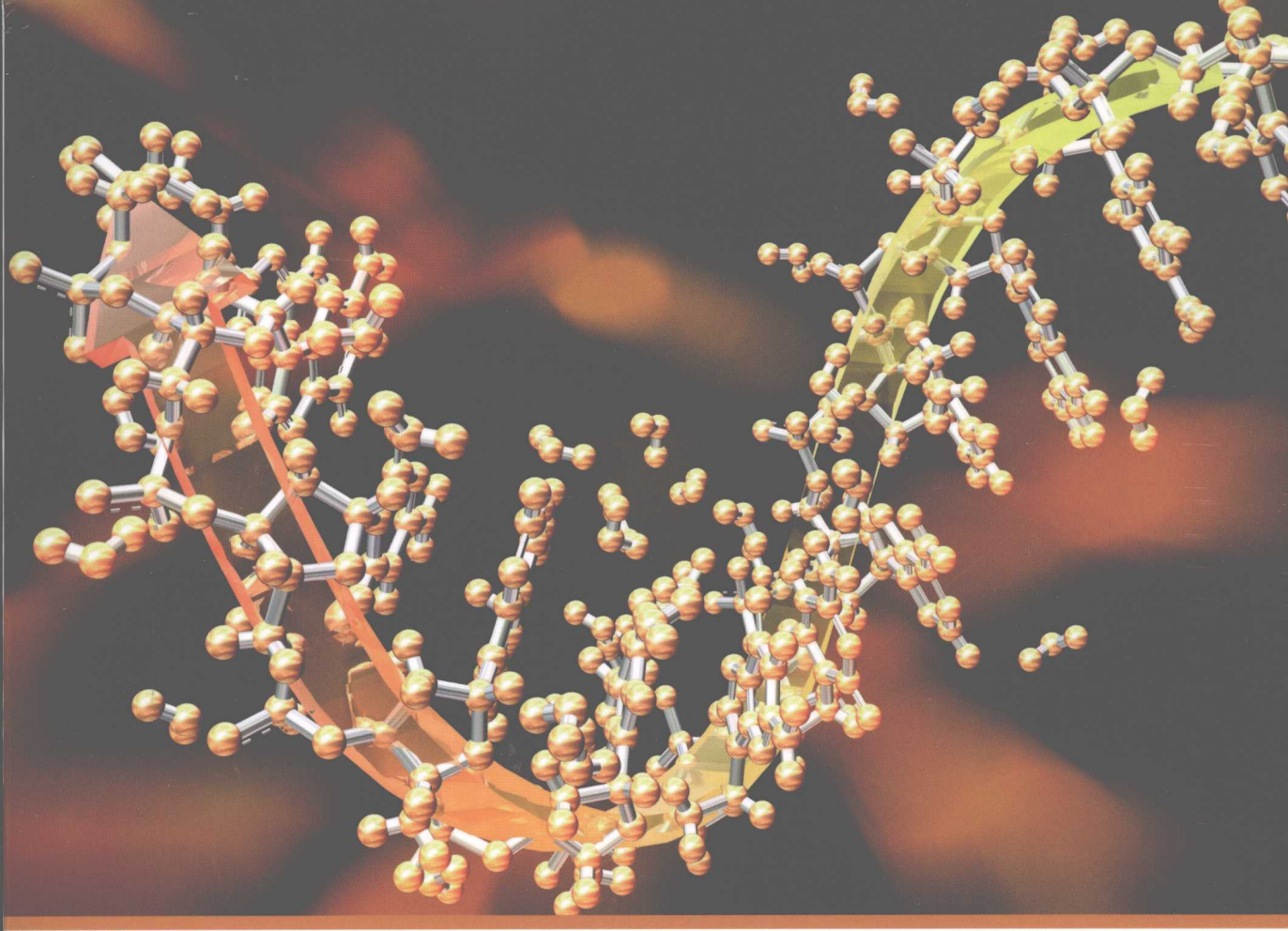
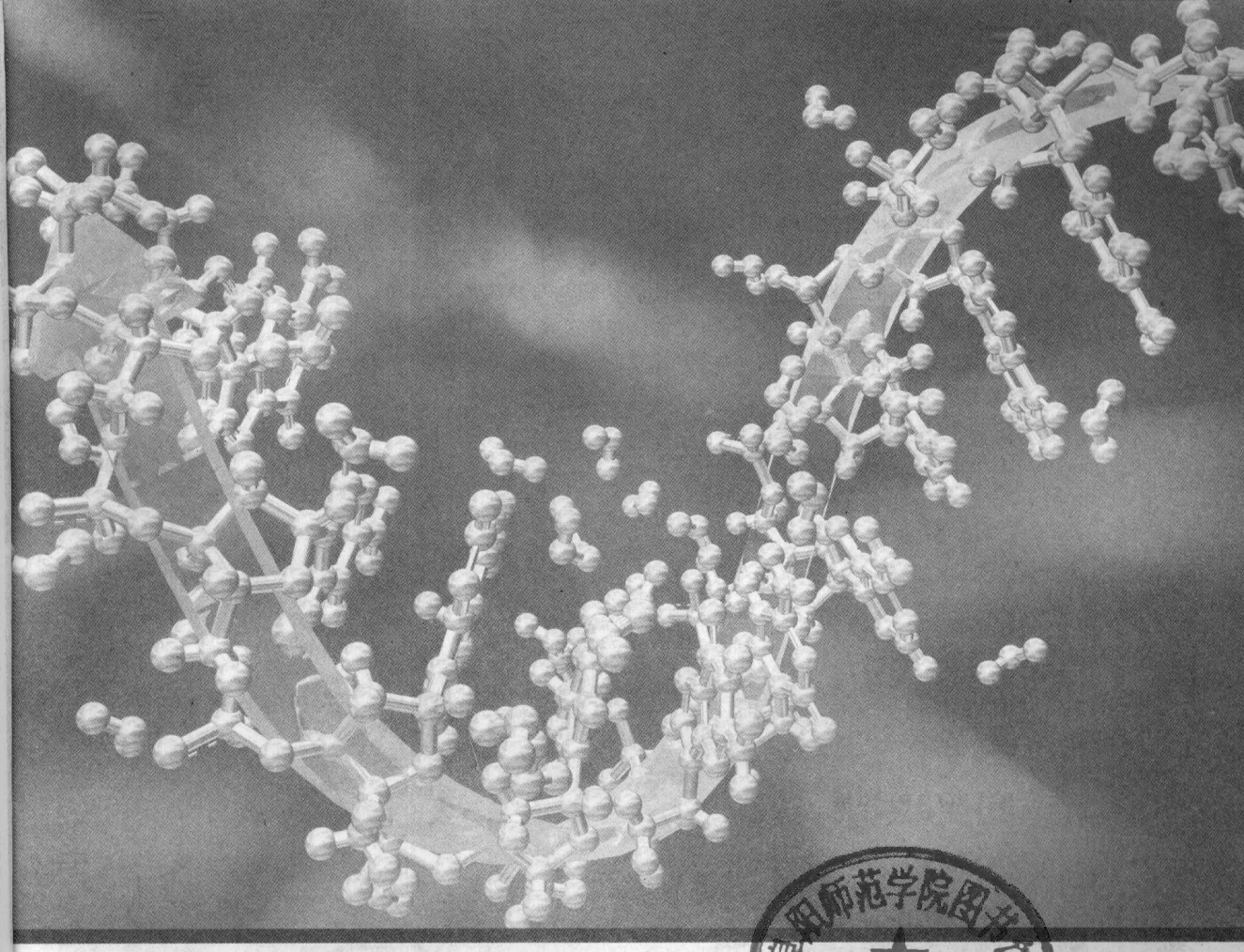


• 以生物化学或分子生物学中的经典实例为素材，使得生物信息学不再仅仅是“枯燥”的计算机问题，而是最熟悉不过的生物学问题； • “step by step”式的工具软件操作流程使得相关软件的使用简明易学； • 花费30分钟实践一个案例，即可了解生物大分子三维结构图像显示与分析技术； • 提供共享软件及操作实例的网页地址(www.bioinfo.sdu.edu.cn)，方便读者下载使用； • 介绍“Origin”和“Reference Manager”软件，详解数据处理和文献分析技术。

生物信息学应用技术

◎ 王禄山　高培基　主编

化学工业出版社
生物·医药出版分社



山东大学研究生教材基金资助

生物信息学应用技术

◎ 王禄山　高培基　主编

化学工业出版社
生物·医药出版分社
·北　京·

本书从生物大分子转化成生物数据（残基序列、原子坐标等）过程开始，介绍了生物信息数据及其存放的格式、数据库的分析工具与检索策略；结合当前生物信息学技术发展趋势，全书按照序列→结构→动力学→功能的思路进行组织，为读者认识与分析生物学规律提供新的思路；背景、原理、方法和分析操作相结合，是一本实用的生物信息学实验手册与操作指南。

本书取材精当，讲述简明，面向生命科学各专业及部分基础医学的读者，可供广大生物信息学入门及提高的读者参考使用。

图书在版编目(CIP)数据

生物信息学应用技术 / 王禄山，高培基主编. —北京：化学工业出版社，2007.9

ISBN 978-7-122-01076-6

Ⅰ. 生… Ⅱ. ①王…②高… Ⅲ. 生物信息论-应用 Ⅳ. Q811.4

中国版本图书馆 CIP 数据核字（2007）第 137433 号

责任编辑：孟 嘉 邵桂林　　装帧设计：潘 峰

责任校对：凌亚男

出版发行：化学工业出版社 生物·医药出版分社

（北京市东城区青年湖南街 13 号 邮政编码 100011）

印 装：北京市兴顺印刷厂

787mm×1092mm 1/16 印张 16¾ 字数 375 千字 彩插 4 2008 年 1 月北京第 1 版第 1 次印刷

购书咨询：010-64518888（传真：010-64519686） 售后服务：010-64518899

网 址：http:// www.cip.com.cn

凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定 价：39.00 元

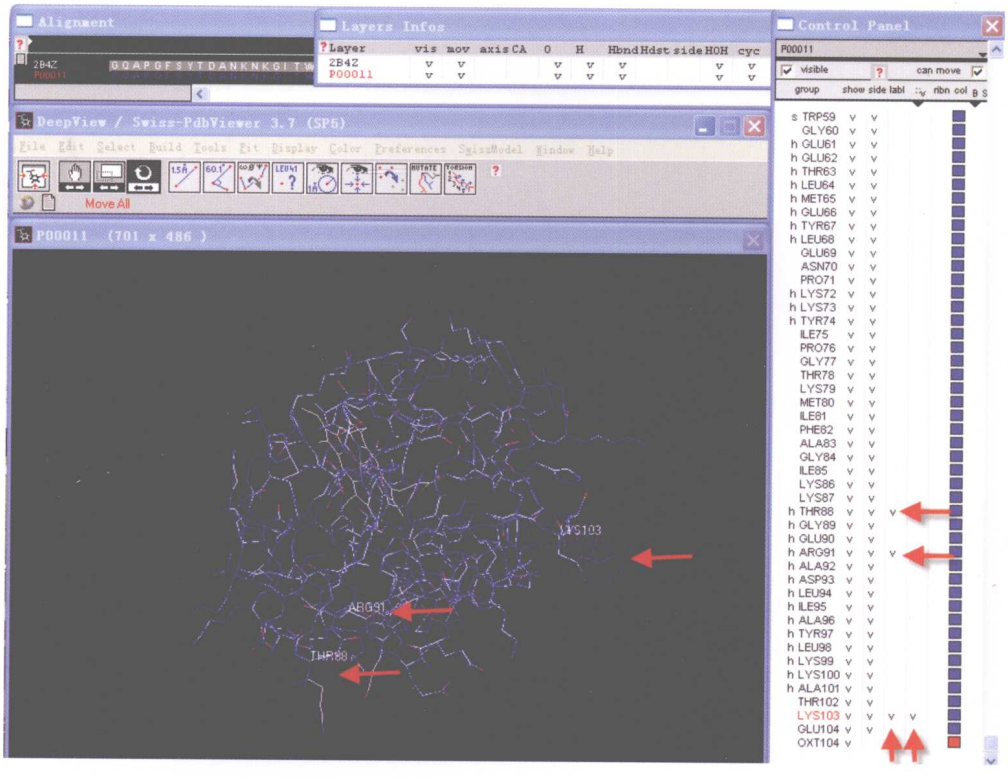


图 8-15　Swiss-PdbViewer 3.7显示界面

编写人员名单

主　编　王禄山　高培基

参编人员（以姓氏汉语拼音为序）

安真仪　高培基　侯　峰

焦绪瑶　孟　静　王禄山

王宇枭　邢　晟　张　正

序

生物信息学或者说生物信息学技术在中国的兴起始于上世纪后期对蛋白质结构功能的分析和理性改造(即在某些项目中所谓的蛋白质工程),盛于近十余年来基因组(特别是人类基因组)和功能基因组的研究和应用。随着在“组学”基础上的系统生物学的迅速发展,生物信息学技术的研究水平和应用范围正在走向新的高度和广度。

为了普及生物信息学技术和培养生物信息学人才,自上世纪末以来,我国出版了不少自编或翻译的专著,其中也不乏以教学为目的的“综述性”教科书。但是,我读王禄山、高培基先生所主编的《生物信息学应用技术》一书,仍有耳目一新之感。

该书并不求大求全,实实在在地为大学学子提供从事生物信息学分析的基本知识与应用技术,甚至包括了图表绘制和文献管理软件的内容。我想,这可能就是得益于著述的各位先生都是亲自从事教学的老师,他们最了解学生的需求,最掌握教学的方法。因此,该书也应该最能得到教学第一线师生的欢迎。

该书虽然并不刻意追求“深奥”和“创新”,却准确地把握了基因组生物信息学分析中的若干关键概念及相应的技术。譬如,对序列比对中“相似”与“同源”概念的差别以及中性进化概念在分子系统分析中关键作用的论述等,都是难能可贵的亮点。

该书并不局限于基因组生物信息学技术,还包括了蛋白质结构分析和模建技术的论述。在这一点上,我清楚地看到了高培基先生的学问和思想。高先生作为中国微生物界德高望重的前辈,数十年如一日地研究纤维素酶,从应用到理论,从微生物学到酶学都作出了杰出的贡献。现在,他老当益壮,提携后进,与青年教师一起,为莘莘学子编写关于生物信息学技术这一创新领域的教科书。我真是由衷钦佩高先生持之以恒、精益求精、不断进步的精神。

我也愿以高先生为榜样,继续努力,为中国生物信息学事业的发展尽自己绵薄之力。为此,不惴浅薄,勉为此序,共图进步。

赵国屏

2007 年 10 月 14 日

于中国科学院上海植物生理生态研究所

前 言

自上世纪50年代，微生物学、生物化学技术的发展对生命科学研究发挥了重要的推动作用。随后建立起的分子生物学技术，使得人们可以对生命系统中遗传信息的传递、储存和表达能进行操作、改造与利用。时至今日，随着计算机技术的引入，大规模、高通量测序与生物信息的存储与管理，使得人类基因组计划能够在新世纪伊始就顺利完成，并可以在全世界范围进行信息的共享与交流，生物信息学技术的快速发展对生命科学研究的巨大推动作用已初见端倪，正方兴未艾。

早在1985年，山东大学就开设了基因、蛋白质序列分析的研究生选修课，并相继开展了这方面的研究工作。1999年在国内高校中山东大学率先开设了生物信息学研究生课程，根据当时的讲义我们编写出版了《基因组与生物信息学》教材，对研究生基因组与生物信息学教学的开展起了一定推动作用。在这些年的教学与实践中，我们体会到生物信息学是一门实践性很强的学科，特别是计算机技术与互联网技术的飞速发展，共享的生物信息数据提供了丰富的实验材料，共享的生物信息学软件提供了可视化的实验方案与流程，共享的专家数据库提供了先进的思路与宝贵的经验，所以对于生物信息学，不仅研究生需要进行系统的学习与研究，本科生只要有一台能够上网的计算机同样可以进行相关理论的学习与研究。

但是，如何对学生进行生物信息学的教学与培训，怎样才能使学生了解生物信息学相关知识与技术仍然值得探索。由于生物信息学还处于飞速发展的时期，各方面的科研进展突飞猛进，新理论、新方法、新发现层出不穷。纵观国内外不同时期的生物信息学专著与教材，读者会发现她的“日新月异”！科研专著可以如此编写，但教材这样组织就会带来很大的问题：因为对那些未了解生物信息学的相关人员或学生来说，他们较缺乏相关知识，不能够判断最新的进展意味着什么；从知识积累的角度看，新的科研进展并不一定会凝结成新的知识，因为知识是共性的，要经得起时间验证的。所以，生物信息学技术的教学要选择基础的、共性的、与其他基础学科有联系有交叉而又有所发展的教学素材，这样读者在学习的过程中才能了解生物信息学所处理的生物学问题、才能从中体会到生物信息学技术处理的优势，从而能够真正激发读者的学习兴趣。按照这个目标，编者在素材的选择上注重利用生物化学或分子生物学教科书中经典的例子进行生物信息学的分析，如：在介绍序列分析时，选用基因工程载体——pUC19质粒进行绘图分析，选用乳糖操纵子进行基因序列的分析；在系统发育分析的例子中，选用了分子水平提出进化“分子钟”假说证据之一：电子传递链中水溶性蛋白——细胞色素c分子的进化分析；在结构显示与分析中选择蛋白酶的活性中心与底物结构中心，等等。这样做的目的是使读者在进行生物信息技术的学习过程中，可以感觉到是在进行生物问题的分析，而不仅是陷入应用计算机对序列进行“比对”

等机械性的演算，同时读者对问题的分析还可以直接参考相关教材，从而使读者体会到生物信息学技术的优势所在。更进一步，读者从经典例子与素材的分析中可以更加深入地了解应用统计方法进行生物信息技术分析时所遇到的问题，对读者以后应用生物信息技术分析解决生物学问题有所启迪。

生命科学在进入分子水平以后，生物与化学、物理学的界限在模糊，在理论上趋于统一。物理、化学中新的计算分析方法（如分子动力学模拟等）也开始应用到生物大分子的分析中，人们可以沿“序列→结构→动力学→功能”过程预测生物学的新功能。新方法、新技术在生物学中的应用大大加速了人们了解生物数据中所蕴藏的生命意义。如何介绍这些新方法、新技术，编者同意斯坦福大学 Russ Alman 教授的观点：“生物信息学方法是基因组革命的一部分，因此在解释生物信息学工具时，必须描述开发这个工具的生物学背景与问题”。所以，在本书总体编写中，我们以“生物信息”如何转变成“生物数据”，到“数据存储”及其“检索与分析”为主线进行组织，在每一章的概述中我们简要介绍相关生物学问题、计算方法与流程及发展现状，列出分析相关共享软件的地址，以及①级编号标注的详细操作流程；同时，我们将菜单操作具体过程以“→”引出方便读者学习与操作。另外，“生物信息学技术”也应该是广义的，应该包括数据处理技术与文献分析技术，所以本书也将有关材料归为两章。编者相信生物信息学的相关技术不仅可促进读者生物信息学相关知识与技能的学习，也可以促进生物化学，分子生物学的科研与教学。

中国科学院院士、国家人类基因组南方研究中心赵国屏研究员欣然作序，中国科学院院士、上海交通大学的邓子新教授，中国科学院海洋研究所的周百成研究员，中国科学院微生物研究所的周培瑾研究员，内蒙古大学物理系的罗辽复教授等给本书提出非常有价值的建议和意见，化学工业出版社对本书的编写给予了热情的支持和帮助，山东大学微生物技术国家重点实验室的曲音波教授、陈冠军教授对本书的编辑与出版也非常关心，程轶喆、李凡、朱涛、孙晓亮、任青措、林瑞婷、张兴敏、吴岳等作为本书的第一批读者给本书提出了许多宝贵的意见，山东大学研究生院为本书提供了出版基金。在此，谨向他们表示由衷的敬意和感谢！同时，由于作者的水平有限，书中肯定存在许多不妥之处，敬请读者批评和指正，宝贵意见请函至 lswang@sdu.edu.cn 或 gaopj@sdu.edu.cn。

编　者

2007 年 9 月

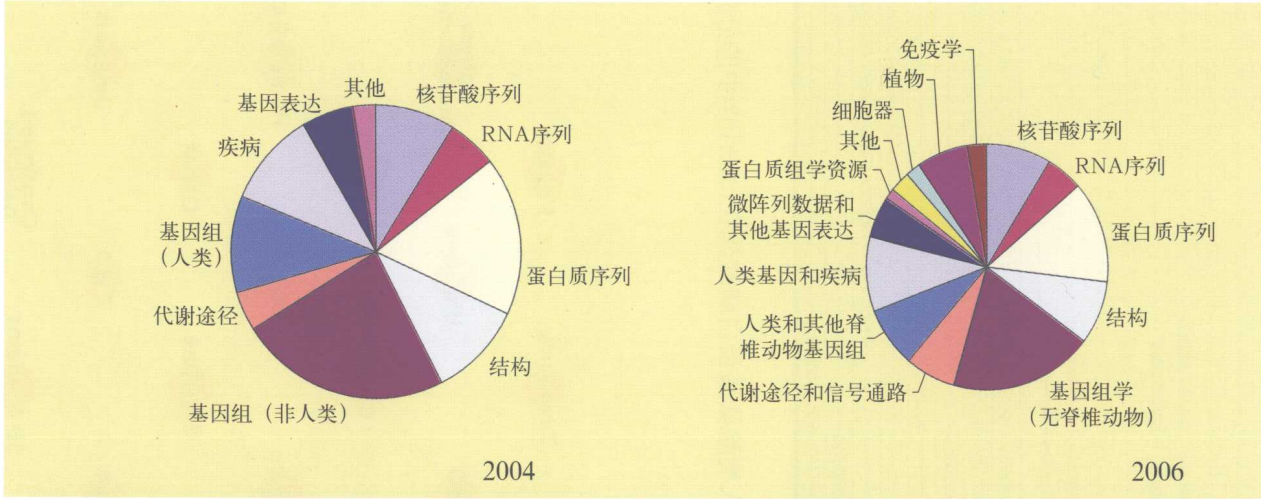


图 3-3　2004年2006年数据库种类数目的比较

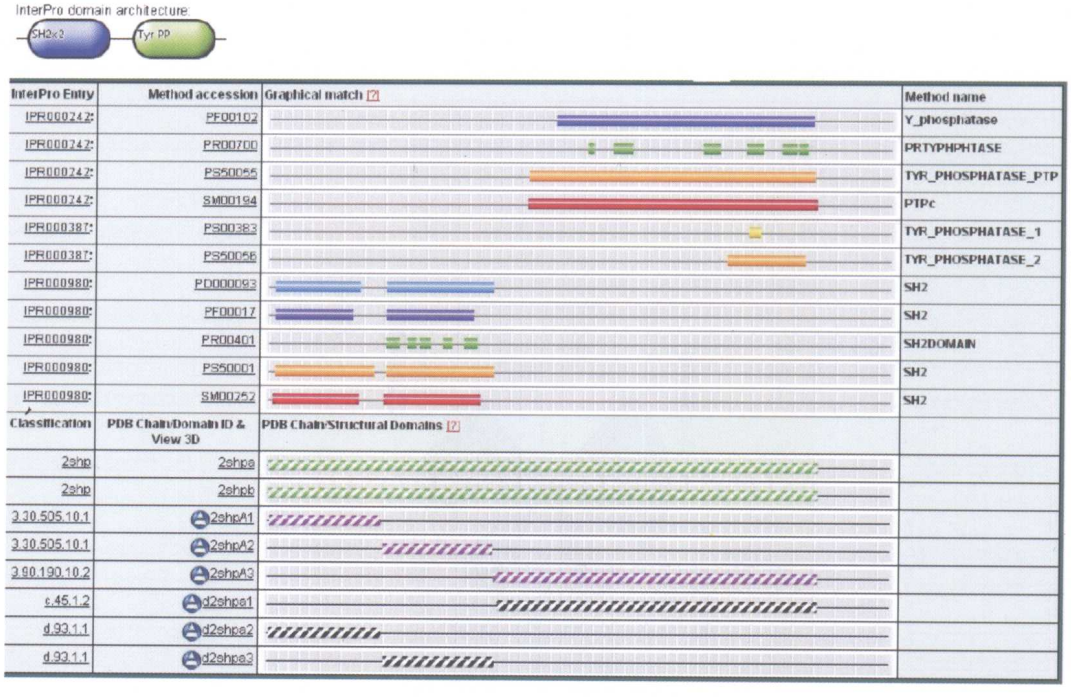


图 3-20　人类酪氨酸蛋白磷酸酶在InterPro中的分析结果

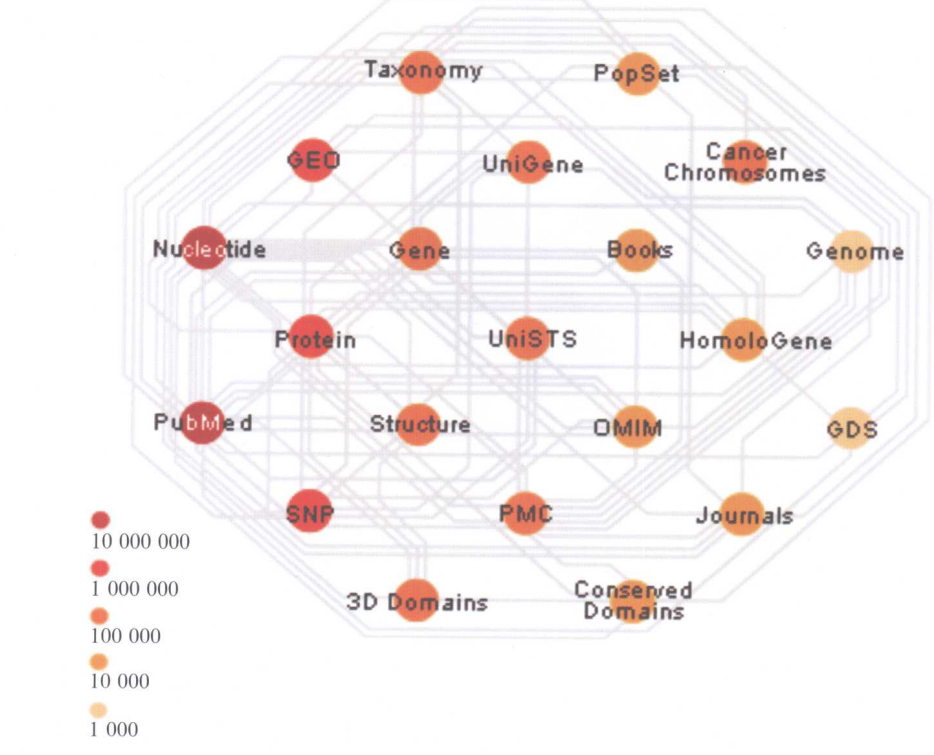


图 4-3　NCBI的数据库资源及其相互链接

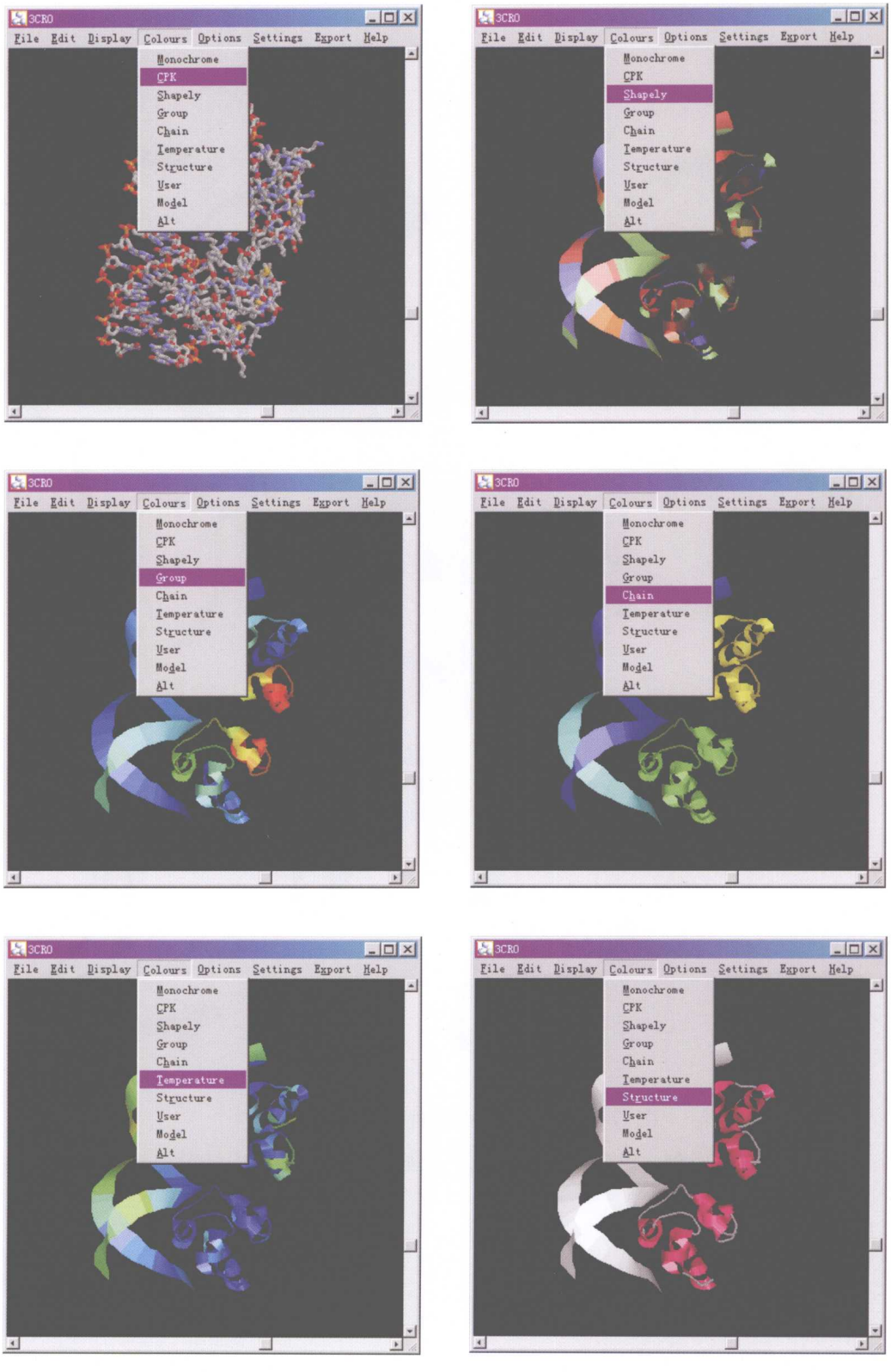


图 7-16　Colours（颜色）菜单

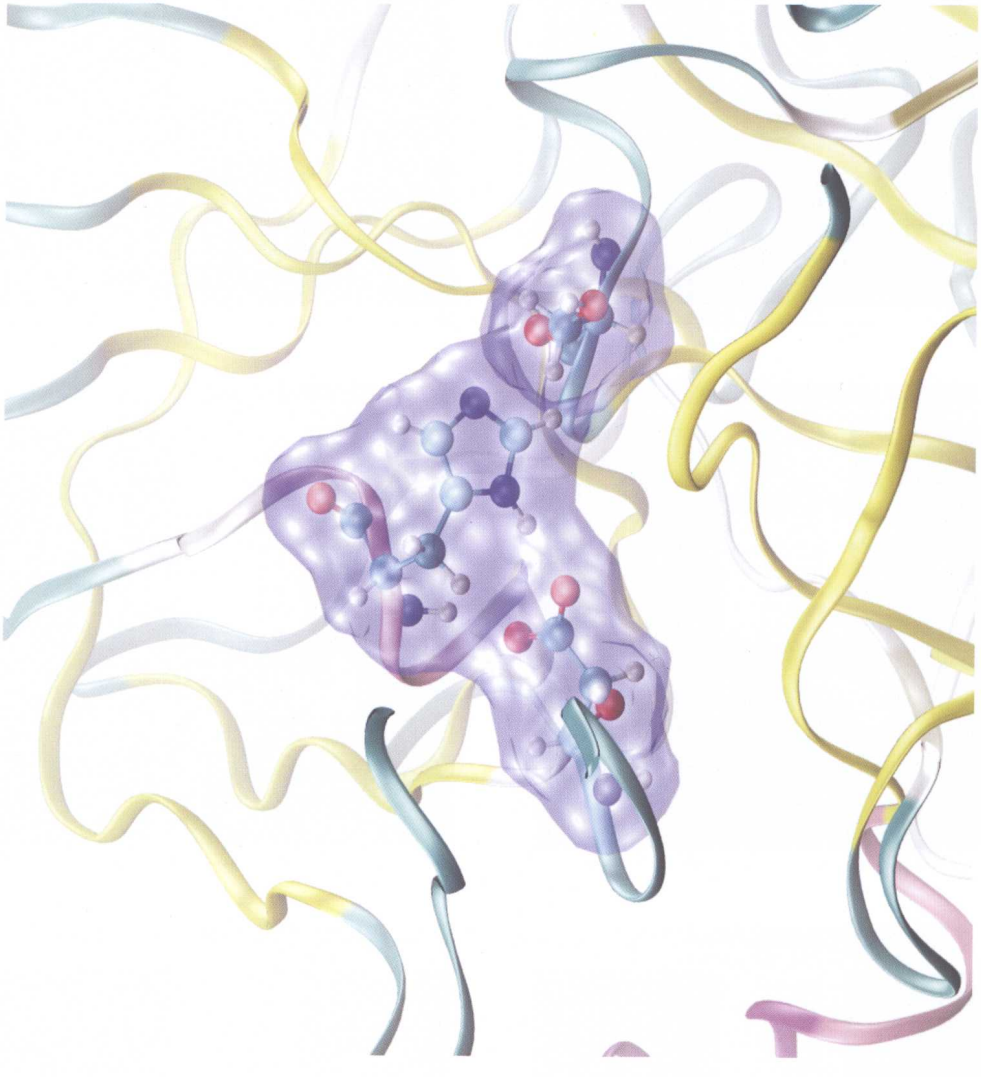


图 7-41　利用POV软件生成的BMP图片
从中可以看到一维序列很远的三个残基，在空间结构中是紧密相联的

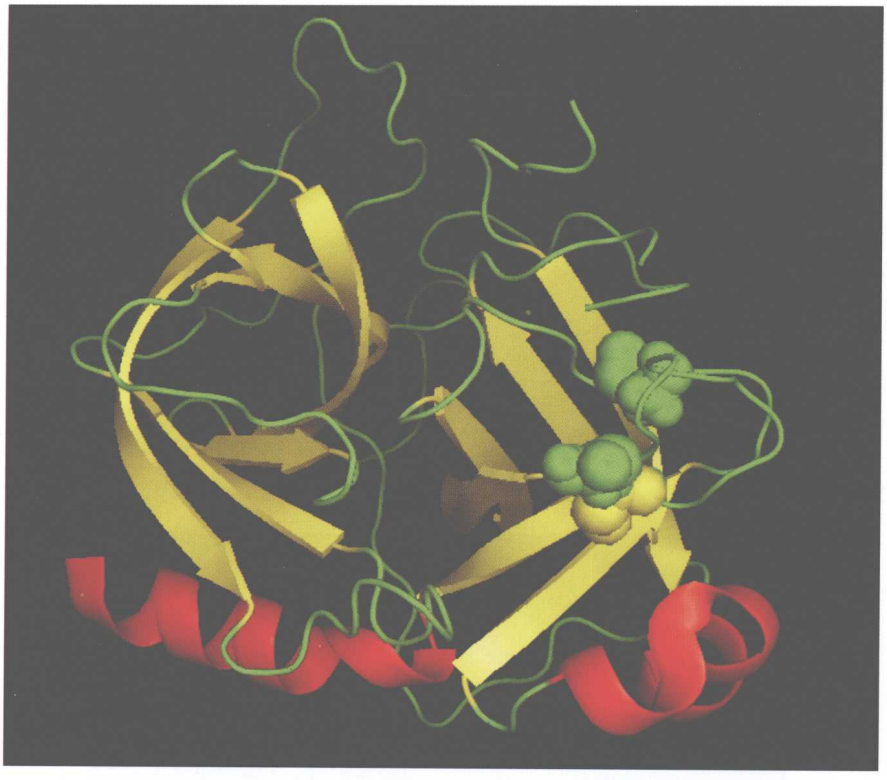


图 7-49　选择球状模型显示

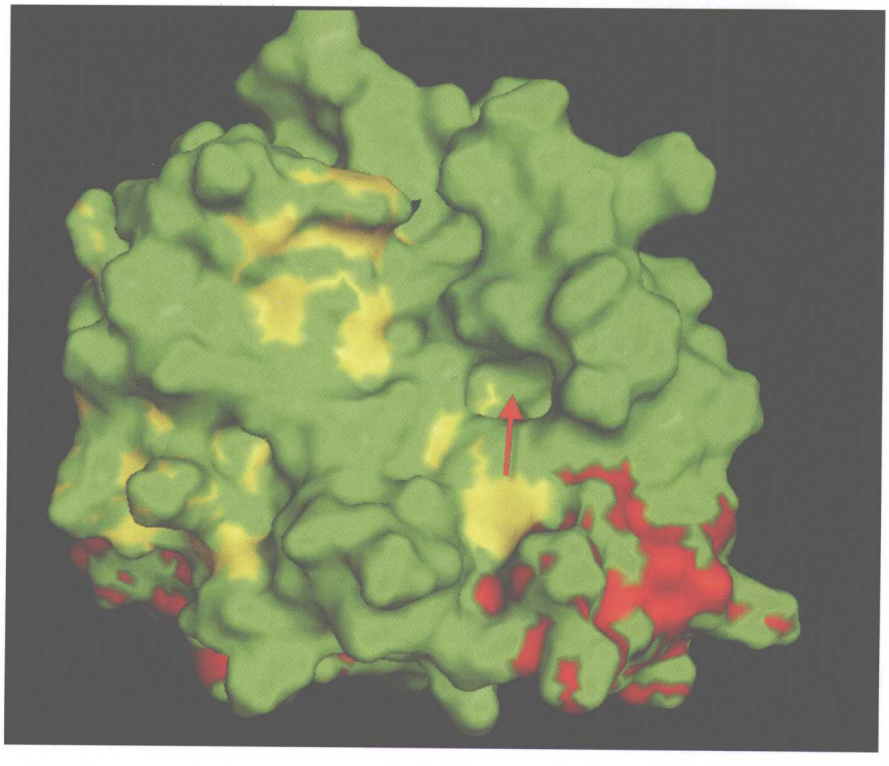
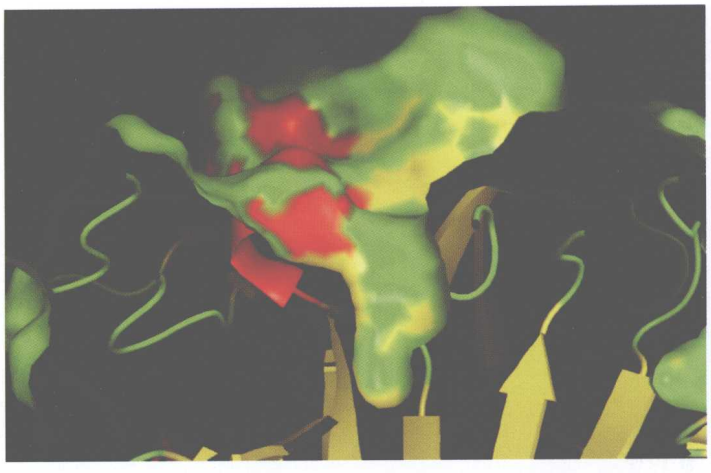
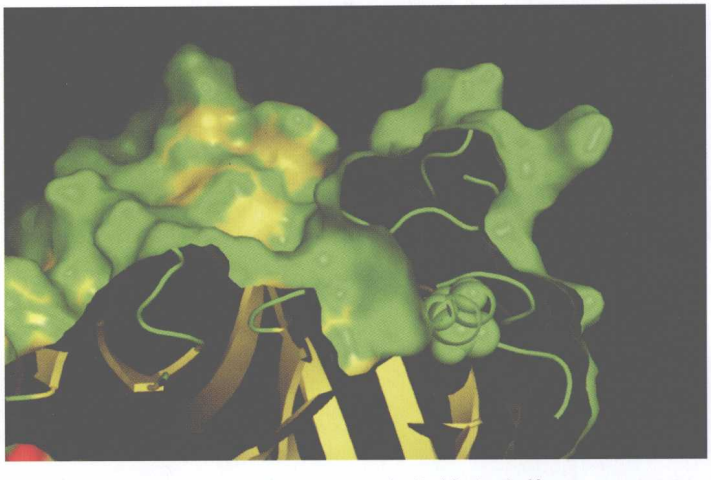


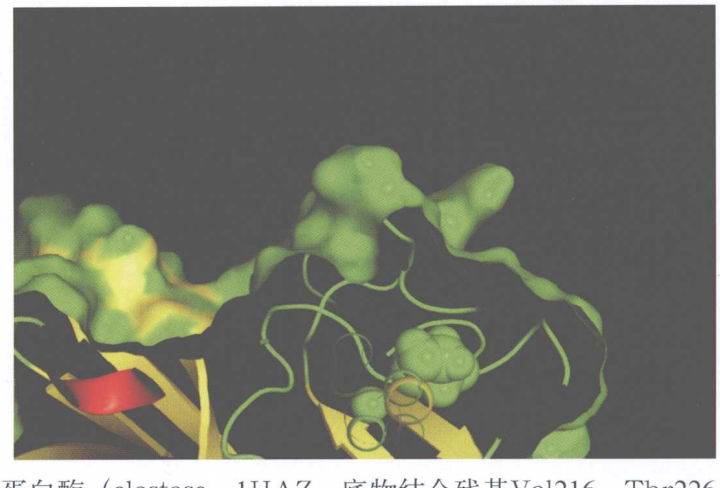
图 7-50　底物结合位点的“口袋”



(a) 胰蛋白酶（trypsin，2BYA，底物结合残基Gly216，Gly226，Asp189）



(b) 胰凝乳蛋白酶（chymotrypsin，2CHA，底物结合残基Gly216，Gly226，Ser189）



(c) 弹性蛋白酶（elastase，1HAZ，底物结合残基Val216，Thr226，Ser189）

图 7-53　通过胰蛋白酶、胰凝乳蛋白酶和弹性蛋白酶底物结合口袋的比较，表明底物的专一性是由酶分子活性中心的结构决定的

目　录

第 1 章　信息学与生物信息学技术 …… 1

1.1　信号、信息、密码与生物信息学技术 …… 1

1.2　生物信息学技术的特点 …… 2

1.2.1　生物信息的储存与传递 …… 2

1.2.2　生物信息传递的密码子 …… 3

1.2.3　信息和密码的层次性 …… 4

1.2.4　信息传递中的“非等价”与多义表达 …… 5

1.2.5　内含子、重复序列与 DNA 多态性 …… 5

1.2.6　基因突变和中性突变理论 …… 6

1.3　生物系统的复杂性和生物信息学技术研究的局限性 …… 6

第 2 章　生物信息数据 …… 8

2.1　核酸的序列与表示方式 …… 9

2.2　蛋白质的序列与表示方式 …… 10

2.3　生物信息数据的存储格式 …… 11

2.3.1　序列最简单注释的 FASTA 格式 …… 11

2.3.2　序列详细注释的 GenBank 格式 …… 13

2.3.3　序列详细注释的 EMBL 格式 …… 18

2.4　生物大分子结构数据的存储格式 …… 19

第 3 章　生物信息数据库 …… 24

3.1　生物信息数据库概述 …… 24

3.1.1　数据库 …… 24

3.1.2　生物信息数据库 …… 24

3.1.3　生物信息数据库分类及发展方向 …… 26

3.2　核酸序列数据库 …… 28

3.2.1　核酸序列基本数据库 …… 28

3.2.2　核酸序列二级数据库 …… 29

3.3　蛋白质序列数据库 …… 30

3.3.1　蛋白质序列基本数据库 …… 30

3.3.2　蛋白质序列二级数据库 …… 34

3.4　生物大分子结构数据库 …… 42

3.4.1 生物大分子结构基本数据库 43
3.4.2 蛋白质结构二级数据库 44

第 4 章 生物信息的检索及策略 48
4.1 生物信息检索概述 48
4.1.1 信息检索的概念 48
4.1.2 检索系统的类型 48
4.1.3 生物信息检索系统 48
4.1.4 信息检索策略 49
4.2 NCBI 数据库检索系统 Entrez 50
4.2.1 美国国家生物技术信息中心 NCBI 50
4.2.2 NCBI 数据库 51
4.2.3 Entrez 简介 52
4.2.4 Entrez 系统检索规则与策略 53
4.2.5 Entrez 检索策略的定制 55
4.3 EBI 的数据库检索系统 SRS 67

第 5 章 序列的分析与相似性搜索 68
5.1 序列分析与相似性搜索概述 68
5.2 序列比对 69
5.2.1 序列比对的原理 69
5.2.2 记分规则 70
5.2.3 序列比对算法 72
5.2.4 多序列比对算法 74
5.3 基于相似性分析的数据库搜索 74
5.3.1 局域比对搜索工具 BLAST 74
5.3.2 BLAST 的检索程序与功能 75
5.4 序列分析软件 76
5.4.1 本地机上进行序列的简单分析 76
5.4.2 利用序列相似性搜索对蛋白质序列的功能注释与分析 86

第 6 章 系统发育分析与分子进化 97
6.1 生物分类与系统发育分析概述 97
6.1.1 生物分类系统 97
6.1.2 系统发育分类学 97
6.1.3 生物进化过程 98
6.1.4 系统发育进化树：进化关系的表示方法 99
6.1.5 系统进化的分析方法 100

6.1.6 分子进化与中性学说 101
6.1.7 分子进化研究的重要意义 104
6.2 分子进化分析的方法与流程 105
6.2.1 系统学的建树方法 105
6.2.2 建树算法比较 107
6.2.3 分子进化分析的流程 107
6.3 分子进化树分析软件 108
6.3.1 多序列比对软件 108
6.3.2 进化分析软件 109
6.3.3 树结构显示软件 109
6.4 系统发育分析/分子进化分析示例 110
6.4.1 基于细胞色素 c 氨基酸序列的真核生物系统发育分析 110
6.4.2 基于 12S rRNA 基因序列的鹳形目鸟类系统发育分析 119
6.5 NCBI 上的系统分类 127
6.5.1 NCBI 的系统分类数据库 127
6.5.2 NCBI 的分类浏览器 128
6.6 进化分析所涉及的 13 种鹳形目鸟类的形态分类简介 128

第 7 章 生物大分子三维结构的可视化分析 133
7.1 分子三维结构可视化概述 133
7.1.1 分子结构模型的建立 133
7.1.2 模型可视化的建立方法 133
7.2 分子结构的显示模型 134
7.2.1 小分子的显示模型 134
7.2.2 生物大分子的结构显示模型 135
7.2.3 生物大分子的分子表面模型 136
7.2.4 光影效果及分辨率 138
7.3 生物大分子结构数据文件的获取 138
7.3.1 生物大分子结构数据的浏览 138
7.3.2 结构数据文件的检索与下载 139
7.4 生物大分子三维结构的可视化分析软件 141
7.4.1 RasMol 软件 142
7.4.2 VMD 软件 153
7.4.3 PyMOL 软件 166

第 8 章 同源模建及分子动力学模拟分析 174
8.1 生物大分子立体结构研究概述 174
8.1.1 生物大分子结构确定方法 174