生物信息学基础

孙啸　陆祖宏　谢建明　编著

清华大学出版社

生物信息学基础

孙啸　陆祖宏　谢建明 编著

清华大学出版社
北京

内 容 简 介

生物信息学是一门新兴的交叉学科。在该领域中，由生物学家和计算机科学家共同研究生物分子信息的获取、管理、分析和利用。生物信息学以计算机、网络为工具，用数学和信息科学的理论、方法和技术去研究生物大分子，研究生物分子信息组织的规律。本书紧紧围绕基因组与后基因组研究，阐述生物信息学的方法、技术、资源及其核心算法，介绍各种信息学方法和技术在生物信息学中的应用。本书首先简要说明生物信息学的研究对象及主要研究内容；然后介绍基本的序列比较算法，介绍各种生物信息学数据资源及主要数据库；接下来以专题形式介绍基因组信息分析、分子系统发生分析及蛋白质结构预测；最后，介绍基因表达数据分析。为了便于计算机和数学研究人员进入生物信息学研究领域，本书还特别介绍了与生物信息学有关的基本分子生物学知识。

本书可以作为高年级大学生或研究生的生物信息学课程教材，也可以作为生命科学工作者、计算机应用人员的参考书。

图书在版编目(CIP)数据

生物信息学基础/孙啸，陆祖宏，谢建明编著. —北京：清华大学出版社，2005.5
ISBN 7-302-10270-8

Ⅰ. 生…　Ⅱ. ①孙…②陆…③谢…　Ⅲ. 生物信息论　Ⅳ. Q811.4

中国版本图书馆 CIP 数据核字(2005)第 002913 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969

组稿编辑：陈国新
文稿编辑：赵从棉
版式设计：肖　米
印 刷 者：北京密云胶印厂
装 订 者：北京市密云县京文制本装订厂
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：21.75　**字数**：515 千字
版　　次：2005 年 5 月第 1 版　2005 年 5 月第1 次印刷
书　　号：ISBN 7-302-10270-8/Q・42
印　　数：1～3000
定　　价：32.00 元

前　言

生物信息学是一门新兴的交叉学科。该领域的工作需要生物学和计算机科学这两门学科高级研究人员的通力合作。这两门学科差别很大，缺乏共同的语言，研究的方法也不一样，因此具有生物学背景的研究人员需要补充信息分析理论和计算机技术，掌握常用的分析工具；而来自计算机科学的研究人员则需要补充生物学知识，了解生物学特别是分子生物学中需要解决的信息分析问题。

随着分子生物学技术的不断进步和基因组研究的不断深入，生物分子数据迅速增长，数据量巨大，其中既有生物分子序列的信息，又有结构和功能的信息；既有生命本质信息，又有生命表象信息，并且数据之间存在着密切的联系。这些生物分子数据具有丰富的内涵，其背后隐藏着人类目前尚不知道的生物学知识。充分利用这些数据，通过数据分析、处理，揭示这些数据的内涵，得到对人类有用的信息，是生物信息学所面临的严峻挑战。生物信息学以计算机、网络为工具，用数学和信息科学的理论、方法和技术去研究生物大分子，发现生物分子信息组织的规律。其研究重点主要落实在 DNA 分子和蛋白质分子两个方面，包括它们的序列、结构和功能。

人才培养和专业技术人员培训是生物信息学目前的一个重要任务，国内迫切需要一本生物信息学专业教材。本书的主要作者早在 1999 年就开设了生物信息学研究生课程，在其他教师的支持下，根据自己的工作积累和国内外生物信息学的发展状况，在参阅了大量国内外资料的情况下，撰写了本书的初稿，形成生物信息学的课程讲义，并在随后几年的教学实践中不断修改，最终形成本书。

编著本书的目的就是为那些对生物信息学感兴趣的高年级大学生或研究生提供一本教科书。当然，对于那些刚刚进入生物信息学领域的研究人员，本书也可以作为基本的参考书。本书主要面向计算机专业的人员，重点介绍生物信息学的核心算法。首先简要介绍生物信息学研究对象及主要研究内容，介绍分子生物学基础，然后介绍基本的序列比较算法，介绍各种生物信息学数据资源，接下来以专题形式介绍基因组信息分析、分子系统发生分析及蛋白质结构预测，最后介绍目前生物信息学研究中的一个热点——基因表达数据分析。

陆祖宏教授是编著本书的倡议者，在本书的编写过程中提出了许多宝贵意见，并进行了全面审核。孙啸教授负责组织本书的编著工作，并编写了本书的大部分章节。谢建明博士编写了第 8 章，并承担了本书的大部分编辑工作。谢雪英博士编写了第 5 章，傅静编写了第 6 章的部分内容，陶怡、汤丽华、韦芬霞和顾珉参加了本书的校对工作。当作者在东南大学将本书作为研究生生物信息学课程教材试用时，许多学生对本书最初的文字、图

表及实例提出了好的建议，在此对他们表示衷心的感谢。

由于生物信息学是一门新兴的交叉学科，对生物学、数学及计算机科学的基础要求非常高，写好这样一本教材非常困难。书中的错误之处在所难免，恳切希望得到广大读者的批评和指正。

编　著

2004 年 5 月于东南大学

目　录

第 1 章　生物信息学引论 …… 1

1.1　引言 …… 1
1.1.1　生物信息学概念 …… 1
1.1.2　生物分子信息 …… 2
1.1.3　生物信息学的研究目标和任务 …… 4
1.1.4　生物信息学的研究意义 …… 6
1.2　生物信息学的发展历史 …… 7
1.3　人类基因组计划和基因组信息学 …… 9
1.3.1　人类基因组计划简介 …… 9
1.3.2　人类基因组计划对生物信息学的挑战 …… 13
1.4　蛋白质结构与功能关系的研究 …… 16
1.5　生物信息学的主要研究内容 …… 18
1.5.1　生物分子数据的收集与管理 …… 18
1.5.2　数据库搜索及序列比较 …… 19
1.5.3　基因组序列分析 …… 20
1.5.4　基因表达数据的分析与处理 …… 21
1.5.5　蛋白质结构预测 …… 21
1.6　生物信息学所用的方法和技术 …… 23
1.6.1　数学统计方法 …… 23
1.6.2　动态规划方法 …… 23
1.6.3　机器学习与模式识别技术 …… 24
1.6.4　数据库技术及数据挖掘 …… 25
1.6.5　人工神经网络技术 …… 26
1.6.6　专家系统 …… 27
1.6.7　分子模型化技术 …… 28
1.6.8　量子力学和分子力学计算 …… 29
1.6.9　生物分子的计算机模拟 …… 29
1.6.10　因特网(Internet)技术 …… 31
1.7　生物信息学目前的发展概况 …… 31

问题与练习 …… 35
参考文献 …… 35

第 2 章　生物信息学的生物学基础 …… 40

2.1　细胞 …… 40
2.2　蛋白质的结构和功能 …… 42
2.2.1　蛋白质的功能 …… 42
2.2.2　蛋白质的分子组成 …… 43
2.2.3　蛋白质的结构层次 …… 44
2.2.4　蛋白质结构与功能的关系 …… 50
2.3　遗传信息载体——DNA …… 51
2.3.1　核苷酸 …… 52
2.3.2　DNA 的结构 …… 53
2.4　分子生物学中心法则 …… 55
2.4.1　DNA 的复制 …… 55
2.4.2　转录 …… 56
2.4.3　翻译 …… 57
2.4.4　mRNA 的反转录与 cDNA …… 59
2.4.5　对遗传信息流的再认识 …… 60
2.5　基因组结构 …… 60
2.5.1　染色体结构 …… 60
2.5.2　基因 …… 62
2.5.3　原核生物基因组 …… 63
2.5.4　真核生物基因组 …… 64
2.6　基因表达调控 …… 69
2.6.1　基因表达调控的层次 …… 69
2.6.2　原核基因调控 …… 70
2.6.3　真核基因调控 …… 70
2.7　新生肽链的折叠 …… 71
2.7.1　新生肽链的加工 …… 72
2.7.2　新生肽链的折叠 …… 72
2.7.3　蛋白质折叠的一般规律 …… 72
2.7.4　帮助新生肽链折叠的生物大分子 …… 73
2.7.5　蛋白质构象病问题 …… 74
2.8　生物大分子结构的测定 …… 74
2.8.1　X 射线衍射结构分析 …… 74
2.8.2　核磁共振结构分析 …… 76
2.9　分子生物学工具 …… 77

问题与练习 …… 79
参考文献 …… 79

第 3 章　序列比较 …… 81

3.1　序列的相似性 …… 81
3.1.1　字母表和序列 …… 82
3.1.2　编辑距离 …… 83
3.1.3　通过点矩阵分析两条序列的相似之处 …… 84
3.1.4　序列的两两比对 …… 86
3.1.5　用于序列相似性的打分矩阵 …… 87
3.2　两两比对算法 …… 92
3.2.1　序列两两比对基本算法 …… 93
3.2.2　子序列与完整序列的比对 …… 96
3.2.3　寻找最大的相似子序列 …… 97
3.2.4　准全局序列比对 …… 98
3.2.5　关于连续空位的问题 …… 99
3.2.6　比较相似序列 …… 102
3.2.7　比对的统计学显著性 …… 103
3.3　序列多重比对 …… 104
3.3.1　SP 模型 …… 105
3.3.2　多重比对的动态规划算法 …… 107
3.3.3　优化计算方法 …… 110
3.3.4　星形比对 …… 112
3.3.5　树形比对 …… 114
3.3.6　其他多重序列比对算法 …… 115
3.3.7　统计特征分析 …… 115
3.4　DNA 片段组装 …… 116
3.4.1　片段组装问题 …… 117
3.4.2　序列片段组装模型 …… 119
3.4.3　序列片段覆盖图 …… 121
3.4.4　贪婪算法 …… 123
3.4.5　非循环图拓扑排序法 …… 124
问题与练习 …… 125
参考文献 …… 126

第 4 章　生物分子数据库 …… 130

4.1　引言 …… 130
4.2　核酸序列数据库 …… 131

4.2.1 GenBank / EMBL-Bank / DDBJ …… 131
4.2.2 基因组数据库 …… 136
4.2.3 表达序列标记数据库 dbEST …… 137
4.2.4 序列标记位点数据库 dbSTS …… 138
4.2.5 面向基因聚类数据库 UniGene …… 138
4.3 蛋白质序列数据库 …… 138
4.3.1 PIR …… 138
4.3.2 SWISS-PROT …… 140
4.3.3 TrEMBL …… 141
4.4 生物大分子结构数据库 …… 142
4.4.1 PDB …… 142
4.4.2 MMDB …… 142
4.5 其他生物分子数据库 …… 143
4.5.1 单碱基多态性数据库 dbSNP …… 144
4.5.2 蛋白质结构分类数据库 SCOP …… 144
4.5.3 蛋白质二级结构数据库 DSSP …… 145
4.5.4 蛋白质同源序列比对数据库 HSSP …… 146
4.5.5 序列模式数据库 PROSITE …… 147
4.5.6 蛋白质指纹数据库 PRINTS …… 147
4.5.7 人类遗传数据库 OMIM …… 147
4.5.8 基因启动子数据库 EPD …… 148
4.5.9 转录调控区域数据库 TRRD …… 148
4.5.10 转录因子数据库 TRANSFAC …… 149
4.5.11 基因本体数据库 GO …… 149
4.5.12 生物、医学文献数据库 PubMed …… 149
4.5.13 目录数据库 DBCat …… 149
4.6 数据库搜索 …… 150
4.6.1 FastA …… 151
4.6.2 BLAST …… 154
4.6.3 VAST …… 158
4.7 数据库集成 …… 159
4.7.1 Entrez …… 160
4.7.2 SRS …… 161
4.7.3 ExPASy …… 162
问题与练习 …… 162
参考文献 …… 163

第 5 章　基因组信息分析 …… 168

5.1　关于遗传语言 …… 168
5.1.1　基因组 DNA 的奥秘 …… 168
5.1.2　探索遗传语言 …… 171
5.1.3　关于生物复杂性 …… 172
5.1.4　基因组学研究带来的希望 …… 172
5.2　原核基因组特点 …… 173
5.2.1　长开放阅读框 …… 173
5.2.2　高基因密度 …… 173
5.2.3　简单的基因结构 …… 173
5.2.4　原核基因组中的 GC 含量 …… 174
5.3　真核基因组特点 …… 174
5.3.1　基因组规模 …… 174
5.3.2　巨大的非编码序列 …… 174
5.3.3　复杂的基因结构 …… 174
5.3.4　复杂的基因转录调控方式 …… 175
5.3.5　可变剪接 …… 175
5.3.6　CpG 岛 …… 176
5.3.7　等值区 …… 176
5.3.8　密码子使用偏性 …… 177
5.4　基因组序列分析 …… 177
5.4.1　基因组序列分析步骤和分析结果评价 …… 177
5.4.2　核苷酸关联分析 …… 179
5.5　基因识别方法 …… 181
5.5.1　最长 ORFs 法 …… 181
5.5.2　基于密码子出现频率的预测方法 …… 182
5.5.3　同源性方法 …… 184
5.5.4　神经网络方法 …… 185
5.5.5　隐马尔可夫模型法 …… 186
5.5.6　模式判别分析法 …… 198
5.5.7　基于动态规划的基因结构预测方法 …… 199
5.5.8　基于剪切比对的基因识别 …… 202
5.5.9　其他基因识别方法 …… 202
5.6　非编码区域分析和调控元件识别 …… 203
5.6.1　调控元件的建模 …… 204
5.6.2　调控元件模式的得分函数 …… 206
5.6.3　模式驱动的调控元件识别 …… 207

5.6.4 序列驱动的调控元件识别…… 208
问题与练习…… 215
参考文献…… 215

第6章 系统发生分析…… 219

6.1 分子系统发生与系统发生树 …… 219
6.1.1 分子系统发生分析…… 219
6.1.2 系统发生树…… 221
6.1.3 距离和特征…… 222
6.1.4 分子系统发生分析过程…… 223
6.2 基于距离的系统发生树构建方法 …… 225
6.2.1 最小二乘法…… 225
6.2.2 连锁聚类方法及非加权分组平均法…… 226
6.2.3 距离变换法 …… 229
6.2.4 邻近归并法…… 230
6.3 基于特征的系统发生树构建方法 …… 232
6.3.1 最大简约法…… 232
6.3.2 快速搜索策略…… 235
6.4 最大似然法 …… 236
6.5 系统发生树的可靠性 …… 238
6.5.1 自举检验…… 238
6.5.2 参数检验…… 239
6.6 全基因组系统发生分析 …… 239
6.6.1 基于多棵系统发生树的方法…… 239
6.6.2 基于基因内容的方法…… 240
6.6.3 基于蛋白质折叠结构的方法…… 240
6.6.4 基于基因次序的方法…… 240
6.6.5 基于连接的直向同源蛋白的方法…… 240
6.6.6 基于代谢途径的方法…… 241
问题与练习…… 242
参考文献…… 243

第7章 蛋白质结构预测…… 245

7.1 引言 …… 245
7.2 蛋白质二级结构预测 …… 249
7.2.1 利用的信息及预测准确性…… 249
7.2.2 Chou-Fasman 方法 …… 250
7.2.3 GOR 方法 …… 252

7.2.4 基于氨基酸疏水性的预测方法…… 255
7.2.5 最邻近方法…… 257
7.2.6 人工神经网络方法…… 258
7.2.7 综合方法…… 261
7.2.8 氨基酸残基之间的距离…… 261
7.3 RNA 二级结构的预测…… 262
7.4 蛋白质空间结构预测 …… 263
7.4.1 同源模型化方法…… 264
7.4.2 线索化方法(折叠识别方法)…… 266
7.4.3 从头预测方法…… 267
7.4.4 预测方法评价…… 272
7.5 蛋白质空间结构比较 …… 273
问题与练习…… 275
参考文献…… 276

第 8 章 基因表达数据分析…… 282

8.1 基因表达数据的获取 …… 283
8.1.1 cDNA 微阵列…… 283
8.1.2 寡核苷酸芯片…… 284
8.1.3 基因表达数据的网络资源…… 285
8.2 基因表达数据预处理 …… 286
8.3 基因表达差异的显著性分析 …… 289
8.3.1 倍数分析…… 289
8.3.2 *t* 检验 …… 290
8.3.3 贝叶斯分析…… 291
8.4 基因表达谱聚类分析 …… 292
8.4.1 相似性度量函数…… 292
8.4.2 聚类方法…… 294
8.4.3 基于模型的聚类方法…… 298
8.4.4 支持向量机…… 299
8.4.5 聚类结果的可视化…… 301
8.4.6 聚类结果的定量评价…… 303
8.5 基因表达数据的分类分析 …… 305
8.5.1 朴素贝叶斯分类法…… 305
8.5.2 *k*-近邻法 …… 306
8.5.3 其他分类法…… 306
8.6 主成分分析 PCA …… 307
8.7 基于基因表达谱的基因调控网络研究 …… 309

8.7.1 布尔网络模型 …… 310
8.7.2 线性组合模型 …… 312
8.7.3 加权矩阵模型 …… 312
8.7.4 数据整合分析 …… 313
问题与练习 …… 314
参考文献 …… 314

附录 1 常用基本词汇表 …… 320

附录 2 生物信息分析工具 GCG …… 333

第 1 章

生物信息学引论

20 世纪是科学技术迅速发展的世纪，物理和化学的发展使我们可以清楚地认识物质的组成，从分子、原子、电子等各层次上深入地了解微观世界；而天文技术、空间技术的发展则使得我们可以了解地球以外的客观世界；以电子信息技术为龙头的工业技术的飞速发展，使得我们可以不断地改造世界，甚至为人类更加舒适地生活创造新的世界。生命科学在 20 世纪同样也得到了发展，生理学、细胞生物学、分子生物学等学科的发展使我们从器官、组织、细胞及生物大分子等各个层次认识了生命的物质基础。生物与其他物质有本质的区别，生物并非只是物质的简单堆积，生物体的生长发育是生命信息控制之下的复杂而有序的过程。目前，我们对生命的奥秘还不甚了解，对生命信息的组织、传递和表达还知之甚少。既然这牵涉到信息的组织、传递和表达，我们就可以用信息科学的方法和技术来尝试认识和分析生命信息。

1.1 引言

传统的生物学是一门实验科学，生物学研究依赖于对实验数据的处理和分析。生物学也是一门发现科学，通过实验发现新的现象、新的生物学规律，经过分析、归纳和总结，提炼出新的生物学知识。在这个过程中，需要对实验数据进行处理和理论分析，并在此基础上解释实验现象，认识导致实验现象发生的本质，探索固有的生物学规律，进而了解和掌握生命的物质基础和生命的本质。随着生物科学和技术的迅速发展，生物数据的积累速度将不断加快，因此，也就对生物数据的科学分析方法和实用分析工具提出了更新、更高的要求。

1.1.1 生物信息学概念

人类为了更深入地了解和认识自身，制定了宏伟的人类基因组计划。人类基因组计划顺利实施，产生了大量的生物分子数据。据权威机构统计，目前生物分子数据量每 15 个月翻一番，生物分子数据发展的速度甚至超过了摩尔定律(即半导体芯片上的晶体管数量每 18 个月翻一番)。这些生物分子数据具有丰富的内涵，其背后隐藏着人类目前尚不知道的生物学知识。充分利用这些数据，通过数据分析、处理，揭示这些数据的内涵，从而得到对人类有用的信息，是生物学家、数学家和计算机科学家所面临的一个严峻的挑战。

生物信息学就是为迎接这种挑战而发展起来的一门新型学科，它是由生物学、应用数学和计算机科学相互交叉所形成的学科，是当今生命科学和自然科学的重大前沿领域之一，也是21世纪自然科学的核心领域之一。

生物信息学(bioinformatics)这个名词有许多不同的定义。从字面上来看，生物信息学是将信息科学和技术应用于生物学。生物信息学广义的概念是指应用信息科学的方法和技术，研究生物体系和生物过程中信息的存储、信息的内涵和信息的传递，研究和分析生物体细胞、组织、器官的生理、病理、药理过程中的各种生物信息，或者也可以说成是生命科学中的信息科学。生物信息学狭义的概念是指应用信息科学的理论、方法和技术，管理、分析和利用生物分子数据。通过收集、组织、管理生物分子数据，使研究人员能够迅速地获得和方便地使用相关信息；通过处理、分析、挖掘生物分子数据，得到深层次的生物学知识，加深对生物世界的认识。在生物学、医学的研究和应用中，利用生物分子数据及其分析结果，可以大大提高研究和开发的科学性及效率，如根据基因功能分析结果来检测与疾病相关的基因，根据蛋白质分析结果进行新药设计。一般提到的“生物信息学”就是指这个狭义的概念，更准确地说，应该是分子生物信息学(molecular bioinformatics)。

生物信息学以计算机、网络为工具，采用数学和信息科学的理论、方法和技术去研究生物大分子，其研究重点主要落实在核酸和蛋白质两个方面，包括它们的序列、结构和功能。生物信息学以基因组DNA序列信息分析作为出发点，破译遗传语言，认识遗传信息的组织规律，辨别隐藏在DNA序列中的基因，掌握基因调控信息，对蛋白质空间结构进行模拟和预测，依据蛋白质结构和功能的关系进行药物分子设计。与生物信息学相关的概念还有计算分子生物学(computational molecular biology)，计算分子生物学主要研究分析方法，开发分析工具，促进生物分子数据的分析。与生物信息学相关的另一个名词是生物计算(biocomputing)，生物计算特指用计算机技术分析和处理生物分子数据。

生物信息学的产生一方面是由于生物科学和技术的发展，另一方面是由于人类基因组计划的实施。其实，早在20世纪50年代生物信息学就已经形成萌芽，20世纪70年代已经产生生物信息学的基本思想，但是生物信息学的真正发展则是在20世纪90年代，在人类基因组计划的推动下，生物信息学才得以迅猛发展。人类基因组计划产生的生物分子数据是生物信息学的源泉，而人类基因组计划所需要解决的问题则是生物信息学发展的动力。

1.1.2 生物分子信息

生物体是一个复杂的系统，生命过程是一个极端复杂的过程，需要物质和能量的支持。生物体同时也是一个信息系统，该系统控制着生物的遗传、生长和发育。所有的信息都存储在生物体内的遗传物质中。在生命科学的研究中，人们已经逐渐认识到，不仅需要用物理、化学和生物学方法研究生命的物质基础、能量转换、代谢过程等，还需要用信息科学方法研究生命信息特别是遗传信息的组织、复制、传递、表达及其作用，否则难以理解生命的工作机制，难以揭示生命的奥秘。从生物学的观点来看，细胞是生命的基本单位，而从信息科学的观点来看，细胞则是存储、复制和传递遗传信息的系统。

生物系统通过存储、修改、解读遗传信息和执行遗传指令形成特定的生命活动，促使

生物体生长发育，产生生物进化。从信息学的角度来看，生物分子是生物信息的载体，生物信息学主要研究两种载体，即 DNA 分子和蛋白质分子。生物分子至少携带着 3 种信息，即遗传信息、与功能相关的结构信息和进化信息。

DNA 是遗传信息的载体。DNA 的核苷酸序列上存储着蛋白质的氨基酸序列编码信息，存储着基因表达调控的信息，存储着遗传信息。遗传信息存储在 DNA 四种字符组成的序列中，生物体生长发育的本质就是遗传信息的传递和表达。因此，可以说 DNA 序列包含着最基本的生命信息。存储在 DNA 中的信息使无活力的分子组织成有功能的活细胞，进而构成能进行新陈代谢、生长和繁殖的生物体。人们已经认识到遗传信息的载体主要是 DNA(在少数情况下核糖核酸即 RNA 也充当遗传信息的载体)，控制生物体性状的基因是一系列 DNA 片段。一方面，DNA 通过自我复制，在生物体的繁衍过程中传递遗传信息；另一方面，基因通过转录和翻译，使遗传信息在生物个体中得以表达，并使后代表现出与亲代相似的生物性状。在基因表达过程中，基因上的遗传信息首先通过转录从 DNA 传到 RNA，然后再通过翻译从 RNA 传递到蛋白质。基因控制着蛋白质的合成，从基因的 DNA 序列到蛋白质序列存在着一种明确的对应关系，而这种对应关系就是我们所知道的第一遗传密码。

蛋白质分子在生物体内执行着各项重要任务，如生化反应的催化、营养物质的输运、信号的识别与传递等。蛋白质的功能多种多样，但是必须注意一点，即蛋白质功能取决于蛋白质的空间结构。要了解和掌握蛋白质的功能必须首先分析蛋白质的结构，对于其他生物大分子也一样。因此，蛋白质结构是一种重要的生物分子信息。然而，蛋白质结构决定于蛋白质的序列(这是目前基本共认的假设)，蛋白质结构的信息隐含在蛋白质序列之中。

作为信息的载体，DNA 分子和蛋白质分子都打上了进化的烙印。通过比较相似的蛋白质序列，如肌红蛋白和血红蛋白，可以发现由于基因复制而产生的分子进化证据。比较来自于不同种属的同源蛋白质，即直系同源蛋白质，可以分析蛋白质甚至种属之间的系统发生关系，推测它们共同的祖先蛋白质。

生物分子信息具体表现为 DNA 序列数据、蛋白质序列数据、生物分子结构数据、生物分子功能数据等。序列数据、结构数据是非常直观的，但是功能数据却是多变复杂的，如关于蛋白质功能的定性描述、蛋白质之间的相互作用描述、基因表达数据、代谢路径、调控网络等。在所有类型的数据中，序列是最基本的数据，而且也是目前最多的数据。

对生物分子数据及其关系的概括见图 1.1。遗传信息从 DNA 序列向蛋白质序列的传递是人类已经基本了解的第一部遗传密码，然而蛋白质序列与蛋白质结构之间也存在着一定的对应关系，蛋白质序列决定蛋白质结构，因此有人将从蛋白质序列到蛋白质结构的关系称为第二部遗传密码。

第一部遗传密码已被破译，但是，对于密码究竟处于 DNA 序列的哪些区域还了解得不全面，对密码的转录过程还不清楚，对大多数 DNA 非编码区域的功能还知之甚少，对 DNA 遗传语言还有待于进一步探索。对于第二部密码，目前则只能用统计学的方法进行分析。无论是第一部遗传密码，还是第二部遗传密码，都隐藏在大量的生物分子数据之中。生物分子数据是宝藏，生物信息数据库是金矿，等待我们去挖掘和利用。

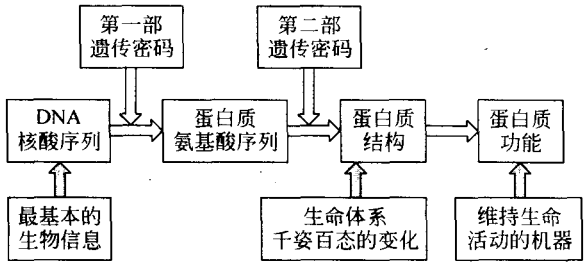

图 1.1　生物分子数据及其关系

与一般信息相比，生物分子信息具有明显的特征。首先，生物分子信息数据量大，例如 DNA 序列以千兆碱基(giga base，Gb)为单位。随着信息处理技术进入现代生物学研究领域，随着互联网在全球的贯通，各种生物信息学数据库迅速发展，生物分子数据的积累速度成倍增长。其次，生物分子信息复杂，既有生物分子序列的信息，又有结构和功能的信息，既有生命本质信息，如基因，又有生命表象信息，如基因表达信息。生物分子信息另一个重要的特征是，生物分子信息之间存在着密切的联系，例如，基因序列与蛋白质序列之间的关系，生物分子序列与结构之间的关系，结构与功能之间的关系，基因变异与疾病之间的关系。

对于生物分子信息，靠人工难以完成数据处理和分析的任务，更谈不上发现隐藏在这些信息之中的内在规律。同时，对于生物分子信息，仅靠某一学科的专家，也无法进行分析研究，因此，在生物信息学研究领域中，要求生物学家、数学家和计算机科学工作者协力合作，发展新的分子生物学计算理论和方法，运用先进的计算机技术收集、集成和分析处理生物信息。

1.1.3　生物信息学的研究目标和任务

揭示生物分子数据的内涵是生物信息学的长远目标。生物分子数据具有深刻的内涵，数据之间存在着复杂的联系，这些数据中蕴涵着丰富的生物学知识和生物学规律。生物信息学的发展将揭示生物分子信息的本质，使人类彻底了解、掌握遗传信息的编码、传递及表达，从而加快人类了解自身的进程。

目前生物信息学的主要任务是研究生物分子数据的获取、存储和查询，发展数据分析方法。主要包括 3 个方面。

第一是收集和管理生物分子数据，使得生物学研究人员能够方便地使用这些数据，并为信息分析和数据挖掘打下基础。生物分子数据来自于生物学实验，应用信息学技术收集和管理这些数据，将各种数据以一定的表示形式存放在计算机中，建立数据库系统，并提供数据查询、搜索和数据通信工具。

第二是进行数据处理和分析。通过数据分析，发现数据之间的关系，认识数据的本质，进而上升为生物学知识。并在此基础上，解释与生物分子信息复制、传递和表达有关的生物过程，解释在生物过程中出现的信息变化与疾病的关系，帮助发现新的药物作用目标，设计新的药物分子，为进一步的研究和应用打下基础。生物分子信息处理流程见图