达是否存在调控网络？

构成和维持一个生物有机体所必备的基本信息包含于其基因组之中，由细胞内进行多种分子生物学反应将这些信息转化为真正的生命现象。基因组中一部分信息转录、翻译成 RNA 和蛋白质，另一部分信息可能参与调控这些大分子的表达。这些物种进行多种复杂反应组成复杂的生命现象，在体内的特定位置上实现其功能。然而，这些过程的大量细节都是在分子生物学研究的实验室里揭示出来的，所形成的大量数据，存储于数据库中。生物信息学是一门深深扎根于全面深入的实验事实和数据的理论生物学，试图从这些数据中提取新的生物学信息和知识。生物信息学的研究内容主要包括以下方面：基因组序列分析和解释、基因多态性分析、基因表达调控、药物分子设计、与疾病相关基因功能验证、基因进化、基因产物结构与功能预报、基于遗传的流行病学等。

因此，生物信息学的研究目标可以总结为：以核酸、蛋白质等生物大分子数据库为主要研究对象，运用数学、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，对庞大复杂的原始数据进行存储、管理、注释、加工，使之成为具有明确生物意义的生物信息。另外，通过对生物信息的查询、搜索、比较和分析，从

中获取基因编码、核酸和蛋白质结构功能、基因调控及其相互关系等理性知识。在大量信息和知识的基础上，探索生命起源、生物进化以及细胞、器官和个体的生长、发育、病变、衰亡等生命科学中的重大问题，弄清这些生物学基本规律和时空联系。总之，生物信息学通过认识生命的起源、进化、遗传和发育的本质，破译隐藏在 DNA 序列中的遗传信息，揭示基因组信息结构的复杂性及遗传信息的根本规律，揭示人体生理和病理过程的分子基础，为人类疾病的诊断、预防和治疗提供最合理而有效的方法和途径。生物信息的研究内容可以总结为以下几个方面。

2.1　基因组测序中的生物信息分析

人类基因组和其他一些生物基因组的大规模测序是生命科学研究很重要的一个里程碑。但是，面对浩瀚的基因组数据，分析方法占很重要的地位。测序仪的采样、分析、碱基读出、载体标识和去除、拼接与组装、填补序列间隙、重复序列标识、读框预测、基因标注等都依赖于信息学的软件和数据库。目前很多来自基因组水平上的分析方法和软件已经提出，能够有效地收集、整理、管理、处理、维护、利用、分析这些基因组数据。另外，比较基因组学是基因组学研究的另一个重要内容，即进行基因组异同的比较，是在整个基因组水平上的比较，比如各物种基因组大小、基因数量、基因是否存在以及其位置和排列顺序等。

2.2　新的疾病易感位点的发现与鉴定

人类基因组序列的测序表明，任何两个无关个体，99.9% 的基因组序列都相似，只有 0.1% 的序列不同，但就是这个比例很小的差异影响疾病以及对疾病的药物治疗。所以研究基因组序列上的变异或者多态性跟疾病的关系，是揭示复杂疾病遗传机制的重要手段。目前人们已经通过实验手段找到影响疾病的重要易感位点，这为疾病的预防预测及治疗提供了重要的依据。但大部分的疾病是复杂疾病，是受多个基因以及基因与环境共同作用的，每个基因单独的作用都比较弱。早在 1996 年，Risch 和 Merikangas 首先提出了常见疾病可能是由于常见基因变异引起的，关联分析通常比连锁分析具有更高的检测效率，而全基因组关联分析（genome-wide association study，GWAS）是发现人类复杂疾病相关遗传变异最有力和最有效的研究方法。随着单体型计划（HapMap）的完成，全基因组关联分析在欧洲开始兴起，短短几年，在《Nature Genetics》相继发表了与疾病及临床连续性状的 GWAS 研究，不作任何假设地寻找与疾病相关的易感位点，大大推动疾病遗传机制的研究。

2.3　功能相关基因表达谱的分析

明确了核酸序列，还应该了解它们是如何发挥功能的，即如何按照特定的时间、空

间进行表达以及表达的量是多少。人类基因组计划完成后，人类基因组研究的重心由结构基因组时代逐渐转向功能基因组时代，即通过对个体在不同生长发育阶段或不同生理病理状态下功能基因表达的横向或纵向分析，研究相应基因在生物体内的功能，阐明不同层次多基因协同作用的机制，进而在人类重大疾病如肿瘤、精神性疾病等的发病机制、分子诊断、药物开发等研究中提供重要的线索。随着各种平台的基因表达芯片的出现，对疾病进行基因表达检测产生海量的数据，如何从这些数据中挖掘出有用的信息，并且进行生物学专业解释，是生物信息在基因表达谱数据中面临的一个重要挑战。

2.4 生物大分子的结构模拟与药物设计

对生物大分子的结构模拟是用计算机对核酸特别是蛋白质的立体三维结构进行预测，包括蛋白质分子设计、RNA 的结构模拟、反义 RNA 的分子设计、蛋白质空间结构模拟和分子设计、不同功能域的复合蛋白质以及连接肽的设计等。药物分子设计结合计算机辅助，以计算机化学为基础，通过计算机的模拟、计算和预算药物与受体生物大分子之间的关系，即先模拟和计算受体与配体的相互关系，对先导化合物进行优化和设计，大致包括活性位点分析、数据库搜索和全新药物设计。因此，对生物大分子的结构设计与模拟、药物分子设计都强烈依赖于生物信息，需要发展新的方法和软件来实现。

2.5 新的生物信息技术方法和软件的研究

随着科学技术的发展，大量的生物数据的出现，急需发展新的分析方法和软件，从而能够有效地处理这些海量的数据。这些技术方法主要包括：第一，开发能处理全基因组测序数据的各种生物信息学分析软件、数据库以及若干数据库工具、电子网络等远程通信工具。第二，改进现有的理论分析方法，如统计方法、参数估计法、降维方法、神经网络方法等。第三，从系统生物学上创建适用于基因组信息分析的新方法、新技术，用于解释基因组的信息。第四，探索 DNA 序列及其空间结构信息的新表征，发展研究基因组完整信息结构和信息网络的研究方法等，发展生物大分子空间结构模拟、分子动态理论模拟和药物分子设计等。

2.6 在基因组及氨基酸水平上研究生物进化

随着基因组和蛋白质结构的测序，大量的 DNA 和氨基酸序列数据的大量增加，目前越来越多的研究转向利用这些序列进行生物之间以及功能基因功能位点的研究。首先，基于不同分子序列发现同一种群所重构出的进化树可能不同。另外，对“水平演化”与“垂直进化”间关系的讨论正逐渐引起人们的重视。也就是近年来发现了基因的“横向迁移现象”，即基因可以在同时存在的种群间迁移。其结果虽可导致序列差异，但这种差异与进化无关。甚至对人类基因组的分析发现，有几十个人类基因只与细菌基因

相似，而在果蝇、线虫中均不存在。另外，通过对功能基因或者基因家族的进化分析，可找到这些基因的正选择、负选择、中性选择位点。功能基因的进化分析将成为研究其功能的一条重要途径。

2.7 系统生物学研究功能基因组

基因在机体内都不是单独起作用的，而是多个基因相互作用，共同影响机体的机能。同时，基因表达的数目在不同的组织相差很大。例如，人脑中有 3 万 ~4 万个转录子，是基因表达的数目最多的组织，而有的组织中仅有几十或几百个基因表达。另外，基因表达在不同的个体生长发育阶段其种类、数量也是不同的，有些基因是在幼年时表达，有些是中年时表达，有些要到老年时才表达。对基因的掌握，需要在基因的序列、功能以及在不同的时间、不同的组织中基因的表达谱进行了解。因此，如何系统地研究这些基因在机体内的作用机制，依赖于生物信息学和实验室研究的结合。

第三节　生物信息学的发展

重点提示:

1. 人类基因组时代的生物信息学是基因组信息学，主要任务是大量核苷酸序列的测定、新基因的发现和鉴定。

2. 在后基因组时代，研究重心从关注生命遗传信息的解释转移到在系统水平上对生物功能的研究。

在生命科学研究中，生物信息学是以计算机为工具对生物信息进行储存、检索和分析的科学，是当今生命科学和自然科学的重大前沿领域之一，同时也将是 21 世纪自然科学的核心领域之一。对于生物信息学的发展历程，可以分为人类基因组时代的基因组信息学和后基因组时代的生物信息学。前者的生物信息学主要是核酸序列上的分析研究，后者主要是功能基因组和蛋白质组水平上的研究。

1　人类基因组计划时代的基因组信息学

随着人类基因组计划的完成，人们发现人类全部 23 对染色体的 3×10^9 个核苷酸有 3 万 ~4 万个人类基因，只是原来估计的 10 万 ~12 万个基因数的几分之一，这些基因是控制人类生命活动的遗传基础。目前认为基因组中约 95% 的核酸序列属于非编码区（所谓“Junk”DNA），它们的作用还不清楚。不过对于基因组 5% 可编码蛋白质的基因已经进行较为全面的研究，有些可用于指导基因药物的生产。HGP 研究的重要成果——序列