卫生部“十一五”规划教材

全国高等医药教材建设研究会规划教材

全国高等学校教材

供8年制及7年制临床医学等专业用

生物信息学

Bioinformatics

主编 李霞

副主编 李亦学 廖飞

人民卫生出版社

PEOPLE'S MEDICAL PUBLISHING HOUSE

全国高等医药教材建设研究会规划教材

全国高等学校教材

供**8年制**及**7年制**临床医学等专业用

生物信息学

Bioinformatics

* 1.《细胞生物学》第2版（含光盘）
* 2.《系统解剖学》第2版（含光盘）
* 3.《局部解剖学》第2版（含光盘）
* 4.《组织学与胚胎学》第2版（含光盘）
* 5.《生物化学与分子生物学》第2版（含光盘）
* 6.《生理学》第2版（含光盘）
* 7.《医学微生物学》第2版（含光盘）
* 8.《人体寄生虫学》第2版（含光盘）
* 9.《医学遗传学》第2版（含光盘）
*10.《医学免疫学》第2版
*11.《病理学》第2版（含光盘）
*12《病理生理学》第2版（含光盘）
*13.《药理学》第2版（含光盘）
*14.《临床诊断学》第2版（含光盘）
*15.《实验诊断学》第2版（含光盘）
*16.《医学影像学》第2版（含光盘）
*17.《内科学》第2版（含光盘）
*18.《外科学》第2版（含光盘）
*19.《妇产科学》第2版（含光盘）
*20.《儿科学》第2版（含光盘）
*21.《感染病学》第2版（含光盘）
*22.《神经病学》第2版（含光盘）
*23.《精神病学》第2版（含光盘）
*24.《眼科学》第2版（含光盘）
*25.《耳鼻咽喉头颈外科学》第2版
*26.《核医学》第2版（含光盘）
*27.《预防医学》第2版（含光盘）
*28.《医学心理学》第2版（含光盘）
29.《医学统计学》第2版（含光盘）
*30.《循证医学》第2版（含光盘）
*31.《医学文献信息检索》第2版（含光盘）
32.《临床流行病学》（含光盘）
33.《肿瘤学》
● 34.《**生物信息学**》（含光盘）
35.《实验动物学》（含光盘）
36.《医学科学研究导论》
37.《医学伦理学》（含光盘）

注：全套书均为卫生部“十一五”规划教材，
画 * 者为普通高等教育“十一五”国家级规划教材

策划编辑　戴薇薇
责任编辑　戴薇薇　欧阳丹
封面设计　郭　森
版式设计　盖　伟

人民卫生出版社网站：

门户网：www.pmph.com　出版物查询、网上书店
卫人网：www.ipmph.com　护士、医师、药师、中医师、卫生资格考试培训

ISBN 978-7-117-12938-1
9 787117 129381 >
定价(含光盘)：90.00 元

- 卫生部“十一五”规划教材
- 全国高等医药教材建设研究会规划教材
- 全国高等学校教材
- 供8年制及7年制临床医学等专业用

生物信息学

Bioinformatics

主　　编　李　霞

副 主 编　李亦学　廖　飞

编　　委　（以姓氏笔画排序）

田　心（天津医科大学）　朱　浩（南方医科大学）
刘建国（河北大学）　许丽艳（汕头大学）
李　霞（哈尔滨医科大学）　李亦学（同济大学）
吴忠道（中山大学）　张　岩（哈尔滨医科大学）
茆灿泉（西南交通大学）　赵雨杰（中国医科大学）
胡福泉（第三军医大学）　童隆正（首都医科大学）
廖　飞（重庆医科大学）　魏冬青（上海交通大学）

学术秘书　汪强虎（哈尔滨医科大学）

人民卫生出版社

图书在版编目（CIP）数据

生物信息学 / 李霞主编. —修订本. —北京：人民卫生出版社，2010.8
ISBN 978-7-117-12938-1

Ⅰ. ①生… Ⅱ. ①李… Ⅲ. ①生物信息论－高等学校－教材 Ⅳ. ①Q811.4

中国版本图书馆 CIP 数据核字（2010）第 105644 号

生物信息学

主　　编：李　霞
出版发行：人民卫生出版社（中继线 010-59780011）
地　　址：北京市朝阳区潘家园南里 19 号
邮　　编：100021
E - mail：pmph @ pmph.com
购书热线：010-67605754　010-65264830
　　010-59787586　010-59787592
印　　刷：北京金盾印刷厂
经　　销：新华书店
开　　本：850×1168　1/16　　印张：31.5
字　　数：932 千字
版　　次：2010 年 8 月第 1 版　2010 年 8 月第 1 版第 1 次印刷
标准书号：ISBN 978-7-117-12938-1/R • 12939
定价（含光盘）：90.00 元
打击盗版举报电话：010-59787491　E-mail：WQ @ pmph.com
（凡属印装质量问题请与本社销售中心联系退换）

第二版出版说明

全国高等学校八年制临床医学专业规划教材自2005年出版以来，得到了教育部、卫生部等主管部门的认可，以及医学院校广大师生的好评。为了进一步满足教学改革与实践不断推进，以及医学科学不断发展的需要，全国高等医药教材建设研究会和卫生部教材办公室在吴阶平、裘法祖、吴孟超、陈灏珠和刘德培院士等的亲切关怀和支持下于2009年启动了该套教材第二轮的修订工作。

第二轮修订过程中仍坚持“精品战略，质量第一的原则，从精英教育的特点、医学模式的转变、信息社会的发展、国内外教材的对比等角度出发，在注重‘三基’、‘五性’的基础上，从内容到形式都‘更新’、‘更深’、‘更精’，为培养高素质、高水平、富有临床实践和科学创新能力的医学博士服务”的编写宗旨，并根据使用过程中的反馈意见与建议，在第一轮的基础上力求做到：学科体系更加完善，增加了《临床流行病学》、《肿瘤学》、《生物信息学》、《实验动物学》、《医学科学研究导论》和《医学伦理学》；相关学科的交叉与协调更为完善，比如《生物化学》与《医学分子生物学》合并为《生物化学与分子生物学》；内容的选材与框架体系的设计更加注重启发性，强调学生创新能力的培养，并适当给学生留下了思维分析、判断、探索的空间；教材的配套更加健全；装帧设计更为精美。

该套书在修订过程中，得到了广大医学院校的大力支持，作者均来自各学科临床、科研、教学第一线，具有丰富临床、教学、科研和写作经验的优秀专家，作者队伍覆盖了目前国内所有开办临床医学专业八年制及七年制的院校。

修订后的第二版仍以全国高等学校临床医学专业八年制及七年制师生为主要目标读者，并可作为研究生、住院医师等相关人员的参考用书。

全套教材共37种，其中36种于2010年8月出版，1种将于2010年年底出版。

八年制教材目录

*1.《细胞生物学》 第2版（含光盘）	主　编 副主编	杨　恬 左　伋　刘艳平
*2.《系统解剖学》 第2版（含光盘）	主　编 副主编	柏树令 应大君　丁文龙　崔益群
*3.《局部解剖学》 第2版（含光盘）	主　编 副主编	王怀经　张绍祥 张雅芳　胡海涛
*4.《组织学与胚胎学》 第2版（含光盘）	主　编 副主编	高英茂　李　和 李继承　陈晓蓉
*5.《生物化学与分子生物学》 第2版（含光盘）	主　编 副主编	贾弘禔　冯作化 屈　伸　药立波　方定志　冯　涛
*6.《生理学》 第2版（含光盘）	主　编 副主编	姚　泰 曹济民　樊小力　王庭槐
*7.《医学微生物学》 第2版（含光盘）	主　编 副主编	贾文祥 陈锦英　江丽芳　黄　敏
*8.《人体寄生虫学》 第2版（含光盘）	主　编 副主编	詹希美 诸欣平　刘佩梅
*9.《医学遗传学》 第2版（含光盘）	主　编 副主编	陈　竺 陆振虞　傅松滨
*10.《医学免疫学》 第2版	主　编 副主编	何　维 曹雪涛　熊思东
*11.《病理学》 第2版（含光盘）	主　编 副主编	陈　杰　李甘地 文继舫　来茂德　孙保存
*12.《病理生理学》 第2版（含光盘）	主　编 副主编	李桂源 吴伟康　欧阳静萍
*13.《药理学》 第2版（含光盘）	主　编 副主编	杨世杰 杨宝峰　颜光美　臧伟进
*14.《临床诊断学》 第2版（含光盘）	主　编 副主编	欧阳钦 吴汉妮　刘成玉
*15.《实验诊断学》 第2版（含光盘）	主　编 副主编	王鸿利 尚　红　王兰兰
*16.《医学影像学》 第2版（含光盘）	主　编 副主编	金征宇 冯敢生　冯晓源
*17.《内科学》 第2版（含光盘）	主　编 副主编	王吉耀 廖二元　黄从新　华　琦
*18.《外科学》 第2版（含光盘）	主　编 副主编	陈孝平 石应康　邱贵兴　杨连粤

*19.《妇产科学》 第2版（含光盘）	主　编 副主编	丰有吉　沈　铿 马　丁　孔北华　李　力
*20.《儿科学》 第2版（含光盘）	主　编 副主编	薛辛东 杜立中　毛　萌
*21.《感染病学》 第2版（含光盘）	主　编 副主编	王宇明 施光峰　宁　琴　李　刚
*22.《神经病学》 第2版（含光盘）	主　编 副主编	吴　江 贾建平　崔丽英
*23.《精神病学》 第2版（含光盘）	主　编 副主编	江开达 于　欣　李凌江　王高华
*24.《眼科学》 第2版（含光盘）	主　编 副主编	葛　坚 赵家良　黎晓新
*25.《耳鼻咽喉头颈外科学》 第2版	主　编 副主编	孔维佳 周　梁　许　庚　王斌全　唐安洲
*26.《核医学》 第2版（含光盘）	主　编 副主编	张永学　黄　钢 匡安仁　李亚明
*27.《预防医学》 第2版（含光盘）	主　编 副主编	孙贵范 凌文华　孙志伟　姚　华
*28.《医学心理学》 第2版（含光盘）	主　编 副主编	姜乾金 马　辛　林大熙　张　宁
29.《医学统计学》 第2版（含光盘）	主　编 副主编	颜　虹 徐勇勇　赵耐青
*30.《循证医学》 第2版（含光盘）	主　编 副主编	王家良 詹思延　许能锋　康德英
*31.《医学文献信息检索》 第2版（含光盘）	主　编 副主编	罗爱静 马　路　于双成
32.《临床流行病学》 （含光盘）	主　编 副主编	李立明 詹思延　谭红专
33.《肿瘤学》	主　编 副主编	郝希山　魏于全 赫　捷　周云峰
34.《生物信息学》 （含光盘）	主　编 副主编	李　霞 李亦学　廖　飞
35.《实验动物学》 （含光盘）	主　编 副主编	秦　川 张连峰　魏　泓　顾为望　王　钜
36.《医学科学研究导论》	主　编 副主编	詹启敏　赵仲堂 刘　佳　刘　强
37.《医学伦理学》 （含光盘）	主　编 副主编	王明旭 尹　梅　严金海

注：全套书均为卫生部“十一五”规划教材，画*者为普通高等教育“十一五”国家级规划教材

八年制教材再版序言

五年来，在大家的热情呵护下，我们共同见证了八年制临床医学教材——这个新生命的诞生与茁壮成长。如今，第二版教材与大家见面，怀纳第一版之精华而不张扬，吞吐众学者之智慧而不狂放，正如医学精英人才所应具备的气质与神韵。在继承中发展，新生才能越发耀眼；切时代之脉搏，思维才能永领潮头。第二版教材已然跨入新的成长阶段，心中唯觉欣喜和慰藉。

回想第一版教材面世之后，得到了各方众多好评，这充分说明了：这套教材将生命科学信息化、网络化以及学科高度交叉、渗透的特点融于一身，同时切合了环境-社会-心理-工程-生物医学模式的转变，诠释了以人为本、协调发展的战略思想。另外，编委构成的权威性和代表性、内容选择、编排体系、印刷装帧质量等，令广大师生耳目一新，爱不释卷。诚然，第一版教材也并非十全十美，比如有的学科仍以介绍知识为主，启发性不强，对学生难以起到点石成金、抛砖引玉的作用，不利于学生创新思维能力的培养；有的学科、章节之间有重复现象，略显冗余，不够干练。另外，随着学科的进展，部分疾病的临床分类、治疗等内容已略显滞后，亟待最新的研究成果加入其中，充实完善。

鉴此，第一版教材的修订工作便提上日程。此次修订，比当初第一版的编纂过程更为艰辛和严谨，从编者的谨慎遴选到教材内容的反复推敲、字斟句酌，可谓精益求精、力臻完美，经过数轮探讨、分析、总结、归纳、整理，第二版教材终于更富于内涵、更具有生命力地与广大师生们见面了。

“精英出精品，精品育精英”是第二版教材在修订之初就一直恪守的理念。主编、副主编与编委们均是各领域内的医学知名专家学者，不仅著作立身，更是德高为范。在教材的编写过程中，他们将从医执教中积累的宝贵经验、体会以及医学精英的特质潜移默化地融入到教材当中。同时，在主编负责制的前提下，主编、副主编负责全书的系统规划，编委会构成团结战斗的团队，各位专家群策群力、扬长补短、集思广益、查漏补缺，为教材的高标准、高质量的修订出版打下了坚实的基础。

注重医学学科内涵的延伸与发展，同时兼顾学科的交叉与融合是第二版教材的一大亮点。此次修订不仅在第一版的基础上增加了《临床流行病学》、《肿瘤学》、《生物信息学》、《实验动物学》、《医学科学研究导论》和《医学伦理学》，同时还合并了《生物化学》与《医学分子生物学》。通过主编顶层设计，相邻学科主编、副主编协调与磋商，互审编写提纲，以及交叉互审稿件等措施，相当程度上实现了突出中心、合理交叉、避免简单重复的要求。

强调启发性以及创新意识、创新思维和创新能力的培养是第二版教材的另一大特色。除了坚持“三基（基础理论、基本知识和基本技能）和五性（思想性、科学性、先进性、启发性和适用性）”，更注重激发学生的思维，让他们成为自己头脑的主人，批判地看待事物，辩证地对待知识，创造性地预见未来。同时，这版教材也特别注重与五年制教材、研究生教材、专科医师培训教材以及参考书的区别与联系。

以吴阶平、裘法祖、吴孟超、陈灏珠为代表的德高望重的老前辈对第二版教材寄予了殷切期望和悉心指导，教育部、卫生部、国家中医药管理局、国家食品药品监督管理局的各位领导的支持是这版教材不断完善的动力之源。在这里，衷心感谢所有关心这套教材的人们！正是你们的关注，广大师生手中才会捧上这样一本融贯中西、汇纳百家的精品。

八年制医学教材的第一版是我国医学教育史上的重要创举，相信修订后的第二版将不负我国医学教育改革的使命和重任，为培养高层次的具有综合素质和发展潜能的医药卫生人才做出更大的贡献。诚然，修订过程虽然力求完美，但纰漏与瑕疵在所难免，冀望各位领导、同道及师生不吝赐教，以便于这套教材能够与时俱进，不断完善。

是为序。

中国工程院院士
中国医学科学院院长
北京协和医学院院长

刘德培

于庚寅端午佳节

二〇一〇年六月十六日

序

生物信息学是20世纪80年代末伴随着基因组研究而产生的一门新的前沿学科。它在获取、加工、储存、分发海量基因组信息的同时，把基因组DNA序列信息分析作为源头，寻找基因组序列中代表蛋白质和RNA基因等功能元件的编码序列，并阐明非编码序列的信息实质，破译隐藏在DNA序列中的遗传语义规律。

基因组研究的初期，生物信息学在大规模基因组的组装与基因标注上发挥了关键的作用。随着基因组研究的深入，不仅DNA序列数据，而且表达序列标签(EST)数据、单核苷酸多态性(SNP)数据、单体型图(HapMap)数据等也大量涌现。进入结构基因组与功能基因组时代后，基因组、蛋白质组、代谢组等组学数据迅速出现。当前，蛋白－蛋白相互作用网络、基因表达调控网络、信号传导网络以及代谢网络的出现与发展，使得生物信息学进入了系统生物学的时代。如果说生物信息学发展的初期面对的主要是序列数据，那么，随着基因组领域研究在质与量上的提高，这一学科面对的数据在类型与本质上都极大地丰富了。为了处理这些数据，生物信息学得到了蓬勃的发展，成为了基因组研究不可或缺的工具。生物信息学自身也成为了一门实用学科。

可以相信，在第二、第三代高通量测序技术的推动下，基因组及相关的成果将很快进入国民经济及人类健康的很多领域，也会直接走到每个人的面前。那时候，生物信息学的很多知识会成为人们生产、生活的基本知识，生物信息学自身也将成为一门基础学科。

在生物信息学不断普及的背景下，李霞和李亦学两位教授组织编写了这本教科书，为生物信息学的教育提供了很好的教材。

本书的特点是：全面地提供了数据库资源，序列对比，表达谱分析，分子进化，芯片数据处理，生物网络，蛋白质结构以及药物设计等几乎生物信息学涉及的所有方面。为了使读者能更好地检验自己掌握程度，每章末还附了习题。由于两位教授长期在生物信息学第一线从事研究与教学，书中也融汇了他们大量的体会与经验。相信这本书能为使用者带来切实的帮助。

陈润生

2010年4月16日

前言

《生物信息学》是在国内资深生物学家与医学专家的倡导下，经全国高等医药教材建设研究会、卫生部教材办公室组织有关专家反复论证后决定组织编写的。在教材的具体筹备和编写过程中，全国高等医药教材建设研究会、卫生部教材办公室先后组织三次会议进行研讨。本书正是根据这些会议精神及专家的指导意见组织全国十余所大学的一线教师和学者共同努力编写而成的。

“21世纪是生命科学的世纪，也是信息科学的世纪”。伴随着人类基因组计划及其他模式生物基因组计划的全面实施，分子生物数据正在呈爆炸性增长。及时、充分、有效地利用不断增长的生物信息资源进行分析和探索，已经成为生物医学领域研究与应用的必备方法。为适应现代生物医学发展和素质教育的需要，我们在编写《生物信息学》教材过程中，突出“三基”(基本知识、基本理论、基本技能)，强调五性(思想性、科学性、先进性、启发性、适用性)，并力求在内容和形式上有所创新。

全书共分为三篇十六章。第一篇生物信息学基础，含DNA、RNA和蛋白质序列信息资源、双序列比对、多序列比对、序列特征分析、分子进化分析、表达序列分析和基因芯片数据分析七章，均系生物医学相关领域发展过程中形成的基础生物信息数据及分析方法；第二篇功能基因组信息学，含基因功能注释、蛋白质分析与蛋白质组学、蛋白质结构分析、转录调控的信息学分析、生物分子网络和计算表观遗传学六章，均系功能基因组研究中颇具特色的生物信息学方法；第三篇生物信息学与人类复杂疾病，含人类复杂疾病与计算系统生物学、单核苷酸多态与人类疾病和miRNA与复杂疾病三章，以及光盘附录内容药物生物信息学，均系近几年发展起来的与复杂疾病有关的重要生物信息学方法。

本书各章节相对独立，每一章都反映了生物信息学某一方向的最新成果与发展趋势。为适应不同读者群的需要，各章的布局是统一的。第一节是引言，以简明易懂的语言介绍该章的主要内容，包括能解决什么问题和解决问题的思路；后面各节介绍基本概念和常用生物信息学方法，着重于生物医学实际应用、操作方法和生物医学意义的解释；小结与主要参考文献放在各章最后。

本书的读者面十分广泛。不论是生物医学领域的学生、教师、研究人员还是生物信息学专业人员，都可以各取所需、各有所获。生物信息学思想和技术是医学研究的有效工具，医学研究者阅读本书，不难知道有哪些现代生物信息学方法可以为之所用、基本思路如何、需要怎样的设计和数据，应用的结果如何解释等；生物信息学专业人员阅读本书，不但可以深入地掌握生物信息学的最新研究成果与未来发展方向，而且还有助于提升研究工作的水平。

本书各章作者都是相关研究方向的专家，每一章都凝聚了他们独特的学术思想、研究心得和研究成果。他们在百忙之中精心组织素材，字斟句酌地编写，付出了大量心血。在此我们对全体编委的无私奉献深表谢意！同时，哈尔滨医科大学生物信息科学与技术学院的老师和研究生们也做了大量的协助工作，在此一并致谢！

本教材得到国家高科技“863”项目和哈尔滨医科大学“211”工程重点学科建设经费的资助，特此鸣谢！

在本教材编写过程中，尽管我们努力跟踪学科的新发展、新技术，并尽力把它们纳入到教材中来，以保持本书的先进性和实用性，但由于时间紧迫、能力有限，直至完稿，仍觉有许多不足之处，希望学术同仁不吝赐教，以便再版时改正。

李　霞　李亦学

2010年4月20日

目 录

第二章 双序列比对 38

CHAPTER 2 PAIRWISE SEQUENCE ALIGNMENT

第三章 多序列比对 58

CHAPTER 3 MULTIPLE SEQUENCE ALIGNMENT

第二篇 功能基因组信息学

第三篇 生物信息学与人类复杂疾病

绪　论
INTRODUCTION TO BIOINFORMATICS

第一节　生物信息学的兴起
Section 1　The Rise of Bioinformatics

20世纪后期，生物科学技术迅猛发展，无论是数量上还是质量上都极大地丰富了生物科学的数据资源。数据资源的急剧膨胀，迫使人们寻求一种强有力的工具去组织这些数据，以利于储存、加工和进一步使用。一方面，海量的生物学数据中必然蕴涵着重要的生物学规律，这些规律将是解释生命之谜的关键，人们需要一种强有力的工具来协助人脑完成对这些数据的分析工作；另一方面，以数据分析、处理为本质的计算机科学技术和网络技术迅猛发展并日益渗透到生物科学的各个领域。于是，一门崭新的、拥有巨大发展潜力的新学科——生物信息学悄然兴起。

早在1956年，在美国田纳西州盖特林堡召开的首次“生物学中的信息理论研讨会”上，便产生了生物信息学的概念。但是，就生物信息学的发展而言，它还是一门相当年轻的学科。直到20世纪80~90年代，伴随着计算机科学技术的进步，生物信息学才获得突破性进展。

1987年，林华安博士正式把这一学科命名为“生物信息学(bioinformatics)”。此后，其内涵随着研究的深入和现实需要的变化而几经更迭。1995年，在美国人类基因组计划第一个五年总结报告中，给出了一个较为完整的生物信息学定义：生物信息学是一门交叉科学，它包含了生物信息的获取、加工、存储、分配、分析、解释等在内的所有方面，它综合运用数学、计算机科学和生物学的各种工具来阐明和理解大量数据所包含的生物学意义。

生物信息学是融合生命科学与数理科学的新兴学科，具体地说生物信息学是以核酸、蛋白质等生物大分子数据库为主要研究对象，以数学、信息学、计算机科学为主要研究手段，以计算机硬件、软件和计算机网络为主要研究工具，对浩如烟海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。并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、核酸和蛋白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、器官和个体的发生、发育、病变、衰亡等生命科学中重大问题，搞清它们的基本规律和时空联系，建立“生物学周期表”。

生物信息学不仅是一门新学科，更是一种重要的研究开发工具。从科学的角度来讲，生物信息学是一门研究生物和生物相关系统中信息内容与信息流向的综合系统科学。只有通过生物信息学的计算处理，人们才能从众多分散的生物学观测数据中获得对生命运行机制的系统理解。从工具的角度来讲，生物信息学几乎是今后有关生物(医药)研究开发所必需的工具。只有根据生物信息学对大量数据资料进行分析后，人们才能选择该领域正确的研发方向。

生物信息学不仅具有重大的科学意义，而且具有巨大的经济效益。它的许多研究成果可以较快地产业化，成为价值很高的产品。

纵观当今生物信息学界的现状可以发现，大部分研究人员都把注意力集中在基因组、蛋白质组、转录组、miRNA组等以及与此密切相关的药物设计上。

一、人类基因组计划

人类基因组计划(human genome project, HGP)是一个国际合作项目，是由美国科学家于1985年率先提出，于1990年正式启动。美国、英国、法国、德国、日本和中国科学家共同参与了这一价值达30亿美元的人类基因组计划。按照这个计划的设想，要把人体内基因的密码全部解开，同时绘制出人类基因的图谱。换句话说，就是要揭开组成人体约3万个基因的30亿个碱基对的秘密。人类基因组计划与曼哈顿原子弹计划和阿波罗计划并称为三大科学计划。

为什么选择人类的基因组进行研究呢？因为人类是在“进化”历程上最高级的生物，对它的研究有助于人类认识自身、掌握生老病死规律等。在人类基因组计划中，还包括对五种生物基因组的研究：大肠杆菌、酵母、线虫、果蝇和小鼠，称之为人类的五种“模式生物”。

HGP的目的是解码生命、了解生命起源、了解生命体生长发育规律；认识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。此外还有测序技术研究，人类基因组序列变异研究，功能基因组技术研究，比较基因组学研究，社会、法律、伦理研究，生物信息学和计算生物学研究，教育培训研究等目的。

HGP的主要任务是人类的DNA测序。测出人类基因组DNA的30亿个碱基对的序列，发现所有人类基因，找出它们在染色体上的位置，破译人类全部遗传信息。利用HGP发展起来的这些技术和资源进行生物学研究的同时促进了人类健康的发展。HGP主要包括下面四张图谱。

1. 遗传图谱(genetic map) 又称连锁图谱(linkage map)，它是以具有遗传多态性(在一个遗传位点上具有一个以上的等位基因，在群体中的出现频率皆高于1%)的遗传标记为“路标”，以遗传学距离(在减数分裂事件中两个位点之间进行交换、重组的百分率，1%的重组率称为1cM)为图距的基因组图。遗传图谱的建立为基因识别和完成基因定位创造了条件。意义：6000多个遗传标记已经能够把人的基因组分成6000多个区域，使得连锁分析法可以找到某一致病的或表现型的基因与某一标记邻近(紧密连锁)的证据，这样可把这一基因定位于这一已知区域，再对基因进行分离和研究。对于疾病而言，找基因和分析基因是个关键。

2. 物理图谱(physical map) 物理图谱是指有关构成基因组的全部基因的排列和间距的信息，它是通过对构成基因组的DNA分子进行测定而绘制的。绘制物理图谱的目的是把有关基因的遗传信息及其在每条染色体上的相对位置线性而系统地排列出来。DNA物理图谱是指DNA链的限制性酶切片段的排列顺序，即酶切片段在DNA链上的定位。因限制性内切酶在DNA链上的切口是以特异序列为基础的，核苷酸序列不同的DNA，经酶切后就会产生不同长度的DNA片段，由此而构成独特的酶切图谱。因此，DNA物理图谱是DNA分子结构的特征之一。DNA是很大的分子，由限制酶产生的用于测序反应的DNA片段只是其中的极小部分，这些片段在DNA链中所处的位置关系是应该首先解决的问题，故DNA物理图谱是顺序测定的基础，也可理解为是指导DNA测序的蓝图。

3. 序列图谱(sequence map) 随着遗传图谱和物理图谱的完成，测序就成为重中之重的工作。DNA序列分析技术是一个包括制备DNA片段化及碱基分析、DNA信息翻译的多阶段过程。通过测序可得到基因组的序列图谱。

4. 基因图谱(gene map) 基因图谱是在识别基因组所包含的蛋白质编码序列的基础上绘制的结合有关基因序列、位置及表达模式等信息的图谱。在人类基因组中，鉴别出的2%~5%长度的全部基因的位置、结构与功能，最主要的方法是通过基因的表达产物mRNA反追到染色体的位置。基因图谱的意义在于它能有效地反映在正常或受控条件中表达的全基因时空图。通过这张图可以了解某一基因在不同时间、不同组织、不同水平的表达，也可以了解一种组织中不同时间、不同基因中不同水平的表达，还可以了解某一特定时间、不同组织中不同基因不同水平的表达。

HGP对人类疾病基因研究具有重要意义，人类疾病相关的基因是人类基因组中结构和功能完整性至关重要的信息。对于单基因病，采用“定位克隆”和“定位候选克隆”的全新思路，促使了亨廷顿

舞蹈病、遗传性结肠癌和乳腺癌等一大批单基因遗传病致病基因的发现，为这些疾病的基因诊断和基因治疗奠定了基础。心血管疾病、肿瘤、糖尿病、神经精神类疾病(老年性痴呆、精神分裂症)、自身免疫性疾病等多基因疾病是目前疾病基因研究的重点。健康相关研究是 HGP 的重要组成部分，并于 1997 年相继提出："肿瘤基因组解剖计划"、"环境基因组学计划"、"国际人类基因组单体型图计划(The International HapMap Project)"。

二、生物信息学与组学

作为新兴的交叉学科，生物信息学的研究重点之一为组学(Omics)研究，即同时研究成千上万个基因或蛋白质集合的生物特性。类比于分子生物学的中心法则(DNA → RNA →蛋白质)，生物信息学在组学研究中的中心法则可以概括为"基因组→转录组→蛋白质组"，即可以从三个层面阐述生物信息学与组学间的关系。

(一) 基因组层面

基因组学(genomics)、结构基因组学(structural genomics)、功能基因组学(functional genomics)这三种密切相关的组学是生物信息学在基因组层面研究的重点内容。基因组学的目标是测定和分析某个(些)物种的全部 DNA 序列(即基因组)；而结构基因组学则可为其提供大量 DNA 及蛋白质数据，是基因组学的有力支撑及基础；功能基因组学的主要任务则是充分、合理利用基因组学及结构基因组学提供的信息，系统的研究基因及其产物的功能。三者间是密切相关，彼此依存的科学体系。

1. 基因组学　出现于 20 世纪 80 年代，具体来说，是研究生物体基因组的组成情况，以及各基因的结构，彼此间关系及表达调控的科学。与过去基因研究相比，其重要特点是具有鲜明的"整体性"，即从基因组的层次阐述基因特点，包括诸如在染色体组上的位置、结构、基因产物的功能及基因与基因间的关系等。其主要工具和方法包括生物信息学、遗传分析、基因表达测量和基因功能鉴定等。

2. 结构基因组　是一门用结构生物学方法在生物体整体水平上(如全生物体、全细胞或整个基因组)对全部蛋白质(主要包括受体蛋白、酶、通道以及与基因调控密接相关的核酸结合蛋白等)，相关蛋白质复合物(如酶和底物、酶与抑制剂、作用原与受体、DNA 与其结合蛋白等)，RNA 及其他生物大分子进行分析，精细测定其三维结构的学科。主要通过基因作图方式构建四种类型的图谱：生物体基因组高分辨率的遗传图谱、物理图谱、序列图谱以及转录图谱，最终获得一幅完整的、能够在细胞中定位以及在各种生物学代谢、生理、信号传导途径中全部蛋白质在原子水平的三维结构全信息图。

3. 功能基因组学　代表基因分析的新阶段，主要利用结构基因组学提供的信息，发展和应用新的实验以及计算方法，通过在基因组或系统水平上全面分析基因功能，使得生物学研究从对单一基因或蛋白质的研究转向同时对多个基因或蛋白质进行系统的研究。其研究内容具体来说主要有如下几方面：①基因组表达及调控的研究：在全细胞的水平下识别基因组表达产物 mRNA 和蛋白质，以及两者间的互作关系；阐明基因组表达在发育过程和不同环境压力下的调控网络。②基因信息的识别：是提取基因组功能信息必备的基础，通过生物信息学方法(如序列比对、基因组比较及基因预测方法等)和生物学实验(如基因芯片技术、基因组扫描、突变检测体系等)手段完成。③基因功能信息的鉴定：主要包括基因突变体的系统鉴定、蛋白质水平、修饰状态的检测等。④基因多样性分析：基因组的差异反映在表型上就形成个体的多样性，而生物信息学中统计遗传方法对于单核苷酸多态性(single nucleotide polymorphism，SNP)的分析则正是这一内容的主要研究手段之一。⑤比较基因组学(comparative genomics)：是在基因组图谱和测序基础上对已知的基因和基因组(如模式生物等)结构进行比较(常用的工具为 FASTA、BLAST 和 CLUSTAL W 等序列比对软件)，以了解未知功能的基因组内在结构、功能、表达机制并可阐明物种进化关系的学科。根据涉及的物种数目可以将其分为两类，即种间比较基因组学和种内比较基因组学。种间比较基因组学主要研究不同物种在基因组结构上的差异，以发现基因的功能、物种的进化关系等；种内的比较基因组学主要研究个体或群体基因组内的变异和多态现象(如 SNP、微卫星等)，这方面的应用与基因多样性分析密切相关。比较基因

组学不但有助于深入了解生命体的遗传机制，也有助于阐明人类复杂疾病的致病机制，揭示生命的本质规律。

（二）转录组层面

转录组学（transcriptomics）是基因组学后面的一门新兴学科。所谓转录组，就是转录后的所有mRNA的总称，这些能被翻译成蛋白质的编码部分以及非编码部分的功能及相互关系的研究就是转录组的任务。人类基因组项目完成测序以后，转录组的研究正在迅速受到科学家的青睐。科学家发现不编码蛋白质的基因不是垃圾，这部分基因中包含非常重要的调控元件，掌握它们之间的关系，可以从根本上提高对生命规律的基本认识。因为DNA序列本身处在不断演化的过程中（本身的自我复制、插入生成的新基因、受环境影响发生基因突变），这是一个复杂系统的演化过程，需要用系统的眼光看问题，把单个基因还原到基础，再综合起来。人类的许多疾病都与基因的调控有关，利用基因调控治愈现阶段重大疾病是人们非常关注的问题。对于非编码DNA调控功能的深入研究，可以为进一步了解这些重大疾病的发生原因以及解决方法带来新的认识。目前，许多有关的研究工作已经展开。

（三）蛋白质组层面

蛋白质组学（proteomics）是研究一个生命体在其整个生命周期中所拥有的全体蛋白质，或者在较小的规模，即在特定的时间和空间（如特定类型的细胞在某一时期经历特定类型刺激时）所拥有的全体蛋白质，包括表达水平、翻译后的修饰、蛋白-蛋白质互作关系等特征，从而在蛋白质水平上获得对于有关生物体生理、病理等过程的全面认识。与基因组学相比，蛋白质组学更为复杂，因为基因组较为稳定，而蛋白质组是动态的，具有时空性和可调节性，即一个生命体在其机体的不同部分以及生命周期的不同阶段，其蛋白表达可能存在巨大差异。因此蛋白质组学能够反映出某基因的表达时间、表达量、蛋白质翻译后加工修饰和亚细胞分布等。蛋白质组学与传统的蛋白质研究的不同之处在于其研究是基于生物体或细胞的整体蛋白质水平进行的。从整体上看，蛋白质组研究包括两个方面：对蛋白质表达模式（即蛋白质组组成）的研究；对蛋白质组功能模式（目前主要集中在蛋白质互作网络）的研究。研究的关键技术包括质谱分析，X-射线晶体衍射，磁共振和凝胶电泳。蛋白质组数据库（如UniProt等）是蛋白质组研究水平的重要标志。生物信息学的发展已经为蛋白质组学的研究提供了方便有效的计算机分析软件，如蛋白质质谱鉴定软件等。

第二节　生物信息学在生命科学中的地位及意义

Section 2　The Significance of Bioinformatics in Life Science

一、生物信息学内涵

生物信息学作为一门新的交叉学科，其研究范畴广泛，涉及各类组学（基因组学、转录组学、蛋白组学、miRNA组学和药物基因组学等）。近年来，随着研究的不断深入，其研究已逐渐从基因组信息学拓展到功能基因组学、药物基因组学和人类复杂疾病研究等。这里仅对生物信息学的部分研究内容做简单介绍。

1. 生物信息学数据库　随着生物实验方法和检测手段的提高和发展，产生了海量生物学数据和成千上万的数据库。生物信息学数据库几乎覆盖了生命科学的各个领域，如核酸序列数据库，蛋白质序列数据库，蛋白质、核酸等三维结构数据库，文献数据库和其他数据库等。数据量的巨大积累往往蕴含着突破性的发现，也更有助于生物学本质的阐明和理解。

2. 序列比对　最常见的比对是蛋白质序列之间或核酸序列之间的两两比对，通过分析序列之间的相似和差异，寻找二者可能的分子进化关系，进一步的比对是将多个蛋白质或核酸同时进行比较，寻找这些有进化关系的序列之间的保守区域、位点，探索产生共同功能的序列模式，此外，还可以把

蛋白质序列与核酸序列相比来探索核酸序列可能的表达框架，把蛋白质序列与具有三维结构信息的蛋白质相比，从而获得蛋白质折叠类型的信息和预测基因的功能。

3. 分子进化分析　在分子水平上，进化是一种伴随突变的自然选择过程。自从20世纪60年代，由于分子遗传学资料的迅速积累，分子进化逐渐成为计算生物学和生物信息学等新兴学科的重要组成部分。分子进化分析着重于研究不同系统发生树分支上基因和蛋白质的变化方式，随着生物数据资源的拓展，其研究方法和研究方向也在不断发展。

4. 基因芯片数据分析　基因芯片技术改变了生物学研究的方法，从单个基因的研究迅速扩展到全基因组的系统生物学研究，基因芯片技术帮助生物学研究进入后基因组时代。基因芯片技术经过近15年的发展已经形成了一个系统的平台，从样品制备、芯片制作、芯片杂交、数据扫描到后期的数据管理，储存以及深度数据挖掘都有了标准化的流程、坚实的理论和实验的支持，成为一个非常稳定可信的实验技术。

5. 基因功能注释　随着后基因组时代的来临，基因组学的研究重心开始从阐明所有遗传信息转移到在整体分子水平对功能进行研究。快速有效的基因功能注释对进一步识别基因，识别基因转录调控信息，研究基因的表达调控机制，研究基因在生物体代谢途径中的地位，分析基因、基因产物之间的相互作用关系，绘制基因调控网络图，预测和发现蛋白质功能等具有重要的意义。

6. 蛋白质结构分析　蛋白质在生命活动过程中有复杂而精细的生物学功能，通常蛋白质的某种精细的局部结构可用于实现某种局部的生物化学功能，蛋白质高级结构的形成和变化都遵循基本的物理化学原理。发掘蛋白质结构的特征信息是理解蛋白质行使其生物功能的机制、认识蛋白质与蛋白质(或其他分子)间相互作用的基础。

7. 转录调控的信息学分析　近年来，随着基因芯片和高通量测序等数据的出现，计算方法在转录因子结合位点的分析中得到了广泛的应用。并且，利用微阵列芯片的海量数据和日益完善的生物信息学分析工具，对基因转录调控区进行详细分析已成为实验手段的重要补充。

8. 生物分子网络　生命活动本身的复杂性和迅速增加的海量数据资源要求生命现象必须要在成千上万个生物分子组成的复杂系统层面上予以认识。因此，系统全面地研究各生物大分子及其间存在的相互作用成为“后基因组”生物学研究的关键目标。为揭示数量巨大的生物大分子及其间的相互作用如何在复杂的生存环境中行使生物学功能，需要研究者采用不同于传统生物学研究手段的新技术。

9. 计算表观遗传学　计算表观遗传学即是把生物信息学的研究策略和方法应用到表观遗传学的研究领域，具有快速、高通量、低成本的特点，可以为当前的表观遗传学的实验研究提供指导；同时，生物学实验可以用来验证运用计算表观遗传学方法推导的结论。结合实验方法和计算表观遗传学方法，是当前表观遗传学研究领域新兴的视角。

10. 人类复杂疾病与计算系统生物学　近30年来，由于生物科学、计算机技术的迅速发展，人们在分子水平积累了大量的实验数据和研究成果，建立了多个复杂疾病相关的数据库，对疾病有了更深刻的认识，应用计算系统生物学的方法揭示复杂疾病的本质，认识其发病机制，寻找到正确的诊断和防治方法是当前乃至未来几十年复杂疾病研究的主要内容。

11. 单核苷酸多态、miRNA与人类疾病SNP作为分子标记的实验设计，可采用统计学、机器学习等方法对风险SNP遗传进行定位，融合功能信息学和系统生物学知识，产生了面向SNP功能和生物学过程研究复杂疾病的基本理论与方法；随着miRNA在复杂疾病中的深入研究，研究者发现其在疾病的发生发展过程中起着巨大的作用，其功能异常能够导致各种人类复杂疾病(例如癌症，心血管疾病等)的发生。

12. 药物生物信息学　人类基因组计划的目的之一在于阐明人的约10万种蛋白质的结构、功能、相互作用以及与各种人类疾病之间的关系，寻求各种治疗和预防方法，包括药物治疗，基于生物大分子结构的药物设计是生物信息学中的极为重要的研究领域。

综上所述，生物信息学是分子生物学与多学科交叉而形成的新兴学科，已经形成了多个研究方向，随着研究的深入和新思想的引入，其内涵将更加丰富。

二、生物信息学在现代生物医学发展中起着重要作用

21 世纪医学模式将发生革命性变化：① 19 世纪末 20 世纪初，以细胞病理学为基础的医学模式，开始向分子医学（以分子生物学、分子细胞学、分子药理学以及现代计算机技术等为基础）模式转变。人类基因组计划正在建立起人类基因与生理、病理之间关系的知识视图。生物领域的新技术（生物芯片、生物信息学）和新的研究方法（功能基因组学、蛋白组学）在临床中逐步得到应用，更新了医学科学基础。②医疗实践以循证医学为主，从基因、蛋白质等大分子水平研究疾病的发病机制，对疾病进行预防、诊断和治疗。其目标是向特异性诊断、个体化治疗发展。21 世纪，遗传信息在临床环境下的集成应用必将导致个性化医疗等新的临床实践。未来 10 年预防性基因检测会变得普遍，首先应用在具有家族遗传倾向的个体化监测中，未来遗传信息将会对临床医学产生普遍影响，医生将通过病人的基因组数据与 Internet 上可获得的数据库（药物、群体数据、临床档案）进行比较来进行疾病诊断及指导病人治疗，临床医师将能够用计算机输出他们病人的遗传构成，从而能够个性化、有针对性地设计给药。

基于遗传信息的决策支持系统、辅助临床医师解释分子标记数据的专家系统、智能化临床决策支持系统等将成为临床医生必不可少的工具。分子水平生物信息检测设备（基因芯片、蛋白质芯片、质谱仪等）将成为医疗领域的新需求。尤其是微流控基因芯片、蛋白质芯片技术将在 21 世纪成熟并应用于临床，因此，生物芯片数据分析技术及分析系统将成为临床医生的常规工具。

伴随着后基因组时代高通量组学（high-throughput omics）技术涌现与生物信息学的飞速发展，出现了大量潜在的生物标记（biomarker）以及这些标记的模式（pattern）。Mendelsohn 和 Brent 的研究表明，其中的一些可以用于诊断癌症等级（grade）与分期（stage）。这些生物标记信息在临床上的应用潜力是巨大的，然而目前仅有少数的标记用于临床实践。如何将这些生物标记应用于临床诊断、疾病风险评估与预防模式、指导个体化治疗、开发新的药物靶点等是目前医学研究的热点问题，也是转化医学的核心内容。

致癌基因、肿瘤抑制基因以及错配修复基因的突变可以作为生物标记，例如，突变的 *KRAS* 基因预测出不同类型肿瘤的转移扩散，而诸如致癌基因 *RAS*、肿瘤抑制基因 *CDKN2A*、*APC* 和 *RB1* 很有可能可用于诊断以及选择治疗方案，其他的 DNA 标记还包括 SNP、线粒体 DNA 失常等。典型的例子是以表皮生长因子受体（EGFR）为治疗靶标的小分子酪氨酸激酶抑制剂吉非替尼（gefitinib）和厄洛替尼（erlotinib）在临床上广泛应用于肺癌、恶性脑胶质瘤等癌症的治疗。

相对于单个分子的 DNA 标记分析，大多数基于 RNA 的生物标记以多个基因的组合表达模式用于临床评估。将基于模式的 RNA 表达分析应用于乳腺癌，成功地发现了先前未知的与生存时间差异相关的分子亚型。这些研究不但增强了评估预后的能力，评估了新辅助疗法的效能，预测了无淋巴结点转移癌症患者发生转移的可能性，而且能正确预测出肿瘤的等级。针对重要药物代谢酶的转录水平分析，已经用于临床前的肺癌和结肠癌化疗效果的预测。

相似的方法，应用于黑素瘤、胶质瘤、前列腺癌等均有新的发现。典型的例子是美国 Duke 大学的 Potti 等采用 Metagene（133 个基因）的方法，将 71 例ⅠA 期的 NSCLC 患者分成生存率完全不同的两组，预后差的一组，其 5 年生存率相当于ⅡB/Ⅲ期，预测的正确性达 93%，高于临床模型预测正确性的 64%。

目前所有美国 FDA 证实的并已经应用于临床的癌症蛋白标记均是单个蛋白，并且其中大部分得自于血液样本。它们在临床上分别发挥这肿瘤的分期、监测、诊断、治疗方案筛选、预后估计等作用。

药物格列卫（gleevec）、赫赛汀（herceptin）和阿莫西芬（amoxifen）在临床治疗中的成功是针对特定

靶标药物开发的典范，这其中，通过分析可以确定哪些病人更适宜于特定靶标的给药。对于给定的治疗方案，鉴别出最适于此的病患也许是最重要的，herceptin 就是一个很好的例子。在未经筛选的乳腺癌患者中，11%~26% 的病患其肿瘤转移得到了抑制，而对于 ERBB2 呈阳性的患者，有效率增至 34%。很明显，herceptin 的成功与鉴别更可能受益的病患紧密相连。尽管在过去的 20 年中，以致癌基因为靶标开发了大量的药物，但其缺陷在于没有针对适宜的病患群体。所以，多数靶向的治疗药物通常效能不佳。可见，应用分子标记区分药物敏感的病患亚群是十分重要的。

综上，复杂疾病的治疗，逐渐走出实验室，迅速进入转化研究阶段，其重要标志，就是依据基因组学或蛋白组学的临床研究。复杂疾病的发生与发展是一个多基因参与的、多步骤的、复杂的生物学过程，仅仅依据病理类型、临床分期以及患者年龄、行为状态等临床特征选择治疗方法已远远达不到个体化数字化治疗的要求。通过生物信息学的方法研究复杂疾病的组学图谱，全面详尽地了解疾病的生物学特性来指导临床治疗，是未来医疗的必由之路。

（李 霞 李亦学）

第一篇 生物信息学基础

第一章 DNA、RNA 和蛋白质序列信息资源

CHAPTER 1 DNA、RNA AND PROTEIN SEQUENCE INFORMATION RESOURCES

第一节 引 言

Section 1 Introduction

随着近年来生物实验方法和检测手段的发展，积累了大量生物学，尤其是分子生物学实验数据。通过对这些数据的分类、收集、整理，产生了成千上万的数据库。生物信息学中的各类数据库几乎覆盖了生命科学的各个领域，如核酸序列数据库，蛋白质序列数据库，蛋白质、核酸、多糖的三维结构数据库，基因组数据库，文献数据库和其他种类数据库。目前，生物信息数据库大致可以分为四类：基因组数据库、核酸和蛋白质一级结构序列数据库、生物大分子（主要是蛋白质）三维空间结构数据库以及由这三类数据库和文献资料为基础构建的二次数据库。前三类数据库是生物信息学的基本数据资源，通常称为基本数据库或初始数据库，也称一次数据库。根据生命科学不同研究领域的实际需要，对基因组图谱、核酸和蛋白质序列、蛋白质结构以及文献等数据进行分析、整理、归纳、注释，构建具有特殊生物学意义和专门用途的二次数据库，也称专门数据库、专业数据库或专用数据库。一次数据库的数据量大、更新速度快、用户面广，但存在过多的冗余数据。二次数据库的容量比较小，更新速度也没有一次数据库那样快，经过筛选后，避免了过多的冗余数据。根据数据库存储的内容可将生物信息学数据库分为核酸、蛋白质、基因图谱、结构、文献等数据库。

第二节 核酸序列数据库

Section 2 Nucleic Acid Sequence Databases

自 20 世纪 80 年代第一个核酸数据库建立以来，核酸数据库迅速发展。在互联网上不仅有核酸序列数据库，还出现了基因组相关数据库、核酸三维结构数据库、基因表达数据库、人类基因突变及疾病相关数据库、进化相关数据库及其他与核酸有关的数据库。

除了本章重点介绍的 GenBank、EMBL 和 DDBJ 三大核酸序列数据库外；还有我国的核酸序列公共数据库（BIOSINO）；特殊类型核酸序列数据库：非编码 RNA 数据库（ncRNA）、表达序列标签数据库（dbEST）、序列标签位点数据库（dbSTS）等，其他还有 miRBase、tRNAdb 等；基因组相关数据库：人类基因组数据库（HGD）、基因组序列数据库（GSDB）、基因组在线数据库（GOLD）等；核酸三维结构数据库：核苷酸三维结构数据库（NDB）、普纳大学核酸结构数据库（BNASDB）等；基因表达数据库：基因表达库（GEO）、斯坦福微阵列数据库（SMD）以及 ArrayExpress、CGED、GXD、BodyMap 等；人类基因突变及疾病相关数据库：人类基因变异数据库（HMGD）、人类遗传双等位基因序列数据库

(HGBASE)、人类孟德尔遗传在线(OMIM)、国际单体型计划(HapMap)、人类单核苷酸多态性数据库(dbSNP)、肿瘤基因数据库(TGDB)、疾病关联数据库(GAD)、癌症基因数据库(CGAP)、人类表观遗传数据库(HEP)、人类 DNA 甲基化与癌症数据库(MethylCancer)等。

一、GenBank 数据库

GenBank(http://www.ncbi.nlm.nih.gov/genbank/)是一个综合数据库,该数据库中包含了已经公开的 30 万余种不同物种生物的核酸序列,这些数据主要来源于全世界不同实验室和大规模测序计划项目。多数序列信息通过网络版的 BankIt 程序或独立的 Sequin 程序提交给 GenBank 数据库,GenBank 工作人员在接收序列数据后,赋予该序列特定的数据库记录号(登录号)。GenBank 数据库每天与欧洲分子生物学实验室的核酸序列数据库(European Molecular Biology Laboratory Nucleotide Sequence Database, EMBL)和日本的 DNA 数据库(DNA Data Bank of Japan, DDBJ)进行数据交换,以保证数据库内容在全世界范围的同步性。通过美国国家生物技术信息中心的检索系统(Entrez)可以进入 GenBank。Entrez 检索程序整合了主要的 DNA 和蛋白质序列数据的分类学、基因组、图谱、蛋白质结构和结构(功能)域信息,还包括相关的 PubMed 的生物医学文献信息。BLAST 程序提供 GenBank 和其他序列数据库中序列相似性搜索服务。通过 FTP 站点可以获得每两个月发布的和每天更新的 GenBank 数据。在 NCBI(http://www.ncbi.nlm.nih.gov/)的主页上提供了进入 GenBank 的路径、相关检索和分析服务。

GenBank 是具有目录和生物学注释的核酸序列综合数据库,由美国国家医学图书馆(the National Library of Medicine, NLM)的国家生物技术信息中心(the National Center for Biotechnology Information, NCBI)构建、维护和管理。该中心位于美国马里兰国家健康研究所(the US National Institutes of Health, NIH)。GenBank 数据库的序列数据来源于序列发现者提交的序列、批量提交的表达序列标签(expressed sequence tag, EST)、基因组勘测序列(genome survey sequence, GSS)和其他测序中心提供的高通量数据,还包括美国专利商标局提供的已发表专利的序列数据。GenBank、EMBL、DDBJ 组成国际核酸序列数据库合作组织(the International Nucleotide Sequence Database Collaboration, INSDC),该组织成员远程合作,每天相互交换数据以保证世界范围内序列信息的一致性和完整性。NCBI 通过网络免费提供 GenBank 数据,通过 FTP 或宽带网提供检索和分析 GenBank 数据服务。

(一) GenBank 数据库结构

自 GenBank 建立以来,数据库的数据量持续呈指数增长,大约每 35 个月翻一番。传统的 GenBank 部分包含 1.08 亿条序列,1.06 千亿个碱基数据。在 2010 年以前的一年内,增加了 1.1 千万条新序列。全基因组鸟枪测序(whole genome shotgun, WGS)项目提供的数据加上传统部分数据库的数据已经超过 2.55 千亿个碱基。全基因组测序数据在数据库中快速、持续增加,GenBank 现有 1000 余种微生物全基因组数据,其中 30% 数据是在过去的一年内提交给数据库的。真核生物基因组的数量也在持续增加,现有数据库中,包括人类参考基因组数据,已有 380 余种全基因组拼接数据。

1. 依据序列的物种来源分类　使用综合序列分类软件(www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy)可以检索分类数据库序列,该软件是由 NCBI、EMBL 和 DDBJ 及合作组织之外的顾问和专家共同合作研发的。在 GenBank 数据库中,现有超过 30 万已命名物种的数据信息,并且还以每月 2500 个新物种的速度增加。在 GenBank 数据库非 WGS 序列中(表 1-1),大约 12% 是源于人类的序列,其中 8% 是表达序列标签。

2. Genbank 记录和分类　每条 GenBank 记录包含简要的序列描述、学名、物种来源分类、参考文献和列举生物意义的特征表(www.ncbi.nlm.nih.gov/collab/FT/),如编码区、蛋白质翻译、转录单位、重复区域、变异和修饰位点等。GenBank 中的数据文件,按惯例被分成独立的“分类”,根据主要分类群分为细菌(BCT)、病毒(VRLA)、灵长类(PRI)和啮齿类(ROD)。近年来增加了测序策略的分类,

表 1-1 在 GenBank 非 WGS 序列记录中，碱基数量居于前列的物种

Organism	Non-WGS bases (billions)	Organism	EST records (millions)
Homo sapiens	13.7	*Homo sapiens*	8.3
Mus musculus	8.4	*Mus musculus*	4.8
Rattus norvegicus	6.3	*Zea mays*	2
Bos taurus	5.3	*Sus scrofa*	1.5
Zea mays	5	*Arabidopsis thaliana*	1.5
Sus scrofa	4.2	*Bos taurus*	1.5
Danio rerio	3.1	*Danio rerio*	1.5
Strongylocentrotus purpuratus	1.4	*Glycine max*	1.4
Nicotiana tabacum	1.2	*Xenopus tropicalis*	1.3
Oryza sativa(Japonica group)	1.2	*Oryza sativa*	1.2

注：引自 Nucleic Acids Res., 2010 January, 38(Database issue): D46–51.

包括表达序列标签、基因组勘测序列、高通量基因组(high-throughput genomic, HTG)、高通量 cDNA (high-throughput cDNA, HTC)和环境样品(environmental sample, ENV)序列等，总共分为 20 个类别。为了方便文件传输，GenBank 数据被分解为多个文件，在每两个月发布的 FTP 站点上现有的数据库被分解为 1800 余个文件。

(1) 表达序列标签(EST)：一直以来，表达序列标签占据了新序列记录和基因序列的主要部分，在 GenBank 发布的第 173 期中包含 340 亿个核酸碱基的 EST 序列。一年内，EST 增加了 14%，在 1800 个物种中检测到了总共 6.28 千万条记录。对于新的 EST 序列，NCBI 使用 BLAST 程序搜索所有数据库中的同源性，并且将其合并加入到伴侣数据库 dbEST(www.ncbi.nlm.nih.gov/dbEST/)中。在 dbEST 中的数据经进一步处理，生成 UniGene(www.ncbi.nlm.nih.gov/unigene)数据库。

(2) 序列标签位点(STS)、基因组勘测序列(GSS)和环境样品序列(ENV)：GenBank 的 STS(www.ncbi.nlm.nih.gov/dbSTS/)部分包含 1.3 百万条序列，包括基于基因组序列的匿名 STSs 以及来源于基因和 ESTs 的 3' 端基因 STSs。这些 STS 记录通常包含图谱信息。GenBank 的 GSS(www.ncbi.nlm.nih.gov/dbGSS/)部分在 2008 年增加了 6%，具有超过 870 个物种的 2.58 千万条记录，数据量超过 1.67 百亿个碱基。GSS 序列来自于 80 余种不同的实验技术。人类 GSS 数据与 STS 记录曾被用于人类基因组计划的 BAC 重叠拼接。GenBank 的 ENV 部分的序列数据采用非 WGS 测序方式，获得环境样品序列的物种来源是未知的。许多 ENV 系列数据来自宏基因组样品，这些微生物样品是从不同动物组织如内脏、皮肤或淡水沉积物、温泉、矿井污水排泄区的特殊环境中获取的。环境样品序列记录中在关键词字段标明“ENV”，并且在来源特征处标明“/environmental_sample”。

(3) 高通量基因组(HTG)和高通量 cDNA(HTC)序列：GenBank 中 HTG(www.ncbi.nlm.nih.gov/HTGS/)部分包含尚未完成的大规模基因组记录，依据这些记录数据的质量分为 0~3 时相，处于 0、1、2 时相的数据记录是尚未完成的记录，在一定时间内，会转变为完成状态。达到 3 时相的记录，就是处于完成状态的数据，完成状态的 HTG 记录被移动到 GenBank 中相应的物种数据库。在 GenBank 发布的 173 期公告中，HTG 部分已包含 14.2 万条序列，达到 2.39 百亿个碱基对数据。

GenBank 中的 HTC 部分由高通量 cDNA 序列组成，HTC 序列是初级质量的序列，可能包含 5' 端和 3' 端非编码区序列，部分编码区和部分内含子序列。完成后的高质量 HTC 序列，被移动到相应的 GenBank 的物种数据库。每个测序计划生成的 HTC 数据都有参考文献加以描述。

(4) 全基因组鸟枪测序序列(WGS)：在 GenBank 中有 1.48 千亿个碱基的 WGS 重叠拼接组序列，它们中许多序列都有某一测序计划来源的注释，已公布的这些序列编号包含 4 字符的计划代码，后跟两个数字版本号码和六个数字拼接代码。因此，编号为“AAAA01072744”的 WGS 记录，表示

计划“AAAA”第一版本拼接编号为“072744”的序列。全基因组鸟枪测序计划向 GenBank 提交了4.8 千万拼接序列，这些原始序列已被用于构建 6.4 百万大规模染色体骨架拼接。在数据库中可以获得人类、黑猩猩、猕猴、马、犬、果蝇、酵母和 1800 余个其他物种的拼接基因组序列以及环境样品的 WGS 计划拼接序列。

尽管 WGS 计划序列可能被注释，但由于有些测序计划正在进行或尚未完成，注释可能还没有从上一个拼接版本顺延到下一个版本或者有些测序计划被认为是准备阶段，许多低覆盖序列并不含有注释内容。通常，鼓励 WGS 序列和基因组序列提交者使用新的格式为“/experimental=text”和“/inference=TYPE:text”证据标签，其中，“TYPE”是多种标准推论类型之一，“text”是由结构化文本组成。

（5）转录组鸟枪组合序列(transcriptome shotgun assembly sequences)：近年来，数量不断增加的测序踪迹序列存放在 NCBI 的踪迹档案(trace archive，TA)中。在 2007 年，鉴于新一代测序技术的出现，如“Roche-454 Life Sciences”、“Illumina Solexa”和“Applied Biosystems SOLiD”，NCBI 建立了序列读取档案(sequence read archive，SRA)。不论 TA 还是 SRA 都是 GenBank 的一部分，但自从发布第 166 期公告后，GenBank 为转录组鸟枪组合序列增加了新的 TSA 数据部分，鸟枪序列的组合被存放在 GenBank 的 TA、SRA 和 EST 的数据部分中。TSA 记录内(如：EZ000001)关键词字段含有“TSA”，并且原始栏提供 TSA 序列组合的碱基范围和序列标识符。

3. 特殊记录类型

（1）第三方注释(TPA)：在 DDBJ/EMBL/GenBank 的第三方注释(third party annotation，TPA)记录允许原始序列记录提交者以外的科学家对公开序列加以注释(www.ncbi.nlm.nih.gov/genbank/TPA.html)。TPA 记录分为三种类型：①experimental，在这种情况下，有直接实验证据表明注释分子的存在；②inferential，在这种情况下，实验证据是间接的；③reassembly，在此类型中主要提供初始阅读序列的较好组合序列。TPA 序列可能是由多条原始序列组合生成的。TPA 记录的格式(如：BK000016)与常规 GenBank 记录相似，但标明“TPA_exp:”、“TPA_inf:”或“TPA_reasm:”，在每条定义行和关键词字段标明“Third Party Annotation，TPA”和“TPA: experimental”、“TPA: inferential”或“TPA: reassembly”。TPA 实验和推论记录如 TSA 记录一样也有原始栏。在权威生物学杂志发表前，TPA 序列不会公开它的记录号、序列数据和注释内容。TPA 可以使用 BankIt 或 Sequin 提交给 GenBank。

（2）GenBank CON 记录：是指较小记录组合记录。尽管许多基因组在 GenBank 中，如细菌基因组，以一条序列表示完整基因组，但是为了便于数据传输和分析，对于部分真核生物基因组，人们希望将非常长的序列分解成较小片段数据一系列叠连群(contig CON)组合。这样，就产生了完整基因组的 CON 分割记录，该记录包含组合指令可以无缝显示和下载全部序列。许多 CON 记录也含有注释内容。

（二）构建数据库

数据库中的数据是由不同的序列发现者分别提交给 GenBank、EMBL 和 DDBJ 的，批量的 EST、STS、GSS、HTC、WGS 或 HTG 序列一般由测序中心提交，GenBank 每天与 DDBJ 和 EMBL 进行数据交换，NCBI 服务器与现有的很多序列资源库每天进行数据更新。

1. 直接电子提交　几乎所有的 GenBank 记录都是通过直接电子提交方式将数据添加到数据库的，绝大多数序列发现者使用 BankIt 或 Sequin 程序提交序列数据给 GenBank。很多杂志要求论文中涉及序列数据的作者，在发表论文之前，必须提交序列数据到公开数据库。

GenBank 几乎每天接收 1600 条记录，工作人员在接到提交序列的两个工作日内，为提交的序列指定记录号。获得记录号后，表明提交已完成，可以在数据库中检索到该序列。序列直接提交后会收到质量确信评审，内容包括：载体污染检查、正确的编码区翻译验证、准确分类验证和文献出处检查。记录在进入数据库之前，GenBank 记录的草稿返回给序列发现者，发现者在论文发表前可以要

求保密。由于 GenBank 规定，序列或记录号在杂志发表后，必须公开存放在数据库中的序列数据，因此，要求作者通知 GenBank 工作人员引证序列论文的发表日期，以便及时公布序列数据。尽管只允许提交序列的科学家修改序列或注释内容，但还鼓励所有用户通过 update@ncbi.nlm.nih.gov，在序列数据公布后，针对该序列进行报道、纠错或删改。

NCBI 与测序中心密切合作，以保证及时公布纳入 GenBank 的批量数据。GenBank 为大规模测序团体提供批量处理程序，如“tbl2asn”，以便于批量数据提交。

（1）使用 BankIt 提交：BankIt(http://www.ncbi.nlm.nih.gov/BankIt/)是 GenBank 为用户提供的三个网络序列数据提交工具之一。提交者可以直接在表格中填写序列信息，加入编码区或 mRNA 特征等生物学注释。BankIt 具有自由格式文本栏、列表栏和下拉菜单，允许提交者不必学习格式规则和描述用语就可以进一步描述序列。在生成提交 GenBank 纯文本格式草稿前，提交者可以回顾检查。BankIt 能够确认提交内容、标记常见错误和使用 Vecscreen BLASTer 软件检查序列中是否包含载体污染序列。BankIt 工具适用于简单提交，特别是只有一条或很少几条记录需要提交时，应该选择 BankIt。另外，BankIt 还可以用于更新数据库中已有的 GenBank 记录。

（2）使用 Sequin 和 tbl2asn 提交：NCBI 还提供了独立的多平台提交程序——Sequin(www.ncbi.nlm.nih.gov/Sequin/index.html)，该程序可以与其他 NCBI 序列检索和分析工具交互使用。Sequin 可以处理 cDNA 简单序列以及分段条目、系统发生分析、族群分析、变异分析、环境样品和序列相似性比较。而 BankIt 和其他网络提交工具不具备上述这些功能。Sequin 具有简便的编辑和复杂的注释功能以及嵌入式的序列质量验证、确认功能。另外，Sequin 还适用于如 5.6Mb 的大肠杆菌基因组那样大的序列处理，通过简单图表可以读入全部注释内容。通过匿名 FTP 站点 ftp.ncbi.nih.gov 中的“sequin”路径可以获得苹果、PC 和 Unix 计算机版本的程序。一旦提交完成，提交者可以发送 Sequin 文件到“gb-sub@ncbi.nlm.nih.gov”信箱。

对于包含基因和注释很多的基因组数据，使用 tbl2asn 提交比较方便。与直接提交不同，使用注释软件可以将注释表转换成 ASN.1(Abstract Syntax Notation One)码记录，提交给 GenBank。

（3）条形码序列提交：生物条形码集团(consortium for the barcode of life，CBOL)是开发使用 DNA 条形码作为工具鉴别物种特征的国际组织，DNA 条形码使用短的，通常为 648 个碱基细胞色素氧化酶亚单位 I 的部分 DNA 序列作为物种鉴别序列。NCBI 与 CBOL 合作，开发在线提交给 GenBank 批量条形码序列的工具(BarSTool)，该工具可允许用户批量更新含有相关资源信息数据的文件。

2. 序列标识符和记录号　每条具有序列和注释内容的 GenBank 记录，都具有特有的标识符，该标识符被称作记录号。该标识符在合作数据库(GenBank，DDBJ，EMBL)中通用，并且在该记录有效期间始终不变，甚至，记录中序列和注释变动，标识符也不改变。在 GenBank 记录中，每个 DNA 序列版本被指定了 NCBI 标识符，称作“gi”，在 GenBank 纯文本文件记录的版本行中跟在记录号之后。在纯文本记录的版本行还有第三个记录版本格式标识符，包含 gi 和记录号所代表的信息。首次在数据库中出现的记录版本标识符在 GenBank 记录编号后面跟“.1”，表明这条记录序列是第一版本。如：

ACCESSION AF000001

VERSION AF000001.1 GI:987654321

当 GenBank 记录中某一序列发生变化时，就会给这条序列赋予新的 gi 号和记录版本标识符增加的版本号，对于原记录的编号全部保持不变，使用旧的记录版本标识符和 gi，仍然可以检索到旧的序列。

利用相似的系统可以跟踪蛋白质翻译的改变，这个标识符在 GenBank 条目的特征部分用来标识 CDS 特征，如：/protein_id=“AAA00001.1”。蛋白质序列翻译也会获得它自己独有的 gi 编号，在 CDS 特征中作为第二个标识符出现，如：

CDS feature:/db_xref='GI: 1233445'

（三）检索 GenBank 数据

1. Entrez 系统　使用 Entrez(http://www.ncbi.nlm.nih.gov/sites/gquery)检索系统可以访问 GenBank 中的序列记录。Entrez 是一个灵活的数据库检索系统，可以检索 35 个数据库。Entrez 数据库包含来源于 GenBank 和其他资源的 DNA 和蛋白质序列，还包括基因组图谱、种群、进化和环境序列数据集、基因表达数据、NCBI 分类学、蛋白质结构域信息和来源于分子模型数据库(molecular modeling database，MMDB)的蛋白质结构数据库，每个数据库经由 PubMed 和 PubMed Central 与相关学术文献关联。

2. 与测序计划相关的序列记录　确认所有 GenBank 记录提交的组织和他们的侧重目标是分析大量序列数据的基础。利用物种或提交者名字定义序列数据集是不可靠的。起始于 NCBI，后由 INSDC 接管的基因组计划数据库(http://www.ncbi.nlm.nih.gov/genomeprj)，允许测序中心登记测序计划，并且赋予其特有的计划标识符，以保证测序计划与该计划产生的数据建立可靠的联系。

在 GenBank 纯文本文件中出现的新的“PROJECT”定义行用于确定测序计划与 GenBank 某条序列记录是关联的。该定义行包含多种格式标识符，如“type”和“value”，分别用分号分开。下面表明了基因组计划与 GenBank 序列记录的关系。

DBLINK Project:18787

基因组计划记录“18787”提供了安乐蜥测序工作的详细进展。

在 Entrez 系统内，这样的序列记录直接连接到相应的基因组计划记录，基因组计划记录也连接到相关联的序列记录。

3. BLAST 序列相似性搜索　序列相似性搜索是 GenBank 数据最基本和使用最多的分析方式。NCBI 提供 BLAST(http://blast.ncbi.nlm.nih.gov/Blast.cgi)系列程序集，用于检测一条查询序列与数据库所有序列的相似性。BLAST 搜索可以在 NCBI 网站上运行，也可以在 FTP 站点下载独立的程序集运行。

4. 用 FTP 获取 GenBank　NCBI 以传统的纯文本文件格式发布 GenBank 数据，并以 ASN.1 格式进行内部维护。通过 NCBI 匿名 FTP (http://ftp.ncbi.nih.gov/genbank)站点可以获得每两个月的全文公告以及每天 EMBL 和 DDBJ 数据更新的内容，还可以从印第安纳大学的镜像站点(ftp://bio-mirror.net/biomirror/genbank/)下载。在“genbank”目录中的“daily-nc”(http://ftp.ncbi.nih.gov/daily-nc/)文件夹中可以获得纯文本的更新数据压缩文件。在 GenBank FTP 站点的“tools”(http://ftp.ncbi.nih.gov/tools/)目录中提供转换每日更新数据集的脚本。

二、EMBL 数据库

欧洲分子生物学实验室(EMBL)核酸序列数据库(http://www.edi.ac.uk/embl/)是欧洲主要的核酸序列收集单位，欧洲生物信息中心(EBI)即 EMBL 在德国海德堡的站点维护这个数据库。

核苷酸数据来自基因组测序中心、世界各地的科学家、欧洲专利局以及与合作伙伴 DDBJ(Japan)和 GenBank(USA)交换的数据。为了达到最佳的同步性，每天 DDBJ/EMBL/GenBank 之间都要交换最新的数据。用户只要进入三者中任意一个数据库都能得到最新数据。这三个数据库之间遵循统一的数据文件指导方针，规范了数据库登录的内容和语法。这种指导方针不仅保证了这些数据库的信息以便捷的格式进行交换，而且与当今的生物信息学软件兼容，反映了分子生物学领域的发展。

EMBL 建立于 1980 年，它起初保存的数据信息是发表在科学文献上序列信息的两倍。现在，主要的测序中心(如：Sanger 测序中心)为 EMBL 数据库提供了比较多的数据。在 2009 年 11 月 21 日发布的公告中宣布，EMBL 数据库已保存了包含 265 969 305 274 个核苷酸的 164 218 403 条记录。通过网址(http://www.ebi.ac.uk/Services/DBstats/)可以看到数据库统计信息。

三、DDBJ 数据库

日本 DNA 数据库(DDBJ)是亚洲唯一的核酸序列数据库，是搜集研究者公认的测定核酸序列的数据库，并且发放给数据提交者国际认证的核酸序列编号。由于 DDBJ 每天将搜集的数据与 EMBL-Bank/EBI 和 GenBank/NCBI 进行交换，使得三个核酸数据库几乎在任何时候都享有相同数据。这种几乎一致的数据库被称作“国际核酸序列数据库(INSD)”。DDBJ 主要收集来自日本研究者获得的序列数据，但也收集数据和发放编号给任何其他国家的研究者。

DDBJ 是由信息生物学中心和国家遗传研究所的日本 DNA 数据库(CIB-DDBJ)共同组建的，90% 来自日本研究者的数据通过 DDBJ 提交。DDBJ 工作的主要任务是共同提高 INSD 的质量。当研究者通过 INSD 公开他们的数据后，全世界将共享这些序列信息，DDBJ 根据 INSD 统一规则尽量详尽地标注这些数据的信息，使用户更好地利用 DDBJ。

四、其他数据库

(一) dbEST

dbEST 是 GenBank 的一个子数据库，包含来源于不同物种的表达序列数据和表达序列标签序列的其他信息。人类表达序列标签(EST)是由随机选择的 600 多个人脑互补 DNA(complementary DNA，cDNA)自动生成的部分 DNA 序列。EST 已被用于人类新基因的发现、人类基因组图谱绘制和基因组序列编码区识别。利用 EST 数据库，人们发现了 337 条代表新基因的序列，包括与其他物种明显相似的 48 条序列，如：酵母 RNA 聚合酶Ⅱ亚单位、果蝇的驱动蛋白、与翅缘发育有关的 Notch 基因和分裂相关的增强子，以及小鼠酪氨酸激酶受体等序列，其中 46 条 EST 是经 PCR 扩增后绘制到染色体图谱上的。这种利用 cDNA 特性的快速方法可以在几年内方便、廉价地为人类绝大多数的基因加注标签，为不同生物学研究领域提供新的基因组标志物资源。

(二) ncRNAdb

非编码 RNA(non-coding RNA，ncRNA)数据库旨在提供非编码 RNA 的序列和功能信息。非编码转录物不编码蛋白质，但在细胞中起调节作用。目前，该数据库包含来源于 99 种细菌、古生菌和真核生物的 3 万多条序列。

该数据库中涉及的原始序列资源是 GenBank 数据库从 FANTOM3(http://fantom.gsc.riken.jp/3/)数据库和 H-Iev 人类基因综合注释数据库(http://jbirc.jbic.or.jp/hinv/ahg-db/index.jsp)3.4 版本获得的小鼠和人类 ncRNA 附加注释信息。基因组图谱信息来源于 UCSC Genome Browser(http://genome.ucsc.edu/)站点。细菌的小细胞质 RNA 序列和注释内容中缺少基因组序列的部分来源于 Rfam(http://rfam.sanger.ac.uk/)数据库。

近年来，为了避免在特殊数据库中过度冗余，在以前版本中的 microRNAs 或 snoRNAs，以及其他管家(基础)RNAs(如：rRNA、tRNA、snRNA、SRP RNA)不包含在 ncRNA 数据库中。该数据可以通过以下四种方式检索：

1. Search ncRNA database：http://biobases.ibch.poznan.pl/ncRNA/。
2. Blast：http://ncrnadb.trna.ibch.poznan.pl/blast.html。
3. Browse Information pages：http://ncrnadb.trna.ibch.poznan.pl/Browser.html。
4. Download：http://ncrnadb.trna.ibch.poznan.pl/download.html。

(三) miRBase

miRBase 序列数据库主要是存放已发表的微小 RNA(microRNA 或 miRNA)序列和注释的数据库。miRBase 使用友好的网络界面，为用户提供 miRNA 数据，允许用户使用关键词或序列检索数据库，通过关联信息链接到 miRNA 的原始参考文献，分析基因组中的定位和挖掘 miRNA 序列间的关系。miRBase 还提供保密的基因命名服务，在新基因发表前指定正式的 miRNA 名称。

在第 14 期公告中，miRBase 序列数据库已经突破 1 万条记录。miRBase 数据库已经将网址转到曼彻斯特大学生命科学院的主机上(http://www.mirbase.org/)。

miRBase 提供如下服务：

1. miRBase(http://www.mirbase.org/)是可以检索已发表的 miRNA 序列和注释内容的数据库，miRBase 序列数据库中每一条记录描述一个预测的 miRNA 转录物发夹部分(在数据库中用术语 mir 表示)，还包含成熟 miRNA 序列的定位和序列信息(用术语 miR 表示)。

2. 通过检索(http://www.mirbase.org/search.shtml)和浏览(http://www.mirbase.org/cgi-bin/browse.pl)可以获得 miRNA 的发夹和成熟序列，还可以使用名称、关键词、参考文献和注释内容查询所有记录，也可以通过网站(http://www.mirbase.org/ftp.shtml)下载所有序列和注释数据。

3. miRBase Registry(http://www.mirbase.org/registry.shtml)为科学家发现新 miRNA 提供专有的名称。

4. miRBase 标签数据库更名为 microCosm，被存放在 EBI 的 MicroCosm Targets 站点中(http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/)，microCosm 资源继续由“Enright group”(http://www.ebi.ac.uk/enright/)管理。目前，miRBase 数据库中 miRNAs 关联到由 microCosm 预测的靶基因，将来 TargetScan(http://www.targetscan.org/)和 Pictar(http://pictar.mdc-berlin.de/)会提供更广泛的靶基因预测综合服务。

第三节 蛋白质序列数据库

Section 3 Protein Sequence Database

随着分子生物学的发展，人们获得了越来越多关于蛋白质序列、结构和功能的信息。世界各国的生物学家和计算机科学家合作利用这些信息构建了蛋白质序列数据库、蛋白质三维结构数据库、蛋白质组数据库(二维凝胶电泳数据库)、信号传导及蛋白质 - 蛋白质相互作用数据库、DNA 和蛋白质相互作用数据库等。常用的蛋白质序列数据库主要有 PIR、MIPS 和 Swiss-Prot 数据库。另外还有与蛋白质功能数据库、结构域和蛋白质家族有关的数据库：PROSITE、InterPro、Pfam、ProDom、SMART 等；蛋白质三维结构相关数据库：PDB、BioMagResBank、SWISS-MODEL Repository、ModBase、CATH、SCOP、ReLiBase、TOPS、SWISS-3DIMAGE 和 BioImage 等；蛋白质二维凝胶电泳数据库：WORLD-2DPAGE、Phoretix links；信号传导及蛋白质 - 蛋白质相互作用数据库：DIP、INTERACT、ProNet、KEGG、CANSITE、SPAD、CSNDB 等；DNA 和蛋白质相互作用数据库：DPInteract；蛋白质翻译后修饰数据库：O-GlycBase、PhosphoBase、RESID 等。下面重点介绍常用的 PIR、MIPS 和 Swiss-Prot 蛋白质序列数据库。

一、PIR 数据库

蛋白质信息库(protein information resource，PIR)(http://pir.georgetown.edu/pirwww/)是一个支持基因组学、蛋白质组学用于系统生物学研究的综合公共生物信息学资源。

PIR 由美国国家生物医学基金会(NBRF)于 1984 年建立，帮助研究者确认和解释蛋白质序列信息的数据库。在此之前，NBRF 首次广泛收集了 1965~1978 年，Margaret O. Dayhoff 和她的研究小组编辑的大分子蛋白质序列和结构图谱。这个研究小组率先用计算机方法比较蛋白质序列，检测远源序列间的相关性和序列内部的重复，依据蛋白质序列相似性比较，推论蛋白质分子进化历史。

40 多年来，从蛋白质序列和结构图谱开始，PIR 免费为科学界提供包括蛋白质序列数据库(protein sequence database，PSD)在内的蛋白质数据库和分析工具。

2002 年，PIR 与国际合作伙伴 EBI(欧洲生物信息学研究所)和 SIB(瑞士生物信息学研究所)在 NIH 资助下建立了世界上唯一的与 PIR-PSD、Swiss-Prot 和 TrEMBL UniProt 数据库一致的蛋白质序

列和结构数据库。

现在，PIR 提供的数据库见图 1-1。

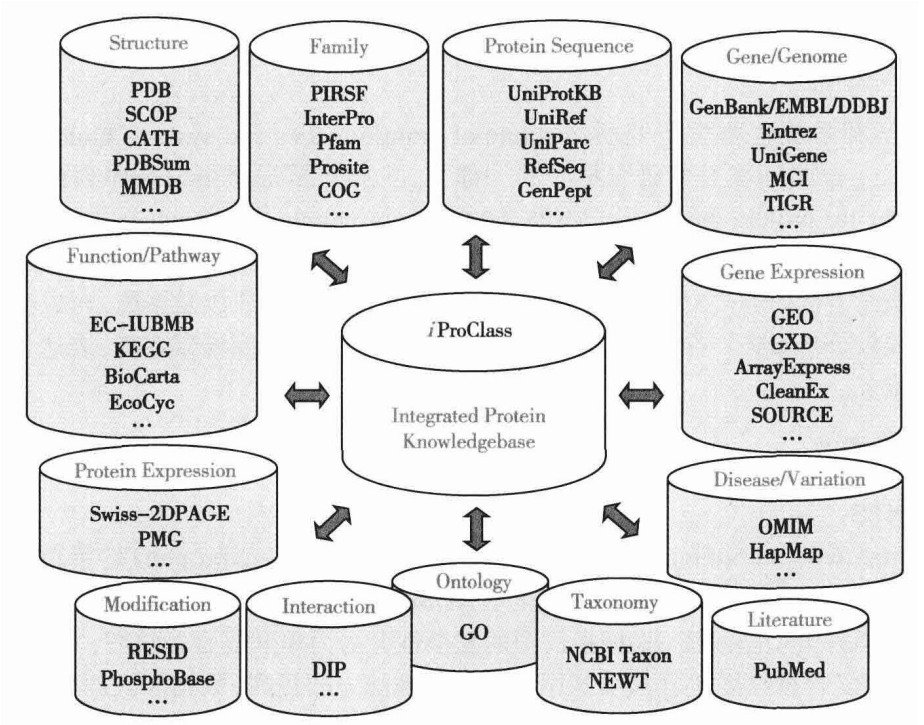


图 1-1 PIR 信息库资源

PIR 主要包括：

1. UniProt- 通用蛋白质资源库 UniProt(http://www.uniprot.org/)是存储和链接其他蛋白质数据库的资源库，并且是蛋白质序列和具有综合功能注释目录的中心资源库。使用 UniprotKB 可以检索准确、可靠的蛋白质综合信息。使用 UniRef 可以减少冗余，加速序列相似性搜索。使用 UniParc 可以检索存档序列和它们来源的数据库。

2. iProClass- 蛋白质知识整合数据库 iProClass(http://pir.georgetown.edu/iproclass/)提供来自 90 多个生物学数据库的大量整合数据，包括蛋白质 ID 图谱服务、UniProtKB 编注蛋白质摘要描述和筛选 UnParc 数据库的蛋白质序列。使用 iProClass 可以检索最新的蛋白质综合信息，包括：功能、转导通路、相互作用、家族分类、基因和基因组、功能注释标准体系(ontology)、文献和分类学信息。使用 iProClass 还可以检索 ID 图谱、蛋白质词典和相关序列。

3. PIRSF- 蛋白家族分类系统 PIRSF(http://pir.georgetown.edu/pirsf/)蛋白家族分类系统是根据超家族到亚家族序列分歧构建的多级网络分类系统，序列分歧反映了全序列蛋白和功能域进化的关系。主要的 PIRSF 分类单位是同源家族，家族成员具有同源性(进化来自共同的祖先)和同拓扑性(具有全长序列相似性和共同的功能域结构)。PIRSF 人工排列家族成员，注释特殊的生物学功能、生物化学活动和序列特征。另外，制定功能位点和蛋白质命名规则，协助传播和规范蛋白质注释标准，系统地检测注释错误。PIRSF 的报告提供研究蛋白质进化相关的独立平台。它概要论述家族的特征，如家族名称、分类分布、分级和功能域结构，以及家族成员，包括功能、结构、传导通路、功能注释标准体系和家族分类，还具有广泛的相关数据库的链接。利用这些信息可以获得所关注蛋白质的准确功能或预测的功能及该蛋白质所属家族成员共有的其他特征。

4. iProLINK- 蛋白质文献、信息和知识整合数据库 iProLINK(http://pir.georgetown.edu/iprolink/)提供有关注释内容的文献、蛋白质名称词典和其他有助于文献挖掘的人文语言处理技术开发的信息、数据库校正、蛋白质名称标记和功能注释标准体系。使用 iProLINK 可以获得描述蛋白质记录的

文本文献资源，在 UniProtKB 记录（生物词典）中加入蛋白质或基因命名的图谱，获得用于开发文本挖掘算法的注释数据集、挖掘蛋白质磷酸化（RLIMS-P）文献和获得蛋白质功能注释标准体系（PRO）信息。

二、MIPS 数据库

生物信息学和系统生物学研究所（institute of bioinformatics and systems biology: IBIS）是慕尼黑亥姆霍兹中心 - 德国环境卫生研究中心的一部分，主办是慕尼黑蛋白质序列信息中心（Munich information center for protein sequences: MIPS）（http://www.helmholtz-muenchen.de/en/mips）。它的重点工作是基因组信息学，特别注重基因组信息系统分析，包括应用生物信息学方法注释基因组、表达分析和蛋白质组学方面研究。MIPS 支持和维护一系列基因组数据库以及系统，可以提供细菌、真菌和植物基因组比较分析服务。在该站点提供基因组分析工具、数据库检索系统、表达分析、蛋白质相互作用等网络服务。

三、其他数据库

（一）PRINTS

PRINTS（http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php）是蛋白质基序指纹图综合数据库，每个指纹图都是使用数据扫描程序 ADSP 或 VISTAS 序列分析软件包反复优化后定义的。数据库中有两种类型指纹图，根据指纹图的复杂性分为简单和复合指纹图：简单指纹图基本上是单一的基序，而复合指纹图包含多个基序。由于识别能力的提高，数据库大部分新数据记录更适合多组合检索，检索结果更容易解释。有两个理由将数据库编排成这种排序数据库：①使 OWL 综合序列数据库中大量数据合理化（如：将序列分成家族、超家族和亚家族）；②为了提高序列分析效率。为了提高序列分析效率采用两种直接方法：①用新序列与现有数据库进行比较分析寻找结构和功能线索；②数据库记录与个别序列或个人数据库进行比较分析。这两种方法基本上是基于知识基础上的，通过预先对全部综合数据库进行检索比较，实现非常快速的判断。

基序是序列比对的保守元素，它是与已知（或未知）功能或结构区域相关的局部序列。如果局部序列保守，并且在任何其他蛋白质序列中具有相同的结构功能区域就是基序。

指纹图是用于预测类似基序的一系列基序。通过反复搜索 OWL 综合序列数据库提炼指纹图，提高它们的识别能力，使用这样排序的基序进行数据库检索是最常用的：例如在最基本的应用中，没有第二个结构信息、相似数据或任何可用于提高识别能力描述的加权算法；一个综合或多基序指纹图包含源于不同多重局部比对的、多个比对基序。在这个系统中，由于识别指纹图的单独元素是互为条件的，所以识别能力得以提高。凭借拥有指纹图的所有元素，很容易识别真正的家族成员。依据所拥有基序的差异，将蛋白质分为亚家族。

任何蛋白家族指纹图归类方法依赖于具有共同的局部序列，这些序列通常与关键功能区域或折叠成核心结构元素有关。使用 SOMAP、XALIGN 或 VISTAS 手动多序列比对获得的结果作为指纹图原始定义，由于该方法自身设计加入了与每个数据库扫描序列比对程序，只有少数需要包含初始序列比对结果。一旦一个基序或基序集已确定，保守区域会以局部序列比对形式被提取。由于不存在基序并列的相关规则，所以不能排除它们会大幅重叠。这样一个基序可能紧连着另外一个基序，基序也可能在序列比对长度范围内的任何距离被分开。单独使用某个序列比对的基序进行数据库扫描，都会产生一系列匹配序列列表，然后分析相关性，用来确定数据库中序列与指纹图元素是完全匹配还是部分匹配。只有完全匹配的序列才被认为是真正匹配。如果搜索运行良好，真正匹配的序列集将包含比原始序列比对更多的序列。然后，用新的真正匹配序列集增加的序列数据生成其他的序列比对基序集。如果被判定的家族过于庞大，再次搜索数据库，在下一次扫描前，从序列比对结果中剔除冗余基序，不断重复这一过程，直到基序不再增加为止，最终获得的比对基序集将加入到

PRINTS 数据库形成精练指纹图。质量好的指纹图在 OWL 中全都可以找到真正匹配，显示确切的切割位点。利用指纹图部分序列特征确认的子序列亚家族可能会导致切割位点难以判定，只使用指纹图中 2 至 3 个基序也会导致切割位点难以判定。但随着使用基序数量增加，分辨能力将会逐渐提高。

（二）Pfam

蛋白质一般是由一个或多个功能区域组成，这些功能区域通常称作结构域(domain)。在不同的蛋白质中结构域以不同的组合出现，形成了蛋白质的多样性。识别出现在蛋白质中的结构域对于了解蛋白质的功能有重要的意义。

Pfam 数据库(http://pfam.sanger.ac.uk/)是一个大的蛋白质域家族集合，每个家族是用多序列比对和隐马模型(HMMs)分析的结果。Pfam 家族有两个质量等级：Pfam-A 和 Pfam-B。Pfam-A 记录来源于原始序列数据库，被称作 Pfamseq，由最新 UniPrort 数据构建。每个 Pfam-A 家族包含一小部分精心筛选的家族代表成员，还包括用隐马模型谱定义检索原始序列数据库的比对序列和由精心筛选的比对序列构建的隐马模型谱和自动生成的全部序列比对序列，还包含所有检测到的属于该家族的蛋白质序列。

Pfam-B 家族是从最新公布的 ADDA 非冗余簇中自动生成的，没有注释并且质量较低。尽管 Pfam-B 家族质量较低，但 Pfam-A 记录中没有的家族，可以在 Pfam-B 家族中识别保守区域。

Pfam 记录使用下列四种方法中的一种进行分类：①家族，即相关蛋白质集合；②域，即在多种蛋白质中被发现的结构单元；③重复，即多拷贝出现可形成稳定结构的，孤立不稳定的短单元；④基序，在球形域(globular domain)以外发现的短单元。相关的 Pfam 记录用序列、结构或隐马模型谱的相似性定义被分为不同家族。

第四节 NCBI 与 EMBL-EBI

Section 4 NCBI and EMBL-EBI

一、NCBI 简介

现代分子生物学迫切需要理解活细胞自然、精美、无声的“语言”。生物基因组由四个代表 DNA 化学亚单位的字符组成的串构成，在这些字符串中表现出生命过程的语法，人类基因组是这种语法最为复杂措辞的体现。破解和使用这个“字符串”所形成的新“单词和短语”是分子生物学领域的核心和焦点。然而，惊人的分子数据量隐藏的秘密以及精细的模式迫使人们需要建立数据库和使用计算机分析工具。目前所面临的挑战是寻找新的处理海量数据和复杂性的方法。利用更好的分析方法和计算工具，能够推进人们对遗传物质和基因在健康和疾病中作用的理解。

美国前任参议员 Claude Pepper 意识到从事生物医学研究的计算机信息处理方法的重要性，于 1988 年 11 月 4 日提出建立国家生物技术信息中心(national center for biotechnology information，NCBI)立法法案，建议该中心为位于国家健康研究所(NIH)的国家医学图书馆(NLM)分支机构。选择 NLM 是由于它具有建立和维护生物医学数据库的经验，也是因为它是 NIH 的一部分，可以构建内部计算分子生物学的检索程序。

（一）NCBI 履行的职责

作为一个国家分子生物学信息资源，NCBI 的使命是开发新的信息技术，帮助理解控制健康和疾病的基本分子和遗传过程。特别是，NCBI 肩负建立存储和分析分子生物学、生物化学和遗传学知识的自动系统；提供研究和医学界使用方便的数据库和软件；努力协调搜集国内外生物技术信息；执行分析生物学重要分子结构和功能的先进研究方法。它履行的职责包括：①使用数学和计算方法在分子水平进行基本生物医学问题的研究；②与 NIH 研究所、学术界、产业界和其他政府机构保持合

作；③通过召集会议、研讨会和系列讲座促进学术交流；④通过院内研究项目支持计算生物学基础和应用研究的博士后培训；⑤通过学术访问项目吸引国际科学界成员参与信息学研究和培训；⑥开发、传播、提供和协调多样化数据库和软件以便于自然科学和医学界使用；⑦开发和推进数据库、数据存储和交换以及生物学系统命名。

NCBI 具有多学科研究团队，该团队由计算机科学家、分子生物学家、数学家、生物化学家、研究医生和致力于计算分子生物学基础和应用研究的结构生物学家组成。这些研究者不仅对基础科学做出重要贡献，而且还为应用研究活动提供新方法。他们共同利用数学和计算方法在分子水平研究基本生物医学问题，这些问题包括基因组成结构、序列分析和结构预测。典型的现行研究项目包括：检测和分析基因的组织结构、重复序列模式、蛋白域和结构元件、构建人类基因组图谱、HIV 感染动力学数学模型、数据库搜索过程中序列错误的效应分析、数据库搜索和多序列比对新算法开发、构建非冗余序列数据库、序列相似性统计学意义评估数学模型和文本检索向量模型，另外，NCBI 研究者与一些 NCBI 内部研究所以及众多学院和政府研究实验室持续保持合作。

1992 年，NCBI 承担了建立 GenBank——DNA 序列数据库的责任，受过高级分子生物学培训的 NCBI 工作人员将来自独立实验室提交的和与 EMBL、DDBJ 数据交换的序列建成数据库，与美国专利商标局协商将专利中的序列数据并入 GenBank 数据库。

除了 GenBank 外，NCBI 也为医学和学术界提供和传播多种数据库，包括在线人类孟德尔遗传(online mendelian inheritance in man，OMIM)、3D 蛋白结构分子模型数据库(molecular modeling database，MMDB)、特有人类基因序列集(UniGene)、人类基因组基因图谱、分类学浏览器和与国家癌研究所合作的癌基因组分析项目(cancer genome anatomy project，CGAP)。

Entrez 是 NCBI 的搜索和检索系统，为用户提供序列、图谱、分类学和结构数据整合的访问程序。Entrez 还提供序列和结构数据图像视图。Entrez 强大而独特的特征是能够检索相关序列、结构和参考文献。通过 PubMed，提供含有期刊索引内容的 MEDLINE 网络检索界面超过一千万以上，并为用户提供期刊文献。为了方便用户获取全文文献，还提供合作出版商网站的链接。

BLAST 是由 NCBI 开发的序列相似性搜索程序，有助于识别基因和基因特征，BLAST 可以在 15 秒内执行对全部 DNA 数据库的序列检索。由 NCBI 提供的其他软件工具包括：开放阅读框搜索(ORF Finder)、电子 PCR 和序列提交工具——Sequin 和 BankIt 等。所有 NCBI 的数据库和软件工具都可以从 HTTP 或 FTP 站点获得，NCBI 还有 E-mail 服务器提供另一种文本搜索或序列相似性搜索路径。

NCBI 通过召集会议、研讨会和系列讲座促进应用于分子生物学和遗传学计算机领域学术交流，设立学术访问项目促进与外单位科学家合作，作为 NIH 内部研究计划的一部分为博士后提供研究职位。其学术指导委员会每年召开两次会议，评审 NCBI 研究活动。

NCBI 的计算生物学部(computational biology branch，CBB)从事分子生物学和遗传学算法、数学和理论问题的基础和应用研究。包括基因组分析、序列比较、序列搜索方法学、大分子结构、动力学和相互作用，以及结构功能预测；设立合作研究项目与 NIH 内部实验室、其他政府机构、学术界和企业的生物学家、化学家、数学家和计算机科学家合作研究计算分子生物学；协商和建议政府部门和研究实验室应用研究分子生物学的计算机分析工具；在计算和理论方法研究方面与分子生物学团队合作，加强理论设计和实验研究。

计算生物学部现有的项目是由资深研究者、终身研究者、科研人员、博士后和学生在进行研究，这些项目重点研究广泛的分子生物学基础问题的理论、分析和应用方法。该团队的专家集中在病毒学统计方法、序列分析、蛋白结构功能分析和基因识别研究方面，研究兴趣跨越计算生物学和信息科学的广泛议题，简单地讲，不仅仅局限于数据库搜索算法、低复杂度序列、序列信号、进化数学模型、化学反应系统的动态特征、统计文本检索算法、蛋白结构和功能预测、比较基因组学、分类树和种群遗传学。

计算生物学部研究者进行的许多基础研究项目加强和提高了NCBI数据库和软件应用工具的公共适用性。NCBI研究者与外界研究者合作，发展了新算法(BLAST、PSI-BLAST、SEG、VAST和COGs)和新的搜索方法(文本邻近)，它改变了计算生物学的领域。目前，开发的算法和应用对于将来促进科学发现具有潜在影响。

计算生物学部成员通过对数据库中数据存储的质量和准确性方面考察，在提高NCBI在线资源的正确性和可靠性，以及用于注释数据信息的准确性方面做出明显的贡献。CBB成员还计划和组织科学公会，为大规模或高通量实验生物学提供最有效使用公共资源方法，为外界提供领导和指导。研究人员积极合作以便准确地解释已知的搜索盲区，并且找到导致这些盲区的机制。

信息工程部(information engineering branch，IEB)执行数据表述和分析的应用研究，包括分子生物学、遗传学和生物化学相关知识的数据存储、管理和计算机检索系统开发；为NIH的数据库，设计代表分子生物学信息(包括核酸、蛋白质和结构信息)不同形式的数据库模式和规格；设计和开发传播软件系统的雏形和可实际运用系统，为研究者提供局域和远程计算服务；通过组建分立和综合数据库，建立与外部数据库链接的方式，协调公众访问序列、遗传、结构和文本信息；与NIH实验室以及外部学术团队设立合作信息学研究项目；为其他政府部门和研究实验室提供先进的软件和数据库设计方法的咨询和设计；发展和促进数据库、数据交换和生物命名的标准化。

信息资源部(information resource branch，IRB)规划、指导和管理NCBI的技术操作，包括用于研究和开发使用的计算机系统以及用于访问公共数据库的计算机系统；为NCBI工作人员提供技术支持和为NCBI以外的用户提供网络服务支持；监控NCBI网络运行，与其他政府部门协调，以便所有国内外用户都可以访问NCBI服务器；组织教育示范和学术交流，培训生物医学界成员使用NCBI信息服务；规划、拟定和管理政府合同和合作协议，以便采购设备和支撑NCBI信息功能的服务器；为用户和机构之间联络，提供、参与基因组计划服务；执行应用研究和开发、提供技术咨询和指导以及明确用户需求，进行调查、评估生物学界用户使用NCBI开发软件的情况；与其他政府部门和生物信息资源协调，以便于NCBI数据库的发展。

(二) 从序列到生存：了解疾病的研究

如今，由于来自世界范围实验室的数据呈爆炸式的增长，生物学正在被改变，所面临的挑战是将数据转变成知识，这些知识将导致更好地理解潜在的健康和疾病生物学过程。对这些知识的追求促使NCBI的研究者们开发综合的、基于计算机的数据分析新方法，以便挖掘大规模和复杂的数据集。这些软件工具一旦被开发，将被研究人员用于回答具体的科学问题。

NCBI网络站点为研究和医学界传播它的资源，使世界范围的科学家有能力将看似离散的数据整合在一起，塑造更多的生物信息意义，进而产生新的知识。下面的实例可以很好地解释这种多步骤过程。

1. 从数据和信息到事实　GenBank和人类基因组：GenBank(http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html)是由NCBI维护的NIH数据库，存储了所有已知的公共DNA序列。序列数据是由全世界的单独科学家以及包括人类基因组计划在内的较大中心提交给GenBank的。存储在GenBank数据库中源自所有物种的DNA序列，目前达到相当大的数量，并且在快速、不断地增长。GenBank是与英国的欧洲信息学研究所和日本的国家遗传研究所合作伙伴合作的国际合作项目。

NCBI研究者利用存储在GenBank的人类基因草图和已测定的核酸序列数据，生成了人类基因组汇编，拼接和排列单独序列是这个项目的关键阶段，它包含许多步骤。采用新数据的更新汇编，填补了现存的空白，增加了整体的准确性，向公众定期发布。NCBI研究者还参与注释或标注人类基因组重要区域的基本过程。这个过程包括将已知人类基因准确放置在基因组中，以及从基因组序列中预测以前未知的基因。首先，主要通过序列比对的方法，将NCBI参考序列(The Reference Sequence，RefSeq)收集的mRNA放置在基因组中。NCBI研究者们首先与外界机构合作，收集多种类型信息，生成序列标准。接着，他们使用计算工具和科学判断进而来确定什么样的序列适当地代表一个基

因，进而将那些数据编排在基因组数据中。

OMIM(http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)是一个含有成千上万条基因和遗传病记录的网络目录，它还作为人类基因组的表型伴侣。OMIM 细胞遗传学和致病图谱，显示文献记载的基因在细胞遗传图谱中定位，还提供按字母顺序排列的 OMIM 中描述的所有疾病列表。

为了验证通过计算机比较分析产生的结果，必须考虑科学文献中报道的真实生物学实验结果，因此，科学数据与文献整合是建立生命科学统一信息资源的必要步骤。为此，提供了从 OMIM 到 PubMed(NCBI 的文献检索系统)的直接链接。

PubMed(http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed)提供网络访问路径，可以访问超过一千万条生物医学期刊文献的引文、摘要和索引条款，它还包含全文期刊链接。目前，每月被搜索约两千万次，每天有 14 万不同用户经由 PubMed 检索资料。

PubMed Central(PMC)，生命科学期刊文献数字文档，建立于 2001 年 1 月，为电子科学通讯和数据检索提供了新模型。PMC 的价值，除了它作为文档的作用，还在于怎样才能把不同存储资源中数据以共同的格式存储在单一的数据库中。当前，PMC 提供免费服务，可以无限制地访问 100 余种生命期刊全文，并且提供全文期刊的种类还在不断增加。

2. 从事实到知识　上述的每个数据库自身都可以提供信息并有应用价值。只有在各个部分相互链接，形成一个综合资源后，存储在每个数据库的信息才能被分析。例如，通过综合特定蛋白质相关的各种信息形式，研究者可以解释以前未知的功能，这样在复杂的生物传导通路中，前所未知的具体步骤变得清楚了。研究者们就可以在这个信息的基础上洞察某一疾病状态信号传导通路中所发生的差错。长远来看，将导致新的诊断和治疗策略的发展。

3. 疾病基因的发现　由 NCBI 开发和传播的综合资源和人类基因组计划已导致许多科学的进展。遗传性非息肉性大肠癌(HNPCC)基因的发现就是这样的例子。

HNPCC 被认为是占所有结肠癌病例的六分之一的因素，尽管大部分种类的癌表现为非遗传性的，有些个体出现某些种类癌的遗传性风险归因于单一变异基因。虽然多年来科学家已经知道 HNPCC 归咎于一个变异的基因，他们几乎找不到变异基因位于何处的线索，并且也证明要想找到它是很困难的。最后，使用人类基因组计划出现的各种工具，一个国际研究小组追踪到了该基因位于 2 号染色体上。几个月后，两个研究小组找到了致病基因。在此三个月后，研究者又识别出第二个位于 3 号染色体的基因，这个基因也与这种类型癌相关。

现在已经知道，大多数 HNPCC 病例中这两个基因一起出现变异。研究者们使用这一信息，开发了筛选这些基因变异的个体血液检查项目。在家族中检测是否出现 HNPCC 的变异基因，以便于内科医生找到基因变异家族成员，通过识别 HNPCC 风险家族成员，医生可以更密切观察患者的疾病发展征兆。家族成员被确定为非携带者就不必忍受大量的医学检查。最为重要的是，携带基因变异的患者，如果表现早期癌征兆便可及时治疗。这是真正的医学进步，因为 HNPCC 得到早期诊断和早期治疗，几乎可以治愈。

二、EMBL-EBI 简介

为了在系统水平理解生物学，大量访问不同类型数据库变得至关重要。诸如基因组测序、微阵列、蛋白质组学和结构基因组学技术已经为许多活生物提供了“零件清单”，现在研究者们正集中研究如何将独立部件组装在一起形成系统，提高每个人的生活质量。高通量技术革命产生的数据的持续增加，需要收集、存储和整理，以便于有效地检索和开发。欧洲生物信息学研究所(EMBL-EBI)，是欧洲分子生物学实验室(EMBL)的一部分，是最重要的生物信息资源，并且拥有履行重要序列分序任务的专业知识的部门之一。

EMBL-EBI 坐落在剑桥郡乡村的 55 英亩园林绿地内的威康信基因组科学园(Wellcome Trust Genome Campus)，这个园区还有威康信桑格研究所，它是世界最大基因组学和生物信息学专业技能、

知识汇集地。EMBL-EBI的前沿工作是为研究者提供公共生物学数据库。尽管它在地理上与海德堡EMBL的总部和其他的格勒诺布尔、汉堡和蒙泰罗通部门相分离，EMBL-EBI是EMBL整体的一部分。为了完成EMBL使命，EMBL-EBI提供一流的研究环境和开发新技术，为欧洲分子生命科学家提供服务和培训。

EMBL-EBI致力于向科学界各方面研究者免费提供现有数据和生物信息学服务，以利于推进科学发展；通过基础生物信息学研究者的深入研究，促进生物学发展；为所有层次科学家，从博士到单独研究者提供先进的生物信息学培训；帮助尖端技术进入产业。

EMBL-EBI起源于EMBL核酸序列数据库(现称为EMBL-Bank)，1980年成立于德国海德堡的EMBL实验室，曾经是世界上第一个核酸序列数据库。其最初目标是建立一个DNA序列中央电脑数据库，而使科学家不必将序列提交给期刊。开始，作为不太重要的从文献中摘录信息工作，很快变成了以直接电子数据提交为主的数据库活动，这就需要高技能的信息学工作人员。随着基因组计划的启动，任务规模不断增加，并且数据的增加与商业部门的研究有关。明显表明，EMBL核酸序列数据库需要更好的经济保障，才能确保长期生存，以适应规模如此庞大的任务。

为了提供服务，还需要与全球同伴合作研究、开发和支持为产业提供协助的项目。为此，1992年EMBL理事会投票决定建立欧洲生物信息学研究所，将它定位在英国威康信基因组科学园，邻近主要以测序为主的桑格研究所。从1992年到1995年，海德堡的作用逐渐过渡，直到1995年9月EMBL-EBI拥有了现在的威康信基因组科学园。

当EMBL-EBI迁到辛科斯顿时，拥有了两个数据库：一个核酸序列数据库(EMBL数据库，现已知的EMBL-Bank)和一个蛋白质序列库(Swiss-Prot-TrEMBL，现已知的UniProt)。以后EMBL-EBI帮助引导了生物信息学革命：以多种形式提供所有主要的分子领域数据资源、拓展了广泛的搜索基础、开发了支持用户的独特方法以及提供了先进的生物信息学培训。

欧洲始终处于生物信息学研究的前沿，但由于EMBL-EBI转向了欧共体的单一“欧洲研究领域”的目标，就更加需要全欧洲生物信息学专家和实验生物学家通力合作，向着共同目标推进生物学研究。出于这一目的，EMBL-EBI与三个欧共体资助的卓越网站合作。

1. BioSapiens(http://www.biosapiens.info/page.php?page=home) 旨在通过建立基因组注释的虚拟研究所，解决欧洲生物信息学过度分散和组建基因组注释培训的欧洲学校。

2. EMBRACE(http://www.embracegrid.info/page.php?page=home) 旨在使生物信息学资源的访问标准化，向数据提供者提供友好界面，将数据提交到标准数据库。这将允许用户生成最离散的数据资源。

3. ENFIN(http://www.enfin.org/)旨在带领实验学家和计算生物学家一起为系统生物学开发下一代生物信息学资源。

基于20多年生物信息学经验，EMBL-EBI维护世界上最广泛的分子数据库。EMBL-EBI是在全球范围内，协调搜集和传播生物学数据的欧洲节点，EMBL-EBI的许多数据库是生物学家们熟知的，包括：EMBL-Bank(DNA和RNA序列)、Ensemble(基因组)、ArrayExpress(基于微阵列的基因表达数据)、UniProt(蛋白质序列)、InterPro(蛋白家族、域和基序)、Reactome(传导通路)和ChEBI(小分子)，新的资源帮助研究者不仅了解构成生物体的分子部件，还了解这些部件是如何组合构成系统的。每个数据库细节可能变化，但它们都遵循相同的服务规则。包括：①可访问性：EMBL-EBI是向社会提供生物学数据的保管者(不是占有者)，并且生物学研究的发展依赖于完全开放地访问这些数据。因而，研究界可以免费获得所有的数据和工具，不加任何限制。②兼容性：EMBL-EBI比世界上任何其他组织都努力去促进生物信息学标准的采用。这些标准的发展推动数据共享。③数据集综合性：现存一系列公开的数据库，EMBL-EBI已通过谈判得到数据共享协议，以保证资源包含综合的和最新数据集。EMBL-EBI还与出版商谈判以保证生物学数据存放在公共数据库作为发表过程的一部分，并与相关出版物交叉引用。④便携性：EMBL-EBI所有的数据库都可以下载获得，在许多情

况下，全部软件系统可以下载，然后在本地安装。⑤保证质量：EMBL-EBI 的数据库通过注释得到增强：存储数据库中的基因或蛋白质的特征是从其他资源定义和解释中提取的，许多是高素质生物学家注释的，EMBL-EBI 做的自动注释是经过严格的质量控制的。

EMBL-EBI 的服务团队使这些资源可以在线获得，还为用户提供帮助。EMBL-EBI 向所有用户保证，他们负责每天超过 250 万人可以访问该网站，EMBL-EBI 资源的主体能满足用户需要的所有信息，不同用户可以选择不同路径进入数据库：①对于生物信息学新手，可以查看 2Can(http://www.ebi.ac.uk/2can/home.html)站点，了解为什么和如何做；②如果不知道哪个服务器能满足你的需要，服务器站点图(http://www.ebi.ac.uk/services/)将帮助你找到正确的资源；③如果需要了解如何获得你需要使用的大部分服务内容，请看帮助网页(http://www.ebi.ac.uk/help/)；④如果在帮助网页找不到你需要的信息，请给 EMBL-EBI 发送 E-mail，EMBL-EBI 将在两个工作日内给您答复。

EMBL-EBI 所有主要资源都是国际合作的产物，EMBL-EBI 与数据提供者以及合作者共同工作，确保 EMBL-EBI 数据库的广泛性和新颖性。EMBL-EBI 积极努力参与制定国际数据标准，例如：EMBL-EBI 的微阵列信息学团队是推动开发 MIAME 标准的推动力量。该标准制定了能清楚描述微阵列实验信息的最少条目。EMBL-EBI 蛋白质组学服务团队协调制定 HUPO's 蛋白质组学标准方案，它是为蛋白质 - 蛋白质相互作用、质谱和普通蛋白质组学开发的数据标准。EMBL-EBI 的计算神经生物学团队是开发生物学建模系统的主干团队，还为开发系统生物学标记语言作出了贡献。EMBL-EBI 积极主动举办欧洲和欧洲以外的计算生物学家活动，提供更多的服务。

EMBL-EBI 为生物信息学研究提供独特环境，对于研究数据库互补资源具有广泛的兴趣。作为 EMBL 整体组成的一部分，是美国以外最高产的研究所，EMBL-EBI 是所有 EMBL 四个中心的最活跃成员，促进了 EMBL 的五个分部之间跨学科研究。

EMBL-EBI 研究团队旨在通过开发新方法解释生物学数据，了解生物学。研究领域包括(括号内为主要研究小组)：进化途径的基因组分析(Paul Bertone)；序列数据进化分析(Nick Goldman)；神经信号计算系统生物学(Nicolas Le Novere)；蛋白质组学：结构、功能和进化(Janet Thornton)；基因组规模调节系统分析(Nick Luscombe)和功能基因组学(Wolfgang Hubert)。

服务团队开发和维护数据资源，还执行开发新服务和强化现有服务的研究。服务团队现研究包括：启动子发现(Ewan Birney)；微阵列数据分析和算法开发(Alvis Brazma)；网络和电子学项目(Peter Rice)和蛋白质序列数据自动注释(Rolf Apweiler)等。

生物信息学是涉及生物学所有纯粹和应用领域的快速扩张学科，EMBL-EBI 为全世界生物信息学家培训工作起了很重要的作用。

三、通过 Entrez Gene 从 NCBI 获取序列信息

Entrez 主要是用于 NCBI 数据库综合的、基于文本的搜索和检索系统。Entrez 综合了科学文献、DNA 和蛋白质序列数据、3D 蛋白质结构和蛋白质域数据、种群研究数据集、表达数据、完整基因组组装和分类学信息，形成一个紧密链接的系统。它是用于搜索 NCBI 链接数据库的检索系统。图 1-2 显示了 Entrez 检索系统所有子数据库，可以执行单个子数据库检索，还可以进行跨库检索。

(一) Entrez Gene 检索

Entrez Gene 检索到的记录提供关键链接，将图谱、序列、表达、结构、功能、索引文献和同源数据链接在一起构成关键链接。用定义序列、已知的图谱定位和从表型信息推测的基因，为基因分配特有标识符。这些标识符在 NCBI 的数据库中通用，可以用于注释更新跟踪和相关信息跟踪。Entrez Gene 用 NCBI 参考序列(RefSeqs)覆盖了基因组，还被整合到 NCBI 的 Entrez 和 E-Utilities 系统的索引、查询和检索中。

检索 Entrez Gene 最简捷的方法是登录到 NCBI(http://www.ncbi.nlm.nih.gov/)的首页，在检索窗口的选择数据库的下拉菜单中选择 Gene 选项，请见图 1-3。

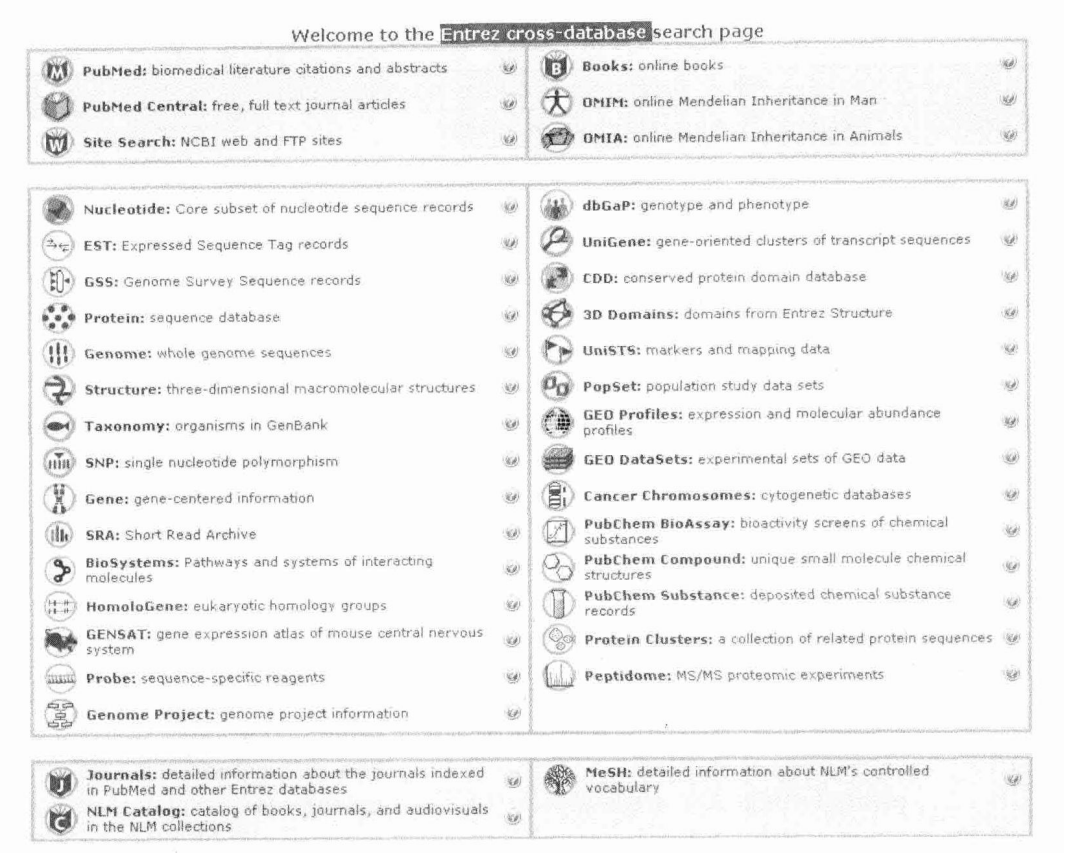


图 1-2 Entrez 检索系统子数据库

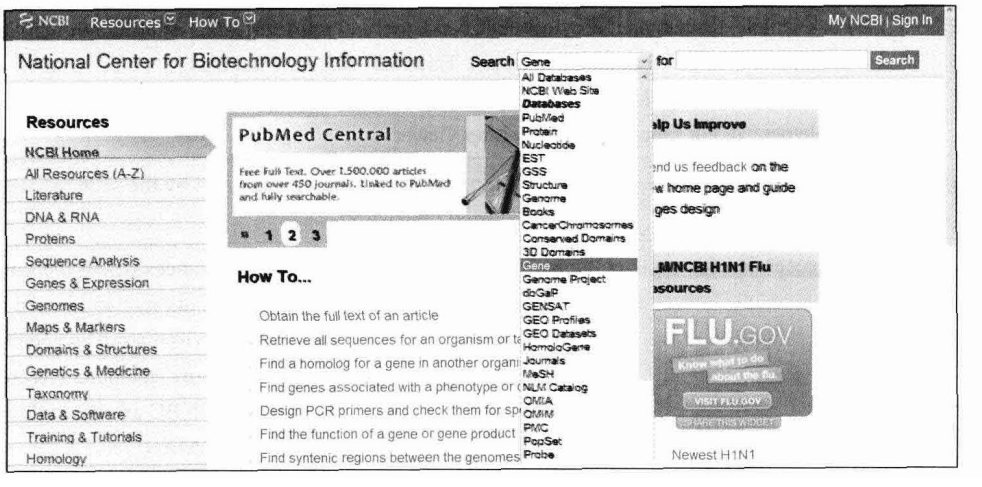


图 1-3 NCBI 检索首页检索窗口的数据库选项下拉菜单

然后，在检索栏输入检索词或词组(图 1-4)。

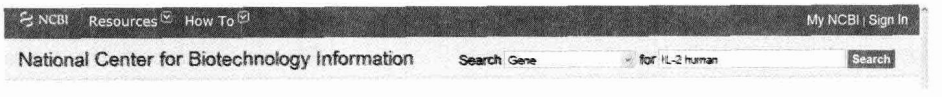


图 1-4 检索栏(for)输入“IL-2 human” —检索人类白介素 2 在 Gene 数据库中的记录

为了准确、快速地在 Gene 数据库中检索到需要的记录，应按表 1-2 中列出的分类信息条目内容输入检索栏，输入的检索词超过两个以上还可以运用布尔逻辑运算检索。

表 1-2 在 Gene 数据库记录中使用的分类信息概述表

Fields, Filters, and Properties in Entrez Gene		
Field name	Definition [including field abbreviations]	Examples
Name subcategory		
Disease name or phenotype of mutants	Disease or phenotype associated with the record [DIS]	Find the genes that contribute to SCID SCID[dis]
Gene name	A symbol for the gene. Includes preferred symbols, aliases, and locus tags [SYM][SYMB][GN][GENE NAME]	Genes with a symbol starting with smt smt*[sym]
Gene/protein name	The short or full name of the gene or any of its protein products (when applicable)	Find genes that have the word kinase in GO annotation but do not have the word kinase in the name kinase[gene ontology] NOT kinase[gene/protein name]
Location subcategory		
Chromosome	Chromosome location of the gene. The value used is according to the convention of the source genome. In other words, if Ⅲ is used, Ⅲ but not 3 will be indexed in this field [CHRM][CHR][CHROMOSOME]	Retrieve records containing the word kinase and the gene is location on chromosome Ⅲ: kinase AND Ⅲ [chr] Retrieve records containing the words zinc and finger that are of human origin but not on chromosome 19: zinc finger NOT 19[chr] AND "Homo sapiens"[orgn]
Default map location	A map location in the units standard for the genome. For example, for human it is cytogenetic band, for mouse it is the MGI map (centiMorgans). This is processed as a text field, so range queries are not implemented. For range queries, use Map Viewer	Rat genes mapped to 18 q: rat[orgn] AND 18q[default map location]
Sequence subcategory In Gene, this means searching by sequence identifier, not by the sequence itself, which is managed by BLAST		
Nucleotide accession	An accession for a nucleotide sequence [NACC]	There are instances where the same accession is applied to both nucleotide and protein sequences. So to restrict an accession to nucleoide, use this field. (Accessions beginning with BC are not in this category) BC052629[NACC]
Nucleotide UID	The gi of a nucleotide sequence[NUID][NUCL_UID][NUCLEOTIDE_UID]	Many integer identifiers have overlapping number spaces, So to find the gene record that corresponds to a given nucleotide gi from gene, use this field 27363473[NUID]
Protein accession	An accession for a protein sequence [PACC][PROT_ACCN]	There are instances where the same accession is applied to both nucleotide and protein sequences. So to restrict an accession to protein, use this field. (Accessions beginning with three letters are not in this category) AAH52629[PACC]

续表

Fields, Filters, and Properties in Entrez Gene		
Field name	Definition [including field abbreviations]	Examples
Protein UID	Rrotein gi [PUID][PROT_UID][PROTEIN UID]	Many integer identifiers have overlapping number spaces, So to find the gene record that corresponds to a given protein gi from gene, use this field 27363473[PUID]
Nucleotide or Protein accession	A sequence accession of any type [ACCN]	Find all the genes encoded in accession AE003828 AE003828
Miscellaneous subcategory (alphabetical)		
Creation date	Date the record was created [cd][cdat][creation date]	Records containing the word xenopus created between February 5, 2004 and February 12, 2004: 2004/2/5:2004/2/12[cd] AND xenopus[orgn]
EC/RN number	Enzyme commission identifier for a product of the gene. Indexed without the EC prefix [ECNO][EC]	Retrieve records where proteins have an E.C. number of 1.9.3.1: 1.9.3.1[ECNO]
Filter	Find records with a relationship to other data in Gene. For more examples of use of filters, see the preview/index section	Retrieve records of mouse kinase genes with expression data stored in GEO mouse[orgn] AND gene_geo[filter] AND kinase
Gene Ontology	GO terms applied to this gene AND the GO identifer as the integer. The terms include the component, function, and process categories [GO][GENE ONTOLOGY]	Rat genes with GO terms starting with 'kinase signaling' kinase signaling*[gene ontology] rat[orgn] Any gene with the GO id of GO:0004872 4872[GO]
LocusLink ID	The gene identifier from LocusLink [LID][LOCUS_ID]	Retrieve the record where LocusID=2: 2[LID]
MIM	Identifier assigned to human genes and phenotypes by OMIM [MIM]	Retrieve records that contain the MIM number 181510: 181510[MIM]
Modification date	Last date the record was modified.[MODDATE] [MDAT][LMOD][DATE][UPDATED][MD]	Retrieve records for genes from eubacterial genomes last modified after March 10, 2004: eubacteria[orgn] AND 2004/3/10:2010/1/1[md] Retrieve records from sea urchins modified in the last 30 days: echinoidea[orgn] AND "last 30 days"[mdat]
Property	An attribute of a Gene record based on its content [prop][property]	Mouse records with transcript variants: mouse[orgn] AND "has transcript variants"[property]
PubMed UID	PubMed ID [PMID]	Many integer identifiers have overlapping number spaces, So to find the gene record(s) that corresponds a paper in PubMed from gene, use this field 12477932[PMID]

续表

Fields, Filters, and Properties in Entrez Gene		
Field name	Definition [including field abbreviations]	Examples
Taxonomy ID	Identifier for the species or strain in the NCBI taxonomy database[TAXID][TID] HINT: txid{value} also works, e.g. txid9606	Find all records in Entrez Gene for the pig: 9823[taxid] alternatively, txid9823
Text Word	Any word in the record [TEXT][WORD][AB][TXT]	Retrieve records that contain '32' in a record that also contains threonine, serine and kinase: serine AND threonine AND kinase AND 32[TEXT]
UniGene cluster number	UniGene cluster including the text prefix [UNIGENE] [UGEN]	Hs.2[UNIGENE]

（二）Entrez Gene 记录显示格式

当进行检索时，检索结果以摘要(summary)格式显示，每页可显示多条记录，摘要显示的每条记录前有一个选择框，可以选择需要显示的记录。显示内容还包括首选名称标志、完整全名、双单词的物种名称(在方括号中)、基因组定位和基因编号。如果基因在已命名的质粒上，那么作为基因定位将给出质粒名称。右侧的 Links 可以关联到显示相关的 Entrez 记录，选择灰色查询条目中的显示选项，可以显示相关记录。

点击选中记录蓝色的名称标志(图 1-5)，进入记录全文报告页面，该页面包括图像和文本显示关于基因已知的信息。通过页面顶部右侧的“Links”目录可以访问 NCBI 其他资源库以及 NCBI 以外资源库的附加信息。全文报告的内容包括：标题、导航菜单、概述、基因组区域、转录物、产物、基因组定位、参考资料、相互作用、等位基因、一般基因信息、一般蛋白质信息、NCBI 参考序列、相关序列和附加链接(图 1-6)。

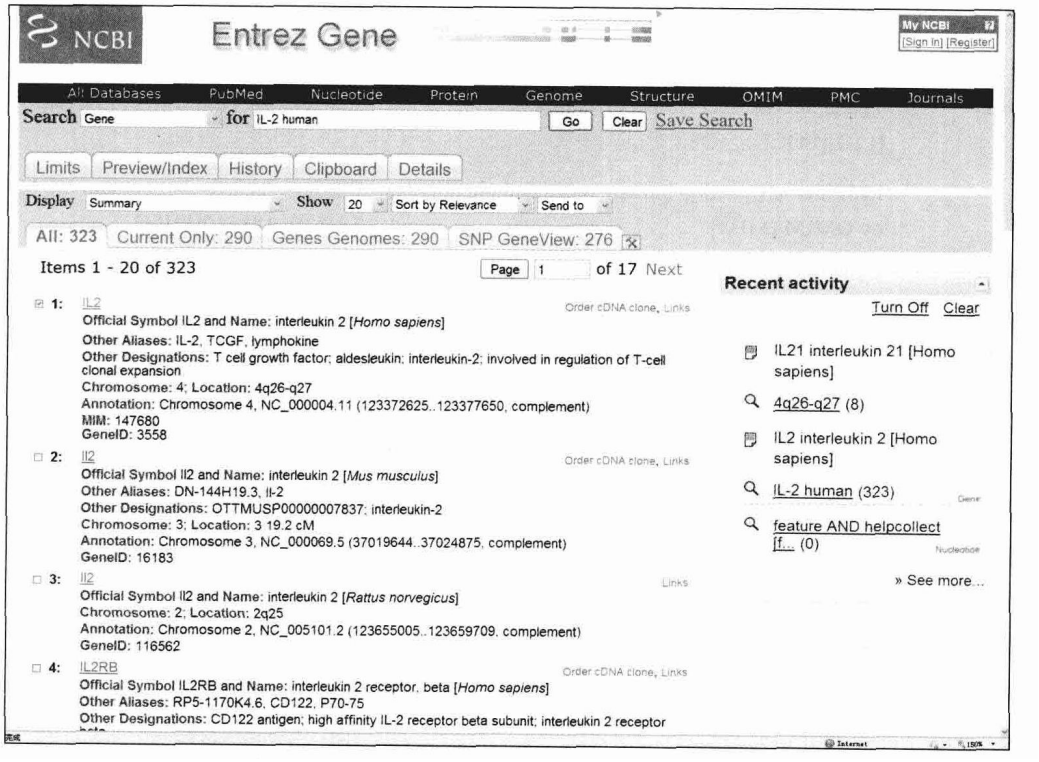


图 1-5 Entrez Gene 检索结果摘要格式显示页面

在全文报告的“Genomic regions, transcripts, and products”部分(图 1-6)的基因模式图,提供了基因组 RefSeq 数据库中的已知基因注释的内含子 / 外显子 / 编码区信息。你可以使用这个模式图查看基因的内含子 / 外显子 / 编码区组成和 RNA 产物,或者将虚拟基因放在基因组参考序列中;确定与任何RNA或蛋白质产物相对应的参考序列和外显子概貌;点击参考序列编号可以浏览基因组、RNA 或蛋白质序列;点击蛋白质编号浏览 Blink 数据库检索蛋白质序列与其他物种相关蛋白质的比较。

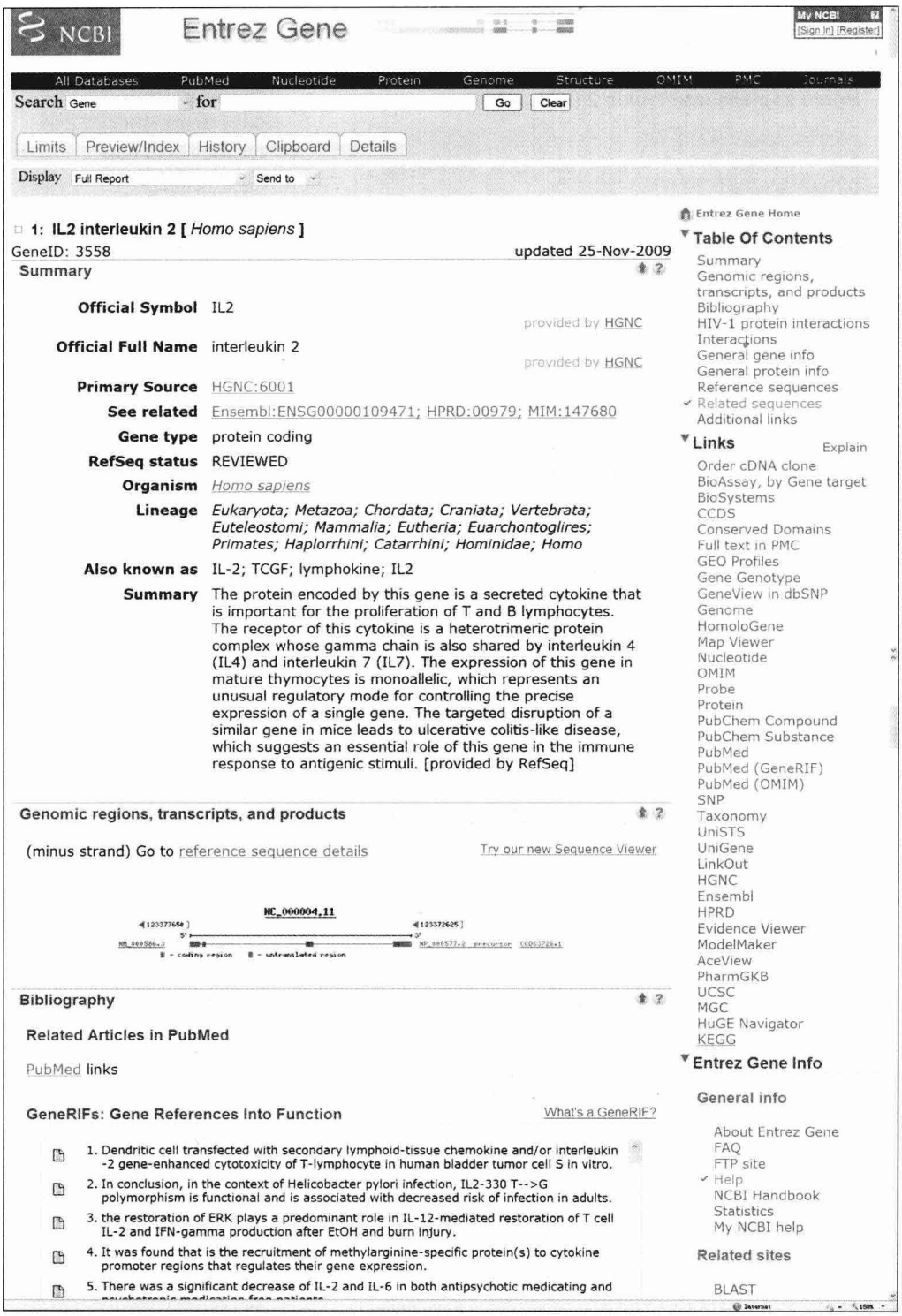


图 1-6 Entrez Gene 全文报告页面

基因模式图的上部、左侧和右侧都显示数据库编号，点击后分别与 Nucleotide、RNA 和 Protein 数据库链接，可以获得相应的序列数据。以图 1-6 中的左侧编码 NM_000586.3 为例，点击后选择 GenBank 格式，显示(图 1-7)Nucleotide 数据库中的 IL2 mRNA 记录。

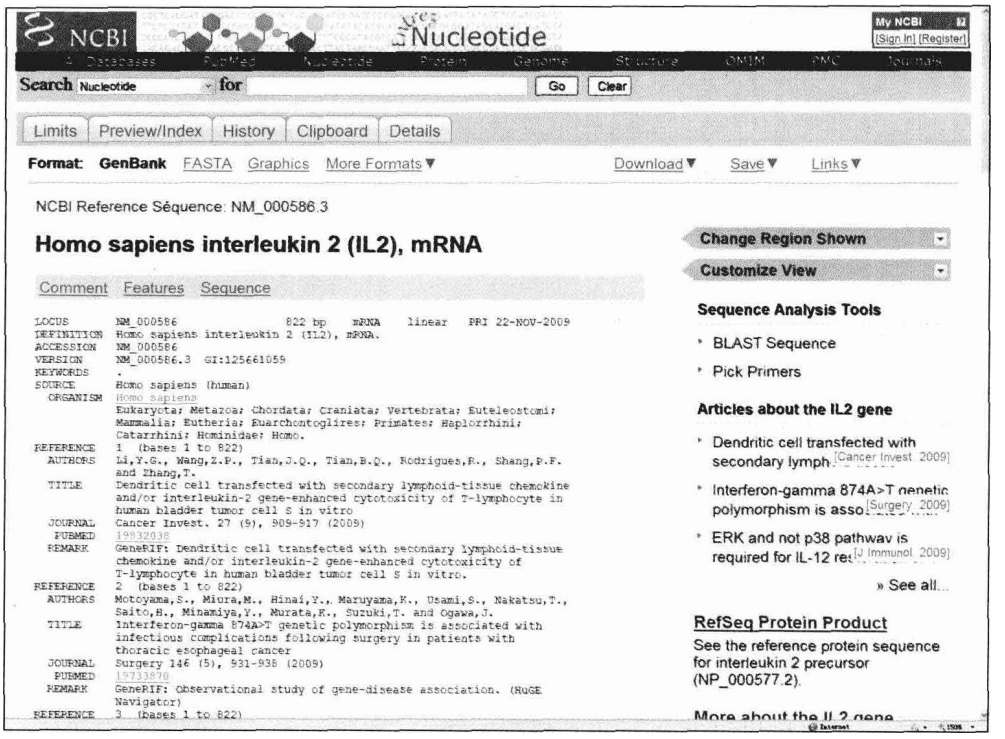


图 1-7 IL2 mRNA 的 Nucleotide 数据库记录检索结果显示界面

(三) Nucleotide 数据库记录显示格式

Nucleotide 数据库记录的显示结果可以用 GenBank、FASTA、Graphics 和 ASN.1 格式显示。序列信息通常用 FASTA 和 GenBank 两种格式显示，FASTA 格式仅包括该序列的简要特征，列出核苷酸序列。GenBank 格式可显示较完整的基因序列记录，反映核苷酸序列的详细信息，其显示字段名及含义见表 1-3。

表 1-3 Nucleotide 数据库中 GenBank 显示格式记录字段注释

含义	字段名	显示信息
基因座位	LOCUS	NM_000586 822 bp mRNA linear PRI 17-JAN-2010
基因定义	DEFINITION	Homo sapiens interleukin 2 (IL2), mRNA
记录编号	ACCESSION	NM_000586
版本	VERSION	NM_000586.3 GI:125661059
关键词	KEYWORDS	
来源物种	SOURCE	Homo sapiens (human)
物种分类	ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
参考文献 1	REFERENCE	1 (bases 1 to 822)
文献作者	AUTHORS	Li,Y.G., Wang,Z.P., Tian,J.Q., Tian,B.Q., Rodrigues,R., Shang,P.F.and Zhang,T.
文献标题	TITLE	Dendritic cell transfected with secondary lymphoid-tissue chemokine and/or interleukin-2 gene-enhanced cytotoxicity of T-lymphocyte in human bladder tumor cell S in vitro

续表

含义	字段名	显示信息
文献出处	JOURNAL	Cancer Invest. 27 (9), 909-917 (2009)
MEDLINE 编号	PUBMED	19832038
评论	REMARK	GeneRIF: Dendritic cell transfected with secondary lymphoid-tissue chemokine and/or interleukin-2 gene-enhanced cytotoxicity of T-lymphocyte in human bladder tumor cell S in vitro
注释	COMMENT	REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from X01586.1 and BC070338.1. On Feb 14, 2007 this sequence version replaced gi:28178860 ……
原始记录	PRIMARY	REFSEQ_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP 1-792 X01586.1 53-844 793-822 BC070338.1 785-814
序列特征	FEATURES	Location/Qualifiers
来源	source	1..822 /organism="Homo sapiens"……
基因	gene	1..822 /gene="IL2"……
外显子	exon	1..202……
编码区序列	CDS	56..517 /gene="IL2"…… /translation="MYRMQLLSCIALSLALVTNSAPTSSSTKKTQLQLEHLLLDLQMIL NGINNYKNPKLTRMLTFKFYMPKKATELKHLQCLEEELKPLEEVLNLAQSKNF HLRPRDLISNINVIVLELKGSETTFMCEYADETATIVEFLNRWITFCQSIISTLT"
序列标签	STS	69..509 /gene="IL2"……
polyA 信号	polyA_signal	779..784 /gene="IL2"……
polyA 位点	polyA_site	799 /gene="IL2"……
序列源	ORIGIN	1 agttccctat cactctcttt aatcactact cacagtaacc tcaactcctg ccacaatgta …… 781 taaatttgat aaatataaaa aaaaaaaaaa aaaaaaaaaa aa
结束符	//	

在一些记录的序列特征(Features)标题下，经常还会出现一些具体的副标题，反映序列的各种其他信息，这些副标题的含义见表 1-4。

表 1-4 Nucleotide 数据库 GenBank 显示格式序列特征副标题含义

副标题	含义	副标题	含义
allele	等位基因	exon	外显子
attenuator	弱化子	GC_signal	真核启动子的 GC- 信号
CAAT-signal	真核启动子的 CAAT- 信号	intron	内含子
CDS	cDNA	LTR	长末端重复序列
conflict	不同测序的差异	Mat-Peptide	编码成熟肽的顺序
enhancer	增强子	mRNA	信使 RNA

续表

副标题	含义	副标题	含义
mutation	突变位点	satellite	卫星重复序列
polyA_site	mRNA 的 polyA 位置	sig_peptide	编码信号肽的序列
precursor-RNA	前体 RNA	snRNA	小核 RNA
prim_transcript	初始转录物	TATA_signal	真核启动子的 TATA- 信号
primer	PCR 引物	terminator	转录终止序列
promoter	启动子	tRNA	转运 RNA
protein_bind	蛋白质结合区	unsure	不能确定的区域
provirus	原病毒序列	-10_signal	原核启动子 Pribow- 信号
RBS	核糖体结合位点	-35_signal	原核启动子的 -35- 信号
rep_origin	双链 DNA 复制起始区	3'UTR	3' 非翻译区
repeat_region	包含重复子序列的区域	5'UTR	5' 非翻译区
rRNA	核糖体 RNA		

四、通过 SRS 从 EBI 中获取蛋白质序列信息

SRS 是世界上主要的生物信息学、基因组和相关数据整合、分析和显示工具。SRS 检索系统是个开放的系统，可以根据不同的需要安装不同的数据库，现在，安装在 EBI 的数据库有 300 多个。SRS 有三种检索方式：快速检索、标准检索和批量检索。

可以通过网址(http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+quickSearch+-id+76e2D1aC0Ri)进入 SRS 开始页面(图 1-8)。在这个页面中，可以开始一个永久项目，在该项目中，允许用户在 SRS 系统中安装用户自己的相关数据库。该页面就是快速检索页面，在页面上方有九个快捷方式按钮，分别是 Quick Search、Library Page、Query Form、Tools、Results、Projects、Views、DataBanks 和 HELP。

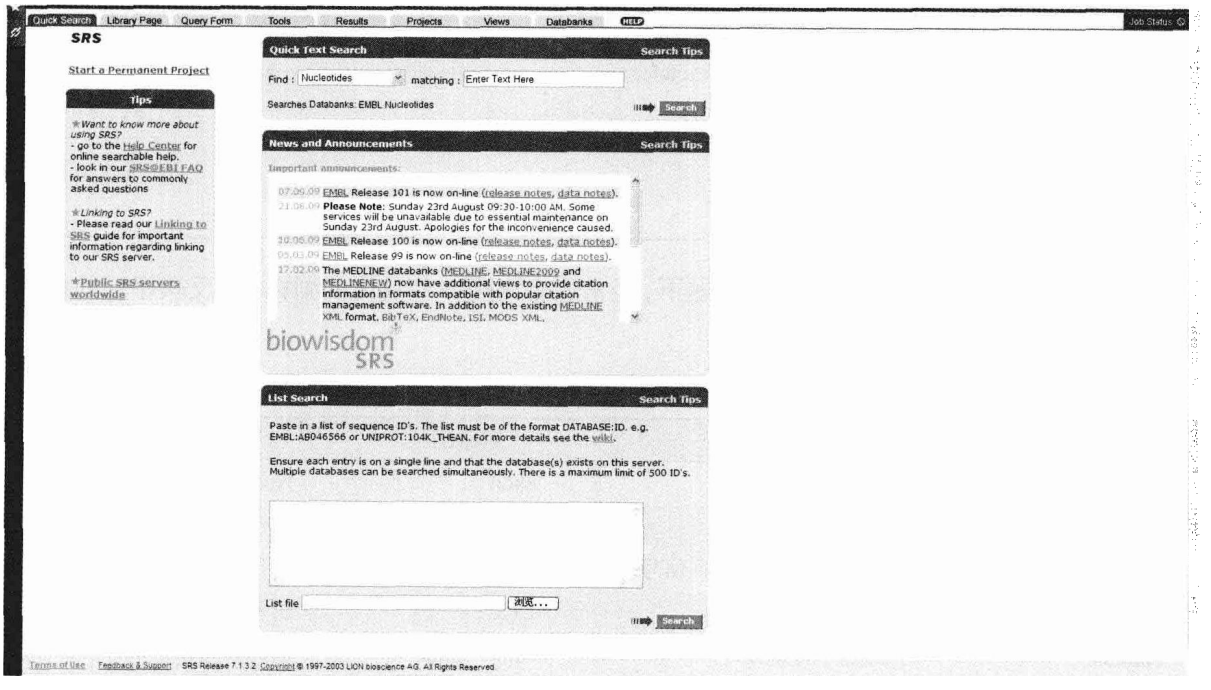


图 1-8 SRS 开始页面

Quick Search 的功能是进入快速检索页面，Library Page 显示备选数据库，Query Format 进入标准检索页面，Tools、Results 显示备选分析工具，Projects 显示已有的用户项目，Views 选择检索结果和分析结果的显示方式。DataBanks 列出所有 SRS 系统安装的数据库，HELP 显示为用户提供的帮助信息。在页面下部的“List Search”窗口允许用户进行批量检索。

当用户开始检索或选择数据库时，服务器将自动为用户生成一个临时项目。在开始页面或点击 Quick 快捷方式按钮，就可以快速检索 EBI 数据库操作，在快速文本检索窗口(图 1-9)的“Find”栏中选择“Protein”选项，在“matching”栏中输入检索词或词组，如输入“KRAS human”。点击“Search”按钮运行检索，然后显示检索结果页面(图 1-10)。

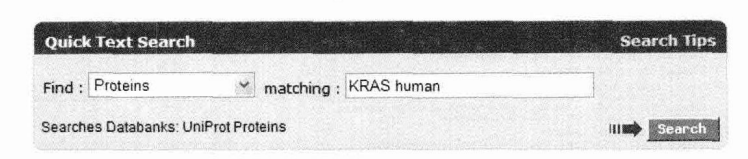


图 1-9 快速文本检索窗口

图 1-10 检索结果页面显示的检索结果

如果用户想进一步查看详细完整的记录内容，可以点击超链接查看。如果想要查询 *KRAS* 基因的产物，在检索结果页面的 GeneName 栏中，找到 *KRAS* 对应的蛋白质记录，通过超链接可检索到查询蛋白质的详细记录(B0LPF9_HUMAN)页面(图 1-11)。

在详细记录页面显示了蛋白质的一般信息、蛋白质来源和描述信息、参考文献信息、交叉链接、关键词和序列信息。点击“Sequence”按钮，查看蛋白质序列(图 1-12)。

由于快速检索是在 SRS 系统的所有数据库中检索，检索到的记录比较多，很多记录不是用户需要查询的内容，因此，可以使用标准检索，可以较快检索到用户需要的记录。点击“Query Form”按钮，进入标准检索页面(图 1-13)，在开始标准检索前，用户必须先点击“Library Page”按钮，选择需要查询的数据库。然后，从检索字段选项的下拉菜单中选择检索范围，在它右侧的文本输入栏输入

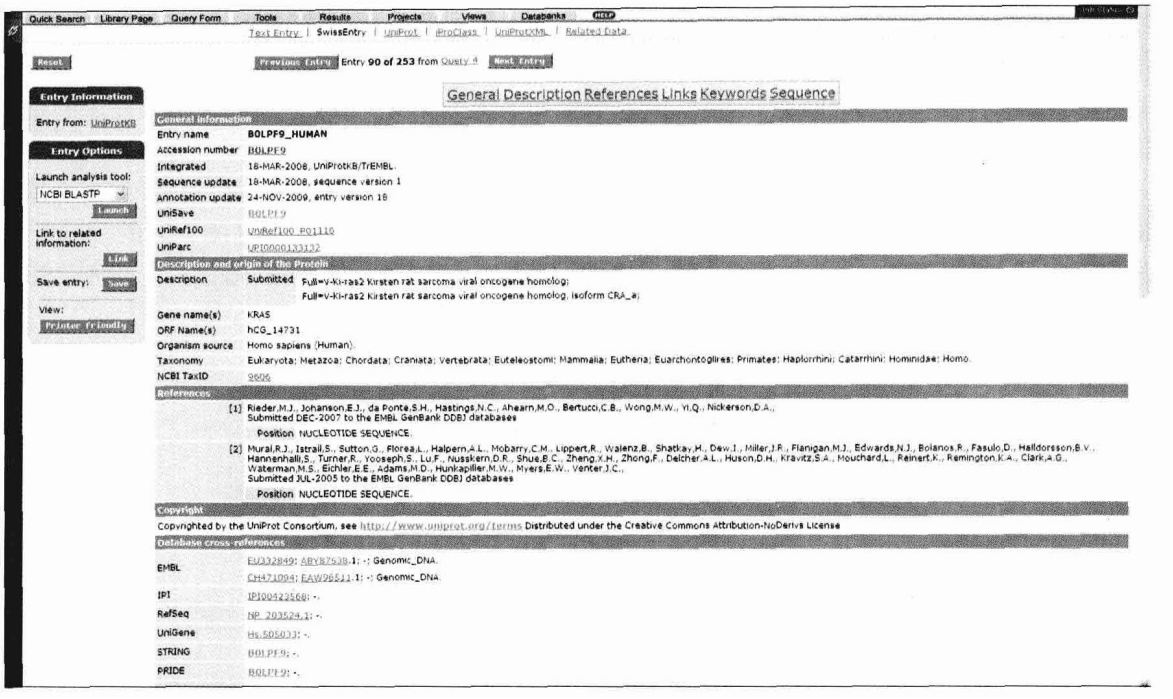


图 1-11 蛋白质记录详细内容页面

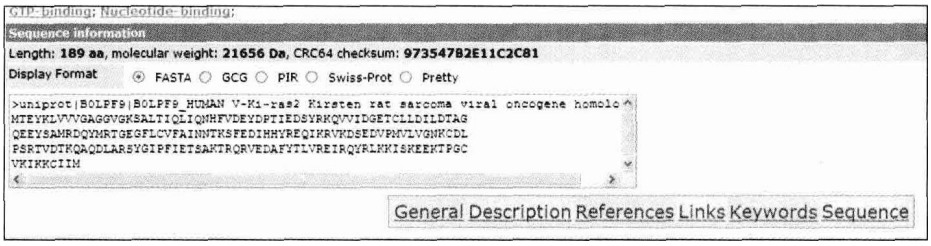


图 1-12 蛋白质序列显示窗口

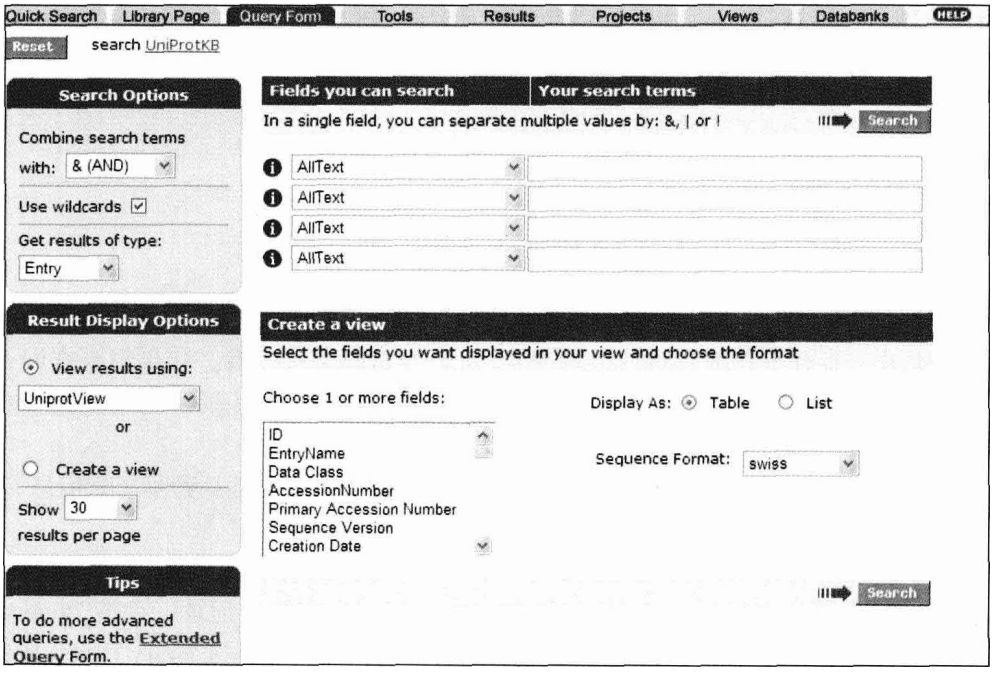


图 1-13 SRS 标准检索页面

检索词。如图 1-14 所示，表明用户想在数据库所有记录中的基因名称字段中，检索包含 KRAS 的记录。点击“Search”按钮，开始检索，显示检索结果页面(图 1-15)。通过超链接可以获得用户选中的查询蛋白质详细记录。

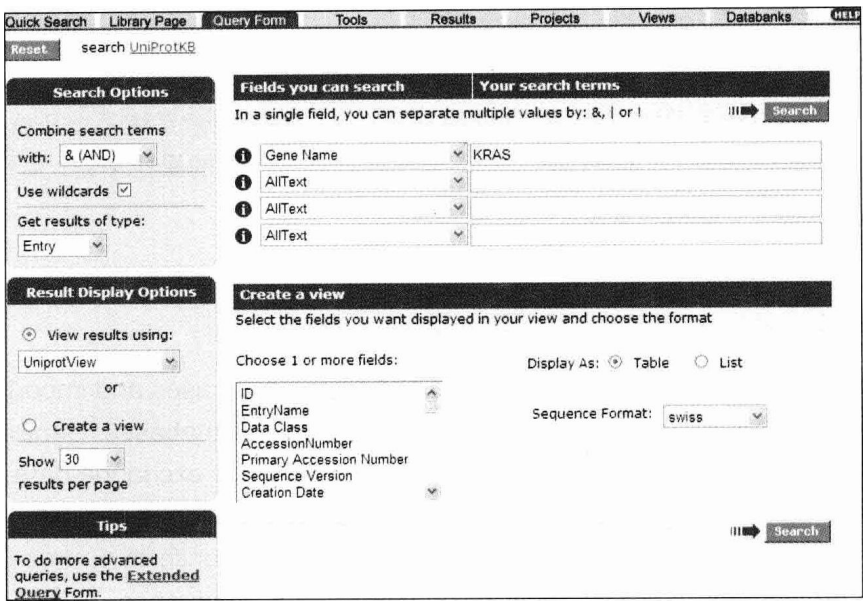


图 1-14 SRS 标准检索页面检索基因名为“KRAS”蛋白质序列输入示意图

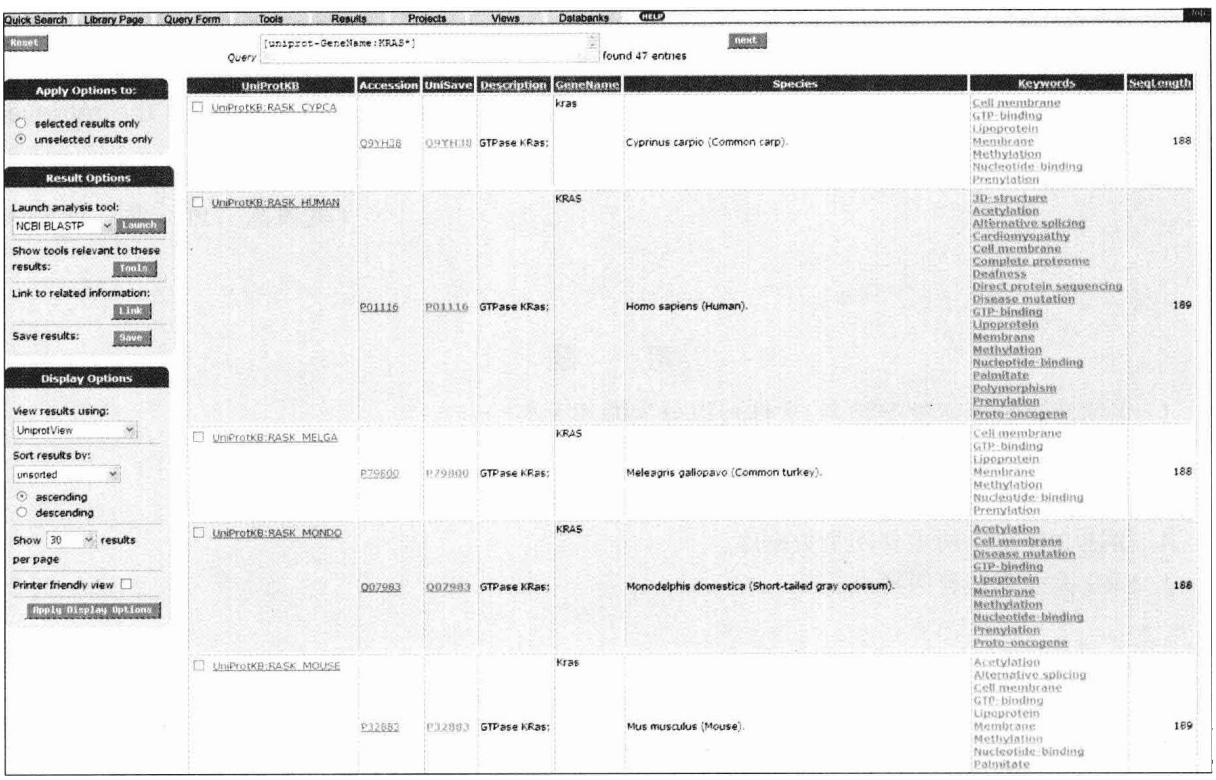


图 1-15 SRS 标准检索结果输出页面

SRS 系统允许用户保存检索结果，以备后期使用。点击“Results”按钮，进入处理检索结果页面。在该页面中选中需要保存的检索结果，点击“Save”按钮，可以保存检索结果到用户的计算机。

小 结

本章介绍了生物信息学常用的数据库和重要网站，重点介绍了三大核酸数据库：GenBank 数据库、EMBL 数据库和 DDBJ 数据库。虽然这三个核酸数据库的网站分别处在不同的地理位置，为了保证数据库内容在全世界范围的同步性，每天这三个数据库进行数据交换。因此，在这三个数据库中检索的数据几乎是一样的。NCBI 的 Entrez Gene 将分类、基因组、图谱、序列、表达、结构、功能、索引文献和同源数据链接在一起，为用户提供了便捷的检索方式。EBI 的 SRS 检索系统是世界上主要的生物信息学、基因组和相关数据整合、分析和显示工具。SRS 检索系统是个开放的系统，可以根据用户不同的需要安装不同的数据库，便于用户开发具有自己特性的操作平台，尤其在数据分析方面，对于检索到的信息可以进行多种方式的分析处理。

Summary

We have introduced the commonly useful bioinformatics databases and important websites in this chapter. We focus on the three major Nucleotide Sequence Databases: GenBank, EMBL and DDBJ. Althouth they distribute in different places in the world, they exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. The data that you retrieve from any of the three databases are nearly uniform. Entrez Gene has been implemented to supply key connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. It provides a nimble retrieval way for users. EBI SRS is the world's premier data integration, analysis and display tool for bioinformatics, genomic and related data. SRS retrieval system is an open system which can be installed different databases according to users special needs. It faciliates the users to develop operation platform with own specific characteristics. Especially on data analysis, the system can process the retrieved information in many ways.

（赵雨杰 何 群 钟连声）

习 题

1. 生物信息数据库种类繁多，大体可以分为哪四大类？
2. 何为生物信息二次数据库？
3. 向 GenBank 提交序列数据软件有几种？各有何特色？
4. 简述 PIR 提供的主要四个数据库。
5. SRS 检索系统有何特点？

主要参考文献

1. Dennis A. B., Ilene K. M., David J. L., et al. GenBank. Nucleic Acids Research, 2008, 36(Database issue): D25-30.
2. Cathy H. W., Lai-Su L. The Protein Information Resource. Nucleic Acids Research, 2003, 31(1):345-347.
3. Hamm G. H., Cameron G. N. The EMBL data library. Nucleic Acids Research, 1986, 14(1):5-9.
4. Sugawara H., Ogasawara O., Okubo K., et al. DDBJ with new system and face. Nucleic Acids Research, 2008, 36(Database issue):D22-24.

5. Lee B., Shin G. CleanEST: a database of cleansed EST libraries. Nucleic Acids Research, 2009, 37(Database issue): D686-689.
6. Szymanski M., Erdmann V. A., Barciszewski J. Noncoding RNAs database (ncRNAdb). Nucleic Acids Research, 2007, 35(Database issue):D162-164.
7. Griffiths-Jones S., Saini H. K., van Dongen S., et al. miRBase: tools for microRNA genomics. Nucleic Acids Research, 2008, 36(Database issue):D154-158.
8. Wu C. H., Huang H., Arminski L., et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. Nucleic Acids Research, 2002, 30(1):35-37.
9. Mewes H. W., Dietmann S., Frishman D., et al. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Research, 2008, 36(Database issue):D196-201.
10. Attwood T. K., Beck M. E., Bleasby A. J., et al. PRINTS--a database of protein motif fingerprints. Nucleic Acids Research, 1994, 22(17):3590-3596.
11. Finn R. D., Mistry J., Tate J., et al. The Pfam protein families database. Nucleic Acids Research, 2010, 38(Database issue):D211-222.
12. Wheeler D. L., Chappey C., Lash A. E., et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, 2000, 28(1):10-14.
13. Brooksbank C., Cameron G., Thornton J. The European Bioinformatics Institute's data resources. Nucleic Acids Research, 2010, 38(Database issue):D17-25.
14. Hamm G. H., Cameron G. N. The EMBL data library. Nucleic Acids Research, 1986,10,14(1):5-9.

第二章 双序列比对
CHAPTER 2 PAIRWISE SEQUENCE ALIGNMENT

第一节 引 言
Section 1 Introduction

当一个研究人员遇到一个 DNA 或蛋白质序列的时候，首先可能会问它包含什么信息，当遇到多个 DNA 或蛋白质序列的时候，则会问它们之间是否存在某种关系。为了有效地比较这些序列进而回答这后一问题，人们需要准确地定义若干概念，其中最主要的是同源、相似和相同的概念，它们是序列比较和分析的基础。

一、同源、相似与相同

如果两个序列享有一个共同的进化上的祖先，则这两个序列是同源(homologous)的。对这个定义需要注意的一点是，同源是个定性的概念，没有“度”的差异。对两个序列，它们或者同源或者不同源，不能说它们 70% 同源或 80% 同源。与同源相关但不同的两个概念是相似(similarity)和相同(identity)，它们都是定量的概念，基于对序列中字符的精确比较，既可以说两个蛋白质序列高度相似，也可以说它们 70% 的氨基酸相同。

在基因组分析中，一个常见而具有挑战性的问题是，对于一个基因或蛋白质，进化可以在产生物种间高度差异的碱基或氨基酸序列的同时保持 DNA 序列、RNA 序列和蛋白质序列二级和三级结构的保守性，这使得同源与相似和相同的关系常常难以确定。相反的例子是，对于某些基因或蛋白质，进化也可以产生物种间高度类似的碱基或氨基酸序列，但它们来自原先完全没有关系的种系，这种进化称为趋同进化(convergent evolution)，它是一种与同源无关的相似或相同，高度类似的碱基或氨基酸序列的产生原因是它们对应于相似或相同的功能，这也给同源性分析带来了困难。再一种与同源无关的相似是，由于氨基酸编码的冗余性，差异相当大的 DNA 序列可产生差异相当小的蛋白质序列。

在基因组测序项目中，同源性是根据数据库搜索和序列比较确定的。如果两个 DNA 或蛋白质序列经比较具有高度相似性，则它们可能是同源的。同源可进一步分作垂直同源(ortholog)和水平同源(paralog)。垂直同源是指在种系形成(speciation)过程中起源于一个共同祖先的不同种系中的 DNA 或蛋白质序列，其关系可用一棵倒置的树说明。水平同源主要是由序列复制事件产生的，例如人 α-1 球蛋白和 α-2 球蛋白是水平同源的，人 α-1 球蛋白和 β- 球蛋白也是水平同源的。一般假定，同源序列具有相同的功能。例如，与血红蛋白同源的人和鼠的肌球蛋白都能在肌肉中运输氧，但应该注意，垂直同源和水平同源基因未必总有相同的功能。图 2-1 以球蛋白基因(globin gene)为例展示了垂直同源和水平同源的区别。

二、相似性的定量描述

相似性可定量地定义为两个序列的函数，即它可有多个值，值的大小取决于两个序列对应位置上相同字符的个数，值越大则表示两个序列越相似。类似地，编辑距离(edit distance)也可定量地定

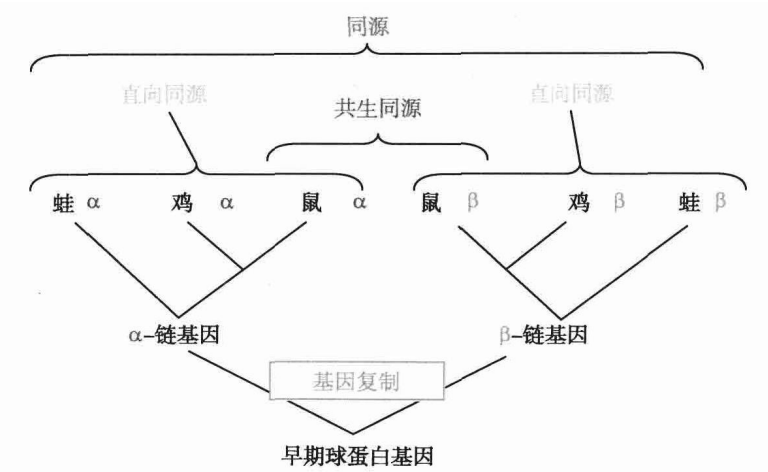


图 2-1 垂直同源与水平同源

义为两个序列的函数，其值取决于两个序列对应位置上差异字符的个数，值越小则表示两个序列越相似。可以看出，相似性和编辑距离是一对相反的定量描述序列相似性的度量。这样，相似性有两种定量表达的方式：编辑距离和相似性得分。编辑距离一般用海明距离表示，对于两条长度相等的序列，它们的海明距离等于对应位置不同字符的个数。例如，图 2-2 是 3 组序列的海明距离，其值分别为 2、3、6。

seq1=	ATC	AGGCT	GCTAGCTA
seq2=	TAC	ACCTT	CGTGAGCA
Hamming Distance (seq1,seq2)=	2	3	6

图 2-2 海明距离

相似性得分是以某种记分规则计算两个序列相似性所得的分值。记分一般是字符位置无关的（字符列无关的），即计算对应字符两两比较的分数，然后将所有字符的分数累加得到两条序列的相似性得分。显然，存在许多不同的记分规则可对两两字符比较进行记分。例如，对于图 2-2 中的 3 组序列，使用不同的记分规则可以得到如图 2-3 所示的不同得分。明显地，除了在两个字符上不同的记分规则可以产生差异，序列间排列的不同也影响相似性得分。例如，如果图 2-2 中 seq1 与 seq2 交错一位再比对，则记分结果将显著受到影响。

		seq1=	ATC	AGGCT	GCTAGCTA
		seq2=	TAC	ACCTT	CGTGAGCA
打分规则1	p(a,a)=1 p(a,b)=0 (a≠b)	相似性得分=	1	2	2
打分规则2	p(a,a)=0.8 p(a,b)=0.2 (a≠b)	相似性得分=	1.2	2.2	2.8
打分规则BLAST	A T C G A 5 −4 −4 −4 T −4 5 −4 −4 C −4 −4 5 −4 G −4 −4 −4 5	相似性得分=	−3	−2	−6

图 2-3 对相似性的记分

有了相似性得分，就可以正式引入比对的概念。双序列比对，或配对比对，是使两个序列产生最高相似性记分的序列排列。这一章的核心是，两个序列间有什么样的排列，以及在什么记分规则下，会产生最大的相似性得分。在下节详细讨论核酸和蛋白质序列比对的记分细节之前，还要介绍有关

比对的第三个重要概念，即空格。

三、空格

在技术上，同源、相似和相同都是通过序列比对确定的。确定两个序列的相似性用双序列比对。由于两个序列罕有完全一致的，在双序列比对时要对序列插入空格(gap)，使得相同的字符能够被挤迫成充分对齐。引入空格的数量和位置对比对结果有显著影响。例如，设有两条序列GACGGATTAG和GATCGGAATAG，如果第一种比对是：

```
GACGGATTAG-
GATCGGAATAG
```

第二种比对是：

```
GA-CGGATTAG
GATCGGAATAG
```

则第二种比对显然比第一种比对更好地揭示了两条序列的相似性，因为它导致了更多字符的对齐。

由于序列的差异都是由突变引起的，双序列比对同时也提供了鉴别突变位点的手段。常见的突变是替换(substitution)、插入(insertion)和删除(deletion)，其中后两者都导致在比对中引入空格。关于突变有下面一些值得注意的要点。一个碱基的替换可能导致也可能不导致相应位置氨基酸的变化，但一个碱基的插入或删除则肯定影响该位置氨基酸的编码，且由于氨基酸是三联编码，它还影响许多氨基酸的编码。另一方面，一个氨基酸的替换、插入或删除对整个序列的影响没有播散性；至于它是否显著地影响一个蛋白质的功能，这取决于它的位置，即是否在关键性的结构域(domain)。

除了对应于单字符插入和删除的空格，比对中还经常用到更大的对应于多个连续字符插入和删除的空格。多个连续字符的插入和删除可由多次独立的单字符插入和删除造成，也可由一次多字符插入和删除造成。尽管单字符突变的发生率高于多字符突变的发生率，从概率上说，引起一次多字符插入和删除的概率要大于引起多次独立单字符插入和删除的概率。后面将会述及，对这些情况的分析会影响比对软件参数的选择，进而影响比对结果。此外，对于长的空格，它们出现在序列的头、中和尾也常常具有不同意义。

第二节 替换记分矩阵

Section 2 Scoring Matrix

上面提到，对于序列中的插入和删除突变，序列比对采用插入空格来处理，使得原本对应的字符仍旧能够对应；而对于序列中的替换突变，需要考虑不同替换的意义。例如在图2-2中，第二列首个字符T对应于第一列首个字符A，这可能是由一起替换突变引起的失配。在双序列比对中对于这类失配应该怎么记分(实际上是罚分)是本节的内容。合理而精确的计分需要考虑替换的各种情形。对于DNA和RNA序列，情况特别简单，施用于4种碱基和6种彼此间替换关系的记分规则可用简单的替换记分矩阵来描述。对于蛋白质序列，因为蛋白质由20种氨基酸构成，且不同的氨基酸具有不同的理化性质，情况较为复杂，存在许多不同的替换记分矩阵。

一、通过点矩阵对序列比较进行记分

“矩阵作图法”或“对角线作图”由Gibb首先提出。将两条待比较的序列分别放在矩阵的X/Y轴上，从下往上和从左到右比较，当对应行与列的字符匹配时，则在矩阵对应的位置上打点。逐个比较所有的字符对，最终形成一个点矩阵。如果两条序列完全相同，则点矩阵的主对角线各位置都被标记，见图2-4(A)；如果两条序列存在相同的子串，则对每一个子串对有一条与对角线平行的由一系列点组成的斜线，见图2-4(B、D)；而对于两条互为反向的序列，则在反对角线方向上有由点组成

的斜线，见图 2-4(C)。这种反映序列比对的方法在直观地揭示多个相配的子串对时尤其有用，一直被使用到现在。

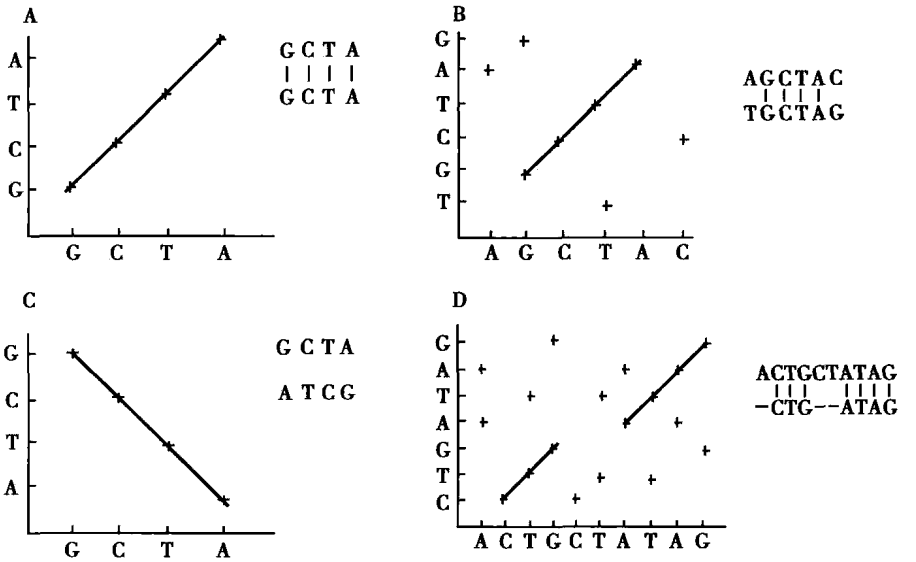


图 2-4 通过点矩阵对序列比对进行记分

A. 两条序列完全相同；B. 两条序列有一个共同的子序列；
C. 两条序列反向匹配；D. 两条序列存在不连续的两条子序列

二、DNA 序列比对的替换计分矩阵

由于替换有多种情形，且可按不同方式罚分，如何精确处理序列中的替换突变十分重要。显然，不同字符间的替换具有不同的概率，也具有不同的意义；同时，不同物种间的替换也有不同的概率和意义。精确地处理替换需要考虑各种情形，而方便地处理替换则要求把不同的处理方法参数化，这些参数就是替换记分矩阵，它们定量地标示了不同替换的意义。

借鉴上面点矩阵的方法，可以为不同字符间的替换建立替换记分矩阵(substitution matrix)，它们或依据相应碱基或氨基酸的理化性质而确定，或依据突变实际发生的概率而确定，因此相当客观和固定。各种替换记分矩阵复杂庞多，下面介绍常见的替换记分矩阵。

1. 等价矩阵(unitary matrix) 等价矩阵(表 2-1)是最简单的一种替换记分矩阵，其中，相同核苷酸间的匹配得分为 1，不同核苷酸间的替换得分为 0。尽管含义清晰明了，由于不含有碱基的任何理化信息和不区别对待不同的替换，在实际的序列比对中较少使用。

2. 转换 - 颠换矩阵(transition-transversion matrix) 核酸的碱基按照环结构特征被划分为两类，一类是嘌呤(腺嘌呤 A、鸟嘌呤 G)，它们有两个环；另一类是嘧啶(胞嘧啶 C、胸腺嘧啶 T)，它们只有一个环。如果 DNA 碱基的替换保持环数不变，则称为转换，如 A → G、C → T；如果环数发生变化，则称为颠换，如 A → C、A → T 等。在进化过程中，转换发生的频率远比颠换高，表 2-2 所示的矩阵用来反映了这种情况，其中转换的得分为 −1，而颠换的得分为 −5。

3. BLAST 矩阵 经过大量实际比对发现，如果令被比对的两个核苷酸相同时得分为 +5，反之得分为 −4，则比对效果较好。表 2-3 是其替换记分矩阵，这个矩阵广泛地被 DNA 序列比对所采用，称为 BLAST 矩阵。BLAST 是目前最流行的核酸序列数据库搜索程序。

三、蛋白质序列比对的替换记分矩阵

当比较蛋白质序列时，简单的替换记分办法，如 +1 表示匹配，0 表示失配，是不够的。构成蛋白质的氨基酸具有不同的生物化学特性，这些特性可影响它们在进化过程中的相互替换。例如，与体

表 2-1 DNA 等价矩阵

	A	T	C	G
A	1	0	0	0
T	0	1	0	0
C	0	0	1	0
G	0	0	0	1

表 2-2 转换 - 颠换矩阵

	A	T	C	G
A	1	−5	−5	−1
T	−5	1	−1	−5
C	−5	−1	1	−5
G	−1	−5	−5	1

表 2-3 BLAST 矩阵

	A	T	C	G
A	5	−4	−4	−4
T	−4	5	−4	−4
C	−4	−4	5	−4
G	−4	−4	−4	5

积差异大的氨基酸相比，体积相似的氨基酸更易于彼此替换。其他特性，如亲水性与疏水性，也影响相互替换的概率。下面介绍多个不同的氨基酸替换记分矩阵。

1. 等价矩阵(unitary matrix) 蛋白质等价矩阵与 DNA 等价矩阵道理相同，是最简单的替换记分矩阵，其中，相同氨基酸间的匹配得分为 1，而不同氨基酸间的替换得分为 0。同样，由于不含有氨基酸的任何理化信息和统计含义，在实际的序列比对中较少使用等价矩阵。

2. 遗传密码矩阵(genetic code matrix，GCM) 遗传密码矩阵通过计算一个氨基酸转变成另一个氨基酸所需的密码子变化的数目而得到，矩阵元素的值对应于代价。如果变化一个碱基就可以使一个氨基酸的密码子改变为另一个氨基酸的密码子，则这两个氨基酸的替换代价为 1；如果需要 2 个碱基的改变，则替换代价为 2(表 2-4)；而 Met 到 Tyr 的转变是仅有的密码子三个位置都需要发生变化的转换。在表 2-4 中，X 代表除 20 种标准氨基酸以外的任何氨基酸。遗传密码矩阵常用于进化距离的计算，其优点是计算结果可以直接用于绘制进化树，但是它在蛋白质序列比对(尤其是相似程度很低的蛋白质序列比对)中很少被使用。

表 2-4 遗传密码矩阵

	A	S	G	L	K	V	T	P	E	D	N	I	Q	R	F	Y	C	H	M	W	Z	B	X
A	0	1	1	2	2	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2
S	1	0	1	1	2	2	1	1	2	2	1	1	2	1	1	1	1	2	2	1	2	2	2
G	1	1	0	2	2	1	2	2	1	1	2	2	2	1	2	2	1	2	2	1	2	2	2
L	2	1	2	0	2	1	2	1	2	2	2	1	1	1	1	2	2	1	1	1	2	2	2
K	2	2	2	2	0	2	1	2	1	2	1	1	1	1	2	2	2	2	1	2	1	2	2
V	1	2	1	1	2	0	2	2	1	1	2	1	2	2	1	2	2	2	1	2	2	2	2
T	1	1	2	2	1	2	0	1	2	2	1	1	2	1	2	2	2	2	1	2	2	2	2
P	1	1	2	1	2	2	1	0	2	2	2	2	1	1	2	2	2	1	2	2	2	2	2
E	1	2	1	2	1	1	2	2	0	1	2	2	1	2	2	2	2	2	2	2	1	2	2
D	1	2	1	2	2	1	2	2	1	0	1	2	2	2	2	1	2	1	2	2	2	1	2
N	2	1	2	2	1	2	1	2	2	1	0	1	2	2	2	1	2	1	2	2	2	1	2
I	2	1	2	1	1	1	1	2	2	2	1	0	2	1	1	2	2	2	1	2	2	2	2
Q	2	2	2	1	1	2	2	1	1	2	2	2	0	1	2	2	2	1	2	2	1	2	2
R	2	1	1	1	1	2	1	1	2	2	2	1	1	0	2	2	1	1	1	1	2	2	2
F	2	1	2	1	2	1	2	2	2	2	2	1	2	2	0	1	1	2	2	2	2	2	2
Y	2	1	2	2	2	2	2	2	2	1	1	2	2	2	1	0	1	1	3	2	2	1	2
C	2	1	1	2	2	2	2	2	2	2	2	2	2	1	1	1	0	2	2	1	2	2	2
H	2	2	2	1	2	2	2	1	2	1	1	2	1	1	2	1	2	0	2	2	2	1	2
M	2	2	2	1	1	1	1	2	2	2	2	1	2	1	2	3	2	2	0	2	2	2	2
W	2	1	1	1	2	2	2	2	2	2	2	2	2	1	2	2	1	2	2	0	2	2	2
Z	2	2	2	2	1	2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	1	2	2
B	2	2	2	2	2	2	2	2	2	1	1	2	2	2	2	1	2	1	2	2	2	1	2
X	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2

3. 疏水性矩阵(hydrophobic matrix) 在相关蛋白质之间，某些氨基酸可以很容易地相互取代而不改变它们的生理生化性质，这些例子包括异亮氨酸(Isoleucine)和缬氨酸(Valine)、丝氨酸(Serine)和苏氨酸(Threonine)。根据20种氨基酸侧链基团疏水性的不同以及氨基酸替换前后理化性质变化的大小，以氨基酸的疏水性为标准制定的疏水性矩阵如表2-5所示。若一次氨基酸替换后疏水特性不发生大的变化，则这种替换得分高，否则替换得分低。该矩阵物理意义明确，有一定的理化性质依据，适用于偏重蛋白质功能方面的序列比对。

表2-5 氨基酸疏水性矩阵

	R	K	D	E	B	Z	S	N	Q	G	X	T	H	A	C	M	P	V	L	I	Y	F	W
R	10	10	9	9	8	8	6	6	6	5	5	5	5	5	4	3	3	3	3	3	2	1	0
K	10	10	9	9	8	8	6	6	6	5	5	5	5	5	4	3	3	3	3	3	2	1	0
D	9	9	10	10	8	8	7	6	6	6	5	5	5	5	5	4	4	4	3	3	3	2	1
E	9	9	10	10	8	8	7	6	6	6	5	5	5	5	5	4	4	4	3	3	3	2	1
B	8	8	8	8	10	10	8	8	8	8	7	7	7	7	6	6	6	5	5	5	4	4	3
Z	8	8	8	8	10	10	8	8	8	8	7	7	7	7	6	6	6	5	5	5	4	4	3
S	6	6	7	7	8	8	10	10	10	10	9	9	9	9	8	8	7	7	7	7	6	6	4
N	6	6	6	6	8	8	10	10	10	10	9	9	9	9	8	8	8	7	7	7	6	6	4
Q	6	6	6	6	8	8	10	10	10	10	9	9	9	9	8	8	8	7	7	7	6	6	4
G	5	5	6	6	8	8	10	10	10	10	9	9	9	9	8	8	8	8	7	7	6	6	5
X	5	5	5	5	7	7	9	9	9	9	10	10	10	10	9	9	8	8	8	8	7	7	5
T	5	5	5	5	7	7	9	9	9	9	10	10	10	10	9	9	8	8	8	8	7	7	5
H	5	5	5	5	7	7	9	9	9	9	10	10	10	10	9	9	9	8	8	8	7	7	5
A	5	5	5	5	7	7	9	9	9	9	10	10	10	10	9	9	9	8	8	8	7	7	5
C	4	4	5	5	6	6	8	8	8	8	9	9	9	9	10	10	9	9	9	9	8	8	5
M	3	3	4	4	6	6	8	8	8	8	9	9	9	9	10	10	10	10	9	9	8	8	7
P	3	3	4	4	6	6	7	8	8	8	8	8	9	9	9	10	10	10	9	9	9	8	7
V	3	3	4	4	5	5	7	7	7	8	8	8	8	8	9	10	10	10	10	10	9	8	7
L	3	3	3	3	5	5	7	7	7	7	8	8	8	8	9	9	9	10	10	10	9	9	8
I	3	3	3	3	5	5	7	7	7	7	8	8	8	8	9	9	9	10	10	10	9	9	8
Y	2	2	3	3	4	4	6	6	6	6	7	7	7	7	8	8	9	9	9	9	10	10	8
F	1	1	2	2	4	4	6	6	6	6	7	7	7	7	8	8	8	8	9	9	10	10	9
W	0	0	1	1	3	3	4	4	4	5	5	5	5	5	6	7	7	7	8	8	8	9	10

4. PAM矩阵 对于氨基酸之间的替换，对实际替换率的直接观察常常是导出合理的记分的好方法，由此产生的一组替换记分矩阵是可接受点突变矩阵(point accepted matrix，PAM)。它们基于氨基酸进化的点突变模型，即如果两种氨基酸替换频繁，说明自然界易接受这种替换，那么这对氨基酸替换得分就应该高。PAM矩阵是目前蛋白质序列比对中最广泛使用的记分方法之一，基础的PAM-1矩阵反映的是进化产生的每百个氨基酸平均发生一个突变的量值。

PAM矩阵的制作步骤是：

(1) 构建序列相似(大于85%)的比对；

(2) 计算氨基酸j的相对突变率m_j(j被其他氨基酸替换的次数)；

(3) 针对每个氨基酸对i和j，计算j被i替换的次数；

(4) 替换次数除以相对突变率(m_j)；

(5) 利用每个氨基酸出现的频度对j进行标准化；

(6) 取常用对数，得到PAM-i(i，j)。

将 PAM-1 自乘 n 次，可以得到 PAM-n(表 2-6)，但这并不意味 n 次 PAM 之后每个氨基酸都发生了变化，因为其中一些氨基酸位置可以经历多次突变，甚至可能会变回到原来的氨基酸。

表 2-6 PAM-250 矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2	−2	0	0	−2	0	0	1	−1	−1	−2	−1	−1	−3	1	1	1	−6	−3	0	0	0
R	−2	6	0	−1	−4	1	−1	−3	2	−2	−3	3	0	−4	0	0	−1	2	−4	−2	−1	0
N	0	0	2	2	−4	1	1	0	2	−2	−3	1	−2	−3	0	1	0	−4	−2	−2	2	1
D	0	−1	2	4	−5	2	3	1	1	−2	−4	0	−3	−6	−1	0	0	−7	−4	−2	3	3
C	−2	−4	−4	−5	12	−5	−5	−3	−3	−2	−6	−5	−5	−4	−3	0	−2	−8	0	−2	−4	−5
Q	0	1	1	2	−5	4	2	−1	3	−2	−2	1	−1	−5	0	−1	−1	−5	−4	−2	1	3
E	0	−1	1	3	−5	2	4	0	1	−2	−3	0	−2	−5	−1	0	0	−7	−4	−2	3	3
G	1	−3	0	1	−3	−1	0	5	−2	−3	−4	−2	−3	−5	0	1	0	−7	−5	−1	0	0
H	−1	2	2	1	−3	3	1	−2	6	−2	−2	0	−2	−2	0	−1	−1	−3	0	−2	1	2
I	−1	−2	−2	−2	−2	−2	−2	−3	−2	5	2	−2	2	1	−2	−1	0	−5	−1	4	−2	−2
L	−2	−3	−3	−4	−6	−2	−3	−4	−2	2	6	−3	4	2	−3	−3	−2	−2	−1	2	−3	−3
K	−1	3	1	0	−5	1	0	−2	0	−2	−3	5	0	−5	−1	0	0	−3	−4	−2	1	0
M	−1	0	−2	−3	−5	−1	−2	−3	−2	2	4	0	6	0	−2	−2	−1	−4	−2	2	−2	−2
F	−3	−4	−3	−6	−4	−5	−5	−5	−2	1	2	−5	0	9	−5	−3	−3	0	7	−1	−4	−5
P	1	0	0	−1	−3	0	−1	0	0	−2	−3	−1	−2	−5	6	1	0	−6	−5	−1	−1	0
S	1	0	1	0	0	−1	0	1	−1	−1	−3	0	−2	−3	1	2	1	−2	−3	−1	0	0
T	1	−1	0	0	−2	−1	0	0	−1	0	−2	0	−1	−3	0	1	3	−5	−3	0	0	−1
W	−6	2	−4	−7	−8	−5	−7	−7	−3	−5	−2	−3	−4	0	−6	−2	−5	17	0	−6	−5	−6
Y	−3	−4	−2	−4	0	−4	−4	−5	0	−1	−1	−4	−2	7	−5	−3	−3	0	10	−2	−3	−4
V	0	−2	−2	−2	−2	−2	−2	−1	−2	4	2	−2	2	−1	−1	−1	0	−6	−2	4	−2	−2
B	0	−1	2	3	−4	1	3	0	1	−2	−3	1	−2	−4	−1	0	0	−5	−3	−2	3	2
Z	0	0	1	3	−5	3	3	0	2	−2	−3	0	−2	−5	0	0	−1	−6	−4	−2	2	3

5. BLOSUM 矩阵(block substitution matrix，BLOSUM) BLOSUM 矩阵是由 Henikoff 首先提出的另一种氨基酸替换记分矩阵，它也是通过统计相似蛋白质序列的替换率而得到的。PAM 矩阵是从蛋白质序列的全局比对结果推导出来的，而 BLOSUM 矩阵则是从蛋白质序列块(短序列)比对推导出来的。基本数据来源于 BLOCKS 数据库，其中包括了局部多重比对，虽然没有使用进化模型，但它的优点在于可以通过直接观察而不是通过外推获得数据。同 PAM 模型一样，也有许多不同编号的 BLOSUM 矩阵，这里的编号指的是序列可能相同的最高水平，并且同模型保持独立性。表 2-7 所示的 BLOSUM 矩阵是由具有 62% 相同比例的序列被组合统计后形成的矩阵。注意，在比对高度相似的序列时使用较高值的矩阵(高至 BLOSUM-90)，在比对差异大的序列时使用较低值的矩阵(低至 BLOSUM-30)。

表 2-7 BLOSUM-62 矩阵

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	4	−1	−2	−2	0	−1	−1	0	−2	−1	−1	−1	−1	−2	−1	1	0	−3	−2	0	−2	−1
R	−1	5	0	−2	−3	1	0	−2	0	−3	−2	2	−1	−3	−2	−1	−1	−3	−2	−3	−1	0
N	−2	0	6	1	−3	0	0	0	1	−3	−3	0	−2	−3	−2	1	0	−4	−2	−3	3	0
D	−2	−2	1	6	−3	0	2	−1	−1	−3	−4	−1	−3	−3	−1	0	−1	−4	−3	−3	4	1
C	0	−3	−3	−3	9	−3	−4	−3	−3	−1	−1	−3	−1	−2	−3	−1	−1	−2	−2	−1	−3	−3
Q	−1	1	0	0	−3	5	2	−2	0	−3	−2	1	0	−3	−1	0	−1	−2	−1	−2	0	3

续表

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
E	−1	0	0	2	−4	2	5	−2	0	−3	−3	1	−2	−3	−1	0	−1	−3	−2	−2	1	4
G	0	−2	0	−1	−3	−2	−2	6	−2	−4	−4	−2	−3	−3	−2	0	−2	−2	−3	−3	−1	−2
H	−2	0	1	−1	−3	0	0	−2	8	−3	−3	−1	−2	−1	−2	−1	−2	−2	2	−3	0	0
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4	2	−3	1	0	−3	−2	−1	−3	−1	3	−3	−3
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4	−2	2	0	−3	−2	−1	−2	−1	1	−4	−3
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5	−1	−3	−1	0	−1	−3	−2	−2	0	1
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5	0	−2	−1	−1	−1	−1	1	−3	−1
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6	−4	−2	−2	1	3	−1	−3	−3
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7	−1	−1	−4	−3	−2	−2	−1
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4	1	−3	−2	−2	0	0
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5	−2	−2	0	−1	−1
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11	2	−3	−4	−3
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	−1	−3	−2
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4	−3	−2
B	−2	−1	3	4	−3	0	1	−1	0	−3	−4	0	−3	−3	−2	0	−1	−4	−3	−3	4	1
Z	−1	0	0	1	−3	3	4	−2	0	−3	−3	1	−1	−3	−1	0	−1	−3	−2	−2	1	4

对于 PAM-n 矩阵，n 越小表示氨基酸变异的可能性越小，高相似序列之间的比对应该选用 n 值小的矩阵，低相似序列之间的比对应该选用 n 值大的矩阵。例如，PAM-250 用于约 20% 相同的序列之间的比对。对于 BLOSUM-n 矩阵，n 越小则表示氨基酸相似的可能性越小，高相似的序列之间比较应该选用 n 值大的矩阵，低相似序列之间的比对应该选用 n 值小的矩阵。例如，BLOSUM-62 用来比较 62% 相似度的序列，BLOSUM-80 用来比较 80% 左右的序列。PAM 与 BLOSUM 编号之间的关系见图 2-5。

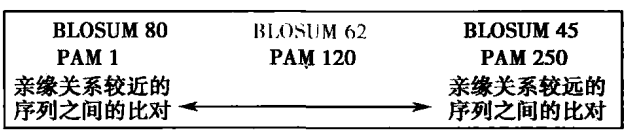


图 2-5 PAM/BLOSUM 矩阵编号与序列亲缘关系的比较

第三节 双序列比对算法

Section 3 Algorithms of Pairwise Sequence Alignment

在生物信息学中，对各种生物大分子的序列进行分析是非常基本的工作。序列的测定和拼接、RNA 和蛋白质的结构功能预测、种系树的构建等都需要对生物分子进行序列相似性的比较。在长期的生命进化过程中，不同物种的 DNA 经历了突变、复制、片段缺失和片段增加等变化，但许多部分仍具有高度的相似。

序列相似性比对在发现生物序列有关功能、结构和进化的信息方面具有非常重要的意义，其主要思想就是运用特定的算法找出两个或多个序列之间产生最大相似性得分的空格插入和序列排列方案。在实际中，序列比对是计算生物学中解决序列装配、进化树重构及分析基因功能等众多问题的第一步。进行序列比对的算法很多，为了找出最优比对，它们大多数基于动态规划算法。根据同时比对的序列数量的不同，一般将序列比对分成双序列比对和多序列比对，后者是前者的推广。一些算法两者皆适用，而一些算法只用于各自的情况。

生物序列(DNA 序列、RNA 序列和蛋白质序列)可以看作是由固定的字母表中的字母所组成的字符串，两条序列 s 和 t 的比对可以简单地表示为：把 s 和 t 这两条序列上下排列起来，在某些位置插入空格，然后依次比较它们在每一个位置上字符的匹配情况，从而找出使这两条序列产生最大相

似度得分的排列方式和空格插入方式。

一、全局比对的经典算法

对于两条序列的比对问题人们提出了很多算法，其中基于动态规划的算法是目前最基本的算法。20 世纪 40 年代，Richard Bellman 最早使用动态规划这一概念表述通过遍历寻找最优决策问题的求解过程。1953 年，Richard Bellman 将动态规划赋予现代意义，该领域后被 IEEE 纳入系统分析和工程中。动态规划算法以递归形式重申一个优化问题。在“动态规划”(dynamic programming)一词中，programming 来自“数学规划”(mathematical programming)，含优化之意，与计算机编程中的 programming 并无联系。

把动态规划算法应用于生物信息学中的序列比对起源于 1970 年，由 Saul Needleman 和 Christian Wunsch 两人首先将其用于两条序列的全局比对，其算法(algorithm)后称为 Needleman-Wunsch 算法。后来，Temple Smith 和 Michael Waterman 两人于 1981 年对双序列的局部比对进行了研究，产生了 Smith-Waterman 算法。下面以 Smith-Waterman 算法为例子分步骤介绍动态规划算法的思想。

1. 动态规划法的思想　首先，对于如下假定的序列：

(1) a，b 是使用某一字符集 ∑ 的序列(DNA 或蛋白质序列)；

(2) m=a 的长度；

(3) n=b 的长度；

(4) S(i,j)是按照某替换记分矩阵得到的前缀 a[1...i]与 b[1...j]最大相似性得分；

(5) w(c，d)是字符 c 和 d 按照替换记分矩阵计算的得分。

可按照规则建立得分矩阵：

$$S(i,0)=0,\quad 0\leqslant i\leqslant m$$

$$S(0,j)=0,\quad 0\leqslant j\leqslant n$$

$$S(i,j)=\max\begin{cases}0\\ S(i-1,j-1)+w(a_i,b_j) & \text{匹配或错配}\\ S(i-1,j)+w(a_i,-) & \text{插入}\\ S(i,j-1)+w(-,b_j) & \text{缺失}\end{cases}\qquad \text{式 2-1}$$

例如，对于序列 a=ACACACTA，序列 b=AGCACACA，记分规则 w(匹配)=+2；w(a，−)=w(−，b)=w(失配)=−1，则获得的得分矩阵如图 2-6 所示。接着，反向搜寻最大得分，同时记下读取路径。为了得到最佳比对，必须从得分最高的位置 S(i,j)开始，在矩阵的(i−1，j)，(i，j−1)或(i−1，j−1)位置中寻找下一个最大得分位置，记下路径(画箭头)，当两个(或三个)位置得分相等时，取对角线方向，依此规则搜寻，直至到起点(0，0)。在本例中，最大得分对应的位置分别为(8，8)、(7，7)、(7，6)、(6，5)、(5，4)、(4，3)、(3，2)、(2，1)、(1，1)和(0，0)，见图 2-7。

$S=$

	−	A	C	A	C	A	C	T	A
−	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

图 2-6　一个得分矩阵实例

$S=$

	−	A	C	A	C	A	C	T	A
−	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

图 2-7　得分矩阵路径实例

最后，是构建最佳匹配。在读取路径中要求：对角线对应匹配(或失配)、上下箭头对应删除、左右箭头对应插入。依此规则，可以得到本例的最佳匹配为：

序列 a = A − C A C A C T A

序列 b = A G C A C A C − A

现在看算法的复杂度。从所使用的数据结构本身及其计算过程来看，序列两两比对基本算法的空间复杂度和时间复杂度都是 O(mn)。

2. 动态规划法的流程　动态规划法大致包括：①按照规则建立得分矩阵；②反向读取最大得分，构建最佳匹配。每一步都包括若干子步骤。按照规则建立得分矩阵的流程是：

```
for i=0 to length(A)
  F(i,0) ←0
for j=0 to length(B)
  F(0,j) ←0
for i=1 to length(A)
  for j=1 to length(B)
  {
    Choice1 ← F(i−1,j−1)+S(A(i),B(j))
    Choice2 ← F(i−1,j)+d
    Choice3 ← F(i,j−1)+d
    F(i,j) ← max(Choice1, Choice2, Choice3)
  }
```

反向读取最大得分，构建最佳匹配的流程是：

```
AlignmentA ← ""
AlignmentB ← ""
i ← length(A)
j ← length(B)
while (i>0 and j>0)
{
  Score ← F(i,j)
  ScoreDiag ← F(i−1,j−1)
  ScoreUp ← F(i,j−1)
  ScoreLeft ← F(i−1,j)
  if (Score == ScoreDiag + S(A(i−1),B(j−1)))
    {
      AlignmentA ← A(i−1)+AlignmentA
      AlignmentB ← B(j−1)+AlignmentB
      i ← i−1
      j ← j−1
    }
    else if (Score == ScoreLeft + d)
    {
      AlignmentA ← A(i−1)+AlignmentA
```

```
        AlignmentB ← "-" + AlignmentB
        i ← i - 1
      }
      otherwise (Score == ScoreUp + d)
      {
        AlignmentA ← "-" + AlignmentA
        AlignmentB ← B(j - 1) + AlignmentB
        j ← j - 1
      }
}
```

二、局部比对的经典算法

有时，人们会遇到这样的情况，即手里有一段序列，想知道这一段序列和另一段人们所关注的序列间有没有同源的子序列。这涉及子序列与完整序列的比对。注意这一短一长两序列间除了人们关注的共同区段外可能并没有太多的相似性(见下面的两个序列)，如果对它们做整体比对则很可能不会有一个高的得分，这使得用于局部比对的各种算法应运而生。

----AGCT----

ATGCAGCTGCTT

处理子序列与完整序列(或短序列与长序列)的比对的一般过程是：设短序列 a 和长序列 b，它们的长度分别为 L_a 和 L_b，比对是在 b 序列中寻找 L_a 长度的 a 序列的过程。这个过程的实现需要对上述动态规划算法做一些改动，它不计算删除序列 a 前缀的得分，也不计算删除序列后缀的罚分，其他行(除最后一行)的计算不变。最后一行的计算是按以下公式：

$$S(i,j)=\max\begin{cases}S(i-1,j-1)+w(a_i,b_j) & \text{匹配或错配}\\ S(i-1,j)+w(a_i,-) & \text{插入}\\ S(i,j-1) & \text{缺失不罚分}\end{cases} \qquad \text{式 2-2}$$

$S_{i,j}$ 依然是最优局部比对的得分，而匹配的子列 b 按如下方式寻找：

$$j=\min\{k \mid S_{i,k}=S_{i,n}\} \qquad \text{式 2-3}$$

然后由位置(i,j)出发，反推比对路径，最终通过斜线(非空位)到达$(0,j)$。

第四节 数据库搜索

Section 4 Database Search

在分子生物学研究中，对于新测定的碱基序列或氨基酸序列，人们往往试图通过数据库搜索找出与其相似的序列，以推测该未知序列是否与已知序列同源，或可能属于哪个基因家族，以及具有哪些生物功能。例如，对氨基酸序列来说，有可能找到已知的同源蛋白质，从而由已知蛋白质的三维结构推测未知蛋白质的空间结构。

数据库搜索是双序列局部比对的特例。新测定的、希望通过数据库搜索确定其性质或功能的序列称作探测序列(probe sequence)，通过数据库搜索得到的和探测序列具有一定相似性的序列称目标序列(subject sequence)。如果数据库搜索的目的是为了确定探测序列是否和某个基因家族存在进化关系，在搜索到相似序列后还需要判断其序列相似性的程度。如果探测序列和目标序列的相似性程度很低，还必须通过其他方法或实验手段才能确定其是否属于同一基因家族。

一、BLAST

基本局部比对搜索工具(basic local alignment search tool, BLAST)是目前最常用的数据库搜索程序。国际上各著名的生物信息中心都提供基于 Web 的 BLAST 在线服务。基本的 BLAST 算法本身很简单,它的要点是片段对(segment pair)的概念。所谓片段对是指两个给定序列中的一对子序列,它们的长度相等,且可以形成无空格的完全匹配。BLAST 首先找出探测序列和目标序列间所有匹配程度(以得分计)超过一定阈值的序列片段对,然后对片段对根据给定的相似性阈值进行延伸,得到一定长度的相似性片段,最后给出高分值片段对(high-scoring pairs, HSPs)。改进后的 BLAST 允许空格的插入。BLAST 实际上是综合在一起的一组程序,不仅可用于直接对蛋白质序列数据库和核酸序列数据库进行搜索,而且可以将探测序列翻译成蛋白质后再进行搜索,以提高搜索结果的灵敏度(表 2-8)。

表 2-8 BLAST 的探测序列和数据库的类型

程序名	探测序列	数据库类型	方法
blastp	蛋白质	蛋白质	用蛋白质探测序列搜索蛋白质序列数据库
blastn	核酸	核酸	用核酸探测序列搜索核酸序列数据库
blastx	核酸	蛋白质	将核酸序列按 6 条链翻译成蛋白质序列后搜索蛋白质序列数据库
tblastn	蛋白质	核酸	用蛋白质探测序列搜索核酸序列数据库,核酸序列按 6 条链翻译成蛋白质
tblastx	核酸	核酸	将核酸序列按 6 条链翻译成蛋白质序列后搜索由核酸序列数据库按 6 条链翻译成的蛋白质序列的数据库

BLAST 是免费软件,除了在线服务,研究人员也可以从 NCBI 等的文件服务器上下载获得,包括 UNIX 和 WINDOWS 系统的版本,安装在本地计算机上使用。需注意的是,本地运行必须有 BLAST 格式的数据库,它们也可以从 NCBI 下载,或利用该系统提供的格式转换工具由其他格式转换而得到。对核酸序列数据库而言,本地运行需要很大的磁盘空间、较大的内存和较快的运算速度,因此必须使用高性能的服务器。

1. BLAST 的算法 BLAST 先找出某些“种子(称作 words)”,即探测序列和数据库序列间非常短的匹配的片段对,它们的比对得分至少是 T,然后向两端不带空格地扩展这些种子,并使用替换记分矩阵计算得分,直到达到最大可能得分。程序并不持续地对种子进行扩展,当得分低于某个既定的阈值时便停止。不带空格的片段比对是标准 BLAST 的一个特征,实际上这也是 BLAST 运行非常之快的一个主要原因。上述的启发式的对片段对进行扩展的技术会产生一个很小的机会,使得正确的最大扩展片段不被找到。对于一个搜索,BLAST 返回探测序列与数据库序列所有得分值大于某阈值 S 的片段对(图 2-8),称为高记分对,S 由系统或用户确定。

2. BLAST 的应用 BLAST 具有非常广泛的应用,包括以下方面:

(1) 确定一个蛋白质或核酸序列有哪些垂直同源或水平同源序列;

(2) 确定哪些蛋白质或基因在特定的物种中出现;

(3) 发现新基因;

(4) 确定一个基因或蛋白质的变种;

(5) 寻找对于一个蛋白质的功能或结构起关键作用的片段。

3. 搜索步骤 使用 BLAST 搜索的步骤是:

(1) 选择感兴趣的序列,可以是 FASTA 格式的序列,也可以是访问编号;

(2) 选择 BLAST 程序,包括 blastp、blastn、blastx、tblastn、tblastx;

(3) 选择数据库;

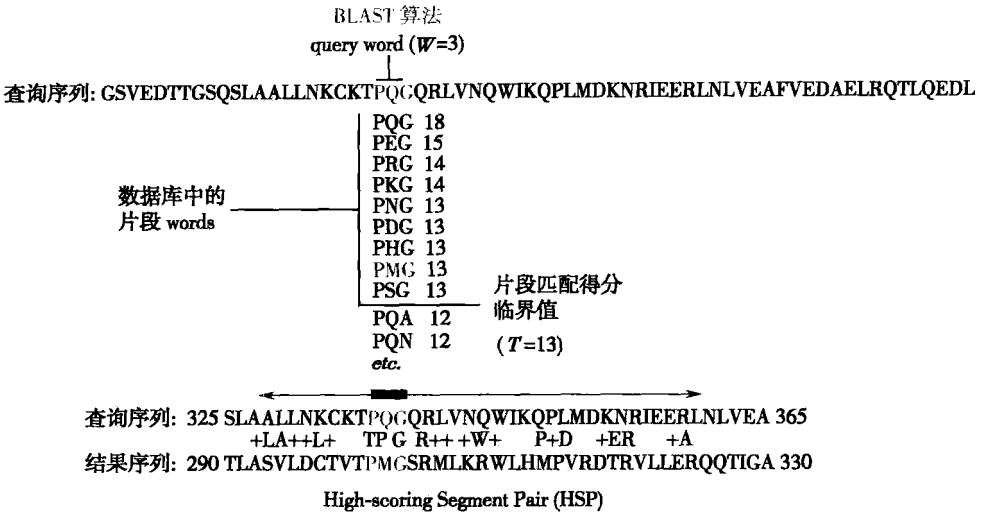


图 2-8 BLAST 算法图示

(4) 选择参数。

4. 选择参数 常用的参数是(不同的版本参数会略有不同):

(1) -p ProgramName

p 代表 program,用来选择程序,可带五个选项之一 blastp、blastn、blastx、tblastn 和 tblastx。

(2) -i QueryFile

用于指定包含探测序列的查询文件。

(3) -d DatabaseName

选择待搜索的数据库,可以选择多个数据库。

(4) -o OutputFileName

数据库搜索输出文件的名称,默认的计算机屏幕。

(5) -e ExpectedValue

E 期望值,这一参数控制搜索的敏感性。

(6) -m SpecifiesAlignmentView

设定搜索结果的显示格式,选项有 12 个,其中 0 是默认参数,显示探测序列和目标序列两两比对的信息。

(7) -F FilterQuerySequence

屏蔽简单重复和低复杂度序列的参数,有 T(选上)和 F(不选)两个选项。

(8) -E CostToExtendGap

给出空位延伸罚分。

二、数据库搜索实例

(一) 多结构域蛋白 H1N1

新型甲型 H1N1 流感病毒是一个来源于不同宿主和不同地域的重组株。HA 蛋白的氨基酸序列分析表明北美毒株不具有高致病性流感病毒的特性;NS1 蛋白和 PB2 蛋白的氨基酸序列分析表明该病毒对人具有明显的亲和性和较低的致病力;M 蛋白的同源建模表明该病毒具有抗金刚烷胺类药物的结构特点;NA 蛋白的序列分析还表明北美毒株仍然对神经氨酸酶抑制剂类药物敏感。下面以 H1N1 流感病毒的红血球凝聚素(Hemaglutinin)为例,说明如何通过 NCBI 网页搜索该蛋白的同源序列和分析相同的结构域,步骤如下:

>gi|224983683|pdb|3GBN|B Chain B, Crystal Structure Of Fab Cr6261 In Complex With The 1918 H1n1 Influenza Virus Hemagglutinin
GLFGAIAGFIEGGWTGMIDGWYGYHHQNEQGSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTAVGKEF
NNLERRIENLNKKVDDGFLDIWTYNAELLVLLENERTLDFHDSNVRNLYEKVKSQLKNNAKEIGNGCFEF
YHKCDDACMESVRNGTYDYPKYSEESKLNREEIDGVSGR

（1）下载蛋白序列；

（2）登录 NCBI 主页 http://www.ncbi.nlm.nih.gov/；

（3）点击“BLAST”；

（4）点击“protein blast”；

（5）对话框中输入 H1N1 流感病毒红血球凝聚素蛋白序列(也可以输入 GenBank 序号，或上传已经存储的序列文件)；

（6）选择蛋白质数据库：Protein Data Bank proteins(pdb)，见图 2-9；

（7）其他参数使用默许参数；

（8）点击 BLAST 按钮，得到数据库搜索的结果。图 2-10 显示了 BLAST 产生的 52 条匹配蛋白质序列(2009 年 10 月 21 日查询)；

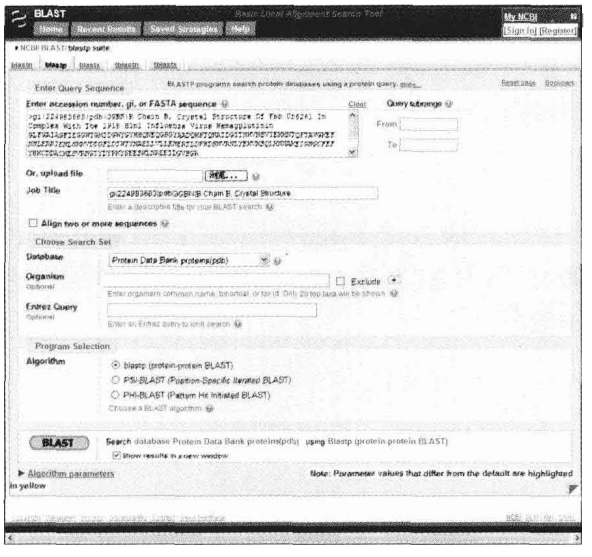


图 2-9 输入 BLAST 查询序列、选择数据库

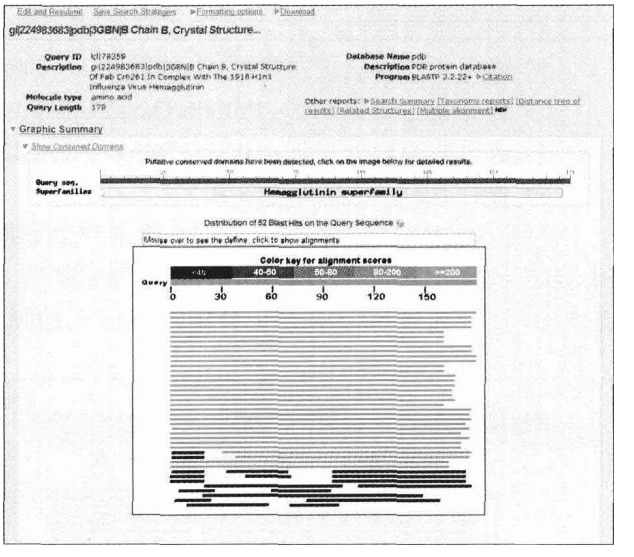


图 2-10 BLAST 查询返回结果图示

（9）点击感兴趣的序列，可以跳到如图 2-11 所示的序列匹配的详情界面；

（10）如果发现该条匹配序列可能是查询的目标序列，可以点击 Protein 3D Structure 链接查看其结构域情况(图 2-12)；

（11）跟踪网页上目标序列的链接，分析其结构域、保守结构域等。

随书光盘中的“演示 blast 蛋白质 H1N1 结构域 .wmv”展示了更多细节。

（二）改变替换记分矩阵的作用

替换记分矩阵在 BLAST 搜索中具有重要作用，可直接影响数据库搜索的结果。下面以脂质运载蛋白序列为例，展示替换记分矩阵的影响。搜索步骤是：

>sp|P31025|LCN1_HUMAN Lipocalin-1 OS=Homo sapiens GN=LCN1 PE=1 SV=1
MKPLLLAVSLGLIAALQAHHLLASDEEIQDVSGTWYLKAMTVDREFPEMNLESVTPMTLT
TLEGGNLEAKVTMLISGRCQEVKAVLEKTDEPGKYTADGGKHVAYIIRSHVKDHYIFYCE
GELHGKPVRGVKLVGRDPKNNLEALEDFEKAAGARGLSTESILIPRQSETCSPGSD

图 2-11 BLAST 查询返回结果序列匹配详情界面

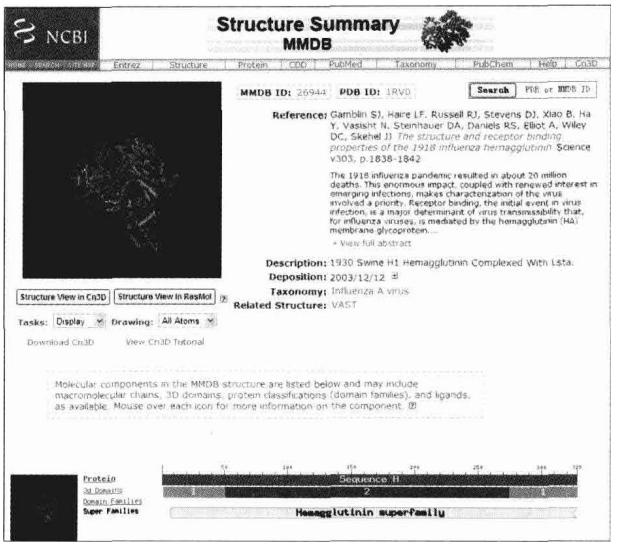


图 2-12 Protein 3D Structure 空间结构分析界面

(1) 下载脂质运载蛋白序列;

(2)~(4) 同实例一;

(5) 输入脂质运载蛋白序列(也可以输入 genBank 序号,或上传已经存储的序列文件);

(6) 选择蛋白质数据库: Protein Data Bank proteins(pdb);

(7) 选择替换记分矩阵: BLOSUM62 或 PAM30;

(8) 查看 BLAST 返回的结果。

使用 BLOSUM62 矩阵,BLAST 搜索产生如图 2-13 所示的 26 条匹配序列;使用 PAM30 矩阵,BLAST 搜索产生如图 2-14 所示的 8 条匹配序列(2009 年 10 月 21 日执行)。更多细节可查看随书光盘中的"演示选择 blast 替换记分矩阵 .wmv" 动画演示短片。

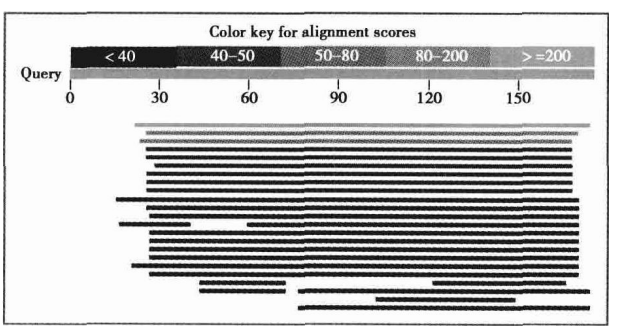


图 2-13 使用 BLOSUM62 矩阵 BLAST 产生的脂质运载蛋白序列搜索结果

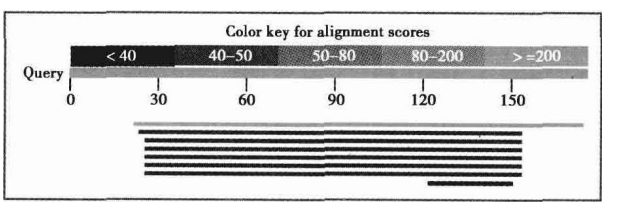


图 2-14 使用 PAM30 矩阵 BLAST 产生的脂质运载蛋白序列搜索结果

第五节 比对的统计学显著性

Section 5 Statistical Significance of Pairwise Alignment

本章前面说过,同源是根据相似性判断的,而两个序列的相似性是用比对算法精确计算的。存在的一个问题是,由比对算法计算的相似性是否可靠? 例如,两个长度为 10 的序列有 50% 的字符一致,另两个长度为 100 的序列有 50% 的字符一致,这两种情况下相似性是否具有同等意义? 实际上,当比对软件产生一个表明两序列高相似的高记分后,人们需要用统计方法检测两序列是真匹配(真

阳性，即两个比对的序列是真正同源的）还是假匹配（假阳性，即它们碰巧相配）。如果两个序列因为比对记分低于某个阈值而被报不匹配，人们也需要知道这是真失配（真阴性，即真正不相关）还是假失配（假阴性，即属于同源序列但比对记分低于假定的同源性的要求）。由于比对结果需要统计检验，在设计比对算法时，一个重要的问题是尽可能提高比对的敏感性（sensitivity）和特异性（specificity）。敏感性是算法正确识别真正相关序列的能力，它等于真阳性的数目除以真阳性与假阴性数目之和；特异性则涉及非同源序列的比对，它等于真阴性的数目除以真阴性与假阳性数目之和。对于数据库搜索，当序列数据库变得十分庞大时，存在相当大的概率使相对短的序列能得到随机的高匹配和产生假阳性。那么，如何评估这种假阳性的产生？

一、全局比对的统计学显著性

当比对两个蛋白质（例如β球蛋白和肌球蛋白）产生一个得分（score）后，人们可以用假设检验（hypothesis testing）来评估这个记分偶然获得的可能性。其做法是，先给定一个无效假设 H_0（亦称检验假设），即两个序列是不相关的，因此根据这一假设，比对得到的得分（score）是偶然出现的；接着再给出假设 H_1，即假定两个序列是真相关的。然后，设定一个统计值 α，通常为 0.05，作为确定统计显著性的阈值。确定比对得分 score 是否偶然获得的一个办法是，将β球蛋白或肌球蛋白与大量非同源的蛋白质做比对，然后将 score 与这些比对的得分进行比较。第二个办法是，把一个序列与一组随机产生的序列进行比对，然后同样将 score 与这些比对的得分进行比较。第三个办法是，随机将两个序列中的一个打乱重组，比如说重组 100 次，并与另一个序列比对，同样得到一组比对的得分。假定由这一群比对得到的得分服从正态分布，则利用下式可计算得分大于或等于 score 的概率：

$$Z=(S-M)/D \qquad \text{式 2-4}$$

假定采用第三个办法，则 M 和 D 分别是100组随机重组序列的比对所产生的得分的平均值和标准差，S 是得分 score。当 Z 值分别为 3.1、4.3 和 5.2 时，得分 score 随机出现的概率分别为 10^{-3}、10^{-5} 和 10^{-7}。因此，可以根据 Z 值判断两个序列相似性得分的显著性。一般假定对于一个高比对得分，当 $Z>5$ 时两条序列在进化上是真正相关的；当 $3\leqslant Z\leqslant 5$ 之间时，如果两者有其他方面的相似性证据（如功能相似），则两条序列也可能是真正相关的；如果 $Z<3$，则表示两条序列未必同源。许多序列比对软件都带有计算 Z 值的程序，可直接用于评价序列比对的显著性。

上述对全局比对的统计学显著性进行检验的方法有一个问题，那就是用于检验的全局比对序列的分布是不知道的，得分的分布情况也是不清楚的。在某些情况下，得分可能不是正态分布。因此，目前对全局比对的统计分析了解尚少。

二、数据库搜索的统计学显著性

局部比对也存在统计学显著性问题，且已经发展了更加严格的统计学检验方法。当用探测序列与一个序列长度统一的随机序列的数据库进行比对时，通常会得到一个符合所谓极值分布的图。与正态分布相比，极值分布是不对称的，向坐标右侧偏移。这一分布的性质使得人们对 BLAST 比对的统计学获得深刻的认识，从而能够估计一个搜索的最高得分随机出现的可能性。对于两个随机序列 s 和 t，随机观察到一个比对得分等于或大于 x 的概率为

$$P(s\geqslant x)=1-\exp(-\mathrm{Kste}^{-\lambda x}) \qquad \text{式 2-5}$$

对于 BLAST 数据库搜索，上式中 s 和 t 分别指探测序列的长度和整个数据库的长度，乘积 st 定义了搜索空间的大小。如前所述，BLAST 返回探测序列与数据库序列所有得分值大于某阈值 S 的高记分片段对，其期望数量为

$$E=\mathrm{Kste}^{-\lambda S} \qquad \text{式 2-6}$$

这就提供了对于假阳性结果的一个估计。另外由此式可看出，E 值与得分 S 和用来度量记分系统的

参数 λ 有关，同时也与探测序列的长度 s 和数据库长度 t 有关。该式具有两个重要特点：①随着 S 的增加 E 值呈指数下降，当 E 值接近零时一个比对随机发生的可能性也就接近零；②数据库的大小以及探测序列的长度将影响特定比对随机发生的可能性。

一个典型的 BLAST 的输出包括 E 值和得分。得分有两种：原始得分(raw scores)和比特得分(bit scores)。原始得分是根据所选择的替换记分矩阵和空格罚分参数计算得到的，比特得分是对原始得分处理后得到的。使用比特得分的好处是，比特得分表明了使用的记分系统并包含了比对的内在信息，它使得不同数据库搜索之间即便使用了不同的替换记分矩阵也可以进行比较。将一个原始得分 S 转换为比特得分 S' 的公式是

$$S'=(\lambda * S-\ln K)/(\ln 2) \qquad \text{式 2-7}$$

这里 λ 和 K 是两个取决于记分系统(替换记分矩阵和空格罚分)的参数。P 值和 E 值是反映比对显著性的两种不同方式。如上面两公式所示，找到一个具有给定 E 值的高记分片段对的概率是

$$P=1-e^{-E} \qquad \text{式 2-8}$$

表 2-9 列出了一些 E 值与 P 值的关系。传统上，人们使用一个低于 0.05 的 P 值来定义统计的显著性，但大部分 BLAST 在线服务使用了 E 值而非 P 值来定义搜索的统计学显著性。当 $E<0.05$ 时，P 值与 E 值接近相同，一个等于或小于 0.05 的 E 值可被认为是统计学上意义显著的。但是当搜索一个很大的数据库时，一些得到高分的比对仍可能是随机发生的，为了确保比对的显著性，这时人们常常将显著性水平下调到一个更小的值，例如 0.01。参数 K 和 λ 可分别被简化地视为搜索步长和计分规则的特征数。

表 2-9 计算的 E 值与 P 值的关系

E	P	E	P
10	0.99995	0.1	0.09516
5	0.99326	0.05	0.04877
2	0.86466	0.001	0.0009995
1	0.63212	0.0001	0.0001

第六节 参数的选择

Section 6 Selecting Scoring Parameters

一、空格罚分参数

对比对的记分由得分与罚分两部分计算而得，不同的算法以及算法的不同运行主要只在于罚分的不同。罚分包含对空格进行罚分和对失配进行罚分。对失配进行罚分是根据第二节所介绍的不同的替换记分矩阵进行计算的，本节简单介绍与空格罚分有关的参数选择。

空格罚分涉及几个问题：①空格罚分是否大于失配罚分；②不同大小空格的罚分；③空格的引入与延伸是否予以不同罚分，这些问题对比对结果可有显著影响。

首先，对于空格罚分是否应大于失配罚分，其确定类似于替换记分矩阵的选择，需根据序列特征而定。如果已知比对的序列包含相当多的进化引起的插入和删除突变，则引入空格可合理地代表这些突变，这可令空格罚分小于失配罚分来实现，使得同源的未突变的字符得以匹配；如果比对的序列少有插入和删除突变，但有许多替换突变，则记分应该忠实地基于这些失配，不应轻易引入空格去破坏虽已变异但仍同源的字符序列，这可令空格罚分大于失配罚分来实现。

其次，对确定不同大小空格的罚分常常缺乏足够的依据。一个插入或删除突变可发生于一个碱

基，也可发生于一段序列，而多个独立发生于相邻单个碱基的删除突变也可造成一段序列的删除，后两者都需要引入一段连续的空格(且称空格段)。尽管一个碱基删除的发生率高于一个片段删除的发生率，从概率上说，相邻碱基 N 次独立删除事件的概率要小于 1 次含 N 个碱基的片段删除的概率，这使得 N 个长为 1 的空格的罚分之和要大于 1 个长为 N 的空格段的罚分，但具体如何确定无定规可循。某些比对软件专门使用了空格段的罚分，而通常的软件是使用两个参数，即引入和扩展一个空格分别用参数 gap_open 和 gap_extend 控制。给 gap_open 一个较低的值使空格容易引入，而给 gap_extend 一个较低的值则使空格容易扩展。上面的记分参数选择主要与全局与局部双序列比对有关。对于使用 BLAST 做数据库搜索，另有几个相关参数需要选择。

二、BLAST 的参数

BLAST 程序的参数有搜索参数，包括字长(word size)、期望值 E、空格罚分、替换记分矩阵、阈值、窗口尺寸(window size)等，以及统计学显著性参数，包括 λ 和 K，其意义与选择用表 2-10 予以说明。

表 2-10 BLAST 的参数

参数	意义与选择
字长(word size)	探测序列和数据库序列间匹配的短片段对的长度(即“种子”的长度)。对蛋白质序列一般为 3，对 DNA 序列一般为 11 或更长些。小的字长产生更多的种子，可提高敏感性，耗费更多的时间，但是否返回更多结果还取决于其他参数；大的字长产生更少的种子，可提高特异性，耗费更少的时间
期望值(Expectation value)	对于 blastn、blastp、blastx 和 tblastn 期望值的默认值是 10，在这个 E 值下随机出现得分等于或大于比对得分 S 的期望数为 10 个。将期望值调小时，返回的数据库搜索结果将变少，匹配被搜索到的概率也变小；反之，增大 E 值将返回更多的结果
引入空格(cost to open a gap)	通常是 11，关于此参数的作用与确定，参见本节第一部分讨论
扩展空格(cost to extend a gap)	通常是 1，关于此参数的作用与确定，参见本节第一部分讨论
替换记分矩阵	对于 blastp 的蛋白质 - 蛋白质搜索常用的氨基酸替换记分矩阵有 PAM30、PAM70、BLOSUM45、BLOSUM62 及 BLOSUM80。通常应该在一次搜索中使用数种矩阵，比较获得的结果
窗口尺寸(Multiple hits window size)	指的是分隔两个独立的种子匹配 / 延伸的间隔，通常是 40。大的参数值产生更少的种子匹配 / 延伸和搜索结果，大的则相反
阈值(Threshold for extending hits)	指的是种子延伸的记分阈值。小的参数值产生更多的种子延伸，大的则相反
λ 值	对于无空格比对通常为 0.32，对于带空格比对通常为 0.267
K 值	对于无空格比对通常为 0.137，对于带空格比对通常为 0.041

三、如何处理太多与太少的数据库搜索返回

如果一次数据库搜索产生了太多的返回结果，可采取如下措施：

(1) 使用参考序列(带“refseq”的)数据库，这样可减少许多冗余结果；

(2) 使探测序列只包含一个结构域，减少多结构域带来的多匹配；

(3) 根据探测序列与数据库序列的关系使用更合适的替换记分矩阵；

(4) 降低 E 值。

如果一次数据库搜索产生了太少的返回结果，可采取如下措施：

(1) 提高 E 值；

(2) 使用更大的 PAM 矩阵或更小的 BLOSUM 矩阵；

(3) 减小字长以及减小阈值。

小 结

同源序列一般是相似的，相似序列不一定是同源的。序列比较的基本操作是比对，两个序列的比对是指这两个序列中各个字符的一种一一对应关系，或字符的对比排列。“矩阵作图法”可以直观地发现比对序列是否存在匹配序列。相似性得分是以某种固定的记分规则规定两两字符比较的分数，然后将序列所有字符的比较得分相加得到两条序列的相似性得分。常用的核酸序列记分规则包括：等价矩阵、转换 - 颠换矩阵、BLAST 矩阵；常用的蛋白质序列替换记分矩阵包括：等价矩阵、遗传密码矩阵、疏水矩阵、PAM 矩阵、BLOSUM 矩阵。动态规划算法普遍应用于生物信息学中的序列比对。为了得到最佳比对，从得分最高的位置 S(i,j)开始，在得分矩阵的($i-1$,j)、(i,$j-1$)或($i-1$,$j-1$)位置中寻找下一个最大得分位置，记下路径，直至起点。

BLAST 是目前常用的数据库搜索程序，许多生物信息中心都提供基于 Web 的 BLAST 服务器。BLAST 程序是免费软件，可以从美国国家生物技术信息中心 NCBI 等文件下载服务器上获得。

Summary

Homology among proteins and DNA is often concluded on the basis of sequence similarity. In evolutionary biology, homology refers to any similarity between characteristics of organisms that is due to their shared ancestry. The aim of a sequence alignment is to match “the most similar elements” of two sequences. This similarity must be evaluated somehow. To quantify the similarity achieved by an alignment, scoring matrices are used: they contain a value for each possible substitution, and the alignment score is the sum of the matrix's entries for each aligned amino acid pair. The matrices of DNA composed of unitary matrix, transition-transversion matrix, BLAST matrix, *etc*. Protein sequences alignment usually use unitary matrix, genetic-code matrix, hydrophobic matrix, PAM and BLOSUM. The optimal alignment is the one which maximizes the alignment score.

BLAST (basic local alignment search tool) programs are used for sequence similarity identification. They identify regions of local alignment to assist in detecting relationships among sequences.

（刘建国 朱 浩 闫蓬勃）

习 题

1. 利用点阵图，完成下列两条序列的比对：

```
PQWIKMSTGG
QWISTGG
```

2. 蛋白质比对替换记分矩阵 BLOSUM、PAM 中序号有什么规律？
3. 对于题 1 中的序列，假设空位不罚分，使用 PAM250 作为替换记分矩阵，最佳比对得分是多少？
4. 遗传密码矩阵(GCM)的设计原理是什么？
5. 动态规划基本算法的空间复杂度是多少？
6. BLAST 搜索中 -e 的含义是什么？
7. 子序列与完整序列的比对需要对动态规划基本算法做哪些改动？

8. 登录NCBI主页，分别使用Non-redundant protein sequences(nr)数据库和Protein Data Bank proteins(pdb)数据库，其他条件均使用BLAST默许条件，搜索如下蛋白质序列：

```
PGKYTADGGKHVAYIIRSHVKDHYIFYCEGELHGKPVRGVKLVGRDPKNNLEALEDFEKAAGARGLSTESILIPRQSE
```

分析获得的序列，哪一个数据库搜索获得的匹配序列多？为什么？

主要参考文献

1. Needleman S. B., Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 1970,48 (3): 443-453.
2. Smith T. F. , Waterman M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.*, 1981, 147: 195-197.
3. Altschul S. F., Gish W, Miller W, et al. Basic local alignment search tool. *J. Mol. Biol.*, 1990, 215 (3): 403-410.
4. Lipman . Rapid and sensitive protein similarity searches. *Science*, 1985, 227 (4693): 1435-1441.
5. Pearson. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 1985, 85 (8): 2444-2448.
6. NCBI Education: http://www.ncbi.nlm.nih.gov/Education/
7. Wikipedia encyclopedia: http://en.wikipedia.org/wiki/Smith-Waterman_algorithm
8. The Stanford Folding@home glossary：http://www.stanford.edu/group/pandegroup/folding/education/h.html
9. 蛋白质数据库 HPDB: http://hpdb.hbu.cn
10. 孙啸，陆祖宏，谢建明. 生物信息学基础. 北京：清华大学出版社；2005年.

第三章 多序列比对
CHAPTER 3 MULTIPLE SEQUENCE ALIGNMENT

第一节 引 言
Section 1 Introduction

多序列比对(multiple sequence alignment)是两个以上DNA序列、RNA序列或蛋白质序列的比对。与双序列比对一样,多序列比对关心的是多个序列中哪些部分相似和哪些部分不同,包括哪些部分发生了什么样的替换、插入和删除。但有所不同的是,多序列比对能有效发掘多个序列中的相似性信息。当两个序列不能很好地比对并由此揭示序列的变化所蕴含的意义时,通过引入更多的序列,多序列比对可有效地使这两个原本难以直接比对的序列合理地关联起来。其次,多序列比对常常用于分析种系距离很大的多个序列,揭示这些序列中保守的和非保守的区段、保守区段的分布特征以及序列变化的进化趋势,这对于研究生物系统的进化是必不可少的。再者,许多预测RNA和蛋白质结构与功能的算法立足于相应的多序列比对,通过比较未知RNA和蛋白质的序列和已知RNA和蛋白质的序列预测未知RNA和蛋白质的结构与功能。因此,多序列比对是基因组分析和蛋白质组分析的最常用手段之一。

对高度相似的多个短序列,多序列比对或许可由手工完成,但如果序列间的距离较大、序列较多或序列较长时,多序列比对立刻成为一个相当困难的问题,因为记分可涉及复杂的替换矩阵、比对的时间开支十分可观和引入空格的数量与位置相当程度地取决于所用的比对方法和软件参数,这些也增加了解释比对结果的困难性。多序列比对主要涉及四个要素:①选择一组能进行比对的序列;②选择一个实现比对与记分的算法与软件;③确定软件的参数;④合理地解释比对的结果。

一、多序列比对具有广泛的应用

与双序列比对比较,多序列比对具有更广泛的重要应用,包括以下几个方面:

1. 获得共性序列 由多序列比对所得到的与所有序列距离最近的序列称为这些序列的共性序列(consensus sequence),共性序列这一特性使之常用于数据库搜索和芯片探针设计,用于识别具有高相似度的序列。

2. 序列测序 如果一个DNA或蛋白质序列被多个机构测序,则测序结果在某些核苷酸或氨基酸上可能存在差异,对这些测序结果进行全局多序列比对可发现这些差异之处,形成的共性序列理论上最为接近真实的序列。其次,对包含重叠区的多个测序序列进行局部多序列比对可发现这些重叠区,实现测序序列的拼接。另外,一个类似的应用是由表达序列标签(expressed sequence tag, EST)组装较长的重叠甚至完整的mRNA。

3. 突变分析 同一种系不同个体的基因组存在因突变而产生的差异,最常见的是单核苷酸多态性分析,它分析同一种系不同个体基因组中单个核苷酸包括置换、缺失和插入在内的变异。这些差异可通过多序列比对进行揭示。

4. 种系分析 相近种系动植物的基因和基因组由于源自共同的直接祖先而具有高度的相似性,

反之，远距种系动植物的基因和基因组由于源自不同的直接祖先而享有更少的相似性，这一事实使得多序列比对常常用于根据基因或基因组序列的差异判断种系关系。多序列比对通常是构造种系树的第一步。

5. 保守区段分析　基因组中功能不同的区段在进化中面对不同的选择压力(selective pression)，即重要的区段不易接受突变而非重要的区段易于接受突变。任何基因组都包含大量不同的在选择压力下保持进化上稳定的保守区段。首先，编码具有重要功能的蛋白质的基因是高度保守的，基因中的外显子尤其保守；其次，大量的基因调节单元，例如启动子和增强子，在不同种系中通常是高度保守的。此外，近年来发现许多非编码 RNA 也是非常保守的。多序列比对是找出进化上保守的这些区段的基本方法。

6. 基因和蛋白质功能分析　分子生物学和发育生物学实验是揭示基因和蛋白质功能的经典方法。在大量基因和蛋白质的功能得以揭示和更多基因和蛋白质的序列得以测定后，根据与功能已知的同源基因和蛋白质进行多序列比对来推断新基因和蛋白质的功能已成为越来越普遍的一个研究手段。

7. RNA 和蛋白质结构分析　类似地，可使用多序列比对考察种系相近的 RNA 和蛋白质家族，通过结构已知的 RNA 和蛋白质推断未知 RNA 和蛋白质的结构。需要注意的是，核苷酸序列和氨基酸序列的进化速度比 RNA 结构和蛋白质结构的进化速度要快，因此仅凭多序列比对仍难以确定 RNA 和蛋白质的结构。例如，人 β 球蛋白(beta-globin)和肌球蛋白(myoglobin)只有 25% 的氨基酸序列相同，但两者的三维结构却几乎相同。

8. 基因组结构分析　多序列比对可施用于整个基因组，揭示基因组的结构特征和进化特征。随着越来越多基因组的测序，多序列比对已频繁地用于基因组结构分析中，最典型的应用是 UCSC 基因组浏览器和 Ensembl 基因组浏览器。

二、多序列比对存在多种种类

从不同角度分类，存在着许多不同的多序列比对。从比对的数据看，存在着 DNA 序列、RNA 序列和蛋白质序列的比对，对于 DNA 序列，还存在基因序列的比对和非基因序列的比对；从比对的特点看，存在着揭示序列整体特性的全局比对和揭示序列局部特性的局部比对；从比对的方法看，存在着基于动态规划法的比对和基于其他方法的比对；从比对的规模看，存在着一个区段的比对和全基因组的比对。本节简要介绍这些不同比对的区别。

1. 蛋白质序列、RNA 序列和 DNA 序列比对　蛋白质序列和 DNA 序列存在明显的差别。首先，三个核苷酸编码一个氨基酸，且编码氨基酸的密码子存在冗余，这使得 DNA 序列的突变(包括替换、插入和删除)与氨基酸序列的突变没有必然和固定的联系。其次，内含子、5'UTR 和 3'UTR 等区段的突变不影响蛋白质编码。这些差别使得蛋白质的多序列比对和 DNA 的多序列比对具有明显不同的特点，某些方法主要用于蛋白质序列而另一些方法则主要用于 DNA 序列。另外，由于许多 RNA 分子具有特殊的次级结构，在进行 RNA 多序列比对时常常考虑已知 RNA 的次级结构，利用次级结构指导多序列比对。

2. 全局比对和局部比对　与双序列比对一样，多序列比对也有全局比对和局部比对。全局多序列比对把整个序列当作一个保守的区段，关心的是序列整体上的可比性和相似度，常常用于外显子序列、RNA 序列和蛋白质序列的比对。为了使序列中的每个字符都得到有效的比对，通常在序列中及序列两端都插入空格，使全部序列具有相同的长度。与全局比对具有明显区别的是局部比对，它们只关心序列中某个保守区段之间的可比性和相似度，通常不考虑序列的长度。局部比对常常用于揭示多个序列中的一个保守区段。

3. 比对包含一个保守区的序列和多个保守区的序列　上述经典的全局比对和局部比对只关注序列中的一个保守区，而越来越多的多序列比对需要处理包含多个保守区的超长序列甚至是整个基

因组序列，这种需求导致了许多多序列比对技术的发展。

4. 揭示种系关系的比对和揭示保守功能的比对　对于一组序列，多序列比对可揭示由进化而导致的序列上的差异并进而根据这些差异重构种系发生过程。因此，对进化上没有关联的序列进行比对一般是没有意义的，而当序列之间的同源性难以断定时使用多序列比对须十分小心。另一方面，在进化过程中存在大量的趋同进化和功能保守性。许多短小的功能区段在无共同祖先的序列中也表现出高度的相似性。例如，在基因调节区中的转录因子结合位点。多序列比对也常用于揭示这种具有功能保守性的区段。

第二节　相似性与距离、计分与罚分、替换矩阵

Section 2　Similarity and Distance, Scoring Matrix and Substitution Matrix

一、相似性与距离是序列相似性的两个主要度量

相似性与距离是两个定量描述多个序列相似程度的度量。使用相似性时，比对计分给出被比对序列间的相似程度，使用距离时，比对计分给出被比对序列间的差异程度。相似性既可用于全局比对也可用于局部比对，而距离一般仅用于全局比对，因为它反映了把一个序列转换成另一个序列所需的字符替换的耗费。在许多情况下这两种度量可通过一个公式相联系，从一个度量转到另一个度量。

使用相似性描述多个序列相似程度的基础是，通过在适当的位置插入空格可使序列中的相同字符对齐。以 k 个序列为例，若 s1，s2，…，sk 之间的比对是由对 s1，s2，…，sk 插入空格而得到的 k 个序列(s1，s2，…，sk)，则比对 A=(s1，s2，…，sk)必须满足：

(1) |s1'|=|s2'|=…=|sk'|；

(2) 移去 si' 中的所有空格得到 si；

(3) 对每个 i，s1[i]，s2[i]，…，sk[i]须有一个不是空格。

这里 |si'| 是 si' 的长度。如果用一个函数 score( )对 s1，s2，…，sk 中的每一对字符进行计分，处理匹配、失配和插入空格三种情况，则对 s1，s2，…，sk 的不同位置插入空格可产生不同的计分，而且对匹配、失配和插入空格进行不同的奖励和惩罚也会产生不同的计分。对于一个比对，不论使用什么计分函数进行计分，相似性被定义为总值等于最大的计分：

$$similarity(s_1,s_2,...s_k)=\max\sum_{i=1}^{|s_1'|}score(s_1'(i),s_2'(i),...s_k'(i)) \qquad \text{式 3-1}$$

使用距离描述多个序列相似程度的基础是，通过字符替换可使一个序列转变为另一个序列。如果在计算中对每个替换操作赋予一个耗费，便能定量地把多个序列之间的距离定义为将每个序列转换为一个共同序列所需的最小耗费。替换操作包括：

(1) 字符 a 替换成 b；

(2) 插入一个空格；

(3) 删除一个空格。

对于上述 k 个序列的例子，如果用一个函数 cos$t$( )对每一列的所有替换操作进行计分，则多个序列之间的距离等值于最小的计分：

$$distance(s_1,s_2,...s_k)=\min\sum_{i=1}^{|s_1'|}\cos t(s_1'(i),s_2'(i),...s_k'(i)) \qquad \text{式 3-2}$$

特别是，如果选用下面的简单函数计算两个字符间的计分：

$$\cos t(x,y)=\begin{cases}0 & \text{if } x=y\\ 1 & \text{if } x\neq y\end{cases} \qquad \text{式 3-3}$$

据此得到的序列间距离则称为编辑距离(edit distance)。如上所述，相似性和距离存在着内在的联系，其中一个用于双序列的将相似性与距离相关联的公式是：

$$similarity(s_1,s_2)+distance(s_1,s_2)=\frac{M}{2}(|s_1|+|s_2|) \qquad 式\ 3\text{-}4$$

这里 M 是一个参数。在某些渐进多序列比对中，计算距离的另一种方式是将相似性通过一个公式转换成距离。

二、存在多种方法对比对进行计分与罚分

接下来考察如何对多序列比对进行相似性计分与罚分。找出两个序列最长公共子序列(longest common subsequence，LCS)的算法假定了一个简单的计分方式，即若 v_i 和 w_j 匹配则奖励 1，若 v_i 和 w_j 不匹配则对插入的空格不作惩罚。为了进行一般化的能够对失配和空格进行不同惩罚的多序列比对，需要使用一个 $(d+1)\times(d+1)$ 的得分矩阵 δ，d 为字母表的大小。例如，若要对不同字符间的失配进行不同的罚分，对于 DNA 序列 d 必须取 4，对于蛋白质序列 d 必须取 20。这样，对于二个序列 v 和 w 比对的情形，比对中的每一个列 $\begin{pmatrix} x \\ y \end{pmatrix}$ 的得分用罚分函数计为 $\delta(x,y)$，对于序列 v 的前 i 个字符和序列 w 的前 j 个字符间的一个最优比对的得分 $s_{i,j}$ 其递归计算过程为：

$$s_{i,j}=\max\begin{cases} s_{i-1,j}+\delta(v_i,-) \\ s_{i,j-1}+\delta(-,w_j) \\ s_{i-1,j-1}+\delta(v_i,w_j) \end{cases} \qquad 式\ 3\text{-}5$$

若对失配和空格的罚分简单地定为常数 $-\mu$ 和 $-\sigma$，而匹配的得分为 +1，则有：

$$s_{i,j}=\max\begin{cases} s_{i-1,j}-\sigma \\ s_{i,j-1}-\sigma \\ s_{i-1,j-1}-\mu, \text{if } v_i\neq w_j \\ s_{i-1,j-1}+1, \text{if } v_i=w_j \end{cases} \qquad 式\ 3\text{-}6$$

与双序列比对相同，在多序列比对中必须对每一列产生一个计分值，然后累加全部列的计分得到整个比对的计分。但与双序列比对不同的是，在多序列比对中一列中不同位置的字符对列计分可能有不同的贡献，这样对一个含有 k 个字符的列进行计分需要一个带 k 个自变量的函数，每个自变量或者是字符或者是空格，而 k 个自变量的组合数量可达 $2k-1$ 之多。更进一步，当序列很长时，不同的列对总体计分可能有不同的贡献，例如，中间的列贡献大而两端的列贡献小，而差异性地考虑不同列的贡献同样大大提高了计算的复杂性。为了简化计算，人们常常采用两个假定。首先，score( ) 函数可以独立于自变量的次序。例如，如果一个列有字符 A，T，A，C，−；另一个列有字符 A，A，T，−，C；则它们应当具有相同的计分。其次，所有的列对计分有相同的贡献。满足这两个特性的一个函数是配对和(sum-of-pairs，SP)，它定义为列中所有字符的配对计分的和。对于上述含字符A，A，T，−，C 的例子，相应列的计分是：

SP-score(A，T，A，C，−) = score(A，T) + score(A，A) + score(A，C) + score(A，−) + score(T，A) + score(T，C) + score(T，−) + score(A，C) + score(A，−) + score(C，−)

对于最长序列长度为 n 的多序列比对，总的计分是：

$$similarity=\sum_{i=1}^{n} SP-score_i \qquad 式\ 3\text{-}7$$

SP 计分方法由于简单有效而被广泛使用(但后面将会提到，在 syntenic 比对里不同的列会对比对有不同的贡献)。需说明的一点是，尽管一列不能仅由空格构成，但一列中可以有两个或多个空格，因此当使用 SP 函数计算计分时需要一个对应于 score(−，−)的值，这在双序列比对中是没有的。

对于有 k 个序列的多序列比对，因为每列要计算 $k(k-1)/2$ 个配对比对的计分，SP 方法需要 O($k2$) 步计算每一列。这样，如果使用动态规划法，全部的运行时间是 O($k2nk2k$)。

三、精确计算失配计分需要使用核苷酸和氨基酸替换矩阵

相比于只由 4 个碱基组成的 DNA 序列，蛋白质序列可由 20 个氨基酸组成。构成蛋白质的氨基酸具有不同的生物化学特性，这些特性会影响它们在进化过程中的相互替换性。例如，与体积差异大的氨基酸相比，体积相似的氨基酸更易于彼此替换。另外，与水的亲和性也影响相互替换的概率。再者，生物学家已观察到天冬酰胺、天冬氨酸、谷氨酸和丝氨酸属于最容易突变的氨基酸，而半胱氨酸和色氨酸则属于最不易突变的氨基酸。例如，丝氨酸突变成苯丙氨酸的概率大约是色氨酸突变成苯丙氨酸的概率的三倍。因此，在比较蛋白质序列时，简单的计分系统(+1 表示匹配，0 表示失配，-1 表示空格)是不够的，必须使用一个能够充分反映氨基酸的相互替换性的计分系统。

影响氨基酸相互替换的因素很多，分析核苷酸如何编码氨基酸是揭示氨基酸相互替换性的手段之一，而由对实际替换率的直接观察可导出反映氨基酸相互替换概率的矩阵。Dayhoff 等研究了 34 个蛋白质家族，包括高度保守的和高度易突变的，根据对氨基酸之间相互替换率的计算得到 PAM 矩阵，即可接受点突变(point accepted mutation)或可接受突变百分比(percent of accepted mutation)矩阵。使用下面的公式：

$$s_{i,j}=10\times\log\frac{PAM_{i,j}}{p_i} \qquad \text{式 3-8}$$

可把 PAM 矩阵转换为一个蛋白质比对计分矩阵，这里 PAMi，j 是 PAM 矩阵中的反映氨基酸 j 被氨基酸 i 替换的概率的单元，pi 是氨基酸 i 在全部蛋白质中出现的频率($\sum_{i=1}^{20}p_i=1$)。PAM 实际上是一个包含多个矩阵的家族，需要不同 PAM 矩阵的原因是，当考虑被比较序列之间的进化距离时，PAM 矩阵必须是这个距离的函数。例如，1-PAM 矩阵用平均每百个氨基酸发生一个突变作为进化单位；当对两个进化距离为 250 单位的序列进行比较时，应使用 250-PAM 而非 1-PAM 矩阵。除了 PAM 系列，Blosum 系列替换矩阵也常用于蛋白质序列的比对。同源性高的序列间的比对使用大数字的 Blosum 矩阵，如 Blosum90，而同源性低的序列间的比对使用小数字的 Blosum 矩阵，如 Blosum45。

四、记分方法可显著影响多序列对比

记分方法以及相应的得分和罚分参数可显著影响多序列比对的记分及比对结果。以上述的 SP 记分为例，全局多序列比对根据 SP 记分最大化匹配字符的数量或最小化失配字符的数量，以此调整列之间的比对，这使得记分方法以及得分和罚分参数对比对产生显著影响。假定六个序列中的三个列如图 3-1 所示，第一列全部由 T 组成，第二列由 5 个 T 和 1 个 C 组成，第三列由 4 个 T 和 2 个 C 组成。如果 T-T 匹配得分 6，T-C 匹配罚分 3，则列 A 的 SP 记分为 90，列 B 的 SP 记分为 45，列 C 的 SP 记分为 9。如果 T-T 匹配得分 6，T-C 匹配罚分 1，则列 A 的 SP 记分为 90，列 B 的 SP 记分为 55，列 C 的 SP 记分为 27。这三个列的得分说明 SP 记分具有两个特性：第一，当一列中出现个别失配时，这些失配会急剧减小整列的记分；第二，当选用不同的匹配得分和失配罚分时，它们对整列的记分也有显著的影响。个别失配对整列记分产生如此大的影响是不符合实际的。为了修正这一点，人们提出了其他的记分方法，它们的基本特点是控制个别失配对整列记分的影响。

五、多序列对比的困难性

多序列比对除了费时，还面临其他的困难性。由于多序列比对涉及多个序列中的匹配、失配、插入、删除和变异，对比对进行合理的记分进而根据记分找到一个最优的比对是一件困难的事。仅当

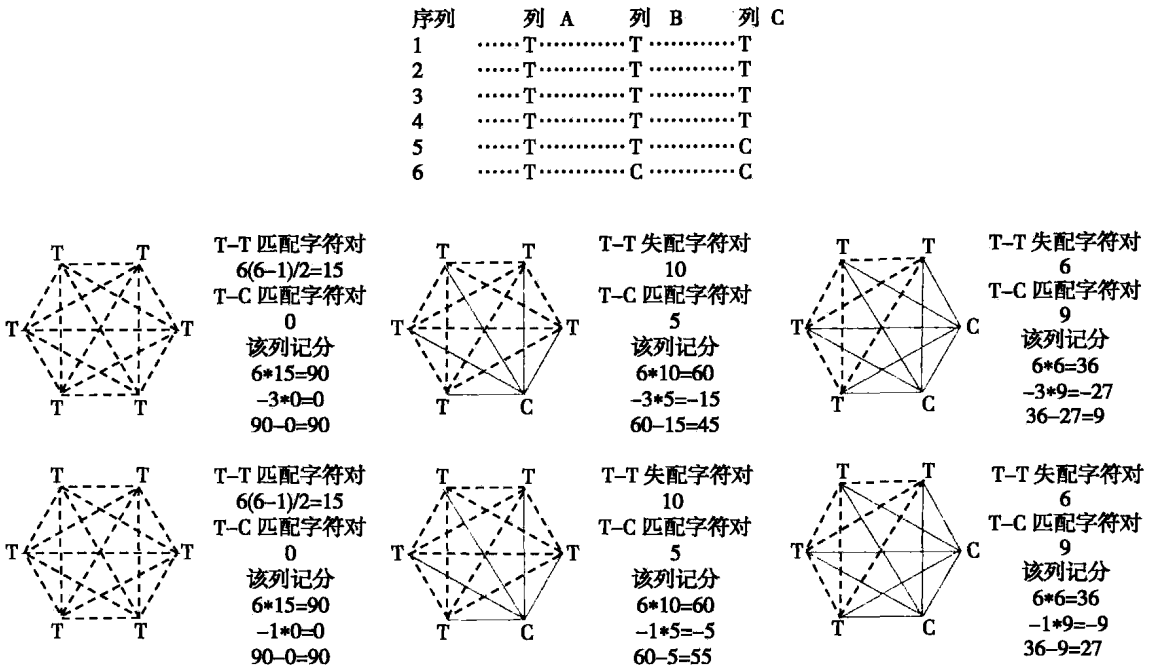


图 3-1 SP 记分及得分和罚分参数对多序列比对有显著影响

选用的记分方法合理时比对结果才有合理性。当多个序列的相似度较高时，记分相对容易，不同的比对方法可产生相差不大的结果；当多个序列的差异度较高时，不同的记分和不同的比对可产生完全不同的比对结果，使对比对结果的评估成为一件困难的事。由于多序列比对在序列数量和序列相似性上会有极大的差异，人们设计了大量的不同算法有针对性地处理不同的多序列比对。对大量蛋白质序列的比对提示，对于氨基酸序列 40% 一致的蛋白质，大多数多序列比对算法产生相似的结果，但对于相距更遥远和一致性更低的蛋白质序列，不同的多序列比对算法可产生非常不同的结果。至于评估哪些算法、参数和结果更合理，对于蛋白质和 RNA 序列，一个办法是比较多序列比对的结果和三维结构比对的结果，对于两者相吻合的，其相应的算法、参数和结果较合理。这种比较需要利用蛋白质结构数据库，且不能检验所有参与比对的序列。

第三节 主要比对方法与软件

Section 3 Methods and Softwares of Multiple Alignment

一、动态规划法

由于动态规划法的时间与空间复杂性太高，人们发展了该算法的多种变体使得它们能够在合理的时间内找到优化比对。变体之一是 S. Altschul 等在 1989 年引入的一个算法，它能极大地缩小 k 维动态规划表的搜索空间，其中心思想如下：首先，对 k 个序列的 $\binom{k}{2}$ 个配对按动态规划法进行配对比对，由于一个 k 序列比对对应于 k 维空间动态规划表中的一个路径(图 3-1)，这些配对比对可看作是 k 维空间中的这个路径在不同的二维空间中的投影。其次，在相应的二维空间中，可以限制投影所可能历经的空间，从而限制在原始的 k 维空间中寻找优化多序列比对历经的路径。第三，每个投影定义了原始 k 维空间的一个子空间，这些子空间的交汇包含了 k 个序列的优化比对。该算法通常采用 SP 函数计分，并使用动态规划法搜索子空间的交汇来找到多序列比对在 k 维空间中的路径。一个关键点是，需要确定一个将多序列比对投影成配对比对的开支上限，该开支上限的选择应能保证动态

规划法找到 k 个序列的最优比对。在使用启发式方法确定配对比对的开支上限时，若比对的质量表明开支上限不够大，则应增大开支上限。不过，一味地增大开支上限并不能持续提高多序列比对的质量。由于该动态规划法的变体本质上属于启发式算法，它有一切启发式算法所固有的缺陷，即当开支上限定的过小时它可能找不到最优比对，而当开支上限定的过大时它所耗费的时间可能与标准动态规划法相差无几。

基于这个算法，S.Altschul 等开发了多序列比对软件 MSA。MSA 使用 SP 计分方法，对空格按常数罚分，可对不同配对比对的记分赋予一个权值以调校它们对多序列比对的贡献。S. Gupta 等在 1995 年对该软件进行了改进，优化了内存使用和时间开销。MSA 用 C 语言写成，运行于 UNIX 系统，其内存要求和时间开销取决于序列个数、序列长度和配对比对的开支上限这一关键参数。比对的主要步骤包括：

（1）计算所有配对比对的记分；
（2）使用配对比对记分构造一棵树；
（3）根据构造的树计算配对比对的权值；
（4）根据构造的树产生一个多序列比对；
（5）计算配对比对在多序列比对中的贡献；
（6）确定优化多序列比对的空间结构；
（7）执行优化的比对。

为了清晰显示不同方法和不同参数所产生的多序列比对的差异，本节给出了六种动物里发状分裂相关增强子 -5(hairy and enhancer slipt 5)蛋白质序列的多序列比对，这些序列在 NCBI 中的访问号(Accession Number)是 NP_001010926(人 Homo sapiens)、NP_034549.1(小鼠 Mus musculus)、NP_062109(大鼠 Rattus norvegicus)、NP_001012713.1(鸡 Gallus gallus)、NP_001037974(非洲蟾蜍 Xenopus tropicalis)和 AAP41832(斑马鱼 Danio rerio)，FASTA 格式的输入数据文件是：

```
>gi|58219048|ref|NP_001010926.1| hairy and enhancer of split 5 [Homo sapiens]
MAPSTVAVELLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPNSKLEKADILEMAVSYLK
HSKAFVAAAGPKSLHQDYSEGYSWCLQEAVQFLTLHAASDTQMKLLYHFQRPPAAPAAPAKEPKAPGAAP
PPALSAKATAAAAAAHQPACGLWRPW
>gi|6754182|ref|NP_034549.1| hairy and enhancer of split 5 [Mus musculus]
MAPSTVAVEMLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPNSKLEKADILEMAVSYLK
HSKAFAAAAGPKSLHQDYSEGYSWCLQEAVQFLTLHAASDTQMKLLYHFQRPPAPAAPAKEPPAPGAAPQ
PARSSAKAAAAAVSTSRQPACGLWRPW
>gi|9506775|ref|NP_062109.1| hairy and enhancer of split 2 [Rattus norvegicus]
MRLPRGVGDAAELRKSLKPLLEKRRRARINESLSQLKGLVLPLLGAETSRYSKLEKADILEMTVRFLREQ
PASVCSTEAPGSLDSYLEGYRACLARLARVLPACSVLEPAVSARLLEHLRQRTVSGGPPSLTPASASAPA
PSPPVPPPSSLGLWRPW
>gi|60593016|ref|NP_001012713.1| hairy and enhancer of split 5 [Gallus gallus]
MAPSALSLEILTPKEKNRLRKPIVEKLRRDRINSSIEQLKLLLEKEFQRHQPNSKLEKADILEMTVSYLK
YSRAFAASAKSLQQDYCEGYAWCLKEALQFLSLHSANTETQMKLICHFQRSQAMPKDSGSPSASTSTHQP
SAKQTPVKPSCNLWRPW
>gi|113205884|ref|NP_001037974.1| hairy and enhancer of split 5, gene 2 [Xenopus tropicalis]
MAPSTDFLDQQKMTPKEKNKLRKPVVEKMRRDRINSSIEQLKGLLETVFHKQQPNVKLEKADILEMTVTY
LRQQTLQIKSEIPHNNDIQMDYKDGYSRCFEEVIDFLSLHQKQPETAKLISHFHSKATASSISSFPIRCS
QSKTANGTGSSSSLWRPW
```

```
>gi|31074173|gb|AAP41832.1| hairy and enhancer of split 5 [Danio rerio]
MAPAYMTEYSKLSNKEKHKLRKPVVEKMRRDRINNCIEQLKSMLEKEFQQQDPNAKLEKADILEMTVVFL
KQQLRPKTPQNAQIEGYSQCWRETISFLSVGSEAVAQRLQQEAQRSAAPELTHTSEAPHQQHTHIKQEPR
AHAPLWRPW
```

使用缺省参数 MSA 比对的结果如下：

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MAPST--VAV-ELLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPN-SKLEKADILEMAV
MAPST--VAV-EMLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPN-SKLEKADILEMAV
MR-----LPR-GVGDAAELRKSLKPLLEKRRRARINESLSQLKGLVLPLLGAETSRYSKLEKADILEMTV
MAPSA--LSL-EILTPKEKNRLRKPIVEKLRRDRINSSIEQLKLLLEKEFQRHQPN-SKLEKADILEMTV
MAPSTDFLDQ-QKMTPKEKNKLRKPVVEKMRRDRINSSIEQLKGLLETVFHKQQPN-VKLEKADILEMTV
MAPAY--MTEYSKLSNKEKHKLRKPVVEKMRRDRINNCIEQLKSMLEKEFQQQDPN-AKLEKADILEMTV

12345678901234567890123456789012345678901234567890123456789012345678901234567890
SYLKHSKAFVAA--AGPKSLHQDYSEGYSWCLQEAVQFLTLHA--ASDTQMKLLYHFQRPPAAPAAPAKE
SYLKHSKAFAAA--AGPKSLHQDYSEGYSWCLQEAVQFLTLHA--ASDTQMKLLYHFQRPPA-PAAPAKE
RFLREQPASVCS--TEAPGSLDSYLEGYRACLARLARVLPACSVLEPAVSARLLEHLRQRTV-------S
SYLKYSRAFA----ASAKSLQQDYCEGYAWCLKEALQFLSLHS-ANTETQMKLICHFQRSQA-------M
TYLRQQTLQIKSEIPHNNDIQMDYKDGYSRCFEEVIDFLSLHQ--KQPETAKLISHFHSKAT--------
VFLKQQ--------LRPKTPQNAQIEGYSQCWRETISFLSVGS---EAVAQRLQQEAQRSAA--------

1234567890123456789012345678901234567
PKAPGAAPPPALSAKATAAAAAA--HQPACGLWRPW
PPAPGAAPQPARSSAKAAAAAVSTSRQPACGLWRPW
GGPPSLTPASASAPAPSPPVPPP----SSLGLWRPW
PKDSGSPSASTSTHQPSAKQTPV---KPSCNLWRPW
--ASSISSFPIRCSQSKTANGTG----SSSSLWRPW
-PELTHTSEAPHQQHTHIKQEPR----AHAPLWRPW
```

二、渐进多序列比对

渐进多序列比对(progressive multiple alignment)以及后面介绍的许多多序列比对方法都基于使用动态规划法的配对比对。渐进比对的思想最早由 W.M. Fitch 和 K.T. Yasunobu 在 1975 年提出，P. Hogeweg 和 B. Hesper 于 1984 年首先将其用于 5S 核糖体 RNA 序列的比对，D.F. Feng 和 R.F. Doolittle 在 1987 年的改进使之广为普及，是得到了最广泛使用的多序列比对。渐进多序列比对首先使用动态规划法构造全部 k 个序列的 $\binom{k}{2}$ 个配对比对，然后以计分最高的配对比对作为多序列比对的种子，按计分高低依次选择序列，逐渐向已构造的多序列比对中加入序列，形成一个树状结构的多序列比对结果。该方法的优点是允许高达数百个序列的比对，缺点是因为最终的结果取决于序列加入的次序，比对的最优性不受保证。渐进多序列比对需要三个步骤：第一，使用动态规划法构造每个序列的配对比对，包括 Cluster W 在内的许多比对算法在这一步使用距离矩阵而不是相似性矩阵来

描述序列间的关联性；第二，由距离矩阵构造一棵指导树(guide tree)，树的两个主要特征是拓扑结构和分支长度，它一般并不被当作是种系树，只反映了参与比对的多个序列如何相关联，用来确定向正在进行的多序列比对加入新序列的次序；第三，以计分最高的配对比对作为多序列比对的种子，根据指导树逐渐向多序列比对中加入序列。需说明的是，在添加序列的过程中需要对序列加入空格，而存在许多引入空格的方法。D.F. Feng 和 R.F. Doolittle 的方法遵循了“一旦引入一个空格则始终保持这个空格”的原则，其合理性在于在配对比对阶段得到比对的两个最接近的序列在决定空格方面应该被赋予更重的权值。Cluster W 还根据空格的不同位置动态决定罚分，使得整个比对形成由块状结构构成的特征，并能有效地使空格数量最少化。

多序列比对的一个特殊问题是对插入和缺失(合称为“插缺” indel)的罚分。双序列比对不区别序列中的插入和缺失，但多序列比对因常常用于种系分析或以种系分析为基础，对序列中的插入和缺失需进行专门的处理。在一般的多序列比对算法里，一个缺失被罚分一次，但一个插入可被过度地罚分多次，这样产生的空格罚分要么太高以至于长的空格从不出现，要么太低以至于序列被许多小空格打散，而所产生的多序列比对常常包含偏少的空格。A.Loytynoja 和 N.Goldman 在 2005 年给出一个算法来区分插入和缺失并校正插入的罚分，他们的算法产生的多序列比对通常包含较多的空格，但更准确，尤其适用于一般 DNA 序列的比对。

渐进多序列比对是一种启发式算法。如前所述，所有的启发式算法都有一个共同的问题，即不保证产生全局最优的比对。首先，渐进多序列比对可能会被一些伪强的、实际上是坏的种子所误导。如果一开始选择的两条序列的配对比对与实际上的最优多序列比对不一致，那么初始的配对比对中的错误在整个多序列比对构造过程中将始终存在并持续传播。其次，在比对的任何阶段出现失配时(例如在配对比对中加入空格)，这些失配不是被纠正而是被传播到最终结果。再者，一个更糟糕的情况是配对比对可能无法组成一个相容的多序列比对(图 3-2)。以上因素使得渐进多序列比对对于距离非常接近的序列效果很好，而当序列间的距离较远时效果不佳。后期的渐进多序列比对软件对这些缺陷进行了改进。

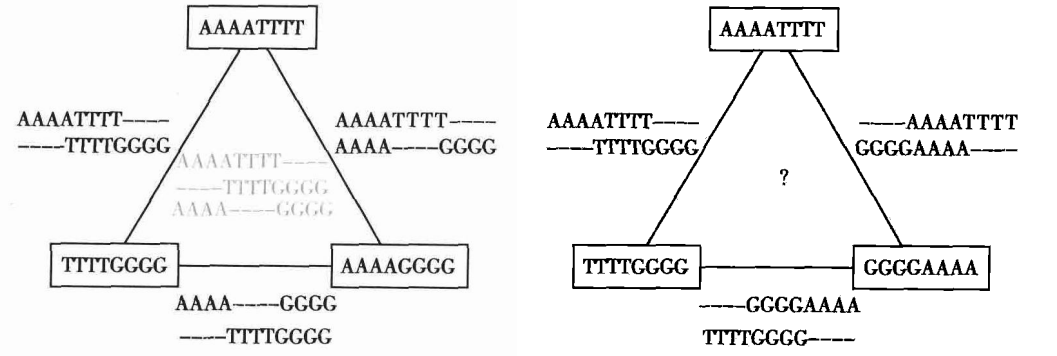


图 3-2 三个序列的配对比对未必总能组合成一个多序列比对

对于接近或超过 100 个序列的多序列比对，渐进多序列比对具有较高效率。最流行的渐进多序列比对软件是 Cluster 家族，最早由 D.G. Higgins 等于 1988 年开发。1994 年他们进一步改进了 Cluster，首先，与 MSA 中的做法类似，在比对中对每个序列赋予一个特殊的权值以降低高度近似序列的影响和提高相距遥远的序列的影响(图 3-3，来自于根的序列其权值等于它到根的距离，而来自于一个分支的序列其权值不超过其到根的距

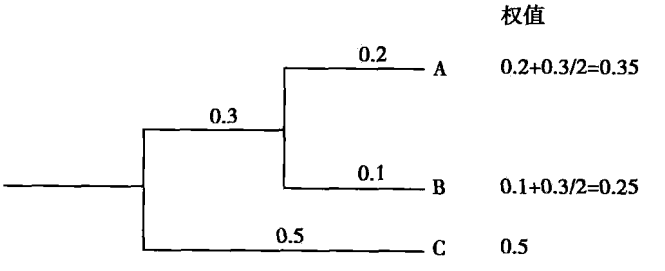


图 3-3 Cluster W 中对序列赋权的方法

离的和),它能更准确地反映在进化中序列所产生的变化;其次,根据序列间进化距离的离异度(divergence)在比对的不同阶段使用不同的氨基酸替换矩阵;第三,采用了与特定氨基酸相关的缺口罚分函数,对亲水性氨基酸区域中的缺口予以较低的罚分;第四,对在早期配对比对中产生缺口的位置进行较少的罚分,对引入缺口和扩展缺口进行不同的罚分。根据这些改进发展的软件称作 Cluster W,它有许多实现和在线服务版本,不同的实现和在线服务版本可能在用户界面和参数设置上略有差异,欧洲生物信息学研究所(EBI)的 Cluster W 在线服务的界面如图 3-4 所示。前述六个动物 hairy and enhancer of split 5 蛋白质序列使用 Cluster W 比对的结果如下(PAM100+ 其余缺省参数):

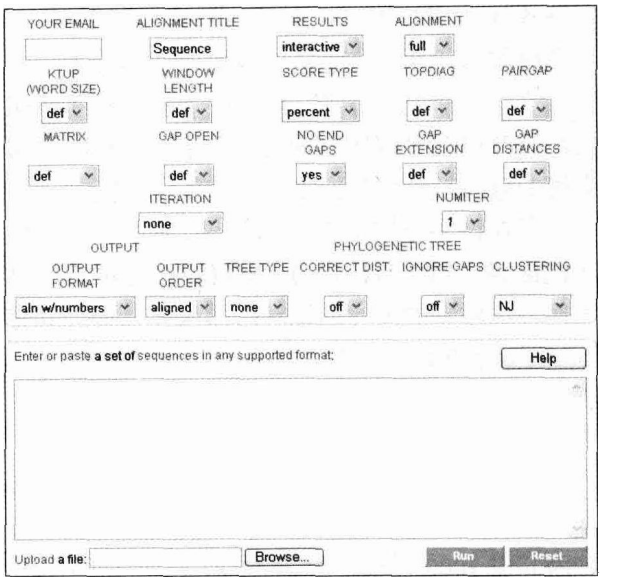


图 3-4 EBI 站点 ClusterW 在线服务的页面

```
1234567890123456789012345678901234567890123456789012345678901234567890
--MAPSTVAVELLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPNSKLEKADILEMAVSY
--MAPSTVAVEMLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEFARHQPNSKLEKADILEMAVSY
-----MRLPRGVGDAAELRKSLKPLLEKRRRARINESLSQLKGLVLPLLGAETSRYSKLEKADILEMTVRF
--MAPSALSLEILTPKEKNRLRKPIVEKLRRDRINSSIEQLKLLLEKEFQRHQPNSKLEKADILEMTVSY
MAPSTDFLDQQKMTPKEKNKLRKPVVEKMRRDRINSSIEQLKGLLETVFHKQQPNVKLEKADILEMTVTY
-MAPAYMTEYSKLSNKEKHKLRKPVVEKMRRDRINNCIEQLKSMLEKEFQQQDPNAKLEKADILEMTVVF

1234567890123456789012345678901234567890123456789012345678901234567890
LKHSKAFVAAAGPKSLHQDYSEGYSWCLQEAVQFLTLHAASDTQMKLLYHFQRPPAAPAAPAKEPKAPGA
LKHSKAFAAAAGPKSLHQDYSEGYSWCLQEAVQFLTLHAASDTQMKLLYHFQRPPAPAAPAKEPPAPGAA
LREQPASVCSTEAPGSLDSYLEGYRACLARLARVLPACSVLEPAVSARLLEHLRQRTVSGGPPSLTPASA
LKYSRAFAASAKSLQQDYCEGYAWCLKEALQFLSLHSANTETQMKLICHFQRSQAMPKDSGSPSASTSTH
LRQQTLQIKSEIPHNNDIQMDYKDGYSRCFEEVIDFLSLHQKQPETAKLISHFHSKATASSISSFPIRCS
LKQQLRPKTPQNAQIEGYSQCWRETISFLSVGSEAVAQRLQQEAQRSAAPELTHTSEAPHQQHTHIKQEP

12345678901234567890123456789
APPPALSAKATAAAAAAHQPACGLWRPW-
PQPARSSAKAAAAAVSTSRQPACGLWRPW
SAPAPSPPVPPPSSLGLWRPW--------
QPSAKQTPVKPSCNLWRPW----------
QSKTANGTGSSSSLWRPW-----------
RAHAPLWRPW-------------------
```

三、迭代法

在渐进多序列比对中,一个序列一经加入构造的比对结果其配对比对便不再重新处理,因此对在渐进比对过程中发现的错误或不适当的记分没有机会进行更正,这提高了比对的运行效率但牺牲

了准确性。当起始的比对处理是较远距离的序列时，其蕴含的错误对多序列比对的影响尤其严重。一类称作迭代法的方法能够克服渐进多序列比对的这个不足。迭代法的基本过程是先用渐进多序列比对产生一个初始结果，再对序列的不同子集进行反复比对并利用这些结果重新进行多序列比对，目标是改进多序列比对的总计分值。迭代法常常使用随机搜索或者通过对比对结果进行重排来寻找更优的解，迭代持续至比对记分值不再提高。

存在许多不同的迭代法软件，例如，分别用于核酸和蛋白质序列的 PRRN 和 PRRP 使用爬山算法优化多序列比对的计分，反复修正比对的权值和局部差异区。当 PRRP 用于优化一个先前用某个快速方法建立的多序列比对时效果尤其好。其他的软件还包括 MAFFT(multiple alignment using fast fourier transform)和 PRALINE(PRofile ALIgNEment)，它们两者都能利用同源序列的信息增进比对的质量，MAFFT 还设置了让用户在速度和准确性之间进行取舍的参数，并允许在配对比对中选用全局或局部比对。

自 2004 年，MUSCLE(multiple sequence alignment by log-expectation)由于其准确性和出色的速度而成为一个流行的用于大量序列多序列比对的软件。据报道，使用桌面计算机 MUSCLE 可以在 21 秒内完成 1000 个长度为 282 的蛋白质序列的比对。MUSCLE 的方法分为三个步骤：首先，使用渐进多序列比对产生一个初始结果，其中含有根据每对序列的相似性计分构造的一棵指导树；其次，重新计算相似性计分，据此改进指导树并再用渐进多序列比对产生一个更新的结果，这一过程迭代地进行；再次，算法根据新计算的 SP 计分值是否增加而决定是接受还是拒绝新产生的比对结果。据报道，与 MAFFT 和 Cluster W 相比，MUSCLE 在包括 BAliBASE 在内的多个基准测试(benchmarking)数据库上表现出更高的准确性。使用 MUSCLE 以及缺省参数对上述六个蛋白质序列的比对结果如下：

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MAPST--VAVELLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEF-ARHQPNSKLEKADILEMAVS
MAPST--VAVEMLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEF-ARHQPNSKLEKADILEMAVS
MR-----LPRGVGDAAELRKSLKPLLEKRRRARINESLSQLKGLVLPLLGAETSRYSKLEKADILEMTVR
MAPSA--LSLEILTPKEKNRLRKPIVEKLRRDRINSSIEQLKLLLEKEF-QRHQPNSKLEKADILEMTVS
MAPSTDFLDQQKMTPKEKNKLRKPVVEKMRRDRINSSIEQLKGLLETVF-HKQQPNVKLEKADILEMTVT
MAPAY-MTEYSKLSNKEKHKLRKPVVEKMRRDRINNCIEQLKSMLEKEF-QQQDPNAKLEKADILEMTVV

12345678901234567890123456789012345678901234567890123456789012345678901234567890
YLKHSKAFVAAAGPKS--LHQDYSEGYSWCLQEAVQFLTLHAA--SDTQMKLLYHFQRPPAAPAAPAKEP
YLKHSKAFAAAAGPKS--LHQDYSEGYSWCLQEAVQFLTLHAA--SDTQMKLLYHFQRPP-APAAPAKEP
FLREQPASVCSTEAPG--SLDSYLEGYRACLARLARVLPACSVLEPAVSARLLEHLRQRT-VSGGPPSLT
YLKYSRAFAASA--KS--LQQDYCEGYAWCLKEALQFLSLHSA-NTETQMKLICHFQRSQ----AMPKDS
YLRQQTLQIKSEIPHNNDIQMDYKDGYSRCFEEVIDFLSLHQK--QPETAKLISHFHSKA----------
FLKQQ------LRPKT--PQNAQIEGYSQCWRETISFLSVGSE--AVAQ-----RLQQEAQRSAAP--EL

123456789012345678901234567890123456789012345
KAPGAAPPPALSAKATAAAA--AAHQPACGLWRPW
PAPGAAPQPARSSAKAAAAAVSTSRQPACGLWRPW
PASASAPAPSPP----------VPPPSSLGLWRPW
GSPSASTSTHQPSAKQ------TPVKPSCNLWRPW
TASSISSFPIRCSQSKTA----NGTGSSSSLWRPW
THTSEAPHQQHTHIK-------QEPRAHAPLWRPW
```

四、基于一致性的方法

渐进多序列比对的基本方法是先产生全部的配对比对，然后根据配对比对的计分高低逐渐构造多序列比对。基于一致性的方法采用了另一种利用序列信息的方式。这里，一致性指的是对于序列x、y和z，如果x_i比对于z_k且z_k比对于y_j，则x_i应比对于y_j。因此，基于一致性的方法的基本特点是充分利用多个序列间的比对信息对配对比对进行更合理的计分。例如，根据x_i和y_j同时比对于z_k而调整x_i和y_j的比对计分，如果序列x中的字符x_i比对于序列y中的字符y_j的似然率(likelihood)为：

$$P(x_i \sim y_j \mid x, y) \qquad \text{式 3-9}$$

则有：

$$P(x_i \sim y_j \mid x, y, z) \approx \sum_k P(x_i \sim z_k \mid x, z) P(y_j \sim z_k \mid y, z) \qquad \text{式 3-10}$$

基于一致性的方法在多序列比对中对每对序列中的每对字符计算如上的似然率。根据基准测试数据的研究，基于一致性方法的多序列比对产生的结果经常比渐进多序列比对产生的结果更准确。

这里介绍一个基于一致性的多序列比对软件ProbCons(Probabilistic Consistency-based Multiple Alignment)。ProbCons分五步进行蛋白质多序列比对：第一，对每对序列中的每对字符计算上述的似然率，得到一个似然率矩阵；第二，用动态规划法计算每个配对比对的预期精度(expected accuracy)，它是得到正确比对的字符数除以较短序列的长度，计分根据上述条件概率公式计算而不采用通常的PAM或BLOSUM矩阵，且空格罚分设为0；第三，根据相关条件概率的计算重新调整配对比对的计分，这一步用到了由多个配对比对揭示的序列中字符的保守性，产生更准确的配对比对的计分；第四，用分层聚类法(hierarchical clustering)构造一棵基于相似性而不是距离的期望准确性指导树；第五，根据该期望准确性指导树对所有的序列进行渐进性比对，方法如同Cluster W。在这些步骤之后，还可进一步用迭代法进行优化。据报道，在BAliBASE和PREFAB等多个基准测试数据库上，ProbCons的性能优于许多多序列比对软件，包括Cluster W、DIALIGN、T-Coffee、MAFFT、MUSCLE和Align-m等。使用ProbCons以及缺省参数对上述六个蛋白质序列的比对结果如下：

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MAPSTV--AVELLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEF-ARHQPNSKLEKADILEMAVS
MAPSTV--AVEMLSPKEKNRLRKPVVEKMRRDRINSSIEQLKLLLEQEF-ARHQPNSKLEKADILEMAVS
MRLPR-----GVGDAAELRKSLKPLLEKRRRARINESLSQLKGLVLPLLGAETSRYSKLEKADILEMTVR
MAPSAL--SLEILTPKEKNRLRKPIVEKLRRDRINSSIEQLKLLLEKEF-QRHQPNSKLEKADILEMTVS
MAPSTDFLDQQKMTPKEKNKLRKPVVEKMRRDRINSSIEQLKGLLETVF-HKQQPNVKLEKADILEMTVT
MAPAYM-TEYSKLSNKEKHKLRKPVVEKMRRDRINNCIEQLKSMLEKEF-QQQDPNAKLEKADILEMTVV

12345678901234567890123456789012345678901234567890123456789012345678901234567890
YLKHSKAFVAA-AGP--KSLHQDYSEGYSWCLQEAVQFLTLHAAS--DTQMKLLYHFQRPPAAPAAPAKE
YLKHSKAFAAA-AGP--KSLHQDYSEGYSWCLQEAVQFLTLHAAS--DTQMKLLYHFQRPPAPAA-PAKE
FLREQPASVCSTEAP--GSL-DSYLEGYRACLARLARVLPACSVLEPAVSARLLEHLRQRTVSGG-PPSL
YLKYSRAFAAS---A--KSLQQDYCEGYAWCLKEALQFLSLHSAN-TETQMKLICHFQRSQAMPK-DSGS
YLRQQTLQIKS-EIPHNNDIQMDYKDGYSRCFEEVIDFLSLHQKQ--PETAKLISHFHSKATASS-ISSF
FLKQQLR-------P--KTPQNAQIEGYSQCWRETISFLSVGSEA---VAQRLQQEAQRSAA-PE-LTHT

123456789012345678901234567890123456
PKAPGAAPPPALS-AKAT-AAAAAAHQPACGLWRPW
```

```
PPAPGAAPQPARSSAKAAAAAVSTSRQPACGLWRPW
TPASASAPAP----------SPPVPPPSSLGLWRPW
PSASTSTHQP---------SAKQTPVKPSCNLWRPW
PIRCSQS-------------KTANGTGSSSSLWRPW
SEAPHQQHT----------HIKQEPRAHAPLWRPW
```

此外，关于多序列比对的算法还包括隐马尔可夫模型、遗传算法、树比对和星比对、基于结构的比对和 RNA 比对等算法。

知 识 拓 展

在本章所讲的多序列比对中，大部分算法都只针对序列的一维内容而不考虑序列可形成的二维结构以及这种二维结构与序列保守性的关系。实际上，大部分 RNA 和氨基酸序列形成特定的二维与三维结构。由于氨基酸由特定的编码产生且担负重要的功能，一维内容的变化和高维结构的变化基本一致，使得保守性、同源性和功能域能比较可靠地由序列比对得到揭示。但 RNA 则相当不同，由于补偿性突变的存在，它同时存在一级序列的高替换性和二级结构的高保守性。补偿性突变指的是当一个核苷酸突变后，与之配对的核苷酸也发生相应的突变，使配对继续得以保持。由于配对的两个核苷酸在序列上可能相距遥远，大量积累的补偿性突变使得许多 RNA 分子在一级结构上呈现高度的变异而同时保留了保守的二级结构。因为许多 RNA 的功能是由其二级结构而非一级序列承担的，通过有效的比对揭示 RNA 的保守性、同源性和功能域是目前是生物信息学的一个重要内容。RNA 分子比对的发展方向是融合结构比对和序列比对，融合的方式有多种。其一是先做序列比对再进行结构比对，由后者改进前者；其二是同时进行结构比对与序列比对，这类算法通常是高度费时的。在结构比对方面，同样存在局部比对与全局比对，且与 BLAST 类似，也出现了基于结构类似性的数据库搜索。近年报道的算法与软件数量呈快速增长，但性能尚欠理想，主要的挑战是在性能与时间 / 空间开销上取得平衡。

五、多序列比对结果编辑器

为了能对多序列比对的结果进行彩色显示和手工编辑，人们开发了各种多序列比对结果编辑器。常用的包括 CINEMA(Colour Interactive Editor for Multiple Alignment)、GDE(Genetic Data Environment)和 GeneDoc，而 MACAW(Multiple Alignment Construction & Analysis Workbench)是一个兼具序列编辑和局部比对的工具。一个 CINEMA 的例子见图 3-5，其下载地址是 www.bioinf.manchester.ac.uk/dbbrowser/CINEMA2.1/。

图 3-5 CINEMA 显示多序列比对结果

第四节 局部比对、glocal 比对和 syntenic 比对

Section 4 Local, Glocal and Syntenic Alignment

一、局部比对

前面介绍的比对多个基因和蛋白质序列的方法大部分都是全局比对，其共同特征是序列中所有对应字符均假定可以匹配，所有字符具有同等的重要性，空格的插入是为了使整个序列得到比对，包括使两端对齐。因此，全局比对适合于比对高度相似且长度相当的序列，相应的基本动态规划法算法是 Needleman-Wunsch 算法。与之不同的是，局部比对不假定整个序列可以匹配，重在考虑序列中能够高度匹配的一个区段，可赋予该区段更大的计分权值，空格的插入是为了使高度匹配的区段得到更好的比对，相应的基本动态规划法算法是 Smith-Waterman 算法。当多个序列长度相当且高度相似时，全局比对和局部比对给出基本相同的结果；否则，全局比对和局部比对可产生非常不同的结果（图 3-6，双序列比对例子），对多序列比对同样如此。

```
--T--CC-C-AGT--TATGT-CAGGGGACACG--A-GCATGCAGA-GAC
  |  ||| ||  | | | |||     || | | |  | ||||  |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG--T-CAGAT--C

          tccCAGTTATGTCAGgggacacgagcatgcagaga
                ||||||||||||
aattccgccgtcgttttcagCAGTTATGTCAGatc
```

图 3-6 对二个序列进行全局和局部比对可得到完全不同的结果

有些多序列比对软件通过参数设定能够既做全局比对也做局部比对，有些多序列比对软件则专用于全局或局部比对。由于局部多序列比对追求对保守区段的比对，且保守区段可能只存在于部分输入序列里，比对的结果在形式上十分不同于全局多序列比对。一个流行的局部多序列比对软件是 Mulan，用于发现进化上保守的功能单元。针对草稿序列（draft sequence）和定稿序列（finished sequence），Mulan 分别用 BLASTZ 和 TBA 的 multi-aligner 进行双序列比对，快速找到序列中的保守区段。另外，Mulan 还结合了 multiTF 软件，专门用于检测进化上保守的转录因子结合位点，且允许交互式地修改参数以针对性地检测远距离种系序列间的保守区段和近距离种系序列间的保守区段。Mulan 的一个不足是未考虑保守区段的次序和朝向。另一个常用的局部多序列比对软件是 CHAOS，它通常与 DIALIGN 组合使用，使后者能进行基于局部比对的快速全局比对，参见第四节中的迭代法。

最近开发的一个用于通过局部多序列比对寻找非编码 RNA 的软件是 SCARNALM。非编码 RNA 基因不像蛋白质编码基因那样具有特定的氨基酸编码，它们在 DNA 序列中所呈现的主要特点是具有一定的进化保守性，且具有一定的二级结构特征，SCARNALM 正是充分利用了这两点。SCARNALM 把局部比对所产生的多个保守区段归类成大致可相互之间进行全局比对的块，然后进一步对这些块进行多序列比对。

二、glocal 比对

无论使用相似性还是距离进行计分，全局双序列比对关注整体的两个序列，每个字符间的一一对应使得序列间具有可转换性，因而不容易产生假同源性的结果。与之不同，双序列局部比对返回两个序列的相似区段，通过发现相似的区段它们能够处理直系同源序列（orthologous sequences）的重排（rearrangement），但因为不考虑整体方面的保守性，可能产生高的假同源性结果。同源区段的重排常常在产生新的种系时发生。随着基因组分析尺度的增大，有效地处理大的区段和准确地比对远距

离种系序列中的保守区段日益重要。为了兼顾几方面的需要，人们发展了 glocal 比对，意指它兼具全局(global)和局部(local)比对的特点，使之在保留了处理序列间整体的可转换性的同时也能一定程度地处理同源区段的重排。目前多数新开发的比对软件或多或少采用了 glocal 比对的策略，充分利用全局与局部比对各自的长处。相比于经典的全局比对仅仅通过字符编辑将一个序列转换为另一个序列，glocal 比对试图以最小的开支通过字符编辑以及区段的倒转、转位和复制将一个序列转换为另一个序列，见图 3-7。

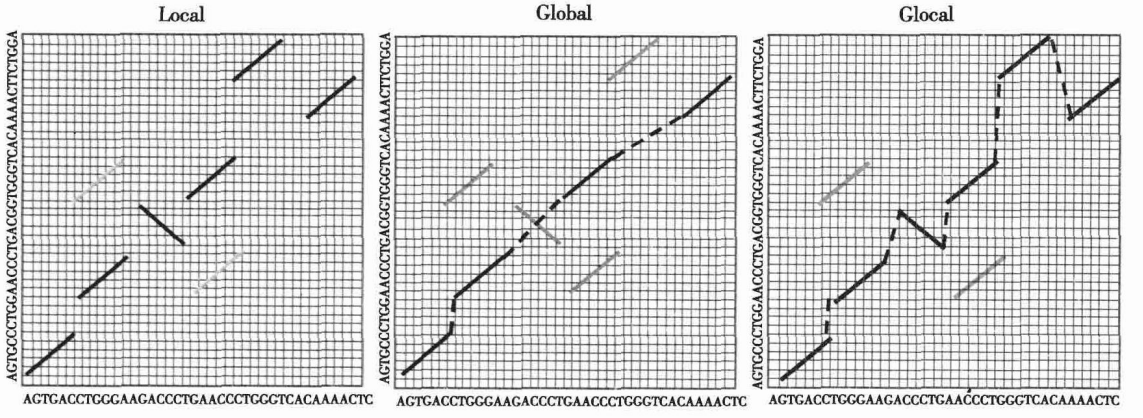


图 3-7 两个序列的局部、全局和 glocal 比对所对应的路径

glocal 比对的一个特点组合是使用全局比对和局部比对的策略。一个用于比对长基因组序列的 glocal 比对软件是 Shuffle-LAGAN，它包含全局比对软件 LAGAN(Limited Area Global Alignment of Nucleotide)和局部比对软件 CHAOS。与在 DIALIGN 中一样，CHAOS 所做的是发现短的局部匹配的种子，以不同方式“串联”种子，计分按不同方式串联的种子；而 LAGAN 所做的是发现局部比对，“串联”局部比对，对各个比对予以计分。已有的测试表明在敏感性和选择性上 Shuffle-LAGAN 都优于单独的全局和局部比对软件。

三、syntenic 比对

使用 Smith-Waterman 算法的局部双序列比对返回两个序列中一个高度匹配的区段。但是，长基因组序列可能包含多个有向和有序的高度保守的区段，如何充分揭示这些区段是大规模基因组分析的重要问题。为了通过多序列比对揭示长基因组序列中的保守区段，人们发展了所谓的 syntenic 比对。syntenic 源自 synteny，原意是“在同一条带(ribbon)上”，在基因组分析中它指的是位于一条染色体上的保守的基因或区段的次序。相比于功能区段，基因组中的非功能区段包含较多的突变、复制和删除，而发生在非功能区段中的这些变化通常并不改变功能区段的朝向和次序。syntenic 比对特别考虑了序列中被非保守区段隔开的多个保守区段。在一个多序列 syntenic 比对软件 Exon-Finder3 里，动态规划法的一个变体被用于进行配对比对，但特殊之处是对相似区段里的失配和空格进行重的罚分而对非相似区段里的失配和空格予以轻的罚分；相应地，它使用两套动态规划表存放两个序列各个前缀中相似区段和非相似区段的计分。配对比对之后，Exon-Finder3 使用星比对进行多序列比对。

另一个具有 syntenic 比对特征的多序列比对软件是 MAP2，它使用 Needleman-Wunsch 算法的一个变体进行配对比对，处理两个序列中被低保守区段(也称差异段)分割开的高保守区段(也称相似块)。多序列比对的结果是由差异段分割开的相似块的一个有序排列，输入序列中所有的差异段直接出现在比对结果中，而对相似块则进行多序列比对。这样，失配和空格罚分只出现在相似块中，而对一个差异段则予以一个常数罚分，其长度与罚分的关系由一个新参数 major_diff 体现。由于对每个相似块进行的是全局比对，MAP2 能够准确地界定相似块和差异段的边界，这对于发现基因组中的

保守功能单元尤其有用。MAP2 对多个序列的比对是用渐进多序列比对方法分两步构造的，在整个过程中相似度高的序列先加入比对，相似度低的序列后加入比对。图 3-8 是一个包含四个 DNA 序列的 MAP2 比对的例子，每个方框是一个差异段，其内的数字是该区段的起始和终止地址，而显示的则是相似块中的序列。

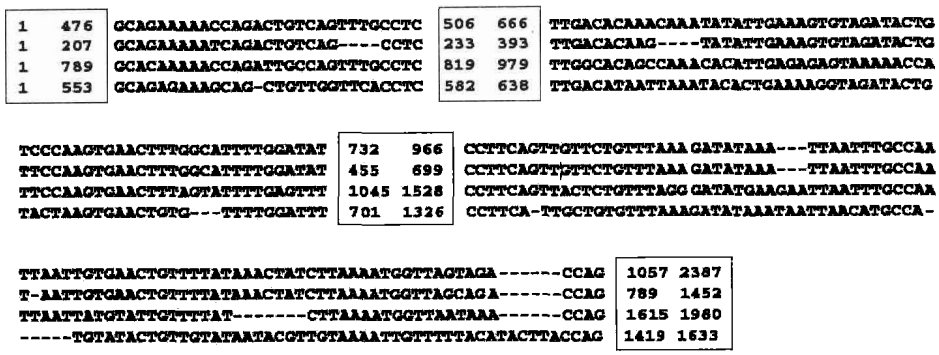


图 3-8 MAP2 产生的多序列比对由得到全局比对的相似块和未进行比对的差异段组成
（摘自：Ye and Huang, Nucleic Acids Research, 2005, 33: 162-170.）

为了比较参数选择对比对的影响，用 MAP2 比对上述六个 hairy and enhancer of split 5 蛋白质序列，当参数是 major_diff=10，mismatch=−1，gap_open=1，gap_extend=1 时 MAP2 识别了 3 个相似块和 2 个差异段，差异段以相应氨基酸的位置标示：

```
12345678901234567890123456789012345678901234567890123456789012345678901234567890
MA-PSTV--AV--EL--L------SPKEKNRLRKPVVEKMRRDRIN-S-SIEQLK-L---LLGAETS--RY
MA-PSTV--AV--EM--L------SPKEKNRLRKPVVEKMRRDRIN-S-SIEQLK-L---LL--EKEFQ--
MRLPRGVGDAA--E---LR---------KS-L-KPLLEKRRRARINESLS--QLKGLVLPLL--ETVF--H
MA-PS----ALSLEI--L------TPKEKNRLRKPIVEKLRRDRIN-S-SIEQLK-L---LL--EKEFQRH
MA-PSTD--F--------LDQQKMTPKEKNKLRKPVVEKMRRDRIN-S-SIEQLK-G---LL--EQEFARH
MA-P------AYMTEYSKL------SNKEKHKLRKPVVEKMRRDRIN-N-CIEQLK-S---ML--EQEFARH

123456789012345678901234567890123456789012345678901234567890
------SKLEKADILEMTVRFL|70   82 |S-L--H----Q-D-YSEGYSWCL-QEA-VQ-FL---T-
-QQDPNAKLEKADILEMTVVFL|70   82 |S-L--H----Q-D-YSEGYSWCL-QEA-VQ-FL---T-
KQQ-PNVKLEKADILEMTVTYL|68   81 |S-L----------DSYLEGYRACLARLARV----LPACSV
--Q-PNSKLEKADILEMTVSYL|70   80 |S-L--Q----Q-D-YCEGYAWCL-KEA-LQ-FL---S-
--Q-PNSKLEKADILEMAVSYL|72   79 |S-EIPHNNDIQMD-YKDGYSRCF-EE--VIDFL---S-
--Q-PNSKLEKADILEMAVSYL|71   71 |QQLRPKTP--QNA-QIEGYSQCW-RET--ISFL---S-

12345678901234567890123456789012345678901234567890123456789012345          6789012345
LHA-ASD-TQMKLLYHF-QR--SGGPPSLTPASAS-AP--APSP---------PVP|143  157|PAC-GLWRPW
LHA-ASD-TQMKLLYHF-QR--S-AAP------ELTHTSEA--PHQQHTHIKQEP|142  158|PAC-GLWRPW
LEP-AVS--AR-LLEHLRQRTVS----------K---AT--A-SSI---SS---F-P|147  147|PSSLGLWRPW
LHS-ANTETQMKLICHF-QR--SQAMP-----KDSGSPS-A-STS---TH--Q-P|141  148|PSC-NLWRPW
LHQ-KQPETA-KLISHF-H---PPAAPAA-PAKEPKAPG-A-AP---------P-P|137  152|-S---LWRPW
VGSEAVA--QR-LQQEA-QR--PPA-PAA-PAKEPPAPG-A-AP---------Q-P|140  143|P----LWRPW
```

这个相似块包含了序列全部的列，因此与前面介绍的全局多序列比对结果本质上无差异。这个结果是可预期的，因为相比于DNA序列，蛋白质序列(尤其是短蛋白质序列)常常是整体上高度保守的。以上例子也说明syntenic比对主要适用于DNA序列。

第五节 全基因组比对

Section 5 Whole Genomic Alignment

一、全基因组多序列比对

随着更多基因组的测序，在基因组分析中多序列比对已开始直接用于整条染色体甚至整个基因组的比对。与syntenic比对类似的是，全基因组比对主要揭示多个序列中保守的和非保守的区段以及这些区段在基因组中的分布特征，但不同的是序列长度为数百万至数千万碱基以上。两个流行的全基因组比对软件是Ensembl基因组浏览器(Ensembl genome browser)和UCSC基因组浏览器(UCSC genome browser)，它们将全基因组多序列比对、全基因组保守区段识别和全基因组浏览捆绑在一起。这里主要介绍UCSC基因组浏览器。

全基因组多序列比对通常使用渐进多序列比对技术，因此其基础是全基因组配对比对。不过，与传统的比对相比，全基因组多序列比对在配对比对和渐进多序列比对两个阶段都有许多特色。首先，全基因组配对比对与传统的配对比对有以下的不同：

1. 全基因组配对比对为了有效处理数百万至数千万碱基以上的序列，将序列分为保守的和非保守的“块”，比对首要考虑发现直系同源的块，然后才是块内序列的详细比对。

2. 除了paralogous直系同源的块，处理一个基因组内数量不同的处在各种位置的旁系同源(paralogous)的块(例如旁系同源基因)也是全基因组配对比对的一个重要问题。

3. 块之间的不对应性使序列中存在更多和更大的缺口，删除事件可同时出现于两个序列中，且可能存在重叠的删除，这些使得引入空格成为一个更加困难的问题。

4. 长基因组序列常常包含高度重复的短序列，它们对测序和比对都是一个困难。

5. 许多块可因复制、删除、反转和转位而发生重排，使得比对更为困难。

由于揭示两个序列中对应的保守区段需要使用局部比对，而对于两个长基因组序列传统的局部比对不能确定具有直系同源关系的多个保守区段的syntenic性质，全基因组配对比对首先需要对所使用的局部比对加以改进。一个软件是由BLAST改进的BLASTZ，它已被广泛用于长基因组序列的配对比对，而另一个专门发展的软件BLAT的做法是，先把一个基因组分解成许多片段，然后根据这些片段在另一个基因组中寻找直系同源块。不过，这些方法仍存有若干不足。

其次，全基因组渐进多序列比对与传统的渐进多序列比对相比有以下不同：

1. 传统的多序列比对方法，尤其是基于动态规划法的方法，不适于处理多个长达数千万的序列。

2. 传统的多序列比对方法难以处理复杂和众多的缺口以及因复制、删除、反转和转位而发生的块的重排。

3. 传统的多序列比对方法没有在两个层次上处理比对，即块之间的比对和块内序列的比对。

4. 不存在评估比对结果的标准测试集，因此对一个基因组多序列比对算法，如何确定敏感性和选择性是一个重要问题。

二、UCSC基因组浏览器

(一) 全基因组多序列比对

针对上述全基因组比对的特点，UCSC基因组浏览器中所采用的多序列比对在多方面作了改进。首先，它采用了参照序列(reference sequence)，使用BLASTZ将每一个序列与参照序列进行局部配对

比对，参照序列中的一个碱基比对另一个序列中的至多一个碱基；其次，依据记分矩阵和两序列的种系关系，对配对比对的结果进行所谓的“串联”(chaining)和“连网”(netting)。一个“串联”的比对(chained alignment)是对多个局部比对序列块的一个有序连接，而对这些序列块的多种不同连接形成所谓的“连网”。“串联”和“连网”有助于识别两个基因组中的直系同源块，且由于“串联”可使许多小的局部比对融合成大的比对，它能极大地减少全基因组比对中序列块的数量和增加序列块的平均长度。在 UCSC 基因组浏览器中，块状部分代表比对的区域，单连线部分代表由删除和插入产生的空格，双连线部分代表更复杂的空格，可包含反转的序列、重叠的删除、频密的突变和未测序区段等。接着，UCSC 基因组浏览器使用 MULTIZ 对多个“串联”的配对比对进行渐进多序列比对，根据已知的种系树首先比对种系关系最近的两个基因组，然后逐渐加入种系关系渐远的基因组。

UCSC 基因组浏览器的另一个重要特征是多序列比对是已预先计算的，其结果存放于数据库中，并根据用户选择的参照序列动态显示。表 3-1 给出了更多与数据库结合的预先计算的(precomputed)多序列比对的例子。

表 3-1 已预先计算的全基因组多序列比对与浏览网站

物种	资源	多序列比对工具	链接
17 个脊椎动物基因组 ENCODE 项目中的区段 9 个昆虫基因组 2 个线虫基因组	UCSC Genome Browser	TBA/MULTIZ	genome.ucsc.edu
9 个脊椎动物基因组	Ensembl Genome Browser	PECAN	www.ensembl.org/index/html
5 个脊椎动物基因组 3 个植物基因组 2 个海鞘基因组	Vista Browser	Shuffle-LAGAN	Pipeline.lbl.gov/cgi-bin/gateway2
3 个线虫基因组	Wormbase	MLAGAN	Wormbase.org
9 个脊椎动物基因组 6 个果蝇基因组	ECR Browser	Blastz	Ecrbrowser.dcode.org
4 个酵母基因组	Broad Institute		www.broad.mit.edu/annotation
12 个果蝇基因组	Eisen 实验室	MAVID, MLAGAN, MULTIZ	Rana.lbl.gov/drosophila

(二) 全基因组多序列比对的显示和序列保守性计算

全基因组多序列比对的主要目的之一是揭示基因组中进化上保守的和非保守的区段以及它们的分布。首先，在碱基层次观察多个全基因组的多序列比对结果和序列保守性非常不方便；其次，许多功能区段的保守性介乎于高度保守和完全不保守之间，不表现为典型的多个序列间的高度匹配，需要用一个单一的指标予以描述；再次，保守区段和非保守区段要能予以明晰的显示。上述三个原因使 UCSC 基因组浏览器使用两个软件 PhastCons 和 PhyloP 把由 MULTIZ 产生的多序列比对结果转换成两种单一的保守性记分和显示，这两个方法都基于已知的种系树结构和利用一个称作 phylo-HMM 的隐马尔科夫模型种系分析方法。PhyloP 只考虑比对的当前列而 PhastCons 同时也考虑比对列的相邻列，这使得 PhastCons 对于保守区段的出现更敏感而 PhyloP 对于保守区段的界定更有效。PhyloP 和 PhastCons 之间的另一个区别是 PhyloP 能够识别加速进化和保守进化的位点，它们分别产生正和负计分，而 PhastCons 的记分总是一个 0 至 1 之间的正值。图 3-9 是一个 UCSC 基因组浏览器多序列比对和保守性计算的例子，所显示的是在多个哺乳及脊椎动物中 Wnt3a 基因的序列。

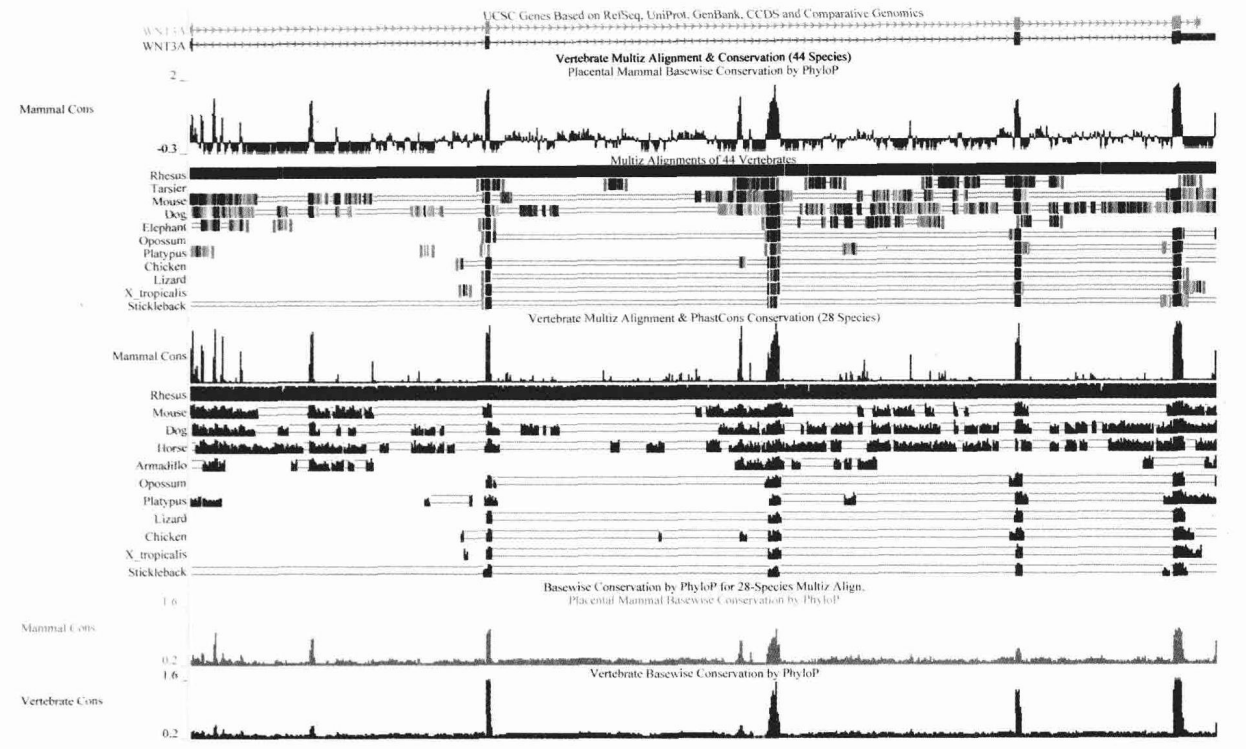


图 3-9 UCSC 基因组浏览器中 Wnt3a 基因的多序列比对和序列保守性

三、其他方法与软件

许多用于长基因组的双序列比对软件被发展成了多序列比对软件，由双序列比对软件 LAGAN 发展而来的多序列比对软件 MLAGAN(Multi-LAGAN)就是一例。与 MULTIZ 有些类似的是，LAGAN 的工作分三个阶段：①产生两序列的所有局部比对，每个赋予一个权值；②对局部比对进行不同的连接，计算具有最大权值的连接；③使用动态规划法是根据局部比对计算最好的全局比对。MLAGAN 的工作则分两阶段：①使用渐进法构造多序列比对；②使用迭代法改进构造的多序列比对。这类方法有一个共同的特点，即利用已知种系关系和使用一个参照基因组，将每一个基因组序列与该参照基因组进行比对。使用参照基因组和已知种系关系的目的是使基因组间的直系同源块获得正确的对应，但存在的两个问题是无法处理仅存在于个别基因组而不存在于参照基因组的某些区段和经历了多次复制的区段。与 UCSC 浏览器类似，MLAGAN 的结果由 VISTA 投影成参照基因组与各个基因组的配对比对。

最近的一些全基因组比对软件放弃了使用参照基因组的做法，其中一个是 Ensembl 基因组浏览器使用的 Enredo。Enredo 对多个基因组进行同步的(simultaneous)多序列比对。由于既不使用参照基因组也不利用基因组间的种系关系，Enredo 对远距离种系的比对可能较为困难。另一个软件是 SuperMap，它基于传统的渐进比对技术，利用基因组的种系关系，但引入了对称(symmetric)配对比对的技术，对所有的序列配对进行两次互为参照的配对比对，能较好地处理出现在任意基因组中的复制事件。在实现上，SuperMap 基于 glocal 比对软件 Shuffle-LAGAN 的“串联”技术(见前面的介绍)，两次运行 Shuffle-LAGAN 使一个配对比对完全对称地处理两个基因组，得到的三组数据是局部比对的结果和对局部比对的两种“串联”，这两种“串联”被转换成一个反映最终比对结果的图。相比于 Shuffle-LAGAN，SuperMap 的主要长处是它能较好地处理两个序列中的复制事件(前面提到一个基因组内处在各种位置的旁系同源块是全基因组比对中的一个重要问题)，它已被扩展成一个基于渐进比对技术的多序列比对软件，并被植入 VISTA 基因组分析软件包中。

第六节 软件、参数和比对质量

Section 6 Softwares, Parameters and Alignment Quality

一、软件的选择

由于存在众多的多序列比对方法和软件，选择合适的软件既十分重要又常常不易。可遵循如下几条原则：第一，序列的种类影响软件的选择。有些软件专用于蛋白质或 DNA 序列，有些软件则两者皆可。比对蛋白质、cDNA 和 RNA 序列时一般选择全局比对，因为整个序列常常是一个功能单元，而比对 DNA 序列时应考虑 glocal 或 syntenic 比对，因为 DNA 序列中常常同时包含保守和非保守的区段。第二，比对的目的影响软件的选择。如果蛋白质和 RNA 序列可能包含多个保守的域（domain），且比对的目的是发现这些域，则应选用 syntenic 比对。发现多个域的典型情形是寻找一个基因中被多个内含子分隔的多个外显子、一个蛋白质中被多个非保守域分隔的多个保守域和一段基因调节区中被多个非保守区段分隔的保守位点。第三，序列的长短影响软件的选择。MSA 不能比对超过 500 字符的序列，比对较长的 DNA 序列可用 MAP2，而比对整条染色体甚至整个基因组时通常使用 UCSC 基因组浏览器和 Ensembl 基因组浏览器。第四，序列保守性的程度可影响软件的选用。在许多 DNA 序列中，保守区段的保守性介乎于高度保守的外显子和完全不保守的 junk DNA 之间，不易由常规的记分机制得以揭示，而 UCSC 基因组浏览器中的 phastCons 和 phyloP 提供了可有效揭示这种中度保守性的计算方法。第五，种系关系的距离影响软件的选择。当序列间种系距离较近时，许多软件会产生大致相同的结果；反之，当序列间种系距离较远时，不同软件产生的结果可能会有相当大的差异，使用基于一致性的方法可充分利用序列间的种系信息。另外，对于比对远距离种系的序列，对敏感性和选择性的取舍十分重要。敏感性与识别尽可能多的同源区段有关，选择性要求识别的同源区段都是真的。不同的软件在这两个彼此矛盾的指标上有不同的取舍。第六，比对种系关系已知的序列时，可使用利用指导树或种系树的算法和软件；比对种系关系未知的序列时，则无法使用这样的软件。对于全基因组序列比对是否使用参照序列以及选用什么序列作为参照序列，这取决于具体序列的特征（包括序列间距离的远近）、对序列的了解（包括对参照序列的了解）、比对的目的（是否主要揭示直系同源区段）以及对比对质量的预估。第七，因为不同算法具有不同的时间和空间复杂度，序列的数量、长度和计算机的性能也影响实际算法和软件的选用。最后，提一下 VISTA 系统，它的多序列比对子系统 mVISTA 包含了三个软件，即用于全局双序列比对的 AVID、用于全局多序列比对的 LAGAN 和用于双序列 glocal 比对的 Shuffle-LAGAN，因此可执行多种多序列比对任务。表 3-2 给出了常见的多序列比对软件。

表 3-2 常见的多序列比对软件及其功能

名称	描述	序列类型	比对类型*	链接	完成时间#
ABA	A-Bruijn alignment	Protein	Global	下载 http://nbcr.sdsc.edu/euler/aba_v1.0/	2004
AMAP	Sequence annealing	Both	Global	在线 http://www.biomath.ucla.edu/msuchard/bali-phy	2006
BAli-Phy	Tree+Multi alignment; Probabilistic/Bayesian; Joint Estimation	Both	Global	在线 + 下载 http://www.biomath.ucla.edu/msuchard/bali-phy	2005 2009
CHAOS/ DIALIGN	Iterative alignment	Both	Local*	在线 http://dialign.gobics.de/chaos-dialign-submission	2003

续表

名称	描述	序列类型	比对类型*	链接	完成时间#
ClustalW	Progressive alignment	Both	Local or Global	在线 http://www.ebi.ac.uk/clustalw/ 下载 http://www.clustal.org/	1994
Codon Code Aligner	Multi alignment; ClustalW & Phrap support	Nucleotides	Local or Global	下载 http://www.codoncode.com/aligner/	2003 2009
DIALIGN-TX, DIALIGN-T	Segment-based method	Both	Local* Global	下载 + 在线 http://dialign-tx.gobics.de/	2005 2008
DNA Alignment	Segment-based method for intraspecific alignments	Both	Local* or Global	在线 http://www.fluxus-engineering.com/align.htm	2005 2008
FSA	Sequence annealing	Both	Global	下载 http://fsa.sourceforge.net/index.html 在线 http://orangutan.math.berkeley.edu/fsa/	2008
Geneious	Progressive/Iterative alignment; ClustalW plugin	Both	Local or Global	下载 http://www.geneious.com/	2005 2009
Kalign	Progressive alignment	Both	Global	在线 http://www.ebi.ac.uk/kalign/	2005
MSA	Dynamic programming	Both	Local or Global	下载 http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html	1989 1995
PRRN/PRRP	Iterative alignment (especially refinement)	Protein	Local or Global	在线 http://align.genome.jp/prrn/	1991
POA	Partial order/hidden Markov model	Protein	Local or Global	下载 + 在线 http://bioinfo.mbi.ucla.edu/poa/	2002
SAM	Hidden Markov model	Protein	Local or Global	在线 http://compbio.soe.ucsc.edu/sam.html	1994 2002
MARNA	Multiple Alignment of RNAs	RNA	Local	在线 http://biwww2.informatik.uni-freiburg.de/Software/MARNA/index.html 下载 http://biwww2.informatik.uni-freiburg.de/Software/MARNA/download/index.html	2005
MAFFT	Progressive/iterative alignment	Both	Local or Global	在线 http://align.bmr.kyushu-u.ac.jp/mafft/online/server/	2005
MAVID	Progressive alignment	Both	Global	在线 http://baboon.math.berkeley.edu/mavid/	2004
MULTALIN	Dynamic programming/clustering	Both	Local or Global	在线 http://prodes.toulouse.inra.fr/multalin/multalin.html	1988
Multi-LAGAN	Progressive dynamic programming alignment	Both	Global	在线 http://genome.lbl.gov/vista/lagan/submit.shtml	2003
MUSCLE	Progressive/iterative alignment	Both	Local or Global	在线 http://www.drive5.com/muscle	2004
Pecan	Probabilistic/consistency	DNA	Global	下载 http://www.ebi.ac.uk/~bjp/pecan/	2008
ProbCons	Probabilistic/consistency	Protein	Local or Global	在线 http://probcons.stanford.edu/index.html	2005
PSAlign	Alignment preserving non-heuristic	Both	Local or Global	下载 http://faculty.cs.tamu.edu/shsze/psalign/	2006
SAGA	Sequence alignment by genetic algorithm	Protein	Local or Global	下载 http://www.tcoffee.org/Projects_home_page/saga_home_page.html	1996 1998

续表

名称	描述	序列类型	比对类型*	链接	完成时间#
StatAlign	Bayesian co-estimation of alignment and phylogeny	Both	Global	下载 http://phylogeny-cafe.elte.hu/StatAlign/	2008
T-Coffee	More sensitive progressive alignment	Both	Local or Global	在线 http://www.tcoffee.org/ 下载 http://www.tcoffee.org/Projects_home_page/t_coffee_home_page.html	2000 2008

注：*指的是主要用于这种比对；#第二个年份指的是最新版的时间。

二、计分等参数的选择

一些软件通过设置参数让用户选择计分办法，而绝大多数软件让用户确定记分参数，这些参数包括匹配分值、失配分值和空格分值，许多软件对引入空格和扩展空格允许予以不同记分。通常的规则是，在任何计分方案中都要保证匹配计分高于失配计分，从而有利于相同字符的比对。另一条规则是，失配有时比使用两个空格好，但这取决于序列间距离的远近、进化的特征以及插入与缺失的情况。若存在较多单个字符缺失，则使用空格比允许失配更合理，因为这些空格能较好地代表这些缺失，使得其余的字符得以正确地匹配；若序列是草稿序列或存在较多字符替换突变，则允许失配可能更合理，因为许多失配可能是由测序错误或替换突变引起的，这些错误和突变没有改变原序列的长度和影响其余字符的对应性。对于对中间和两端的空格是否予以不同的罚分，全局比对和局部比对有不同的做法。当用全局比对比较多个长度相似的序列时，对所有空格进行同样的罚分是合适的，当用局部比对比较多个长度不一的序列时，对两端的空格一般不予罚分。对于 glocal、syntenic 和基因组比对，通常对序列中间的空格也有不同的罚分，即对高保守区和低保守区中的空格减罚不同的分值。

一些参数，例如序列的权值和替换矩阵的选用，依赖于对序列的了解。对于启发式算法中控制时间开销的参数，其确定取决于序列的个数和长度、计算机的性能和容许的运行时间。

三、控制比对质量

可组合使用如下措施提高多序列比对的质量：

1. 剔除太差的序列　一般而言，使用较多的序列能提高多序列比对的质量，但当一个序列与其他序列差异太大并因此对多序列比对产生强的负面影响时，剔除该劣质序列可显著提高其余序列的比对质量。

2. 对序列要有一定的了解　例如，当用 MAP2 进行比对时，对差异段的长度要有大致的了解，由此设定合理的 major_diff 参数。当序列中包含较多失配或较多插入 / 删除时，对失配和空格要予以合适的记分。

3. 选用合适的算法　根据比对的序列和比对的目的选择合适的算法和软件对比对质量非常重要。某些算法如隐马尔科夫模型和遗传算法对某些类型的比对尤其适用，而基于迭代法的软件常常能比渐进比对软件产生更好的结果。

4. 使用多个算法和软件　多个算法和软件能够彼此检验比对的结果。如果多个软件给出相同或相近的结果，比对结果较为可信。其次，当序列较长和较复杂时，可先用 glocal、syntenic 和全基因组比对软件进行粗线条的比对，再用基于动态规划法的软件对更小的区段进行仔细地比对。

5. 选用合适的记分和参数　不同的参数可使一个软件产生相当不同的结果，这从第四节中的例子可得到反映。一些算法，如隐马尔科夫模型，对参数相当敏感，选择合适的参数对运行这类软件十分重要。

6. 尝试多种记分和参数 当对序列缺乏了解时，尝试多种记分和参数相当重要，它可能是产生有意义的比对的关键。

一般认为，算法、软件和参数完全决定了比对的质量，因此当选定了它们之后比对的质量就完全地被决定了。但是，最近发展的全基因组序列比对由于涉及识别多个区段的倒转、转位、复制和缺失，人们提出了在比对中动态控制比对质量的概念。其办法是，在比对过程中对质量进行动态统计，当某些区段的比对质量低于一定的阈值时，清除并重新执行这些比对，或将这些区段拆散为更小的区段进行比对。具备这种功能的软件能够产生更高的比对质量。

四、注意事项

（一）蛋白质多序列比对的矩阵

对于蛋白质序列的比对，一个重要的问题是不同的氨基酸具有不同的彼此间的替换能力。一般而言，相比于性质不同的氨基酸之间的匹配，化学或物理性质相似的氨基酸之间的匹配应具有更高的分值，因此需要对不同的匹配和失配给予特定的计分。在这种情况下，仅仅基于符号一致的计分系统是不够的，需使用专用于蛋白质序列比对的矩阵，最常用的是 PAM 矩阵。

（二）全基因组比对中参照序列的选择

对于使用参照序列的全基因组配对比对，一般要求参照基因组的长度和复杂性不低于被比对基因组的长度和复杂性，否则后者中的许多区段可能无从得以比对。对于哺乳动物全基因组多序列比对，通常用人基因组作为参照基因组，对于果蝇全基因组多序列比对，通常用黑腹果蝇 D. melanogaster 基因组作为参照基因组。

（三）多序列比对中有一些共同的问题需要注意

首先，比对非同源的序列是没有意义的，并且这种比对一般会产生很差的结果。当一组序列中有 1~2 个序列与其他序列不同源时，它们可能会显著影响多序列比对的结果。其次，当序列的一致性较低时，基于不同算法的软件可能会产生非常不同的结果，这时判断哪些结果更正确经常是一件困难的事。在这种情况下，有必要运行多个软件并检验结果的一致性。第三，紧凑的比对（也就是包含的空格较少）未必最符合进化的实际，因此也未必是最正确的比对。不过，如何判断一个紧凑的比对是否正确却相当困难，因为不知道一段序列在进化过程中是字符缺失还是字符替换是造成序列产生差异的主要原因。在这种情况下，采用基于一致性的方法且使用更多的序列或有帮助。最后，当比较不同软件产生的比对结果时，还要考虑到参数的不同和计分方法的不同，在很多情况下比对的计分没有直接的可比性。

（四）对多序列比对的结果解释

要注意的是，如果比对是为了进行种系分析，对判断多个序列是否有共同起源须十分小心。尽管从概率的角度说优化比对不太可能随机出现，进一步将比对所获得的计分与完全无关序列的比对所获得的计分进行比较也经常是必要的。如果输入序列比对的计分显著高于无关序列比对的计分，则表明比对所揭示的序列间的相似性并非偶然。另外，也可使用多种方法进行比对以检查结果是否一致。

小 结

多序列比对是双序列比对的自然推广，许多算法是基于双序列比对的，采用的计分方法也类似，因此本质上的创新不多。不过，在实际的基因组分析中，所做的大部分序列比对是多序列比对，这使得多序列比对技术得到了更多样化的发展[1]。即便所关心的只是两个序列间的进化关系和功能单元的对应性，如果这两个序列种系跨度较大，采用更多的相关的或介乎于它们之间的物种的序列连同进行多序列比对也常常是必不可少的，这些序列所包含的信息能使这两个序列之间的关系通过多个

序列的比对而得以更准确、更可靠和更充分地揭示。

多序列比对算法大致可分为三类：第一类是动态规划法的各种变形和简化版本，包括使用动态规划法进行配对比对的多种渐进多序列比对算法；第二类是特殊比对算法的多序列比对版本，如遗传算法和隐马尔科夫模型；第三类是多序列比对特有的算法，如星比对和树比对。对于单纯的全局比对或局部比对，如果不考虑时间开销，基于 Needleman-Wunsch 算法和 Smith-Waterman 算法的动态规划法是最精确的算法。动态规划法后来有两个主要的改进，一个是 1989 年 Stephen Altschul 等引入的方法，通过修改配对比对来约束 k 维的搜索空间[2]，基于这一改进的软件是 MSA。不过，对于许多超过 20 个序列的多序列比对 MSA 仍然不够实用。另一个更大的改进是使用动态规划法进行配对比对的渐进多序列比对，DF Feng 和 RF Doolittle 于 1987 年提出了一个有效的算法[3]，而典型的例子是 1988 年 DG Higgins 等开发的 Cluster[4]。1994 年 DG Higgins 等对其做了重要的改进，产生了今天广为使用的 Cluster W[5]。

多序列比对之所以发展了种类繁多的算法和软件，原因有如下四个：第一，对精确算法的不同的简化适用于具有不同特征的序列和不同目的的比对。例如，种系距离近的序列和种系距离远的序列通常要求使用不同的算法。第二，比对蛋白质序列和 DNA 序列对算法有不同的要求[6]。由于许多蛋白质的序列和功能是高度保守的，由多序列比对揭示未知蛋白质的功能是研究中的重要手段。蛋白质序列比对几乎都是多序列比对，且为此发展了多种数据库。第三，随着大量基因组的测序，许多动物间的种系关系已通过多种方式得以确定，这使得一大类多序列比对可在已知的种系关系的指导下进行，并将比对结果直接集成到基因组浏览器中[7]。第四，多序列比对如今越来越多地用于大规模基因组分析，典型的全局比对和只关注序列间一个保守区段的局部比对均已十分少见，更多地用到的是所谓的 glocal 比对和 syntenic 比对。大规模基因组分析使多序列比对技术得到了进一步的发展[8]。

一些多序列比对软件如 LAGAN 可用于比对超长的 DNA 序列，但针对全基因组序列比对也发展了其他专门软件。在新的算法与软件发展方面，一个用于长基因组序列 glocal 比对的软件是 Shuffle-LAGAN[9]，它特别考虑了发现序列中区段的重排。与 glocal 比对十分类似的是 syntenic 比对[10]，一个具有 syntenic 比对特征的多序列比对软件是 MAP2[11]，它的特点是忽略序列中高度差异的区段而对高度相似的区段进行局部的全局比对，因此能产生明确的差异区段与保守区段的边界。另一个流行的软件是 Mulan[12]，它特别考虑了发现保守的基因表达调节单元。关于全基因组多序列比对，UCSC 基因组浏览器所使用的方法最具代表性[13,17]。UCSC 基因组浏览器提出了一些新的概念用于复杂和大规模基因组序列的比对[13]，其中包括使用参照序列。目前许多全基因组比对算法使用参照序列，两个没有采用这种做法的软件是 Ensembl 和 VISTA[14]。

关于多序列比对的质量，动态的质量控制是一个重要方向，它对于复杂的基因组序列比对尤其重要，在这方面已有了若干的进展[15,16]。

Summary

Multiple sequence alignment (MSA) is a natural extension of pairwise sequence alignment, and many algorithms, including scoring methods, are therefore based on the latter. In practical sequence analysis, pairwise alignment is rarely used. Even to compare just two sequences, if the evolutionary distance between them is large, aligning them together with more sequences instead of aligning them alone, are preferred or even required, because the inclusion of more evolutionarily related sequences can make conserved elements more accurately and more reliably revealed. Different demands for MSA have stimulated the development of diverse variants of algorithms [1].

Algorithms of MSA fall into three classes. The first are based on either the dynamic

programming method or its simplified variants. Various progressive multiple alignments, with pairwise alignment being conducted using the dynamic programming method, are typical cases. The second are the MSA version of some special algorithms. For example, genetic algorithms and hidden Markov models, with slight modifications, are often used to make MSA. The third, however, are unique to MSA. Star alignment and tree alignment are specifically developed for aligning multiple sequences. In a traditional MSA that aligns either the whole sequences or an element within them, as long as the cost of time is not concerned, the dynamic programming methods - Needleman-Wunsch algorithm for global alignment and Smith-Waterman algorithm for local alignment - produce the best results. To reduce time consumption, two simplified variants have been developed respectively. In 1989, Stephen Altschul et al. modified the dynamic programming based pairwise alignment in order to reduce the k-dimensional search space when aligning k sequences [2]. Upon the simplified algorithm the software *MSA* was developed, to enable aligning multiple sequences with the features of dynamic programming. However, if there are over 20 sequences and sequences are long, *MSA* remains too slow. In 1987, a more fundamental modification was made by DF Feng and RF Doolittle. The key idea is, first, to use dynamic programming to make pairwise alignment, and second, to progressively add aligned sequences into the multiple alignment [3]. Based on this modification, the popular Clustal was developed by DG Higgins et al. in 1988 [4]. They further significantly revised it in 1994, releasing the Clustal W, probably the most widely used MSA software so far [5].

Four reasons explain why numbers of algorithms and softwares of MSA have been developed. First, what behind various variants of dynamic programming is the demand of keeping the key features of dynamic programming while enabling alignment to be done in reasonable time. For example, when handling evolutionarily close and distant sequences, algorithms with different features are needed. Second, protein sequences consisting of 20 amino acids, DNA sequences consisting of 4 neucleotides, and RNA sequences containing secondary structture, should be aligned quite differently [6]. Third, with more and more genomes being sequenced, the phylogenetic relationship among them has been much revealed, which allows sequences to be aligned more accurately under the guidance of the phylogenetic tree [7]. Finally, MSA are now widely applied to extra long sequences or even the whole genomes. In large-scale genome analysis, MSA is often neither typically global nor local. Rather, what should be revealed are multiple conserved elements flanked by highly evolved regions. To effectively identify such conserved elements, glocal and syntenic alignments are much more required. This has significantly stimulated the development of MSA techniques in the recent years [8].

To align extra long sequences or the whole genomes, old softwares such as LAGAN may be workable, but more new versions have been specifically developed. A special one for glocal alignment is Shuffle-LAGAN [9], with the strength of handling rearrangement of sequence blocks. Similarly [10], a software called MAP2 is developed to perform syntenic alignment of multiple sequences [11]. It applies global alignment to highly conserved regions but neglects the regions that are highly species-specific. Another popular software is Mulan [12], specifically developed for identifying conserved gene regulatory elements. As for whole genome alignment, UCSC Genome Browser is a representative. It combines MSA with conservation computation and genome browsing [13,17]. Aligners in UCSC use several new concepts and techniques to effectively align one or several genomes to a reference genome. To make use of annotated genome, to use a

reference genome in whole genome alignment has been widely adopted, but Ensembl Genome Browser and VISTA adopt a different approach [14]. Instead of aligning each genome to a reference genome in pairwise alignment, Ensembl Genome Browser simultaneously aligns multiple genomes. VISTA, somewhat similarly, makes twice, symmetric pairwise alignment for each pair of sequences before constructing multiple alignment.

A critical issue of MSA is the control of alignment quality, which is especially essential for whole genome MSA. Instead of judging the result when alignment is finished, to dynamically control alignment quality during stages of alignment is an important direction [15,16].

（朱 浩）

习 题

1. 使用 Cluster W 比对一个基因的 DNA 序列、mRNA 序列和其蛋白质的氨基酸序列，并仔细观察所产生的结果。

2. 使用 MAP2 比对一个基因的 DNA 序列、mRNA 序列和其蛋白质的氨基酸序列，并仔细观察所产生的结果。

3. 分别使用 MSA 和 Cluster W 比对 5、6、7、8 个物种中一个基因的 DNA 序列和其蛋白质的氨基酸序列，并仔细观察所需要的时间。

4. 分别使用 UCSC Genome Browser、Ensembl Genome Browser 和 VISTA Genome Pipeline 查看多个物种中的一个基因及其邻近序列。

主要参考文献

1. Batzoglou S. The many faces of sequence alignment. Brief Bioinformatics, 2005, 6: 6-22.
2. Lipman D. J., Altschul S. F., Kececioglu J. D. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA., 1989, 86: 4412-4415.
3. Feng D. F., Doolittle R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of Molecular Evolution, 1987, 25: 351-360.
4. Higgins D. G., Sharp P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 1988, 73: 237-244.
5. Thompson J. D., Higgins D. G., Gibson T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research, 1994, 22: 4673-4680.
6. Edgar R. C., Batzoglou S. Multiple sequence alignment. Curr. Opin. Struct. Biol., 2006, 16: 368-373.
7. Kuhn R. M., Karolchik D., Zweig A. S., et al. The UCSC genome browser database: update 2007. Nucleic Acids Research, 2007, 35: D668-673.
8. Blanchette M. Computation and analysis of genomic multi-sequence alignment. Annu. Rev. Genomics Hum. Genet, 2007, 8: 193-213.
9. Brudno M., Malde S., Poliakov A., et al. Glocal alignment: finding rearrangements during alignment. Bioinformatics, 2003, 19: I54-62.
10. Adi S. S., Ferreira C. E. Gene prediction by multiple syntenic alignment. Journal of Integrative Bioinformatics, 2005, 2: 246-250.

11. Ye L., Huang X. MAP2: multiple alignment of syntenic genomic Sequences. Nucleic Acids Research, 2005, 33: 162-170.
12. Ovcharenko I., Loots G. G., Giardine B. M., et al. Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. Genome Research, 2005, 15: 184-194.
13. Blanchette M., Kent W. J., Riemer C., et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Research, 2004, 14: 708-715.
14. Dubchak I., Poliakov A., Kislyuk A., et al. Multiple whole-genome alignments without a reference organism. Genome Research, 2009, 19: 682-689.
15. Loytynoja A., Milinkovitch M. C. SOAP, cleaning multiple alignment from unstable blocks. Bioinformatics, 2001, 17: 573-574.
16. Prakash A., Tompa M. Statistics of local multiple alignment. Bioinformatics, 2005, 21: I344-350.
17. Kent W. J., Baertsch R., Hinrichs A., et al. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. USA., 2003, 100: 11484-11489.

第四章 序列特征分析
CHAPTER 4 ANALYSIS OF SEQUENCE CHARACTERISTICS

第一节 引 言
Section 1 Introduction

DNA 序列是遗传信息的源泉，蛋白质是组成生物体的基本物质，是生命活动的主要承担者。对 DNA 序列和蛋白质序列进行序列特征分析，能够从分子层面上解读基因的结构特点，了解与基因表达调控相关的信息，明确 DNA 序列与蛋白质序列之间的编码关系，阐明蛋白质序列与蛋白质空间结构之间的关系和规律，为进一步研究蛋白质功能与蛋白质结构之间的关系提供理论依据。

基因是 DNA 分子中含有特定遗传信息的一段核苷酸序列，是遗传物质的最小功能单位。基因的概念是随着遗传学、分子生物学、生物化学等领域的发展不断完善的。从分子生物学的角度讲，基因是负载特定生物遗传信息的 DNA 分子片段，在一定的条件下能够表达这种遗传信息，产生特定的生理功能。

原核生物基因的结构非常简单，其典型结构如图 4-1 所示。一个完整的原核生物基因结构是从基因的 5' 端启动子区域开始，到 3' 端终止区域结束。基因的转录开始位置由转录起始位点确定，转录过程直至遇到转录终止位点结束，转录的内容包括 5' 端非翻译区(5'UTR)、开放阅读框(open reading frame，ORF)及 3' 端非翻译区(3'UTR)。基因翻译的准确起止位置由起始密码子和终止密码子决定，翻译的对象为介于这两者之间的开放阅读框。在原核基因组中，基因分布的密度非常高，其中 DNA 分子的绝大部分是用来编码蛋白质的，只有非常小的一部分不转录，这点与真核生物的 DNA 分子不一样。

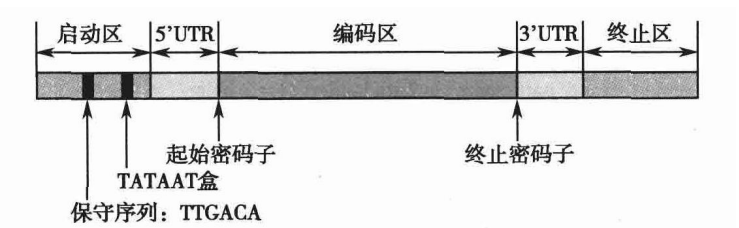


图 4-1 原核基因的典型结构

真核生物基因远比原核生物基因复杂，其典型结构如图 4-2 所示。大多数真核生物基因都是由蛋白质编码序列(外显子，exon)和非蛋白质编码序列(内含子，intron)两部分组成，在一个结构基因

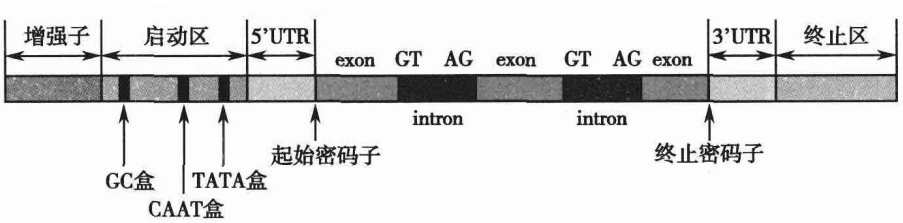


图 4-2 真核生物基因的典型结构

中，编码某一蛋白质不同区域的各个外显子被长度不同的内含子所隔离，形成镶嵌排列的断裂方式，所以真核生物基因有时被称为分裂基因。每个外显子和内含子的连接区域都是一段高度保守和特异的碱基序列，内含子的 5' 端是 GT，3' 端是 AG，这种连接方式称为 GT-AG 法则，普遍存在于高等真核生物基因中，这一保守序列(conserved sequence)与剪切机制密切相关，它是RNA剪切的信号序列，有助于对编码区的识别。

一个完整的基因，不仅包括编码区域，还包括 5' 端和 3' 端两侧长度不等的特异性序列，尽管这些序列不编码氨基酸，但在基因表达过程中起着重要的作用。因此，严格的“基因”分子生物学定义是：产生一条多肽链或功能 RNA 所必需的全部核苷酸序列。

4 种字符字母 A、T(U)、G、C 表示核酸序列中蕴藏着的生命信息，蛋白质则执行着生物体内各种重要的工作。蛋白质序列由相应的核酸序列所决定，通过对基因的转录和翻译，将原来 4 种字符的 DNA 序列根据三联密码有规律地翻译成 20 种字符的蛋白质氨基酸序列。

蛋白质是一种生物大分子，蛋白质中相邻的氨基酸通过肽键形成一条伸展的肽链，这条链称为蛋白质的一级结构，不同蛋白质其肽链的长度不同，肽链中不同氨基酸的组成和排列顺序也各不相同。肽链上的氨基酸残基形成局部的二级结构，各种二级结构在空间卷曲折叠形成特定的三维空间结构。有的蛋白质由多条肽链组成，每条肽链称为亚基，亚基之间又有特定的空间关系，称为蛋白质的四级结构。一般认为，蛋白质的一级结构决定二级结构，二级结构决定三级结构。图 4-3 是钙调素蛋白质的序列及结构，图 4-3(A)是其氨基酸序列；图 4-3(B)是其蛋白质的二级结构，其中 H 表示 α 螺旋，E 表示折叠，B 表示 β 桥，G 表示 310 螺旋，I 表示 π 螺旋，T 表示氢键转角，S 代表转向；图 4-3(C)是其蛋白质的空间结构。

```
>1GGZ:A|PDBID|CHAIN|SEQUENCE
ADQLTEEQVTEFKEAFSLFDKDGDGCITTRELGTVMRSLGQNPTEAELRDMMSEIDRDGNGTVDF
PEFLGMMARKMKDTDNEEEIREAFRVFDKDGNGFVSAAELRHVMTRLGEKLSDEEVDEMIRAADT
DGDGQVNYEEFVRVLVSK
```

A

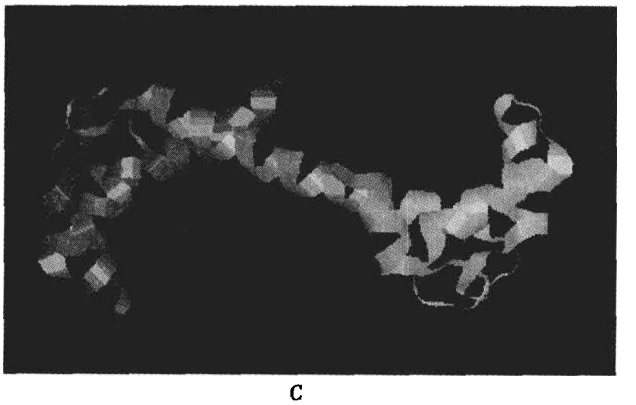
```
ADQLTEEQVT EFKEAFSLFD KDGDGCITTR ELGTVMRSLG QNPTEAELRD MMSEIDRDGN
     HHHHH HHHHHHHHH  TT SSEE HH HHHHHHHHTT     HHHHHH HHHTT TT S

GTVDFPEFLG MMARKMKDTD NEEEIREAFR VFDKDGNGFV SAAELRHVMT RLGEKLSDEE
SSEEHHHHHH HHHHHHHHHH HHHHHHHHHH HH TT SSEE HHHHHHHHHH HH     HHH

VDEMIRAADT DGDGQVNYEE FVRVLVSK
HHHHHHHH T TSSSSEEHHH HHHHHH
```

H = alpha helix; B = residue in isolated beta-bridge; T = hydrogen bonded turn;
E = extended strand, participates in beta ladder; G = 3-helix; I = 5 helix; S = bend

B



C

图 4-3 钙调素蛋白质的序列及结构

A. 氨基酸序列；B. 二级结构；C. 空间结构

蛋白质的生物学功能在很大程度上取决于蛋白质的空间结构，但蛋白质的空间结构又取决于蛋白质一级结构中的氨基酸组成和排列顺序，蛋白质结构构象的多样性导致了不同的生物学功能。蛋白质分子只有处在自己特定空间结构的情况下，才能获得特定的生物活性，空间结构稍有破坏，就很可能会导致蛋白质生物活性的降低甚至丧失，因为它们特定的结构允许结合特定的配体分子。例如，血红蛋白和肌红蛋白与氧的结合、酶和它的底物分子、激素与受体以及抗体与抗原等。知道了基因密码，科学家们可以推演出组成某种蛋白质的氨基酸序列，却无法绘制蛋白质的空间结构。因此，揭示出人类每一种蛋白质的空间结构，已成为后基因组时代的制高点，也是结构基因组学的基本任务。

第二节 DNA 序列特征分析

Section 2 Analysis of DNA Sequence Characteristics

一、利用 GENSCAN 识别基因开放阅读框

开放阅读框是指从 5' 端开始翻译起始密码子(ATG)到终止密码子(TAA、TAG、TGA)的蛋白质编码碱基序列。每个序列都有 6 个可能的开放阅读框，其中 3 个开始于第 1、2、3 个碱基位点并沿着给定序列的 5' → 3' 方向进行延伸，另外的 3 个开始于第 1、2、3 个碱基位点但沿着互补序列的 5' → 3' 方向进行延伸。在开始这项工作之前，并不知道 DNA 双链中哪一条单链是编码链，也不知道准确的翻译起始位点在何处，由于每条链都有 3 种可能的开放阅读框，2 条链共计 6 种可能的开放阅读框，可以从这 6 个可能的开放阅读框中找出一个正确的开放阅读框，通常情况下选择中间没有被终止密码子隔开的最大读码框作为正确的结果。根据这个开放阅读框翻译得到的氨基酸序列才是真正表达的蛋白质产物。

真核生物的开放阅读框不仅含有编码蛋白的外显子，还有内含子，并且内含子将开放阅读框分割为若干个小片段，开放阅读框的长度变化范围非常大，因此真核生物的基因预测远比原核生物困难。但是，在真核生物的开放阅读框中，外显子与内含子之间的连接在绝大部分情况下满足 GT-AG 规律：内含子序列 5' 端起始的两个核苷酸总是 GT，并且其 3' 端最后的两个核苷酸总是 AG，即：5'-GT……AG-3'，这个规律有助于真核生物开放阅读框的识别。

GENSCAN 是美国麻省理工学院 Chris Burge 于 1997 年开发的人类(或脊椎动物)基因预测软件，它是根据基因组 DNA 序列来预测开放阅读框及基因结构信息的开放式在线资源，尤其适用于脊椎动物、拟南芥和玉米等真核生物。GENSCAN 的网址为 http://genes.Mit.edu/GENSCAN.html。图 4-4 是 GENSCAN 的在线操作页面。

1. 主要参数设置

(1) 数据输入：有两种方法提交要预测的序列，一种方法是通过“浏览”按钮上传DNA序列文件，另一种方法是直接将序列粘贴到指定的框中，然后点击 “Run GENSCAN”。

(2) Organism：根据要分析的序列来源在下拉框中选择不同的物种，如 Vertebrate(脊椎动物)、Arabidopsis(拟南芥)、Maize(玉米)。

(3) Suboptimal exon cutoff：这个参数用于定义非确定外显子阈值，其设置范围为 0.01~1.00。在一般情况下，0.10 为一个合适的设置值，低于此数值的设置可能会导致大量非确定外显子的产生，而其中大部分是无意义的；高于此数值的设置则可能会导致丢失有意义的记录。

(4) Print options：根据预测的内容进行选项，预测缩氨酸选择 Predicted peptides only；预测编码区和缩氨酸则选择 Predicted CDS and peptides。

2. 运行程序、输出结果 点击 “Run GENSCAN” 按钮运行程序，输出结果见图 4-5 和图 4-6。

3. 举例 以人类 cosmid 序列为例，其在 GenBank 中的编号为 AC002390。从 GenBank 中下

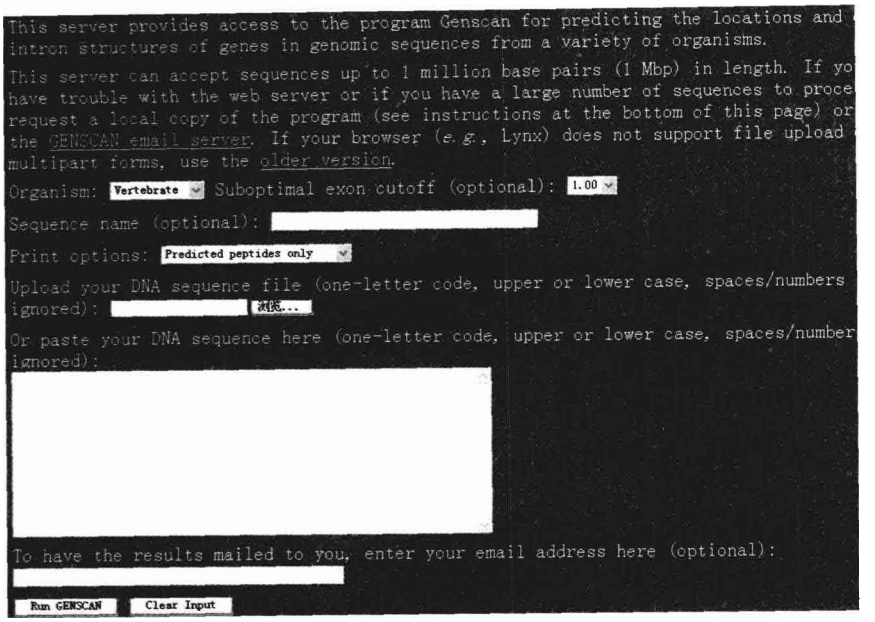


图 4-4 GENSCAN 的在线操作页面

载 DNA 序列并粘贴到指定的框中，Organism 参数选择“Vertebrate”，Suboptimal exon cutoff 参数选择 0.10，Print options 参数选择“Predicted peptides only”，最后点击“Run GENSCAN”按钮运行程序，返回结果如图 4-5 和图 4-6。图 4-5 中各个参数的含义如下：

（1）Gn.Ex: gene number, exon number (for reference)

（2）Type: Init = Initial exon (ATG to 5' splice site)

Intr = Internal exon (3' splice site to 5' splice site)

Term = Terminal exon (3' splice site to stop codon)

Sngl = Single-exon gene (ATG to stop)

Prom = Promoter (TATA box/initation site)

PlyA = poly-A signal (consensus: AATAAA)

Gn.Ex	Type	S	Begin	End	Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	532	657	126	0	0	66	105	46	0.633	2.88
1.02	Intr	+	1399	1459	61	0	1	90	94	20	0.688	1.11
1.03	Intr	+	3269	3349	81	0	0	118	94	76	0.606	10.81
1.04	Intr	+	6557	6649	93	0	0	42	80	77	0.503	2.24
1.05	Intr	+	10004	10093	90	0	0	66	53	84	0.861	2.67
1.06	Intr	+	11990	12019	30	0	0	135	115	37	0.954	9.20
1.07	Intr	+	12099	12173	75	1	0	128	44	90	0.339	8.09
1.08	Intr	+	15414	15459	46	1	1	130	109	21	0.433	5.87
1.09	Intr	+	27955	28151	197	1	2	77	98	122	0.487	11.16
1.10	Intr	+	46659	46791	133	1	1	112	38	68	0.244	3.90
1.11	Term	+	51762	51783	22	0	1	101	38	8	0.025	-5.12
1.12	PlyA	+	52398	52403	6							1.05
2.00	Prom	+	59901	59940	40							-2.16
2.01	Init	+	68711	68764	54	1	0	89	92	66	0.282	8.38

图 4-5 用 GENSCAN 预测 AC002390 序列的基因 / 外显子

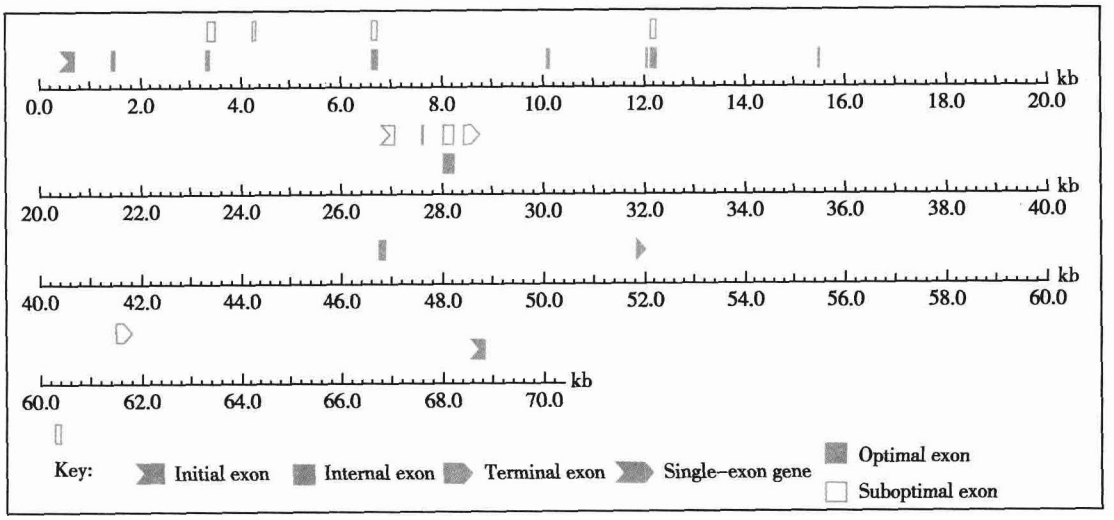


图 4-6 用 GENSCAN 预测 AC002390 序列的基因 / 外显子的位置图

(3) S: DNA strand (+ = input strand; − = opposite strand)

(4) Begin: beginning of exon or signal (numbered on input strand)

(5) End: end point of exon or signal (numbered on input strand)

(6) Len: length of exon or signal (bp)

(7) Fr: reading frame (a forward strand codon ending at x has frame x mod 3)

(8) Ph: net phase of exon (exon length modulo 3)

(9) I/Ac: initiation signal or 3' splice site score (tenth bit units)

(10) Do/T: 5' splice site or termination signal score (tenth bit units)

(11) CodRg: coding region score (tenth bit units)

(12) P: probability of exon (sum over all parses containing exon)

(13) Tscr: exon score (depends on length, I/Ac, Do/T and CodRg scores)

GENSCAN 可以对序列中的多个基因同时进行识别，且对由序列中识别出的基因按顺序进行编号，如图 4-6 所示，本例中共识别出 2 个基因，同时，对每一个基因中的特征序列也进行编号(1.01，1.02，…，2.01)；结果中，Type 参数给出了外显子的性质，P 值表示分析结果为外显子的可能性，当 $P>0.99$ 时为可能性极高的外显子，预测结果几乎与实际完全吻合；当 $0.5<P<0.99$ 时为中等可能性的外显子，预测结果在大多数情况下与实际吻合；当 $P<0.5$ 时为低可能性的外显子，预测结果不可靠。

对图 4-6 所示的结果进行分析，用 GENSCAN 预测 AC002390 序列共识别出 2 个基因。第一个基因的起始外显子(Init)从 582 碱基开始到 707 碱基结束，共有 126 个碱基；接着有 9 个中间外显子(Intr)；终止外显子(Term)从 51 812 碱基开始到 51 833 碱基结束，共有 22 个碱基；其后有 PlyA 信号，从 52 448 碱基开始到 52 453 碱基结束。第二个基因起始外显子的前端有启动子区域，从 59 951 碱基开始到 59 990 碱基结束，共 40 个碱基。由于提交的序列长度只有 70 311 个碱基，所以第二个基因的结构预测并不完整。

GENSCAN 是进行基因预测的首选工具，但是它也有缺陷，就是过分估算了基因数目，例如它对人类基因组进行预测的结果有 45 000 个基因，相当于现在普遍认可数目的两倍，即使这样也比预测出太少的基因要好，可以对过剩的基因进行去假阳性分析。

二、利用 POLYAH 预测分析转录终止信号

转录终止信号是在 mRNA 序列的 3' 端终止密码子下游位置上的加尾信号(tailing signal)。前体 mRNA 3' 端多聚腺苷酸化是真核细胞内 mRNA 转录后处理的三个最主要步骤(包括 5' 帽子结构的形成、内含子的剪切及 3' 端的多聚腺苷酸化)之一，与 mRNA 稳定性的调节、mRNA 的细胞内转运、翻译的起始以及一些其他的细胞机制和疾病机制有重要关系。真核生物前体 mRNA3' 端的多聚腺苷酸化包括两个步骤：①特异性的核苷酸内切酶在 PolyA 位点处进行断裂；②腺苷酸聚合酶在断裂位点处添加 PolyA 尾巴，其主要标志为 AATAAA 或 ATTAAA 两种序列，称为多聚腺苷酸信号(polyadenylation signal)，简称 PolyA 信号序列，也称为转录终止信号。在 3'UTR 区存在多个潜在 PolyA 位点，因此对 PolyA 位点的准确识别，对于预测基因结构、理解 mRNA 的形成机制及某些疾病的分子机制具有巨大的作用。

SoftBerry 网站的 POLYAH 软件是识别 3' 端剪切和 PolyA 区域的在线工具，其网址为 http://linux1.softberry.com/berry.phtml?topic=polyah&group=programs&subgroup=promoter，图 4-7 是 POLYAH 的在线操作页面。

仍以人类 cosmid 序列为例，其在 GenBank 中的编号为 AC002390。从 GenBank 中下载 DNA 序列并粘贴到 POLYAH 指定的框中，点击 PROCESS 按钮提交序列后，网页返回结果见图 4-8。在使用 POLUAH 软件时不需要设置任何参数。用 POLYAH 预测 AC002390 序列的转录终止信号的结果见

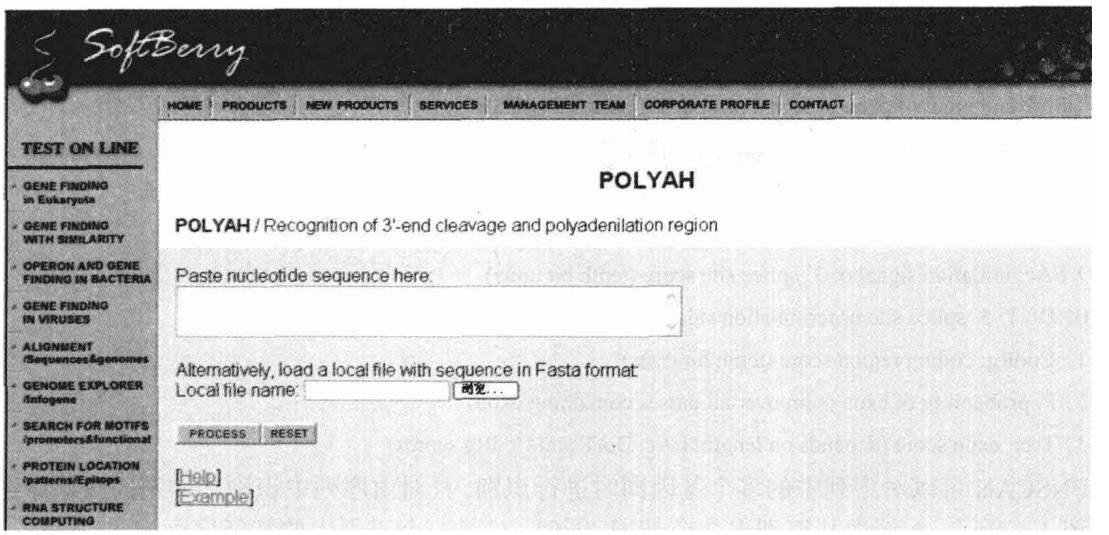


图 4-7 POLYAH 的在线操作页面

图 4-8，结果中列出了 AC002390 序列所有可能的 50 个 PolyA 位点的位置(Pos)和权重(LDF)。例如，在 52 398 碱基处有 PolyA 信号，权重 2.54，这与前面 GENSCAN 预测的结果(图 4-5)一致。值得注意的是，真核生物基因组序列本身存在大量的重复序列，当以 PolyA 位点预测基因终止信号位点时会出现比较大的假阳性。

```
> test sequence
 Length of sequence-     70313
    50 potential polyA sites were predicted
 Pos.:     122 LDF-   3.38
 Pos.:    5057 LDF-   6.18
 Pos.:    6060 LDF-   3.62
 Pos.:    6064 LDF-   4.24
 Pos.:    6076 LDF-   6.17
   .        .
   .        .
   .        .
 Pos.:   44580 LDF-   6.78
 Pos.:   50627 LDF-   4.15
 Pos.:   50635 LDF-   2.84
 Pos.:   52398 LDF-   2.54
 Pos.:   56541 LDF-   5.73
 Pos.:   56546 LDF-   5.64
 Pos.:   56551 LDF-   2.71
   .        .
   .        .
   .        .
```

图 4-8 用 POLYAH 预测 AC002390 序列的转录终止信号的结果

三、利用 PromoterScan 预测分析启动子区域

启动子是基因的一个组成部分，是位于结构基因 5' 端上游区域的 DNA 序列，控制基因表达(转录)的起始时间和表达的程度。启动子本身并不控制基因活动，而是通过与称为转录因子的蛋白质结合而控制基因活动。转录因子就像一面“旗子”，指挥着 RNA 聚合酶的活动。如果基因的启动子部分发生突变，则会导致基因表达的调节障碍，这种突变常见于恶性肿瘤中。

Bioinformatics and Molecular Analysis Section 网站的 PromoterScan 软件是预测分析启动子区域的在线工具，其网址为 http://www-bimas.cit.nih.gov/molbio/proscan/，图 4-9 是 PromoterScan 的在线操作页面。

仍以人类 cosmid 序列为例，其在 GenBank 中的编号为 AC002390。从 GenBank 中下载 DNA 序列并粘贴到 PromoterScan 指定的框中，点击 submit 按钮提交序列后，网页返回结果见图 4-10。在使用 PromoterScan 软件时不需要设置任何参数。用 PromoterScan 预测 AC002390 序列启动子区域的

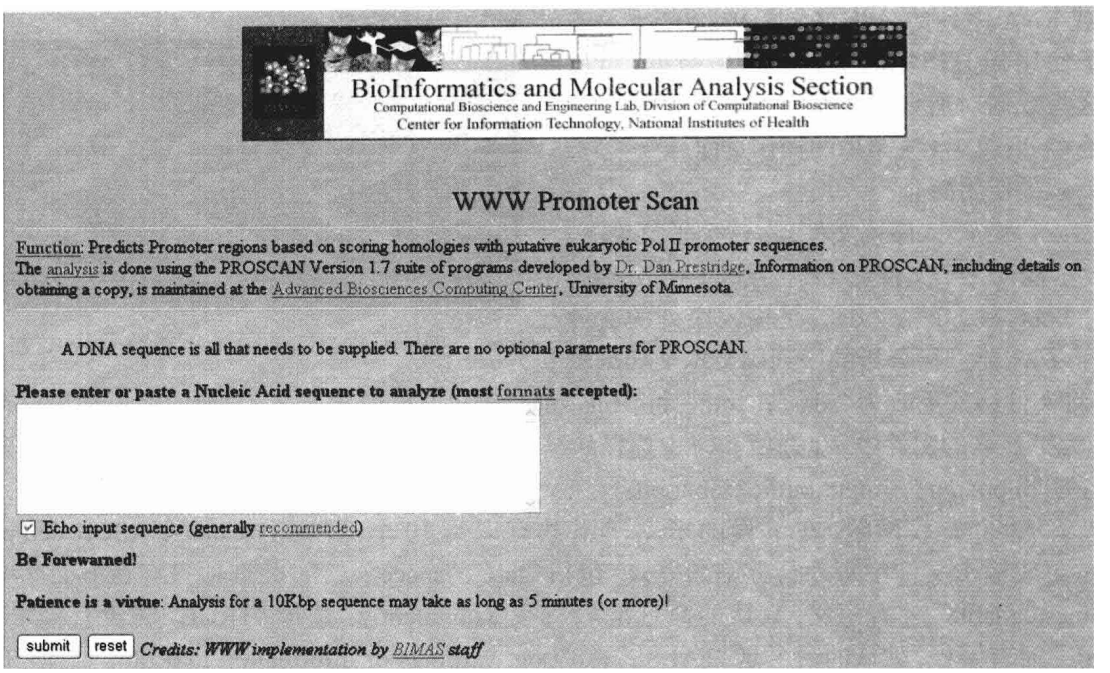


图 4-9 PromoterScan 的在线操作页面

结果见图 4-10，PromoterScan 以单元的形式列出了所有可能的启动子区域，图 4-10 只是结果的一部分。结果中给出了可能的转录因子名称，在 Ghosh TFD database 中的 ID 号，序列所处的正负链、位置及权重。如果在启动子区域中发现 TATA 盒，将给出转录起始位点(transcription start site，TSS)位置的预测，如图 4-10 中的第二个结果单元，在 55449 碱基的位置上发现 TATA 盒，则转录起始位点(TSS)位置从 55479 开始。

值得注意的是，因为转录因子长度较短，无论同源匹配还是模式识别，预测结果的假阳性比例都很高，所以需要结合外显子 / 内含子预测以及 GpG 岛预测的结果来综合判断。此外，并非所有基因的上游区域都符合已知启动子结构的模式，不能单凭 PromoterScan 的预测结果来否定外显子 / 内含子预测或 CpG 岛的预测结果。

```
Promoter region predicted on forward strand in 47985 to 48235
Promoter Score: 57.71 (Promoter Cutoff = 53.000000)

Significant Signals:
 Name                TFD #    Strand  Location  Weight
PEA1                 S01595     +       48087   1.539000
AP-1                 S01426     -       48093   1.513000
TFIID                S01540     +       48111   1.971000
TFIID                S00087     +       48111   2.618000
AABS_CS2             S01612     +       48199   1.012000
Sp1                  S00952     +       48224   50.000000
Sp1                  S01542     -       48233   3.608000

Promoter region predicted on forward strand in 55226 to 55476
Promoter Score: 60.49 (Promoter Cutoff = 53.000000)
TATA found at 55449, Est.TSS = 55479
Significant Signals:
 Name                TFD #    Strand  Location  Weight
Sp1                  S00802     +       55260   3.292000
Sp1                  S00978     -       55265   3.361000
UCE.2                S00437     +       55315   1.278000
NFI                  S00281     +       55377   1.948000
CTF                  S00780     -       55383   1.448000
JCV_repeated_sequenc S01193     -       55408   1.658000
TFIID                S01540     +       55450   1.971000
TFIID                S00087     +       55450   2.618000
TFIID                S00615     +       55450   2.920000
```

图 4-10 用 PromoterScan 预测 AC002390 序列的启动子区域结果

四、利用 CodonW 分析密码子偏好性

密码子使用偏好性是指生物体中编码同一种氨基酸的同义密码子的非均匀使用现象。这一现象的产生与诸多因素有关，如基因的表达水平、翻译起始效应、基因的碱基组分、某些二核苷酸的出现频率、G+C 含量、基因的长度、tRNA 的丰富度、蛋白质的结构及密码子——反密码子间结合能的大小等。所以对密码子的使用偏好性进行分析具有重要生物学意义。

CodonW 是美国 DEC 公司开发的对密码子使用偏好性进行分析的免费软件工具。此软件建立在大量的统计学分析基础上，为简化在线分析的复杂性而开发的，它可以在 Windows 环境下运行，并且可以同时处理 2000 条以上的序列，其网址为 ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z。通过对

DNA 或 RNA 序列的分析，CodonW 会产生关于密码子使用相关指标的统计学分析数据，可以利用这些数据对所要了解的序列进行分析。图 4-11 是 CodonW1.4 主菜单的操作页面。

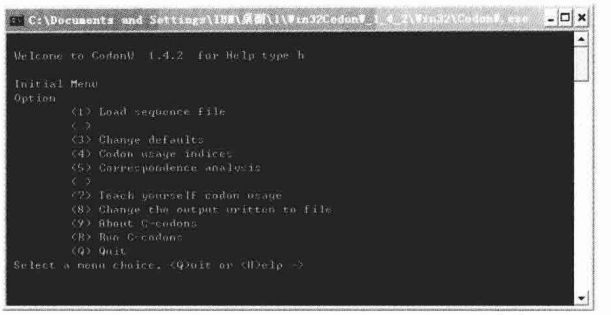


图 4-11 CodonW1.4 主菜单的操作页面

1. 软件使用简介

选项 <1>：输入序列文件，在这个子菜单下，根据提示输入三个文件名，其扩展文件名分别为“.dat”，“.out”，“.blk”。“.dat”是输入序列的文件扩展名，把需要进行分析的序列用文本文档编辑好，再将文件格式更改为“.dat”；“.out”和“.blk”都是输出文件的扩展名，一般的输出结果保存在以“.out”为扩展名的文件中，在选项 <8> 中选择的结果会保存在以“.blk”为扩展名的文件中。默认的文件名为“input.dat”，“input.out”，“input.blk”。

选项 <3>：参数设置，在下设的子菜单下共有以下 10 个选项：①Change ASCII delimiter in output：更改输出文件中每一行分割的方式，包括“tab”、“space”、“,”，此选项只有在下面⑧中选择“machine-readable”才有意义，默认的命令是“,”；②Run silently：选择“TRUE”时没有警告，选择“FALSE”时有警告，默认的命令是“FALSE”；③Log warnings or information to a file：选择“TRUE”时，所有的警告和错误都会被存到后缀名为“.log”的文件中，便于分析，选择“FALSE”时，所有的警告和错误会直接在屏幕上报错，当同时处理大量序列时，选择“TRUE”可以让运行过程更加简洁，同时也可以对运行过程中存在的错误进行系统的分析，默认的命令是“FALSE”；④Number of lines on screen：按照提示输入数字确定界面上同时显示的错误信息行数，如输入“24”表示界面上会同时显示 24 条错误信息；⑤Change the genetic code：在子菜单下有 8 种可供选择的遗传密码子，它们分别是 Universal Genetic code、Vertebrate Mitochondrial code、Yeast Mitochondrial code、Filamentous fungi Mitochondrial code、Insects and Plathyhelminthes Mitochondrial code、Nuclear code of Cilitia Nuclear code of Euplotes、Mitochondrial code of Echinoderms、默认的是 Universal Genetic code；⑥Change the Fop or CBI values：Fop(Frequency of Optimal Codons)和 CBI(Codon Bias Index)是用来衡量密码子使用频率和偏向性的指标，在这个子菜单下，有 8 个物种的名称，这 8 个物种的最优密码子已经确定，运行时需要选择其中一个，默认的物种是 E.coli；⑦Change the CAI values：CAI(Codon Adaptation Index)是用来分析密码子适应性的指标，子菜单中列了 3 个物种的名称，默认的物种是 E.coli；⑧Toggle human or machine-readable output：选择“Human-readable”时输出的数据比较庞大但便于阅读，选择“Machine-readable”时输出的数据是根据①选择的分割方式显示的，默认的选项是“Human-readable”，当处理的序列比较多时，推荐选择“Machine-readable”；⑨Toggle output for each or all genes：选择“Individual genes”会对每一条序列进行单独的分析，选择“Concatenate”会把输入文件中的序列全部链接起来当作一条序列处理，注意当选择“Concatenate”时，COA(Correspondence analysis)不能生成。默认的命令是“Individual genes”；⑩Correspondence analysis defaults：此选项下设菜单和选项 <5> 的下设菜单是一样的。

选项 <4>：选择衡量密码子使用偏好的指标。表 4-1 是 11 个指标的全称和缩写，这些指标可以单选或者多选。

选项 <5>：Correspondence analysis(COA)是一种发掘非随机使用同义密码子的方法，该方法已经在许多物种中得到证实，如 E. coli、S. cerevisiae、M. tuberculosis。以下三种选项每次运行只能选择一个或者不选：①COA on codon usage：选择这个命令，程序只会对同义密码子(synonymous codons)进行分析，产生的文件中只有对同义密码子的分析数据；②COA.on RSCU：选择这个命令，程序只会对相对同义密码子使用(relative synonymous codon usage-RSCU)进行分析，结果也是对应 RSCU 的分析；③COA on Amino Acid usage：选择这个命令，程序只会对氨基酸的使用进行分析。

表 4-1 11 个密码子使用的指标

序号	全称	缩写
1	Codon Adaptation Index	CAI
2	Frequency of Optimal Codons	Fop
3	Codon Bias Index	CBI
4	The effective number of codons	ENc
5	G+C content of the gene	G+C
6	G+C content at 3rd position of synonymous codons	GC3s
7	Silent base composition	
8	Number of silent sites	LSil
9	and amino acids	LAA
10	Hydrophobicity of protein	GRAVY
11	Aromaticity score	Aromo

选项 <7>：自学有关密码子使用的知识。

选项 <8>：选择存入文件 ".blk" 的选项。

选项 <9>：关于 CodonW 的版权信息。

选项 <R>：运行程序：运行后的会自动生成 "summary.coa"，"eigen.coa"，"amino.coa or codon.coa"，"genes.coa"，"cusort.coa"，"hilo.coa"，"cai.coa"，"fop.coa"，"inertia.coa" 等文件，其中 "summary.coa" 文件中整合了其他所有文件的数据信息。

选项 <Q>：退出程序。

2. 由于 CodonW 不能直接显示分析的结果，只能生成存储分析数据的文件，因此需要借助其他的数据统计软件打开 CodonW 生成的文件，可以制成图表以便于对比分析研究。常用的软件有 StatsView，Excel，Minitab，SAS，SPSS，Harvard Graphics，gnuplot。

3. 举例 对玉米、水稻等 7 种植物的 *waxy* 基因的序列进行密码子使用偏好性分析，*waxy* 基因序列来源于 GenBank，其在 NCBI（http://www.ncbi.nlm.nih.gov）上的登录号、基因功能及分子类型见表 4-2。

表 4-2 *waxy* 基因的序列

序号	genebank 登录号	物种	基因功能
1	AY094405	Arabidopsis haliana	granule bound starch synthase I mRNA
2	AF486514	Hordeum vulgare	granule bound starch synthase I mRNA
3	X03935	Zea mays	glucosyl transferase
4	X62134	O.sativa	granule bound starch synthase I mRNA
5	X88789	P.sativum	mRNA for starch synthase
6	U23945	Sorghum bicolor	granule-bound starch synthase precursor(Wx) mRNA
7	X57233	Wheat	waxy mRNA for granule-bound starch synthase

将序列存放在文件 "waxy.dat" 中，打开 CodonW 软件，选择选项 <1>，提示输入文件 "waxy.dat"，"waxy.out"，"waxy.blk"，将文件 "waxy.dat" 载入软件中。回到主菜单，其他选项按照默认的设置，选择选项 <R> 运行软件，CononW 会自动生成相关的文件，图 4-12 是文件 "hilo.coa" 的结果。图中显示了高密码子使用偏向性的 RSCU 值和个数，以及低密码子使用偏向性的 RSCU 值和个数。RSCU 值表示频率值。

		High RSCU	Bias CU	Low RSCU	Bias CU			High RSCU	Bias CU	Low RSCU	Bias CU
Phe	UUU	0.00	(0)	1.33	(2)	Ser	UCU	0.00	(0)	0.41	(2)
	UUC	0.00	(0)	1.67	(20)		UCC	0.00	(0)	2.07	(10)
Leu	UUA	0.00	(0)	0.00	(0)		UCA	0.00	(0)	0.00	(0)
	UUG	0.00	(0)	0.24	(2)		UCG	0.00	(0)	0.41	(2)
	CUU	0.00	(0)	0.71	(6)	Pro	CCU	0.00	(0)	0.13	(1)
	CUC	0.00	(0)	3.06	(26)		CCC	0.00	(0)	1.81	(14)
	CUA	0.00	(0)	0.12	(1)		CCA	0.00	(0)	0.13	(1)
	CUG	0.00	(0)	1.88	(16)		CCG	0.00	(0)	1.94	(15)
Ile	AUU	0.00	(0)	0.21	(2)	Thr	ACU	0.00	(0)	0.15	(1)
	AUC	0.00	(0)	2.79	(26)		ACC	0.00	(0)	2.37	(16)
	AUA	0.00	(0)	0.00	(0)		ACA	0.00	(0)	0.15	(1)
Met	AUG	0.00	(0)	1.00	(18)		ACG	0.00	(0)	1.33	(9)
Val	GUU	0.00	(0)	0.00	(0)	Ala	GCU	0.00	(0)	0.43	(6)
	GUC	0.00	(0)	1.74	(23)		GCC	0.00	(0)	2.14	(30)
	GUA	0.00	(0)	0.00	(0)		GCA	0.00	(0)	0.29	(4)
	GUG	0.00	(0)	2.26	(30)		GCG	0.00	(0)	1.14	(16)
Tyr	UAU	0.00	(0)	0.22	(2)	Cys	UGU	0.00	(0)	0.00	(0)
	UAC	0.00	(0)	1.78	(16)		UGC	0.00	(0)	2.00	(12)
TER	UAA	0.00	(0)	0.00	(0)	TER	UGA	0.00	(0)	3.00	(1)
	UAG	0.00	(0)	0.00	(0)	Trp	UGG	0.00	(0)	1.00	(8)
His	CAU	0.00	(0)	0.00	(0)	Arg	CGU	0.00	(0)	0.18	(1)
	CAC	0.00	(0)	2.00	(8)		CGC	0.00	(0)	2.00	(11)
Gln	CAA	0.00	(0)	0.11	(1)		CGA	0.00	(0)	0.00	(0)
	CAG	0.00	(0)	1.89	(18)		CGG	0.00	(0)	1.45	(8)
Asn	AAU	0.00	(0)	0.09	(1)	Ser	AGU	0.00	(0)	0.21	(1)
	AAC	0.00	(0)	1.91	(22)		AGC	0.00	(0)	2.90	(14)
Lys	AAA	0.00	(0)	0.05	(1)	Arg	AGA	0.00	(0)	0.18	(1)
	AAG	0.00	(0)	1.95	(38)		AGG	0.00	(0)	2.18	(12)
Asp	GAU	0.00	(0)	0.23	(4)	Gly	GGU	0.00	(0)	0.30	(4)
	GAC	0.00	(0)	1.77	(31)		GGC	0.00	(0)	2.22	(30)
Glu	GAA	0.00	(0)	0.16	(3)		GGA	0.00	(0)	0.52	(7)
	GAG	0.00	(0)	1.84	(34)		GGG	0.00	(0)	0.96	(13)

图 4-12 用 CodonW 分析 *waxy* 基因的 RSCU 值和个数

第三节 蛋白质序列特征分析

Section 3 Analysis of Protein Sequence Characteristics

一、利用 ProtParam 分析蛋白质的理化性质

蛋白质是由氨基酸组成的大分子化合物，对组成蛋白质的氨基酸进行理化性质统计分析是对未知蛋白质进行分析的基础。蛋白质的理化性质包括蛋白质的分子量、氨基酸的组成、等电点、消光系数、亲水性和疏水性、跨膜区、信号肽、翻译后修饰位点等。

ExPASy(Expert Protein Analysis System)是由瑞士生物信息学中心维护，并与欧洲生物信息学中心(EBI)及蛋白质信息资源(protein in formation resource，PIR)组成 Universal Protein Knowledgebase (Uniprot)联盟。ExPASy 数据库提供了一系列蛋白质理化分析工具，以便于检索未知蛋白质的理化性质，并基于这些理化性质鉴别未知蛋白质的类别，为后续实验提供帮助。其中 ProtParam(physico-chemical parameters of a protein sequence)就是计算氨基酸理化参数常用的在线工具，其网址为 http://expasy.org/tools/protparam.html，图 4-13 是 ProtParam 的在线操作页面。

以成纤维细胞生长因子(FGF)蛋白质序列为例，其在 GenBank 中的编号为 G00016。从 GenBank 中下载此蛋白质序列并粘贴到 ProtParam 指定的框中，点击 Compute parameters 按钮提交序列，蛋白质序列理化性质的分析结果见图 4-14。在使用 ProtParam 软件时不需要设置任何参数。

从图 4-14 可以看出用 ProtParam 工具分析 G00016 蛋白质序列的结果包括：氨基酸残基数(number of amino acids)、分子质量(molecular weight)、理论等电点(theoretical pI)、氨基酸组成(amino acid composition)、负电荷氨基酸残基总数(total number of negatively charged residues)、正电荷氨基酸残基总数(total number of positively charged residues)、原子组成(atomic composition)、分子式(formula)、原子总数(total number of atoms)、消光系数(extinction coefficients)、半衰期(estimated half-life)、不稳定系数(instability index)、脂肪系数(aliphatic index)、总平均疏水性(grand average of hydropathicity)。消

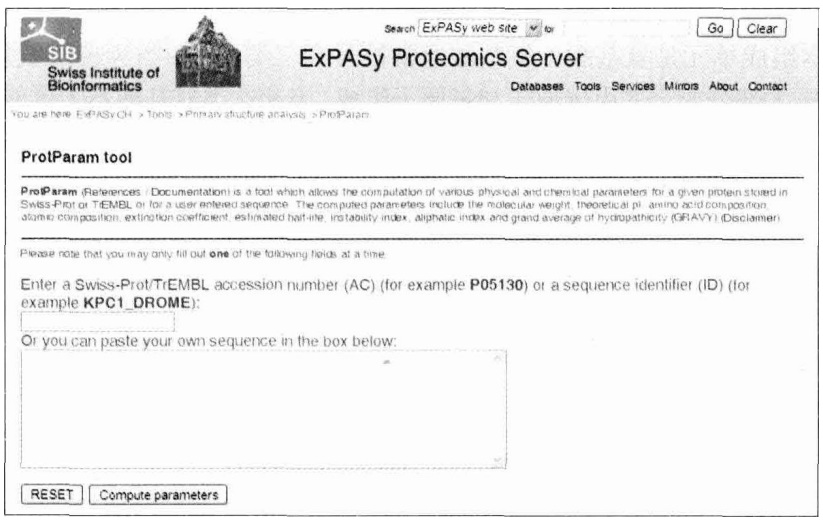


图 4-13 ProtParam 的在线操作页面

```
Number of amino acids: 157 ← 氨基酸残基数
Molecular weight: 18191.9 ← 分子质量
Theoretical pI: 8.43 ← 理论等电点
Amino acid composition: [CSV format] ← 氨基酸组成
Ala (A)   12    7.6%
Arg (R)   11    7.0%
  :
Val (V)   11    7.0%
Total number of negatively charged residues (Asp + Glu): 19 ← 负电荷氨基酸残基总数
Total number of positively charged residues (Arg + Lys): 21 ← 正电荷氨基酸残基总数
Atomic composition: ← 原子组成
Carbon      C             807
Hydrogen    H            1269
Nitrogen    N             223
Oxygen      O             234
Sulfur      S              11
Formula: C807H1269N223O234S11 ← 分子式
Total number of atoms: 2544 ← 原子总数
Extinction coefficients: ← 消光系数
Extinction coefficients are in units of M-1 cm-1, at 280 nm measured in water.
Ext. coefficient    26025
Abs 0.1% (=1 g/l)   1.431, assuming ALL Cys residues appear as half cystines
Ext. coefficient    25900
Abs 0.1% (=1 g/l)   1.424, assuming NO Cys residues appear as half cystines
Estimated half-life: ← 半衰期
The N-terminal of the sequence considered is E (Glu).
The estimated half-life is: 1 hours (mammalian reticulocytes, in vitro).
                            30 min (yeast, in vivo).
                            >10 hours (Escherichia coli, in vivo).
Instability index: ← 不稳定系数
The instability index (II) is computed to be 52.82
This classifies the protein as unstable.
Aliphatic index: 82.61 ← 脂肪系数
Grand average of hydropathicity (GRAVY): -0.400 ← 总平均疏水性
```

图 4-14 用 ProtParam 分析 G00016 序列理化性质的结果

光系数反映了蛋白质在特定波长下吸收可见光或不可见光的能力，该参数可用来测定蛋白质纯度；不稳定系数反映了蛋白质在试验中的稳定程度，按 ProtParam 定义，不稳定系数分值小于 40 时，预测的蛋白质在试验中比较稳定，反之则不稳定。脂肪系数由计算球状蛋白脂肪族氨基酸侧链所占的相对体积得到，反映了蛋白质的热稳定性，带有脂肪族侧链的氨基酸有四种：丙氨酸、缬氨酸、异亮氨酸和亮氨酸。总平均疏水性反映了蛋白质的亲疏水性，它是根据 Hphob.Kyte 和 Doolittle 标度来计算的，其值越高，说明蛋白质的疏水性越强。

二、利用 ProtScale 分析蛋白质的亲水或疏水性

蛋白质的基本组成单元是氨基酸。氨基酸通常被分为三类：第一类为疏水氨基酸(hydrophobic amino acid)，其侧链大部分或全部由碳原子和氢原子组成，因此这类氨基酸不太可能与水分子形成氢键；第二类为极性氨基酸(polar amino acid)，其侧链通常由氧原子或氮原子组成，它们比较容易与水分子形成氢键，因此也称为亲水氨基酸；第三类为带电氨基酸(charged amino acids)，这类氨基酸在生物 pH 环境中带有正电或负电。氨基酸的亲疏水性是构成蛋白质折叠的主要驱动力，一般通过亲水性分布图(hydropathy profile)反映蛋白质的折叠情况。蛋白质折叠时会形成疏水内核和亲水表面，同时在潜在跨膜区出现高疏水值区域，据此可以测定跨膜螺旋等二级结构和蛋白质表面氨基酸分布。

ExPASy 的 ProtScale 程序是计算蛋白质亲疏水性分析的在线工具，其网址为 http://expasy.org/tools/protscale.html，图 4-15 是 ProtScale 的在线操作页面。氨基酸标度(amino acid scale)被定义为指派给每种氨基酸类型的一种数值，最常用的标度是氨基酸的亲疏水性和二级结构构象参数，而氨基酸的其他标度都依赖于氨基酸不同的理化性质。目前 ProtScale 网站上提供了 57 种标度，分析序列前要对氨基酸标度进行选择，图 4-15 列出了其中的 14 种标度，它们分别是分子质量(molecular weight)、密码子数(number of codon)、膨胀度(bulkiness)、极性(polarity)、折射系数(refractivity)、识别因子(recognition factor)等。

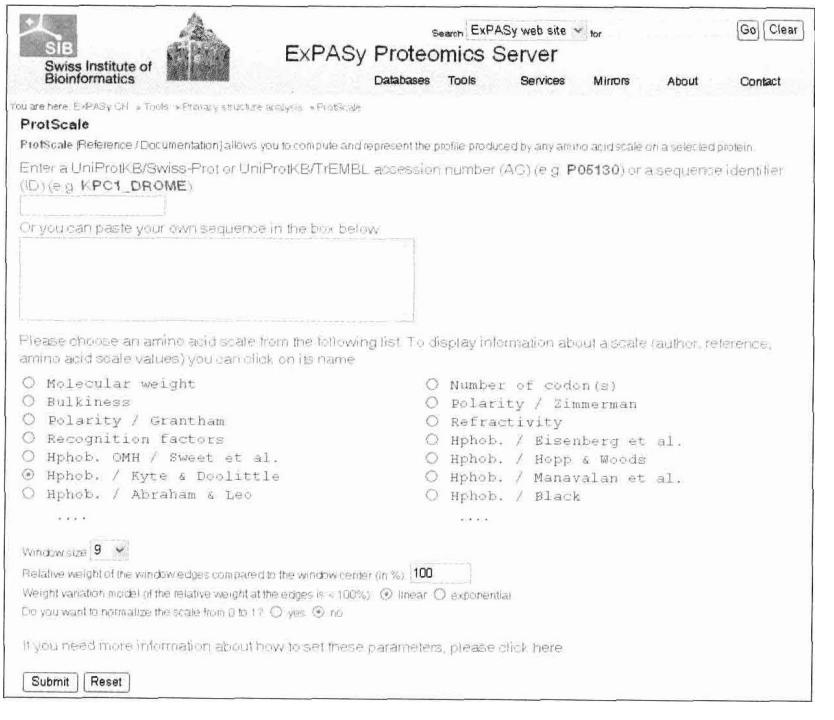


图 4-15 ProtScale 的在线操作页面

1. 主要参数设置

（1）数据输入：有两种方法可以提交要分析的序列，一种方法是在规定的框内输入查询序列在UniProtKB 数据库、Swiss-Prot 数据库、UniProtKB 数据库或 TrEMBL 数据库中的编号；另一种方法是直接将一个蛋白质序列粘贴到指定的框中，选择其他参数的选值，点击“Submit”键。

（2）氨基酸标度的选择：在 ProtScale 网站提供的 57 种标度中选择一种标度作为本次分析蛋白质序列的参数，可以直接点击所选择的氨基酸标度，以便了解这种标度的作者、参考文献、氨基酸的标度值。

（3）Window size：可以把它称为计算窗口，这个参数的大小规定了每次计算得分时所截序列内氨基酸的个数，并且每个氨基酸在窗口内的位置不同，其标度的权值也不同，根据计算模型的要求，一般这个参数选择奇数。

（4）Relative weight of the window edges compared to the window center(in %)：这个参数是计算窗口内最边缘氨基酸的标度权值，其大小决定了计算窗口内相邻氨基酸之间权值的变化比例，一般中心位置氨基酸标度的权值为 100%，计算窗口边缘的氨基酸标度的权值为小于 100% 的权值。

（5）Weight variation model：权值变化模型，这个参数有两个选值：linear(线性)和 exponential(指数型)，根据上一步计算窗口内最边缘氨基酸的标度权值设定和权值变化模型选择，程序可以计算出计算窗口内相邻氨基酸之间权值的变化比例。

（6）Do you want to normalize the scale from 0 to 1：这个参数规定了是否将标度值标准化，它有两个选择：yes(是)和 no(否)。如果要对不同标度的计算结果进行比较，参数要选择“yes”。

2. 把序列粘贴到指定的框中，选择好参数后，点击“Submit”按钮提交序列和参数值，输出结果见图 4-16 和图 4-17。

Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:
(The values in parentheses are the original values, the normalized values have been used in the computation.)

```
Ala:   0.700 ( 1.800)   Arg:   0.000 (-4.500)   Asn:   0.111 (-3.500)
Asp:   0.111 (-3.500)   Cys:   0.778 ( 2.500)   Gln:   0.111 (-3.500)
Glu:   0.111 (-3.500)   Gly:   0.456 (-0.400)   His:   0.144 (-3.200)
Ile:   1.000 ( 4.500)   Leu:   0.922 ( 3.800)   Lys:   0.067 (-3.900)
Met:   0.711 ( 1.900)   Phe:   0.811 ( 2.800)   Pro:   0.322 (-1.600)
Ser:   0.411 (-0.800)   Thr:   0.422 (-0.700)   Trp:   0.400 (-0.900)
Tyr:   0.356 (-1.300)   Val:   0.967 ( 4.200)   :   0.111 (-3.500)
:   0.111 (-3.500)   :   0.446 (-0.490)
```

图 4-16 Hohob./Kyte & Doolittle 标度

Weights for window positions 1,..,13, using **linear weight variation model**:

1	2	3	4	5	6	7	8	9	10	11	12	13
0.10	0.25	0.40	0.55	0.70	0.85	1.00	0.85	0.70	0.55	0.40	0.25	0.10
edge						center						edge

图 4-17 用 Window size=13 时计算窗口内每个位置上氨基酸的标度权值

3. 举例 以牛视紫红质蛋白(bovine rhodopsin)为例，其在 GenBank 中的编号为 P02699。从 GenBank 中下载蛋白质序列并粘贴到指定的框中，氨基酸标度选择“Kyte & Doolittle”，Window size 参数选择 13，Relative weight of the window edges compared to the window center(in %)参数选择 10，Weight variation model 参数选择“linear”，Do you want to normalize the scale from 0 to 1 参数选择“no”。图 4-16 是“Kyte & Doolittle”标度中每个氨基酸的标度值，对于每个氨基酸而言，括号内的数值是原标度值，括号外的数值是标准化的标度值。图 4-17 显示了计算窗口内每个位置上氨基酸的标度权值。图 4-18 是 P02699 序列的疏水性分析的结果图形显示，另外还有不同格式的结果，可以点击相应的位置查看。从 ProtScale 分析 P02699 序列的结果来看，该蛋白存在 7 个高疏水性区域，分别分布在 40~60 区域、75~90 区域、125~135 区域、155~170 区域、205~230 区域、255~275 区域、285~

295 区域；而 4 个主要的最小分值区域则位于 67、147、196、247 氨基酸位点附近，这些区域为高亲水性。在进行蛋白质亲疏水性分析时，可以选择不同类型的氨基酸标度来增强结果信号并去除假阳性信号，还可以调节计算窗口的大小来去除“噪声峰谷”。

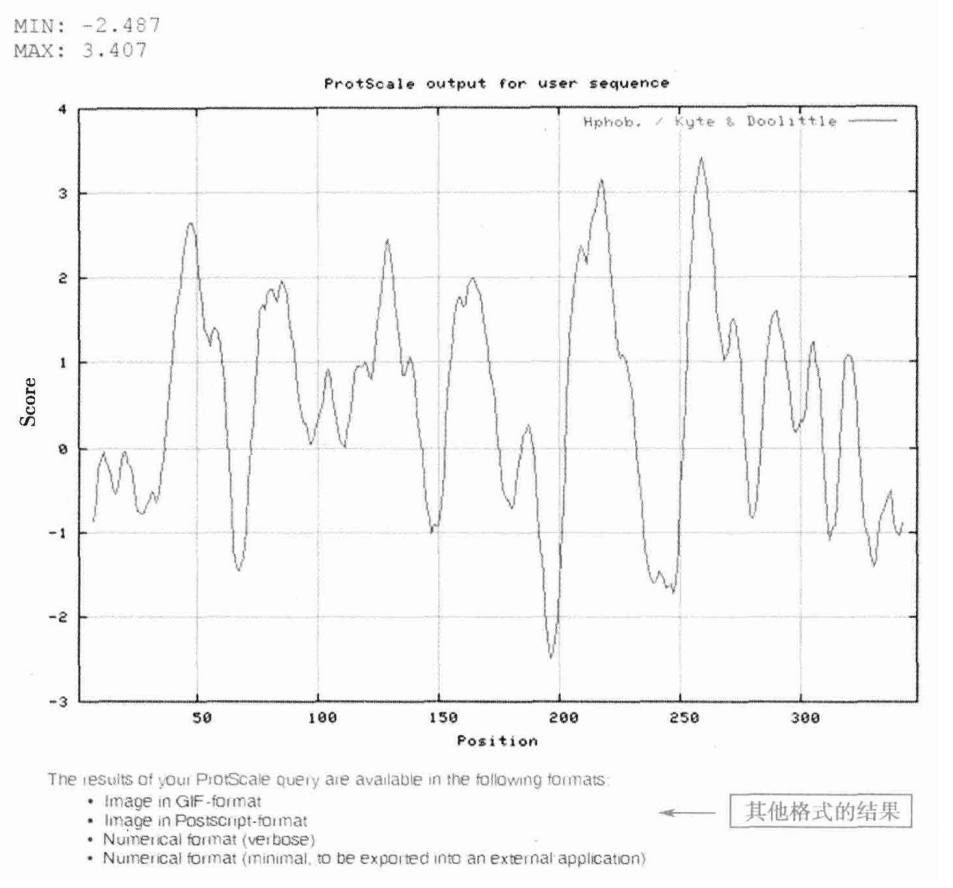


图 4-18 用 ProtScale 分析 P02699 序列疏水性结果的图形显示

三、利用 TMpred 分析蛋白质的跨膜区

生物膜所含的蛋白质叫膜蛋白，是生物膜功能的主要承担者。根据蛋白质分离的难易及在膜中分布的位置，膜蛋白基本可分为两大类：外在膜蛋白和内在膜蛋白。外在膜蛋白约占膜蛋白的 20%~30%，分布在膜的内外表面，主要在内表面，为水溶性蛋白质，它通过离子键、氢键与膜脂分子的极性头部相结合，或通过与内在蛋白的相互作用间接与膜结合；内在蛋白约占膜蛋白的 70%~80%，是双亲媒性分子，可不同程度的嵌入脂双层分子中。有的贯穿整个脂双层，两端暴露于膜的内外表面，这种类型的膜蛋白又称跨膜蛋白。内在膜蛋白露出膜外的部分含较多的极性氨基酸，属亲水性，与磷脂分子的亲水头部邻近；嵌入脂双层内部的膜蛋白由一些非极性的氨基酸组成，与脂质分子的疏水尾部相互结合，因此与膜结合非常紧密。

TMpred 是 EMBnet 开发的分析蛋白质跨膜区的在线工具，其网址为 http://www.ch.embnet.org/software/TMPRED_form.html 。TMpred 基于对 TMbase 数据库的统计分析来预测蛋白质跨膜区和跨膜方向。TMbase 来源于 Swiss-Prot 库，并包含了每个序列的一些附加信息，如：跨膜结构区域的数量、跨膜结构域的位置及其侧翼序列的情况。Tmpred 利用这些信息并与若干加权矩阵结合进行预测。图 4-19 是 TMpred 的在线操作页面，用户将一个蛋白质序列输入查询序列文本框，并可以指定预测时采用的跨膜螺旋疏水区的最小长度和最大长度。输出结果包含四个部分：可能的跨膜螺旋区、相关性列表、建议的跨膜拓扑模型以及表示相同结果的图。

以 G 蛋白偶联受体蛋白质序列为例，其在 GenBank 中的编号为 P51684。从 GenBank 中下载此蛋白质序列并粘贴到 TMpred 的查询序列文本框中，选择预测时采用的跨膜螺旋疏水区的最小长度和最大长度分别为 17 和 33，输出文件格式选择“html”，输入序列格式选择“Plain Text”(纯文本格式)，按“Run TMpred”按钮，可得到 TMpred 对 P51684 序列的分析结果。图 4-20 是用 TMpred 分析 P51684 序列得到的 7 个可能的跨膜螺旋区，由膜内到膜外(inside->outside)的跨膜螺旋有 7 个，分别为：47~69，83~104，123~141，166~184，219~236，255~276，300~319；由膜外到膜内(outside->inside)的跨膜螺旋有 7 个，分别为：55~74，84~104，120~141，166~185，212~235，252~274，299~319。另外图中还给出了每个跨膜螺旋的得分及中心位点，它们的得分都大于 500，因为通常认为只有得分大于 500 的跨膜螺旋才是有意义的。图 4-21 是用 TMpred 分析 P51684 序列得到的 7 个可能的跨膜螺旋区的相关性列表，结果中给出了这 7 个跨膜螺旋在某个方向的偏好性，符号“+”表示这个跨膜螺旋在此方向上有偏好性，符号“++”表示这个跨膜螺旋在此方向上有很强的偏好性。图 4-22 用 TMpred 分析 P51684 序列所得到的 7 个可能的跨膜螺旋区建议跨膜拓扑模型，结果中给出了两个可能的跨膜拓扑模型，在第一个跨膜拓扑模型中：55~74 是从膜外到膜内的跨膜螺旋，83~104 是从膜内到膜外的跨膜螺旋，120~141 是从膜外到膜内的跨膜螺旋，166~184 是从膜内到膜外的跨膜螺旋，212~235 是从膜外到膜内的跨膜螺旋，255~276 是从膜内到膜外的跨膜螺旋，299~319 是从膜外到膜内的跨膜螺旋，这个跨膜拓扑模型的总得分为 14 211，即为各个跨膜螺旋得分之和。同样的方法可以分析结果中的第二个跨膜拓扑模型，其总得分为 12 004。图 4-23 是用 TMpred 分析 P51684 序列得到的 7 个可能的跨膜螺旋区的图形显示结果。

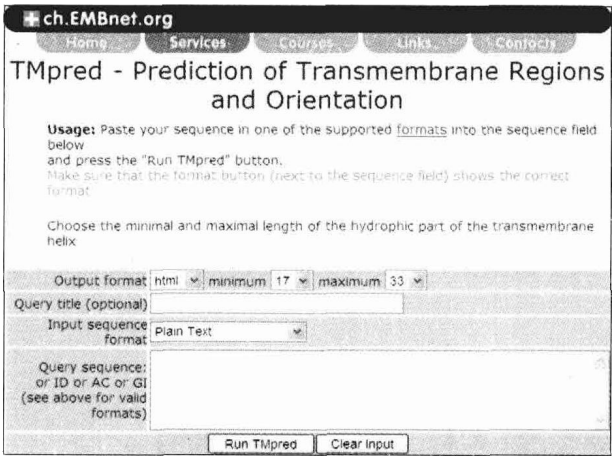


图 4-19 TMpred 的在线操作页面

1.) Possible transmembrane helices

The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

```
Inside to outside helices :   7 found
     from         to     score center
  47 (  51)  69 (  69)   2494     61
  83 (  86) 104 ( 104)   1914     94
 123 ( 123) 141 ( 139)   1352    131
 166 ( 168) 184 ( 184)   2170    176
 219 ( 219) 236 ( 236)   2453    227
 255 ( 255) 276 ( 273)   2140    265
 300 ( 300) 319 ( 319)    915    309

Outside to inside helices :   7 found
     from         to     score center
  55 (  55)  74 (  71)   2707     63
  84 (  86) 104 ( 104)   1470     94
 120 ( 123) 141 ( 139)   1451    131
 166 ( 166) 185 ( 185)   1934    176
 212 ( 214) 235 ( 232)   2530    224
 252 ( 258) 274 ( 274)   1386    266
 299 ( 299) 319 ( 319)   1299    309
```

图 4-20 用 TMpred 分析 P51684 序列所得到的可能的 7 个跨膜螺旋区

2.) Table of correspondences

Here is shown, which of the inside->outside helices correspond to which of the outside->inside helices.

Helices shown in brackets are considered insignificant.
A "+"-symbol indicates a preference of this orientation.
A "++"-symbol indicates a strong preference of this orientation.

```
      inside->outside  |  outside->inside
 47- 69 (23) 2494      |    55- 74 (20) 2707 ++
 83- 104 (22) 1914 ++  |    84- 104 (21) 1470
123- 141 (19) 1352     |   120- 141 (22) 1451  +
166- 184 (19) 2170 ++  |   166- 185 (20) 1934
219- 236 (18) 2453     |   212- 235 (24) 2530
255- 276 (22) 2140 ++  |   252- 274 (23) 1386
300- 319 (20)  915     |   299- 319 (21) 1299 ++
```

图 4-21 用 TMpred 分析 P51684 序列所得到的 7 个可能的跨膜螺旋区的相关性列表

3.) Suggested models for transmembrane topology

```
2 possible models considered, only significant TM-segments used

-----> STRONGLY prefered model: N-terminus outside
 7 strong transmembrane helices, total score : 14211
 # from   to length score orientation
 1   55   74 (20)    2707 o-i
 2   83  104 (22)    1914 i-o
 3  120  141 (22)    1451 o-i
 4  166  184 (19)    2170 i-o
 5  212  235 (24)    2530 o-i
 6  255  276 (22)    2140 i-o
 7  299  319 (21)    1299 o-i

------> alternative model
 7 strong transmembrane helices, total score : 12004
 # from   to length score orientation
 1   47   69 (23)    2494 i-o
 2   84  104 (21)    1470 o-i
 3  123  141 (19)    1352 i-o
 4  166  185 (20)    1934 o-i
 5  219  236 (18)    2453 i-o
 6  252  274 (23)    1386 o-i
 7  300  319 (20)     915 i-o
```

图 4-22 用 TMpred 分析 P51684 序列所得到的 7 个可能的跨膜螺旋区的建议的跨膜拓扑模型

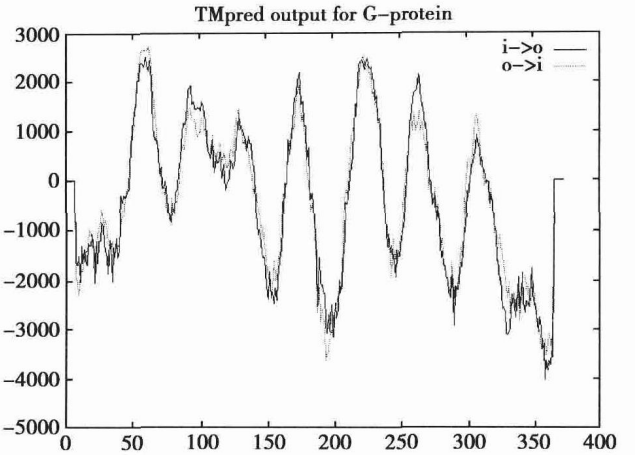


图 4-23 用 TMpred 分析 P51684 序列所得到的 7 个可能的跨膜螺旋区的图形显示结果

四、蛋白质序列分析软件包 Antheprot

对蛋白质的研究是生物化学领域非常重要的一个部分。随着人类基因组计划的实施和完成，得到了大量的蛋白质序列数据，对众多的蛋白质序列数据进行分析工作是一个非常困难的工作，用人工的方法无法完成如此大量的分析工作。运用计算机，利用一定的运算规则，进行蛋白质序列分析是唯一的方法。蛋白质序列分析软件包 Antheprot 正是这样的一个程序，该程序是位于法国的蛋白质生物与化学研究院(Institute of Biology and Chemistry of Proteins)用十多年时间开发出的蛋白质研究软件包，它包括了蛋白质研究领域所包括的大多数内容，功能非常强大。应用此软件包，能进行各种蛋白质序列分析与特性预测，Antheprot 的网站：http://antheprot-pbil.ibcp.fr/，主程序名为 Anthepro.exe，双击主程序名就可以打开 Antheprot_2000 的主窗口，可以输入蛋白质序列，对序列进行编辑、打印、拷贝、改变设置等操作，更重要的是，可以在此调用各种分析工具，对蛋白质序列进行分析。

1. 工具栏

打开一个具有正确格式的序列后，Methods 菜单被激活，相应的快捷键出现在工具栏中，如图 4-24 所示。

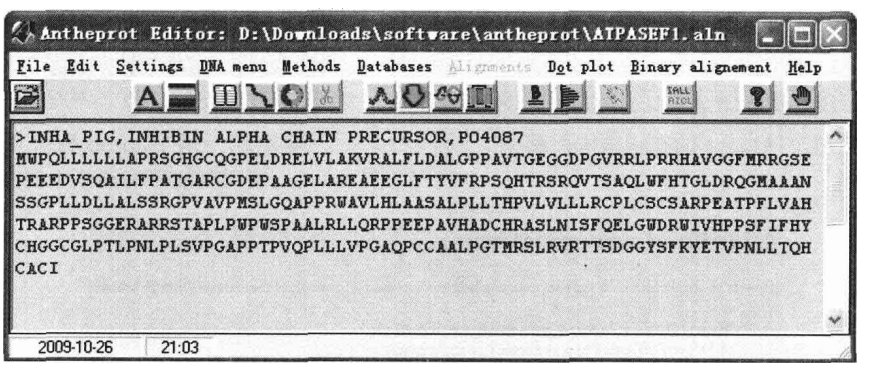


图 4-24 Antheprot 的主窗口

这些按键的含义为：

Change text color，更改选定区域内文本的颜色；

Sequence information，序列信息，计算蛋白质序列的分子量、比溶、各氨基酸残基的百分比组成；

Titration curve，滴定曲线，计算蛋白质序列滴定曲线与等电点；

Helical wheel projection，选定序列的一个片段后，绘制 Helical wheel 图；

Prediction of cleavage site for signal peptide，预测信号肽的剪切位点；

Secondary structure prediction by all，预测蛋白质序列的二级结构；

PROSITE site/signature detection，在蛋白质序列中查找符合 PROSITE 数据库的特征序列；

Physico-chemical profiles，绘制蛋白质序列的理化特性曲线；

Pridict transmembrane region，预测跨膜区；

Similarity search with Blast，用 Blast 方法在选择的数据库中查找相似序列；

Similarity search with Fasta，用 Fasta 方法在选择的数据库中查找相似序列；

Dot Matrix Plot，进行点阵图分析；

Binary alignment(BINALIGN)，在当前蛋白质序列中查找符合 Prosites 数据库的特征序列。

2. 基本功能

(1) Edit(编辑)：包括了常见的剪切(Cut)、复制(Copy)、粘贴(Paste)以及在当前序列中寻找感兴趣的子序列(Find Text)等操作。

(2) Setting(参数设置)：Colors 子目录下可以进一步定义背景颜色(BackColor)、前景颜色(ForeColor)、螺旋(Helix)、折叠(Sheet)、转角(Turn)和卷曲(Coil)，选择 Default 恢复默认颜色；Font 子目录下可以对屏幕和打印文件进行字体、字型和大小的设置；Printer 可对序列文本或图片进行打印；Symbols 子目录可对螺旋(Helix)、折叠(Sheet)、转角(Turn)结构的表示符号进行设置。

(3) Methods(方法选择)：Secondary structure prediction 子目录下可以进行蛋白质二级结构的预测，预测的方法有 Garnier、Gibrat、DPM、Levin、Predator、SOPMA、PHD；Profiles 子目录下可以进行蛋白质疏水性、跨膜区及螺旋的预测；在下载相应的 Prosite 数据库后，Site detection 子目录下可以进行本地的序列模式分析；Similarity search 子目录可以进行本地或网上的 FASTA 及 BLAST 相似性搜索；AA Composition，Specific Vol，M.Weight 子目录可以进行氨基酸组成、分子量等多种物理性质的预测；Helical wheel 子目录可以绘制蛋白质序列的 Helical wheel 图；Amphiphilicity 子目录可以对蛋白质序列进行极性分析；Potential cleavage site of signal pept 子目录可以对真核生物和原核生物序列潜在的信号肽剪切位点进行分析；Titration curve 子目录可以计算蛋白质序列滴定曲线与等电点。上述这些操作可通过相应的快捷键方便进行。

(4) Database(数据库)：可以对 SWISSPROT、SWISSPROT-TrEMBL 及本地数据库(local database)进行序列模式检索。

3. 蛋白质序列编辑功能

Antheprot 的主窗口即为序列编辑窗口，有两种方法可以将序列输入到编辑窗口：

(1) 直接用键盘输入蛋白质序列到编辑窗口，或将序列复制后粘贴到编辑窗口。Pearson/Fasta 格式是最常用的蛋白质序列输入格式，为文本格式文件，带有 > 开始的为序列蛋白的名称，其后是蛋白质序列，如图 4-24 所示。

(2) 可以使用菜单中的 File/Open 命令或 Open file 快捷键打开各种序列文件，Antheprot 可识别的文件类型包括：①单序列文件格式：*.SEQ，序列格式可以是 DNA/Strider 格式、EMBL 格式、NBRF 格式、Pearson/Fasta 格式、PIR 格式或 IG/Stanford 格式；②含有多个蛋白质序列的蛋白质数据库文件：*.BAS，序列格式以 Pearson/Fasta 格式添加序列，上限为 30 条序列或 32K 文件大小；*.ALN，含有 ClustalW 格式的多序列比对；*.MUL，含有多队列的 Multalin 4.1 格式文件；*.SIT 文件，在 Prosite 蛋白质数据库中查寻出 Site 结果的文件；*.PDB 文件，蛋白质原子空间结构的 PDB 格式文件；*.CNS 文件，使用 IBCP(http://www.ibcp.fr/predict.html)服务器预测蛋白质二级结构获得的结果文件格式。

4. 蛋白质基本分析功能

除了通过单击快捷按键进行各种蛋白质序列分析，也可以通过菜单进行各种操作。下面介绍主要分析工具的功能。

(1) 点阵图(Dot plot): 只有当输入一条序列时，才可以激活 Dot Plot 程序。在 Antheprot 主窗口下选择菜单命令 Dot plot 或单击 Dot Matrix Plot 键，运行 Dot Plot 程序，将弹出对话窗口要求用户选择进行比对的第二条序列，然后确定 Window size 和 Similarity threshold 参数，选择替代矩阵，单击 "Ok" 键，将输出分别以两条序列 X、Y 轴进行同源性比较的 Dot plot 图。在图中可以进行以下分析：①在一条序列内查找重复序列；②在两条序列中不同位置查找相似片段；③在不同序列中查找同源性。

(2) 相似性搜索(Similarity search): 此程序用来在一个蛋白质数据库中，找出所有与查询序列具有同源性的序列。在 Antheprot 主窗口下选择菜单命令 Methods/Similarity search/WWW or local Fasta，或者 Methods/Similarity search/WWW-Blast at PBIL NPS@ server 便可实现在蛋白质数据库中对相似性蛋白质序列的查找。如果是在本地蛋白质数据库中查找和分析，需要先下载 Prosite 蛋白质数据库和 SwissProt 蛋白质数据库，解压后放置在 Antheprot 同一目录下，便可实现本地分析检索功能。

(3) 序列位点查询(Site/Signature): 这个功能使用之前，要从 Prosite 数据库下载 Prosite.DAT 和 Prosite.DOC 两个文件，将其存放到 Antheprot 目录下，在 Antheprot 主窗口下单击 PROSITE site/signature detection 键或选择菜单命令 Methods/Site detection(Prosite)，将查询出当前序列中存在的 Prosite 数据库中已定义的序列位点和模式，其检索的缺省设置为 100% 匹配，也可设定为 Mismatch 1 或 Mismatch 2(1~2 个字符不匹配)。需要明确的是，在蛋白质中找到的特征序列并不一定具有生物学意义，但是查找的结果对寻找潜在的有意义蛋白质片段具有很大帮助。

(4) 在序列数据库中查找自定义特征序列(Pattern Search): 在 Antheprot 主窗口下选择菜单命令 Databases/Pattern search，便开始在网上的蛋白质数据库或本地蛋白质数据库中查找含有自定义特征序列的蛋白质，定义的特征序列必须符合 Prosite 规定的语法。语法规则如下：

- 位置分隔符；
- [ ] 允许此位置为括号内的任何一个残基；
- { } 允许此位置为除了括号内所包括的任何一个残基；
- x 代表任何残基；
- x(3) 代表任何 3 个氨基酸残基。

根据以上语法，如果要查找的特征序列为 N-[PT]-{GM}-x(2)-[ILVM]，则序列 N-P-K-G-H-V 和 N-T-L-K-G-M 将能被找到，而序列 N-L-K-G-H-V 和 N-T-G-K-H-V 将不能被找到。查找结束后，所有找到的序列将以 Pearson 格式存储在 SEQ_x.BAS 文件中。

(5) 理化特性预测：在 Antheprot 主窗口下，单击 Physico-chemical profiles 键可以显示预测的蛋白质序列的理化特性曲线。在这个功能下可以分析显示蛋白质的六种理化特性曲线：结合抗原性曲线(Antigenicity)、亲水曲线(Hydrophobicity)、抗原性曲线(Antigenicity)、疏水曲线(Hydrophilicity)、螺旋跨膜区域曲线(Helical membranous regions)、溶剂可及性曲线(Solvent Accessibility)。

(6) 蛋白质二级结构预测：在 Antheprot 主窗口中，单击 Secondary structure prediction by all 键或选择菜单 Methods/Secondary structure prediction，再选择不同的预测方法，就可以对蛋白质序列进行二级结构的预测。蛋白质二级结构预测方法有：Garnier, Gibrat, DPM, Levin, Predator, SOPMA, PHD。

(7) 计算蛋白质序列的分子质量、比容与各氨基酸残基组成百分比例：在 Antheprot 主窗口中，单击 Sequence information 按键或选择菜单 Methods/AA Composition, specific vol, M. Weight，便可以计算当前蛋白质序列的分子量、比容以及各蛋白质残基百分组成的理化特性。

(8) 计算蛋白质序列的滴定曲线与等电点：在 Antheprot 主窗口中，单击 Titration curve 键或选择菜单 Methods/Titration curve 可计算当前蛋白质序列的滴定曲线与等电点。

(9) 绘制 Helical wheel 图：在 Antheprot 主窗口中，可以选定一个蛋白质序列片段绘制其 Helical Wheel 图。用鼠标左键定义起始点，用鼠标右键定义结束点，然后单击 Helical wheel projection 键或选择菜单 Methods/Helical wheel 便可进行计算与绘图。在 Helical Wheel 图形框中单击鼠标左键放大

图形，单击鼠标右键缩小图形。

（10）预测潜在的信号肽剪切位点：在 Antheprot 主窗口下，可以对当前蛋白质序列进行潜在的信号肽剪切位点的预测。单击 Prediction of cleavage site for signal peptide 键或选择菜单 Methods/Potential cleavage site of signal pept，再选择原核序列（Procaryotic sequence）或真核序列（Eucaryotic sequence）便可进行预测。

（11）结果的输出以及与其他应用程序进行数据交换：Antheprot 输出的所有文字、结果图形与曲线，均可以打印；也可以拷贝到剪贴板，粘贴到其他应用软件，如 WORD；或者输出 *.BMP 格式的文件，使用其他软件进行再编辑。程序的缺省设置将输出图形背景色设为黑色，在所有图形输出窗口，均可在 Options 菜单由用户定义背景与前景色颜色，可以将黑色背景反转为白色。

第四节 序列综合分析

Section 4 Sequence Analysis Software

一、EMBOSS 软件包

1. EMBOSS 简介

EMBOSS（European Molecular Biology Open Software Suite，EMBOSS）软件包是一个开源的序列分析软件包，该软件包源于 1988 年开始开发的 EGCG 系统，整合了目前可以获得的大部分序列分析软件，并有一套专门设计的 C 语言库函数。该软件包含 160 多个小型程序，能够自动识别处理不同存储格式的数据，可以通过互联网提取数据，能很好地进行序列模体挖掘、关键词同源性数据库搜索、序列比较、进化分析、序列二级结构分析、限制性酶切图谱分析、引物设计、序列模式识别与翻译、片段拼接等工作。同时它提供了一个扩展库，以方便科学家编制软件。使用 EMBOSS，可以将系列分析工作进行无缝整合，弥补了很多其他软件功能分散、分析效率低下的缺陷。EMBOSS 遵照 GPL 协议，打破了商业软件包发展的传统模式，使科研工作者自由、免费的使用功能强大的分析工具。EMBOSS 的主页网址为 http://emboss.sourceforge.net/。

2. EMBOSS 的运行环境

EMBOSS 软件包主要运行于 linux 操作系统和 Mac 操作系统。现在基于 Windows 操作系统的 EMBOSS 也能自由免费使用。需要说明的是基于 Windows 操作系统时，主要采用 staden 进入 EMBOSS，在使用时，需要安装 Embosswin 软件。Embosswin 的下载网址是：ftp://emboss.open-bio.org/pub/EMBOSS/wEMBOSS Explorerindows/。

3. EMBOSS 的使用界面

EMBOSS 程序码完全公开，其核心程序的基本设计与各种开发平台相兼容，可以供研究人员作为开发应用程序的平台，不同的机构也因此开发了各种 EMBOSS 使用界面，包括 Jemboss、Other GUIs、Web interfaces、Workfolws、Ports and packages 等，下面介绍两种常用的使用界面。

（1）JEMBOSS：JEMBOSS 全名为 java-EMBOSS，即 java 界面的 EMBOSS 程序图形使用界面，由英国 HGMP-RC（Human Genome Mapping Program Resource Center）开发，采用友好的 java 窗口执行 EMBOSS 程序，并加入档案管理功能，使用者下载 java 执行程序后，可在自己的电脑上执行 emboss 程序并储存分析结果，使用方便。

图 4-25 是 JMBOSS 的主界面，主要分为两块区域，左边为程序列表区，右边为程序执行窗口。左边程序列表区上方是按程序类别进行分类，中间可输入程序名称搜索程序，下边就是详细的程序列表。

JEMBOSS 使用的主要程序有：①最重要的程序，Wossname：根据关键字查找序，Showdb：显示所有整合的数据库；②序列编辑，Revseq：将序列反转并互补，Seqret：序列格式转换；③两个序列相

似性图形表达，Dottup：精确匹配，Dotmatcher：近似匹配；④双序列比对，Needle：全局比对，ater：局部比对；⑤多序列比对，Emma：clustalW，alview：多序列编辑；⑥寻找 SNP，Deffseq；仅限于双序列比对中；⑦其他，Potorf，Getorf：翻译，Imep：等电点预测，Tap：跨膜区预测，Pepinfo：蛋白质性质，Patmatmotifs：Motif 搜索。

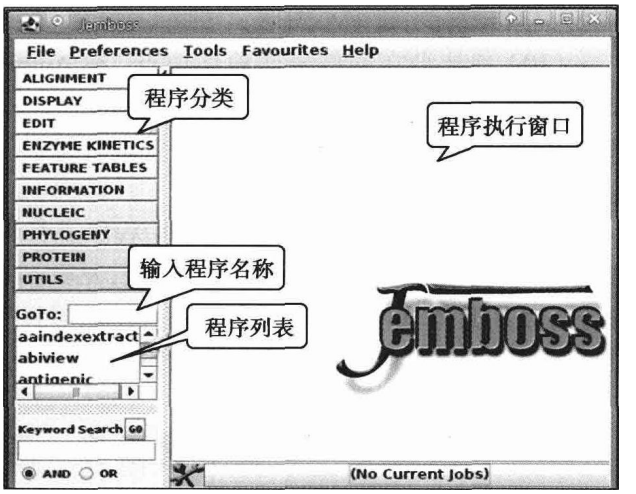


图 4-25 Jemboss 的主界面

（2）EMBOSS Explorer：EMBOSS Explorer 是 Web interfaces 的一种界面，利用 EMBOSS Explorer 可将 EMBOSS 软件从单机版变为网络版，可以让更多的人享用已有的分析平台。这个版本适合安装在服务器上，提供在线分析。图 4-26 是 EMBOSS Explorer 的主界面，左边的区域是程序列表，右边的区域是执行程序区域。

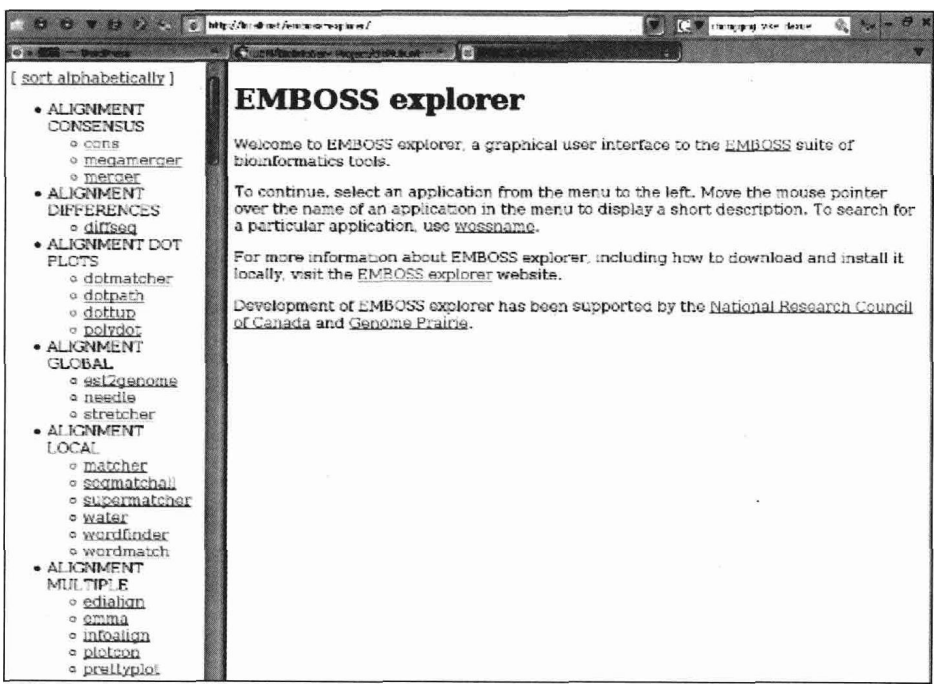


图 4-26 EMBOSS Explorer 的主界面

二、DNAStar 软件包

1. DNAstar 软件包简介 DNAstar 软件包可在计算机上进行 DNA 和蛋白质序列分析，可进行分子生物学中的小规模序列分析和多序列比对。DNAstar 软件包有 PC Windows 和 Macintosh 两种版本，它的主要功能是有 7 种程序可以针对不同的应用和需要进行选择。有关 DNAstar 软件包更详细的信息查询网站：http://www.dnastar.com。

2. DNAstar 的安装环境 该软件包可以在苹果机(Macintosh)或 PC 机(Windows)上安装和升级。建议至少 30Mb 的硬盘，32Mb 的 RAM。

3. DNAstar 软件包中七种程序的介绍

（1）EditSeq：EditSeq 是能够迅速、正确地输入，并且修改 DNA 或蛋白质序列的工具。每个 EditSeq 文件都可以分为三个可编辑的部分，第一部分为序列文件，第二部分是评论，第三部分是序

列的注释。EditSeq 能读取大部分的序列格式，包括 FASTA、GenBank、ABI、GCG 和 ASCII 格式。序列输入的方法有三种：①用菜单命令或拖拽方式输入序列文件；②键盘输入；③复制、粘贴。

（2）GeneQuest：GeneQuest 可以发现和注释 DNA 序列中的基因，并能分析生物学所关心 DNA 的其他特征：包括开放阅读框 ORFS、拼接点连接，转录因子结合位点、重复序列、限制性内切酶酶切位点等。GeneQuest 也提供了整合的 BLAST 和 Entrez 搜索功能，如果知道 Genbank 序列的登录号或名称，就可以直接打开序列。GeneQuest 能直接打开 DNASTAR、ABI 和 GenBank 格式文件，其他格式的序列文件也可以使用 EditSeq 改为 DNASTAR 格式。

（3）MapDraw：根据实验设计、实验分析和实验结果展示需要的不同，MapDraw 可以制作六种酶切图。从简单的线性图到有注释的环形图，在展示限制性酶切位点的同时，还可以同时展示序列的其他特征、六个阅读框及其翻译。MapDraw 可以按照位置、酶切频率等来排列酶切位点；另外，还可以手工选择任何酶切位点进行结合。MapDraw 工具能使使用者规划酶切位点和克隆实验，产生详细的结果。

（4）MegAlign：MegAlign 提供 6 种比对方法进行 DNA 和蛋白质序列比对和多序列比对。多序列比对可以在 MegAlign 的工作区进行查看和编辑。可以根据比对结果制作进化树，根据有关序列距离的数据和残基替代数据可以容易地制成表格。一般多序列比对的结果展示在比对窗口，相似性和差异性用彩色的直方图展示。

（5）PrimerSelect：PrimerSelect 能够设计 PCR、测序和杂交实验所使用的引物和探针。输入 DNA，RNA 或反向翻译的蛋白质模板序列后，PrimerSelect 就可以在 pentamer 窗口计算序列的溶解温度、自由能和末端自由能。PrimerSelect 可以通过控制包括引物浓度、盐分和 GC 计算温度等参数来限定计算结果。在模板处理后，PrimerSelect 按照用户定义的参数确定引物的位置，并给出引物评分，然后筛选出模板序列上的最佳引物序列。通过引物工作栏可以修改引物，检查参数编辑对翻译阅读框、二级结构、错误的引物位置和限制性酶切位点的影响。如果选择并优化了引物，PrimerSelect 可以建立有关模板和引物的文件。

（6）Protean：Protean 可以使用多种方法分析、预测蛋白质结构，并以图形化的方式展示出来。各种方法按照科学概念进行分类。使用者可以按照任何顺序在 Protean 文件上展示各种方法计算的结果。另外，Protean 可以输入来自蛋白质数据库中标注的序列特征，同时允许注释新特征。

（7）SeqMan II：SeqMan II 可以将成千上万个序列装配成重叠群。在装配前，SeqMan II 可以修整质量差的序列以及从序列中清除噪声数据，还提供完善的编辑和输出功能。SeqMan II 可以使用 DNAstar 软件包独特的跟踪品质评价策略，通过对自动测序仪产生的跟踪数据进行评价，自动生成最精确的一致序列。

三、Omiga 2.0 软件包

1. Omiga 2.0 简介　Omiga 2.0 是一款强大的蛋白质、核酸分析软件，可以实现对核酸序列和蛋白质序列分析的大部分功能，同时它还兼有引物设计的功能。

2. Omiga 2.0 的主要功能

（1）实现核酸序列与其互补链之间的转化，序列的拷贝、删除、粘贴、置换以及转化为 RNA 链，以不同的读码框、遗传密码标准翻译成蛋白质序列。

（2）查找核酸限制性酶切位点、序列模式及开放阅读框，设计并评估 PCR、测序引物。

（3）查找蛋白质的水解位点(proteolytic sites)、序列模式、二级结构等。查寻结果以图谱及表格显示，表格设有多种显示形式。

四、Vector NTI 软件包

1. Vector NTI 简介　Vector NTI 是由 Informax 公司开发的一种高度集成、功能齐全的分子生

物学应用软件，可以对 DNA、蛋白质分子进行分析和操作。有关 Vector NTI 更多的信息可以登录 Vector NTI 的官方网站(http://www.informaxinc.com/)查询。

2. Vector NTI 主要功能

(1) DNA 序列的开放阅读框、序列模式、功能区搜索、限制酶图谱、蛋白质翻译；

(2) PCR 引物、测序引物、杂交探针的设计和评价；

(3) DNA 测序片段的拼接；

(4) 同源比较和系统发育树构建；

(5) 蛋白质结构预测：三维结构、化学键、翻译后修饰位点、结构域等；

(6) 模拟电泳：琼脂糖电泳、PAGE。

3. Vector NTI 中附带的资料库　程序中附带的资料库(Vector NTI database)包括：DNA/RNA 序列、蛋白质序列、限制酶、寡核苷酸、电泳 marker。此外程序还提供资料库的开发功能，使用者可以自己修改、添加、拷贝感兴趣的各类资料库。

建立新的序列资料有四种方法：

(1) 用 GenBank/GenPept，EMBL/SWISS-PROT 或 FASTA，ASCII 等格式输入 DNA 或氨基酸序列；

(2) 以 Copy/Paste 方式贴入，然后保存到资料库中；

(3) 从其他序列文档、载体中剪切拼接成新序列；

(4) 从 DNA 或 RNA 序列翻译成蛋白质序列。

使用 GenBank/GenPept，EMBL/SWISS-PROT 或 FASTA 等格式输入的序列，因为内含注释，所以可以直接显示。如果自己粘贴的序列，没有内部特征信息，如 CDS、motif 等，需要自己进行编辑。

小　结

本章主要介绍了有关生物序列特征分析的一些方法和应用软件。本章共分 4 节：第一节，主要介绍了原核生物和真核生物的基因结构特点、蛋白质结构特点及进行生物序列特征分析的意义；第二节，主要介绍了 DNA 序列特征分析的方法，包括：开放阅读框的识别、转录终止信号的预测分析、启动子区域的预测分析、密码子使用偏好性分析的介绍；第三节，主要介绍了蛋白质序列特征分析的方法，包括：蛋白质理化性质的分析、蛋白质亲疏水性的分析、蛋白质跨膜区的分析及蛋白质分析软件包 Antheprot 的介绍；第四节，主要介绍了序列综合分析软件包：EMBOSS、DNAStar、Omiga、Vector NTI。

Summary

This chapter mainly introduces some methods and application software concerning analysis of characteristics of biological sequences. This chapter contains four sections. The first section describes the structural characteristics of genes in prokaryote and eukaryote, characteristics of protein structure and the significance of biological sequence analysis. The second section is devoted to some approaches about DNA sequence analysis, including identification of open reading frames (ORFs), prediction and analysis of transcription stop signals, prediction and analysis of promoter region, analysis of code usage bias and introduction of analysis software. The third section includes some approaches about protein sequence analysis, including the physical and chemical characters of Protein, hydrophilic and hydrophobic of protein, prediction of protein transmembrane regions and protein analysis software--Antheprot. The fourth section

covers the integrated software package of sequences analysis, such as EMBOSS、DNAStar、Omiga、Vector NTI.

（田 心 王兆月 周 猛）

习 题

1. 简述原核生物和真核生物基因结构特点。

2. 简述蛋白质的结构特点。

3. 在 GenBank 中查找一条脊椎动物的 DNA 序列，利用 GENSCAN 软件进行序列的基因开放阅读框的分析预测；设置不同的参数值，对结果进行比对研究。

4. 在 GenBank 中查找一条 DNA 序列，利用 POLYAH 软件预测分析转录终止信号。

5. 在 GenBank 中查找一条 DNA 序列，利用 PromoterScan 软件预测分析启动子区域。

6. 从网站 ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z 下载 CodonW 软件，安装到本地计算机，然后从 GenBank 中查找一组 DNA 序列，利用 CodonW 软件分析这组 DNA 序列密码子的使用偏性。

7. 在 GenBank 中查找一条蛋白质序列，利用 ProtScale 软件分析这条蛋白质序列的疏水性，设置不同的参数值，对结果进行比对研究。

8. 在 GenBank 中查找一条蛋白质序列，利用 SignalP 软件分析预测这条蛋白质序列的跨膜区域；设置不同的参数值，对结果进行比对研究。

主要参考文献

1. Deleage G., Combet C., Blanchet C., et al. ANTHEPROT: An integrated protein sequence analysis software with client/server capabilities. COMPUTERS IN BIOLOGY AND MEDICINE, 2001, 31(4): 259-267.
2. Cai M. S., Cheng A. C., Wang M. S., et al. Characterization of Synonymous Codon Usage Bias in the Duck Plague Virus UL35 Gene. Intervirology, 2009, 52(5): 266-278.
3. John F., Peden B.Sc., M.Sc. Analysis of Codon Usage. The University of Nottingham for the Degree of Doctor of Philosophy, 1999.
4. Burge C. B., Karlin S. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol., 1998, 8(3): 346-354.
5. Rice P., Longden I., Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics, 2000, 16(6): 276-277.
6. Prestridge D. S. Prediciting pol Ⅱ promoter sequence using transcription factor binding sites. J. Mol. Biol., 1995, 249(5): 923-932.
7. Gasteiger E., Hoogland C., Gattiker A., et al. Protein Identification and Analysis Tools on the ExPASy Server. Methods Mol Biol., 1999, 112: 531-52.
8. Hofmann K., Stoffel W. TMbase - A database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler, 1993, 374(47): 166.
9. Bendtsen J. D., Nielsen H., von Heijne G., et al. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol., 2004, 340(4): 783-795.
10. Nielsen H., Engelbrecht J., Brunak S., et al. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering, 1997, 10(1): 1-6.
11. Lupas A., Van Dyke M., Stock J. Predicting Coled Coils from Protein Sequences. Science, 1991, 252(5010): 1162-1164.
12. 薛庆中. DNA 和蛋白质序列分析工具. 北京：科学出版社；2009 年.

13. 张成岗，贺福初. 生物信息学方法与实践. 北京：科学出版社；2002 年.
14. 李军，张莉娜，温珍昌. 生物软件选择与使用指南. 北京：化学工业出版社；2008 年.
15. 胡松年，薛庆中. 基因组数据分析手册. 杭州：浙江大学出版社；2003 年.
16. 张阳德. 生物信息学. 北京：科学出版社；2004 年.
17. 孙啸，陆祖宏，谢建明. 生物信息学基础. 北京：清华大学出版社；2005 年.
18. 李衍达，孙之荣. 生物信息学基因和蛋白分析的实用指南. 北京：清华大学出版社；2000 年.

第五章 分子进化分析
CHAPTER 5 MOLECULAR EVOLUTION ANALYSIS

第一节 引 言
Section 1 Introduction

进化是一种不断改进的过程，在《物种起源》中这样描述："每个生物每时每刻都在为生存进行反复的斗争，如果在复杂甚至多变的生存条件下该生物仍然能够不断改进自己，那么其将有较大的生存可能性并被自然选择所保留。根据严格的遗传法则，任何被自然选择保留下来的物种都倾向于繁殖其已经被改进的新的生命形式。"尽管自然选择在形态形成和行为进化方面似乎普遍存在，但在某些基因和基因组进化中所起的作用也有其他看法。分子进化的中性学说认为，种内和种间大多数可见差异不是自然选择，而是适合度很小的随机突变的固定所决定的。

在分子水平上，进化是一种伴随突变的过程。自从20世纪60年代，由于分子遗传学资料的迅速积累，分子进化逐渐成为计算生物学和生物信息学等新兴学科的重要组成部分。分子进化分析着重于研究不同系统发生树分支上基因和蛋白质的变化方式，其研究方法和研究方向也在不断发展，最大似然法、模式识别等很多机器识别的方法也被广泛用于系统发生树构建和同源基因识别。

人类基因组和多种生物基因组测序计划的完成，推动了分子进化的跨越式发展，基因表达和生物网络的进化等研究内容不断出现在最新的研究中，扩展了分子进化分析的研究范畴。许多研究者认为基因表达调控的差异可能对物种内和物种间的表型差异有重要的作用。基因的进化可能不是独立进行的，而是受到蛋白质互作或通路的限制，是一个协同进行的过程，这些研究拓展了分子进化的深层发展，此外促进了多个基因共同进化或者以模块的形式研究进化关系，以及从整个网络的层面实现进化的研究。在本章下面的内容中，将对分子进化的基本知识和研究进程进行介绍。

第二节 系统发生分析与重建
Section 2 Phylogeny Reconstruction

一、核苷酸置换模型及氨基酸置换模型

（一）DNA序列进化分析

由于DNA序列包括多种不同类型的区域，如蛋白质编码区、非编码区、外显子、内含子、侧翼区、重复DNA序列和插入序列等。因此DNA序列的进化演变比蛋白质序列的演变更复杂。因此，弄清所研究的DNA类型和功能是十分重要的。即便单独考虑蛋白质编码区，密码子第一、二、三位的核苷酸替代样式也不尽相同。而且，某些区域比其他区域更易受到自然选择的影响，因此DNA不同区段呈现不同的进化模式。这里主要研究蛋白质编码区和RNA编码区，这些区域的进化相对简单，但通过它们来理解进化的一般规律极为重要。

1. 两个序列间的核苷酸差异　同一祖先序列传衍的两条后裔序列，它们的核苷酸差异随时间增长而增加。一个简便的描述序列分歧大小的测度是两条后裔序列中不同核苷酸位点的比例。

$$\hat{p} = n_d / n \qquad \text{式 5-1}$$

这里，n_d 和 n 分别为所检测的两序列间不同核苷酸数和配对总数。在以下的内容中，将此估计称为核苷酸间的 p 距离。

2. 核苷酸替代数的估计　如同氨基酸替代，当序列间亲缘关系较近时，p 距离可用来估计每个位点上的核苷酸替代数。然而，当 p 较大时，因为没有考虑回复突变和平行突变，替代数将被低估。由于核苷酸在序列中只有四种状态，这个问题对核苷酸序列比对氨基酸序列估计更为严重。

估计核苷酸替代数，一般应用核苷酸替代的数学模型。为此，许多学者提出了不同的替代模型，其中两个模型以替代率矩阵的形式列在表 5-1 中。

表 5-1　核苷酸替代模型

	(A)Jukes-Cantor 模型					(B)Kimura 模型			
	A	T	C	G		A	T	C	G
A	--	α	α	α	A	--	β	β	α
T	α	--	α	α	T	β	--	α	β
C	α	α	--	α	C	β	α	--	β
G	α	α	α	--	G	α	β	β	--

（1）Jukes-Cantor 方法：这个最简单的核苷酸替代模型由 Jukes 和 Cantor 提出。该模型假定任一位点的核苷酸替代都是以相同频率发生的，且每一位点的核苷酸每年以 α 概率演变为其他 3 种核苷酸中的一种。因此，一个核苷酸演变为 3 种其他核苷酸的任何一种的概率为 $\gamma=3\alpha$，γ 为每年每个位点的核苷酸替换率。

在这个模型中，假设每对核苷酸的替代率相同，所以 A、T、C 和 G 的期望频率是 0.25。因此，应用公式(5-1)是不需要假定核苷酸频率不随时间变化的。

（2）Kimura 两参数法：在实际数据中，转换替代速率常高于颠换速率。Kimura 考虑到这种情况，提出一种估计每个位点核苷酸替代数的方法。该模型中，位点转换替代率(α)不同于颠换替代率(2β)。

用 Kimura 模型，每个核苷酸的平衡频率为 0.25。因此，无论核苷酸初始频率为何，均可应用。这一点和 Jukes-Cantor 模型类似，使得这两个模型较其他模型应用范围更广。

【例 5-1】 人与猕猴的细胞色素 b 基因间的核苷酸替代数估计

动物线粒体 DNA 中的细胞色素 b 基因是高度保守的，因此常被用于研究亲缘关系较远的动物的进化关系。表 5-2 列出了人与猕猴的细胞色素 b 基因的 10 种不同类型核苷酸对的数目，并分别以密码子第 1、2 和 3 位点列出。

表 5-2　人和猕猴线粒体细胞色素 b 基因 DNA 序列中观察到的 10 种核苷酸对

密码子的位置	转换		颠换				相同对				总数	
	TC	AG	TA	TG	CA	CG	TT	CC	AA	GG	n_d	n
第 1	21	22	5	1	5	4	68	93	100	56	58	375
第 2	20	3	6	1	0	2	140	87	71	45	32	375
第 3	60	16	6	5	49	2	11	122	102	2	138	375
合计	101	41	17	7	54	8	219	302	273	103	228	1125

表 5-3 列出了两种不同方法得出的核苷酸替代数估计值 $\hat{d}$。对第 2 密码子来说，两种方法所获得的两种 $\hat{d}$ 值十分接近，$\hat{p}$ 仅略低于相应的 $\hat{d}$ 值。这表明当 $\hat{p}$ 不大时，不论运用何种方法，同一位点

上多重替代的校正实际上并不影响$\hat{d}$值。第1密码子上由两种方法获得的两个估计值$\hat{d}$彼此也相似，虽然它的$\hat{d}$值已接近第2密码子$\hat{d}$值的2倍。然而，在第3密码子上，$\hat{p}$值已充分大，因此多重替代的校正变得不重要。

表5-3 人和猕猴的线粒体细胞色素b基因中第一、第二和第三密码子位置上每位点的替代数估计值

密码子的位置	p	Jukes-Cantor	Kimura
第1	15.5±1.9	17.3±2.4	17.8±2.5
第2	8.5±1.4	9.1±1.6	9.2±1.7
第3	32.8±2.5	50.6±4.9	52.3±5.4

（二）氨基酸序列进化分析

1. 氨基酸差异和不同氨基酸的比例　蛋白质或肽链的进化演变研究开始于两个或多个氨基酸序列的比较。这些不同序列分别来自不同的物种。图5-1显示了六种脊椎动物的血红蛋白α链的氨基酸序列。图中，不同的氨基酸分别用不同的单字母代表。

```
人     V-LSPADKTN  VKAAWGKVGA  HAGEYGAEAL  ERMFLSFPTT  KTYFPHF-DL  SHGSAQVKGH  60
马     .-..A.....  .....S...G  ..........  .....G....  .......-..  ........A.
牛     .-..A...G.  .........G  ..A.......  ..........  .......-..  ..........
袋鼠   .-..A...GH  ...I.....G  .....A..G.  ..T.H.....  .......-..  ......IQA.
蝾螈   MK..AE..H.  ..TT.DHIKG  .EEAL.....  F...T.L.A.  R....AK-..  .E..SFLHS.
鲤鱼   S-..DK..AA  ..I..A.ISP  K.DDI.....  G..LTVY.Q.  ....A.WA..  .P..GP....

人     GKKVA-DALT  NAVAHVDDMP  NALSALSDLH  AHKLRVDPVN  FKLLSHCLLV  TLAAHLPAEF  120
马     .....-.G..  L..G.L..L.  G...D..N..  ..........  .........S  ...V...ND.
牛     .A...-A...  K..E.L..L.  G...E.....  ..........  ......S...  ...S...SD.
袋鼠   ...I.-...G  Q..E.I..L.  GT..K.....  ..........  ..........  .F....GDA.
蝾螈   ....M-G..S  .....I..ID  A..CK...K.  .QD.M...A.  .PK.A.NI..  VMGI..K.HL
鲤鱼   ....IMG.VG  D..SKI..LV  GG.AS..E..  .S......A.  ..I.ANHIV.  GIMFY..GD.

人     TPAVHASLDK  FLASVSTVLT  SKYR  144
马     ..........  ..S.......  ....
牛     ..........  ...N......  ....
袋鼠   ..E.......  ...A......  ....
蝾螈   .YP..C.V..  ..DV.GH...  ....
鲤鱼   P.E..M.V..  .PQNLALA.S  E...
```

图5-1 六种脊椎动物血红蛋白α链的氨基酸序列

一个简单的测度是两序列间的氨基酸差异数(n_d)。如果所有序列的氨基酸数目相同(n)，上述差异数就可用来比较不同序列对间的分歧程度。实际上，当比较很多序列时，氨基酸序列常含有插入或缺失(图5-1)。在这种情况下，计算n_d时一定要删除所有的插入或缺失(间隔)。否则，不同的序列对之间相比较时计算出来的n_d是没有意义的。

实际上，不同蛋白质间序列分歧更方便的测度是两个序列间有差异的氨基酸所占的比例。即使n随不同序列而变化，该比例值(p)也可用于比较分歧程度(表5-4)。公式为：

$$\hat{p} = n_d / n \qquad \text{式 5-2}$$

这一比例值也可称为p距离。假如所有氨基酸位点都以相等概率替代，则n_d遵循二项分布。

在图5-1所给出的例子中，删除所有间隔后可比较的总氨基酸位点数为140。因此，在此例中$n=140$。n_d值出现在表5-4对角线上部，可以很容易地计算出$\hat{p}$，列于对角线下部。当所比较的物种亲缘关系很远时(如人和鸡)，$\hat{p}$值较大。这说明随着两个物种的分歧时间增大，氨基酸的替代数也增大，但p并不严格与分歧时间(t)成比例(图5-2)。

表 5-4 不同脊椎动物血红蛋白 α 链中不同氨基酸的数目(上对角线)及不同氨基酸的比例(下对角线)

	人	大鼠	牛	小鼠	鸡
人		17	17	26	61
大鼠	0.121		17	29	66
牛	0.121	0.121		25	63
小鼠	0.186	0.207	0.179		66
鸡	0.436	0.471	0.450	0.471	

注：计算排除了缺失和插入，使用的氨基酸总数为 140。

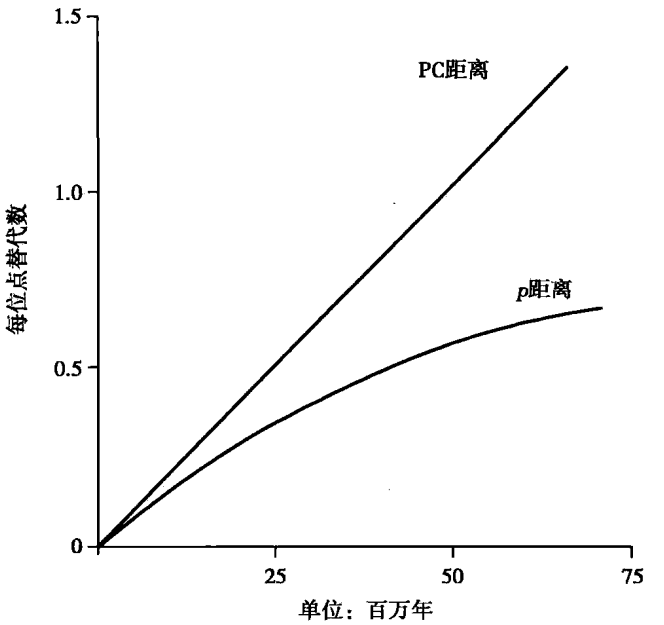


图 5-2 p 距离和泊松校正(PC)距离随分歧时间(t)变化的关系

2. 泊松校正(Poisson correction, PC)距离 p 与 t 的变化呈现非线性关系，原因之一是当多个氨基酸替代出现在同一位点时，n_d 偏离实际氨基酸的替代数将会逐渐增加。运用泊松分布能够更精确估计替代数的方法之一是运用泊松分布的概念。令 r 为一个特定位点每年的氨基酸替换率(简便起见，假设所有位点的 r 都相同)，在 t 年后，每个位点氨基酸替代的平均数为 rt。在一个给定位点氨基酸替代数 k(k=1, 2, 3, …)的发生频率遵循泊松分布，即，

$$P(k;t)=e^{-rt}(rt)^k/k! \qquad \text{式 5-3}$$

因此，在某一位点氨基酸不变的概率是 $p(0;t)=e^{-rt}$。如果多肽链的氨基酸为 n，不变氨基酸的期望值为 ne^{-rt}。

实际上，人们并不知道祖先物种的氨基酸序列。因而，只能对已有 t 年分化的两个同源序列进化比较来估计氨基酸的替代数。由于一个序列的氨基酸无替代概率为 e^{-rt}，因而两个序列同源位点均无替代的概率是：

$$q=(e^{-rt})^2=e^{-2rt} \qquad \text{式 5-4}$$

此概率可用 $1-\hat{P}$ 来估计，而 $q=1-p$。公式中 $q=e^{-2rt}$ 是近似的，因为回复突变和平行突变(在两个不同进化系内出现所导致的同源氨基酸发生同一种突变的情况)，并未加以考虑。当然，除非 $\hat{P}$ 相当大(如 >0.3)，上述突变的作用一般可以忽略。

如果应用公式(5-4)，则两个序列间每个位点氨基酸替代总数($d=2rt$)为

$$d=-\ln(1-p) \qquad \text{式 5-5}$$

分子进化研究中，常常需要知道氨基酸的替代率(r)。如果从其他生物学信息中已弄清了两个序列间的分化时间 t，此速率的估计值为：

$$\hat{r} = \hat{d}/(2t)$$

注意，此处$\hat{d}$被 $2t$ 而不是 t 所除，因为该速率指一个进化系的速率。

3. 自展法的方差和协方差　可以有若干种方法来估计两个序列间氨基酸替代数。实际上，每个模型都是对真实情况的模拟，仅仅提供了氨基酸的近似替代数。因此，前述的估计距离方差的分析公式也是近似的。用最小二乘法估计多个序列构建的系统树的分支长度时，也需要获得不同序列间的距离方差和协方差的估计值。解决这一问题的一个简便途径是应用自展法(bootstrap)计算多种距离测度的方差和协方差。自展法不要求关于$\hat{d}$值分布的假设，只要求每一个位点是独立进化。

假定有 3 个是有进化关系的且均含 n 个氨基酸的序列：

$$x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, \ldots, x_{1n}$$
$$x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, \ldots, x_{2n}$$
$$x_{31}, x_{32}, x_{33}, x_{34}, x_{35}, \ldots, x_{3n}$$

这里，x_{ij} 表示第 i 个序列第 j 个位点上的氨基酸。对序列 1、2，序列 1、3 以及序列 2、3 分别计算$\hat{q}$值，即$\hat{q}_{12}$、$\hat{q}_{13}$ 和$\hat{q}_{23}$。把$\hat{q}_{ij}$ 代入公式，便获得序列 i 和 j 的 PC 距离($\hat{d}_{ij}$)。

在自展法计算方差和协方差时，具有 n 个氨基酸的 3 个序列的随机样本是从原始数据集中产生的。随机样本以伪随机数从原始的数据集中按列有放回随机抽取，形成自展重复抽样数据集。一旦获得了随机样本，便能对 3 对序列的每一对计算出距离的估计值。如此重复 B 次，便能产生 B 个距离值$\hat{d}$。以$\hat{d}_b$表示第 b 次自展重复抽样的$\hat{d}$值，然后可用式(5-6)计算自展方差：

$$V_B(\hat{d}) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{d}_b - \bar{d})^2 \qquad \text{式 5-6}$$

这里，$\bar{d}$是所有重复抽样$\hat{d}_b$的平均值。一般来说，计算 $V_B(\hat{d})$可做约 1000 次重复抽样($B=1000$)。

自展法通常基于一个假设，即所有位点都是独立进化。在位点总数低时，这一假设是不成立的。但如果位点总数很大($n>100$)，如本例，此假设可以成立，因为以不同速率替代的大多数位点在每次自展样本上都会出现。

自展法的一个优点是，在没有数学公式可用时，也能算出方差和协方差，而且能比近似的数学公式提供更好的估计。它能方便地以同样的标准统计公式对任何距离测度计算出方差和协方差。但是，当原始样本太小且存在偏倚时，这种偏倚不能被自展法消除。在这种情况下，解析法将得到比自展法更准确的方差和协方差。

【例 5-2】 由解析法和自展法获得的 PC 距离标准误

表 5-5 列出了由解析式和自展法算出的 PC 距离($\hat{d}$)的标准误，自展法重复了 1000 次。它们是血红蛋白 α 链数据。表 5-5 列出了上述数据集的$\hat{d}$值。显然，由上述两种方法所获得的标准误基本是一致的。对 p 和 Γ 距离，用上述两种方法也可以获得几乎相等的标准误。因此，用自展法估计进化距离的标准误是合适的。

表 5-5　解析法估算的 PC 距离的标准误(下对角阵)及自展法估算的 PC 距离的标准误(上对角阵)

	人	马	牛	袋鼠	蝾螈	鲤鱼
人		0.031	0.031	0.039	0.078	0.083
马	0.031		0.030	0.043	0.083	0.081
牛	0.031	0.031		0.038	0.080	0.079
袋鼠	0.040	0.043	0.039		0.081	0.084
蝾螈	0.074	0.080	0.076	0.080		0.090
鲤鱼	0.082	0.081	0.079	0.086	0.089	

4. Γ 距离 以上所介绍的进化距离都有一个假定，即所有核苷酸位点的替代速率相同。事实上，速率可因位点不同而变化。在蛋白质编码基因中，密码子的第 1、第 2 和第 3 个位置上的替代率是不同的。蛋白质活性中心的氨基酸功能制约也对氨基酸位点间的速率差异有重要影响。在 RNA 编码基因上也观察到速率差异现象，主要是由于 RNA 功能限制及二级结构的影响。不同位点替代速率的统计分析指出，速率变异近似地遵循 Γ 分布。

鉴于上述原因，许多学者致力于发展适用于核苷酸替代的 Γ 距离。一般而言，Γ 距离比非 Γ 距离更符合实际，但前者比后者方差更大。有鉴于此，除非所使用的核苷酸数目非常大，否则 Γ 距离不一定对构建系统树有更优的结果。

二、系统发生树的基本概念及搜索方法

在研究从病毒到人类的各种生物的进化历史中，DNA 或蛋白质序列的系统发育分析已经成为一个重要的工具。由于不同的基因或 DNA 片段的进化速率存在较大的差异，可以通过这些基因或 DNA 片段来估计几乎所有水平上的有机体间的进化关系。系统发育分析对于阐明多基因家族的进化关系，以及理解在分子水平上的适应性进化过程也是十分重要的。

（一）系统发育树的种类

1. 有根树和无根树 基因或生物体的系统发育关系常常用有根或无根的树形结构来表示，即有根树和无根树。树的分支样式称为拓扑结构。对一定规模的分类群（任何分类学单位：属、种、群体和 DNA 序列等），可能的有根树和无根树的拓扑结构数目很大。如果一个类群数为 m 的有根二叉树，其可能的拓扑结构数为：

$$1\cdot 3\cdot 5\cdot\cdots\cdot(2m-3)=[(2m-3)!]/[2^{m-2}(m-2)!],(m\geq 2)$$

若 $m=10$，则有 34 459 425 种有根二叉树。无根树可能的拓扑结构的计算来用 $m-1$ 替换公式中的 m 即可，即 $m=10$ 时，结果为 2 027 025 种。在大多数情况下，大部分可能的拓扑结构可以通过明显不可能的进化关系或其他信息排除。

2. 基因树和物种树 进化学家常常对代表一个物种或群体进化历史的系统发育树感兴趣，这种树称为物种树或种群树。然而，当一个系统发育树由来自各个物种的一个同源基因构建时，得到的树将不完全等同于物种树。当某一坐位出现等位基因多态性时，从不同物种取样的基因分离的时间将比物种分歧时间长。根据基因构建的树的分支结构也可能不同于物种树，称这种树为基因树。同样需要注意的是，如果检测的氨基酸或核苷酸数目较少，重建的基因树和物种树的分支式样也可能不同。因此，可以通过检测大量的氨基酸或核苷酸来避免这种错误。

当所研究的基因属于一个多基因家族时，有可能出现问题。因为构建一个不同物种的系统发育树，应当使用直系同源而不是旁系同源，因为只有直系同源才代表物种形成事件。然而，事实上，要区分直系同源基因和旁系同源基因是很难的。

3. 期望树与现实树 在推断系统发育的理论中，常常假设所研究的 DNA 或蛋白质序列非常长（理论上无限长），从中获得的大量核苷酸或氨基酸均是随机取样。一个用无限长的序列或每一分支的替代数的期望值构建的树称为期望树，建立在实际替代数基础上的树称为现实树，由所观察到的序列数据构建的树称为重建树。期望树、现实树和重建树通常是不同的。大多数构建树的方法的目的是重建现实树，这一类方法包括邻接法、最大简约法和最大似然法等。

当选择构建树的 DNA 序列不同，重建树的拓扑结构和分支长度也将不同，因此，评价物种树或种群树时，应尽量使用多基因。

4. 拓扑距离 两个不同的树之间的拓扑距离通常可以用序列分割的方法来测量。对于无根二叉树，这个距离是有差异内部分支数的两倍。如果两个 8 序列的树具有相同的拓扑结构，则 $d_T=0$，若所有内部分支均产生不同的分割，则 $d_T=10$。然而，如果比较的两个树具有多歧点，则上述规则不起作用，这种情况下，可以使用 Rzhetsky 和 Nei 的普遍性公式计算：

$$d_T=2[\min(q_1,q_2)-p]+|q_1-q_2| \qquad \text{式 5-7}$$

这里，q_1 和 q_2 分别是树 1 和树 2 的内分支树，p 是使两树产生相同序列的分割树。当包含多歧点时，q_1 和 q_2 可能不同；但对于二叉树，q_1 和 q_2 一般是相同的。

（二）基于距离法构建系统发生树

构建系统发生树通常使用的方法分为两大类：距离法和简约法。

构建树的方法一般包括两个过程：拓扑结构的判断和一个既定的拓扑结构分支长度的估计。当拓扑结构已知时，估计分支长度可以用多种统计学方法，如最小二乘法和最大似然法等，问题在于如何判断或重建一个拓扑结构。

系统发育重建的方法具有很大的争议，曾经从事通过形态学特征来研究系统发育的研究者倾向于使用假设条件较少的简约法；从事分子生物学工作的研究者倾向于使用分析法；数学家和统计学家试图建立各种复杂的数学模型，而较少地考虑实际应用。

距离方法　距离方法涉及两个步骤：计算物种对之间的遗传距离以及从距离矩阵重建一棵系统发育树。下面介绍两种不需要分子钟假设的方法：最小二乘法(least-squares，LS)和邻接法(neighbor-joining，NL)。

（1）最小二乘法：最小二乘法(图 5-3)将成对距离矩阵作为给定数据，通过匹配那些尽可能近的距离来估计一棵树上的分支长度，即对给定的和预测的距离差的平方和最小化。预测距离是沿连接两个物种的通路的分支长度总和计算的。距离差的平方和的最小值则是树与数据(距离)相似测度，它可用作树的分值。

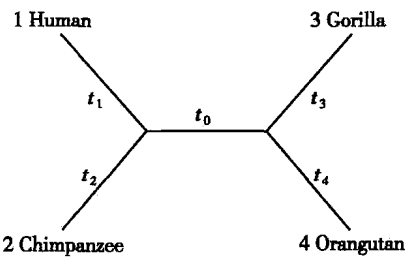


图 5-3　估计支长的最小二乘法标准的示意图

设物种 i 和 j 之间的距离为 d_{ij}，树上物种 i 到 j 间通路的支长和为 $\hat{d}_{ij}$。LS 方法对所有独立的 i 和 j 对求距离差的平方 $(d_{ij}-\hat{d}_{ij})^2$ 的最小值，使得这棵树与距离之间的拟合尽可能地近。例如，对 Brown 等的线粒体数据在 k80 模型下计算成对距离(表 5-6)作为观测数据。现在，考虑树上人、黑猩猩、大猩猩、猩猩及它们的 5 个支长 t_0、t_1、t_2、t_3、t_4。

表 5-6　线粒体 DNA 序列的成对距离

	1. 人	2. 黑猩猩	3. 大猩猩	4. 猩猩
1. 人				
2. 黑猩猩	0.0965			
3. 大猩猩	0.1140	0.1180		
4. 猩猩	0.1849	0.2009	0.1947	

在这棵树上，人与黑猩猩之间的预测距离是 t_1+t_2，人与大猩猩之间的预测距离是 $t_1+t_0+t_3$，依此类推。则距离差的平方和为：

$$\begin{aligned}S&=\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2\\&=(d_{12}-\hat{d}_{12})^2+(d_{13}-\hat{d}_{13})^2+(d_{14}-\hat{d}_{14})^2+(d_{23}-\hat{d}_{23})^2\\&\quad+(d_{24}-\hat{d}_{24})^2+(d_{34}-\hat{d}_{34})^2\end{aligned}$$

S 是 5 个未知支长 t_0、t_1、t_2、t_3、t_4 的函数。最小化 S 的支长值为 LS 估计：$\hat{t}_0=0.008\,840$，$\hat{t}_1=0.043\,266$，$\hat{t}_2=0.053\,280$，$\hat{t}_3=0.058\,908$，$\hat{t}_4=0.135\,795$ 对应的树的分值为 $S=0.000\,035\,47$。对其他两棵树，可以进行类似的计算。的确，其他两棵二元树都趋向于星状树，内分支长估计值为 0。具有最小 S 的树为人，黑猩猩、大猩猩、猩猩称为 LS 树，它是真实系统发育关系的 LS 估计。

用最小二乘法标准确定的树采用同样的标准估计分支长(表 5-7)。如果对支长没有什么约束,就有解析解,可以通过解线性方程获得。非约束方法可以是树重建的一种良好的方法,但是对支长没有明确定义。一些模拟研究建议约束支长为非负值,将改善树重建效果,大多数计算机程序在现实 LS 方法时不采用约束。值得注意的是,当所估计出的支长为负值时,它们多数时候其实是接近于 0。

表 5-7 K80 模型(Kimura, 1980)下的最小二乘法

树	t_0	t_1	t_2	t_3	t_4	S_j
τ:[$(H, C), G, O$]	0.008 840	0.043 266	0.05 328	0.058 908	0.135 795	0.000 035
τ:[$(H, C), C, O$]	0.000 000	0.046 212	0.056 23	0.061 854	0.138 742	0.000 140
τ:[$(H, C), C, O$]	同上					
τ:[$(H, G), C, O$]	同上					

(2) 邻接法:对树进行比较(特别是距离法中)所用的一个标准是以树的支长总和来度量进化总量,支长总和最小的树称为最小进化树(minimum evolution tree)。

邻接法是基于最小进化标准的一种聚类算法。由于它计算快,又能产生合理的树,因而得以广泛应用。它从一个星状树开始,然后加入两个节点,选择能达到树长减少最大的一对。随后,产生一个新节点来替代两个加入的节点将矩阵的维数减少了一次。重复这一过程,直到完全解出这棵树,该算法的每一步都要更新树的支长以及树长。

(三) 基于字母特征构建进化树

最大简约法 在采用等位频率来重建人类种群间的关系时,研究者建议进化树的合理估计为进化总数的最小值,这种方法在应用于离散数据时被称为简约法,而最小进化法在今天被看作是对重复突变进行修正后支长总数最小化的方法。

在一个位点上性状变化的最小数目常常被称作性状长度(character length)或位点长度(site length)。对序列上的所有位点而言,性状长度之和是对整个序列所需要变化的最小数目,称为树长(tree length)、树分值(tree score)或简约分值(parsimony score)。具有最小树分值的树是真实树的估计,称为最大简约树。多棵树是等价最佳树的情况经常见到,尤其是序列非常相似时。

假设在某个特定位点,四个物种的数据是 AAGG,且考虑图 5-4 给出的两棵树所需的最小变化数目。通过将性状状态标注到灭绝的祖先状态节点来计算这个数目。

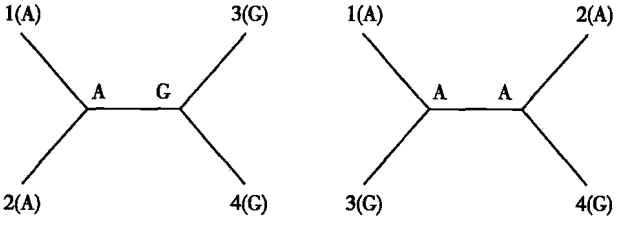


图 5-4 最大简约发建树示意图

对第一棵树,可以通过标注 A 和 G 到两个节点来做到这一点,内支只需要一次变化(A-G)。对第二棵树,可以将 AA(已显示)或 GG(未显示)标注到两个内节点,任何一种情况下,最少都需要两次变化。注意,某位点上被标注为祖先状态的一组性状状态被称为祖先重建(ancestral reconstruction)。对于具有$(n-2)$个内节点的 n 物种的二元树而言,在每个位点重建的总数为 $4(n-2)$(核苷酸)或 $20(n-2)$(氨基酸)。达到变化最小数目的重建称为最简约重建(most parsimonious reconstruction)。因此,对第一棵树,只有一个单一的最简约重建,而对第二棵树,两个重建是等价最简约。

一些位点对树的判别并无贡献,因而是没有信息的。例如,一个恒定位点,即所有物种在该位点具有相同的核苷酸,对任何树都不影响。类似地,单变位点,即两个观察的性状中有一个只出

现一次(例如 TTTC 或 AAGA)——对每棵树只需要一次变化,因而也不是信息位点。一个性状为 AAATAACAAG(对十个物种)的位点也是非信息的,因为对任意树只要对所有祖先节点标注 A 都需要三次变化。对一个简约信息位点(parsimony-informative site)而言,至少要有两个状态被观测到,每一个至少两次。注意,信息位点和非信息位点的概念仅仅只用于简约法。而在距离法或似然法中,所有位点(包括不变位点)都影响计算,应当被包括在内。

常常将所有物种在某个位点上观察到的性状状态看作是位点构型(site configuration)或位点模式(site pattern)。这意味着对四个物种而言只有三种位点式样是有信息的,它们是 xxyy、xyxy 和 xyyx,这里 x 和 y 是任意两个不同状态。很明显,这三种位点式样分别"支持"三棵树,分别是 T1:(1、2),3,4; T2:(1、3),2,4 和 T3:(1、4),2,3。设具有这些位点式样的位点数分别是 $n1$、$n2$ 和 $n3$,如果 $n1$、$n2$ 或 $n3$ 是三个中最大的,则 T1、T2 和 T3 是最简约树。

(四)用于系统发育重建的距离测度

1. 当每个位点的核苷酸替代数目的 Jukes-Cantor 估计值小于 0.05 时,应当使用 p 距离或 Jukes-Cantor 距离,而不管是否存在转换 / 颠换,不管替代速率是否因核苷酸位点而异。

2. 当 $0.05<d<1$,且检验的核苷酸较多时,用 Jukes-Cantor 距离,除非转换 / 颠换比较高($R>5$)。但此比率较高且检测的核苷酸数目很多时,要使用 Kimura 距离。

3. 对于很多序列来说,$d>1$ 时构建的系统树会因为某些原因而不可靠(如存在对位排列错误),因此,建议尽量避免使用这些数据。可以淘汰进化很快的那部分基因区域(如去除免疫球蛋白的超变区基因),仅使用进化速度慢的区域。

4. 当距离很大而 n 很小时,用来估计每个核苷酸位点替代数据的很多距离方法不能使用,在这种情况下,p 距离可以获得相对可靠的拓扑结构。

5. 当一个系统树是通过一个基因的编码区构建时,同义与非同义替换之间的差别就很重要,可以用 dS 来构建树。

6. 普遍地,如果两种距离测度对于同一数据获得相同的距离值(或极为相近)时,应该使用简单的测度,因为它的方差较小。

三、分子钟假说

(一)概述

分子钟(molecular clock)假说认为 DNA 或蛋白质序列的进化速率随时间或进化谱系保持恒定。在 20 世纪 60 年代初期,人们就观察到不同物种中蛋白质序列的差异,如血红蛋白、细胞色素 C 及血纤肽中大致与物种分歧时间成正比。通过这些观察,提出了分子进化钟的概念。

首先需要澄清几点:第一,分子钟应当被看作是氨基酸或核苷酸突变的随机性所导致的随机钟。它不像普通钟表以固定时间间隔跳动,而是以一个随机间隔跳动。第二,不同蛋白质间或蛋白质的不同区域间进化速率的差异很大,因而分子钟假说允许不同蛋白质间进化速率不同,或者说每个蛋白质有其自身固有的分子钟,以不同的速率跳动。第三,速率恒定性未必对所有物种适用,很有可能只存在于某一类群中。例如,可以说就某个特定基因而言,分子钟假说在灵长类中成立。

分子进化的中性学说(neutral theory of molecular evolution)提出之时,分子进化的"似钟特性"被认为"可能是该学说最有力的证据"。中性学说强调相对适应度接近于零的中性或近中性突变的随机固定。分子进化的速率则等于中性突变率,而与环境变化或种群大小等因素无关。如果突变率相似而蛋白质功能在同一类群中保持不变,以至于中性突变比例相同,那么根据中性学说的预测,进化速率将是恒定的。蛋白质间的速率差异则被解释为由于不同蛋白质具有不同的功能限制,因而中性突变的比例不同。

近年来,考古学数据被用来校订分子钟,即将序列间的距离转换成绝对地质时间和置换率。病毒基因分析涉及类似的情况,其进化非常迅速,以至于数年之内就可以观测到变化。人们可以用病

毒被隔离的时间来校正分子钟，并使用与这里讨论基本相同的方法来估计分歧时间。

（二）相对速率检验

最简单的分子钟假设检验是采用第三个物种 C（外类群）来检验两个物种 A 和 B 是否以相同的速率进化。这一检验称为相对速率检验（relative-rate test），其实几乎所有的分子钟检验比较的都是相对速率而不是绝对速率。如果分子钟假说为真，那么从祖先节点 O 到物种 A 和 B 的距离应当相等：dOA = dOB 和 a = b。同理，人们可以得出 dAC = dBC。

（三）内部分支检验

1. 正态偏离（Z）检验　如前所述，推断树的可靠性是通过检验其每个内部分支的可靠性来完成的。这个检验（内部分支检验）适用于由距离法构建的树。考虑 5 序列树，在 5 序列的情况下，有 15 种可能的无根二分歧树，每个树由 5 个外部分支和 2 个内部分支组成。假设拓扑结构 A 是正确的，而其他的都是不正确的，则表明正确拓扑结构的所有分支长度估计的期望值是 0 或者正值，而不正确拓扑结构中至少有一个内部分支长度为负值，且该分支产生了序列间的一个不正确分区。只要使用无偏距离估计而分支长度用 LS 方法估计，则对于任何数目的序列构造的树进行检验似乎都是正确的。因此，如果一个树的某个内部分支估计值被确定为负值，该树的拓扑结构很可能就是错误的。

上述的零假设检验能相当方便地应用于由距离法（特别是由 NJ 或者 ME 方法）获得的树的分析上，因为只有正确树的所有内部分支才可能是正值的。但在 MP 和 ML 树中，不管拓扑结构如何，所有内部分支都为正值，因此，就很难建立出一种检验零假设的分析方法。然而，使用自展法可以检验零假设。

2. 自展内部分支检验　另一种用于距离树内部分支检验的是自展内部分支检验。这种方法是检验一个给定树的每个内部分支的可靠性。与自展检验法相似，从原始序列中随机抽样形成与原始数据数目相同的核苷酸（或者氨基酸），再用从原始序列数据获得的树拓扑结构来计算所有分支长度，并对同一种拓扑结构重复数百次。一个内部分支的长度估计 b 将随着重复次数变化而不同，且可能为负值。可以计算 b 的平均数以及标准误，并进行 Z 检验。

该检验结果通常与上述分析方法获得的结果非常相似。但是该方法优于解析法，即无需分别计算每个替代模型 b 的标准误；所有替代模型的标准误可用同样的方法计算。因此计算时间不会随序列数增加而迅速增加。这个方法比解析法更易运用。然而，当核苷酸或者氨基酸数目小时，该方法可能会给出 Pc 的有偏估计，这是因为如果原始样本有偏差，则此偏差在重复抽样时不能被除去。在这种情况下，解析法要好得多。

第三节　核苷酸和蛋白质的适应性进化

Section 3　Adaptive Evolutions of Nucleotide and Protein

基因和基因组的适应性进化最终决定形态、行为和生理上的适应，以及物种分歧和进化创新（evolutionary innovation）。因此，在分子进化研究中，分子适应是一个令人振奋的课题。尽管自然选择在形成形态和行为进化方面似乎普遍存在，但它在基因和基因组进化中所起的作用尚存在争议。分子进化的中性学说认为，种内和种间大多数可见差异不是由自然选择，而是由适合度很小的随机突变的固定决定的。近 40 年来人们发展了一系列中性检验方法，本节介绍正选择和负选择的基本概念以及分子进化的主要理论，还将简要介绍几种群体遗传学中发展起来的常用的中性检验方法。另外引入应用范围比较广的 dN/dS 检验，并且详细地介绍了其计算方法。

一、中性与近中性理论

在群体遗传学中，一个新突变基因 a 与野生型显性基因 A 的相对适合度由选择系数 s 来度量。设基因型 AA、Aa 和 aa 的相对适合度分别为 1、1 + s 和 1 + 2s，则 s < 0、= 0 及 > 0 分别对应负选择

(negative selection)或净化选择(purify selection)、中性进化和正选择(positive selection)。新突变基因的频率各世代高低不同,既受自然选择又受随机漂变的影响。究竟是随机漂变还是自然选择决定了突变的命运取决于 Ns(N 为有效群体的大小)。若 $|Ns|\gg1$,则自然选择决定基因命运;若 $|Ns|$ 接近于 0,则随机漂变的作用非常重要,而且该突变为中性或近中性。

按照中性理论,今天观察到的遗传变异——无论是种内多态性还是种间分歧,均不取决于自然选择所驱动的有利突变的固定,而是取决于那些事实上没有适合效应(即中性的)突变的随机固定。下面是该理论的一些观点和预测。

(1) 大多数突变是有害的,会被净化选择所清除。

(2) 核苷酸置换率等于中性突变率(即总突变率乘以中性突变所占比例)。如果物种间中性突变率恒定(或者日历时间或者世代时间),则置换率也是恒定的。这个预测为分子钟假说提供了解释。

(3) 功能较重要的基因或基因区域进化较慢。在具有较重要作用或处于较强功能约束下的一个基因中,中性突变比例较小,使得核苷酸置换率较低。现在,功能重要性和置换率之间的负相关在分子进化中是一个普遍现象。例如,替代置换率几乎总是比沉默置换率低;密码子第 3 位比第 1 和第 2 位进化更快;具有相似化学性质的氨基酸比不相似的氨基酸更容易相互替代。如果自然选择在分子水平上驱动进化过程。那么可想而知,功能重要的基因的进化速率比功能不重要的基因要高。

(4) 种内多态性和种间分歧是中性进化同一过程的两个阶段。

(5) 形态特征(包括生理、行为等)的进化的确是自然选择所驱动的。中性学说关注的是分子水平上的进化。

围绕中性理论的争论已产生很多的群体遗传理论和分析工具。下面将讨论其中几种。

二、基因适应性进化的统计学检验方法

以下几个是典型的统计学研究适应性进化的方法,已经形成了稳定的软件。根据输入数据的不同可以检验相应基因的选择强度。

1. Tajima 的 D 检验　在随机交配的群体中,一个中性基因上保持的遗传变异量由 $\theta=4N\mu$ 决定,这里 N 为(有效)群体大小,μ 为每一代的突变率。从每个位点的角度定义 θ,它也是从群体中随机抽取的每条序列的期望位点杂合度。例如,在人类非编码 DNA 中,$\hat{\theta}\sim0.0005$,意味着两条随机的人类序列间大约 0.05% 的位点不同。群体数据一般很少有变异,所以通常采用无限位点模型,假定每个突变都发生在 DNA 序列的不同位点上,且无须校正多重命中。注意,群体规模大和突变率高都会导致群体中保持更高的遗传变异。

两种从群体中随机抽取 DNA 序列的简单方法可以用来估计 θ。第一种是包含 n 条序列的样本中的多态性位点数 S,期望值 $E(S)=L\theta a_n$,这里的 L 为序列中的位点数,$a_n=\sum_{i=1}^{n-1}1/i$,故 θ 可由 $\hat{\theta}_S=S/(La)$ 估计。第二种方法是对 n 条序列所有成对比较的核苷酸差异的平均比例值的期望为 θ,将 θ 作为一个估计值,则记作 $\hat{\theta}_\pi$。这两种 θ 的估计在中性突变模型下均无偏,即假定无选择、无重组、无群体分化或大小变化,以及突变和漂变之间平衡。然而,如果模型的假设不成立,则不同因素对 $\hat{\theta}_S$ 和 $\hat{\theta}_\pi$ 有不同影响。例如,若轻微有害突变在群体中保持较低频率能显著增加 S 和 $\hat{\theta}_S$ 值,但对 $\hat{\theta}_\pi$ 几乎没有影响。θ 的两个估计量可以为了解造成严格中性模型失效的因素和机制提供信息。因此,Tajima 构建了以下的检验统计量:

$$D=\frac{\hat{\theta}_\pi-\hat{\theta}_S}{SE(\hat{\theta}_\pi-\hat{\theta}_S)} \qquad \text{式 5-8}$$

这里,SE 为标准误差。

在无效中性模型下,D 的均值为 0,方差为 1。Tajima 建议采用标准正态分布和 β 分布来确定 D 是否显著不同于 0。

Tajima 的 D 检验的统计显著性可能与几种不同的解释相容，而且难以区分它们。正如前面所讨论的，一个负 D 值表明存在净化选择或群体中分离的轻微有害突变。然而，负 D 值也可能是由群体扩张造成的。在一个扩张群体中，可能分离出许多新的突变，且它们在数据中以单元(singletion)形式出现，即其他所有序列在此位点上都相同，只有一条序列不同。单元增加了分离位点的个数并导致 D 值为负。类似地，D 值为正可解释为平衡选择将突变维持在居中频率。然而，一个收缩的群体也能够导致 D 值为正。

2. Fu 和 Li 的 D 检验与 Fay 和 Wu 的 H 检验　在 n 条序列的一个样本中，一个多态位点上突变核苷酸的频率为 $r=1,2,\cdots,n-1$。样本中观察到的突变的这种分布成为位点频谱(site-frequency spectrum)。通常，采用亲缘关系很近的外类群来推断祖先的和衍生的核苷酸状态。例如，若在一个 $n=5$ 的样本中观察到的核苷酸为 AACCC，而外类群中为 A(假定的祖先状态)，则 $r=3$。Fu 设 r 为突变规模。如果祖先状态未知，则不可能区分突变规模是 r 还是 $n-r$，使得那些突变被划为同一类，位点频谱则被认为是折叠的，折叠构象提供的信息远少于非折叠构象，因而，采用外类群来推断祖先状态应当增加检验效力，但缺点是该检验可能会受到祖先重建中误差的影响。

Fu 和 Li 区分了内部突变和外部突变，即分别在系谱树内支或外支上发生的突变。设这两类突变的个数分别为 η_I 和 η_E，注意 η_E 为单突变的个数，他们构建了以下的统计量

$$D=\frac{\eta_I-(a_n-1)\eta_E}{SE(\eta_I-(a_n-1)\eta_E)} \qquad \text{式 5-9}$$

这里，$a_n=\sum_{i=1}^{n-1}1/i$，SE 为标准误差。与 Tajima D 检验相类似，该统计量也是作为中性模型下 θ 的两个估计值间的差异来构建的。Fu 和 Li 认为群体中分离的有害突变倾向于近期产生，位于树的外支，且对 η_E 起作用；而内支上的突变多为中性，且影响 η_I。

3. McDonald-Kreitman 检验和选择强度估计　中性学说认为种内多样性(多态性)和种间分歧是同一进化过程的两个阶段，即两者都是由中性选择突变的随机漂变所致。因而，如果同义和非同义突变都是中性的，则种内同义和非同义多态性的比例应与种间同义和非同义差异的比例相同。

近缘物种蛋白质编码基因中的可变位点可依位点是否具有多态性或固定差异，以及该差异是同义还是非同义的，划分为一个 2×2 列表中的 4 类(表 5-8)。假设从物种 1 中抽取 5 条序列，从物种 2 中抽取 4 条序列，若某位点在物种 1 中数据为 AAAAA，在物种 2 中为 GGGG，则该差异被称为固定差异。若某位点在物种 1 中的数据 AGAGA，而在物种 2 中为 AAAA，则该位点被称为多态性位点。注意，无限位点模型无需对隐藏变化进行校正。如果数目不多，则中性无效假设等价于列表的行和列之间独立并可被 χ^2 分布或 Fisher 精确检验验证。McDonald 和 Kreitman 测定了果蝇 3 个亚群的乙醇脱氢酶基因(Adh)序列，获得了表 5-8 中列出的数据。P 值小于 0.006，说明与中性期望有显著偏差。种间替代突变远多于种内替代突变。他们将此模式认作驱动种间差异的正选择证据。

表 5-8　果蝇 Adh 基因中存在沉默突变、置换突变以及多态性位点个数

变化类型	固定差异	多态性
置换(非同义)	7	2
沉默(同义)	17	42

注：数据来自 McDonald and Kreitman，1991。

为了弄清这个解释后面的推论，假定同义突变是中性的，考虑选择对物种分歧之后出现的非同义突变的影响。人们预期有利替代突变会很快固定下来并成为种间的固定差异。因而，若固定的替代突变过剩(如同在 Adh 中观察到的)，则表明存在正选择。

人们在哺乳动物线粒体基因中已观察到过剩的替代多态性，表明净化选择下存在轻微有害替代突变。有害突变被净化选择清除，而且不会在种间比较中看见，但在种内还是会分离。

三、d_N 或 d_S 检验

蛋白质编码序列区分为同义置换和非同义置换，对理解自然选择的作用来说，这比内含子或非编码序列优越得多。若将同义置换率作为基准点，可以推断自然选择在非同义置换固定过程中是推动还是阻碍作用。非同义 / 同义置换率的比率($\omega=d_N/d_S$)可以在蛋白质水平度量选择压力。如果选择对适合度没有影响，则非同义突变将以与同义突变相同的速率被固定，使得 $d_N=d_S$ 及 $\omega=1$。如果非同义突变是有害的，则净化选择将降低其固定速率，使得 $d_N<d_S$ 及 $\omega<1$。如果非同义突变受到达尔文选择的青睐，则其被固定的速率将高于同义突变，致使 $d_N>d_S$ 及 $\omega>1$。因此，非同义突变率显著高于同义突变率即为蛋白质适应性进化的证据。

然而，可以预料一个功能蛋白上的大多数位点在大部分进化时间都是受约束的。即使发生正选择，也只能影响几个位点，且只有偶尔发生。因此，这种成对平均方法很少检测到正选择。近期研究着重检测影响系统发育关系中特定谱系或蛋白质中单个位点的正选择。

对编码蛋白质的 DNA 序列，同义和非同义置换被定义为平均每个同义位点上的同义置换数(d_S 或 K_S)以及平均每个非同义位点上的非同义置换数(d_N 或 K_A)。

本节主要使用记数法计算，计数方法类似于 JC69 等核苷酸置换模型下的距离计算，有三个步骤：①对同义和非同义位点计数；②对同义和非同义差异计数；③计算差异比例并校正多重命中(multiple hit)。将位点和差异都计数后，就可以区分同义和非同义这两种类型间的差异了。

1. 位点计数　每个密码子都有 3 个核苷酸位点，分成同义和非同义两类。以密码子 TTT(Phe)为例，由于 3 个密码子位置上每个核苷酸都可以转变为另外 3 种核苷酸，该密码子就有 9 个直接邻居：TTC(Phe)、TTA(Leu)、TTG(Leu)、TCT(Ser)、TAT(Tyr)、TGT(Cys)、CTT(Leu)、ATT(Ile) 和 GTT(Val)。其中，密码子 TTC 和密码子 TTT 编码同一个氨基酸。因此，对密码子 TTT 而言，就有 $3\times1/9=1/3$ 个同义位点，$3\times8/9=8/3$ 个非同义位点(表 5-9)。在计数过程中，不计入变为终止密码子的突变。将该方法用于序列 1 中的所有密码子，并将计数结果相加以获得全序列中同义和非同义位点的总数。然后，对序列 2 重复该过程并计算两条序列间的平均位点数目，分别计为 S 和 N，有 $S+N=3\times L_c$，这里 L_c 为序列中的密码子的数目。

表 5-9　密码子 TTT(Phe)中的位点计数

目标密码子	突变类型	置换率($\kappa=1$)	置换率($\kappa=2$)
TTC(Phe)	同义	1	2
TTA(Leu)	非同义	1	1
TTG(Leu)	非同义	1	1
TCT(Ser)	非同义	1	2
TAT(Tyr)	非同义	1	1
TGT(Cys)	非同义	1	1
CTT(Leu)	非同义	1	2
ATT(Ile)	非同义	1	1
GTT(Val)	非同义	1	1
总和		9	12
同义位点数		1/3	1/2
非同义位点数		8/3	5/2

注：κ 为转换 / 颠换置换率比率。

2. 变异计数 第二步是对两条序列间的同义和非同义变异进行计数。换言之，在两条序列间所观测的差异可按同义和非同义划分。再按密码子逐一处理。很明显，如果两个所比较的密码子相同（如 TTT 对 TTT），则同义和非同义变异数目为 0；如果两个所比较的密码子间仅在一个位置上存在差异（TTC 对 TTA），就很容易发现这种单一的变异是同义的还是非同义的。然而，如果两个比较的密码子间在 2~3 个位置上都存在差异（如 CCT 对 CAG 或 GTC 对 ACT），则有 4~6 条进化途径能使一个密码子变成另一个密码子。多条途径中可能涉及同义和非同义差异数不同。大部分计数方法对不同途径赋予同等权重。

例如，密码子 CCT 和 CAG 间存在两条途径（表 5-10）。第一条途径要通过中间密码子 CAT 转换，涉及两个非同义变异；而第二条途径通过中间密码子 CCG 转换，涉及一个同义变异和一个非同义变异。如果对这两条途径赋予相同权重，则两个密码子间有 0.5 个同义变异和 1.5 个非同义变异。如果同义突变率高于非同义突变率，如同几乎所有基因中表现的一样，第二条途径应该比第一条途径的可能性更大，预先不知道 d_N/d_S 比率和序列分歧度，就很难对不同途径赋予合适的权重。不过，计算机模拟结果表明加权对估计值的影响很小，尤其是当序列的分歧度并不是很大时。

表 5-10 密码子 CCT 和 CAG 间的两条途径

途径	差异	
	同义	非同义
CCT(Pro)↔CAT(His)↔CAG(Gln)	0	2
CCT(Pro)↔CCG(Pro)↔CAG(Gln)	1	1
平均	0.5	1.5

计数沿着序列密码子逐一进行，将差异数相加得到两条序列间总的同义和非同义差异数，分别记为 S_d 和 N_d。

3. 多重命中校正 现在有：

$$p_S = S_d / S$$
$$p_N = N_d / N \qquad \text{式 5-10}$$

分别是同义和非同义位点上的差异比例，它们等同于针对核苷酸的 JC69 模型下的差异比例。因此，套用 JC69 中对多重命中的校正。

$$d_S = -\frac{3}{4}\log\left(1-\frac{4}{3}p_S\right)$$
$$d_N = -\frac{3}{4}\log\left(1-\frac{4}{3}p_N\right) \qquad \text{式 5-11}$$

当只关注同义位点和差异时，每个核苷酸并不存在 3 个其他核苷酸来突变的情况。实际上，对多重击中校正的作用很少，至少在序列分歧度不高时如此，故校正公式带来的偏差也就不是非常重要了。

4. rbcL 基因应用实例 应用上述方法来估计黄瓜和烟草中叶绿体蛋白 1，2- 二磷酸核酮糖羧化酶 / 加氧酶大亚基（rbcL）基因间的 d_S 和 d_N。黄瓜（Cucumissativus）rbcL 基因的 Genbank 序列号为 NC_007144，烟草（Nicotiana tabacum）为 Z00044。在黄瓜和烟草基因中分别有 476 个和 477 个密码子，对位排列后的序列则有 481 个密码子。删除了任意一个物种对位排列时出现的间隔密码子，这样序列中就剩下 472 个密码子。

表 5-11 列举了数据的一些基本统计值，它们是对 3 个密码子位置分别进行分析后获得的。碱基组成不等，第三个密码子富含 A/T。3 个密码子位置的转换 / 颠换置换频率的比率估计值大小依次为 $\hat{\kappa}_3 > \hat{\kappa}_1 > \hat{\kappa}_2$。序列距离的估计值也是同样的顺序 $\hat{d}_3 > \hat{d}_1 > \hat{d}_2$。这类模式在蛋白编码基因中很常见，反

映了遗传编码结构以及基本上所有氨基酸都处于选择压力之下，同义置换率高于非同义置换率。当对密码子逐一进行检测时，两个物种间有 345 个密码子是一致的，115 个密码子在一个位置上有差异，其中 95 个是同义的，20 个是非同义的。10 个密码子在两个位置上有差异，2 个密码子在 3 个位置上均不相同。

表 5-11 黄瓜和烟草 rbcL 基因的基本统计量

位置	位点	π_T	π_C	π_A	π_G	$\hat{\kappa}$	$\hat{d}$
1	472	0.179	0.196	0.239	0.386	2.202	0.057
2	472	0.270	0.226	0.299	0.206	2.063	0.026
3	472	0.423	0.145	0.293	0.139	6.901	0.282
总计	1416	0.291	0.189	0.277	0.243	3.973	0.108

随后，1416 个核苷酸位点被分为 $S=343.5$ 个同义位点以及 $N=1072.5$ 个非同义位点。在两条序列间观察到 141 个差异，这些差异分为 $S_d=103.0$ 个同义差异和 $N_d=38.0$ 个非同义差异。因此，在同义和非同义位点上的差异比例分别为 $p_S=S_d/S=0.300$ 和 $p_N=N_d/N=0.035$。使用 JC69 校正后得到 $d_S=0.383$ 和 $d_N=0.036$，其比值 $\hat{\omega}=d_N/d_S=0.095$。根据这一估计，该蛋白质处于强烈的选择压力之下，在群体中发生一个非同义突变的概率只有同义突变的 9.5%。

四、适应性进化基因

基于 ω 比率检验获得的大多数正选择基因可分为以下三类。第一类包括针对病毒、细菌、真菌和寄生虫攻击的防御机制或免疫作用中的宿主基因，以及与破坏宿主防御机制有关的病毒或病原基因。例如，前者包括主要组织相容性复合体、淋巴细胞蛋白 CD54、植物中与识别病原有关的 R 基因及哺乳动物中反转录病毒抑制剂 TRIM5α；后者包括病毒表面或包膜蛋白、疟原虫细胞膜表面抗原以及由植物天敌(如细菌、真菌、卵菌、线虫和昆虫)产生的多糖。可见，病原基因由于受到正选择进化出不被宿主防御机制识别的新类型，同时宿主也必须适应并识别出病原，这就激发了一场进化“军备竞赛”，驱动新的替代突变在宿主和病原中固定。蛇或蝎子毒液中的毒素用于捕获猎物，也处于类似选择压力下，因而进化速率很快。第二类主要包括与生殖有关的蛋白质或信息素。一些研究已检测到有关精 - 卵识别的蛋白质及雄性或雌性生殖其他方面的快速进化。这些基因上的自然选择也可能加速或导致新物种形成。第三类正选择基因与上述两类有所重叠，包括基因复制后获得新功能的基因。基因复制是基因、基因组和遗传系统进化的初级驱动力，被认为在新基因功能进化中起引领作用。复制基因的命运由能否为机体带来选择优势所决定，多数复制基因被清除或因有害突变失去功能而退化为假基因。由于亲代基因需要不同功能，有时新拷贝会在适应进化驱动下获得新功能。已检测到许多基因在基因复制后经历加速蛋白质进化，其中包括灵长类 DAZ 基因家族、灵长类绒毛促性腺蛋白。群体遗传检验也表明正选择在复制核基因早期进化动态中的重要作用。

还有很多其他基因也被检测处于正选择之下，尽管它们不如那些参与到进化军备竞赛中的基因(如宿主 - 病原拮抗作用及生殖)那么多。这也许是基于 ω 比率的检验方法的局限性所致，即可能错过一次性的适应性进化。在这种进化中，一个有利突变出现并迅速在群体中扩散开来，接踵而至的就是净化选择。若要检测到更多正选择，也许需要改进能检测影响某个谱系上少数位点的插曲式或局部的进化方法。

统计检验不能证明基因是否真正经历适应性进化。具有信服力的例子也许要建立在实验验证和功能检验上，两者在观察到的核酸变化与蛋白质折叠以及表型变化(如催化化学反应的效率不同)之间建立直接联系。

第四节 分子进化与生物信息学

Section 4 Molecular Evolution and Bioinformatics

一、基因组进化概述

基因组学(genomics)是一门只有10多年历史的新兴学科，发展极为迅速，并产生了许多分支学科。随着研究的不断深入，它已从结构基因组学(structural genomics)进入到功能基因组学(functional genomics)。利用基因组学研究的方法和成果来研究生物进化，也就是进化基因组学(evolutionary genomics)所要研究的问题，越来越受到进化生物学研究者的关注。

目前，尽管进化基因组学还没有正式列在基因组学的议事日程上，但也已经有了不少相关的研究，比较基因组学(comparative genomics)就是其中之一。对不同生物基因组结构的异同及其特点进行比较，除了在功能基因组学的研究上很有意义外，还有可能在一定程度上了解基因组的进化，特别是基因组的结构特征与生物复杂性的关系。例如，通过比较，发现基因组中蛋白质和功能RNA基因的密度与生物的复杂程度有一定的负相关。在细菌基因组中，基因的平均密度是1个基因/1kb；在酵母中是1个基因/2kb；而线虫是1个基因/5kb；果蝇是1个基因/13kb；到人类则是1个基因/40kb。这种密度的变化显然是与基因组进化中调控元件和“非基因序列”的扩增有关。

比较基因组学的研究还表明，基因和基因组是由并非很多的基本结构单位(构件)构成的，而这些构件在进化中被反复使用(重组)。以形成新的基因和基因组，这就像为数不多的化学元素可以组成无数的化学物质(分子)那样。新的化学分子是通过已有元素或分子之间的化学反应产生的，所以，基因组的进化有可能以化学反应作为其动态模型，即新基因组的产生是通过已有基因或基因组的重组、重排、重新建立新的关系而达成。要充分认识这种类比的意义，就必须开展进化基因组学的研究。

基因组的进化与基因组的三维结构之间显然也有很重要的关系。人与黑猩猩DNA序列的相似程度达99%，两者的差异很可能是在其基因组的三维结构(包括三维调控关系)上。因此，进化基因组学必将深入进行这方面的研究。

为了解基因组及其发展变化的本质，当然还要研究与生命起源有关的最原始的基因和基因组的起源，以及其后的进化模式与过程，这样就有可能在分子水平上认识生物进化的分段途径。总之，进化基因组学将是基因组学中最触及事物本质的一个分支。

二、病毒基因组分析

(一)病毒基因组

对生物的分类应该体现其系统演化。对病毒来说，它的生命是相对脆弱的，很难达到像古细菌、细菌和真核生物那样综合全面的程度。病毒也受突变和自然选择的影响，并且病毒基因组的进化速度远远超过其他细胞的基因组。有很多证据证明，早在一万年前病毒就已经存在，这些证据包括人类的骨骸残骸，历史记录和遗物。然而，远古病毒的DNA或RNA还没有被找到。

RNA病毒基因组的RNA聚合酶一般缺乏校正能力。这导致基因组的突变率比DNA基因组高100万~1000万倍。对于DNA病毒，其突变率一般比宿主细胞高10~1000倍。除了高突变率，许多病毒的复制速度也是非常惊人的。单个细胞能产生10 000个脊髓灰质炎病毒颗粒，而一个被艾滋病病毒感染的个体一天能产生10亿个病毒颗粒。许多病毒的基因组由相对独立的多个片段组成。这些片段能够在病毒复制过程中随机重组，从而在子代病毒中产生大量不相同的子类。流感病毒几乎每年都能引起大范围的疾病流行就是这个原理的体现。病毒经常处于强大的选择压力下，如宿主的免疫反应或抗病毒药物作用。因此，艾滋病病毒快速的突变和复制确保某些病毒株通过突变产生对抗病毒药物的抗性，而且会经受环境的选择而存活下来。

病毒经过漫长的进化历程已经能够侵入系统发生树中所有物种：古细菌、细菌和真核生物。植物病毒(番茄丛矮病毒)、动物病毒(如SV40病毒，鼻病毒和脊髓灰质炎病毒)以及噬菌体(如噬菌体ΦX174)的衣壳蛋白中都有“β-折叠桶”或“果冻卷”折叠结构。除非发生了显著的趋同进化，否则这种现象一般说明这些病毒是同源的。感染植物和动物的反转录病毒具有双链RNA基因组以及封装它的特殊衣壳体。有一类噬菌体(Φ6)也具有这种特征，也说明了感染不同物种的病毒之间具有同源性。在对这些病毒基因组以及蛋白质的分析中并没有发现序列相似性，再次凸显了病毒基因组高速进化的特点。病毒基因组的高度多样性使人们无法根据其序列数据绘制出涵盖所有病毒的全面完整的系统发生树，这反映了病毒基因组形成历程中复杂的分子进化事件。

(二) 病例研究：SARS流行病的系统发生分析

2003年2月28日，曾暴发一场大规模的流行系疾病，经确认，命名为急性呼吸系统综合征(SARS)。同年3月15日，WHO发布全球警告，称SARS为“世界范围的健康威胁”。他们警告可能的地点包括加拿大、印度尼西亚、菲律宾、新加坡、泰国和越南。

流行病的起源：尽管SARS的起源和原因还不知道，但应该离人们知道的时间不远，通过分析多个SARS基因组就可以知道这个疾病是怎样发生和它的起源以及如何在许多国家扩散的。在2003年3月的第3周，美国、加拿大、德国，以及中国香港分别独立的从SARS患者身上分离出新的冠状病毒(SARS-CoV)。

通过分析大量的完整病毒基因组数据集，可以回答很多重要的问题。下面将提供一些工具来回答这些问题中的一部分。是怎样一种病毒导致了这样一场流行病？这种病毒的原始宿主是什么？跨越物种障碍的时间和地点？是怎样一个关键突变让这种转换成为可能？

为了回答这些问题，首先要了解一些系统发生分析关键算法，这些在前面章节中已经提到过，然后把这些算法应用于2003年获得的SARS数据(所有这些数据都可从Genbank获得)。

1. SARS基因组　SARS-CoV基因组是在2003年4月由加拿大团队获得的，29 751bp的单链RNA序列。可以通过GenBank获得这个数据(查询编号为AY274119.3)。在图5-5中提供了该病毒的基因图谱。其GC含量大概是41%，是已经公布的冠状病毒基因组GC含量范围之内的。并且由一个典型的冠状病毒结构，按照一定的顺序排列5个或者6个基因。

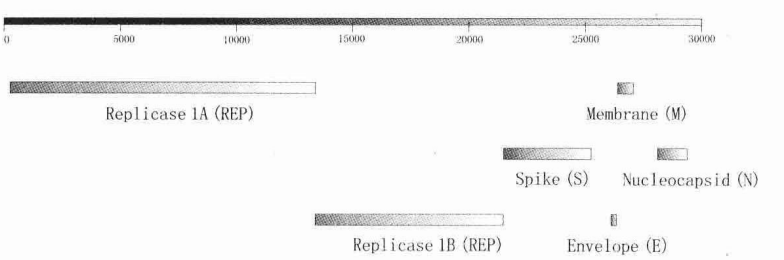


图5-5　SARS病毒基因图谱

2. SARS流行发生重构　在SARS流行病发生的时候，有关其起源和本质等许多关键的问题都可以通过基因组序列分析来获得。在2003年早期多个团体就已经获得和发布了SARS的序列，并以此作为基础探寻流行病起源和扩散，现在可以用GenBank中的许多病毒序列来研究这次流行病。选取了13条已知获取时间和地点的序列，然后展示如何用这些序列来挖掘这次流行病的信息。

鉴定宿主：SARS病毒在早期被认为是冠状病毒，和已知的其他冠状病毒有相同序列的基因。然而，它又是完全不同于其他已知人类冠状病毒，因此，很可能是从其他动物中起源的。使用多种动物冠状病毒的蛋白质构建了邻接树，其中包括了在果子狸中发现的冠状病毒。SARS看起来和果子狸冠状病毒最相近，和人类其他冠状病毒都比较远。

使用表5-12的13个基因组，用邻接法构建了系统发生树(图5-6)，这种疾病并不是通过鸟类携带的，而是起源于果子狸，后在人类中传播。这个距离矩阵是通过Jukes-Cantor模型计算的距离矩阵

表 5-12 SARS 病毒检验时间及地方

Table	Name, location, and sampling date of SARS virus isolates used in our case study		
Name of isolate	Acc.number	Date	Location
GZ01	AY278489	DEC-12-2002	Guangzhou
ZS-A	AY394997	DEC-22-2002	Zhongshan
ZS-C	AY395004	JAN-04-2003	Zhongshan
GZ-B	AY394978	JAN-24-2003	Guangzhou
HZS-2A	AY394983	JAN-31-2003	Guangzhou
GZ-50	AY304495	FEB-18-2002	Guangzhou
CUHK-W1	AY278554	FEB-21-2003	Hong Kong
Urbani	AY278741	FEB-22-2003	Hanoi
Tor 2	AY274119	FEB-27-2003	Toronto
Sin2500	AY283794	MAR-01-2003	Singapore
TW1	AY291451	MAR-08-2003	Taiwan
CUHK-AG01	AY345986	MAR-19-2003	Hong Kong
Palm civet	AY627048		

并且用核苷酸序列做全局比对进行校正作为遗传距离。

从这棵树上，能够了解这次流行病的整个过程。如果把果子狸作为外类群，可以看到所有早期的病例都是发生在广州，并且 Hotel Metropole 冠状病毒几乎和它们中的一条序列是完全一致的。

因为已经知道每个测序的 SARS 病毒收集的时间，这样就能观察到经过若干时间突变的过程。方便起见，使用了 spike 蛋白质对应的开放读码框。相对于从果子狸获得的序列，看到其遗传距离随着时间在粗略按线性模式逐渐提高(x 轴表示时间，原点代表 2003 年 1 月 1 日)。如果在这些数据中插入最小二乘法的拟合曲线，就可以估计这次流行病起源的大概时间。任何一个在零点附近日期都可能是开始的时间，估计在 2002 年 9 月 16 日到 2003 年 1 月 1 日之间。这种方法是比较粗糙的，而且其中很多假设还都没有证实，但仍然给人们一个很可能的时间点，最早的病例报道可以追溯到 2002 年的下半年。

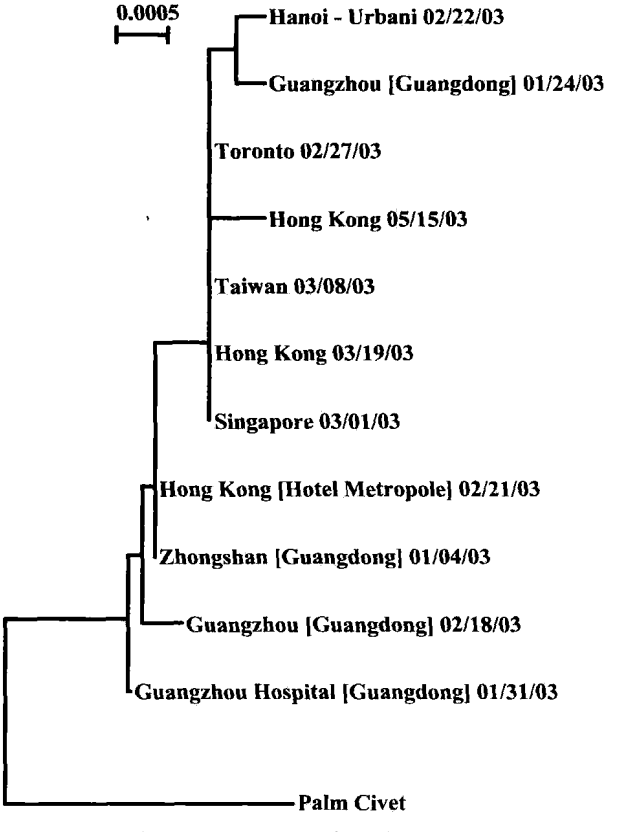


图 5-6 SARS 病毒系统发生树

三、原核生物基因组比较

(一) 与人类疾病相关的细菌分类

细菌和真核生物已经相互“交战”几百年了。细菌为了繁殖需要占据人体这个营养丰富的环境。典型的细菌“殖民地”包括皮肤、呼吸道、消化道(口腔、大肠)、尿道和生殖系统等。据估计每个人身上的细菌数目超过自身的细胞数目。大多数情况下，这些细菌对人类是无害的。然而，有些细菌在

一定条件下能够导致感染，甚至带来灾难性的后果。最近一些年，由于广泛使用抗生素导致了细菌抗药性的增强，因此亟须找到细菌的毒性因子，然后找到相应的接种疫苗。对这个问题的一个解决办法就是比较细菌的致病株和非致病株。

（二）原核生物基因组比较数据库

NCBI 提供了一个非常有效的基因组比较工具，并且使用起来非常容易。从基因组查询页面上，选择 Drosophila melanogaster(果蝇)就得到图 5-7 所示的页面。选择 TaxPlot，就能够将两个基因组和一个参考基因组，(如 caenorhabditis elegans 和 saccharomyces cerevisiae)进行比较。在这个图上，每一个点都代表参考基因组中的一个蛋白质。x 坐标和 y 坐标显示了被比较蛋白质组中每个蛋白质最佳匹配的 BLAST 分值。如果蛋白质都在图的对角线上，表明它们在参考蛋白和输入蛋白中的分值相同(或者几乎相同)。然而，也有值得注意的异常值，代表了两种生物不同表型的重要基因。这些点是可以点击的(图中带圆圈的数据点)。TaxPlot 还能根据 COG 分类系统规则在图上标注颜色。

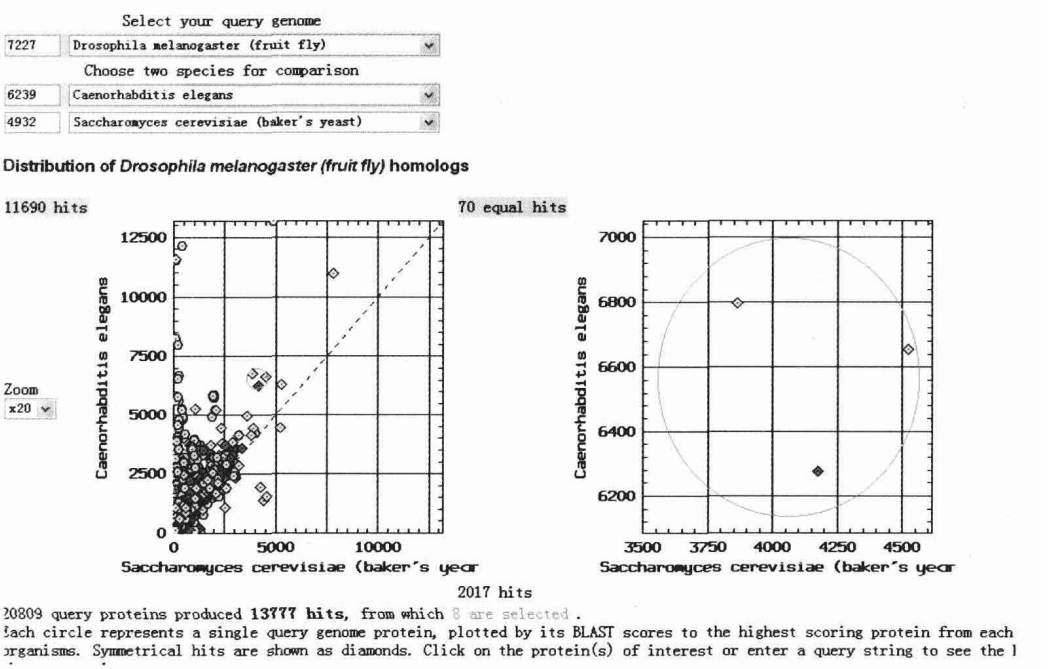


图 5-7 TaxPlot 界面示意图

在整个微生物基因组的比对中最大的挑战就是利用动态程序，比对上百万的碱基对所需要的大量时间。然而对于基因组比对来说，这些工具还比较初级。MUMmer 软件包提供了一个对微生物基因组进行快速准确的比对方法。最近，经过对算法改进后，也能够对真核生物序列进行比对。

MUMmer 将两条序列作为输入。这个算法找到了所有的长于一个设定的最小长度值 k 并且很好匹配的子序列。根据定义，这些匹配序列是最小的，因为如果将它们向任意方向延长一点就会导致不匹配。

MUMmer 的输出结果(图 5-8)由点阵图组成，该结果以最小比对长度 150bp 为序，显示了两个基因组序列的比对结果。结果包括如下内容：SNP；比单个 SNP 更加分散的序列区域；大的插入片段(例如，经过转座、序列逆转和水平基因转移)；散在重复片段(例如一个基因组中的复制)；片段串联重复(拷贝数)。

大肠杆菌 K12 和大肠杆菌 O157：H7(在受污染的食品中有这个菌株，会导致如出血性结肠炎之类的疾病)。在大约 45 亿年前发生分支。测序并比较两个基因组，发现大肠杆菌 O157：H7 大约比大肠杆菌 K12 长了 859 000 个碱基对。这两个细菌有大约 4.1Mb 的共同基因组骨架，大肠杆菌 O157：H7 有另外 1.4Mb 的序列(大部分通过水平基因转移得到)。MUMmer 的输出结果对于找出两

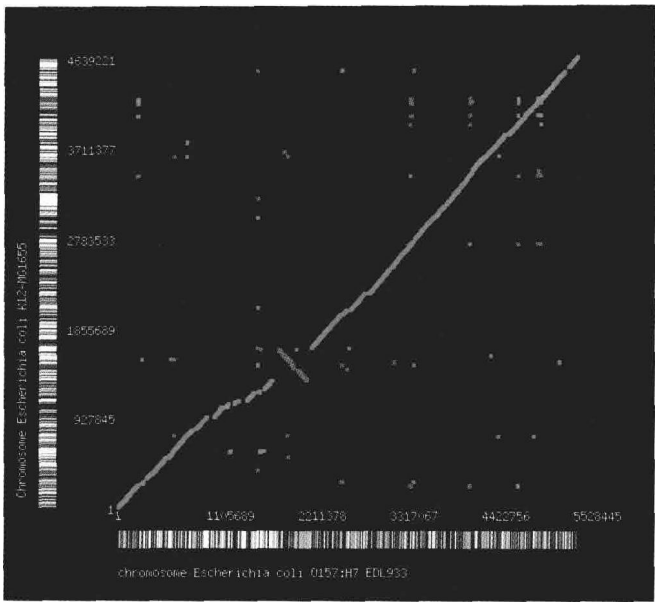


图 5-8 MUMmer 输出结果

个基因组中的共同区域和反向重复区域非常有用。

四、蛋白质互作网络进化

近年来，随着鉴别蛋白质互作关系的高通量实验技术（如酵母双杂交、免疫共沉淀、基于质谱的串联亲和纯化等），以及生物信息学方法在预测蛋白互作领域的发展与应用，越来越多的蛋白质互作数据涌现出来，为进化研究提供了新的视角。

对蛋白质互作网络的进化分析可分为五个层面：蛋白质个体、蛋白质互作对（protein interaction pair）、模体（motif）、网络模块（network module）以及整个网络。即按照包含蛋白质的数目将网络进化问题分层：第一层是仅包含一个蛋白的蛋白质个体；第二层为包含两个蛋白的蛋白互作对；第三层为网络模体一般包含 3~5 个蛋白质；第四层为网络模块，相对于之前的三层包含的蛋白数目更多，且可能由模体组成；第五层则是整个网络的进化分析，探究网络的发生发展过程。

（一）网络中的蛋白质个体进化

蛋白质互作网络对蛋白质个体进化性质的影响，即蛋白质互作是否会减慢蛋白质进化速率，是在蛋白质个体层面上研究网络进化的主要问题。

由于研究者选择的研究对象多数为酵母，尽管所选的互作数据不同，采用的进化速率评估方法、寻找直系同源蛋白的方法及所统计分析方法等不尽相同，但从现有的研究成果可以得出如下结论：蛋白连接度同其进化速率之间可能存在较弱的负相关关系。因为影响蛋白质进化速率的因素很多，除了与网络拓扑性质相关的蛋白连接度（由互作数目定义）、蛋白中心性（由介数定义）外，还有可能与蛋白表达水平、蛋白必要性、蛋白质功能及其参与的生物学过程、蛋白质丰度、密码子适应指数等有关，并且这些因素之间存在错综复杂的依赖关系。

（二）网络中的蛋白互作对进化

互作的两个蛋白质在进化上是否趋向具有相似的性质？在分子水平上是否趋向共进化？这是网络中蛋白互作对进化研究要回答的问题。

多年来，研究者开发了许多预测蛋白质互作的方法，如比较基因组学方法、利用系统发育树相似性进行预测的方法、利用基因表达水平相关性进行预测的方法和同源预测方法等，这些方法多是基于相互作用蛋白共进化的思想。这些预测算法的成功，从另一个角度为互作蛋白具有共进化的现象提供了有力证据。目前学术界普遍认同的观点是：互作的蛋白质倾向于具有更相似的进化速率，且

网络中的蛋白互作对在表达水平等层次上也可能存在微弱的共进化现象。对于这一观点的解释主要有两种，一种假设为，共进化是施加在互作的蛋白对上相似进化压力的结果。相似的进化压力可能来源于作用在这两个互作蛋白对上的相似调控机制，如协同转录和调控等。这种假设不仅适用于解释发生直接物理互作蛋白对间的共进化，对共享一个生物学关系的一组蛋白质的共进化现象也同样适用。另一种假设为，共进化直接与互作蛋白的共适应相关。即当蛋白质序列上直接或者间接通过影响蛋白质折叠而参与互作的位点发生有害突变时，与其互作的蛋白质通过发生互补的改变来维持两个蛋白质的互作关系，进而保持功能。综合两种假设，即两种共进化推动力可能是在不同程度，不同水平和不同情况下发挥各自的作用。

（三）网络中的模体进化

网络模体是指复杂网络中在不同位置重复出现的特定的相互连接模式，在数量上显著地高于随机期望，一般含有 3~5 个节点。对于网络模体进化的研究主要集中在探讨模体是否对其成员蛋白质进化具有约束作用。研究表明，模体成员蛋白质要比非模体成员蛋白质在进化上更具有保守性。在不同拓扑结构模体中，成员蛋白质的保守性不同，可能的原因是不同的模体模式所承受的进化约束显著不同。

（四）网络中的模块进化

蛋白质互作网络(图 5-9)具有层次模块化特性。功能模块的最显著特点是其往往表现出可能在功能和拓扑上互相联系，在蛋白互作网络中主要以蛋白质复合物的形式存在。目前的研究成果表明，网络的模块化对蛋白质进化可能有约束作用，成员蛋白之间在进化速率、表达水平等方面表现出共进化特性。类似蛋白质互作预测领域，许多功能模块预测算法(如比较基因组学方法)都是基于模块成员蛋白共进化的思想，其成功也反过来支持了功能模块成员蛋白的共进化特点。

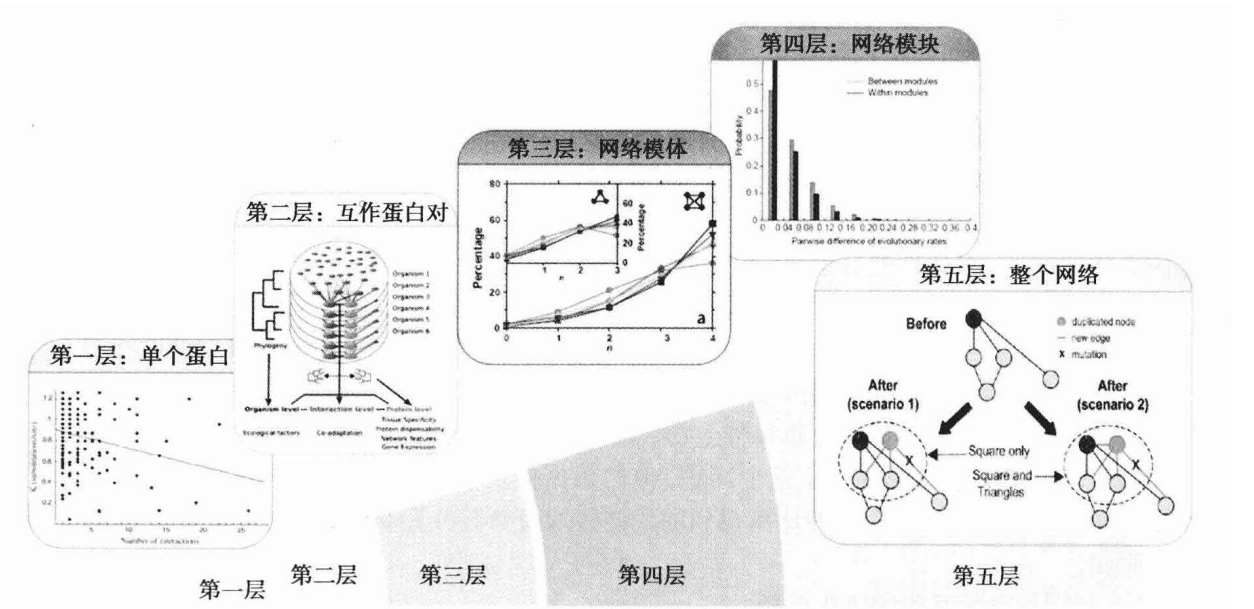


图 5-9 蛋白质互作网络进化图

第一层表示网络中的蛋白质个体进化，表明蛋白连接度同其进化速率之间存在较弱的负相关关系；第二层表示网络中的蛋白互作对进化，揭示出互作的蛋白质倾向于具有更相似的进化速率可能由多种因素导致；第三层表示网络中的模体进化，模体成员蛋白更具有保守性；第四层表示网络中的模块进化，成员蛋白之间在进化速率上表现出共进化特性；第五层表示网络的整体进化中的复制 - 分歧模型

（五）网络的整体进化

研究蛋白质互作网络整体进化的最主要问题是蛋白质互作网络的起源。随之而来的问题是蛋白质互作网络具有的无标度(scale free)分布，小世界(small world)性质和模块化结构等是如何起源和

进化的？这些特性的存在是生物体长期进化过程中自然选择的结果，还是存在内在约束机制使其发生成为不可避免的趋势？

多年来，学者们先后提出了多个无标度和小世界网络的进化模型。目前应用最为广泛的是优先连接模型和复制 - 分歧模型。优先连接模型描述网络的生长是通过不断向网络中添加新的节点来实现的，而新添加的节点倾向于优先与原有网络中度高的节点连接。这一模型揭示的问题是蛋白质年龄与连接度之间存在的强烈而显著的关系，即蛋白质起源越早，其连接度越高。并且当控制表达水平后，这种关系并没有被显著地削弱。在复制 - 分歧模型中，网络中的初始蛋白质被随机选择并复制，且伴随该蛋白质参与所有互作。随后，基因突变导致副本和原蛋白逐渐发生分歧，表现为它们参与的互作发生改变。从生物信息学的角度，则可以理解为基因组层面上的改变在网络拓扑结构变化上的体现。有研究表明，酵母中至少有 40% 的蛋白质互作来源于复制事件。而对于蛋白质复合物的起源和进化研究显示，有相当一部分复合物是通过逐步的部分复制而进化来的，并且被复制的复合物仍然保持原复合物的核心功能，但具有不同的绑定特异性和规则。

五、代谢网络进化分析

各种高通量技术和代谢通路数据库的发展使得分析代谢网络进化（metabolic network evolution）成为可能。一般说，生物网络具有稳健性和进化性的一个主要原因归功于其模块化组织。模块定义为一组连接非常紧密的基因或酶的集合，功能相对独立，而模块与模块之间的连接较为稀疏。从仅有几个基因的简单网络能够利用计算机模拟的手段构建出具有几百个节点上千条边的大网络。另外，有些研究通过比较多个物种的拓扑结构对代谢网络的进化机制进行探讨，发现不同代谢通路的拓扑特征提供不同的系统发育信息。

（一）代谢网络模块性的进化分析

一个生物网络中的模块包含很多元素（例如蛋白质或反应），这个模块形成了一个结构上的子系统，并且有其独特的功能。在代谢网络中，存在很多小的，高连接度的模块，这些模块又分层组合成为大的单元。对于模块的进化，目前主要有两个假设：一是模块倾向于正选择，因为已经限定好的模块能维持细胞的功能，通过模块的进化变化能够提升其可进化性；二是尽管模块不能直接通过选择进化，但模块之间在进化上存在一致性，还能通过其他可以被选择的性质，例如由水平基因转移引起的基因聚类的加速，多效性的最小化，和对新环境的适应性等。

由于生物之间的遗传相关，其代谢网络也存在着一定的相似性，所以系统发育相近的生物代谢网络模块也应该是相近的。伴随模块内变异逐渐增多，物种之间的差异也就越大，相反亦然。如果对不同物种代谢模块统计相应得分，就可以根据这个得分构建生物代谢系统发育树。但对模块的变异量化研究存在一定难度，如何计算每种生物代谢网络的得分是研究关键。

Anat Kreimer 等成功地解决了这个问题，他们根据模块的特性，使用 Newman 的算法计算代谢网络中模块的得分，根据每个物种计算得到的代谢模块分数建立距离矩阵，形成了如图 5-10 所示的系统发育树。

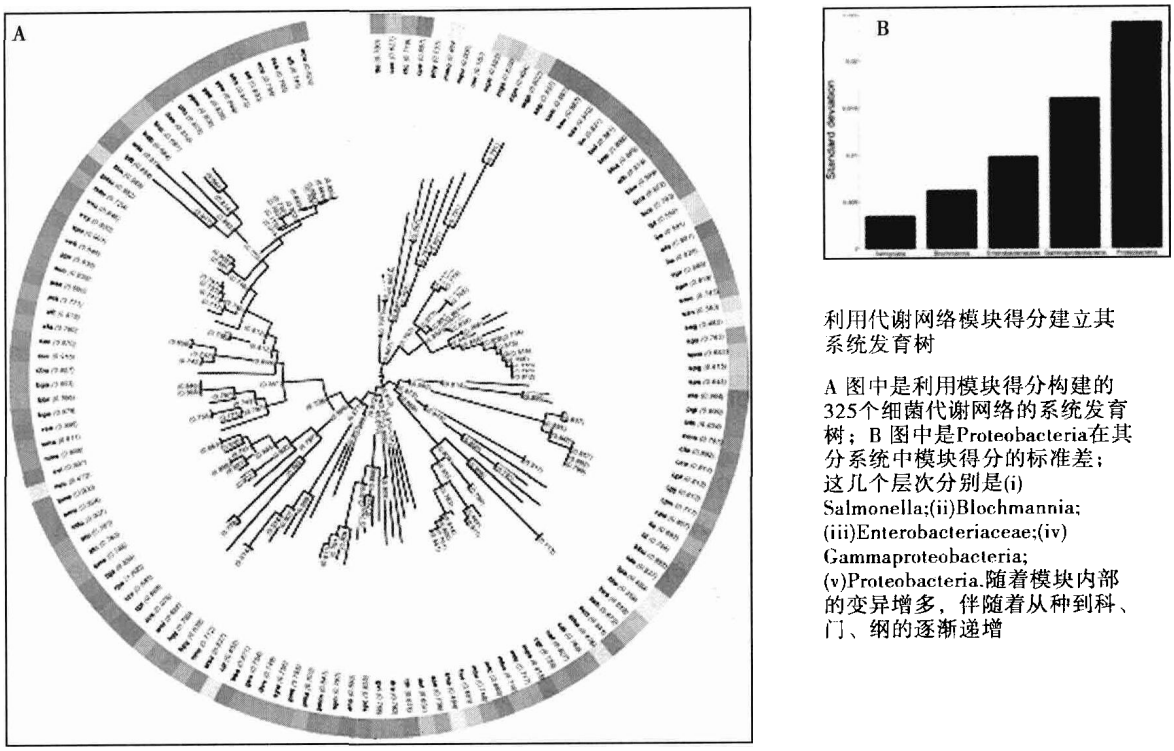
（二）代谢与环境互作的进化分析

代谢网络一般是在一定的生化环境下行使功能，同时通过吸收和分泌各种有机和无机的化合物来与环境发生互作（图 5-11）。例如，在网络内部新陈代谢流动性的分布或生命体的增长率都是通过这种作用来完成。

和环境的这种相互作用在一定程度能够在代谢网络的结构进化上反映，所以这些代谢网络不应只是单单推断代谢功能，还应当能够观察到物种和环境互作进化的现象。在分析代谢网络的拓扑结构时，有一类化合物是通过外源获得，这类化合物定义为“种子集合”。如果一个物种的环境能够决定其代谢反应，那么这些“种子集合”就是代谢网络与外界环境之间一个很好的代理。

每种生物的代谢网络种子集合是不同的，根据集合中的基因在这种生物是否存在可以构造进化

的距离矩阵。因为在进化过程会有新的化合物以种子或者非种子的身份加入到代谢网络中，如果是以种子的身份被整合到代谢网络中，这个种子存在的状态可能不会太长，要么从代谢网络中被拿掉，要么快速的变为非种子化合物。



利用代谢网络模块得分建立其系统发育树

A 图中是利用模块得分构建的325个细菌代谢网络的系统发育树；B 图中是Proteobacteria在其分系统中模块得分的标准差；这几个层次分别是(i) Salmonella;(ii)Blochmannia;(iii)Enterobacteriaceae;(iv) Gammaproteobacteria;(v)Proteobacteria.随着模块内部的变异增多，伴随着从种到科、门、纲的逐渐递增

图 5-10 利用代谢网络模块得分建立其系统发育树

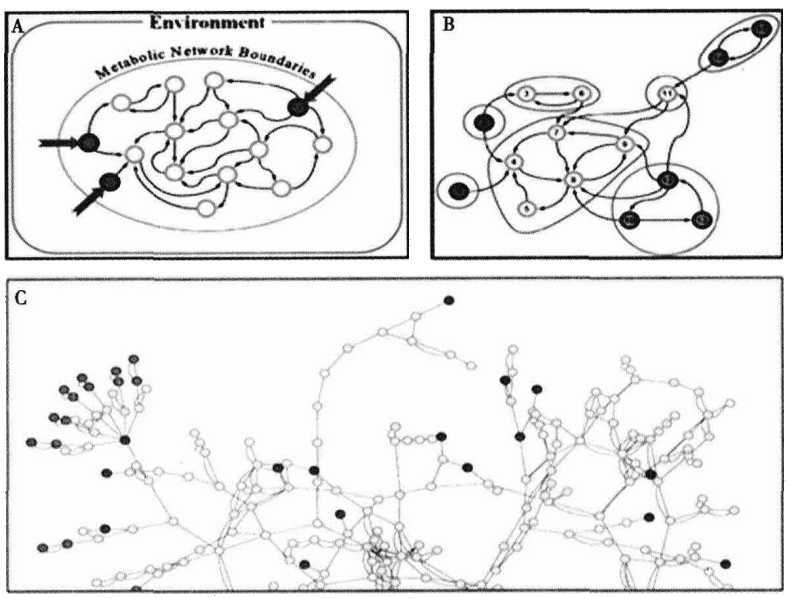


图 5-11 代谢与环境互作的进化分析示意图

在代谢网络中鉴定种子复合 A 代谢网络与环境相互作用的示意图，种子是用红色标记 B 代谢网络中种子获得过程 网络首先用 kosaraju 的强连通组分(SCC)的方法分解，子网中的源组分就是要找的种子。图中的源组分是用红色表示的，节点颜色的饱和程度代表种子的置信程度 C Buchnera 代谢网络图，红色为种子复合物

小 结

近年来，由于序列数据的爆炸性增长，分子进化领域得到了快速发展。基因组的大规模数据也需要更有效的统计方法去分析和解释，这无论是在概念上还是计算上都非常具有挑战性。本节讨论了经典的分子进化统计方法，也涉及最新前沿进展。基因表达的进化、蛋白质互作网络的进化、共进化等一系列新的概念都成为这一结合领域研究热点。分子进化分析已经从对单一基因、蛋白质的进化分析扩展到蛋白质网络的进化和表达的进化。蛋白质的进化率不但和其必要性有明显的相关性，从蛋白质互作网络来看，网络中的度也和进化率存在相关性。比较基因组学也给网络的动态性提供了新的数据，这对于理解分子进化的数量进化给予了新的方向。当然这还需要很多关于网络动态性和结构的新理论。

Summary

The field of molecular evolution has experienced explosive growth in recent years due to the rapid accumulation of genetic sequence data, continuous improvements to computer hardware, and the development of sophisticated analytical methods. The increasing availability of large genomic data sets requires powerful statistical methods to analyze and interpret them, generating both computational and conceptual challenges for the field. At the same time, a new field, combination of bioinformatics and molecular evolution, quickly emerges with the vigorous development of bioinformatics. Evolution sometimes requires the cooperative action of several genes, and conversely, a single gene evolution may participate in different functional contexts. Networks evolution research has become a hot area, which include signal transduction, protein interaction, and metabolism etc. Most current work is devoted to addressing the coevolution researches. Comparative genomics provides new data on the dynamics of genetic networks, opening a promising research direction towards a quantitative understanding of molecular evolution. Of course, this is an area in need of new theoretical concepts linking structure and dynamics of these networks.

（李 霞 张绍军 李亦学）

习 题

1. 核苷酸置换模型主要有哪几种类型，各自有什么特点？
2. 氨基酸替代中为什么需要泊松校正模型？
3. 构建基因的系统发育树有哪几种常见的方法？
4. 如何利用序列数据判断基因存在非中性进化？
5. 为什么病毒在进化研究中有很重要的作用？
6. 可以哪些方面来研究蛋白质互作网络的进化？

主要参考文献

1. Borenstein E., Kupiec M., Feldman M. W. et al., Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proc Natl .Acad. Sci. U S A., 2008, 105(38): 14482-14487.

2. Chen K., Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet, 2007, 8(2): 93-103.
3. Cristianini N., Introduction to Computational Genomics A Case Studies Approach. Cambridge: Cambridge university press; 2006.
4. Fraser H. B., Hirsh A. E., Steinmetz L. M., et al., Evolutionary rate in the protein interaction network. Science, 2002, 296(5568): 750-752.
5. Juan D., Pazos F., and Valencia A. Co-evolution and co-adaptation in protein networks. FEBS Lett, 2008, 582(8): 1225-1230.
6. Kreimer A., Borenstein E., Gophna U., et al. The evolution of modularity in bacterial metabolic networks. Proc. Natl. Acad. Sci. U S A., 2008, 105(19): 6976-6981.
7. Nei M. Molecular Evolution and Phyolgenetics. London :Oxford university press; 2000.
8. Pevsner J. Bioinformatics and Functional Genomics. 2nd Edition. Wiley-Blackwell. 2009.
9. Wuchty S., Oltvai Z. N., and Barabasi A. L. Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet, 2003, 35(2): 176-179.
10. Yang Z. Computational molecular evolution Oxford Series in Ecology and Evolution. London: Oxford university press; 2006.

第六章 表达序列分析
CHAPTER 6 ANALYSIS OF EXPRESSED SEQUENCES

第一节 引 言
Section 1 Introduction

表达序列(expressed sequence)是指由基因组表达为RNA的序列,其中绝大部分是mRNA分子。它们可以进一步翻译为蛋白质序列;少部分表达为构成核糖体的rRNA或负责转运氨基酸的tRNA,rRNA和tRNA即为基因表达的终产物,不再翻译为蛋白质。

表达序列标签(expressed sequence tag,EST)是通过从cDNA文库中随机挑选克隆,进行一轮单向测序所获得的序列,通常为几十至500bp左右,它们大多不是完整的基因序列,但携带了表达基因的部分遗传序列。虽然EST不能与基因划等号,但由于EST来自基因的代表性标签序列,更由于EST的获得十分快速、简便且廉价,故EST已成为广大科学工作者研究基因表达最有用的材料,其价值得到了越来越多科学家的认可,已成为基因组学研究领域的主要内容。1993年NCBI网站专门设立了EST数据库dbEST(database of EST),系统地收集和保存来自全世界研究者们提交的EST序列。在dbEST中记录了每一个EST的登记号、测序引物、碱基序列、cDNA文库构建方法、组织来源等详细信息,是非常有用的表达序列数据库。截至2010年1月22日,dbEST数据库已收录了64 679 625条EST数据,其中大部分来自于人和小鼠。

EST具有广泛的用途,主要用于以下诸方面:

1. 基因组物理图谱的绘制 物理图谱是以特异的DNA序列为标记的基因组图谱,标记之间的距离以物理距离如碱基对(bp、kb、Mb)表示。用于作图的序列也称为"序列标签位点"(sequence-tagged site,STS),故物理图谱也称为STS图谱。EST是用于绘制物理图谱最常用的序列标签位点。1995年WhiteHead研究所的Hudson发表的人类基因组物理图谱即含有15 086个STS,其中大多数是EST,平均密度为1个标记/199kb。第二年在这份图谱上又添加了大量STS,从而将许多蛋白质编码基因定位在物理图谱上。这份图谱的密度为1个标记/100kb,达到了人类基因组计划最初确立的目标。

2. 基因识别 当一个物种的全基因组测序完成之后,首要的任务就是要对基因组中所包含的全部基因进行预测。迄今为止基因预测软件不可能百分之百地准确预测出全部基因,因此对预测基因的验证就至关重要。EST是来源于基因组中转录出来的mRNA,每一条独特的EST代表的是特定发育阶段或某种病理、生理状态下表达出来的基因的部分序列,因此,将预测基因与同物种的所有EST进行比对,有助于基因识别的验证。以某物种的EST为训练集,可以提高基因预测算法的灵敏度和准确度。

3. 基因表达谱的构建 基因表达谱是反映生物个体在特定组织、特定器官、某一特定发育时期、某种病理生理状态或某种治疗状态下,所有基因表达水平的图谱。基因表达谱(gene expression profile)可以用来比较不同物种、不同组织、不同器官、不同发育阶段或不同病理生理状态下基因表达水平的差异,鉴定出与某一发育阶段、某一代谢途径(或过程)或病理生理状态相关的表达基因。基

因表达谱的绘制在功能基因组学研究、基因诊断、药物治疗、药靶基因寻找等方面有着重要意义。

4. 发现新基因 由于EST是一段表达序列标签，必然来自某一基因。将获得的某个EST序列在数据库中搜索，如果没有注释为已知基因，那么该序列就很可能来自一个新基因；如果在数据库中发现了某个基因的相似性序列，但又不能定义为同一基因的话，那么，这个EST很可能是某个基因家族的新成员。

5. 电子PCR克隆 EST来自cDNA文库的测序。但在cDNA文库构建过程中，许多基因都缺少5'端的序列信息，因此大多数情况下，一个EST仅仅是某个基因的部分序列。在基因鉴定及其功能研究中，获得全长cDNA克隆是前提条件。在EST数据库中，存在着大量同一基因的EST冗余序列，通过聚类拼接，即可获得较长、甚至全长基因的cDNA序列。这一拼接过程是通过计算机完成的，故又称为电子PCR克隆(e-PCR clone)。

6. SNP发现 单核苷酸多态性(single nucleotide polymorphism，SNP)是基因某一核苷酸位点上的变异导致的遗传多态性。SNP可能与基因功能、基因失活、基因的表达调控、疾病等遗传学特性相关。SNP的鉴定及其与某些生物学特性或疾病的相关性研究已成为热点研究领域。数据库中的EST来自全世界各实验室，通过对冗余EST进行序列组装联配，有可能发现基因组中存在的SNP。但值得注意的是，EST数据来自单轮测序，序列中可能存在错误。因此，从EST数据库发现的SNP还需要进行实验验证，排除假阳性结果。

本章主要介绍cDNA文库构建与EST数据的实验获取、EST数据库dbEST、数据库UniGene、EST数据分析方法、基因表达系列分析(SAGE)的原理、方法及其应用。

第二节 EST数据分析

Section 2 Analysis of EST data

一、cDNA文库构建与EST数据的实验获取

知识拓展

由于基因表达的差异变化是调控细胞生命活动过程的核心机制，通过比较同一类细胞在不同生理条件下或在不同生长发育阶段的基因表达差异，可以为分析生命活动过程提供重要信息。研究基因差异表达的主要技术有差别杂交(differential hybridization)、扣除(消减)杂交(subtractive hybridization of cDNA，SHD)、mRNA差异显示(mRNA differential display，DD)、抑制消减杂交法(suppression subtractive hybridization，SSH)、代表性差异分析(represential difference analysis，RDA)、交互扣除RNA差别显示技术(reciprocal subtraction differential RNA display)、基因表达系列分析(serial analysis of gene expression，SAGE)、电子消减(electronic subtraction)和DNA微列阵分析(DNA microarray)等。

从根本上讲，生物信息学分析是建立在实验数据基础上的，没有核苷酸序列及蛋白质氨基酸序列，就不可能有生物信息学。因此，本节在讨论EST数据分析之前，有必要简要叙述EST数据的实验获取。前面提到，EST数据是通过构建cDNA文库，从文库随机挑取克隆，通过单轮测序获得的。下面介绍cDNA文库构建和获得EST数据的方法。

(一) cDNA文库有非标准化和标准化文库之分

非标准化cDNA文库(unnormalized cDNA library)是指对用于建库的mRNA未进行任何预处理而直接用于构建cDNA文库。它反映了组织中所有基因的表达水平，适用于基因表达谱研究。但由

于文库可能存在大量高丰度表达基因，在随机挑取克隆时，可能多次挑到相同的克隆，导致 EST 数据冗余度较高，因此测序的成本较高。

标准化 cDNA 文库(normalized cDNA library)是指通过杂交的方法如扣除杂交(subtractive hybridization)或抑制性扣除杂交(suppression subtractive hybridization)去除一些中、高表达丰度的 mRNA 拷贝数后再形成的文库。这样，中、高表达丰度基因表达产物 mRNA 冗余度较少，会相应地对低丰度基因表达产物起富集作用，有助于检测到低丰度表达基因，但不能用于基因表达谱研究。

(二) 获得总 mRNA 后，需要将它们反转录为 cDNA，可以用两类不同引物达到此目的

1. Oligo d(T)引物　由于真核生物 mRNA 的 3' 端总是有一段长度不等的 polyA，故可使用 Oligo d(T)作为引物反转录获得 cDNA，用于构建 cDNA 文库。使用 Oligo d(T)作为反转录的引物的好处是可以减少基因组或其他种类的 RNA(如 rRNA 及 tRNA)对文库的污染概率，但缺点是从这类文库中难以获得较大基因的全长 cDNA。而且，由于长度不等的 polyA 的存在，会给 EST 的 3' 端测序的成功率带来不利影响。

2. 随机引物　利用随机引物可以在基因组的多个不同位点同时启动第一链的反转录，从而获得 cDNA。利用这种引物建库的好处是可以获得靠近 5' 端的 EST。靠近 5' 端的 EST 具有更多的信息含量，因而在基因功能鉴定和构建 EST 数据库时有利于基因聚类和阅读框的寻找，有利于寻找同源基因。原核生物的 mRNA 的 3' 端缺乏 polyA，故只能采用随机引物来建立 cDNA 文库。随机引物 cDNA 文库适用于发现和注释新基因及基因表达研究。它的缺点是不能避免其他种类 RNA(rRNA 及 tRNA)对文库的污染，同时，这种引物反转录产生的 cDNA 的 3' 端位置是不固定的，导致 EST 聚类时假阳性比例增加。

cDNA 文库建好后，随机从文库中挑取克隆，抽提质粒，只需对每个克隆进行一轮测序，即可获得长短不同的 EST 数据。这些 EST 数据可提交到 EST 数据专用数据库(如 dbEST)，供全球的研究者使用。

二、EST 数据库

科学实验获得的大量 EST 需要以数据库的形式进行存储和管理，以方便用户浏览和检索，从中获取所需要的 EST 数据。常用的 EST 数据库见表 6-1。

表 6-1　常用的 EST 数据库

数据库名称	网址	说明
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/	综合
UniGene	http://www.ncbi.nlm.nih.gov/unigene	综合
Gene Indices	http://compbio.dfci.harvard.edu/tgi/	综合
REDB	http://redb.ncpgr.cn/index.php	水稻
Mendel-ESTS	http://ukcrop.net/perl/ace/search/Mendel-ESTS	植物
MAGEST	http://www.genome.jp/magest/	海鞘类
ChickEST	http://www.chickest.udel.edu/	鸡
COGEME	http://cogeme.ex.ac.uk/	真菌
PEDE	http://pede.dna.affrc.go.jp/	猪
CR-EST	http://pgrc.ipk-gatersleben.de/cr-est/index.php	农作物
NEMBASE	http://www.nematodes.org/nematodeESTs/nembase.html	寄生虫
OSESTDB	http://vmd.vbi.vt.edu/EST/	霉菌

其中，dbEST、UniGene 和 Gene Indices 是三个数据量最丰富、最常用的 EST 数据库，因此下面专门进行介绍。

(一) dbEST 数据库

dbEST(http://www.ncbi.nlm.nih.gov/dbEST/)是 GenBank 数据库的一部分,其中收集了大量物种的一轮单向测序的 cDNA 序列或 EST 序列数据以及其他相关信息,是目前最大的一个公共表达序列数据库。dbEST 由 NCBI 在 1992 年建立,截至 2010 年 1 月 22 日,dbEST(版本号 012210)中共有数据条目 64 679 625 条,其中数据量居前 10 位的物种如表 6-2 所示。由表 6-2 可以看出,dbEST 中来自于人类和一些重要的模式生物如小鼠的 EST 数据占据了很大的比重,这反映人们对于这些物种的表达序列研究得更为广泛和深入。

表 6-2 dbEST 中数据来源前 10 位的物种及其对应的数据量(版本: 012210)

物种名称	dbEST 中的数据条目
Homo sapiens(human)	8 300 249
Mus musculus+domesticus(mouse)	4 852 146
Zea mays(maize)	2 018 798
Bos taurus(cattle)	1 558 493
Sus scrofa(pig)	1 538 636
Arabidopsis thaliana(thale cress)	1 527 299
Danio rerio(zebrafish)	1 481 930
Glycine max(soybean)	1 422 982
Xenopus(Silurana) tropicalis(western clawed frog)	1 271 375
Oryza sativa(rice)	1 249 110

1. 向 dbEST 提交数据 很多 EST 研究项目都会获得大量的 EST 序列,这些序列通常被成百上千条打包并分批次提交到 GenBank 或者 dbEST。这些数据的引用、提交者和文库信息等往往存在大量的冗余。为了改进向数据库中提交这类数据的效率,NCBI 设计了专门的精简提交程序和数据提交格式。目前,向 dbEST 或 GenBank 中提交 EST 数据的方式有两种: 使用 E-mail 或使用序列提交软件。

(1) 通过 E-mail 提交 EST 序列: 对于测序得到的序列数据和图谱数据,可以通过 E-mail 的形式提交到 dbEST 数据库(从 2009 年开始,图谱数据不再向 dbEST 中提交)。准备提交的 EST 数据文件需要以一定的格式进行制作,并且必须是纯文本格式,不然就可能无法提交成功。EST 序列数据包含了四个文件: Publication 文件、Library 文件、Contact 文件和 EST 文件。每一个文件由多项标签组成,每一项标签名大写,后面一个冒号,冒号后面是数据。数据区域和对应的标签应该在同一行,不能自动换行,可以例外的是: Publication 文件中的 "TITLE" 项和 "AUTHORS" 项; Library 文件的 "Description(DESCR)" 项; EST 文件的 "CIATTION"、"COMMENT" 和 "SEQUENCE" 项。上述例外的项中,数据可以从对应的标签行或者下一行开始,允许自动换行。每一个文件中某些项是必需的。例如,任意一个数据条目中,"TYPE" 项都是必需的,甚至在一个文件中一个给定的 "TYPE" 下可以有多个条目。如果一个非必须项的数据不可用,那么这一项可以忽略,或者该项的标签后面的数据区域保持为空。不能使用 '*'、'-' 等字符表示空数据。文件中每一条记录的最后应该以双竖线作为结束标识符 '||'。

1) 文献(Publication)文件: 文献文件包括有效标签和一个短的描述符 '||',通常格式及相关说明如下:

TYPE: Pub(表示该文件为文献文件)
TITLE: (文献标题)
AUTHORS: (作者)
JOURNAL: (文献发表期刊)
VOLUME: (卷号)

ISSUE: (期号)
PAGES: (页码)
YEAR: (出版年份)
STATUS: (发表状态)
|| (结束标识符)
例:
TYPE: Pub
MEDUID: 92347897
TITLE: Expressed sequence tags and chromosomal localization of cDNA clones from a subtracted retinal pigment epithelium library
AUTHORS: Gieser, L.; Swaroop, A.
JOURNAL: Genomics
VOLUME: 13
ISSUE: 2
PAGES: 873-6
YEAR: 1992
STATUS: 4
||

其中,"TYPE"项在每一个数据条目的开头都是不可缺少的,即使一个文件中有多个同一类型的条目。该项为一个不限格式的字符串,但是要求与EST文件的"CITATION"项内容一致。"MEDUID"为"MEDLINE"记录标识符,通常不要求提供这一项。"STATUS"项用不同的值表示不同状态的文献:1-未出版,2-投稿中,3-待出版,4-已出版。

2)文库(Library)文件:该文件通常要求的格式及相关说明如下:

TYPE: Lib(表示该文件为文库文件)
NAME: (文库名称)
ORGANISM: (来源物种)
STRAIN: (菌株)
CULTIVAR: (培养种或栽培种)
SEX: (性别)
ORGAN: (器官)
TISSUE: (组织)
CELL_TYPE: (细胞类型)
CELL_LINE: (细胞系)
STAGE: (细胞所处发育阶段)
HOST: (宿主)
VECTOR: (载体)
V_TYPE: (载体类型)
RE_1: (载体位点1的限制性内切酶)
RE_2: (载体位点2的限制性内切酶)
DESCR: (对文库准备方法、载体等其他需要说明的相关信息)
|| (结束标识符)
例:
TYPE: Lib

NAME: Rat embryonic day 17 post-fertilization Library
ORGANISM: Rattus norvegicus
STRAIN: Sprague-Dawley
SEX: male
STAGE: embryonic day 17 post-fertilization
TISSUE: aorta
CELL_TYPE: vascular smooth muscle
DESCR:
||

其中，“TYPE”项是必需的，位于每一条目的开头。“NAME”项最好小于等于48个字符，最大不超过128个字符，超过48个字符时，创建用于BLAST搜索的fasta格式的文件标志行将被自动转换成48个字符。“NAME”项必须与EST文件的“LIBRARY”项内容一致。“ORGANISM”项用来描述序列的来源物种，因此是最重要的一项。文库的“NAME”要求与单个的EST的物种信息连接起来。“DESCR”项应该包含足够的关于mRNA/cDNA的出处的详情。

3）联系人(contact)文件：该文件通常要求的格式及相关说明如下：

TYPE: Cont(表示该文件为联系人文件)
NAME: (提交者姓名)
FAX: (传真)
TEL: (电话)
EMAIL: (电子邮件)
LAB: (提供EST的实验室)
INST: (研究机构)
ADDR: (联系地址)
|| (结束标识符)

例：

TYPE: Cont
NAME: Sikela JM
FAX: 303 270 7097
TEL: 303 270
EMAIL: tjs@tally.hsc.colorado.edu
LAB: Department of Pharmacology
INST: University of Colorado Health Sciences Center
ADDR: Box C236, 4200 E. 9th Ave., Denver, CO 80262-0236, USA
||

其中，“TYPE”项是必需的，位于每一个条目的开头。在一个联系人文件中，一个给定的“TYPE”项下可以有多个条目。除“TYPE”外，其他项都不是必需的，但通常要求有联系人的姓名(“NAME”项)。一般尽可能填写更多的项，以便提供给用户关于该EST来源的联系信息和其他信息。EST文件中联系人姓名项(“CONT_NAME”)必须与联系人文件中“NAME”项的内容一致，提交程序将会自动匹配二者的信息。该文件内容在提交后不允许修改，因此这些信息应该准确无误。

4）EST文件：EST文件通常要求的格式及相关说明如下：

TYPE: EST(表示该文件为EST文件)
STATUS: (状态)
CONT_NAME: (联系人姓名)

CITATION: （引用文献标题）
LIBRARY: （联系人的实验室名称）
EST#: （EST 号，由联系人的实验室提供）
CLONE: （克隆号）
SOURCE: （克隆的来源，如 ATCC）
SOURCE_DNA: （以纯 DNA 作为克隆的来源 id 号）
SOURCE_INHOST: （贮存在宿主中的克隆来源 id 号）
PCR_F: （正向 PCR 引物）
PCR_B: （反向 PCR 引物）
INSERT: （插入碱基长度）
ERROR: （对插入碱基长度错误的估计）
PLATE: （平板号码或代码）
ROW: （行数或字母）
COLUMN: （列数或字母）
SEQ_PRIMER: （测序引物描述或序列）
P_END: （从哪一端测序，如 5'）
HIQUAL_START: （高质量测序起始位置）
HIQUAL_STOP: （高质量测序结束为止）
DNA_TYPE: （DNA 类型，默认为 cDNA）
PUBLIC: （发表日期）
PUT_ID: （提交者给序列假定分配的鉴别号）
POLYA: （序列是否具有 polyA 尾巴结构）
COMMENT: （对该 EST 的评论）
SEQUENCE: （序列）
|| （结束标识符）

例：

TYPE: EST
STATUS: New
CONT_NAME: Kerlavage AR
EST# : HHC189f
CLONE: HHC189
SOURCE: ATCC
SOURCE_INHOST: 65128
OTHER_EST: HHC189r
CITATION:
Complementary DNA sequencing: expressed sequence tags and human genome project
SEQ_PRIMER: M13 Forward
P_END: 5'
HIQUAL_START: 1
HIQUAL_STOP: 285
DNA_TYPE: cDNA
LIBRARY: Hippocampus, Stratagene (cat. #936205)

PUBLIC:
PUT_ID: Actin, gamma, skeletal
COMMENT:
This is a comment about the sequence.
It may span several lines.
SEQUENCE:
AATCAGCCTGCAAGCAAAAGATAGGAATATTCACCTACAGTGGGCAC
CTCCTTAAGAAGCTGATAGCTTGTTACACAGTAATTAGATTGAAGATA
ATGGACACGAAACATATTCCGGGATTAAACATTCTTGTCAAGAAAGG
GGGAGAGAAGTCTGTTGTGCAAGTTTCAAAGAAAAAGGGTACCAGCA
AAAGTGATAATGATTTGAGGATTTCTGTCTCTAATTGGAGGATGATTC
TCATGTAAGGTTGTTAGGAAATGGCAAAGTATTGATGATTGTGTGCTA
TGTGATTGGTGCTAGATACTTTAACTGAGTATACGAGTGAAATACTTG
AGACTCGTGTCACTT
||

其中，“TYPE”项是必需的，位于每一条目的开头，在一个 EST 文件中，一个给定的“TYPE”下面可以有多个条目。“STATUS”项的有效值是‘New’(对于新条目)或者‘Update’(对已存在的 EST 条目的更新)。当更新一个 EST 文件时，仅仅具有描述文本的项会被改变。更新文件中的文本将完全取代当前版本中对应项的旧文本。“DNA_TYPE”项对于 cDNA 将省略，除非 DNA 类型与此不同。序列将从“SEQUENCE”标签这一行或者下一行开始。为了实现自动匹配，下列项后的字符串必须完全一致：EST 文件的“CONT_NAME”项与 Contact 文件的“NAME”项；EST 文件的“LIBRARY”项与 Library 文件的“NAME”项；EST 文件的“CITATION”项与 Publication 文件的“TITLE”项。这些项将被自动进行匹配，因此它们的内容、拼写、大小写、空格等必须完全一致。1999 年 4 月后增加了 4 个新项：“POLYA”，“TAG_LIB”，“TAG_TISSUE”，“TAG_SEQ”。其中，“POLYA”项代表 EST 序列后是否有一个 polyA 尾巴，‘Y’表示有，‘N’表示没有。

如果所有的 EST 共享相同的 Publication、Library 和 Contact 信息，那么这三个文件仅仅需要各准备一份就够了，然后对于每一个序列完成一个单独的 EST 文件。如果每个 EST 文件都具有不同的 Publication、Library 和 Contact 信息，那么必须对于每个 EST 文件都完成一个新的 Publication、Library 和 Contact 文件。一旦特定的 Publication、Library 和 Contact 信息被输入数据库，则不需要再重新发送数据文件。

编辑好各个文件以后，就可以通过电子邮件提交序列文件到 dbEST 数据库。将完整的文件发送到邮箱 batch-sub@ncbi.nlm.nih.gov，既可以在一个邮件内批量附上所有的文件，还可以将其写在一个邮件文件内，但必须保证以纯文本格式发送。此外，可以将 Publication、Library 和 Contact 数据放在一个文件内提交，甚至可以和 EST 文件作为同一个文件发送，服务器分析程序将通过“TYPE”项来进行区分。

提交数据后，提交者将从 dbEST 管理员邮箱收到给 EST 序列分配的 dbEST ID 号以及 GenBank Accession 号码。如果你希望在文献出版前不公开你的序列，你可以在 EST 文件的“PUBLIC”项中指定数据发布的日期。你的序列将在指定的日期公布，或者在 Accession 号码或序列数据被发表的时候公布，任意一种情况都可优先。一旦你的序列在公共数据库中发布，它们将可以在 GenBank 中检索到，利用 EntreZ 检索系统可以方便地进行检索。

(2) 通过软件提交 EST 序列：如果 EST 序列经过了拼接和注释，也就是说序列的生物学特性已经明确，那就不能提交到 dbEST，而应该提交到相应的核酸数据库。提交到 NCBI 核酸数据库的常用软件为 BankIt 和 Sequin。前者以 web 的形式提交，后者通过 Sequin 这个软件把序列整理为要求

的格式以后发送 E-mail 进行提交。对于序列数量较少，注释较简单的序列采用 BankIt 更方便、迅速(http://www.ncbi.nlm.nih.gov/BankIt/)。而对于序列数量较多、注释较为复杂的序列提交任务，可以选择使用 Sequin(http://www.ncbi.nlm.nih.gov/Sequin/)。

若想了解关于提交 EST 拼接序列的更多信息，可以通过电子邮件联系(gb-admin@ncbi. nlm.nih.gov)。从 2009 年开始，来自新一代测序平台(罗氏 454、Illumina、Applied Biosystems SOLiD、Helicos Biosciences HeliScope 等)的序列，应该提交到 Short Read Archive(SRA)(访问网址为 http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi，了解相关信息可以联系 sra@ncbi.nlm.nih.gov)。不能通过 EST 提交程序提交的序列类型包括：线粒体序列、rRNA、病毒序列、载体序列。载体以及连接区域序列应该在提交前从 EST 序列上移除。

(3) 更新 dbEST 数据：更新 EST 序列的过程和提交新序列基本一致。将原序列文件需要更新的项更新，然后将“STATUS”项由‘New’改为‘Update’，然后提交。除了需要更改的项，更新文件还需要包括“TYPE”、“STATUS”、“EST#”和“CONT_NAME”项。若对 Publication、Contact、源数据、“EST#”或者“CONT_NAME”进行更改，则需要发送一个邮件对这些更改进行说明(batch-sub@ncbi.nlm.nih.gov)。

2. 利用 dbEST 数据库获取 EST 表达数据　EST 序列收录在 GenBank 数据库的 EST 子库中，可以匿名登录 NCBI FTP 获取，也可以使用 EntreZ 检索系统来检索。此外，dbEST 库中的序列还可以使用 BLAST 电子邮件服务来进行搜索。例如，使用 tblastn 程序可以查询一个氨基酸序列，该程序将 dbEST DNA 序列通过六种读框方式翻译为氨基酸序列，并与查询序列进行比较。

EST 序列的 fasta 格式的 flat 文件数据可以通过 FTP 下载，这些数据存放在 ftp.ncbi.nih.gov/repository/dbEST 目录下。

使用 EntreZ 来检索 EST 数据的步骤如下，首先进入 EntreZ 检索页面：http://www.ncbi.nlm.nih.gov/sites/gquery，在检索主页面的搜索输入栏内输入你想检索的 EST 数据的关键词信息，然后点击“go”按钮，开始检索。检索结果将在数秒到数十秒内出现。例如，如果要检索人类血红蛋白 β 亚基的相关 EST 数据，可以在输入栏内输入组合关键词“HBB Human”，然后点击“go”开始检索，检索结果的页面如图 6-1 所示。可以看到，在各子数据库的标签前面都出现了一个数字，这一数字表示

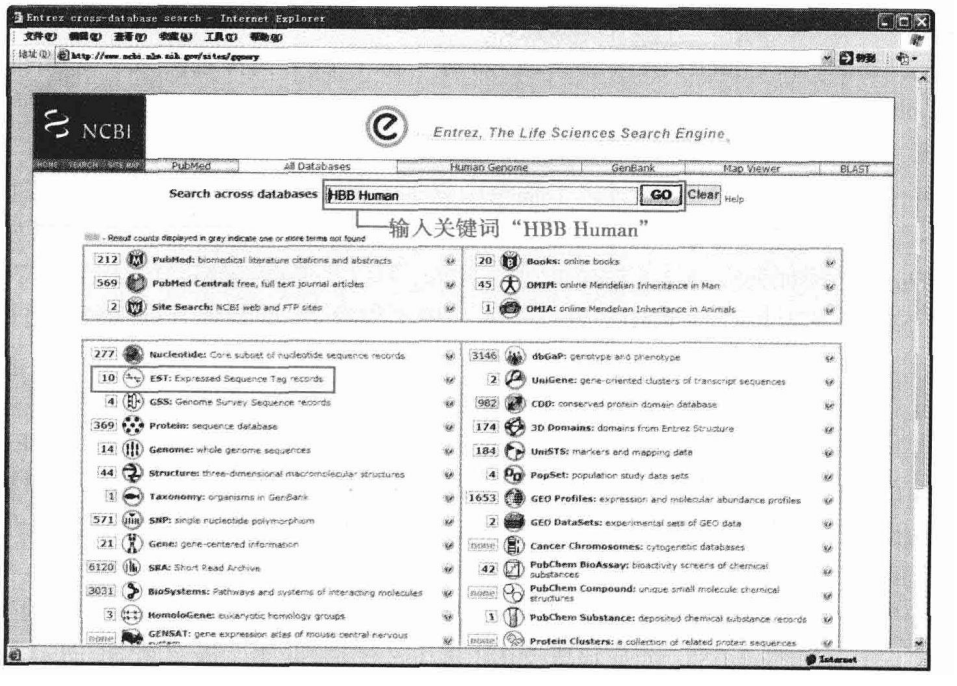


图 6-1　在 Entrez 中以关键词“HBB Human”检索人类血红蛋白 EST 数据

在该数据库中检索到的数据量。“EST”数据库标签前面的数字为10，表示在dbEST数据库中检索到了10条记录。此外，也可以直接在dbEST数据库的主页面的搜索框内输入“HBB Human”进行搜索(图6-2)。搜索结果如图6-3所示。

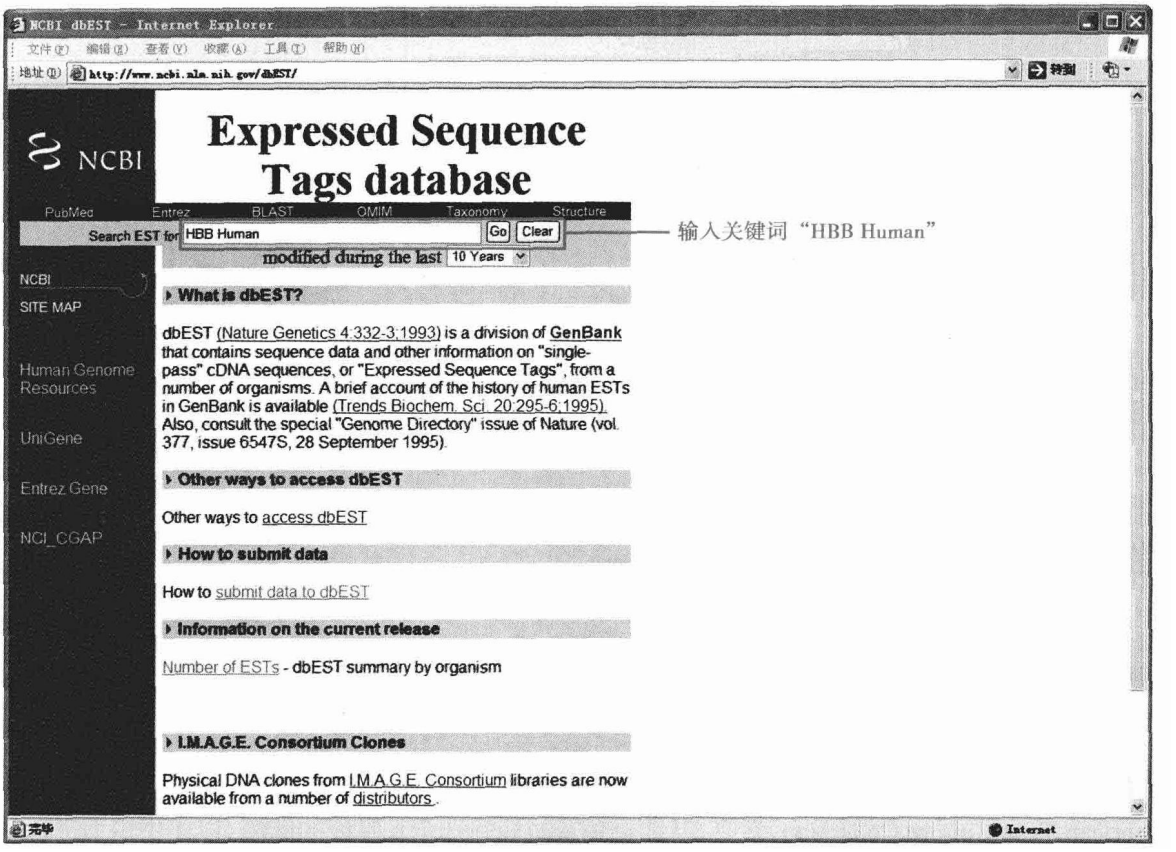


图6-2 在dbEST中以关键词“HBB Human”检索人类血红蛋白EST数据

每条记录前面用数字表示该记录的序号，序号前面的复选框可以用来选择多个记录。以第一条数据记录“DN991377”为例，“DN991377”表示该EST数据的GenBank访问号(Accession Number)；该条记录第二行的文本“TC119811 Human adult whole brain, large insert, pCMV expression library Homo sapiens cDNA clone TC119811 5- similar to Homo sapiens hemoglobin, beta (HBB), mRNA sequence”，是对该数据的简要描述；该条记录的最后一行为“gi|66251208|gb|DN991377.1|[66251208]”，其中“gi|66251208”表示该数据的gi号为66251208，“gb”表示该数据来源于GenBank，“DN991377.1”表示该EST数据的版本号。

在搜索结果页面上，可以点击每一条记录前面的复选框选择多条记录，然后点击“Display”下拉框选择所需要显示的格式来显示选定的数据。若要保存选择的数据，可以在“Send to”下拉框中选择保存的方式(保存为文本、单独文件、剪贴板或其他)。点击每一条数据编号的链接可以进入该EST数据的主页面查看其详细信息，默认显示格式为EST格式，还可以显示为GenBank格式或fasta格式。

3. I.M.A.G.E.协定简介及cDNA物理克隆的获取 许多EST所对应的cDNA克隆可通过基因组及其表达的整合分子分析(intergrated molecular analysis of genomes and their expression，I.M.A.G.E.)协定免费索取，这与电子基因克隆相辅相成，I.M.A.G.E.协定由美国LLNL国家实验室主持，宗旨是共享排列好的cDNA文库中的克隆，大规模的EST测序项目如人类EST项目等都加入了I.M.A.G.E.协定。当研究者通过另外的途径得到基因的部分序列，并通过同源性检索后发现该片段

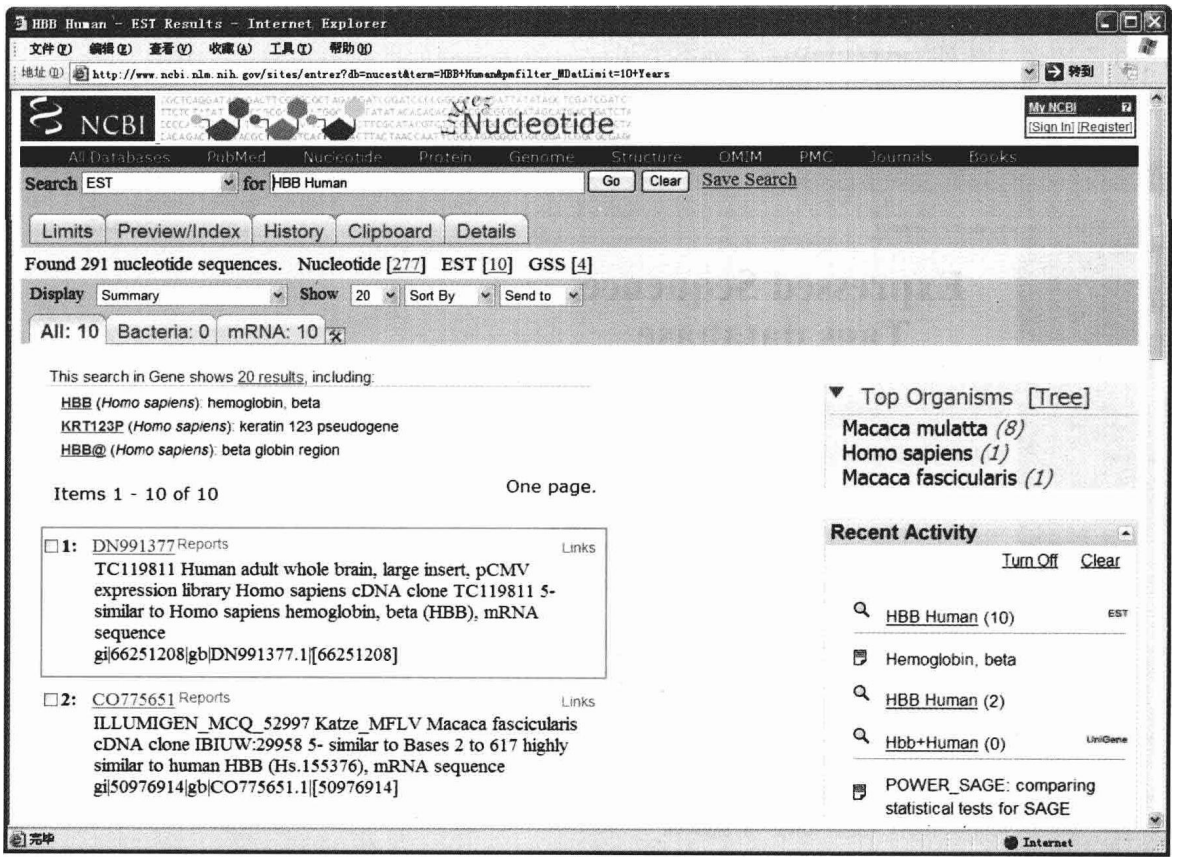


图 6-3 dbEST 中关键词“HBB Human”的检索结果

与加入 I.M.A.G.E. 协定的 EST 序列高度同源时，便可免费索取其原始克隆。可通过美国的 ATCC（American Type Culture Collection）索取，从而避免或减轻筛选全长基因的麻烦，以集中精力进行基因的功能研究。ATCC 的官方网址为 http://www.atcc.org/，用户可以利用其提供的搜索服务来搜索所需要的克隆（http://www.atcc.org/ATCCAdvancedCatalogSearch/SearchforClones/tabid/460/Default.aspx）。

（二）UniGene 数据库

1. UniGene 简介 UniGene 是 GenBank 的一部分，访问地址为 http://www.ncbi.nlm.nih.gov/unigene。UniGene 通过计算机程序对 GenBank 中的序列数据进行适当处理，剔除冗余，将来自同一基因的相关序列，包括 EST 序列片段搜集到一起，构成一个基因聚类（gene cluster）。UniGene 除了包括人的基因外，也包括小鼠、大鼠等其他模式生物的基因。一个 UniGene Cluster 包含代表单一基因的各种序列和相关信息，例如基因表达的组织类型和图谱定位信息等。UniGene 构建基因聚类时结合使用有监督的和无监督的方法，而且在聚类过程中使用了不同水平的严格度，聚类的算法为 megablast（BLAST 的一种改进算法），数据库不产生一致性序列（contig）。

2. 查询 UniGene 可以通过两种方法查询 UniGene 中的数据：第一种方法是通过 NCBI FTP 下载，UniGene 数据的 FTP 存放目录为：ftp://ftp.ncbi.nih.gov/repository/UniGene/；第二种方法是使用 EntreZ 搜索 UniGene 数据库获取所需要的数据，具体使用步骤是：首先，登录 UniGene 主页面，在搜索栏内输入你要检索的数据的相关信息。例如，要检索人类血红蛋白 β 亚基的 UniGene 数据，可以在搜索栏内输入关键词“HBB Human”。然后，点击“go”按钮进行搜索，如图 6-4 所示。由搜索结果可以看到，在 UniGene 数据库中搜索到了 2 条记录（图 6-4）。这些记录按序号排列，每一条记录均有名称并对应着该数据的链接，点击链接可以进入查看该数据记录的详细信息。例如，点击第一条记录名称“Hemoglobin，beta”的链接，可以进入新的页面显示该条 UniGene 记录的详情。

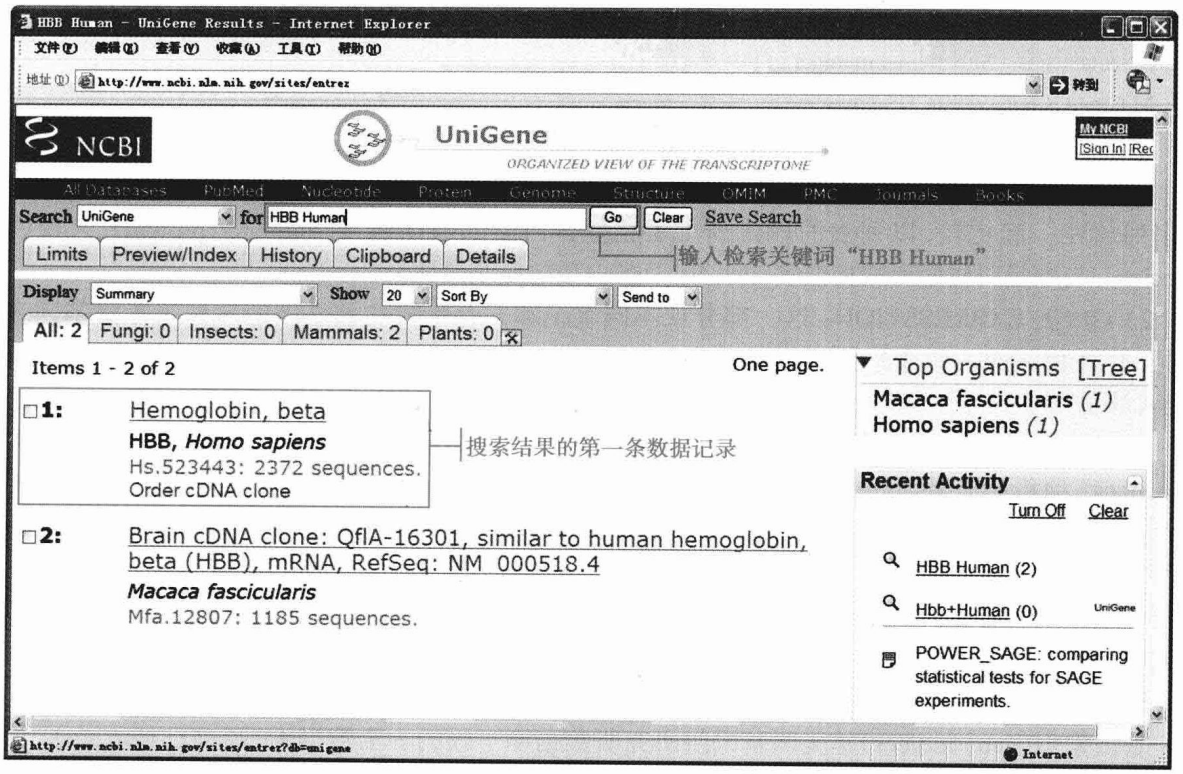


图 6-4 在 UniGene 中以关键词"HBB Human"检索人类血红蛋白 β 亚基的数据

3. UniGene 记录的解读 以上面提及的人类血红蛋白 β 亚基的 UniGene 记录为例，其 UniGene 数据文件的内容如下：

Hemoglobin, beta (HBB)

SELECTED PROTEIN SIMILARITIES

Comparison of sequences in UniGene with selected protein reference sequences. The alignments can suggest function of a gene.

……

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

……

MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

……

SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

……

mRNA sequences (25)

……

EST sequences (10 of 2336)

……

Key to Symbols

P Has similarity to known **P**roteins (after translation)

A Contains a poly-Adenylation signal

S Sequence is a Suboptimal member of this cluster

M Clone is putatively CDS-complete by **M**GC criteria

可以看到，UniGene 的数据文件记录主要分为以下几部分：

1）SELECTED PROTEIN SIMILARITIES：这一部分将与该基因编码的蛋白质相似的蛋白质序列放在一起比较，以列表的形式列出了相似蛋白质的数据链接（reference protein）、来源物种（species）、与原蛋白质序列的相似度（Id%）以及蛋白质的氨基酸数量（length）。

2）GENE EXPRESSTION：这一部分从不同组织和发育阶段的该基因的序列来考察基因表达情况，并与相关的一些表达信息链接起来，包括 EST 表达谱、GEO 表达谱（实验的表达数据）以及 cDNA 来源等。

3）SEQUECNES：与该基因相关的序列数据，主要有 mRNA 序列以及 EST 序列等。其中“mRNA sequences”项将与该基因相关的 mRNA 序列数据的信息以列表的形式给出，包括 mRNA 序列的数据库标识符、mRNA 序列的说明以及序列特点；“EST sequences”项则将与该基因相关的 EST 序列以列表的形式给出，包括 EST 序列的数据库标识符、Clone I.M.A.G.E. 中的 ID 号、表达的组织、阅读方向以及序列的特点等。EST 数据列表后面有一个“Download Sequence”按钮，点击它可以将所有 EST 序列数据以 fasta 格式下载到本地。

（三）Gene Indices 数据库

基因索引（gene indices）是美国基因组研究所数据库下属的一个数据库，目前已经收录了 42 类动物、47 类植物、15 类原生生物和 10 类真菌的详细的基因组信息。每一类生物的基因索引内容包括该索引的产生背景、版本信息，并提供同源序列的 BLAST 检索和 EST、TC、文库等序列报告；此外还包含基因表达、基因发生、代谢途径等功能注释和分析；有的还包括其他的信息和附加信息。总体来说基因索引数据库集中了许多在该物种基因组研究上有重要作用的机构的数据，内容丰富，检索方便，同时这些数据还在不断的更新和补充。该数据库的访问地址为：http://compbio.dfci.harvard.edu/tgi/。

基因索引数据库还提供了一个大规模 EST 数据聚类工具 TGICL（TGI clustering tools），该工具可在 linux 或者 sunOS 系统下运行，能够用于对来自实验的 EST/mRNA 序列进行聚类和拼接组装。

三、EST 数据分析方法

特定物种（或组织）的 EST 序列代表了随机取样地各种转录产物 mRNA，因此可能会有多个 EST 代表同一个转录产物。通过 EST 数据聚类分析，可以将代表同一个转录产物的 EST 序列归为一类，然后使用序列拼接程序装配成更长的、更高质量的序列，同时也减少了 EST 的冗余。这对于 EST 的功能识别、剪接产物的区分、基因表达谱的分析都有很大的帮助。经聚类和装配得到的唯一序列（unique sequence），代表了对所研究的物种（组织）在某一特定时间所表达的基因的随机采样的结果。

在对 EST 序列进行聚类和拼接组装之前，往往需要对 EST 数据的质量进行修订，也就是对数据进行预处理。因此，EST 数据分析主要包含四个步骤：EST 数据预处理、EST 数据聚类、EST 数据拼接、拼接结果的注释。而前三个步骤是对序列进行注释的前提和基础。

（一）EST 数据预处理

测序得到的 EST 数据往往是原始的峰图文件，或者是包含载体序列和杂质的低质量的序列，因此若需要得到高质量的 EST 序列，对原始数据进行预处理是必要的，预处理主要有以下几个方面：

1. 提取序列　若原始 EST 数据来自测序仪的峰图文件，则需要从峰图文件中提取序列。从峰图文件得到的序列文件可以保存为 fasta 格式或者其他序列聚类和拼接软件所支持的格式，然后去除其中低复杂度的区域。

2. 屏蔽赝象序列(artifactual sequences) EST 数据中夹杂着一些不属于表达的基因的序列，称为赝象序列，包括表达载体序列、重复序列以及外源的污染序列(如核糖体 RNA、细菌或其他物种基因组 DNA 等)。可以使用 BLAST、RepeatMasker 或 Crossmatch 等软件进行自动分析来实现屏蔽这些序列的目的，也可以人工除去这些序列。

3. 去除嵌合克隆(chimeric clone)序列 在克隆过程中，DNA 分子的两个片段可能会发生融合并像单一克隆那样进行复制，这种克隆称之为嵌合克隆。EST 序列的嵌合克隆是在文库构建过程的反应中产生的，其序列特征表现为：序列的中间有很长的 polyA 序列，或载体序列。因此，对 EST 数据进行质量检查的时候需要识别这类型的序列，并将嵌合的序列分离开。

4. 去除过短的序列 通常把那些小于 100bp 的序列去除掉，不参加后续的聚类拼接和注释分析。

(二) EST 数据聚类

聚类分析是一种通过将相似的或者相关联的数据划分到特定的组(类)中以简化大规模数据集的方法。EST 聚类的目的在于将属于同一个基因的具有重叠部分(over-lapping)的 EST 数据聚在一起，整合至单一的簇(cluster)中。对 EST 数据进行聚类分析有助于产生更长的一致性序列(contigs)；有助于降低数据的冗余性，更正数据的错误；有助于发现同一基因的不同剪切形式。常见的收录 EST 聚类的数据库有：①UniGene(http://www.ncbi.nlm.nih.gov/UniGene)；②TIGR Gene Indices(http://www.tigr.org/tdb/tgi/)；③STACK(http://www.sanbi.ac.za/Dbases.html)。这些数据库采用自己的聚类方法将各种 EST 数据进行了归类和注解。

EST 聚类分为不严格的聚类(loose clustering)和严格的聚类(stringent clustering)。

不严格的聚类：不严格的聚类系统产生大的、“松散”的类。在所形成的每一类中，表达基因 EST 数据的覆盖率高，含有同一基因不同的转录形式，如各种选择性剪接体、由选择不同的多腺苷酸位点(polyadenylation site)而产生地不同的转录本等。其主要缺点在于每一类中可能包含旁系同源基因(paralogous expressed gene)的转录本，信噪比低，序列的忠实性低。这种系统的代表，如 STACK 采用的基于字(word)的聚类算法，省略了序列比对过程，其核心在于识别并计算序列间有多少长度为 n 的字能够匹配，而且并未采用有关克隆的来源及注释的信息，称之为 d2_cluster 聚类方法。

严格的聚类：严格的聚类系统产生高度相关的聚类成员，因此更加可靠。但是表达基因 EST 数据的覆盖率低，因此所含有的同一基因的不同转录形式少。这种系统的代表，如 TIGR 的 Gene Indices 所采用的类似于 BLAST 和 FASTA 的序列比对程序 FLAST。这类聚类方法对序列比对搜索的结果进行分析并按照用户定义的标准判断两个序列是否为一类。

(三) EST 数据拼接

EST 数据聚类和拼接通常是一个连续的过程，称为 EST 序列组装(EST Sequence Assembling)。EST 聚类后属于同一个类的序列进行拼接，可以组装为更长的一致性序列(contig)。用于 EST 序列的聚类和组装的软件有多种，下面介绍常用的几种。

1. Phrap Phrap(phragment assembly program)是由华盛顿大学分子生物技术学院的 Phil Green 和 Brent Ewing 开发的 Phred/Phrap/Consed 软件包的一部分，它基于 swat 算法构建，主要用于 shotgun 测序序列的组装，在 unix/linux 系统下运行。它的独特之处在于有较高的精确度，可以利用整个 reads(测序得到的原始 EST 序列)，而不仅是修正过的高质量的部分来组装；它寻找序列间的重叠(overlap)部分，将高质量嵌合匹配的片段组装成 contig，最后生成完整的 DNA 序列。Phrap 还能帮助查找组装中的问题，处理大规模的数据集，是目前应用比较成功的序列拼接软件之一。Phrap 程序以及相关文档可通过电子邮件 phg@u.washington.edu 直接向作者索取。用户在邮件中需要附上尊重作者版权的相关协议、承诺和个人信息，书写内容和方式可参考网站：http://www.phrap.org/phredphrapconsed.html。

(1) 程序安装：上传 Phrap 的程序文件到本地 linux/unix 服务器，编译安装即可。如果编译器不

识别 -O2 最优标记，可将 makefile 文件中 CFLAGS=-O2 改为 CFLAGS=-O，删除所有扩展名为 .o 的文件后重新编译。如果数据集多于 64 000 条序列，或者序列中含有长于 64 000bp 的序列，则需要使用 phrap.manyreads 文件或 phrap.longreads 文件，这两个程序编译命令为：$ make manyreads。

（2）程序运行：程序运行的基本命令为 phrap seq_file [-option1 value] [-option2 value] ……，其中，seq_file 为输入的 fasta 格式的核酸序列文件名，方括号内为可选项，用于设置各种参数。如有质量文件，则需和序列文件放在同一目录下，且名为“序列文件名 .qual”。例如序列文件名为 test.seq.screen，质量文件名必须为 test.seq.screen.qual。质量文件不需要包含在命令行中，并且质量文件中的序列和序列文件中的序列必须一一对应，包括阅读顺序和碱基个数。以人类视黄醇结合蛋白基因（RBP4）的 UniGene 条目中收录的 318 条 EST 序列（截至 2009 年 12 月）为例，将该 318 条 EST 的 fasta 格式的序列保存在 test.seq.screen 文件中，使用 Phrap 进行聚类和组装的操作命令为：

phrap test.seq.screen –view –new_ace –minmatch 20 –minscore 40>phrap.out

该命令对序列文件 test.seq.screen 中的序列进行聚类和拼接，要求聚类的序列间最小重叠区域长度为 20bp，最小比对分值为 40。运算完成后，程序运行目录会产生一系列相关文件，分别为：

*.contigs 文件。组装好的 contig 序列，格式为 fasta 格式。其中包括单个 read 的 contig（这类 reads 和其他 contig 有比对上的部分，但达不到与之连接的标准）。

*.contigs.qual 文件。contig 组装的质量文件，格式为 fasta 格式。此文件记录每个 contig 的碱基质量信息。

*.singlets 文件。包括所有没有拼接到任何一个 contig 中的单独序列，格式为 fasta 格式。

*.log 文件和 *.problems 文件。包括各种诊断信息和问题的归纳，对使用者基本没用。

*.ace 文件。当使用参数 -new_ace 或 -old_ace 时才会产生的文件，用 consed 查看组装结果时需要该文件。

*.view 文件。当使用 -view 参数时产生的文件，用 phrapview 查看组装结果时需要该文件。

phrap.out 文件。记录组装过程信息和组装结果（哪些 EST 序列被组装到哪一条 contig，以及重复区域碱基配对结果等）。

（3）参数设置：详细的参数列表可以查看 Phrap 帮助文档，一些基本参数设置项及默认值如下：

1）比对打分

-penalty　用于 swat 比对时碱基不匹配（替换）罚分，默认值为 22；

-gap_init penalty　gap 罚分，默认值为 −2；

-gap_ext penalty　扩展 gap 罚分，默认值为 −1；

-ins_gap_ext gap_ext　插入罚分；

-del_gap_ext gap_ext　删除罚分；

-matrix　打分矩阵；

-raw　只使用原始的 Smith-Waterman 打分。

2）结合搜索

-minmatch　最小匹配长度，默认值为 14；

-maxmatch　最大匹配长度，默认值为 30；

-max_group_size　最大组限制，默认值为 20；

-bandwidth　在 swat 的片段比较中设定的 1/2 片段宽度（全长 2×bandwidth+1），减小片段宽度会缩短比较时间，但是会降低灵敏度。默认值为 14。

3）比对过滤参数

-minscore　最小比对分值，默认值为 30；

-vector_bound　序列开始部分可能的载体碱基数目，默认值为 80；

-masklevel　值为 0 时只报告单条最高分值的比对，为 101 时报告所有的比对，默认值为 0。

4）输入相关

-default_qual 没有质量文件时的碱基默认质量值，默认值为 15；

-subclone_delim 克隆名称的分隔符号；

-trim_start 序列开头去掉的碱基数，默认值为 0。

5）拼接相关

-forcelevel 在最后的 contig 合并周期中用于控制的严格度参数，可以从 0(最严格)到 10(最不严格)之间变化，默认值为 0；

-bypasslevel 合并时对于不一致序列的控制严格度，默认值为 1；

-maxgap 合并 contig 时允许的最大的不匹配区域的长度，默认值为 30；

-repeat_stringency 控制匹配的严格度，小于 1(最严格)而大于 0(最不严格)，默认值为 0.95；

-revise_greedy 在弱结合部位打断，并尝试重新结合；

-shatter_greedy 打断弱的结合但不尝试重新结合；

-preassemble 组内序列先结合。

6）一致性序列构建参数

-node_seg 最小线段的大小(用于通过加权有向图)，默认值为 8；

-node_space 加权有向图节点之间的间距，默认值为 4。

7）输出相关

-tags 标准输出时选择行的标签，有助于配对；

-old_ace 产生旧格式的 ace 文件；

-new_ace 产生新格式的 ace 文件；

-ace 同参数 -new_ace；

-view 产生适用于 phrapview 的 ".view" 文件；

-print_extraneous_matches 打印 contigs 间的非局部比对信息。

8）其他

-retain_duplicates 保留完全相同的序列，而不是去除；

-max_subclone_size 最大克隆长度，默认值为 5000；

-trim_qual 定义序列高质量部分的质量值，默认值为 13。

(4) 注意事项：通常情况下 EST reads 数量(即原始序列数量)不要超过 15 万。如果覆盖度不是很高并且重复序列很少，Phrap 能完成 50 万条以下序列的拼接；如果覆盖度很高或者重复序列很多，Phrap 处理序列的难度将大大增加。对于重复序列，可以通过对序列的统计分析去掉高重复的 reads，只保留具有唯一序列的 reads 进行拼接。此外，如果程序运行提前终止，并给出以下错误信息提示：fatal error: requested memory unavailable，则表示计算机内存不足；若程序运行时间过长，则可以试着提高参数 -minmatch 的值。

2. CAP3 CAP3 是一个用于 DNA 序列聚类和拼接的程序，属于 CAP 程序的第三版，较以前的版本有了一些改进，能够消除 3' 端和 5' 端的低质量区域；在计算 reads 之间的重叠时使用碱基质量信息来进行多序列比对，从而产生一致序列；也可以利用正反向关系来确认和连接 contigs。该软件可以通过邮件向作者索取(huang@mtu.edu)，也可以从网址下载原程序和相关文档(http://seq.cs.iastate.edu/)。下载该程序的源码前首先需要填写个人相关资料信息，下载页面提供了 windows、linux/unix 等不同平台的二进制源代码的下载，用户可以根据自己的需要进行选择。例如 windows 平台下，将对应的程序压缩包下载解压缩以后，将 cap3 文件重命名为 cap3.exe，即为 CAP3 的可执行程序。

以 windows 平台为例，简要介绍 CAP3 的使用。Windows 版本的 CAP3 是通过命令行进行操作的，其程序的基本运行命令为：cap3 File_of_reads [options]。

例：cap3 test.seq.screen –o 21 >cap3.out

File_of_reads 是输入的 fasta 格式的 EST 序列文件。CAP3 和 Phrap 一样，质量文件也是可以选择的，其文件名必须和序列文件名一致(如 test.seq.screen 对应的质量文件名 test.seq.screen.qual)。[options] 为可选择的参数设置项。CAP3 的一些主要的参数设置项如下：

-a N 带宽扩展长度，要求 N>10，默认值为 20；
-b N 不同碱基质量的切除值，要求 N>15，默认值为 20；
-c N 末端碱基质量切除值，要求 N>5，默认值为 12；
-d N 在不同时的最大 qscore 的总和值，要求 N>20，默认值为 200；
-e N 不同数目间的清除值，要求 N>10，默认值为 30；
-f N 任意重叠区域的最大空位长度，要求 N>1，默认值为 20；
-g N 空位罚分，要求 N>0，默认值为 6；
-h N 最大突出的长度的百分数，要求 N>2，默认值为 20；
-m N 匹配得分，要求 N>0，默认值为 2；
-n N 不匹配得分，要求 N<0，默认值为 −5；
-o N 重叠的长度阈值，要求 N>20，默认值为 40；
-p N 重叠一致的百分比，要求 N>65，默认值为 80；
-r N 反向值，要求 N≥0，默认值为 1；
-s N 重叠相似度打分阈值，要求 N>400，默认值为 900；
-t N 词匹配的最大数量，要求 N>30，默认值为 300；
-u N 修改约束的最小数量，要求 N>0，默认值为 3；
-v N 连接约束的最小数量，要求 N>0，默认值为 2；
-w N 剪切信息的文件名，默认值为 none；
-x N 输出文件的前缀字符串，默认值为 cap；
-y N 剪切范围，要求 N>5，默认值为 250；
-z N 剪切位置好的 reads 的编号的最小值，要求 N>0，默认值为 3。

执行程序运行命令 cap3 test.seq.screen –o 21>cap3.out，对 test.seq.screen 中的序列进行组装后，其输出结果中各文件的含义如下：

cap3.out：记录 contig 的组装信息；
test.seq.screen.cap.contigs：组装成的 contigs 文件，fasta 格式；
test.seq.screen.cap.contigs.qual：contigs 的质量文件；
test.seq.screen.cap.singlets：收录没有参与组装的单独序列(singlets)，fasta 格式；
test.seq.screen.cap.info：信息文件，主要为 reads 的切除信息；
test.seq.screen.cap.contigs.links：contigs 的关系文件；
test.seq.screen.cap.ace：用于 Consed 程序的输入。

在 test.seq.screen.cap.contigs 中可以看到来自 RBP4 的 318 条 EST 序列拼接成了 4 条 contigs；在 test.seq.screen.cap.singlets 中可以看到有 8 条序列没有参加组装。

CAP3 除了单机版软件外，还有基于 web 在线使用的版本，其访问地址为：http://pbil.univ-lyon1.fr/cap3.php。登录该页面，将待组装的序列文件以 fasta 格式输入到文本输入框中，然后点击“SUBMIT”按钮提交就可以了。需要注意的是，序列文件的大小不能超过 50kb。组装结果将由服务器返回并显示在原网页上，包括“contigs”(记录拼接得到的 contigs 序列)、“Single sequences”(未参与拼接的序列)、“Assembly details”(组装详细情况)、“Your sequence file”(用户提交的序列信息)等。

3. TIGR Assembler　TIGR Assembler 是用于将大批量 shotgun 测序获得的 DNA 片段组装成一致性序列的工具，由 TIGR(The Institute for Genomic Research)于 1995 年研发。TIGR Assembler 的

工作原理与步骤如下：①对所有的数据进行快速初步的两两比对(类似于 blast)，生成一个潜在的重叠序列列表；②利用潜在的重叠序列的分布来标记一条序列是否含有重复区域；③组装从无重复区域存在的序列开始(如果不存在无重复区域的序列，则对重复序列进行组装)，直到没有序列存在时结束；④利用潜在重复序列列表提供的信息对现存的序列集和无重复的序列进行组装。TIGR Assembler 的相关使用文档可以从以下网站下载获取：http://www.jcvi.org/cms/publications/listing/abstract/article/tigr-assembler-a-new-tool-for-assembling-large-shotgun-sequencing-projects/。

(1) 在线版 TIGR Assembler 的使用：在线版的 TIGR Assembler 的访问地址为：http://nbc11.biologie.uni-kl.de/framed/left/ menu/auto/right/tigr_assembler/。用户登录该页面，可以在文本输入框内直接输入 fasta 格式的 EST 序列，也可以上传本地计算机上的序列文件，设置好参数后点击“Assemble!”按钮提交任务，计算完成以后序列组装的结果将显示在页面上。该 web 版本的程序提供三个参数设置项：“Minimum length”项为被考虑进行组装的两个 DNA 序列的最小重复序列长度，最小值为 32，默认值为 40；“Maximum Error 32”项为同一个组装中两个 DNA 序列重复区域相邻的 32 个碱基配对中允许的最大错误数量(空位或错配)加 1，默认值为 8；“Num Conflicts”项为如果一个好的克隆配对在同一个组装中不存在导致一条具有一致序列的原序列被 GDE.align 文件丢弃，属于这类情况的最大质量冲突数量，默认值为 2。

(2) 单机版 TIGR Assembler 的使用：单机版的 TIGR Assembler 可以通过下列 FTP 地址下载：ftp://ftp.tigr.org/pub/software/assembler/。单机版程序应用于 unix/linux 平台，基本的运行命令为：TIGR_Assembler [options] scratch_file <input_file> output_file。

例：TIGR_Assembler -f test.cluster.fasta –l 20 scratch_file <test.seq.screen> tigr.out

该命令对 test.seq.screen 中的序列进行聚类和拼接，结果保存在 tigr.out 文件中。程序的主要参数设置项如下：

input_file：输入文件名，fasta 格式；

-a：将组装生成的各个 contig 文件输出至指定的目录，默认为 result 目录；

-C：contig_file 是否包含组装信息的文件；

-q：输入质量文件，为组装结果提供质量信息；

-g：若将 max_err_32 定义为两个重叠的序列间连续 32 个配对的碱基内部可以存在的最大的错配(或空位)数量，则该项值设为 max_err_32+1；

-l：两条 DNA 序列在组装中重叠部分的最小限制长度；

-A：组装结果以 contig 的形式列出，同时注有质量信息和 contig 的组成信息；

-f：输出一个多 fasta 序列文件，将拼接成的 contigs 和不能拼接成 contigs 的 singlets 一起写入该文件；

output_file：输出文件，包含组装结果的各种信息。

将待组装的序列文件放在当前目录下，运行程序命令，即可进行组装。

4. Staden Package　Staden Package 是一个用于测序项目管理的整合软件包，包括序列组装、突变检测、序列分析、序列峰图以及对 reads 文件操作等功能。其中的序列组装程序包括了所有类型的组装所需要的工具和方法，同时又具有许多自己的特点，比如友好的用户界面。这个程序包除了 unix/linux 系统版本之外，还包括一个 windows 版本，这为不熟悉 linux 系统的用户提供了更好的选择。该软件包及相关信息可以通过下列网址获得：http://staden.sourceforge.net/。

Staden Package 软件包中用于序列组装的程序有两个：pregap4 和 gap4。pregap4 是 gap4 的前处理程序，可以处理原始的峰图文件，以及对序列进行载体和污染检查，还可以进行序列组装，经 pregap4 处理得到的结果可以通过 gap4 进行查看和编辑。

(1) pregap4：pregap4 的输入可以是原始的各种测序峰图文件，也可以是独立的 fasta 格式的序列文件，其程序界面如图 6-5 所示：

pregap4 的主界面有三个标签，分别是“Files to Process”、“Configure Modules”和“Textual Output”。在“Files to Process”标签页面上点击“Add files”可以将序列文件或峰图文件导入，或者点击“Add files of filenames”导入一个序列文件列表。

文件导入以后，点击“Configure Modules”标签，在该标签页面设置各种拼接参数(图 6-6)。“General Configuration”栏内有多个参数设置项，使用鼠标点击每一项前面的方括号可激活或关闭该项。参数项关闭时字体为灰色，激活时字体为黑色，只有激活时右侧的参数输入和编辑操作才有效。pregap4 主要的参数设置项有：

Estimate Base Accuracies：从原始的峰图文件中读出碱基，并评估质量；

Trace Format Conversion：拼接后输出文件格式；

Initialise Experiment Files：默认模块，生成包含 ID、EN、LN、LT、AQ 和 SQ 信息的实验文件，为后续的载体标记、屏蔽和装配模块的运行提供必需的信息；

Augment Experiment Files：有两个 database 可以选，用来增加实验文件信息；

Quality Clip：去除低质量的数据，其中，如果输入的是 fasta 格式的序列文件，则 Clip Mode 项选择“by base calls”，如果输入的是峰图文件，则选择“by confidence”；

Sequencing Vector Clip：在 reads 中标记克隆载体序列，如果为 PCR 产物测序，则无需选择此模块，拼接前，先将载体序列放到 vector-primer 文件中；

Screen for Unclipped Vector：将 reads 中的载体去除；

Cloning Vector Clip：对 BAC to BAC 方法进行测序的 reads，检查是否有克隆载体序列；

Gap4 shotgun assembly：对拼接的结果数据库文件取名，选择“Creat new database”则创建一个新的数据库，选择“Append to existing database”则将结果文件添加到已有的数据库中。

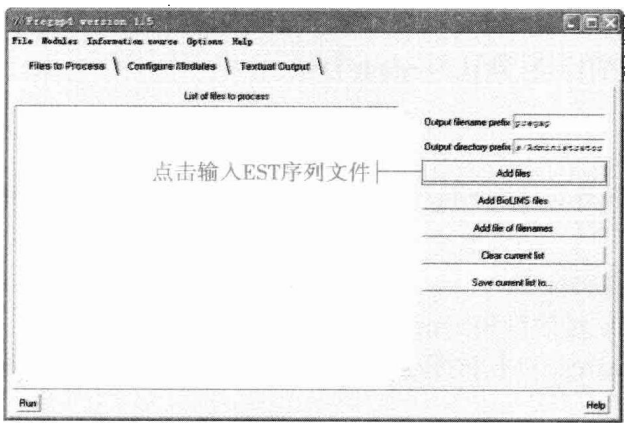


图 6-5 pregap4 的文件输入界面

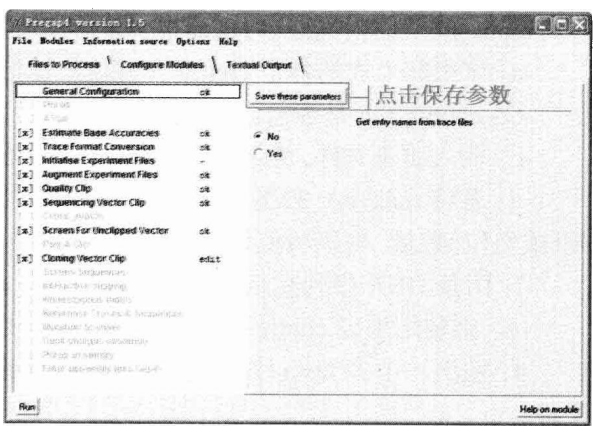


图 6-6 pregap4 的参数设置界面

各种参数设置好以后，点击窗口左下角的“RUN”按钮，程序便开始运行，并自动转到“Textual Output”标签页面中(图 6-7)，该标签内的“Output window”窗口中显示拼接的进度，拼接过程中的错误信息会显示在下面的“Error Window”窗口内。拼接得到的一系列文件一般保存在输入数据的存放目录下，这些文件分别是：

*.list.passed：通过处理的序列列表；

*.list.failed：未通过处理的序列列表；

*.list.log：日志文件；

*.list.report：报告文件；

*(设定的数据文件名).0.aux：用于启动数据库的文件；

*(设定的数据文件名).0：和 *.0.aux 共同构成了拼接结果数据库；

*.0.log：对该数据库所执行的操作的记录。

（2）gap4：pregap4 的拼接结果可以使用软件包中的 gap4 程序进行检查。双击拼接结果数据库文件“*.0.aux”，就可以启动 gap4 来检查其中的拼接结果（如图 6-8，以程序安装目录内自带的 demo 为例）。Gap4 启动后会出现两个窗口，其中一个为主窗口，显示序列拼接的结果相关信息，另一窗口为“contig Selector”窗口，其中所有的 contig 用一根横线来表示，各 contigs 之间用竖线间隔，当鼠标移动到任一条 contig 的位置时，窗口底部将显示该 contig 的名称及包含的 reads 的数目。在窗口的“View”菜单下可以查看拼接相关的其他信息，比如查看 reads 在 contig 上的相对位置、显示 contig 的限制性内切酶位点、搜索终止密码子等。

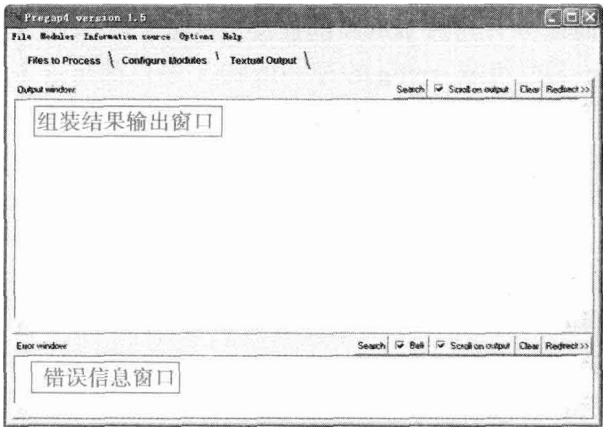


图 6-7 pregap4 的运行结果显示界面

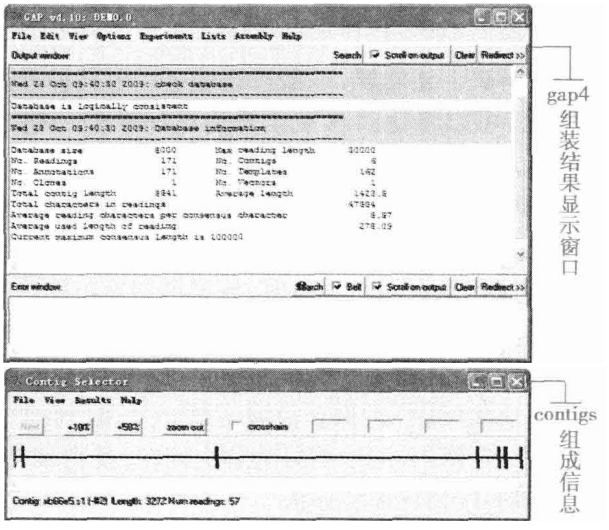


图 6-8 EST 序列组装信息在 gap4 中的显示界面

5. EST 聚类软件比较及常见问题 通过以上介绍，对各种不同的 EST 数据聚类和拼接软件有了初步的了解，用户可以根据自己的需要来选择合适的软件。对上述 4 种软件进行一个简单比较，如表 6-3 所示。

表 6-3 4 种 EST 组装软件比较

	phrap	CAP3	TIGR Assembler	Staden Package
应用平台	unix	unix/windows	unix	unix/Windows
可获得性	学术用户取得认证后可免费下载使用	需要联系作者获取	免费下载	免费下载
输入数据	海量数据，长短 reads 皆可	大量数据	大量数据	大量数据
用户界面	命令行	命令行	命令行	命令行 / 图形界面
主要应用	基因组、EST	EST	EST	基因组、EST

对习惯 unix/linux 系统的用户而言，四类软件都适合用于对 EST 序列数据进行聚类和拼接，其中 phrap 对于不同长度和规模的序列的适应性最为出色；对习惯 windows 的用户而言，CAP3 和 Staden Package 都可以提供很好的 EST 数据分析功能，而 Staden Package 具有图形化界面，人机交互更加方便。

需要注意的是，虽然有各种各样的软件可用来对 EST 数据进行聚类和组装，但是由于 EST 数据量庞大加上生物学数据的复杂性，使得这些软件在某些情况下也会出现一些问题，主要有：

1. 错拼 错拼指的是不应该拼接在一起的 EST 被拼接到了一起。造成错拼的原因主要有：①3’端 EST 序列的 ployA 尾巴的存在；②嵌合克隆 reads 的存在，连接了两条（或多条）本不相关的

EST；③基因家族中序列相似性高的序列（重复序列）无法被软件很好地识别，从而拼接在一起；④其他类型的重复序列，也会导致错拼。

2. 漏拼 漏拼就是遗漏了某些 EST，使得应该拼接在一起的 EST 没有拼接在一起。产生漏拼的原因主要有：①EST 序列末端质量较低，在较为严格的拼接参数下这些序列相似性不能达到标准，从而没有被拼接在一起；②嵌合克隆 reads 的存在，连接了两条（或多条）本不相关的 EST，从而使本该连接的 EST 序列不能连接到对应的 EST 上；③各种重复序列的存在，也同样可以导致漏拼。

3. 选择性剪切引起的错误 在大多数真核生物转录后的修饰和调控中，都存在选择性剪切现象，这在高等动物中所占的比例尤其明显（人类中超过 50%）。从不同的剪切产物所得到的 EST 序列，它们之间既有部分相似，又有部分不同，这样序列同样也会导致在拼接中出现如嵌合克隆存在时出现的问题，当然这样产生的差错不能简单地说成是错拼或漏拼。

要解决上述问题，一般可以通过以下几种方法来进行检查和处理：①调整拼接参数，如对 polyA 的情况，可以将比对参数设置得更加严格一些，从而区分开这些 polyA 的连接；而对 EST 末端质量较低造成的漏拼，则需要将参数设置得宽松一些；②将组成 contig 的每一条 EST 序列都与这个 contig 进行比较，那些不完全比对上的 EST 有可能是错误拼接的，也可能是选择性剪切造成的结果；③与已知核苷酸和蛋白质数据库进行 Blast 比对，如果出现同一个 contig 比对上两个不同的 contig 或 singlet 比对上同一个基因，则说明很有可能这些 EST 是可以聚在一起的；④挑选出那些经检查后可能拼接在一起的序列进行单独的拼接，避免那些重复或嵌合克隆等造成的影响。

总而言之，当出现拼接问题时，要综合考虑可能造成这些问题的各种情况，搞清楚是参数问题还是序列质量问题，是嵌合克隆问题还是选择性剪切问题，是基因家族还是其他重复问题造成的。多种方法结合使用，可以更好地解决问题，使得聚类结果更加准确。

（四）序列注释和分析

由 EST 序列组装得到 contigs 或基因序列，可以进一步的注释和分析，从而更深入了解基因表达的相关特征。对序列的注释和分析通常包含下列几个方面：

1. 一级序列同源性比对 为了寻找 EST 序列或基因序列在其他物种中的同源序列，可以通过序列比对的方法，利用序列比对工具如 BLAST，在相应物种的数据库中搜索可能的同源序列。一般来说，利用核酸和蛋白质一级数据库中的 blastn 或者 blastx 程序，能够准确地搜索到 EST 来源的基因序列或编码的蛋白质序列在相应物种中的同源序列。由已知同源序列的注释信息可以推测所研究基因的功能。

2. 蛋白质结构域和功能位点搜索 在蛋白质家族数据库（Pfam）或者其他蛋白质功能库（如 Interpro）中可以搜索 EST 来源的基因编码的蛋白质的家族、结构域、作用位点等信息。

3. 基因功能分类 对基因进行功能分类是基因分析的重要环节。利用标准基因词汇体系 Gene Ontology（GO），可以进行近似的分类。Gene Ontology 包含了基因参与的生物过程，所处的细胞位置，发挥的分子功能三方面的功能信息，并将粗细不同的功能概念组织成有向无环图（directed acyclic graph，DAG）的结构。Gene Ontology 是一个使用有控制的词汇表和严格定义的概念关系，以有向无环图的形式统一表示各物种的基因功能分类体系，从而较全面地概括了基因的功能信息，纠正了传统功能分类体系中常见的维度混淆问题。在基因表达谱分析中，Gene Ontology 常用于提供基因功能分类标签和基因功能研究的背景知识。利用 Gene Ontology 的知识体系和结构特点，可以发掘与基因差异表达现象关联的单个特征基因功能类或多个特征功能类的组合。Gene Ontology 数据库的访问地址为 http://www.geneontology.org/，利用该数据库，通过浏览、查询和 BLAST 搜索等方式可以获得基因的功能分类信息。

4. 表达量比较分析 表达量比较分析主要是在不同的组织或者不同情况的表达序列之间进行量化比较。将来自不同组织的 EST 序列进行比较，可以了解同一基因在不同组织内的表达水平。

5. 通路（pathway） 分析使用 EST 数据，结合利用，如 KEGG（http://www.genome.jp/kegg/）、

BioCarta(http://www.biocarta.com/)等通路信息数据库资源，可以进行相关通路分析、差异表达基因的功能分类、所属信号通路分类等。

6. 可变剪切分析 真核基因的可变剪切是一种重要的表达形式。根据 EST 序列本身以及基因组和 mRNA 序列可以预测相关的可变剪切形式。例如，将 EST 序列比对到基因组或者 mRNA 的对应位置上，可以了解基因是否存在可变剪切的表达。

【例 6-1】 家猪脑组织 EST 序列分析

首先获得来自不同发育阶段的家猪脑组织 EST 序列，文库信息如表 6-4 所示。

表 6-4 家猪脑组织 EST 序列文库数据

Library name	cbe	ece	fce	ecc	fcc	ebs	fbs
Tissue	Cerebellum			Cortex cerebrum		Brain stem	
Develop-mental phase	adult	Foetus 50d	Foetus 100d	Foetus 50d	Earlyborn 107d	Foetus 50d	Newborn 115d

第一步：数据预处理采用 crossmatch 等软件去除载体序列、细菌基因组序列等污染序列，使用 RepeatMasker 等软件屏蔽重复序列，丢弃低复杂度区域和小于 100bp 的序列。预处理后得到 46 011 条高质量序列，其长度和质量统计信息如图 6-9 所示。

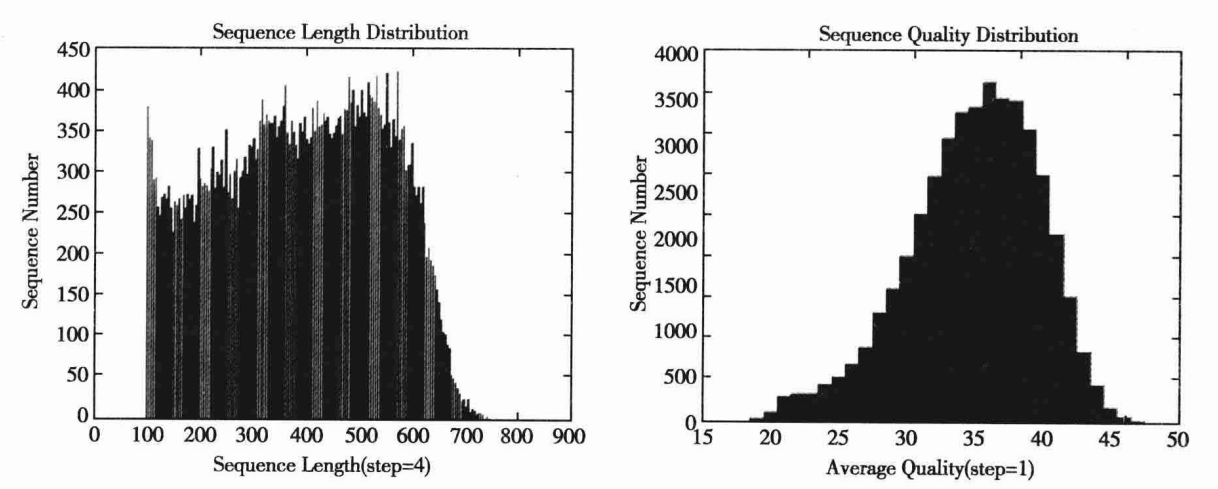


图 6-9 家猪脑组织 EST 序列的长度和质量统计图

第二步：聚类和拼接分别使用 phrap 和 CAP3 对 46 011 条 EST 序列进行聚类和拼接(使用默认参数)，结果如表 6-5 所示。

表 6-5 phrap 和 CAP3 对家猪脑组织 EST 序列的组装结果比较

软件	高质量序列	Contigs	Singlets
phrap	46 011	5740	10 763
CAP3	46 011	5176	13 459

可以看到，CAP3 采用的聚类和拼接方法更加严格一些，得到更少的 contigs 数量。

第三步：注释和分析。

1. 同源性搜索 使用 BLAST 在人类 EST 库中搜索 contigs 序列的相似序列，发现 76% 的序列存在匹配结果，24% 的序列没有匹配结果，这反映出猪脑组织和人脑组织表达序列的差异程度。同样使用 BLAST 程序在人类基因组中进行比对搜索，提取 E 值低于 1e-5 的序列，在人类各条染色体

上命中的目标数量如图 6-10 所示。从图上可以看到猪脑组织表达序列在人类中的同源序列在各染色体上分布不均衡，一号染色体上同源序列数量最多。

2. 基因功能分类　按照 Gene Ontology 的三个标准——分子功能(molecular function)、生物学过程(biological process)和细胞组分(cell component)对序列进行注释，结果分别如图 6-11、6-12、6-13 所示。

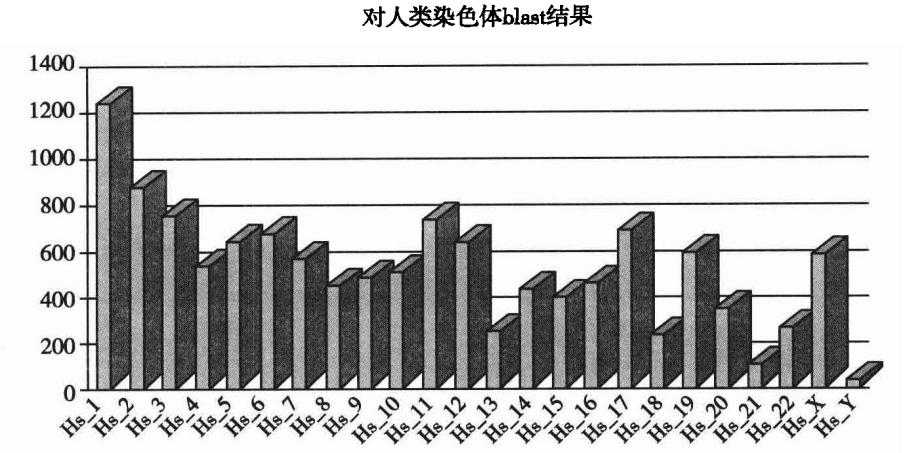


图 6-10　家猪脑组织 EST 在人类染色体上的 BLAST 搜索结果

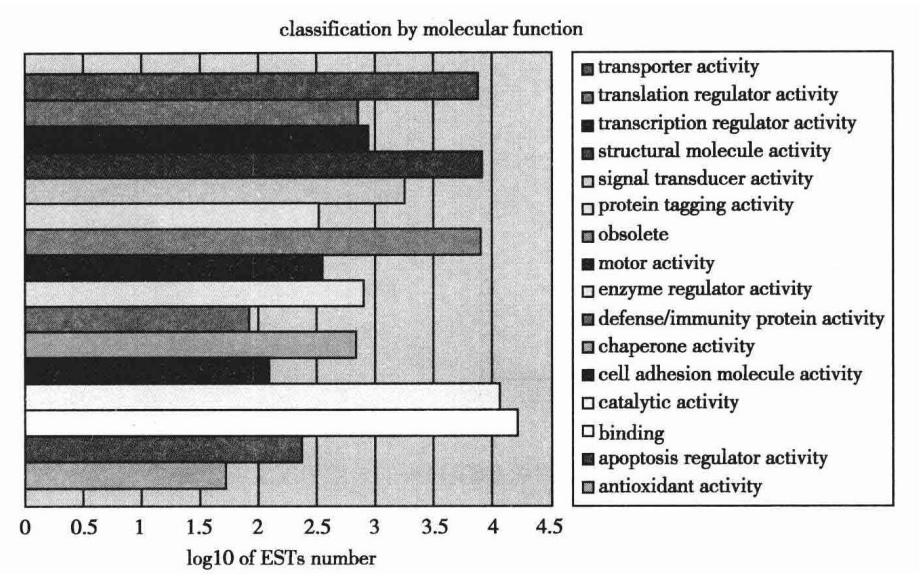


图 6-11　家猪脑组织表达序列的分子功能分类图

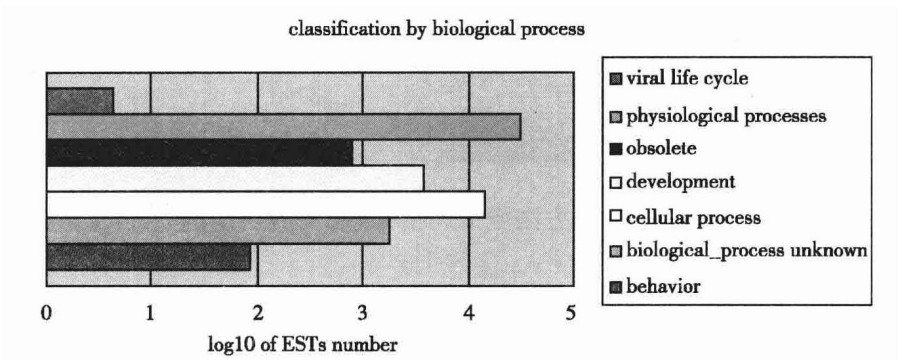


图 6-12　猪脑组织表达序列的生物学过程分类图

3. 表达量比较分析 对来自猪脑不同组织的EST数据中的翻译控制肿瘤蛋白(translationally controlled tumor protein, TCTP)表达序列的比例进行统计和比较，如图6-14所示。该图显示TCTP在不同组织的不同文库中的表达量存在明显差异，例如来自小脑组织的cbe文库中，TCTP的表达量明显高于其他组织。

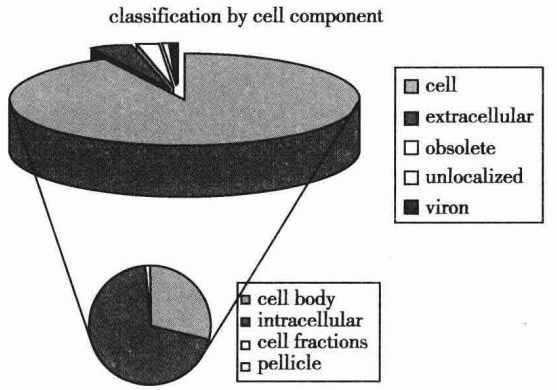


图6-13 家猪脑组织表达序列的细胞组分分类图

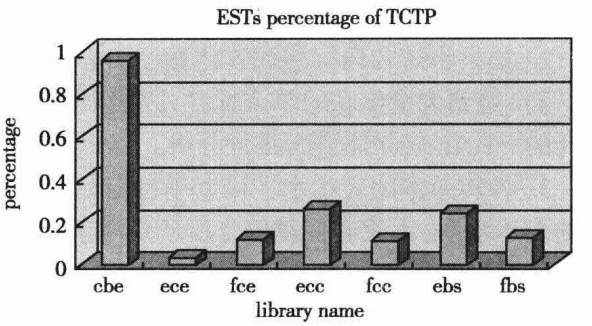


图6-14 家猪脑组织中TCTP的表达量比较

第三节 基因表达系列分析

Section 3 Serial Analysis of Gene Expression

知 识 拓 展

目前用于转录组数据获得和分析的方法主要有基于杂交技术的芯片技术包括cDNA芯片和寡聚核苷酸芯片，基于序列分析的基因表达系列分析(serial analysis of gene expression, SAGE)和大规模平行信号测序系统(massively parallel signature sequencing, MPSS)。SAGE可以帮助获得完整转录组图谱、发现新基因及其功能、作用机制和通路等信息。MPSS是对SAGE的改进，它能在短时间内检测细胞或组织内全部基因的表达情况，是功能基因组研究的有效工具。因其需要配套的软硬件较为昂贵，目前国内外的相关应用报道不多。MPSS技术对于致病基因的识别、揭示基因在疾病中的作用、分析药物的药效等都非常有价值。

一、SAGE技术原理简介

基因表达系列分析(SAGE)是1995年由Velculescu提出的一种快速分析基因表达谱信息的技术，发表在著名杂志Science上。SAGE技术在理论上来说可以检测到一个细胞内所有表达的转录本，而且可以给每一个转录本定量，不管它是低丰度还是高丰度。SAGE和基因芯片技术一样，具有高通量、平行性检测细胞内基因表达谱的特点。但它可在未知任何基因或EST序列的情况下对靶细胞进行表达谱研究，这一点是基因芯片技术所不具备的。SAGE区别于差异显示、消减杂交等其他技术的主要特点是可用于寻找那些较低丰度的转录物，最大限度地体现基因组的基因表达信息，这使之成为从总体上全面研究基因表达、构建基因表达图谱的首选策略。并在此基础上，可发现新的基因。

SAGE的核心是以高通量方式快速检测能独特代表每个基因转录本序列的标签(tag)，标签的长度约9~12bp，同一tag在某组织中出现的频度反映的是该tag所代表基因在这种组织中的表达丰度。目前，NCBI通过Internet操作平台提供了多种来源的肿瘤组织、细胞系及相应正常组织近100多个SAGE文库数据，为能找到代表性SAGE tag的基因进行多组织、不同细胞间表达谱的定量分析、比

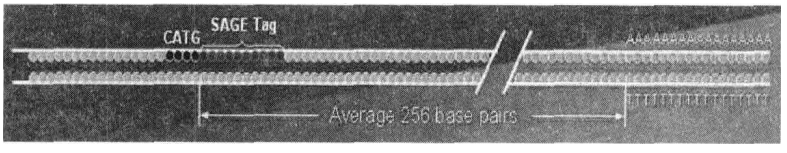
较以及途径(或过程)的功能实现的科学描述提供了良好的实验材料和环境。因此,该技术是试图从系统生物学层面来回答生物学现象的强有力技术之一。下面简介其技术原理。

SAGE 技术得以成立的理论基础有三个基本要点:

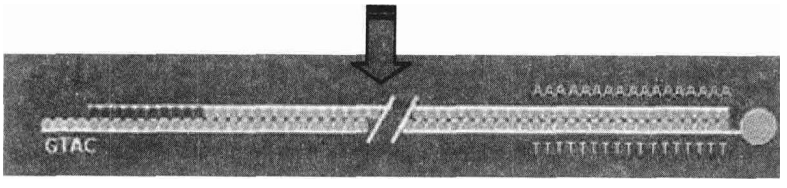
1. 一段来自任一转录本特定区域的长度仅 9~14bp 的短核苷酸序列"标签"(Tag),就已包含了足够的信息来特异性地确定该转录本。理论上讲,一个 9 碱基的序列能有四的 9 次方(即 262 144)种不同的排列组合,而人类基因组据估计仅编码约 80 000 种不同转录本,因此在理论上每一个独特 9 碱基标签就能够代表一种转录本的特征序列。

2. 如果将短片段标签相互连接,形成串联的 DNA 分子,再对该标签串联体分子进行克隆和测序,就可得到大量串联的单个标签,并能以连续的数据形式进行处理,这样就可快速、廉价地对数以千计的 mRNA 转录本进行批量分析。

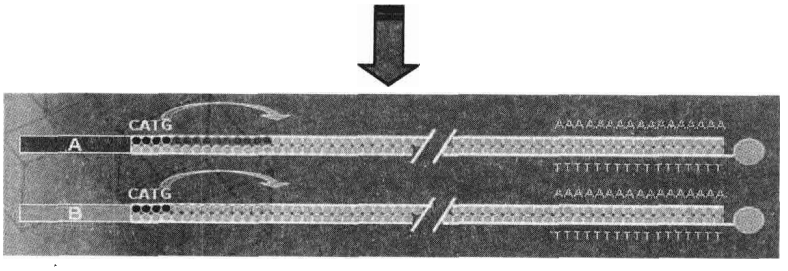
3. 各转录本的表达水平可以用特定标签被测得的次数来定量。因此该技术可以精确地定量基因的表达水平,不论它们是低丰度还是高丰度的。下面的示意图将有助于理解 SAGE 技术的原理及其操作流程。



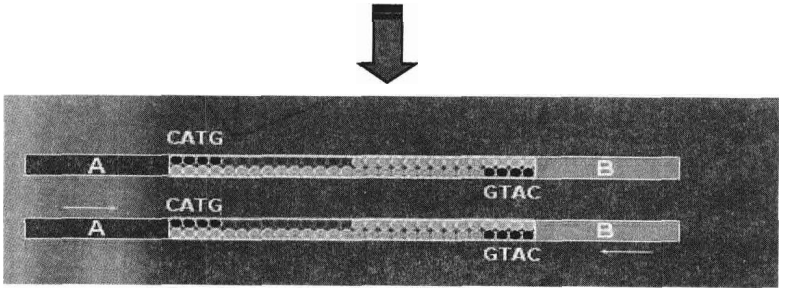
提取总 mRNA,用生物素标记的 Oligo d(T)引物反转录,生成 cDNA。用锚定酶(anchoring enzyme)NlaⅢ酶切。该酶的识别位点是 CATG。(理论上,每 256 个碱基内就会有一个 CATG 位点)。



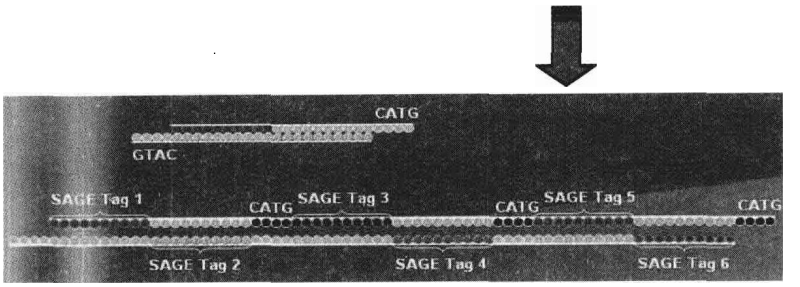
用 NlaⅢ酶切后产生的小片段"标签"(tag)。将其分两份,分别连接 A、B 接头。接头内含有标签酶 BsmF1 识别位点序列和 PCR 引物 A、B。



BsmF1 能在识别位点下游 13~14bp 处(跨越 CATG 后约 9~10bp)切割 DNA 双链,产生平末端,得到长约 50bp 的标签。



混合A、B片段，用连接酶连接，得到双标签(ditag)，用引物A、B通过PCR扩增双标签。



用锚定酶消化双标签，产生带CATG的末端；再用连接酶连接双标签可以得到一个长的DNA分子。对此长DNA分子进行克隆与测序。用SAGE软件分析标签数据。测得的每一个不同序列特异性的标签，代表的是每一个不同的基因。特异性标签的种类体现的是有多少种不同的基因在表达。同一种标签测到的次数就是该标签代表的基因的表达程度。

二、SAGE技术方案简介

SAGE技术方案主要包括三个阶段：

(一) SAGE文库的构建

1. 双链cDNA的获得　从细胞或组织中抽提mRNA，用5'端有生物素修饰的Oligo d(T)为引物，通过mRNA反转录生成双链cDNA。

2. 生物素化cDNA 3'末端片段的获得　用锚定酶(anchoring enzyme)NlaⅢ酶切上述双链cDNA。NlaⅢ能够识别CATG位点，切割后的产物用链霉亲和素包被(streptavidin-coated)的磁珠进行亲和纯化，获得生物素化cDNA的3'末端片段。

3. 带接头的SAGE标签的获得　分离得到的cDNA的5'端被补平，并分成两部分，分别接上接头(linker)A或B。该接头包含有限制性内切酶BsmFI的识别序列和一个PCR引物的序列(引物A或B)。

4. 用标签酶(tagging enzyme)　BsmFI切割上述连有接头的cDNA BsmFI在其识别位点3'端下游的14~15bp处进行酶切，从而每条cDNA释放出一个带有接头的SAGE标签(tag)。BsmFI酶切的位置，20%在其识别序列的下游14bp处，80%在其识别序列的下游15bp处。

5. SAGE双标签的获得　将带有接头A、B的两种SAGE标签，分别用DNA聚合酶(Klenow酶)进行末端补平、混合，并用连接酶(ligase)连接，得到带接头的双标签(lingker-adapted ditag)。用引物A、B进行PCR扩增。再用NlaⅢ酶切，PAGE胶分离，得到纯化的带有CATG突出末端的SAGE双标签(SAGE ditag)。

6. SAGE双标签多聚体的获得　用T4 DNA连接酶连接SAGE双标签，获得双标签形成的多聚体(concatemer)。

(二) 多聚体分子的克隆与测序

选择合适长度的多聚体片段，克隆至高拷贝的克隆载体。这样，得到的克隆其插入序列是由一系列20~22bp长的SAGE双标签组成，每两个双标签中间由4bp的NlaⅢ酶切位点CATG分隔开。利用质粒载体上的通用引物，对插入片段进行单向测序。SAGE要求高质量而且长度长的序列。

(三) 标签序列的提取

标签序列的提取可以使用SAGEnet提供的SAGE提取软件包(可以从http://www.sagenet.org/上免费下载)来完成，也可以使用NCBI提供的高度用户化的unix操作系统和C程序(网址是：http://www.ncbi.nlm.nih.gov/sage)来完成。两者的基本处理过程相同，包括下述步骤：

1. 在双标签多聚体序列中定位NlaⅢ酶切位点(即CATG)；

2. 提取 CATG 位点之间的 20~26 个碱基长的双标签序列；

3. 去除重复出现的双标签序列，包括在反向互补方向上重复的双标签序列；

4. 截取每个双标签序列最靠近两头末端的 10 碱基，即为标签序列；

5. 去除接头序列(即 TCCCCGTACA 和 TCCCTATTAA)，同时去除含有不确定碱基(即除 A、C、T、G 四种碱基以外的碱基)的标签；

6. 计算每个标签的出现次数，以列表的形式给出一个包含每个标签及其表达丰度的报告。

三、SAGE 技术的缺陷与改进

(一) 样本用量问题

常规 SAGE 需要的样本量约为 50~100ng mRNA 或 5~50μg 总 RNA，这显然比 cDNA 文库用量少得多。这个量需要从 2×10^5~2×10^6 个细胞中提取，这对于来源有限的某些特殊组织(骨髓间质干细胞、视网膜色素上皮细胞)来说，仍然是一个问题。

1999 年，Datson 对 SAGE 进行改进，建立了 MicroSAGE 方法。该方法使用的组织量为常规 SAGE 的 1/5000~1/500。而且，从 RNA 的提取到获得标签的这一系列操作均在同一试管内进行，且在操作过程中不使用苯酚 - 氯仿抽提或乙醇沉淀，避免了更换试管与一系列纯化过程造成的样本丢失，因而只需 10^5 个细胞就可以满足实验的需要。MicroSAGE 技术的问世，使得对于诸如骨髓间质干细胞的分化研究、视网膜周围细胞与黄斑区细胞、视网膜色素上皮细胞基因表达谱研究得以实现。

2000 年，Ye 等建立了 miniSAGE 技术，仅需使用 1μg 的总 RNA。该技术进行了三方面改进：①使用了相位锁定(phase lock)凝胶，增加了每次苯酚抽提后 DNA 的回收率和纯度；②减少了连接子的用量，因而减少了连接子在标签二聚体扩增时的干扰，增加了标签二聚体的产量；③使用了 mRNA capture 试剂盒，使 mRNA 提取、cDNA 合成、锚定酶切 cDNA、生物素化的 cDNA 与链亲和素的连接、连接子与 cDNA 结合及 cDNA 标签的释放均在同一试管内进行，大大地减少了样本量的丢失。此方法有利于进一步拓展 SAGE 的应用。

·Johns Hopkins 肿瘤中心与 Howard Hughes 医学研究所将 microSAGE 技术的详细方案在 http://www.sagenet.org/protocol 上公布，推荐使用的细胞数为 5×10^4~2×10^6，以确保实验的成功。

(二) 标签确认困难问题

应用 SAGE 研究，获得的标签有三种情况：①唯一匹配，可以检索到唯一与之匹配的基因或 EST，这些标签一般不须进行其他检测；②多重匹配，部分标签在现有数据中检索到多个与之匹配的表达序列。这些表达序列除了具有相同的 SAGE 标签以外，无其他相似性，也就是说对于某一特定的组织，根据这些 SAGE 标签很难确定哪个基因为该组织的表达基因，需另行相关检测；③未知基因标签，现有数据库中无匹配基因 EST，可能为潜在的新基因，但必须使用相应手段对其进行鉴定。后二者是 SAGE 数据处理与分析过程中面临的两大难题。

Chen 等建立了 longSAGE 技术 GLGI(Generation of longer cDNA fragments from serial analysis of gene expression (SAGE) tags for gene dentification, GLGI)，试图解决 SAGE 标签的识别难题。该技术在一定程度上弥补了 SAGE 在标签确认中的缺陷，特别有助于得到未知表达基因的 3' 端序列。

得到长片段的标签，增大标签的特异性，似乎成为减少多重匹配与鉴定未知标签的好办法。随着标签长度的增大，标签的特异性增高，当使用 21bp 的标签时，其特异性达到 99.83%。但 longSAGE 显然要增加测序工作量，只有在拥有大规模测序、分析的实验室内才可以进行。要得到合适长度的标签，选用适当的锚定酶与标签酶是这项技术改进的重要环节。

在 Ryo 等进行的 longSAGE 研究中，使用 *Rsa*Ⅰ(识别位点为 GATC)作为锚定酶，*Bsm*FⅠ为标签酶，得到 18bp 的 SAGE 标签。这种标签特异性较强，且能较好地用于未知标签的基因检测。Saha 等介绍了能获得 21bp 长的 SAGE 标签 longSAGE 技术，21bp 中含有 17bp 的独特片段、4bp 的

ⅡS 型限制性内切酶 *Mme*Ⅰ酶切位点，并对相应的 SAGE 步骤进行改良，在预测的 16 000 个已知基因中，75% 以上能得到一个独特的 21bp 的标签，而常规 SAGE 得到的 14bp 标签唯一匹配率仅在 1/3 左右。

未知标签的鉴定，需进行 RT-PCR、以标签为引物的快速 cDNA 末端扩增法(RACE)、以标签作为寡核苷酸探针杂交或计算机预测附近区域的外显子等方法来实现。总之，Long-SAGE 标签在鉴别新基因方面具有明显的优势，标签的多重匹配现象也明显降低，是 SAGE 技术演变后较为完善的研究手段。

(三) 实验误差问题

任何实验都可能存在实验误差。但在 SAGE 技术中，获得的是一段序列的序列标签，依据这种标签寻找匹配基因的时候，需要特别注意实验误差带来的影响。在 SAGE 操作过程中，多个环节都可能带来实验误差。

理论上，能识别 4bp 的限制性内切酶均可以作为 SAGE 的锚定酶，以获得 3' 末端的序列。4bp 锚定酶在平均 256bp 范围内会出现一个酶切位点。但实际操作中，仍有部分基因不具有酶切位点，也就是基因出现锚定酶漏切的现象。究竟有多少基因被漏切无法预测。克服基因被漏切的解决办法是，使用另外一种不同的 4bp 限制性内切酶对 ds-cDNA 进行酶切。使用双锚定酶几乎可以使每一基因均可被切割到，但是无疑增加了实验操作的复杂性，样本用量也会大大增多。目前 SAGE 技术使用的标签酶基本上是 *Bsm*FⅠ，该酶的识别区域为 GGGAC，大约每隔 500bp 区域有一个酶切位点，但同样存在漏切问题。因此，使用两组不同的锚定酶与标签酶，构建 2 个独立的 SAGE 标签库，应该是 SAGE 研究的理想办法，但是，这样实验工作量无疑加大一倍。

另外，*Bsm*FⅠ并不是将连接序列精确地切成 14bp 的片段，常出现较长或较短的片段，一般为 12~16bp。不同长度的标签带来的问题是：无法确定构建出来的标签二聚体是(12+16)bp、(13+15)bp 还是(14+14)bp 构成方式。为了避免这种情况的出现，需要研究者们进一步探索出某种条件下精确切割标签成单一长度的ⅡS 型限制性内切酶。

DNA 测序误差是目前测序分析中出现的不可避免的现象，一次性地测序分析，单一碱基测序错误率约为 1%，按每个标签 10bp 计算，标签的错误率可达 10%。即使双向测序或多次测序可以减少这种测序的错误率，但实际应用中，标签测序分析往往是一次完成。目前还没有哪个实验室对 SAGE 标签实施多次测序的报道。

鉴于上述种种可能出现的实验误差，在确认标签对应的基因时，对结果的解释与判断需要小心，必要时需要实验来验证。

四、SAGE 技术的应用前景

前面提到过，从 EST 出发，可以发现新基因。但 EST 技术难以获得低拷贝的基因。在真核生物中，高丰度表达的基因只是少数，但它们却大大干扰了对于大多数低丰度基因的分析。SAGE 技术克服了转录本丰度的影响，在新基因分析中具有独特的地位。许多研究提供了大量找不到对应表达序列的“无匹配标签”(unmatched tags)，它们绝大多数是低拷贝标签。这些无匹配标签很可能来源于一些尚未鉴定的新转录本。这意味着还有许多新基因尚待发现和鉴定。Chen 等建立的 GLGI 技术，从 SAGE 生成较长的 cDNA 片段，然后与 GenBank 中的数据比对，从而发现和鉴定新基因。同时，Saha 等建立的 long-SAGE 技术，生成 21bp 长的 SAGE 标签，直接与人类基因组计划的数据库比对，一次发现了大量新基因。这些方法的建立，大大加速了对全基因组的诠释。

人类基因组计划的最终目标是绘制出一张完整的基因图，1998 年 Deloukas 等基于辐射杂交作图原理构建了 GeneMap'99。此图为 45 059 个 EST 和基因作了染色体定位，且将这些 EST 和基因归类为 24 106 个 UniGene 簇(UniGene cluster)，每一 UniGene 簇包含着一群序列相同或相近的基因或 EST。但 GeneMap'99 并不能提供给有关基因表达方面的信息。但如果将其与 SAGEmap 所提供

的基因表达定量和定性资料相结合，则能得到人类的全基因组转录图谱。基于这一目的，Caron 等于 2001 年将来自于公共 SAGE 数据库的 231 万个标签和他们自己从神经母细胞瘤分离得到的 16 万个标签信息按照组织来源分成 12 组，再依据一定运算法则将这 12 组 SAGE 标签与 UniGene 簇中已做染色体定位的基因进行对应，得到了一张可大致反映基因表达丰度的全基因组转录图谱。虽然 Caron 等的全基因组转录图谱还很不完整，但却显现出重要意义。例如，发现染色体上存在高表达的基因簇区域(region of increased gene expression，RIDGE)，这提示基因组的结构是高度有序的。RIDGE 具有较高的基因密度，在 1cR(centiary)的染色体长度上含有 6~30 个基因，而在低转录区域仅有 1~2 个基因。位于 RIDGE 中的基因的转录水平要比染色体的平均水平高 7 倍，是低水平转录区域基因密度的 20~200 倍，这是人们事先不曾预想到的。

SAGE 技术是描述表达谱的强有力工具。研究者们利用 SAGE 技术分析了多种细胞、组织的转录谱特征。全面分析比较正常和疾病状态下各种细胞、组织或器官的表达谱，不仅可以发现细胞、组织或器官功能的维持机制，也有可能解释参与疾病的发生、发展的特异性基因、信号途径或疾病治疗的新靶点。采用 SAGE 技术，分析胚胎细胞或干细胞中的大量尚未鉴定的极低拷贝基因的表达谱，不仅有助于了解参与分化、增殖的基因，了解分化、发育的分子基础，而且可能为人们获得基因组完整的图谱提供不可忽略的基因来源。

多种免疫细胞(单核细胞、巨噬细胞、树突状细胞)表达谱的获得，将有助于揭示免疫细胞发挥各种免疫功能如抗原提呈、炎症反应等的分子基础，深化人们对免疫调控复杂机制的了解，同时也为疾病的诊断、治疗奠定理论基础。

SAGE 技术对于寻找肿瘤特异性相关基因、发现肿瘤组织特异标志物、全面分析肿瘤组织基因表达谱和揭示肿瘤发生的分子机制等方面都有着重要意义。近年来，利用 SAGE 研究各种肿瘤，取得了很大的进展。随着肿瘤基因组解剖计划(cancer genome anatomy project，CGAP)的进行，有 5 百多万转录物标签从 100 多种人类细胞中获得。Porter 等为了确定乳腺肿瘤发生和进展过程中分子的变化，利用 SAGE 技术分析了正常乳腺上皮细胞、原位癌、浸润癌和转移癌基因表达的全部轮廓。通过配对比较和分层聚类分析 8 个 SAGE 库，他们得出下面的结论：许多在正常乳腺上皮细胞中高表达的基因，在肿瘤细胞中却没有表达。这些基因主要是编码分泌、胞质分裂和化学激动作用蛋白质的基因。这就提示异常的旁分泌和自分泌在肿瘤的发生和发展中可能起重要的作用。他们发现在正常组织向原位癌转变的过程中涉及许多编码分泌、细胞非自控因子的基因，从而提示正常组织向原位癌转变的过程有可能是最有希望的肿瘤预防和治疗的突破口。随着 SAGE 技术被广泛用于分析正常组织和癌症组织基因表达谱，CGAP 计划建立了肿瘤 SAGE 数据专用数据库(http://cgap.nci.nih.gov/SAGE)，为将 SAGE 技术深入、广泛地用于研究肿瘤提供了方便。

以往的研究注重于某个基因对细胞及机体的影响，如往往通过基因敲除(knockout)或基因敲入(knockin)来认识某个基因的功能，这些方法对于科学的贡献是毋庸置疑的。但是，基因的功能表现往往又是许许多多基因共同参与的网络信息调控的结果，人们过去的观察很可能是“瞎子摸象”所得到的结果，而不清楚这个基因给整个基因网络所带来的改变。因此，过去人类对于许多基因功能的了解可能仅仅是某个侧面。在功能基因组时代、系统生物学到来的时代，人们需要对基因的功能进行全面的了解和重新注释，但这需要有相应的技术作为支撑。SAGE 技术作为一种大规模的、高通量的定性与定量的研究技术，它可以被认为是系统生物学时代的标志性技术之一。

五、SAGE 数据库和分析软件

(一) NCBI SAGE 数据库

1. GEO 数据库简介　近年来，利用高通量杂交阵列和基于测序技术的分子生物学实验已非常普及，这些技术要么被单一使用，要么被联合使用来评估大量 mRNA 和基因组 DNA 分子的信息。促成这种普及的主要因素是这些技术的平行化、高通量特性及其伴随在时间上的高度保守性，即在

极为相似的条件下同时(或者几乎同时)进行大量的分子样品实验获得信息资源。当研究成果在科学文献上发表后,通过公共的高通量数据库,可以实现对相关数据的进一步挖掘。因此,建立高通量数据的公共数据库平台是非常必要的。

GEO 数据库(http://www.ncbi.nlm.nih.gov/geo/)是最早建立的基因表达数据公共贮存库之一,于2000 年 7 月在 NCBI 上首次发布。GEO 存储大量的基因表达 / 分子丰度信息,支持符合 MIAME(minimum information about a microarray experiment)标准的数据提交,是目前最重要的对基因表达数据进行浏览、查询和检索的在线资源。GEO 可接受和存贮广泛的高通量试验数据,包括单通道和双通道的微阵列实验(用于测量 mRNA、基因组 DNA 和蛋白质丰度)、非阵列技术如基因表达系列分析(SAGE)、蛋白组学质谱分析数据等。NCBI 原有的单独存储的 SAGE 数据已经被合并到了 GEO 里面。

截至2010年2月2日,GEO中已收录了6945个GPL(GEO platforms)数据,396 729个GSM(GEO samples)数据,15 474 个 GSE(GEO series)数据。

一个 GEO 仓库包括四个基本实体:提交者(Submitter)、平台(Platform)、系列(Series)和样本(Sample)(图 6-15)。

(1) 提交者:数据提交者的联系和登录信息,与许多平台、许多样本和许多系列有关。

(2) 平台:关于用于以高通量方式检查样本的物理试剂的信息,与一个提交者、许多样本有关。

(3) 样本:关于被检查的 mRNA 样本,实验条件,和实验产生的基因表达测量数据信息,与一个提交者、一个平台和许多系列有关。

(4) 系列:样本收集,样本是如何相关的,如何排序的,分析是如何进行的,和聚类数据是如何获得的信息,与一个提交者、许多样本有关。

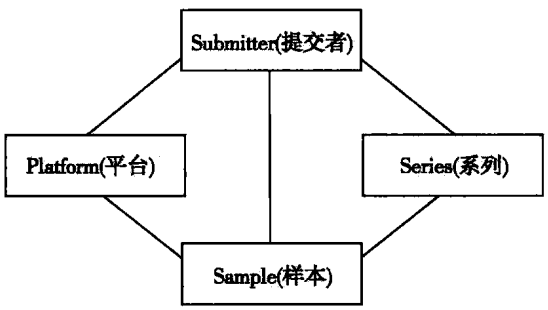


图 6-15 GEO 数据库的四个基本实体

为了能达到一个开放灵活的设计理念,以方便用户储存和检索不同类型的数据,GEO 中的数据并不完全压缩集中在同一数据库中。相反,而是用一种图表分隔的 ASCII 表格形式,来对每一个平台和样本的数据进行保存。这种表格包含有多项专栏,并在表格的上面伴有专栏名称。当前摘录表格中的数据主要是为了索引,但是,为了方便用户更广泛地搜索和检索,这些数据还可被进一步、更深一层次地摘录和提炼。另外,数据投放者自己也添加了一些专栏,以用于贮存附加的、被他们定义的相关信息。从本质上说,平台是描述一连串在特定实验中被检测或被定量分析的因素。比如寡核苷酸探针组、cDNA、SAGE 标签、抗体等。平台登录号的首字母为“GPL”。样本是指以一个平台为基础、描述某个杂交实验或者实验条件的所有特征因素的大量测量信息。每个样品有一个,而且只有一个必须先前被确定的亲代平台。样本登录号的首字母为“GSM”。系列是把构成某个实验的相关样本集中到一个有生物意义的数据集,同时可能还收集一些已被递呈者注明的重要基因或者分析结果纲要。一个系列中的样品是通过某一共同的属性联结在一起的,系列登录号的首字母为“GSE”。GEO 平台和样本的数据格式不像元数据格式那样被保存在一个指定的数据库格式字段区域内,也不是完全的高度集中,而是以文本的形式保存。这种设计理念能使 GEO 保持适应不断发展的技术趋势,同时也允许在被保存数据的数量和类型方面达到最佳。

为了能够有效地检索、显示和分析数据,GEO 中开发了一些新的工具和特性。为了创建这些工具,GEO 数据首先被组合为可比较的集合,或称为 GEO 数据集(GEO datasets, GDS)。NCBI 中建立了两个新的数据库来查询这些数据,均收录在 Entrez 综合数据库检索系统中,这两个数据库分别为 GEO 表达谱(GEO profiles)和 GDS。GDS(数据编号格式为 GDSxxx, x 表示阿拉伯数字)是当前的 GEO 样品数据库。GDS 内的样品代表相同的平台,也就是它们分享一组共同的探针。GEO 表达谱

数据库贮存个体基因表达和由基因表达库组成的分子丰度图，可以通过基因注解或预处理的图谱特征寻找感兴趣的特殊图谱。

2. GEO 数据库查询　GEO 数据可以使用 Entrez GEO 数据集和 Entrez GEO 表达谱进行查询。Entrez GEO 数据集查询所有的实验注解，查询主页面网址为：http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds。和其他的 NCBI Entrez 数据库查询方式一样，可以使用布尔短语并限定特征字段来进行查询。对于感兴趣的实验，通过属性限定可提高查询效率，如基因名、GEO 登录号、关键词、变异性、组织、创建日期和平台等。例如，使用检索词 “dual channel[Experiment Type] AND metastasis AND human[Organism]” 寻找人类新陈代谢的所有的双通道核苷酸微点阵实验数据组。检索结果信息显示了数据组标题、简短实验说明、分类法、实验变量类型和原始平台的链接、相关系列记录和完整 GDS 记录。一旦确定相关数据集，可进一步研究感兴趣基因的表达图谱。Entrez GEO 表达谱可以查询预处理的基因表达 / 分子丰度图谱，即样品和系列记录，其查询页面网址为 http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo。查询可以使用属性限定，如关键词、平台和样品类型、提交者、组织、发表日期和补充文档类型等。例如，利用检索词 “Type 1 diabetes[GDS Text] AND apolipoprotein[Gene Description] NOT Homo sapiens[Organism]”，检索到所有在非人类的物种中Ⅰ型糖尿病相关数据集中的载脂蛋白相关的基因资料。检索结果中每个表达谱数据显示报告人提交的注解、简短实验信息、分类法以及该图谱的条形索引图。这个索引图对于快速、大量的文档扫描和比较非常有用，单击索引图像可显示图谱的详细内容。因为样品通常组合为系列内有意义的数据组，所以对一个系列及其相关样品和平台的检索更具有说明性。

除了上述在 Entrez 系统中检索数据的方式，还可以在 GEO 数据库主页面中（http://www.ncbi.nlm.nih.gov/geo/）直接检索数据（图 6-16）。GEO 主页面的向导图（GEO navigation）显示了三项基本

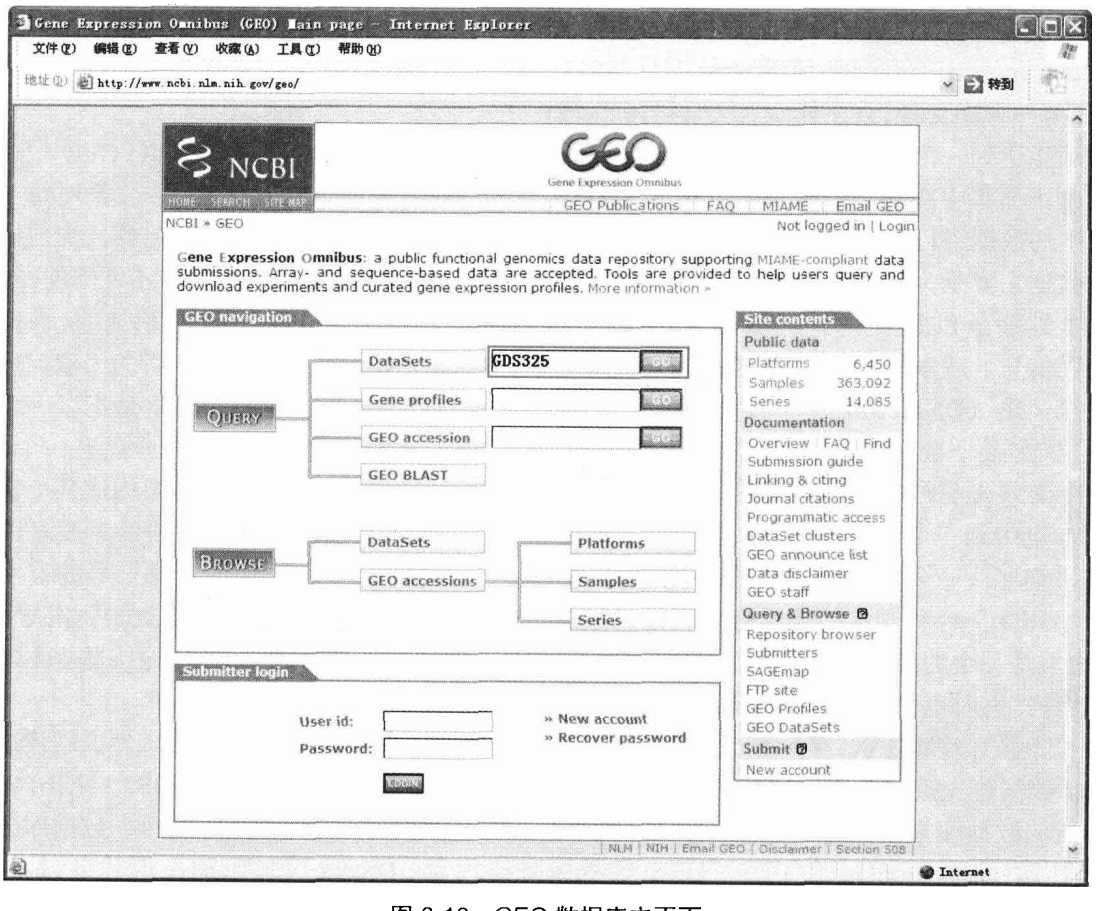


图 6-16　GEO 数据库主页面

功能：查询(query)、浏览(browse)和数据提交(submit)。查询功能提供四类查询方式：Datasets 查询栏可以查询感兴趣的系列、数据集或平台数据；Profiles 查询栏可以定位和观察感兴趣的基因表达谱数据；GEO accession 查询栏可以输入一个特定的数据记录号来查询对应的数据记录，如 GPLxxx、GSExxx、GSMxxx、GDSxxx；GEO BLAST 查询项可以通过 BLAST 搜索核酸序列的相似序列表达谱数据。浏览功能项则为用户提供浏览数据集、平台、样本、系列等数据的入口。

在 GEO 中检索某一感兴趣的数据系列的例子如下：在“Query”项的 GDS 检索框输入数据编号，如“GDS325”，单击“GO”按钮进行搜索。搜索结果显示为检索到的 GDS 编号及相关的平台与样品，单击“GDS325 record”，进入显示该数据信息的页面(图 6-17)。

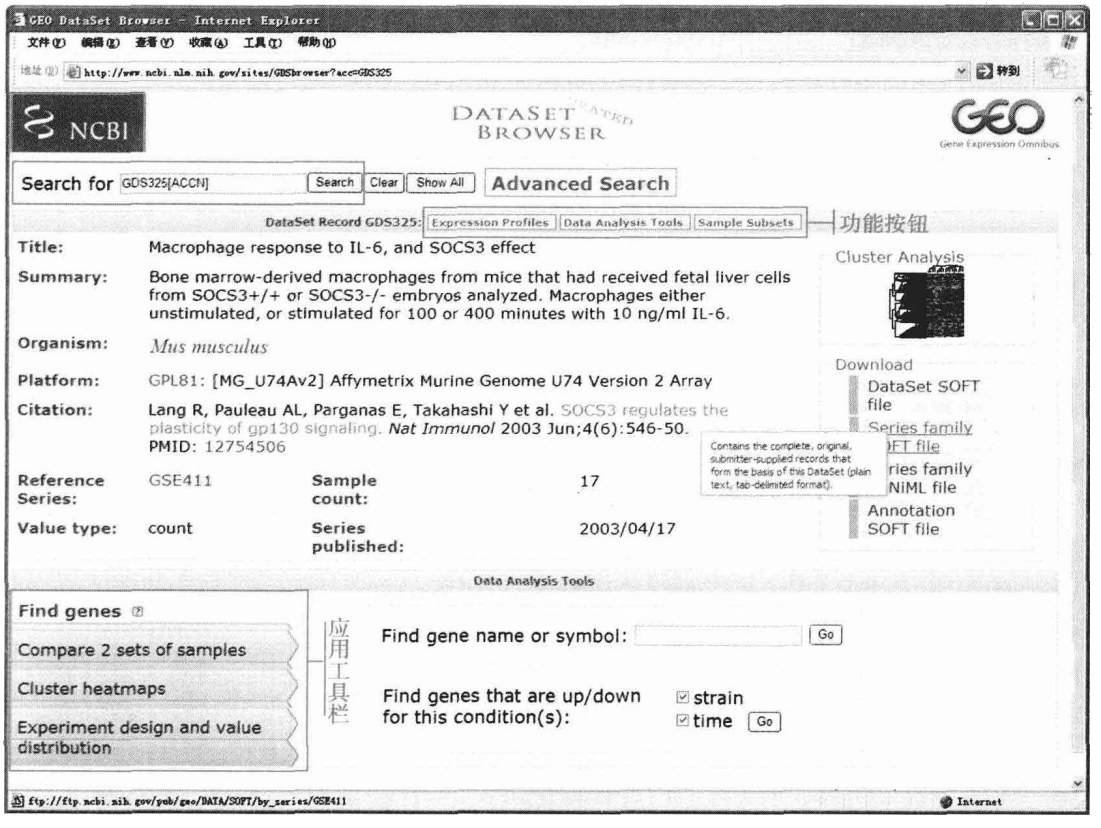


图 6-17 GEO 中“GDS325”的数据记录页面

该页面显示这一数据记录的名称为“Dataset Record GDS325”，其数据信息包括“Title”、“Summary”、“Organism”、“Platform”等项，其中“Title”项是该数据的标题，“Summary”项是对该数据的简要描述；“Organism”项是该数据的来源物种；“Platform”是对应的平台数据。此外，还记录了引用文献、参考系列、样本统计等相关信息。数据的标题栏中给出了三个链接，点击“Expression Profiles”链接，则打开一个新页面显示与 GDS325 数据集相关的表达谱数据；点击“Data Analysis Tools”链接，则在页面下方显示数据分析工具栏；点击“Sample subsets”，在页面下方显示相关的样本子集。页面右侧还显示了“Cluster Analysis”和“Download”的相关信息，其中点击“Cluster Analysis”可以在新页面中显示在数据库中该数据的聚类分析结果；而“Download”下提供了数据集 SOFT 格式文件、系列家族 SOFT 格式文件、系列家族 MINiML 格式文件以及注释信息的 SOFT 格式文件的下载，用户可以根据自己的需要来下载不同的数据文件。

3. GEO 数据分析 在 Entrez GEO 表达谱搜索得到的结果中，通过每一条数据记录旁的“Profile neighbors”、“Chromosome Neighbors”、“Sequence neighbors”、“Homologs”、“Links”等工具，可以找到感兴趣的相关数据。“Profile neighbors”检索与原基因相似类型数据组的其他基因 / 分子，由此可

以推断某些普通功能元件或调控元件。“Chromosome Neighbors”显示与原基因来自同一染色体的基因或分子数据；“Sequence neighbors”基于核苷酸序列相似性在所有 GEO 数据库寻找相关基因，因此可以用于鉴别同源序列如基因家族，或用于物种间对照。“Links”可以通过 GEO 数据库链接到其他 Entrez 数据库的相关记录，包括 GenBank、PubMed、Gene、UniGene、OMIM、HomologGene、Taxonomy、SAGEMap 以及 MapViewer。

除了 Entrez 查询系统以外，GEO 还提供了几个辅助工具来协助增强对数据的挖掘和可视化等分析。例如，在每一条数据记录的主页面的“Data Analysis Tools”标签中，就提供了多种数据分析工具，可以方便地对该数据进行分析(图 6-17)。

(1) Find genes 工具：基因发现工具。这一工具提供给用户快速寻找指定基因的功能。可以通过点击数据记录主页面的“Data Analysis Tools”标签，在页面下方的标签栏内选择“Find genes”，然后在搜索框内输入基因名称或符号，设定搜索条件，点击按钮“GO”进行搜索。

(2) Cluster heatmap 工具：聚类图分析工具。大多数据集都提供了样本和基因等级聚类图，用户可以选择浏览这些聚类图，并选择感兴趣的多聚类部分，然后进行放大、下载、制作线性图表或直接链接到 Entrez GEO Profile 记录。GEO 提供 9 种预处理的分层聚类类型和用户指定的 K 均值和 K 中值聚类。以 GDS325 为例，在该数据记录的右侧点击“Cluster Analysis”标签，就可以打开聚类图的显示页面(或者在该数据记录主页面的“Data Analysis Tools”标签中选择“Cluster heatmaps”)。在聚类图页面中(图 6-18)，改变“Display Options”栏内的颜色选项可以改变不同表达水平的显示颜色；而“Clustering”一栏则提供了对聚类参数的调节选择，用户可以通过改变“Distance”和“Hierarchical”的计算方式来获得所需要的聚类结果。使用鼠标双击聚类图的相关区域，可以将图放大查看。通过页面上对应的按钮可以进行数据下载、绘制选择基因的表达谱、在 Entrez GEO 中搜索表达谱数据以及存储自己的选择等操作。

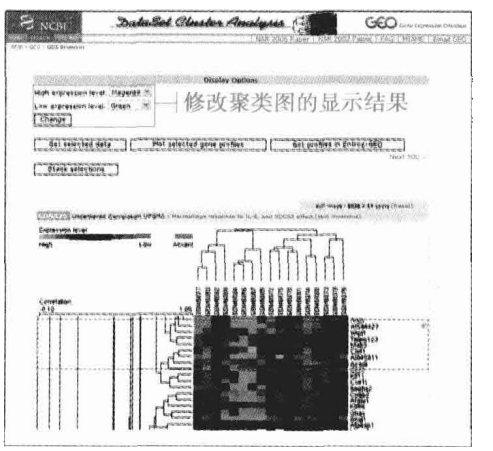


图 6-18 GEO 中“GDS325”的聚类分析图

(3) Query Group A versus B 工具：两个子集比较下的查询工具。这个功能的特性是通过计算一个数据集内、不同实验子集间的平均秩次或值的差别，来鉴别感兴趣的基因表达谱。要使用这一工具，可以通过点击数据记录主页面的“Data Analysis Tools”标签，在页面下方的标签栏内点击“Compare 2 sets of samples”来进行。使用步骤分为三步：第一步，选择试验水平参数；第二步，选择 A 子集和 B 子集中的样本，通过鼠标即可选择 Group A 和 Group B 中需要比较的样本；第三步，点击“Query Group A vs. B”来进行查询。

(4) Experiment design and Value distribution：实验设计和数据分布查看工具。一个数据集中的每个样本均会有对应的数据图，可以大概了解一个数据集的数值分布状态。用户可以在“Data Analysis Tools”标签中的“Experiment design and value distribution”一栏内，看到一个数据缩略图，点击该缩略图可以打开新的页面，其中会显示该图的详情。

(5) GEO BLAST：GEO 序列比对工具，该工具提供给用户使用 BLAST 来搜索感兴趣的核苷酸序列的相似序列的 GEO 基因表达谱。GEO BLAST 数据库包含了所有 GenBank 中的序列，而且，使用 NCBI BLAST 输出标准的 BLAST 比对结果，并且在适当的位置显示“E”(GEO)图标链接，点击“E”图标即可直接链接到 GEO Profiles 数据。要使用该功能，可以在 GEO 数据库主页面的向导图中点击“GEO BLAST”按钮，进入 BLAST 搜索页面，输入查询序列、设置搜索参数，然后进行搜索。

(6) Subset effects：子集效应：如果不同子集间的基因表达值或秩次存在显著性差异，那么这

些表达谱就会被自动标记。通过这个特性可以检索到所有相关的表达谱。换句话说，对于某个特别的实验变量，如“年龄”或“血型”，一旦出现有意义的表达谱，该表达谱即会被标记。例如，查询GDS186或者GDS187的表达谱及子集效应，可以在Entrez GEO Profiles的搜索栏内输入“(GDS186 OR GDS187) AND "value subset effect"[Flag Type]”来进行查找。

4. 数据提交与更新　GEO提供了一个基本的结构，以方便数据投放者能以MIAME兼容的形式提交他们的数据。提交到GEO的数据会遵循一些被限定的条件并服从一些基本规则，而在数据的构成形式上得到确认，以确保所得到的记录包含一定意义的信息，并且能被正确地归纳和组织。数据的提交者保持对数据进行管理和编辑的权限，同时也要对他们记录的内容和质量负责，这些记录的概要将会在微阵列基因表达数据协会委员会(microarray gene expression data society，MGED)以公开信件的形式发布。很明显，GEO不进行单独实验或分析，提交数据的可靠性、价值、质量或生物学意义依赖于数据投放者。一旦数据投放者建立了他们自己私人的GEO账号，他们将有三条途径来储存他们的数据：①交互式网络格式。对每个平台和样本的投稿，均会有一个文本和图表分隔的数据表格文件被上传和验证。这个程序在提交相对少的数据时非常直接和实用。也可以用相同的交互式网络格式对单个数据记录进行更新。②直接用单一的综合性文本格式(simple omnibus format)即SOFT的格式提交。SOFT是专门为快速批量提交数据而设计，这样的文件很容易从普通的表格程序和数据库应用软件生成。单一的SOFT文件可同时包括多平台、样本和系列的数据，且能被直接上传到数据库，批量更新也可以用SOFT格式快速完成。关于SOFT格式的详细信息可在GEO网站获得(http://www.ncbi.nlm.nih.gov/geo/info/soft2.html)。③数据投放者还可以用有效的微阵列基因表达标志语言(microarray and gene expression markup language，MAGE-ML)格式，以FTP的形式把文件上传到GEO。

向GEO中提交数据的基本步骤如下：

(1) 创建GEO账号：用户可以在GEO数据库主页面的导航图的“Submit”一栏中找到“Create a new account”按钮，点击该按钮进入账号申请页面，正确填写相关信息即可创建自己的GEO账号。

(2) 选择提交方式：在“Submit”一栏中提供了“Direct Deposit/Update”和“Web Deposit/Update”两种提交方式。“Direct Deposit/Update”允许用户批量提交多个数据记录，支持SOFT、MINiML、GEOarchive、SOFTmatrixz等格式的文件；同时还可以批量更新已经存在的数据记录。提交过程非常直观易懂。“Web Deposit/Update”允许用户通过简单的、分步的网络交互方式提交单个数据记录，或者更新已经存在的单个数据。

(3) 准备数据，执行提交：以“ect Deposit/Update”方式为例，用户将数据按照被接受的格式编辑成数据文件，然后在“Direct Deposit/Update”的主页面中选定文件格式，指定上传文件的路径，指定是提交新数据还是更新数据，选择发布数据的日期(指定提交日之后的某个日期发布数据，或者选择立即发布)，填写其他相关项信息，点击“Submit”按钮即可自动提交数据。提交的数据将被GEO服务器审核，通过审核的数据才能被接受发布。

(二) SAGEnet

SAGEnet是一个收录SAGE技术方法、文档、资讯以及SAGE实验数据的网络资源库(图6-19)，访问地址为http://www.sagenet.org/。该网站由约翰霍普金斯大学医学院建立和维护，主要提供下列信息和数据。

FINDNGS：这一项主要介绍SAGE技术的基本原理，以及SAGE技术的一些应用。

RESOURCES：这一项包括网站资源列表以及一些相关网站的链接。资源列表中列出了三类资源：SAGE数据；基因图谱的SAGE标签；SAGE实验方案和软件。SAGE数据中收录了来自于人类结肠癌组织以及胰腺癌组织的SAGE数据，此外，还收集了小鼠增生细胞以及酵母的SAGE数据；基因图谱的SAGE标签收集了人、小鼠、大鼠的一些数据；SAGE实验方案和软件中提供了一些SAGE实验方案的技术文档，以及关于如何获得SAGE软件的说明。

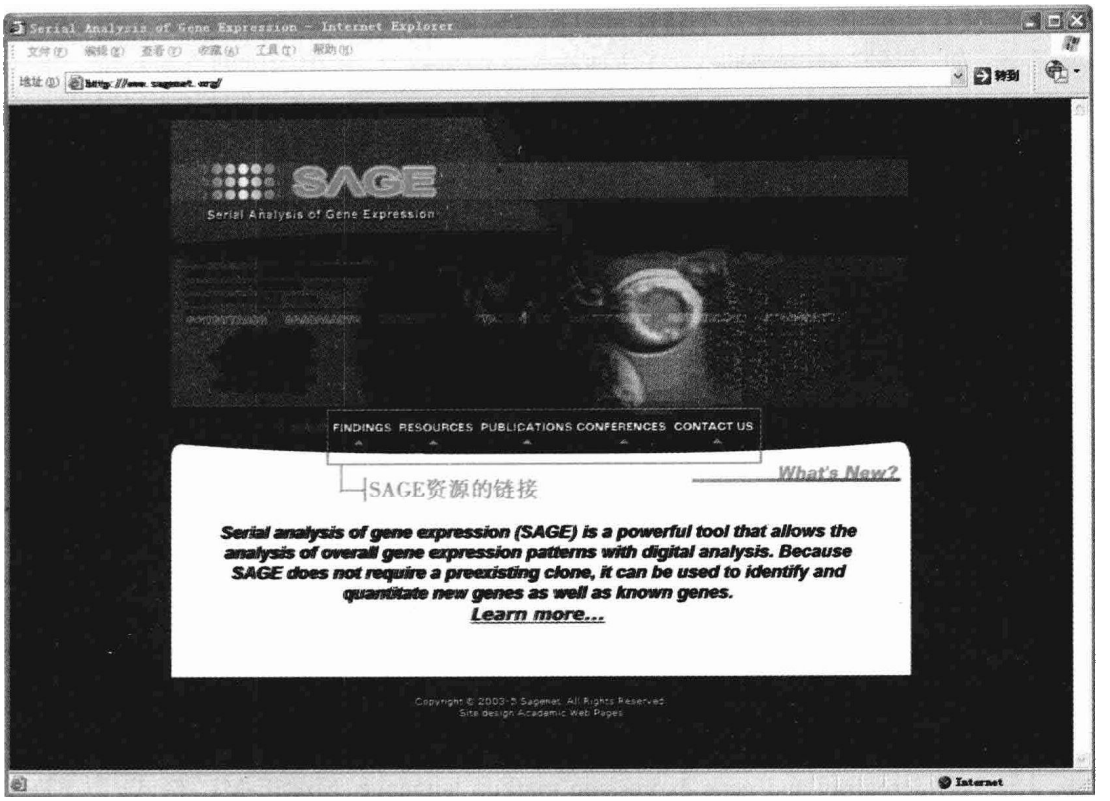


图 6-19 SAGEnet 网络资源库的主页面

PUBLICATIONS：提供了 SAGE 技术的原理和应用的一些研究文献的列表。

CONFERENCES：提供 SAGE 技术相关学术会议的信息。

CONTACS US：官方的联系方式，学术研究用户可以与之联系免费获取 SAGE 相关的实验方案和软件，但是不能用于商业用途。

（三）SAGE Genie

SAGE Genie 数据库是美国癌症基因组分析计划的一部分，由 NCBI 建立和维护，访问地址为：http://cgap.nci.nih.gov/SAGE。该数据库将人和小鼠已知基因的长度为 10~17 个核苷酸的 SAGE 标签进行严格分析，提供了这些基因的高度直观和可视化的表达信息。近年来 SAGE Genie 还合并了与 SNP 关联的可选择标签的参考数据库。SAGE Genie 数据库在以前的 SAGEmap 数据库的基础上，提供了更为友好的界面；通过查询 SAGE Genie 数据库，可以获知基因在正常组织和癌变组织中的相对表达量。

此外，SAGE Genie 还提供了一系列工具，用于方便地查看和分析相关数据。例如，在人类 SAGE Genie 工具集中，SAGE Anatomic Viewer 可以查看给定基因在人的正常组织和癌变组织中的相对表达量；SAGE Digital Gene Expression Displayer 可以用于研究两个 SAGE 文库之间的基因表达谱的显著性差异；SAGE Genie Downloads 可以 FTP 的形式查看和下载人类基因、标签、数据集和图谱等数据文件；SAGE Tag Extraction 可用于在用户上传的序列文件中提取 10bp 或 17bp 的 SAGE 标签。在小鼠 SAGE Genie 工具集中，mSAGE Expression Matrix 可以查看某一基因在小鼠不同发育阶段的相对表达量；mSAGE Digital Gene Expression Displayer 可用于研究两个小鼠的 SAGE 文库中，基因表达谱差异的显著性；mSAGE Genie Downloads 可通过 FTP 来查看和下载小鼠的基因、标签、数据集等数据文件等。

（四）小鼠 SGAE 数据库

小鼠 SGAE 数据库(The mouse SAGE Sit，http://mouse.img.cas.cz/sage/)是由捷克科学院分子遗

传研究所构建的一个数据库，收录了各种小鼠组织和细胞系的公开的 SAGE 文库。该数据库将多个小鼠 SAGE 数据库组合在一起，并开发了基于 WEB 的工具提供给用户浏览数据。用户除了浏览和查询数据以外，还可以利用文库比较工具，通过设定一定的参数来比较两个 SAGE 池中的数据；还可以使用 SAGE 标签鉴别工具来对提交的标签序列进行鉴定等。

（五）其他 SAGE 数据库

一些 SAGE 研究项目构建了专门的数据库来管理 SAGE 数据，如小鼠内脏在不同发育阶段的 SAGE 数据库 GutSAGE(http://genome.dfci.harvard.edu/GutSAGE/) 和 StormSAGE(http://genomedfci.harvard.edu/StomSAGE/)、精子发育的 SAGE 数据库 GermSAGE(http://germsage.nichd.nih.gov/germsage/home.html)等。

（六）SAGE 数据分析

对 SAGE 数据分析主要包括从原始的序列中得到标签列表，比较来自不同组织细胞或不同生理状态乃至不同物种的标签及其出现频率，在相应数据库中搜索匹配序列，进行基因功能的分析或发现新的基因等。

目前，用于 SAGE 数据分析的应用软件很多，并且在不断发展，主要有：

1. SAGE300　SAGE300 是约翰霍普金斯大学 SAGE 研究计划开发的 SAGE 数据分析软件，与 SAGEnet 提供的 SAGE 实验方案配套使用，对于学术用户免费。学术用户要获得该软件，可以在 SAGEnet 网站下载申请获得该软件的协议表，填写相关信息后传真到约翰霍普金斯医学院技术转让部门，协议文件和传真地址可从下列网址获得：http://www.sagenet.org/protocol/index.htm。

2. WEBSAGE　WEBSAGE 是一个基于 web 的 SAGE 数据分析工具(http://www2.mnhn.fr/websage/)，可以用于对 SAGE 数据进行统计分析，鉴别差异表达的标签，绘制分析结果的散点图等。该软件是一个免费软件，且使用简单。用户登录其网址提交需要分析的 SAGE 数据，程序将自动进行分析并返回分析结果。

3. ATCG　ATCG 是一个在线的表达序列数据分析工具，可以从标签序列来构架基因表达图谱，支持 SAGE、MPSS、SBS 数据。该软件的访问地址为：http://retina.med.harvard.edu/ACTG/。该软件接受 10bp 的短 SAGE 标签、17bp 的长 SAGE 标签、13bp 的 MPSS 标签、16bp 的 MPSS 或 SBS 标签。用户可以在标签数据提交页面上输入标签序列或者上传标签序列文件，选择数据来源(人或小鼠)和类型，选择合适的数据库，然后运行 ATCG 进行分析。此外，该软件所在网站还提供了人和小鼠的标签到基因的映射数据的下载。

4. POWER-SAGE　POWER-SAGE 是一个 SAGE 实验辅助分析工具，可以对不同大小的样本和不同使用频率的标签的组合进行“虚拟”的 SAGE 实验分析，用以确定最好的实验方案。POWER-SAGE 能够适应大规模转录和不同的样本大小，是规划 SAGE 实验的一个有用的工具。获取该软件可以联系作者，联系邮箱为：michale.man@pfizer.com。

5. 其他 SAGE 数据分析软件　还有其他一些研究者开发了相关的 SAGE 数据分析工具，如 Vitural-SAGE、eSAGE、USAGE 等。这些工具的开发都进一步丰富了 SAGE 数据的分析方法。

小　结

表达序列是指由基因表达为 RNA 的序列。表达序列标签(EST)是从 cDNA 文库中随机挑取克隆，测序后获得的序列，通常为几十至 500bp 左右，它大多不是完整的基因序列，只携带了表达基因的部分遗传序列。这些序列存放在相关 EST 数据库中，最常用的三个数据库是 dbEST、UniGene 和 Gene Indices。EST 数据可以用于构建基因组物理图谱、基因识别、研究基因组表达谱、发现新基因以及发现 SNP 位点等。

SAGE 技术是基因表达系列分析。SAGE 技术的实现包括 SAGE 文库的构建、多聚体分子的

克隆与测序和标签序列的提取。SAGE 技术获得的数据主要存放在 NCBI 的 GEO、SAGEnet 以及 SAGE Genie 等数据库。SAGE 数据反映的是特定细胞内、特定时期(特定阶段)、特定处理后的所有表达转录本(包括低丰度转录本)序列,而每一个转录本只测序其中的十几个或数十个碱基序列。它不仅可反映某基因是否表达,而且可反映出基因的表达强度,是从总体上全面研究基因表达、构建基因表达图谱的首选策略,被认为是一种大规模的、高通量的定性与定量获取基因组表达信息的技术,是系统生物学时代的标志性技术之一。本章中介绍了 SAGE 技术的实现方案、相关数据库以及 SAGE 数据的分析方法。

Summary

Expressed sequences are referred to those RNA sequences transcribed from genomic DNA. An expressed sequence tag (EST) is a short sub-sequence of a transcribed cDNA sequence, which is usually obtained from end sequencing of random clone in cDNA library. EST is not a completed gene with only a part of genetic sequence, usually 100~500 bp in length. EST data is deposited in international database, the most commonly used being dbEST, UniGene and Gene Indices. EST data are usually used for construction of genome physical map, gene identification, gene expression profile study, discovery of new gene and SNP and so on.

SAGE is a special biotechnique for Serial Analysis of Gene Expression. Performance steps of SAGE include construction of SAGE library, cloning and sequencing of polymeric mRNA sub-molecules, extraction and analysis of tag sequence. SAGE data is deposited in database GEO of NCBI, SAGEnet and SAGE Genie. SAGE data reveal global expression profiles of all genes under given space and phase conditions or special treatments. It detects all transcripts in a cell or tissue, including low abundance transcript. It reflects not only what kind of genes express but also magnitude of expression. So it is the first choice for global expression and construction of expression profiles. SAGE is considered as a qualitative or quantitative high-throughput technique for acquisition of gene expression information, and also a marked technique in systemic biology epoch. In this chapter SAGE procedure, databases relative to SAGE and analysis method had been described.

(邹凌云 胡福泉)

习 题

1. 什么是表达序列,什么是表达序列标签(EST)?
2. EST 分析具有哪些用途?
3. EST 数据是怎样获得的?
4. 国际上常用的EST数据库有哪些?怎样向EST数据提交数据?怎样从EST数据中获取数据?
5. 常用 EST 数据分析方法有哪些?怎样实现对 EST 数据的分析?
6. 什么是 SAGE 技术? SAGE 技术基于什么基本原理?
7. SAGE 技术的实现方案包括哪些步骤?
8. SAGE 技术具有哪些应用前景?
9. SAGE 技术获取到的数据存放在哪里?它的相关数据库的访问与使用方法有哪些?

主要参考文献

1. Velculescu V. E. Tantalizing transcriptomes--SAGE and its use in global gene expression analysis. Science, 1999, 286(5447): 1491-1492.
2. Velculescu V. E., Zhang L., Vogelstein B., et al. Serial analysis of gene expression. Science, 1995, 270(5235): 484-487.
3. Tremain N., Korkko J., Ibberson D., et al. MicroSAGE analysis of 2353 expressed genes in a single cell-derived colony of undifferentiated human mesenchymal stem cells reveals mRNAs of multiple cell lineages. Stem Cells, 2001, 19(5): 408-418.
4. Sharon D., Blackshaw S., Cepko C. L., et al. Profile of the genes expressed in the human peripheral retina , macula , and retinal pigment epithelium determined through serial analysis of gene expression (SAGE). Proc. Natl. Acad. Sci. USA, 2002, 99(1): 315-320.
5. Datson N. A., van der Perk-de Jong J., van den Berg M.P., et al. microSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. Nucleic Acids Res., 1999, 27(5): 1300-1307.
6. Ye S. Q., Zhang L. Q., Zheng F., et al. miniSAGE: gene expression profiling using serial analysis of gene expression from 1 microg total RNA. Anal. Biochem., 2000, 287(1): 144-152.
7. St Croix B., Velculescu V. E., Zhang L., et al. MicroSAGE detailed protocol. http://www.sagenet.org/Protocol.
8. Chen J. J., Rowley J. D.,Wang S. M. Generation of longer cDNA fragments from serial analysis of gene expression tags for gene identification. Proc. Natl. Acad. Sci. USA, 2000, 97(1): 349-353.
9. Chen J. J., Lee S., Zhou G. L., et al. High throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. Genes Chromosomes Cancer, 2002, 33(3): 252-261.
10. Ryo A., Kondoh N., Wakatsuki T., et al. A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. Anal. BioChem., 2000, 277(1):160-162.
11. Saha S., Sparks A. B., Rago C., et al. Using the transcriptome to annotate the genome. Nature Biotech., 2002, 19(5): 508-512.
12. Anisimov S. V. Serial Analysis of Gene Expression (SAGE): 13 years of application in research. Curr. Pharm. Biotechnol., 2008, 9(5): 338-350.
13. Datson N. A. Scaling down SAGE: from miniSAGE to microSAGE. Curr. Pharm. Biotechnol., 2008, 9(5): 351-361.
14. Matsumura H., Krüger D. H., Kahl G., et al. SuperSAGE: a modern platform for genome-wide quantitative transcript profiling. Curr. Pharm. Biotechnol., 2008, 9(5): 368-374.
15. Matsumura H., Reuter M., Krüger D. H., et al. SuperSAGE. Methods Mol. Biol., 2008, 387: 55-70.

第七章 基因芯片数据分析
CHAPTER 7 MICROARRAY DATA ANALYSIS

第一节 引 言
Section 1 Introduction

基因芯片或称微阵列(microarray)是20世纪90年代随着计算机技术和基因组测序技术的发展而产生的一种新型的生物技术，它能够平行、高通量地检测成千上万基因转录本的表达水平，为系统地检测细胞内mRNA分子的表达状态并推测细胞的功能状态提供了可能。同时高通量数据也为数据的分析任务提出了新的挑战。应用基因芯片可以比较正常和异常细胞中基因的表达，帮助识别疾病相关基因和药物作用靶标，分析复杂疾病的致病机制，为个性化诊断和治疗提供指导，也可以揭示基因间的表达调控关系，同时它在制药和临床研究中也有重要的作用。

第二节 常见的芯片平台与数据库
Section 2 General Microarray Platform and Database

知 识 拓 展

基因芯片技术改变了生物学研究的方法，从单个基因的研究迅速扩展到全基因组的系统生物学研究，基因芯片技术帮助生物学研究进入后基因组时代。基因芯片技术经过近15年的发展已经形成了一个系统的平台，从样品制备、芯片制作、芯片杂交、数据扫描到后期的数据管理，储存以及深度数据挖掘都有了标准化的流程、坚实的理论和实验的支持，成为一个非常稳定可信的实验技术。近几年基于该技术新开发的SNP芯片可用于比较不同个体间基因组SNP位点差异及基因组拷贝数变异，为从基因组变异的角度研究疾病的发生机制提供了研究基础；新开发的微小RNA芯片可以检测微小RNA在发育过程或人类疾病发生和发展过程中的表达变化，从而为干细胞研究和癌症研究提供了研究平台；新开发的DNA甲基化芯片及CHIP-chip实验平台为表观遗传学研究和转录调控机制的研究提供了新的手段。

基因芯片的制备原理类似于Northen杂交，它也是基于碱基互补配对的原理测量细胞内mRNA表达丰度的实验方法。但是，与Northen杂交显著不同的是基因芯片可以同时检测成千上万个基因的表达水平。这种高通量技术的主要特点为平行性、微型化和自动化。平行性是指它对基因的检测可以做到时空的一致性，例如，可以使用一张芯片检测细胞内所有基因在某个组织中、某个时间状态下的表达状态，从而在后期的分析中不会受到时空因素的影响。微型化指基因芯片非常小巧，携带方便。基因芯片的自动化指芯片探针的制备固定、探针与实验样本的杂交、信号的提取过程等都依赖于计算机自动完成。

根据探针制备原理的不同可将基因芯片分为预先合成然后点样芯片、原位合成芯片和新一代的光纤微珠芯片(BeadArray)。预先合成然后点样芯片根据探针类型的不同又可分为 cDNA 芯片和寡核苷酸芯片，前者的探针是全长 cDNA 序列，后者的探针是运用传统的 DNA 合成仪合成的寡核苷酸序列，预先设计的探针运用点样机器人以高密度分布于硝酸纤维膜或经过处理的玻片上。而原位合成芯片直接在固体基质上用 4 种单核苷酸合成所需的寡核苷酸片段。GeneChipTM 是高密度寡核苷酸微阵列原位合成的代表，制造工艺采用原位光刻合成。其他原位合成制造工艺还有光敏抗蚀层并行合成法、微流体通道在片合成法、喷印合成法及分子印章在片合成法。光纤微珠芯片是新一代基因芯片产品，它是一种以光导纤维和纳米材料(硅珠)为主要组成元件的芯片。与前两种芯片的最大差别在于，光纤微珠芯片的探针序列不是固定在平板上，而是固定在球形的硅珠表面，从而提高了杂交的均匀性和效能。

一、cDNA 微阵列芯片

cDNA 芯片(cDNA microarray)实验的工作流程如图 7-1 所示，通过克隆的方法获得目标 cDNA 序列并将其作为探针，并通过点样机器人将探针高密度固定在表面通过特殊处理的基质(如玻片)上，制备成 cDNA 芯片。cDNA 芯片为双通道双染色芯片，即一张芯片运用两种荧光标记可同时检测两种不同条件下基因的表达水平。

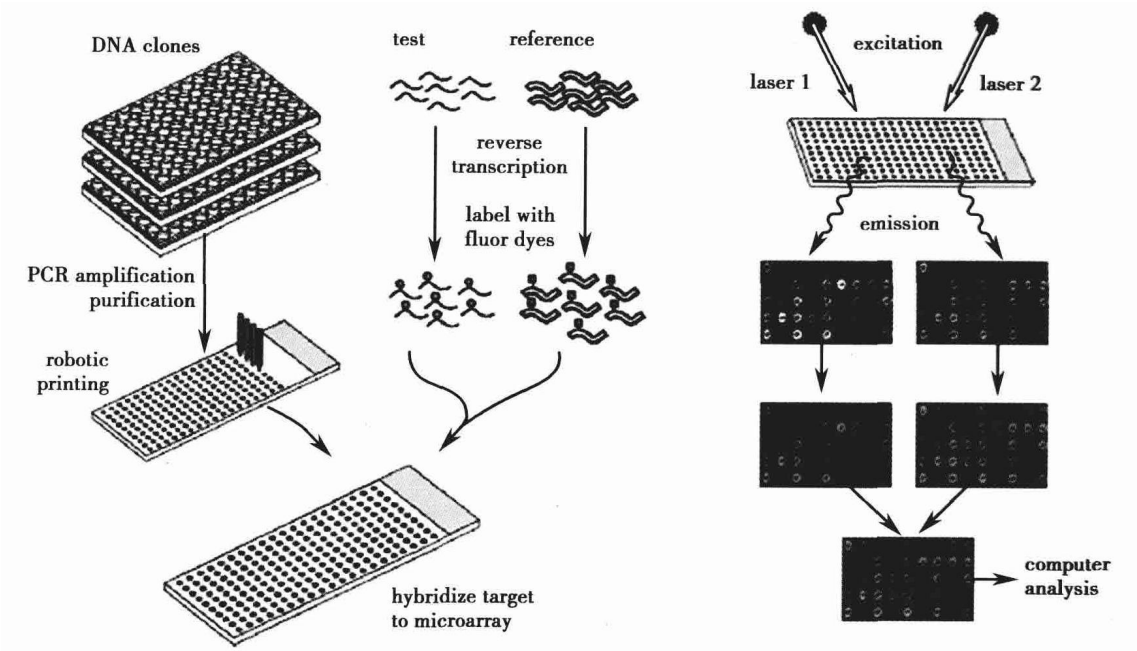


图 7-1 cDNA 芯片实验的工作流程

(引自：http://www.biox.cn/content/20050420/11491.htm)

图中实验组(test)为某实验条件下提取的总 mRNA，对照组(reference)为对照条件下提取的总 mRNA(mRNA 反转录成 cDNA，因为 RNA 本身不稳定，而 cDNA 保存时间更长)。两组样本分别用不同的荧光染料进行标记，这里实验组用 cy5 荧光分子进行标记，对照组用 cy3 荧光分子进行标记。两组样本等量混合后在一定的实验条件下与芯片上的探针进行杂交。杂交结束后洗脱那些没有与探针互补结合的 cDNA 片段，然后将芯片置入黑箱中，分别用 cy3 和 cy5 荧光染料对应波长的激发光对芯片进行激光共聚集扫描，获取两种不同条件下芯片上每个探针杂交后的荧光信号强度，运用该荧光强度推测各种基因的相对表达水平。在杂交结果的可视化处理中，通常 cy3 通道的杂交信号用绿色荧光显示，cy5 通道的杂交信号用红色荧光显示，两通道的荧光进行重叠后，基于荧光颜色可以对每个基因的表达情况进行初步判断。例如，黄色荧光表示红绿通道荧光信号对等，基因在两种不

同条件下的表达量相当；偏红色的荧光表示基因在实验组的表达有上调；偏绿色的荧光表示基因在对照组的表达有上调。由于 cDNA 芯片探针来源于获得的 cDNA 克隆，因此探针的长短大小不一，需要的杂交条件也不同。但是在进行芯片杂交实验时，只能设定一个固定的杂交条件(如 T_m 值等)，这就出现了由实验体系本身导致不能很好地解决非特异性杂交和杂交效能等问题，因而可靠性和重复性不是很理想。

二、寡核苷酸芯片

寡核苷酸芯片(oligonucleotide microarray)类似于 cDNA 芯片，但是在探针的设计上优于 cDNA 芯片。它的探针并不是来源于 cDNA 克隆，而是预先设计并合成的代表每个基因特异片段的序列，长度约为 50bp。然后将其点样到特定的基质上制备成芯片，从而克服了探针序列太长导致的非特异性交叉杂交和探针杂交条件变化巨大导致的数据结果的不可靠。但由于 oligo 芯片的探针预先一次性合成，并且每次芯片的使用很少，在一批实验过程中系列芯片的制备存在时间差，因此早期合成的 Oligo 探针存在着降解而导致检测质量下降的情况，这样除非重新合成探针，否则后期芯片的检测质量会降低很多。同时，如果实验目标发生变动，则合成的探针无法继续使用，造成浪费。

三、原位合成芯片

原位合成芯片(chip)在制备时采用的并不是上述两种芯片的点样技术，而是光引导聚合技术。它是将光平版印刷技术(photolithographic approach)运用到 DNA 化学合成中，利用固相化学、光敏保护基及光刻技术得到位置确定、高度多样性的探针簇。该技术的原理如图 7-2 所示。

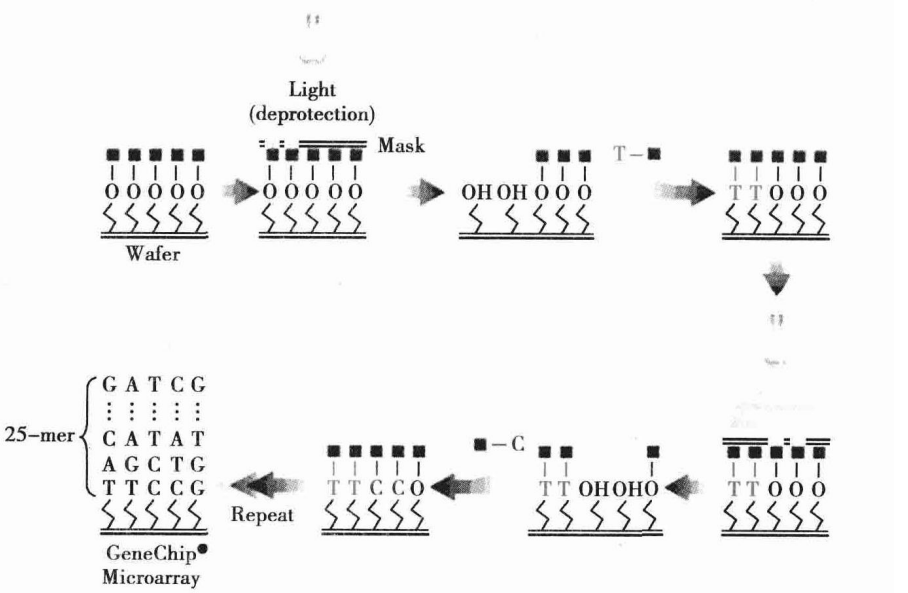


图 7-2 原位合成芯片的探针制备过程

(引自：http://www.dkfz.de/gpcf/affy_technology.html)

该法利用光敏保护基来保护碱基单位的 5' 羟基。第一步利用光照射使固体表面上的羟基脱保护，然后固体表面与光敏保护基保护、亚磷酰胺活化的碱基单体接触，合成只在那些脱保护基的地方进行。光照区域就是要合成的区域，该过程通过一系列掩膜来控制。如此循环以合成寡核苷酸，直到达到设定的寡核苷酸长度。其长度一般为 15~25 个碱基。

芯片实验的工作流程如图 7-3 所示。芯片为单通道单染色芯片，即一张芯片运用一种荧光标记检测一种条件下基因的表达水平。与 cDNA 芯片类似，首先从组织或细胞中提取总 mRNA，mRNA 反转录成双链 cDNA。当样本需要与芯片进行杂交时，cDNA 在体外又转录成 RNA(cRNA)，并用生

物素(Biotin)标记。荧光标记后的 cRNA 被随机打成 30~400 碱基长度的片段，然后与芯片探针进行杂交。洗脱不能与探针互补的样本序列，运用能与生物素结合的荧光分子 cy5 标记与探针序列结合的样本 cRNA。最后扫描芯片，提取荧光信号。

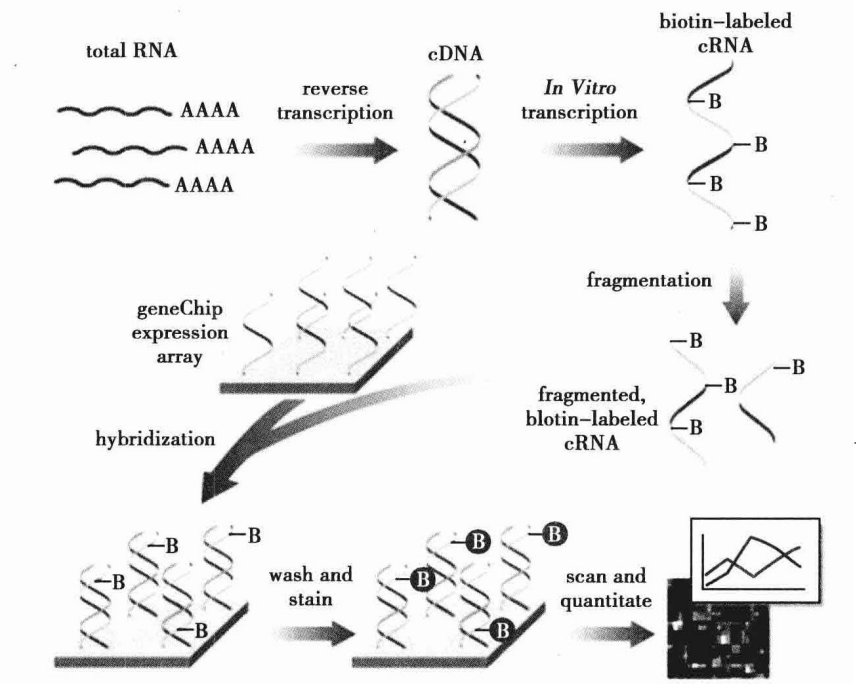


图 7-3 芯片实验的工作流程

(引自 http://www.dkfz.de/gpcf/affy_technology.html)

这种方法的优势在于同一批芯片的所有探针都是在一个条件下完成的，因此同一批芯片的探针浓度的均一性很好，进而使得检测数据的重复性很好。同时，由于原位合成芯片中探针合成和芯片制备的过程同时进行，探针不需要预先合成，避免了传统点样芯片探针的降解情况，从而保证了实验的重复性。芯片平台的另一个优势是在进行探针设计时考虑了重复和对照的原则，如图 7-4 所示。

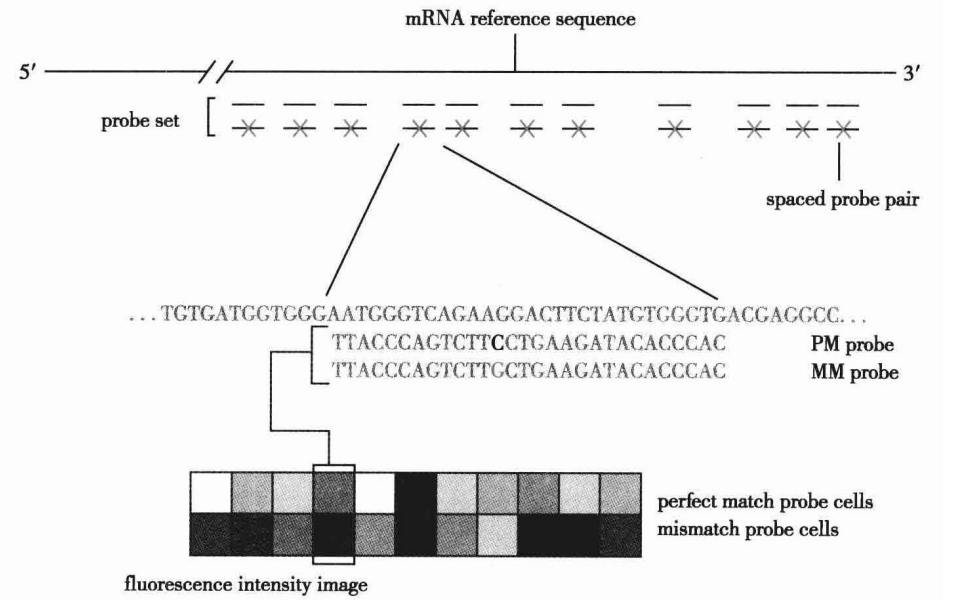


图 7-4 原位合成芯片的探针设计

(引自 http://www.dkfz.de/gpcf/affy_technology.html)

对于某个待检测的基因，该平台设计了多个重复的寡核苷酸探针构成探针集，从基因序列的3'端以覆瓦的方式设计，每个寡核苷酸探针还设计了对照的错配探针(mismatch，MM)，即在该探针的中间位置发生一个核苷酸的改变，该对照探针能检测到的荧光强度为探针非特异性杂交信号强度。与对照探针相对应的寡核苷酸探针称之为完美匹配探针(perfect match，PM)。这种独特的PM-MM探针设计，能有效解决芯片非特异杂交问题，真正提高检测效能。

四、光纤微珠芯片

光纤微珠芯片是利用独特的微球阵列(beadarray)技术生产的芯片，是新一代基因芯片产品。光纤微珠芯片以光导纤维和纳米材料(硅珠)为主要组成元件。探针连接在硅珠上(图7-5)，每个探针由两部分组成，即地址序列和探针序列。地址序列由23bp组成，特异地对应于某个基因的微珠，它可以对每种类型的微珠进行编码，只在解码过程中使用。探针序列为代表每个基因特异片段的约50bp长度的序列。每种类型微珠上可连接100万左右相同的探针，对于某个基因的检测，在一个芯片平台上平均每种类型的微珠有30个拷贝，从而保证了探针数目的充裕。

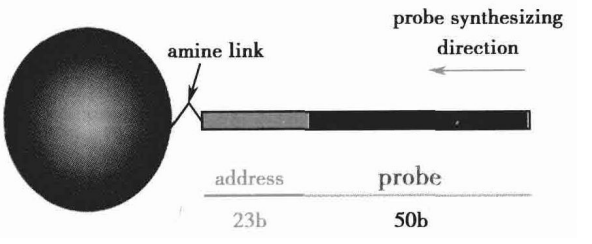


图7-5 光纤微珠芯片的探针设计

不同类型的微珠混合到一起形成“微珠池”。图7-6为光纤微珠芯片的制备流程。在上述微珠池中插入若干束极微小的光纤，5万根光纤组成一束，每根光纤的末梢有一个用化学方法蚀刻的微孔，每个微孔恰可容纳一个直径为3μm的微珠，当接触到微珠池时，每根光纤会“拾起”一粒微珠并用末端的开口将其牢牢夹住。经过此番处理后，即可将这些光纤束组装成书本大小的功能芯片平台。微珠以“无序自组装”的方式随机进入光纤束，从激光扫描仪上发出的激光通过光纤传递给荧光素，后者发出的荧光又通过光纤传递给检测器。采用解码流程对芯片上微珠的类型、位置、数量、信号强弱进行解读，不合格的微珠信道将被关闭。解码过程既能完成对芯片信息的收集确认，又实现了芯片生产过程中的100%质控。

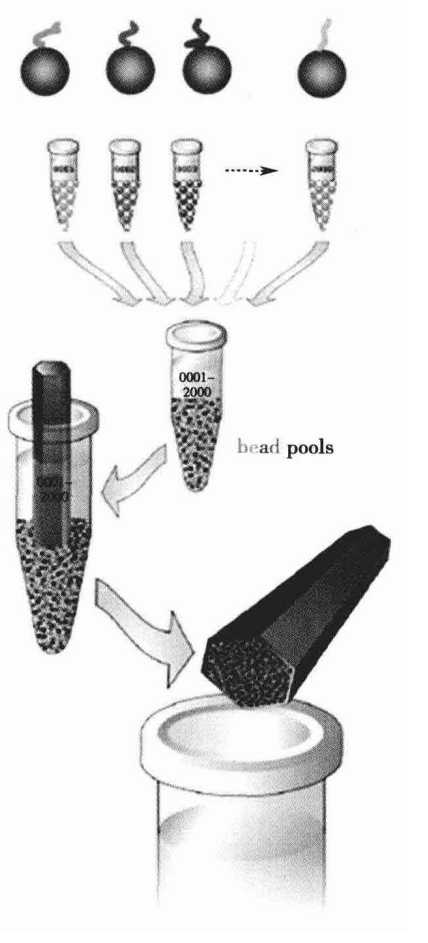


图7-6 光纤微珠芯片的设计

基因芯片产生了大量的基因表达数据，这些数据为功能基因组研究提供了重要的资源。为了能共享及使用这些数据，需要对这些数据定义共享的数据结构，包括数据注释标准、通用的数据交换格式以及基因表达数据的公共数据库。基因芯片数据注释标准为MGED(the Microarray gene expression data)组织开发的微阵列实验最小信息(the minimum information about a microarray experiment，MIAME)，它是解释和验证结果所必需的微阵列实验的最小信息描述。同时，MGED组织还开发了微阵列基因表达标记语言(microarray gene expression -markup language，MAGE-ML)，它是一种用来描述基于实验的微阵列信息的通讯语言。MAGE-ML基于XML，可以描述微阵列设计、制造、实验组织和实施信息、基因表达数据等。美国NCBI的gene expression omnibus，GEO、英国EBI的ArrayExpress数据库都采用了该标准，斯坦福微阵列数据库(stanford microarray database，SMD)也正在兼容该标准。

五、基因表达数据库

基因表达数据库(GEO)是由美国NCBI管理和维护的公共数据库，是经过专家整理和核对的在线的基因表达数据库资源，主要储存基因表达数据，提供基因表达数据的浏览、查询和检索。GEO数据库中的记录是根据其不同的数据类型以不同的登录号编写方式进行区分。由芯片生产厂家递交的芯片平台数据以登录号GPL*** 表示；单张芯片描述的原始数据及处理后的荧光强度数据以登录号GSM*** 表示；单个实验(包括一系列芯片)的数据以登录号GSE*** 表示；由NCBI整理的具备相似实验条件，有生物学意义，在统计学上具有可比性的不同的芯片实验构成的实验集组，以登录号GDS*** 表示。GEO数据库提供一些简单的数据分析功能，例如差异基因筛选、聚类分析等。GEO定义了两个数据库：Datasets和Profiles。Datasets存储了以“实验为中心”的芯片数据。Profiles存储了以“基因为中心”的单个基因表达的数据。GEO数据库中的数据存储方式主要有：芯片原始数据，如cel文件或cDNA芯片的扫描图像文件；以MIAME兼容的形式递交的数据；矩阵形式存储的txt文件。

六、斯坦福微阵列数据库

斯坦福微阵列数据库(SMD)是由美国斯坦福大学管理和维护的基因表达的公共数据库。SMD存储微阵列实验得到的原始和标准化数据，提供网页界面进行数据检索、分析和可视化。SMD充分利用了许多公共数据库资源，例如SGD、YPD、WormPD Unigene、dbEST和SWISS-PROT，将表达数据和相关的其他生物信息建立联系。

七、其他常用基因表达数据库

ArrayExpress是由EBI管理和维护的基因表达数据的公共数据库。它由两部分组成：一部分收录了运用MIAME标准支持的基因芯片实验数据，另一部分是基于第一部分数据重新统一注释后的基因表达谱数据。基因芯片实验数据可以通过关键字、物种、芯片平台、作者、杂志或登录号进行检索查询。基因表达谱数据可以通过基因的名称和其他属性进行查询。

CGED(cancer gene expression database)是包括基因表达谱和临床信息的癌症基因表达数据库，包括乳腺癌、结肠癌、肝癌、食道癌、甲状腺癌和胃癌。该数据库的数据由日本奈良生命科学与技术学院、大阪大学医学院、京都大学医学院和大阪癌症和心血管疾病医学中心合作提供。

第三节 基因芯片数据的预处理

Section 3 Preprocessing of Microarray Data

由于获取的芯片原始数据来自不同的芯片平台，数据信息会有差异。往往需要前期的数据预处理以后才能进行深层次的数据挖掘，这种预处理主要包括数据提取、数据对数转化、数据过滤、补缺失值和标准化处理等。

一、基因芯片数据的提取

对于双通道的cDNA微阵列芯片和寡核苷酸芯片，扫描后的一张芯片图像及将某个荧光点(spot)放大后的图像如图7-7所示。扫描后的一张芯片图像中，红色的荧光点表示该点所检测的基因在两种实验条件下相比表达有上调，绿色的表示表达有下调，黄色的表示表达无改变。放大后的图像中主要包含的信息有：通道1的前景荧光强度值*CH1I*代表第一种条件下基因的表达值，通道1的背景荧光强度值*CH1B*代表第一种条件下非特异的荧光强度背景值；通道2的前景荧光强度值*CH2I*代表第二种条件下基因的表达值，通道2的背景荧光强度值*CH2B*代表第二种条件下非特异的荧光强度背景值。该基因在两种条件下的荧光强度比值为$Ratio=(CH1I-CH1B)/(CH2I-CH2B)$。

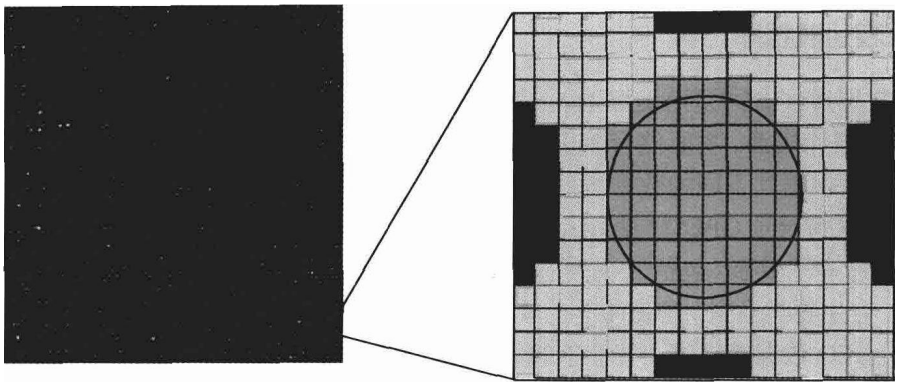


图 7-7 cDNA 微阵列芯片荧光信号

圈内红色像素为前景信号，一般用圈内像素的中值或均值表示前景荧光强度值 *CHI*；灰色像素为背景信号，一般用圈外灰色像素的中值或均值表示背景荧光强度值 *CHB*，该基因在某种条件下的荧光强度值为两者之差 *CHI* - *CHB* 黑色像素为邻居荧光点

高通量的荧光信号需要有专业的软件将其转化成数字信号。不同的生产商在生产双通道的 cDNA 微阵列芯片和寡核苷酸芯片时，由于芯片扫描系统的图像处理软件不同(例如，GenePix 软件、Feature Extraction 软件)，从而提取的基因芯片原始数据内容和格式也有差异，但数据中最基本的双通道的前景和背景信息 *CH1I*、*CH1B*、*CH2I* 和 *CH2B* 是不变的。

图 7-8 为单通道芯片(左图)及扫描后的基因芯片荧光图像(中图)和放大后荧光图像(右图)，右图中黑色的荧光块(feature)表示无荧光强度，即该荧光块对应的基因没有杂交信号，荧光强度水平按照颜色从低到高依次为蓝黑、蓝、高蓝、绿、黄、橙、红、白。荧光强度越高表示与探针杂交的 RNA 越多，从而基因的表达量越高。

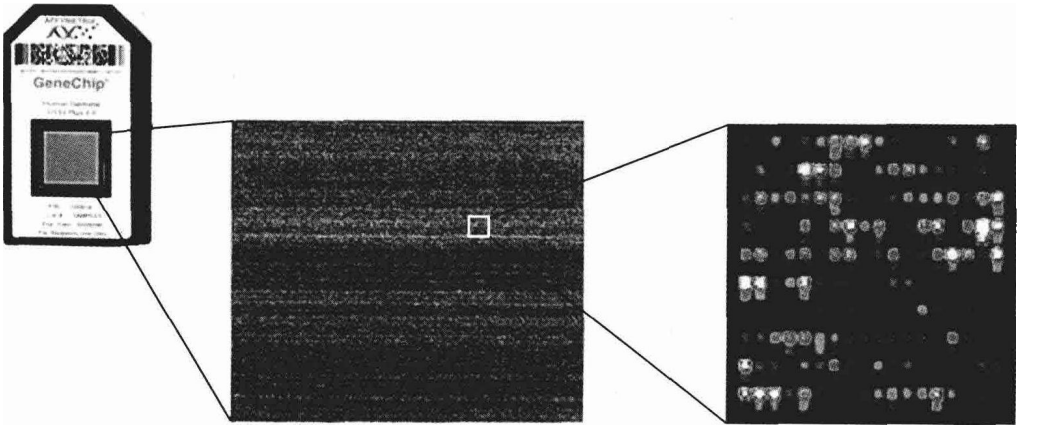


图 7-8 芯片外观及扫描后的荧光图像

芯片对于某个待检测的基因设计了探针集进行检测，因此芯片检测的探针数远大于基因数，例如 Human Genome U133 芯片包含了 100 万个不同的寡核苷酸探针，代表了 33 000 个人类基因，所以芯片扫描系统的图像处理软件不但包括将荧光信号转化成数字信号的数据提取，也包括基于探针集的基因表达值汇总提取。运用数据提取软件提取后的探针水平的数据以“扩展名 .cel”的文件格式进行保存。而通常以文本形式存储的原始数据是经过汇总和标准化后的基因表达信息，包括定性和定量信息。定性信息以 P/A/M(Present/Absent/Marginal)表示，说明某基因在某条件下的表达判断有无或不确定。定量基因是基于探针集汇总后的基因水平的荧光信号强度值。

提取后的大规模基因表达数据通常可以用矩阵形式表示，行代表基因，列代表样本，矩阵中的元素代表基因在样本中的表达水平。例如，采用点有 p 个基因探针的 DNA 芯片检测 n 个样本的表达

谱数据可由 $p \times n$ 矩阵 $X=(x_{ij})$ 表示，其中 x_{ij} 可代表第 i 个基因 g_i 在第 j 个样本 X_j 的表达水平。则样本集 $X=\{X_1, X_2, \cdots, X_n\}$ 中的每个样本 X_j 为一个 p 维向量；基因集 $g=\{g_1, g_2, \cdots, g_p\}$ 中的每个基因 g_i 为一个 n 维向量。基因表达谱中蕴含着丰富的信息，许多生物信息学的研究都致力于挖掘其中有意义的信息。

二、数据对数化处理

芯片原始数据一般呈偏态分布，如果对数据做对数化转换后，数据可近似正态分布，从而为后续的数据分析带来方便。

三、数据过滤

数据过滤的目的是去除表达水平是负值或很小的数据或者明显的噪音数据。通过简单的数据处理软件得到的基因表达谱数据，每个点的荧光信号强度通常为前景信号值减去背景信号值。在某些情况下，例如过闪耀现象，由于邻近基因背景的强信号辐射得到了较大的背景信号值，而该点对应基因的表达量很低或没表达得到了较小的前景信号值，从而导致该点基因的荧光信号值为负；或者芯片存在物理因素导致的信号污染，如划伤、手指印等。由于这些因素导致的不真实数据会给后期的处理带来噪音，所以需要对数据进行过滤处理。

四、补缺失值

基因表达谱中的数据缺失大致分为两种类型：一种是非随机缺失，在这种情况下数据缺失跟基因的表达丰度有关，例如基因的表达丰度过低，背景值超过前景信号值；或基因的表达丰度过高，高表达基因的荧光强度值超过了能检测的最大信号强度阈值(图 7-9)。对于这种情况，目前的数据补缺方法还无法有效的处理。另一种是随机缺失，即基因表达谱中的数据缺失与基因表达值的高低无关，而是与其他的因素，例如杂交效能低、物理刮伤、指纹、灰尘、图像污染等，数据补缺处理对于这种情况比较有效。

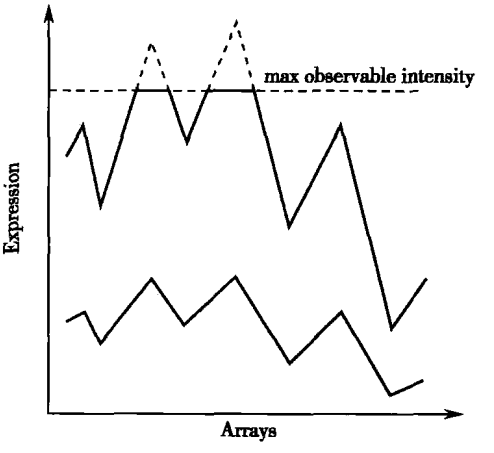


图 7-9 高表达基因的数据缺失
(引自：Shaffer J: Missing Value Imputation for Microarray Data. Power Point 2006)

设基因表达谱矩阵 X 中第 i 个基因在第 j 个样本下表达值 x_{ij} 缺失，对于缺失值的处理有两种方法：一是删除含有缺失值的行或列；二是数据补缺。前种方法的处理会丢失一些有用信息，适用于行或列中包含的缺失值较多的情况。而数据补缺的方法有以下几种。

(一) 简单补缺法

用 0、1、每行或每列的均值作为缺失值的可能信号值。一般用 0 值补缺时，基因表达数据为一种条件下的荧光信号值或两种条件下的荧光信号比值的对数值，用 0 补缺则认为该基因在某种条件下无表达或在两种条件下的表达无差异；用 1 值补缺时，基因表达数据为两种条件下的荧光信号比值，用 1 补缺认为该基因在两种不同条件下无差异表达；用每行或每列的均值补缺时，则认为某基因在某样本中表达的缺失值估计为该基因在其他样本中表达的平均水平或所有基因在该样本中表达的平均水平。

(二) k 近邻法

k 近邻法的基本思想是用在总样本空间中与待补缺基因距离相近的 k 个邻居基因在缺失条件下的表达值推测缺失值。首先确定含有缺失值的基因 i 的 k 个邻居基因，设 $x_{1j}, x_{2j}, \cdots, x_{kj}$ 分别为基因 i

的 k 个邻居基因在第 j 个样本中的表达值，常用的定义邻居基因的距离函数有欧氏距离或相关系数；然后运用邻居基因在该样本中信号值的加权平均估计缺失值：

$$x_{ij} = \sum_{g=1}^{k} w_g x_{gj} \qquad \text{式 7-1}$$

这里 w_g 为权重系数，由邻居基因 g 与基因 i 的距离决定，距离越近 w_g 越大。

（三）回归法

回归法与 k 近邻法相似，区别在于 k 近邻法用邻居基因对应表达值的加权平均估计缺失值，而回归法用回归模型预测缺失值，然后再加权平均。回归法的基本步骤为：

1. 首先确定含有缺失值的基因的 k 个邻居基因，设 $X_1, X_2 \cdots X_k$ 为基因 i 的 k 个邻居基因在 n 个样本中的表达向量。

2. 具有缺失值的基因 X_i 较之邻居基因分别作线性回归模型，基于回归模型预测缺失值：

$$\begin{aligned} x_{ij}^{\ 1} &= a_1 + b_1 x_{1j} \\ x_{ij}^{\ 2} &= a_2 + b_2 x_{2j} \\ &\vdots \\ x_{ij}^{\ k} &= a_k + b_k x_{kj} \end{aligned} \qquad \text{式 7-2}$$

3. k 个缺失值的加权平均为最终的缺失值估计值：

$$x_{ij} = \sum_{g=1}^{k} w_g x_{ij}^{\ g} \qquad \text{式 7-3}$$

这里 w_g 为邻居基因的权重，若邻居基因与第 i 个基因的距离近，权重大，反之权重小。

（四）其他方法

包括奇异值分解、因子分析回归等方法。

五、数据标准化

预处理过程最主要的一个步骤是数据标准化(normalization)。由于基因芯片数据中存在有不同来源的变异，主要包括两方面：感兴趣的变异和混杂变异，前者指生物来源的变异，例如正常组样本和疾病组样本基因转录本表达的差异，而后者指在芯片实验过程中引入的变异，例如在样本的染色、芯片的制作、芯片的扫描过程中引入的系统误差(偏倚)，只有运用正确合理的标准化方法去除这些系统误差才能发现真正的生物学变异，确保后期数据分析的可靠性。

不同芯片平台的制作原理不同，引入的系统误差不同，标准化的方法也有差异。下面以双通道的 cDNA 芯片和单通道的芯片为例介绍标化的基本方法。重点介绍 cDNA 芯片的标化方法。

（一）cDNA 芯片

1. 系统误差来源　对于 cDNA 芯片而言，整个实验过程可能引入的系统误差主要来源于以下几个主要方面：染料的物理属性(Cy3 染料和 Cy5 染料的热光敏感性、半衰期不同)、染料的结合效率、探针的制备、探针和样本的杂交过程、数据收集时的扫描过程、不同芯片间的差异、不同芯片杂交条件的差异等。

2. 标准化过程的参照物　在对芯片数据进行标化处理时，涉及参照物的选定，一般以具有稳定表达的基因作为参照物，即已知有一些基因在不同的条件下其转录本的表达量相等，那么该基因通过芯片测得的荧光强度值的差异主要由系统误差造成，从而可以估计该系统误差的大小，作出相应的纠正。

稳定表达的基因主要包括以下几类：持家基因(housekeeping genes)、外源性的或人工合成的控制基因(controls)、芯片上大部分稳定表达的基因(所有基因)、相对稳定基因子集(invariant set)等。

用控制基因作为参照基因，即在实验样本和对照样本中都加入表达量相等的参照基因，特别要

保证控制基因在芯片上的探针具有特异性，不与其他的基因进行杂交。运用持家基因和控制基因作为参照基因进行标化时，由于实验误差导致它们本身在两组样本中是否真正相等表达，以及杂交的特异性问题，尤其是持家基因实际上并不像人们想象的那样在不同的实验条件下稳定表达，而是倾向于高表达，所以用若干持家基因或控制基因作为参照基因进行标化，对于标化结果的可靠性存在着一定的挑战。相对而言，运用大部分稳定表达的基因作为参照基因更为可靠。因为对于表达谱数据而言，真正与条件相关的表达异常的基因只有一小部分，而大部分基因在不同的条件下都是稳定表达的，所以运用这大部分稳定表达的基因要比运用若干个认为是稳定表达的基因作为参照基因的可靠性要大得多，但是它的标化方法相应也要复杂。所以这里主要介绍使用所有基因作为参照基因的标化方法，用持家基因和控制基因作为参照基因的标化方法仅简单提及。

另外有人在确定用个别稳定基因(持家基因、控制基因)还是大部分稳定基因作为参照物时选择了一个折中方案，即相对稳定基因子集，这组基因的确定方法是这样的：选择一组基因子集 $g_1, g_2, \cdots, g_m$，它们对应的探针子集 $p_1, p_2, \cdots, p_k$，cDNA 芯片 $m=k$，Oligo 芯片 $m<k$。在两组样本中这组探子集具有相同的次序，即 $p_1<p_2<\cdots<p_k$，这部分探针作为稳定表达的探针。

前面提及的都是以基因作为参照物进行标准化处理，在应用中还有一些方法以重复样本互相作为参照物，这对于相同条件下的重复芯片实验尤其适用，即如果在没有系统误差的情况下，重复芯片的结果应该尽可能吻合，即使有一点差异也是由于随机误差造成的，那么标化的目的就是使重复芯片的一致性提高。下面介绍的有一种非线性标化方法就是以此为理论基础的。

3. 标准化方法　标化的方式不是唯一确定的，根据实验设计的不同，这里将标化分为三部分：片内标化、染色互换实验的自标化和多片间的标化。

(1) 片内标化：在进行标化之前，首先要对数据进行对数转换，即将芯片上所有基因的 Cy5 染料标记(红光)的荧光强度跟 Cy3 染料标记(绿光)的荧光强度相除后取对数值(称为 log-Ratios 值)，经过这种处理后芯片上所有的基因基本满足正态分布，这为以后的数据分析带来方便。片内标化对于一个实验中包含的不同芯片独立操作，主要方法如下：

全局标化(global normalization)：全局标化假设红光的荧光强度(R)和绿光的荧光强度(G)相差一个常数 k，即 $R=k\cdot G$，由于芯片上的大部分基因都是稳定表达的，且芯片上基因的荧光强度值经对数转换后基本满足正态分布，所以芯片上所有基因的 log-Ratios 值的密度分布应该如图 7-10 的黄色曲线所示：

而实际上由于红光和绿光的荧光强度存在差异，即使是具有相同表达水平的两个基因经 Cy3 和 Cy5 两种不同的染料标记后所测得的荧光强度也不一致，黄线的峰值会偏离 0 的位置，由于通常 Cy3 的荧光强度值高于 Cy5，所以峰值会向左偏移，如图中红色曲线所示。全局标化的目的就是要将实际测得的 log-Ratios 值分布的峰值位置移至 0 处：

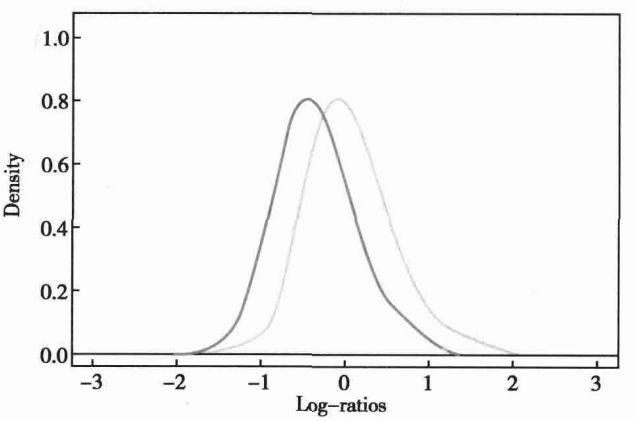


图 7-10　全局标化前后 log-Ratios 值分布图

(引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.)

$$\log_2 R/G \rightarrow \log_2 R/G - c = \log_2 R/(kG) \qquad \text{式 7-4}$$

这里，位置参数 $c=\log_2 k$，它的值是芯片上所有基因的 log-Ratios 值的中值或均值。

全局标化法由于纠正了染料偏倚(dye bias)及其标化方法的简单可行而被普遍应用，但是它并没有考虑芯片的空间差异带来的偏倚和荧光强度依赖的染料偏倚。这种方法对以相对稳定基因子集、持家基因或控制基因作为参照基因时同样适合，只不过在估计位置参数时仅采用相对稳定基因子

集、持家基因或控制基因来估计，在其他方法中如合适也可以考虑类推。

荧光强度依赖的标化(intensity dependent normalization)：在许多情况下，染料偏倚的大小依赖于荧光强度，Yang 等对荧光强度与染料偏倚的关系作过如下的研究，即以 log-Ratios 值 $M=\log_2 R/G$ 作为纵坐标，以平均荧光强度 $A=\log_2\sqrt{RG}$ 作为横坐标，根据芯片上所有基因对应的 M 值与 A 值作散点图，结果如图 7-11 所示。

这说明对于不同 A 值处的大部分基因的 M 值偏离 0 的幅度不同，对它们进行校正时也应该区别对待。荧光强度依赖的标化的目的就是要将不同 A 值对应的 log-Ratios 值分布的峰值位置移到 0 处，经过标化后的 M 值与 A 值的散点图中散点应该分布于 $M=0$ 的轴周围，见图 7-12：

$$\log_2 R/G \rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A)G) \qquad \text{式 7-5}$$

这里 $c(A)$是 M 对 A 的拟合曲线对应的函数，由于大部分基因是稳定表达的，所以认为少数差异的基因不会影响曲线的拟合。

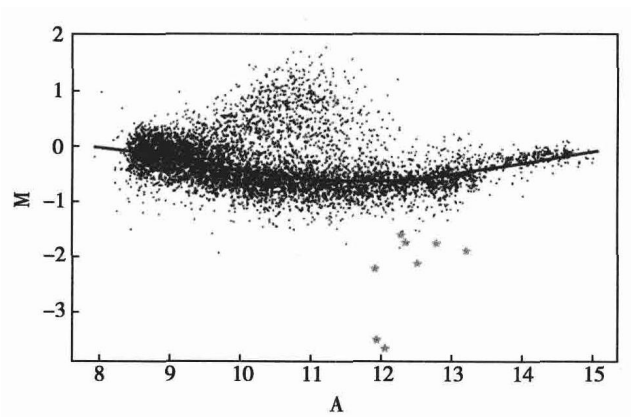


图 7-11 荧光强度依赖的标化前的 A-M 散点图
(引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.)

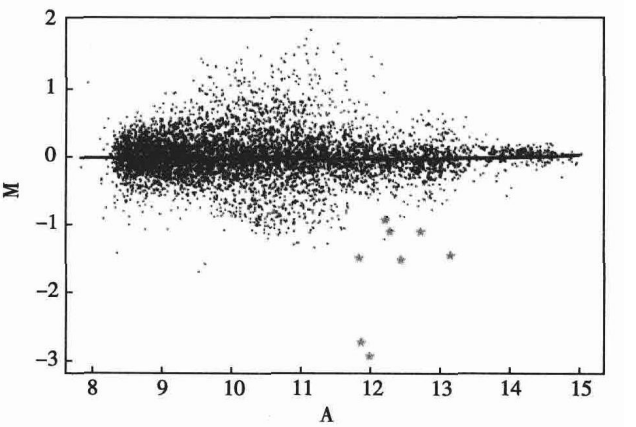


图 7-12 荧光强度依赖的标化后的 A-M 散点图
(引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.)

点样针标化(within-print-tip-group normalization)：一张芯片可以分成几个栅格(grid)，一个栅格内的探针采用同一根点样针点样，不同栅格采用不同的点样针。由于不同点样针针尖的长短粗细、磨损程度等存在细微差异，导致在不同的栅格间存在系统误差。图 7-13 中不同颜色的拟合曲线对应于不同的栅格。

点样针标化实际上是考虑了点样针差异情况下的荧光强度依赖的标化：

$$\log_2 R/G \rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G) \qquad \text{式 7-6}$$

这里 $c_i(A)$指对应于第 i 个栅格的拟合曲线的函数，$i=1,2,\cdots,I$，I 为栅格数。

双参数标化：以上提到的都是单参数标化法，即标化法仅调整了 log-Ratios 值，但是同时人们发现来自不同栅格的基因其 log-Ratios 值具有不同的离散度，即 log-Ratios 值的方差不同。图 7-14 为经过 log-Ratios 值单参数标化后的不同栅格的 log-Ratios 值分布箱式图。

双参数标化法就是兼顾了这两者的标化方法。具体的操作可以有所不同，例如：经过点样针标化法调整后，不同栅格的基因都被调整至峰值对应处的 log-Ratios 值为 0 的水平，然而来自不同栅格的基因的 log-Ratios 值可能具有不同的离散度，可以用求得的每个栅格中基因的 log-Ratios 值的标准差 σ_i 作为尺度，相应的每个基因的 log-Ratios 值除以其所在栅格的尺度就完成了离散度调整的过程。另一种好的方法是通过中位数求得尺度 $\hat{a}_i$，这种方法对于异常或者两端的 log-Ratios 值不敏感。通

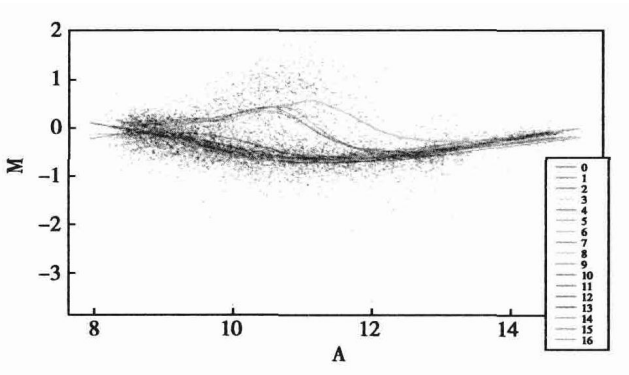


图 7-13 标化前不同栅格的 A-M 散点图

(引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.)

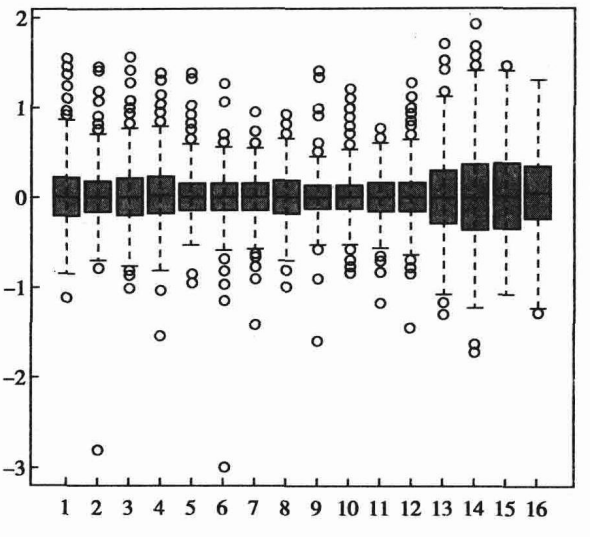


图 7-14 不同栅格的 log-Ratios 值分布盒状图

(引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.)

过一定的数学假设，可以推导出：

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}} \qquad \text{式 7-7}$$

$$MAD_i = median_j\left\{|M_{ij} - median_j(M_{ij})|\right\} \qquad \text{式 7-8}$$

这里 $i=1,2,\cdots,I$，I 为栅格数，j 为基因，$median_j(M_{ij})$ 第 i 个栅格中所有基因的 log-Ratios 值的中位数。求出尺度后就可以作相应的纠正了。

(2) 染色互换标化(Paired-slides normalization, dye-swap)：这种标化方法被应用在特殊的实验设计——染色互换芯片实验中，实验设计如下所示：

	实验组	对照组
芯片 1	cy3	cy5
芯片 2	cy5	cy3

即与普通的 cDNA 芯片相比，每张芯片都会作相应的重复实验，除了实验组和对照组的染色作互换以外，其他的实验条件都保持不变。

这样对于芯片 1，采用 $\log_2 R/G-c$ 作标化，而对于芯片 2，采用 $\log_2 R'/G'-c'$ 作标化。这里 c 和 c' 分别表示标化函数，它可以由上面提及的任何一种片内标化方法获得。由于这种特殊的实验设计，那么结果标化以后的 *log-Ratios* 值应该满足以下等式：

$$\log_2 R/G - c \approx -(\log_2 R'/G' - c') \qquad \text{式 7-9}$$

由于芯片 1 和芯片 2 实验是在两种相同的实验条件下进行的，所以假定 $c \approx c'$，那么标化函数 c 的求法就可以写作：

$$c \approx \frac{1}{2}\left[\log_2 R/G + \log_2 R'/G'\right] = \frac{1}{2}(M + M') \qquad \text{式 7-10}$$

染色互换的标化方法简单，但是相对其他的实验设计该方法的成本翻了一倍，另外在作 $c \approx c'$ 的前提假设时，一定要根据实验获得的数据作相应的分析，如图 7-15，黑色和蓝色的散点分别来自于两张重复实验的芯片，只有当两种散点的拟合曲线相似时才支持假设 $c \approx c'$，从而才能运用此种标化法

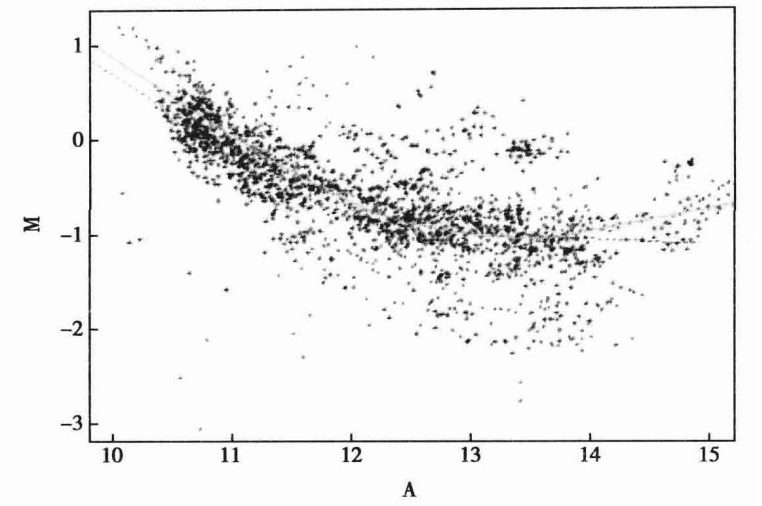


图 7-15 染色互换实验 A-M 散点图比较

（引自：Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4):e15.）

进行标化。

（3）片间标化：线性标化法（linear Scaling methods）不管采用何种片内标化法处理，log-Ratios 值的峰值将会移至 0 处。片间标化的目的是去除不同芯片间的系统误差，使片间的 log-Ratios 值具有可比性。来自不同芯片的基因的 log-Ratios 值具有不同的离散度，所以片间标化可采用上述在双参数法中提到的离散度标化。通过求 $\hat{a}_i$ 或 σ_i 进行离散度标化的方法都是线性标化法。

非线性标化法（Non-linear methods），例如，sACE（simulataneous alternating conditional expectation），该方法通过对芯片数据进行非线性转换优化数据，优化目标是最大化两张重复芯片的相关性，所以这种非线性标化法尤其适合于重复实验。

分位数标化法（quantile normalization）的目的是让所有芯片中的每张芯片所测的数据都具有相同的分布。这种标化法来自于 quantile-quantile plot 思想，即如果 quantile-quantile plot 在一条对角线上则两个数据向量的分布相同，否则不同。这种思想可以延伸至处理 n 个数据向量，那么 n 个数据向量的分位数在 n 维空间中可用单位向量$(\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}})$表示，这说明 n 个数据向量具有相同的分布。

令 $q_k=(q_{k1},q_{k2}\cdots q_{kn})$为 n 张芯片的 k 分位数向量，这里 $k=1,2,\cdots,p$。$d=(\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}})$为单元对角阵。为了将 n 张芯片的 k 分位数向量通过某种转换排列在对角线上，可以作如下的 q_k 到 d 的映射：

$$proj_d q_k=(\frac{1}{n}\sum_{j=1}^{n}q_{kj},\cdots,\frac{1}{n}\sum_{j=1}^{n}q_{kj}) \qquad \text{式 7-11}$$

这表明采用 k 分位数的均值代替原始数据就能够保证每张芯片具有相同的分布。

具体的算法如下：

1）给定长度为 p（p 个分位数）的 n 个数组（n 张芯片）构成 $p\times n$ 的矩阵 X；

2）矩阵 X 的每列都作排序后构成新矩阵 X_{sort}；

3）求得 X_{sort} 每一行的均值，并将每一行均值分配给该行中的每个元素，得到 X'_{sort}；

4）将 X'_{sort} 的每列按照原始矩阵 X 的顺序重新排序得到标化后的矩阵 $X_{normalized}$。

（二）单通道芯片

1. 系统误差来源　单通道芯片不同于 cDNA 芯片，它不是两组样本的竞争性杂交，而只是采用一种染料标记后的一组样本与芯片上探针的杂交，所以它不存在 cDNA 芯片中所涉及的染料偏倚；

另外这种芯片采用的是原位合成法非点样法，所以它也不存在 cDNA 芯片中不同栅格带来的系统误差。cDNA 芯片通常将标化的过程分为两大块：片内标化和片间标化，片内标化主要是为了纠正染料和点样针带来的系统误差，而对于单通道芯片不存在这两种系统误差，所以对于此类芯片的标化没有片内标化和片间标化的明确区分。此类芯片的系统误差主要是由不同芯片间的差异带来的。

2. 标准化过程的参照物　标化过程中使用的参照物跟 cDNA 芯片大致相同。

3. 标准化方法　单通道芯片要比 cDNA 芯片复杂，它设计了两类探针：与目标样本完美匹配(perfect match，PM)的探针及对应的在完美匹配的探针序列中央发生一个碱基替换(mismatch，MM)后的探针，这两类探针构成了一个探针对。对于一个基因而言，通常会设计 16~20 个这种探针对，使它们构成一个探针集。所以对于单通道芯片，除了标准化处理外还要基于探针集进行汇总分析得出基因转录物表达的信号估计。理论上 PM 的荧光强度应高于 MM，实际上由于存在 MM 的探针可能会与其他的转录物进行杂交，导致 MM 高于 PM，这种情况下的 MM 值没有意义，即无信息。图 7-16 是一个 10 个探针对组成的探针集的示意图。

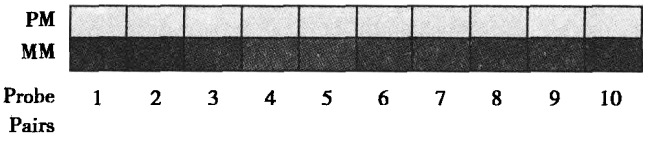


图 7-16　探针集杂交结果示意图

每个探针集中的探针将共同决定某基因杂交信号，包括定性和定量的。定性的信号包括有 Present、Marginal 和 Absent，定量的信号为该基因实际的荧光强度值(Real Signal)。不管是定性还是定量的信号都是综合了该基因对应的所有探针对的结果，表示该基因在某种条件下的表达情况。

单通道芯片的标化方法与 cDNA 芯片的标化方法有所差别，但它不是应用于两种不同的染色通道，而是应用于两张芯片。

第四节　差异表达分析

Section 4　Analysis of Differentially Expression Gene

标准化处理就是要过滤非生物学来源的混杂变异，即经过这种标化处理以后是否可以发现真正的生物学变异，并不至于把非生物学变异归为生物学变异，即差异表达基因和非差异表达基因的识别。差异基因的筛选方法有很多，最简单的是阈值法，用倍数分析基因表达水平差异，即计算基因在两个条件下表达水平的比值，确定比值的阈值，将绝对值大于此阈值的基因判断为差异基因，这种阈值法比较武断，人为因素太大且不严谨。另有些方法包括统计学的 t 检验法、方差变异模型和 SAM 等方法。

一、倍数法

运用倍数 f 值估计每个基因在实验条件下较之对照条件下表达量的倍数差异值。阈值的确定具有一定的困难。

$$f = \frac{x_1}{x_c} \qquad \text{式 7-12}$$

当 f 值约等于 1 时，表明该基因在两种不同条件下的表达没有差异，反之，当 f 值明显大于 1 或小于 1 时，表示基因在条件 I 下的表达有上调或下调。f 值越偏离 1，差异表达越显著。但是对于不同的数据集，具体阈值的确定很困难，通常以 2 倍差异为阈值，但这带有很大的人为因素，不具有统计学意义。在芯片数据分析的早期被应用，目前通常被用于基因的大规模初筛。

二、t 检验法

运用 t 检验法可以判断基因在两种不同条件下的表达差异是否具有显著性。

零假设为 $H_0:\mu_1=\mu_2$，即假设某基因在两种不同条件下的平均表达水平相等，与之对应的备选假设是 $H_1:\mu_1\neq\mu_2$。t 检验的计算公式为：

$$t=\frac{\overline{x_1}-\overline{x_2}}{\sqrt{s_1^2/n_1+s_2^2/n_2}} \qquad \text{式 7-13}$$

其中均值

$$\overline{x_i}=\sum_{j=1}^{n_i}x_{ij}/n_i \qquad \text{式 7-14}$$

方差

$$s_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{x_i})^2 \qquad \text{式 7-15}$$

n_i 为某一条件下的重复实验次数，x_{ij} 为某基因在第 i 个条件下第 j 次重复实验的表达水平测量值。根据统计量 t 值，得到 p 值，设定假设检验水准 α，若 $p<\alpha$，则拒绝零假设，认为某基因在两不同条件表达差异具有统计学意义；反之，则接受零假设，认为某基因在两不同条件下表达无差异。

由于芯片实验成本较高，n_i 较小，从而对总体方差的估计不很准确，t 检验的检验效能降低。

为了解决这个问题，随机的方差模型法对总体方差的估计进行了修改。这种模型的前提假设为，不同的基因具有不同方差，但这些方差可以看作是来自同一分布的独立样本，方差的倒数满足参数为 a，b 的 λ 分布，其中 $1/ab$ 为期望方差，那么 t 统计量的计算公式中的分母，即合并方差的估计修改为：

$$s^{2'}=\frac{(n_1+n_2-2)s^2+2a(1/ab)}{(n_1+n_2-2)+2a} \qquad \text{式 7-16}$$

其中

$$s=\sqrt{s_1^2/n_1+s_2^2/n_2} \qquad \text{式 7-17}$$

三、方差分析

方差分析可用于基因在两种或多种条件间的表达量的比较，它将基因在样本之间的总变异分解为组间变异和组内变异两部分。组间变异体现了不同条件带来的基因表达的差异，组内变异体现了包括个体差异和测量带来的随机误差。通过方差分析的假设检验判断组间变异是否存在，如果存在则表明基因在不同条件下的表达有差异。分别计算总变异、组间变异和组内变异：

$$SS_{总}=\sum_i\sum_j(x_{ij}-\bar{x})^2 \qquad \text{式 7-18}$$

$$SS_{组间}=\sum_i n_i(\bar{x_i}-\bar{x})^2 \qquad \text{式 7-19}$$

$$SS_{组内}=\sum_i\sum_j(x_{ij}-\bar{x_i})^2 \qquad \text{式 7-20}$$

其中 x_{ij} 为某基因在第 i 种条件第 j 个样本中的表达值；$\bar{x}$ 为该基因在所有样本中的平均表达值；$\bar{x_i}$ 为该基因在第 i 种条件下样本中的平均表达值，n_i 为该条件下的样本数。

将变异除以自由度计算均方，消除了自由度的影响：

$$MS_{组间}=\frac{SS_{组间}}{v_{组间}} \qquad \text{式 7-21}$$

$$MS_{组内}=\frac{SS_{组内}}{v_{组内}} \qquad \text{式 7-22}$$

$$F=\frac{MS_{组间}}{MS_{组内}} \qquad \text{式 7-23}$$

其中 $v_{组间}=k-1$，$v_{组内}=N-k$，$v_{总}=N-1$，N 为样本的总个数，k 为条件数。

根据统计量 F 值，得到 p 值。设定假设检验水准 α，若 $p<\alpha$，则拒绝零假设，认为某基因在不同条件下的表达差异具有统计学意义；反之，则接受零假设，认为某基因在不同的条件下表达无差异。

四、SAM 法

在进行统计学假设检验时，最后作出的推断结论不管是拒绝 H_0 或是不拒绝 H_0，均可能发生错误，即Ⅰ型错误或Ⅱ型错误。Ⅰ型错误（假阳性）即在假设检验作推断结论时，拒绝了实际上正确的检验假设 H_0，即将无差异表达的基因判断为差异表达。Ⅱ型错误（假阴性）即接受了实际上不正确的 H_0，即将有差异表达的基因判断为无差异表达。

在运用 t 检验和方差分析进行差异基因筛选时，存在多重假设检验的问题。若芯片检测了 n 个基因，整个差异基因筛选过程需要做 n 次假设检验，若每次假设检验犯假阳性的概率为 p，则在这个差异基因筛选过程中至少有一个基因是假阳性的概率为 $P=1-(1-p)^n$，由于芯片检测的基因个数 n 较大，从而导致假阳性率 P 的增大。对于这种多重假设检验带来的放大的假阳性率，需要进行纠正。常用的纠正策略有 Bonferroni 校正，控制 FDR（false discovery rate）值等。

SAM（significance analysis of microarrays）算法就是通过控制 FDR 值纠正多重假设检验中的假阳性率。计算相对差异统计量 d：

$$d=\frac{\overline{x_1}-\overline{x_2}}{s+s_0} \qquad \text{式 7-24}$$

统计量 d 衡量了基因表达的相对差异，是 t 统计量的修正。

计算所有基因的 d 值，这些 d 值的分布应该独立于基因的表达水平。然而在低表达丰度情况下，由于 s 值较小，d 值的方差较大。为了确保 d 值的方差独立于基因表达水平，在分母上加上一个小的正常量 s_0。通过窗口法确定 s_0 值，该 s_0 值能使 d 值的变异系数最小。

扰动实验条件，模拟基因在两组间无表达差异的表达向量，计算扰动后的基因表达的相对差异统计量 d_p，随机扰动 $|P|$ 次，计算所有扰动的平均相对差异统计量，见图 7-17。

$$d_E=\frac{1}{|P|}\sum d_p \qquad \text{式 7-25}$$

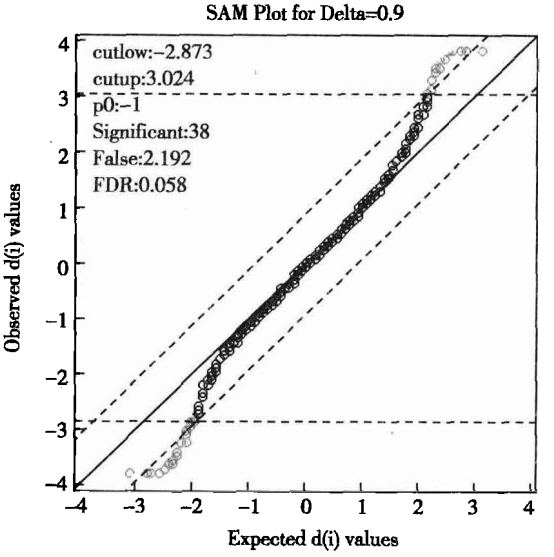


图 7-17 d 对 d_E 散点图

当 FDR=0.058 时，阈值大概在±3 外，落在阈值以外的绿色标记的基因即为差异表达基因

（引自：Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad .Sci. U S A .,2001, 98(9):5116-5121.）

确定差异表达基因阈值：以最小的 d 正值和最大的 d 负值作为统计阈值 $d(t)$，运用该阈值，统计在 d_E 值中超过该阈值的假阳性基因个数，估计假阳性发现率 FDR(false discovery rate，FDR)值，FDR 值为在所有判断为差异表达的基因中假阳性基因的比例。

$$FDR = \frac{\sum \frac{\#of(d_p > d(t))}{|P|}}{\#of(d \geq d(t))} \qquad \text{式 7-26}$$

通过调整 FDR 值的大小得到差异表达的基因。

五、信息熵

与上述差异基因筛选方法不同的是运用信息熵进行差异基因筛选时，不需要用到样本的类别信息，所以运用信息熵找到的差异基因并非指在两种不同条件下表达有差异的基因，而是指在所有条件下表达波动比较大的基因。

首先对每个基因进行离散化处理，然后计算该基因的信息熵。

$$H = -\sum_{i=1}^{m} p_i \log p_i \qquad \text{式 7-27}$$

其中 p_i 表示某个基因表达值在某一段取值的概率(这里用某一段的频数值近似代替概率值)，m 为离散的区段数。H 值越高，说明该基因在这些条件下表达值的变异程度越大，揭示该基因为差异表达基因。

第五节 基因芯片数据的聚类分析

Section 5 Cluster Analysis of Microarray Data

知 识 拓 展

癌症异质性是当前癌症诊断与治疗所面临的最大困难，在临床上同一种癌症经过相同的治疗，其效果与预后往往有明显差异，其主要原因之一是对癌症异质性(heterogeneity)的认识还十分有限。到目前为止，运用传统的聚类分析方法，癌症的异质性已经在 mRNA 表达层面得到证实，相关的研究包括对淋巴瘤、乳腺癌、肺癌和前列腺癌等亚型的分析。然而癌症的异质性程度远远不仅于此，甚至在同种肿瘤组织中，癌细胞的大小、形态、抗原表达、增殖能力、细胞间互作和治疗敏感性等方面都存在巨大的差异。目前表观遗传学的研究，以及癌症干细胞领域的研究可能为癌症的诊断和治疗提供新的思路。

无监督的聚类分析是基于研究对象属性的相似性对研究对象进行分组，使组内样本相似，组间样本有差异。

当聚类分析应用于基因表达谱数据分析时，可以解决以下两方面的问题：第一，如果将研究对象定为样本，则衡量样本相似性的属性即是基因，基于基因表达的相似性可以将 mRNA 表达相似的样本聚为一类。对样本进行聚类可以进行实验样本的质量控制，即检测实验样本的杂交效能；检查样本根据它们的已知类别是否聚到一处；识别样本的新亚型。其中识别样本的新亚型是聚类分析应用于肿瘤样本的最重要的用途。肿瘤为高度异质性的疾病，基于临床病理学诊断判断为相同的肿瘤往往具有不同的分子机制，运用基因芯片数据的聚类分析可以进一步区分肿瘤的新亚型，从而开发肿瘤的个性化诊疗新途径。第二，如果将研究对象定为基因，则衡量基因相似性的属性即是样本，基于基因在样本空间中表达的相似性可以将基因进行聚类，基因“类”通常涉及功能上相关的基因，或参与同一个代谢通路，或编码蛋白质复合物的成分等。聚在同一类的基因可以找到共表达模式的分子

机制，如基因上游保守序列分析，进一步构造基因调控网络模型。

聚类分析中最主要的两个因素是评价研究对象相似性程度的距离（或相似性）尺度和将研究对象分组的聚类算法。

一、聚类分析中的距离（相似性）尺度函数

距离尺度函数的选择取决于研究者想发现哪种类型的关系。常用的表达相似性尺度有几何距离、线性相关系数、非线性相关系数和互信息等。

（一）几何距离

几何距离可以衡量研究对象在空间上的距离远近关系，如图7-18所示，空间上相近的物体运用几何距离可以判断为同一类，而空间上较远的物体则判断为不同类。

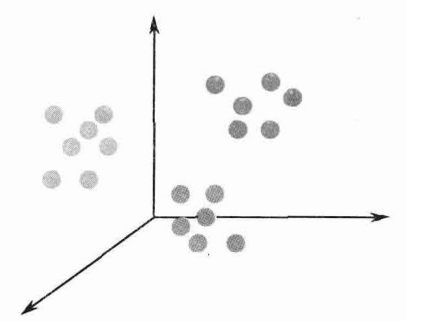


图7-18 基于几何距离衡量的物体在空间上的相似性

常见的几何距离函数有明氏距离：

$$d(x,y)=\{\sum|x_i-y_i|^{\lambda}\}^{\frac{1}{\lambda}} \qquad \text{式 7-28}$$

其中 x 和 y 分别为样本向量或基因向量，x_i 和 y_i 为对应的第 i 个分量，明氏距离通过综合考查各分量的差异来衡量两物体的远近关系。

当 $\lambda=1$ 时，明氏距离即为马氏距离（*Manhattan*）；

当 $\lambda=2$ 时，明氏距离即为欧氏距离（*Eulidean*）；

当 $\lambda=\infty$ 时，明氏距离即为切氏距离（Chebyshev），即：

$$d(x,y)=\max_i|x_i-y_i| \qquad \text{式 7-29}$$

明氏距离在考查两物体的相似性时没有考虑不同分量量纲差异的影响，所以用明氏距离作相似性尺度时应该先对数据进行标准化处理，消除不同分量之间的量纲差异。

Camberra 距离则不需要考查各分量量纲差异的影响：

$$d(x,y)=\sum_{i=1}^{p}\frac{|x_i-y_i|}{|x_i+y_i|} \qquad \text{式 7-30}$$

（二）线性相关系数

几何距离比较适合于衡量样本间的相似性，或者基因在样本空间（如不同组织间）的相似性。当基因表达数据是一系列时间点数据时，运用几何距离会丢失重要信息。如图7-19所示，图中描述了三个基因在五个时间点的基因表达水平波动，如果用几何距离进行衡量，则基因2和基因3相似性高，而基因1在空间上离基因2和基因3较远会判断为相似性差。实际上，基因1的表达水平在不同时间点与其他两基因具有相似的波动趋势，很有可能与基因2和基因3具有功能上的相关性，但是用欧氏距离找不到这种具有生物学意义的基因关系。

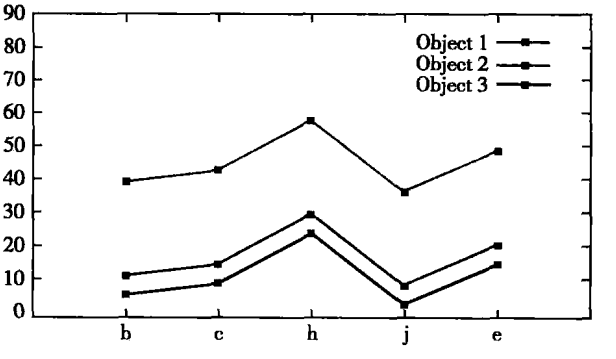


图7-19 三基因在五时间点的表达值波动图

（引自：Haixun Wang WW, Jiong Yang and Philip S.Yu: Clustering by Pattern Similarity in Large Data Sets. In: International Conference on Management of Data 2002: 2002: 394-405.）

运用皮尔森相关系数则可以发现这种在系列时间点上具有相似波动模式的相关关系。

$$r=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) \qquad \text{式 7-31}$$

其中$\bar{x}$为基因向量x的期望值，σ_x为x的标准差；$\bar{y}$为基因向量y的期望值，σ_y为y的标准差，n为向量的维数，即时间点数。

根据皮尔森相关系数的正负号，分别可以发现正相关和负相关的基因相关关系(图 7-20，图 7-21)，而负相关的相关关系如果用几何距离进行衡量，往往会因为距离上较远而忽略这种关系。

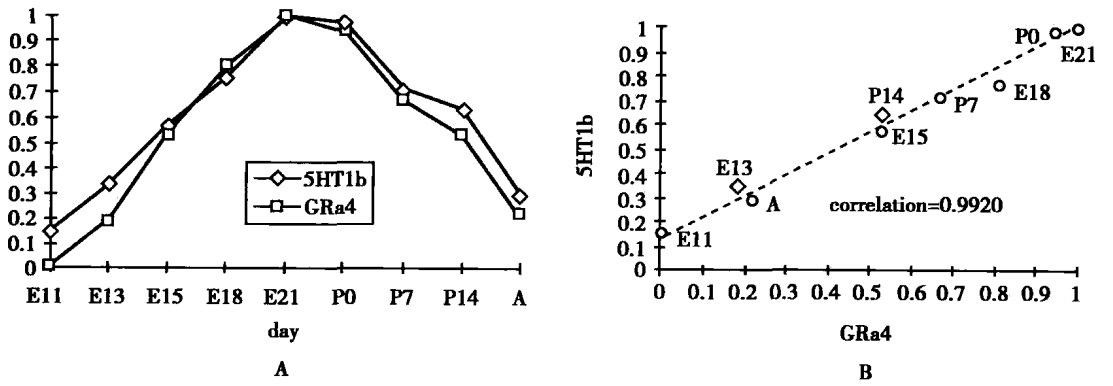


图 7-20 基因间正相关的线性相关关系

(引自：Haixun Wang WW, Jiong Yang and Philip S.Yu: Clustering by Pattern Similarity in Large Data Sets. In: International Conference on Management of Data 2002; 2002: 394-405.)

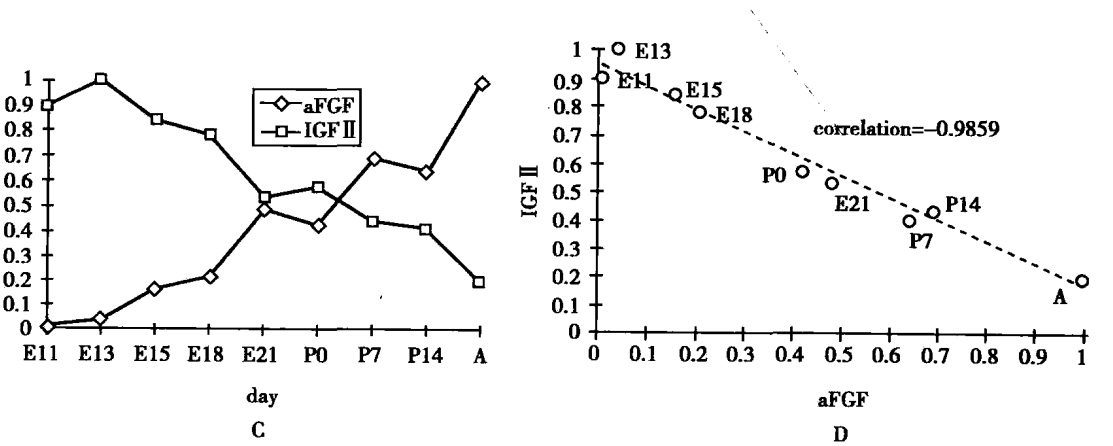


图 7-21 基因间负相关的线性相关关系

(引自：Haixun Wang WW, Jiong Yang and Philip S.Yu: Clustering by Pattern Similarity in Large Data Sets. In: International Conference on Management of Data 2002; 2002: 394-405.)

(三) 非线性相关系数

某些在功能上有相关关系的基因虽然在表达上不具有严格的线性相关关系，但在时间点的波动趋势上却是相似的。如图 7-22 所示，两基因的表达具有同升或同降变化趋势，但明显不具有线性相关关系。

这种非线性相关关系模式可以用斯皮尔曼秩相关系数进行衡量：

$$\gamma = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad \text{式 7-32}$$

其中d为每对观察值x_i与y_i的秩次之差，n为时间点数。

(四) 互信息

线性与非线性相关系数都只能衡量基因间的单调相关关系，对于那些在整个时间序列上基因间的表达没有单调升降关系的，如图 7-23 所示。在前阶段两基因间是正相关关系，而在后阶段两基因间是负相关关系，两基因间的关系具有非单调性的特点。

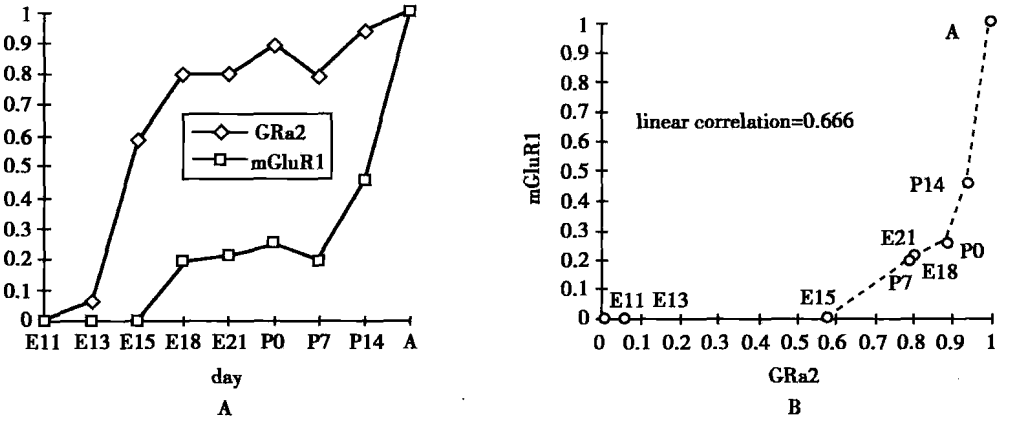


图 7-22 基因间非线性相关关系

（引自：Haixun Wang WW, Jiong Yang and Philip S.Yu: Clustering by Pattern Similarity in Large Data Sets. In: International Conference on Management of Data 2002; 2002: 394-405.）

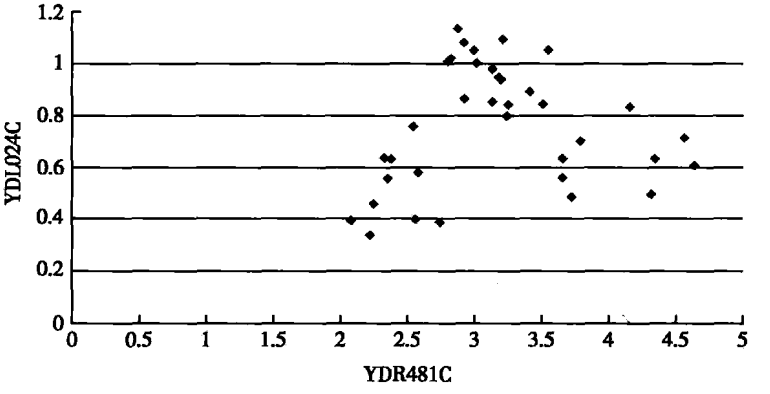


图 7-23 基因间的非单调相关关系

（引自：Wang H, Wang Q, Li X, Shen B, Ding M, Shen Z: Towards patterns tree of gene coexpression in eukaryotic species. Bioinformatics 2008, 24(11):1367-1373.）

对于这种非单调的表达相似关系，可以用互信息进行衡量：

$$\gamma = H(x) - H(x \mid y) \qquad \text{式 7-33}$$

其中 $H(x)$表示 x 的熵，$H(x|y)$表示 x 的条件熵。当 x 和 y 为离散型向量时，条件熵的计算方式为：

$$H(x \mid y_J) = -\sum_{I=1}^{n} p(x_I \mid y_J) \log p(x_I \mid y_J) \qquad \text{式 7-34}$$

$$H(x \mid y) = -\sum_{I=1}^{n}\sum_{J=1}^{m} p(y_J) p(x_I \mid y_J) \log p(x_I \mid y_J) \qquad \text{式 7-35}$$

$p(\cdot)$为概率密度函数，可以由频数估计。n 和 m 分别为离散化 x 和 y 时的离散化单位。在计算互信息时采用的离散化方式会造成一定的信息损失，一般离散化单位的估计由向量 x 和 y 的长度决定。

$$n \leqslant \log_2 size(x) \qquad \text{式 7-36}$$

$$m \leqslant \log_2 size(y) \qquad \text{式 7-37}$$

（五）其他

对于某些具有时间延迟模式的基因相似性关系，需要采用动态规划等算法去发现。

二、聚类分析中的聚类算法

聚类算法主要包括有：分割算法(如 k 均值聚类、SOM 聚类等)、分层算法(如层次聚类等)、基于密度算法、基于网格算法等。这里主要介绍基因芯片数据中常用的层次聚类、k 均值聚类、SOM 聚类，以及基于子空间内的相似性进行基因和样本耦合的双向聚类算法。

(一) 层次聚类

层次聚类算法将研究对象按照它们的相似性关系用树形图进行呈现，图 7-24 呈现的是白血病的两种亚型的层次聚类图。进行层次聚类时不需要预先设定类别个数，树状的聚类结构可以展示嵌套式的类别关系。

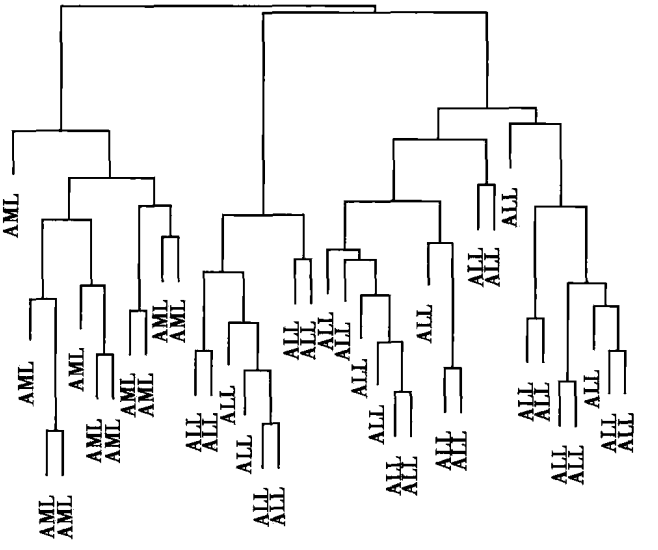


图 7-24 树状层次聚类图

(引自：Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999, 286(5439):531-537.)

根据层次的形成方式，层次聚类可以分为凝聚法(agglomerative)和分裂法(division)。凝聚法，也称为自底向上的方法，一开始将每个研究对象作为单独的一个类，然后不断地合并相近的对象或类。分裂法，也称为自顶向下的方法，一开始将所有的对象置于一个类中，然后一个类被不断地分裂为更小的类。

在层次聚类中，类的合并和分解按照一定的距离函数度量。在对含非单独对象的类进行合并或分裂时，常用的类间度量方法有：最小距离(single linkage)、最大距离(complete linkage)、平均距离(average linkage)和质心距离(centroid linkage)。如图 7-25 所示，最小距离以两类间距离最近的两对象的距离作为两类的距离；最大距离以两类间距离最远的两对象的距离作为两类的距离；平均距离遍历两类中所有对象之间的距离，然后取平均值作为两类的距离；质心距离为分别计算两类的质心，然后以质心间的距离作为两类的距离。

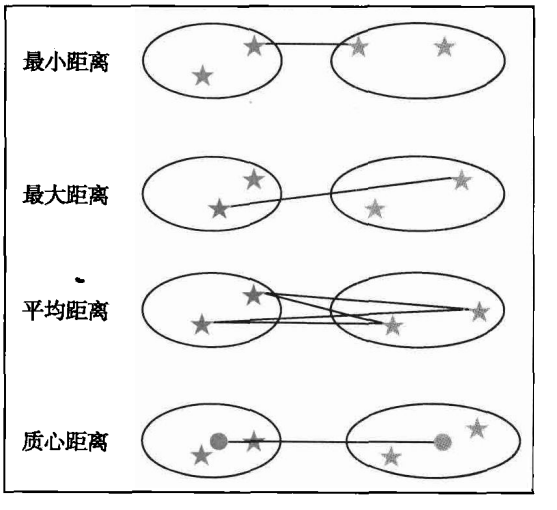
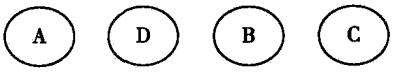


图 7-25 类间相似性度量方法

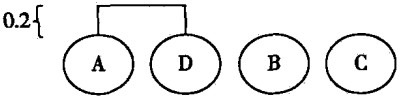
下面以一个例子说明自底向上的层次聚类算法的过程，该算法采用了欧氏距离衡量样本间的相似性，最小距离衡量待合并的两类间的相似性。

1. 设有四个样本 A、B、C 和 D，每个样本自成一类，运用欧氏距离计算它们两两之间的相似性得出距离矩阵。



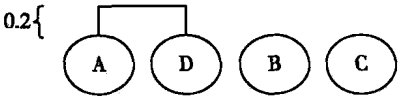
距离	A	B	C	D
A		2	0.7	0.2
B			1	2.5
C				0.3
D				

2. 由于 A 与 D 样本的距离最小，最先合并 A 与 D 样本。



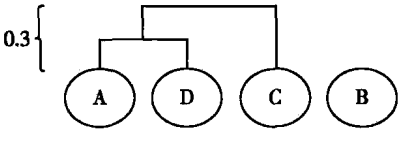
距离	A	B	C	D
A		2	0.7	(0.2)
B			1	2.5
C				0.3
D				

3. 合并后的类别数为三类，调整距离矩阵，即分别运用最小距离法计算 B 样本、C 样本与 AD 类的距离。



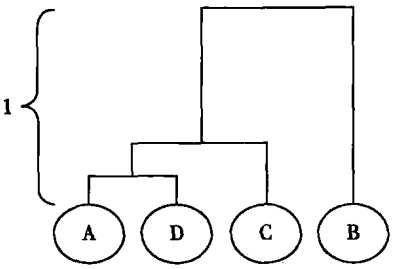
距离	AD	B	C	
AD		2	0.3	
B			1	
C				

4. 基于新的距离矩阵，需合并 AD 类与 C 样本。



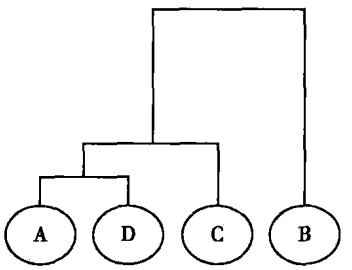
距离	AD	B	C	
AD		2	(0.3)	
B			1	
C				

5. 继续调整距离矩阵，目前的类别数是两类。



距离	ADC	B		
ADC		1		
B				

6. 合并 ADC 类与 B 样本，得出最后的树状图。



距离	AB CD			
ABCD				

7. 根据聚类结果和表达值可以用 treeview 等软件生成可视化的聚类结果，从而对聚类结果有直观认识。图 7-26 中红色表示基因上调，绿色表示基因下调。

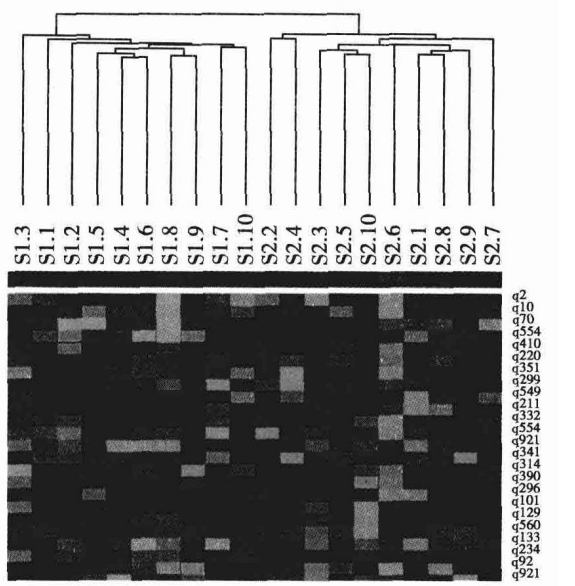


图 7-26 基因表达谱数据聚类结果可视化

下面介绍一个基于基因表达谱芯片数据，运用层次聚类算法发现肿瘤亚型的典型例子。弥漫性大 B 细胞淋巴瘤(DLBCL)是非霍奇金淋巴瘤的一种常见亚型，具有临床异质性，40% 病人治疗效果好、生存期长，其他的 60% 则相反，这种临床上的异质性可能反映了肿瘤在分子层面的致病机制差异。2000 年 Alizadeh 等运用基因芯片数据，基于层次聚类算法证实了 DLBCL 肿瘤病人在 mRNA 层面确实存在两种亚型(图 7-27)。

第一种亚型表达的基因具有生发中心 B 细胞中基因表达的特点，命名为“germinal centre B-like DLBCL(GC B-like DLBCL)”，第二种亚型表达的基因通常有体外诱导激活的外周血 B 细胞基因表达的特点，命名为“Activated B-like DLBCL”。子图 A 表示运用生发中心 B 细胞的表达特征可区分两种亚型；子图 B 为在大规模基因表达层面上发现分别在 GC B-like DLBCL 和 Activated B-like DLBCL 亚类中选择性表达的基因(橙色、蓝色)，橙色区域的基因为生发中心 B 细胞中表达的基因，蓝色区域的基因为体外诱导激活的外周血 B 细胞中表达的基因。子图 C 为仅采用这些选择性表达的基因进行层次聚类的结果。GC B-like DLBCL 亚型较之 Activated B-like DLBCL 亚型具有明显长的生存期，与临床上的差异非常吻合。从而基因表达层面的分子分型识别了具有临床意义的 DLBCL 肿瘤亚型，为 DLBCL 的个性化诊疗提供了新途径。

(二) k 均值聚类

k 均值聚类是根据聚类中的均值进行聚类划分的分割算法。具体的分析流程(图 7-28)为：

1. 初始化类中心，随机选定 k 个类中心，例如可选取 k 个研究对象作为类中心。

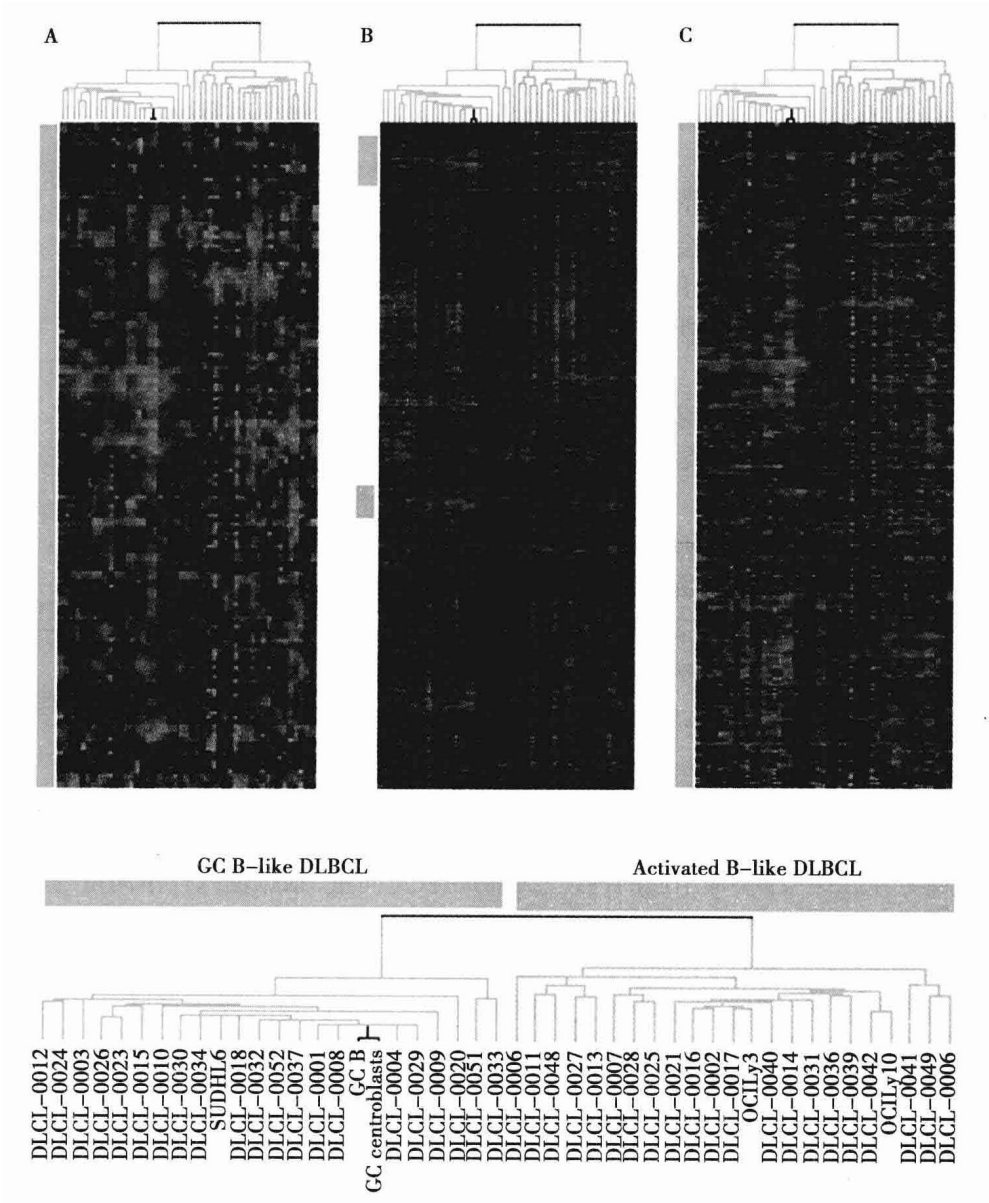


图 7-27 DLBCL 在基因表达层面的分子分型

（引自：Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X et al: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature.2000. 403(6769):503-511.）

2. 计算每个对象与这些类中心的距离，并根据最小距离重新对相应对象进行划分。

3. 重新计算每类样本的均值，作为更新的类中心。

4. 循环上述流程 2 至 3，直到每个聚类不再发生变化。

k 均值聚类可以看作是个优化问题，它的优化目标是最小化类内样本两两间的距离之和：

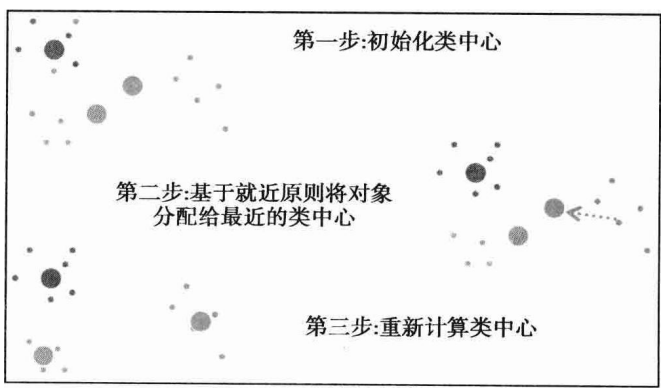


图 7-28 k 均值聚类的分析流程

$$w(C)=\frac{1}{2}\sum_{c=1}^{k}\sum_{C(i)=C(j)=c}d_E(x_i,x_j)^2 \qquad \text{式 7-38}$$

这里 x_i 和 x_j 分别是属于同一个类别中的样本，$d_E(\cdot)$为欧氏距离函数，$C(i)$和 $C(j)$分别是样本 x_i 和 x_j 的类别，k 为类别数，C 为类结构。

k 均值聚类算法的聚类结果依赖于初始化的类中心，选取不同的类中心可能会有不同的聚类结构。为了克服这个问题，可以采用多个初始化方式，选定具有最小 $w(C)$对应的聚类结果作为最佳的类结构。

另外，k 均值聚类需预先指定类别个数，但是很多情况下实际上不知道真正的类别数，一些启发式的方法可以帮助确定 k 的取值。例如，假设有八个研究对象，遍历八个对象可能的聚类类别数，每选定某个 k 的取值时，得到 $w(C)$值，根据 $w(C)$值的变化确定最佳的类别数，如图 7-29 所示，箭头所指处为最佳的类别数，因为当 k 为 4 时，$w(C)$值有明显的下降。

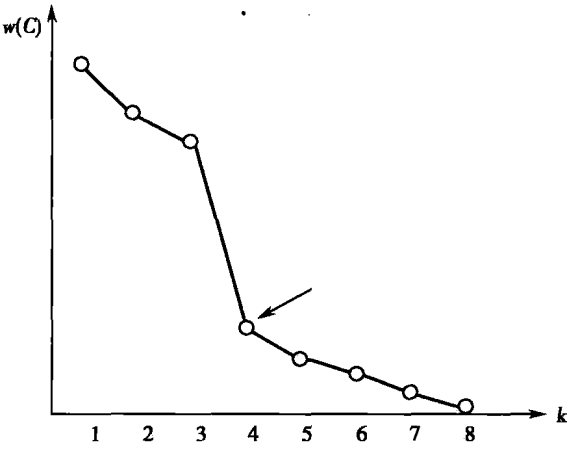


图 7-29 k-$w(C)$波动图确定最佳的类别数

（三）自组织映射聚类

自组织映射聚类(SOM-Self Organization Mapping)与 k 均值聚类相似，也属于分割算法，需要预设类别个数。如图 7-30 所示，在 SOM 神经网络中，预设类别个数为 6，输出层的神经元 1、2、3、4、5、6 以栅格方式排列于二维空间，输出层的神经元有初始权重向量，根据输入样本向量与输出层神经元的距离，找到具有最短距离的神经元作为兴奋神经元，其他神经元根据与该兴奋神经元的距离确定不同的兴奋度，然后根据兴奋度的不同对神经元权重进行调整，完成一个学习过程，随着样本的继续输入，不断进行这种学习过程。最后神经元可以根据输入样本向量的特征，以拓扑结构展现于输出空间，如图中黑点表示学习样本，在不断的学习过程中，输出层的神经元根据输入样本的特点进行权重调整，最后拓扑结构发生了改变。

（四）双向聚类

上述的聚类算法都是基于基因表达谱行和列的全局相似性，但是从生物学角度讲，一组基因表达上的相似性可能只限制在某些实验条件内，运用所有实验样本对基因进行聚类会因为引入噪音而影响基因表达相似性的度量，而样本的相似性也常常不需要运用所有基因来计算，例如层次聚类中的 DLBCL 亚型的区别，仅用橙色和蓝色区域基因的子类就可以区别。所以双向聚类的目的就是识别基因表达谱矩阵中同质的子矩阵(图 7-31)，运用特定的基因子类识别样本子类。

下面介绍一种双向聚类方法，寻找疾病样本和致病基因簇之间的对应关系，该方法按样本和基因两个方向同时进行迭代聚类。

设基因表达谱矩阵 M，定义初始的样本集和基因集分别为 S_1 和 G_1，$S_j(G_i)$表示以 G_i 为特征对样本集 S_j 聚类的结果。同理，$G_i(S_j)$表示以 S_j 为特征对基因集 G_i 聚类的结果。其详细的分析流程如下：

1. 初始化过程　首先以芯片上所有的基因 G_1 为特征，对 S_1 聚类：$S_1(G_1)=(S_j)$，$j=2,3,\cdots$；再利用数据集中所有的样本 S_1 作为特征对所有基因 G_1 进行聚类：$G_1(S_1)=\{G_i\}$，$i=2,3,\cdots$，此时聚类深度(Cluster Depth)为 0。

2. 识别稳定的样本类和基因类　发现稳定的基因簇 $G_i(i=2,3,\cdots)$和稳定的样本子集 $S_j(j=2,3,\cdots)$，进一步计算 $S_j(G_i)$(包括 S_1)和 $G_i(S_j)$(包括 G_1)，这样又得到许多样本子集 $S_j(G_i)$，和基因簇 $G_i(S_j)$，此时聚类深度为 1。

3. 重复步骤 2 过程，直至达到一定的阈值(聚类深度)或没有新的稳定基因簇或样本子集出现。

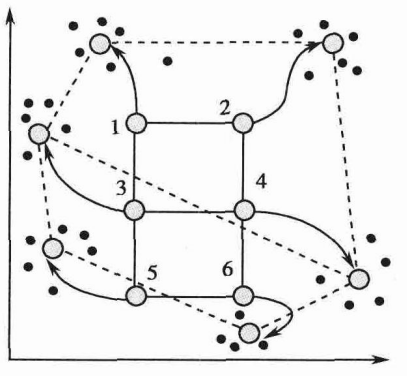


图 7-30 SOM 映射学习过程
（引自：Tamayo et al. 1999）

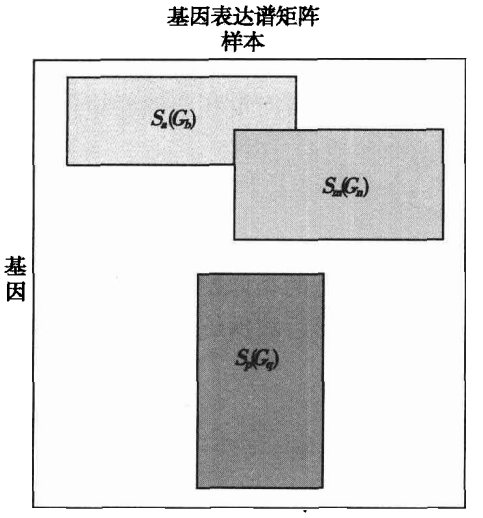


图 7-31 双向聚类识别同质的子结构

总之，聚类分析方法在基因表达谱数据中具有重要的应用，即使没有类别结构的随机样本也可以得到类别结构。一方面聚类证实方法可以检测聚类发现的类别是否为潜在的分组；另一方面，对于基因表达谱数据而言，mRNA 分子层面的分型只有与临床差异相吻合才更具有临床的诊断治疗意义。聚类分析应用于基因表达谱数据，为复杂疾病的亚型识别、致病机制及分子标记的识别提供了有效的工具。

第六节 基因芯片数据的分类分析

Section 6 Classification of Microarray Data

对于基因芯片数据，无监督的聚类分析可同时对样本和基因进行聚类，从而完成不同的分析任务。而有监督的分类分析一般是单向的，即以基因为属性，构建分类模式对样本的类别进行预测。因此，分类分析可以构建 mRNA 分子层面的预测模型，从而为疾病的预测提供新的手段；另外，参与分类模型的基因往往是对样本判别有重要作用的基因，所以在分类过程中还可以同时进行疾病相关基因的挖掘。

常用的分类方法有线性判别分析（如Fisher线性判别）、k近邻分类法、支持向量机（SVM）分类法、贝叶斯分类器、人工神经网络分类法、决策树与决策森林法，以及基因芯片数据分析中常用的 PAM 分类器。下面主要介绍 Fisher 线性判别、k 近邻分类法与决策树。

一、Fisher 线性判别

线性判别函数是最简单的判别函数，相应的分类面是超平面 $g(x)$：

$$g(x)=w^T x+b\begin{cases}>0, L_1\\<0, L_2\end{cases} \qquad \text{式 7-39}$$

其中 w 是分类面的法向量，b 是分类面的偏移，L_1 和 L_2 分别是两类别的类别标签。设计线性分类器的关键是估计 w 和 b，选择 w 就是寻找最佳投影方向，投影后变成一维数据的分类问题，见图 7-32。

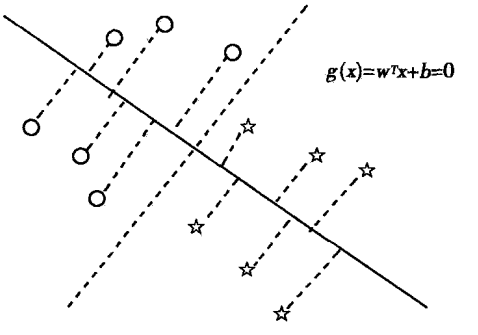


图 7-32 线性判别函数的分类思想

Fisher 线性判别的基本思想是寻找一个最佳的投影方

向，使得样本在投影后的一维空间内满足类间离散和类内紧致的特点，投影后的数据分别运用离散度和均值衡量类内和类间的数据特点。

投影前数据的均值向量和离散度矩阵分别为：

$$m_i = \frac{1}{n}\sum x \quad i=1,2 \qquad \text{式 7-40}$$

$$S_i = \sum((x-m_i)(x-m_i)^T) \quad i=1,2 \qquad \text{式 7-41}$$

其中 m_1 和 m_2 分别是两类原始数据的均值向量；S_1 和 S_2 分别是两类原始数据的离散度矩阵。

原始数据与投影后数据统计量之间的关系是：

$$\mu_i = w^T m_i \qquad \text{式 7-42}$$

$$\begin{aligned}\sigma_i^2 &= \sum(w^T x-\mu_i)^2 \\ &= w^T\sum(x-m_i)(x-m_i)^T w \\ &= w^T S_i w\end{aligned} \qquad \text{式 7-43}$$

其中 μ_1 和 μ_2 分别是两类投影后数据的均值；σ_1 和 σ_2 分别是两类投影后数据的离散度。

Fisher 准则函数为：

$$J_F(w) = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2} \qquad \text{式 7-44}$$

Fisher 准则函数分母衡量了总类内离散度，分子衡量了类间距。找到最佳的投影方向使得 $J_F(w)$ 最大，从而使投影后的样本满足类间离散和类内紧致的特点。

$$w_{opt} = \arg\max J_F(w) \qquad \text{式 7-45}$$

$J_F(w)$ 只与投影方向有关，求解 w 的最优解 w_{opt}，通过一系列的计算得到：

$$w_{opt} = (S_1+S_2)^{-1}(m_1-m_2) \qquad \text{式 7-46}$$

以两类均值的中点作为分类阈值 b：

$$b = -\frac{\mu_1+\mu_2}{2} \qquad \text{式 7-47}$$

或投影后数据的均值作为分类阈值 b：

$$b = -\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2} \qquad \text{式 7-48}$$

对于样本 x，若 $w^Tx+b>0$，则判断为 L_1 类；若 $w^Tx+b<0$，则判断为 L_2 类。

二、k 近邻分类法

k 近邻分类法的分类思想是：给定一个待分类的样本 x，首先找出与 x 最接近的或最相似的 k 个已知类别标签的训练集样本，然后根据这 k 个训练样本的类别标签确定样本 x 的类别。

如图 7-33 所示，三角形样本为待分类的样本 x，当邻居数 k 为 1 时(左图)，与它最近的样本为圆形样本，从而可将圆形样本对应的类别标签赋予 x；当邻居数 k 为 3 时(中图)，与它最近的样本有两个圆形样本，一个星形样本，占多数的圆形样本对应的类别标签赋予 x；当邻居数 k 为 5 时(右图)，与它最近的样本有四个圆形样本，一个星形样本，占多数的圆形样本对应的类别标签赋予 x。

k 近邻分类法的算法步骤为：

1. 构建训练样本集合 X。

2. 设定 k(k 为奇数)的初值。k 值的确定没有一个统一的方法(根据具体问题选取的 k 值可能有较大的区别)。一般方法是先确定一个初始值，然后根据实验结果不断调试，最终达到最优。

3. 在训练样本集中选出与待测样本 x 最近的 k 个样本，假定样本 x 检测的基因个数为 n，即 $x\in R^n$，x_i 为样本 x 的第 i 个基因的表达值，样本之间的“近邻”一般由欧式距离来度量。那么两个样本 x 和

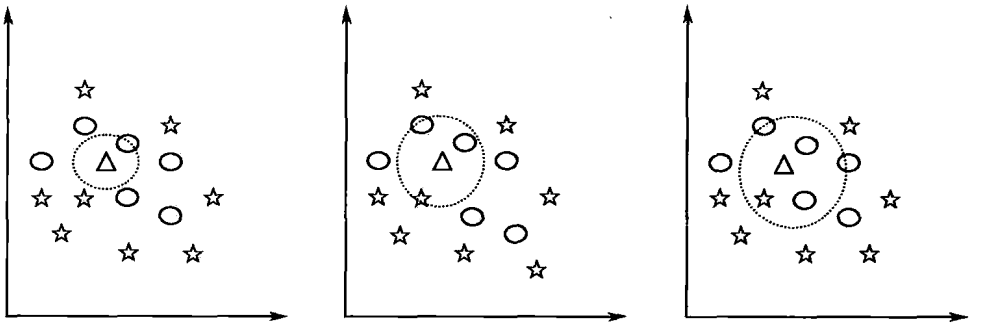


图 7-33 k 近邻分类法的分类思想

y 之间的欧式距离定义为：

$$d(x,y)=\{\sum|x_i-y_i|^2\}^{\frac{1}{2}} \qquad \text{式 7-49}$$

4. 设 $y_1, y_2\cdots y_k$ 表示与 x 距离最近的 k 个样本，k 个邻居中分别属于类别 $L_1, L_2\cdots, L_l, \cdots L_c$ 的样本个数为 $n_1, n_2\cdots, n_l, \cdots n_c$，判别函数 $g_l(x)=n_l$，如果 $g_l(x)=\max_l(n_l)$，则将 x 的类别定为 L_l 类。

5. L_l 即是待测样本 x 的类别。

三、决策树

决策树是一种多级分类器，利用决策树分类可以将一个复杂的多类别分类问题转化成若干个简单的分类问题来解决。决策树分类器呈一个树状的结构，内部节点上选用一个属性进行分割，每个分叉都是分割的一个部分，叶子节点可表示样本的一个分布。

图 7-34 为一棵二叉分支的决策树，根节点 1 中包含 40 个肿瘤样本和 22 个正常样本，运用基因 *M26383* 进行分割，当 *M26383* 的基因表达水平大于 60 时，样本被分至右子节点 3，否则被分至左子节点 2，左子节点中包含 14 个正常样本，0 个肿瘤样本，表示该节点内样本已经分纯，不需要再继续进行分割，定义为叶子节点。节点 3 的样子继续进行分割，运用基因 *R15447* 进行分割，当 *R15447* 的表达水平大于 290 时，样本被分至节点 5，否则被分至节点 4，节点 5 已分纯，不需要再进行分割。节点 4 继续用基因 *M28214* 分割，得到最后两个叶子节点 6 和 7。

所以，构造决策树的方法是采用自上而下的递归分割，采用贪婪算法，从根节点开始，如果训练集中的所有观测是同类的，如都为正常样本，则将其作为叶子节点，节点内容即是该类别标记。否则，根据某种策略选择一个属性（如基因），按照属性的各个取值，把训练集划分为若干个子集合，使得每个子集上的所有例子在该属性上具有同样的属性值。然后再依次递归处理各个子集，直到符合某种停止条件。

在构造决策树的过程中最重要的一点是在每一个分割节点确定选择哪个基因，以及该基因的哪种分割方式对样本进行分割，这需要通过分割准则衡量使用哪个基因更合理。分割准则主要包括有 *Gini* 指数、信息增益等。

1. *Gini* 指数变化（$\Delta Gini$） *Gini* 指数是用来测量节点纯度的指标，对于某节点 N 的 *Gini* 指数定义为：

$$Gini(N)=1-\sum_{j=1}^{k}p_j^2 \qquad \text{式 7-50}$$

其中 p_j 是指第 j 类在某节点中的概率，即某节点中属于第 j 类的样本的频率。k 指分类变量的类别。一个完全纯的节点 *Gini* 指数为 0，*Gini* 指数越大说明节点越不纯。

如果结点 N 分成两子节点 N_1 和 N_2，则 *Gini* 指数变化：

$$\Delta Gini=Gini(N)-(\frac{n_1}{n}Gini(N_1)+\frac{n_2}{n}Gini(N_2)) \qquad \text{式 7-51}$$

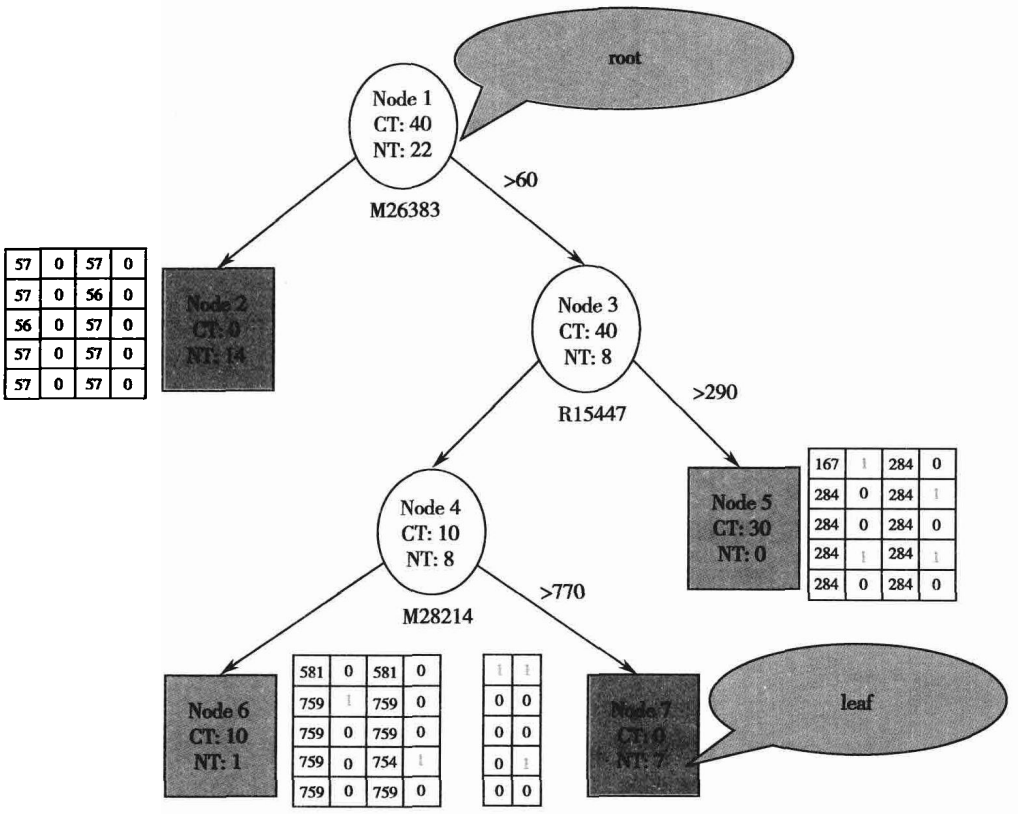


图 7-34 决策树应用于肿瘤基因表达谱的分类分析

（引自：Zhang H, Yu CY, Singer B, Xiong M: Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci USA 2001, 98(12):6730-6735.）

其中 $Gini(N_1)$和 $Gini(N_2)$为子节点 N_1 和 N_2 的 *Gini* 指数，n 为节点 N 中样本的个数，n_1 和 n_2 分别为节点 N_1 和 N_2 中样本的个数。选取 $\Delta Gini$ 最大的作为分割的基因及对应的分割方式。

2. 信息增益　该指标运用分割前后熵值的变化衡量节点纯度的变化。对于某节点 N 信息熵的定义为：

$$H(N) = -\sum_{i=1}^{k} p_i \log_2 p_i \qquad \text{式 7-52}$$

其中 p_i 是指第 i 类在某节点中的概率。k 指分类变量的类别。熵值越大说明节点越不纯。

如果结点 N 分成两子节点 N_1 和 N_2，则信息增益为：

$$Gain = H(N) - \left(\frac{n_1}{n} H(N_1) + \frac{n_2}{n} H(N_2)\right) \qquad \text{式 7-53}$$

选择信息增益最大的作为分割的基因及对应的分割方式。

通过上述方法生成的决策树对训练集的准确率往往可能达到 100%，但其结果却会导致过拟合（对信号和噪声都适应），建立的树模型不能很好地推广到总体中的其他样本，因此需要对树进行剪枝。剪枝方法主要有前剪枝和后剪枝。前剪枝即在树的生长过程中通过限定条件停止生长；后剪枝即在长成一棵大树后，从下向上进行剪枝。

四、分类模型的分类效能评价

在分类的过程中，运用重抽样方法（re-sampling）把样本集合分为训练集（training set）和检验集（testing set）。训练集用于分类模型的构建，检验集用来检验分类模型的分类性能，评价分类效能的好坏。

（一）重抽样方法有：

1. n 倍交叉验证(n-fold cross validation) 随机将样本集分为近似的 n 等份，选取一份作为检验集，余下的 $n-1$ 份作为训练集，循环 n 次。这种方法产生不相重叠的训练集和检验集。

2. Bagging(Bootstrap aggregating) 在原训练集上采用有放回抽样，每次随机抽取小于或等于原训练集大小的集合(称这种集合为原训练集的副本)，随机抽样的数目与原训练集大小一致时每一副本训练集理论上包含原训练集的 63.2% 的样本，其余的为重复抽取的样本。由该副本作为训练集，余下的样本作为检验集。

3. 无放回随机抽样 每次抽取样本集的 $1/n$ 作为检验集，余下的样本集作为训练集。

4. 留一法交叉验证(leave-one-out cross validation，LOOCV) 该方法每次随机留出一个样本作为检验集，余下的作为训练集。

（二）分类效能

分类模型的分类性能指标有以下几种：

敏感性(sensitivity)：$\frac{TP}{TP+FN}$

特异性(specificity)：$\frac{TN}{TN+FP}$

阳性预测率(positive predictive value，precision)：$\frac{TP}{TP+FP}$

阴性预测率(negative predictive value)：$\frac{TN}{TN+FN}$

均衡正确率(balanced accuracy)：$\frac{1}{2}(\frac{TP}{TP+FN}+\frac{TN}{TN+FP})$

正确率(correct or accuracy)：$\frac{TP+TN}{TP+TN+FP+FN}$

其中 TP，TN，FP，FN 分别表示真阳性(true positive)，即样本标签为阳性类，分类模型也正确地将之判断为阳性类的样本个数；真阴性(true negative)，即样本标签为阴性类，分类模型也正确地将之判断为阴性类的样本个数；假阳性(false positive)，即样本标签为阴性类，而分类模型却将之判断为阳性类的样本个数；假阴性(false negative)，即样本标签为阳性类，而分类模型却将之判断为阴性类的样本个数。

总之，当分类分析应用于基因芯片数据时，可以构建疾病预测模型，从分子层面对复杂疾病进行诊断。然而，由于复杂疾病的发生并不是单个基因的改变，而是由环境因素与遗传因素共同作用的结果，在疾病的发生发展过程中涉及的基因较多，同种疾病往往分子机制也存在很大的异质性。因此即使是针对同种疾病，运用不同芯片数据进行分类分析时，其构建的分类模型中参与的基因往往重复性较差，这使得预测模型不具有代表性，目前很难推广到临床的诊断中。

第七节 基因芯片数据的其他分析

Section 7 Complementary Analysis of Microarray Data

一、降维处理

基因芯片表达谱数据的一个最重要特点是在基因表达谱数据信息获取的过程中，检测基因的数目往往高达几千甚至几万个，而样本获取的数目，由于成本和样品来源等方面的原因，一般只有数十或百计。样本获取的数目比检测基因的数目小许多，并且随机干扰因素较多，检测误差较大，是典型的高维、高噪问题。另一方面，由于功能相似的基因其表达水平高度相关，从分类学的角度看，存在

大量的对分类无意义的基因，即冗余基因，因此，对基因芯片表达谱数据进行降维处理是必要的。降维处理的方法包括特征选择和特征提取。

特征选择是按某一评价准则从基因表达谱的 D 个基因中挑选 d 个基因的最优特征子集，这 d 个基因有明确的生物学意义，可能是与疾病相关的基因，因此可以为从分子水平解释疾病的发病机制，疾病的诊断和治疗提供一定程度的指导。运用基因芯片数据解决的一个重要问题是挖掘与复杂疾病相关的特征基因，并进行功能鉴定。复杂疾病相关特征基因不仅为疾病诊断提供分子标记，而且可作为候选的药物靶点，对癌症等复杂疾病的分型、诊断及病理学研究有非常重要的实际意义。在第四节的差异表达分析中进行的即是特征基因选择过程，第六节基因芯片数据的分类分析中，某些分类方法在进行分类的同时也进行特征基因选择，例如决策树方法等，即参与分类模型的基因是具有类别判别能力的基因，如果分类模型是与疾病相关的，则这些基因提示可能与疾病有关。

特征提取是对数据进行变换，其中对变量作线性组合是一种简单而有效的方法，例如主成分分析，它把高维数据投影到低维子空间，从而提取出包含数据尽可能变异特征的低维数据。通过特征提取后得到的样本向量的各分量是原来基因的线性组合，生物学意义不明确，但通过特征提取的降维处理可以消除数据中的噪音，进一步做聚类分析等后续研究。

二、时间序列的表达谱数据分析

当表达谱数据的实验条件是时间序列时，其包含了更多的信息，如果用普通的分析方法对时间序列的表达谱进行分析会造成信息的损失。例如，假设基因表达谱数据检测了不同的时间点 $t_1, t_2 \cdots t_n$ 的基因表达水平，通过基因的聚类分析可以识别典型的时间表达模式。通常的距离函数忽略了时间点的先后顺序，如果将时间点交换顺序，即将表达谱列的排列顺序打乱后再进行聚类分析得出的结论也会是相同的。对于这种问题一个很简单的解决方式是，考虑两个邻近时间点的差异值 $x_{i(j+1)}-x_{ij}$，其中 $x_{i(j+1)}$ 为基因 i 在第 $j+1$ 个时间点的表达值，x_{ij} 为基因 i 在第 j 个时间点的表达值。将该差异值添加到基因表达谱中，产生新的基因 i 的表达向量，x_{i1}，$(x_{i2}-x_{i1})$，$x_{i2} \cdots x_{ij}$，$(x_{i(j+1)}-x_{ij})$，$x_{i(j+1)} \cdots x_{in}$，然后可以基于新产生的扩大以后的基因表达谱矩阵进行分析。

基因之间的调控具有时间延迟效应，运用非时间序列的表达谱分析无法发现这种关系，从而在建立基因表达调控网络时无法考查动态的调控方式，但是基于时间序列的表达谱数据，运用合适的算法就可以发现基因间不同时间延迟的调控模式。例如基因 i 和 j 的表达向量分别用 $x_i=(x_{i1}, x_{i2} \cdots x_{in})$ 和 $x_j=(x_{j1}, x_{j2} \cdots x_{jn})$ 表示，考虑时间延迟的共表达关系：若基因 i 比基因 j 延迟 t 个时间点，则 $x_i=(x_{it}, x_{i(t+1)} \cdots x_{in})$，$x_j=(x_{j1}, x_{j2} \cdots x_{j(n-t+1)})$。若基因 j 比基因 i 延迟 t 个时间点，则 $x_j=(x_j, x_{j(t+1)} \cdots x_{jn})$，$x_i=(x_{i1}, x_{i2} \cdots x_{i(n-t+1)})$。运用相关系数计算延迟后的基因表达向量有无统计学意义的相关性。

改变 t 的取值，遍历所有可能的延迟共表达模式，选择最有统计学意义的作为最有可能的时间延迟方式。

三、基因转录调控网络分析

在生命科学领域，遗传网络（如基因调控网络、基因相关网络和蛋白质互作网络等）作为一种系统的、整体的研究方法正在受到重视。该方法建立在分子生物学、数学和信息学等多学科交叉的基础上，具有复杂性、稳定性和层次性等一系列特征，同时，遗传网络能够以图形的方式形象地反映分子间复杂的互作关系。通过基因表达信息，结合一定的分析和计算方法，可以构建合适的遗传网络拓扑结构来模拟系统的行为。基因转录调控网络分析在第十一章和第十二章有详细的介绍。

四、功能富集性分析

通过基因表达数据的分析，可以得到几十个，甚至上百个基因在不同的条件间有差异的表达，这些基因与表型相关。如果将这些基因表达水平的改变上升到生物体内功能水平的改变，则可以更好

地将表型相关的功能改变定位于若干生物功能或功能通路，为进一步细致的分析提供思路。功能富集性分析的基本思想是将基因映射至某个功能框架，如 KEGG、Gene Ontology 等数据库，然后通过统计学的方法挑选差异基因非随机富集的通路，这些通路提示与表型有关。功能富集性分析在第八章有详细介绍。

第八节 常用表达谱分析软件

Section 8 General Microarray Analysis Software

一、ArrayTools

BRB-ArrayTools 是基因芯片数据分析的集成软件包。BRB-ArrayTool 能够处理不同芯片平台，单、双通道的表达谱数据，该软件基本功能有数据可视化、标准化处理、差异基因筛选、聚类分析、分类预测、生存期分析、基因富集性分析等。BRB-ArrayTools 还可以通过基因的 CloneID、GenBank 号、UniGene 号连接至 NCBI 数据库，或者通过芯片的 ProbesetID 连接至 NetAffy 站点获取探针的详细信息，进行基因的功能注释。ArrayTools 以 Excel 插件的形式呈现，用户界面友好，计算由 Excel 外部的分析工具完成。ArrayTools 软件可以通过 http://linus.nci.nih.gov/~brb/download.html 下载安装。

二、DChip

Dchip(DNA-Chip Analyzer)是一款主要进行基因表达芯片和 SNP 芯片探针水平和高水平分析的软件，其他芯片分析平台的基因表达数据和 SNP 数据也可以分析。探针水平的分析为基于统计模型提取表达信息、处理交叉杂交和图像污染信息，包括数据输入、可视化处理、标准化处理、表达值提取、奇异芯片分析等。高水平分析包括基因过滤、样本比较、层次聚类、分类分析、通路分析、LOH 和 SNP 芯片的拷贝数分析等全面的数据分析。Dchip 软件可以通过 http://groups.google.com/group/dchip-software/web/downloading-dchip 网页下载安装，下载后的可执行文件直接双击运行。

三、SAM

SAM(Significance Analysis of Microarrays)是差异基因筛选的统计学方法。与该方法对应的软件由 Balasubramanian Narasimhan 和 Robert Tibshirani 编写。SAM 的输入为基因表达谱矩阵及其每个实验对应的反应变量。反应变量可以是两种类别信息，例如治疗前和治疗后；也可以是多类别信息，例如乳腺癌、淋巴瘤、大肠癌等；可以是定量变量，例如血压；或者是癌症的生存期信息等。对于每个基因 i，SAM 计算统计量 d_i，d_i 是衡量基因表达与反应变量之间的关联强度，SAM 还采用重复扰动数据集判断这种关联强度的统计学意义。判断基因是否差异的阈值由调节参数 δ 决定，差异基因筛选的假阳性率可指导 δ 值的确定。用户也可以通过选择倍数差异(fold change)阈值保证挑选出的差异基因的倍数差异至少满足预先指定的阈值。SAM 的输出为差异基因表(表达上调基因和表达下调基因)、δ 值表和样本大小评价表。

SAM 软件可以通过 http://www-stat.stanford.edu/~tibs/SAM/ 网页下载，安装后以 Excel 插件的形式运行。

四、Cluster 和 TreeView

Cluster 是对 DNA 芯片数据进行聚类分析的软件。TreeView 是对 Cluster 的聚类结果进行交互式可视化呈现的软件。Cluster 软件的功能有：数据过滤、标准化处理、层次聚类、均值聚类、SOM 聚类和主成分分析。

五、R 语言和 BioConductor

Bioconductor(http://www.bioconductor.org/)是一个开源和开放式的软件开发项目，主要由哈佛医学院 / 哈佛公共卫生学院的 Dana Farber 癌症研究所生物统计组开发，该项目的目标是建立多方面的、强有力的基因组数据的统计与图形分析方法，促进各种生物数据的集成、推动数据的综合分析和利用，促进形成高质量的文档，同时发展各种开放式软件，并加强基因组数据分析方法和技术的培训。Bioconductor 的应用功能主要以包的形式组织。处理、分析和注释 DNA 微阵列实验数据是 Bioconductor 目前的主要应用领域和重要组成部分。包括对微阵列数据的预处理、数据可视化、基因表达数据的分析和注释。其中 affy 是用于单通道寡核苷酸芯片数据的读取、作图和标准化分析等预处理的工具包。marray 是用于双通道(cDNA)微阵列数据的读取、作图和标准化分析等预处理的软件包。Limma 是使用线性模型进行实验分析和差异表达分析(如倍数差异、t 检验等)的软件包，对双通道和单通路的所有芯片类型都适用，另外它还提供针对双通道点样芯片的标准化分析和其他分析功能。

六、Bioinformatics Toolbox

Bioinformatics Toolbox 是基于 MATLAB 环境开发的基因组和蛋白质组分析工具箱。该工具箱包括以下功能：数据格式与数据库，即与基因组和蛋白质组相关的数据库建立连接，进行数据的获取和多种数据格式间的转换；序列分析，即序列的特征分析、两序列对位排列、多序列对位排列、基于隐马尔可夫链的模式识别等；进化分析；微阵列数据分析，即芯片数据的读取、标准化、可视化等处理；质谱数据分析；统计学习，即运用统计学习工具分类或识别数据中的特征；在 MATLAB 环境中调用其他的生物信息学软件(例如 BioPerl 或 BioJava)。

小　结

细胞的生物过程是由许多分子共同参与的复杂过程，对于这种过程的理解需要从全局的系统的角度出发。由于基因芯片技术能同时检测成千上万个基因在某种条件下的表达丰度，这为系统地分析提供了可能。然而高通量的芯片数据存在各种来源的噪音，为了获得可靠的基因表达数据，一方面需要完善芯片技术平台，另一方面需要优秀的数据分析算法。通过扫描仪采集后的基因表达的荧光信号，必须经过数据的过滤和预处理消除数据中可能存在的错误和系统误差后才能进行下一步的数据分析。数据分析的方法通常包括差异基因筛选、有监督学习和无监督学习等。通过这些常用的数据分析方法可以进行疾病潜在生物路标的发现、肿瘤异质性分析、肿瘤分型诊断等。随着芯片数据的高速增长，需要对这些数据进行收集、存储和管理，目前基因表达数据相关的公共数据库包括有 GEO、SMD 和 ArrayExpress 等。MGED 还组织开发了微阵列基因表达标记语言(MAGE-ML)，它用来组织微阵列实验的数据和相关信息，从而提供微阵列数据表示和交流的有效手段。目前微阵列技术在甲基化芯片、SNP 芯片、miRNA 芯片、CHIP-on-Chip 实验等领域都有新的应用。

Summary

Biological processes are composed of many molecules to participate in the celluar society, understanding of such a kind of complicated system, therefore, should be conducted from a global perspective. As the gene microarray technology can simultaneously detect the expression abundances of thousands of genes under certain conditions, which just meets the requirement of systematic analysis. However, the various sources of noise in these high-throughput microarray data need to be filtered both by improving the chip technology platform and the excellent data

analysis algorithms. Microarray raw data collected through the scanner have to pass through the data filtering and preprocessing to eliminate potential data errors and system errors before the next step in data analysis. Data analysis methods typically include Analysis of Differentially Expression Gene aiming to seek for the potential biological marks, Cluster Analysis of Microarray Data aiming to unravel the tumor heterogeneity, and Classification of Microarray Data aiming to develop the new approaches of disease diagnosis. With the rapid growth of microarray data, public databases including GEO, SMD and the ArrayExpress are developed to meet the need for such data collection, storage and management. MGED has developed the microarray gene expression markup language (MAGE-ML), which is used to orgnize microarray experimental data and related information, thus providing microarray data representation and exchange of effective means. At present microarray chip technology has bee applied in some new areas such as DNA methylation, SNP detection, miRNA expression, CHIP-on-Chip.

（李亦学 王海芸 吕飒丽）

习 题

1. 以下哪些是 cDNA 芯片数据的主要系统误差来源：
   A. 不同染料的物理性质差异
   B. 染料结合效能
   C. 点样针差异
   D. 不同芯片间的差异
   E. 实验条件的差异
2. 简述多重假设检验对假设检验结果的影响，以及如何校正这种影响。
3. 特征选择和特征提取的异同是什么？芯片数据分析中常用到的是哪种方法？为什么？
4. 聚类分析的机器学习方法在芯片数据分析中的应用是什么？
5. 聚类分析中有哪些相似性函数，采用不同的相似性函数对结果会有什么影响？
6. 对于有监督的分类分析，如何评判分类效能的好坏？
7. 基因富集性分析方法中常用的假设检验方法有以下哪些：
   A. χ^2 检验
   B. 超几何分布
   C. 二项分布
   D. t 检验
8. 下列哪些数据库是基因表达数据库：
   A. SMD
   B. Gene
   C. GEO
   D. ArrayExpress
   E. CGED
   F. GO
9. 论述基因芯片数据的应用。
10. 从 GEO 数据库中查找一套肿瘤数据，然后运用所学的方法对其进行分析，讨论数据揭示的生物学意义。

主要参考文献

1. 孙啸，谢建明，周庆，等. R 语言及 Bioconductor 在基因组分析中的应用(第一版). 北京：科学出版社；2006 年.
2. 边肇祺，张学工. 模式识别(第二版). 北京：清华大学出版社；2000 年.
3. 蒋知俭. 医学统计学. 北京：人民卫生出版社；1997 年.
4. Oliphant A., Barker D. L., Stuelpnagel J. R. et al. Enabling an accurate, cost-effective approach to high-throughput genotyping. Biotechniques, 2002, Suppl: 56-58.
5. Yang Y. H., Dudoit S., Luu P., et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res., 2002, 30(4): e15.
6. Boes T., Neuhauser M. Normalization for Affymetrix GeneChips. Methods Inf. Med., 2005, 44(3): 414-417.
7. Li C., Wong W. H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl. Acad. Sci. USA ., 2001, 98(1): 31-36.
8. Tusher V. G., Tibshirani R., Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA., 2001, 98(9): 5116-5121.
9. Alizadeh A. A., Eisen M. B., Davis R. E., et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 2000, 403(6769): 503-511.
10. Tibshirani R., Hastie T., Narasimhan B., et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA., 2002, 99(10): 6567-6572.
11. Zhang H., Yu C. Y., Singer B., et al. Recursive partitioning for tumor classification with gene expression microarray data. Proc. Natl. Acad. Sci. USA., 2001, 98(12): 6730-6735.
12. Wang H., Wang Q., Li X., et al. Towards patterns tree of gene coexpression in eukaryotic species. Bioinformatics, 2008, 24(11): 1367-1373.

第二篇 功能基因组信息学

第八章 基因注释与功能分类
CHAPTER 8 GENE ANNOTATION AND FUNCTIONAL CLASSIFICATION

第一节 引 言
Section 1 Introduction

随着后基因组时代的来临，基因组学的研究重心开始从阐明所有遗传信息转移到从整体分子水平对功能进行研究。这种转变的一个重要标志是产生了功能基因组学(functional genomics)。如果说生物信息学在人类基因组计划中的研究重点是基因组序列的话，那么在功能基因组学中，生物信息学的研究重点则是序列的生物学意义，以及基因组编码序列的转录、翻译的过程和结果，着重分析基因表达调控信息、基因及其产物的功能。功能基因组学的主要任务之一是进行基因组功能注释(genome annotation)，了解基因的功能，认识基因与疾病的关系，掌握基因的产物及其在生命活动中的作用等。在使用全局方法进行研究时，研究人员往往同时检测大量基因的表达水平，从而在整体水平上获得关于基因功能及基因之间相互作用的信息，如何应用生物信息学方法，高通量地注释这些基因的生物学功能是一个重要的工作。快速有效的基因注释对进一步识别基因，识别基因转录调控信息，研究基因的表达调控机制，研究基因在生物体代谢途径中的地位，分析基因与基因产物之间的相互作用关系，绘制基因调控网络图，预测和发现蛋白质功能，揭示生命的起源和进化等具有重要的意义。

本章主要介绍当前常用的基因及其产物的功能注释体系和工具，以及在此基础上发展起来的基因集功能富集分析、基因产物功能预测等方法。

第二节 基因注释数据库
Section 2 Gene Annotation Database

目前，研究人员已经掌握了大量的全基因组数据，同时关于基因、基因产物以及生物学通路的数据也越来越多，解释生物学实验的结果，尤其从基因组角度，需要系统的方法。某个物种的基因组包括成千上万的基因甚至更多，它们在分子水平的复杂网络中相互作用。这些分子网络趋于模块化，相近的模块再形成一种组合的单元发挥功能。进一步，这些模块可以按照进化时间组装成层级结构来发挥更高级的功能。描述单一的蛋白质功能已经十分复杂，要是在基因组范围内进行描述就会更加复杂，可能最好的工具就是计算机程序。因此提供一个结构化的标准生物学模型，便于计算机程序进行分析，成为从整体水平系统研究基因及其产物的一项基本需求。本节主要介绍当前应用较为广泛的基因及其产物注释数据库。

一、基因本体数据库

（一）简介

基因本体(gene ontology，GO)数据库是GO组织(GO consortium)在2000年构建的一个结构化的标准生物学模型(图8-1)，旨在建立基因及其产物知识的标准词汇体系，涵盖了基因的细胞组分(cellular component)、分子功能(molecular function)、生物学过程(biological process)三个方面，目前已经成为应用最广泛的基因注释体系之一。GO数据库最初收录的基因信息来源于3个模式生物数据库：果蝇、酵母和小鼠，随后相继收录了更多数据，其中包括国际上主要的植物、动物和微生物基因组数据库(表8-1)。GO术语在多个合作数据库中的统一使用，促进了各类数据库对基因描述的一致。

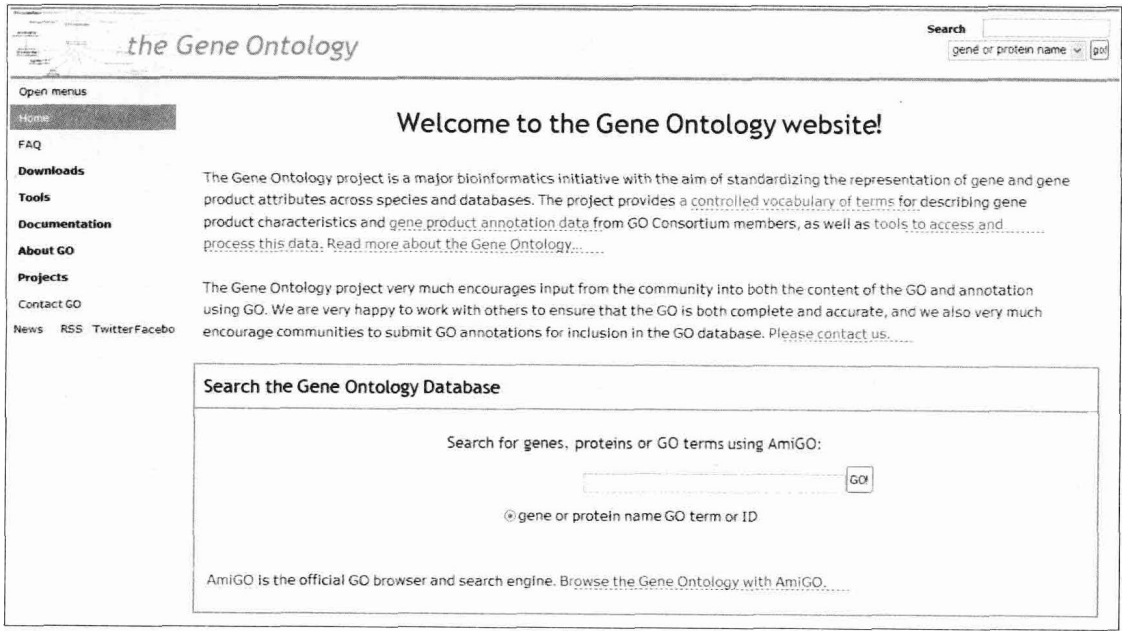


图8-1 GO数据库主页

表8-1 GO数据库收录的基因组数据列表

机构简称	收录的基因组数据	网站
BBOP	果蝇	http://www.berkeleybop.org
BHF-UCL	心血管基因	http://www.cardiovasculargeneontology.com
dictyBase	粘菌盘基网柄菌	http://dictybase.org
EcoliWiki	大肠杆菌	http://ecoliwiki.net
FlyBase	果蝇	http://flybase.bio.indiana.edu
GeneDB	裂殖酵母 恶性疟原虫 硕大利什曼原虫 布氏锥虫	http://www.genedb.org
GOA	UniProt和InterPro注释	http://www.ebi.ac.uk/GOA
Gramene	农作物基因数据库	http://www.gramene.org
MGD and GXD	小家鼠	http://www.informatics.jax.org
RGD	褐家鼠	http://rgd.mcw.edu
Reactome	生物过程知识库	http://www.genomeknowledge.org
SGD	芽殖酵母 酿酒酵母	http://www.yeastgenome.org

续表

机构简称	收录的基因组数据	网站
TAIR	拟南芥	http://www.arabidopsis.org
IGS	基因组研究的工具和数据	http://www.igs.umaryland.edu
JCVI	若干种细菌基因组数据库	http://www.jcvi.org
WormBase	线虫	http://www.wormbase.org
ZFIN	斑马鱼	http://zfin.org

GO 通过控制注释词汇的层次结构使得研究人员能够从不同层面查询和使用基因注释信息。从整体上来看 GO 注释系统是一个有向无环图(directed acyclic graphs),包含三个分支,即:生物学过程(biological process),分子功能(molecular function)和细胞组分(cellular component)。注释系统中每一个结点(node)都是基因或蛋白质的一种描述,结点之间保持严格的关系,即 "is a" 或 "part of"(图 8-2)。因此,一个基因或蛋白质可从三个层面得到注释,即基因或蛋白质参与的生物学过程,在细胞内的特定组分,以及分子功能上所扮演的角色。随着生命科学研究的逐步深入,GO 注释数据库正在不断积累和更新。目前 GO 已经成为生物信息领域中一个重要的资源和工具,并正在逐步改变着人们对各种生物学数据的组织和理解方式,它的存在极大地加快了生物数据的整合和利用。

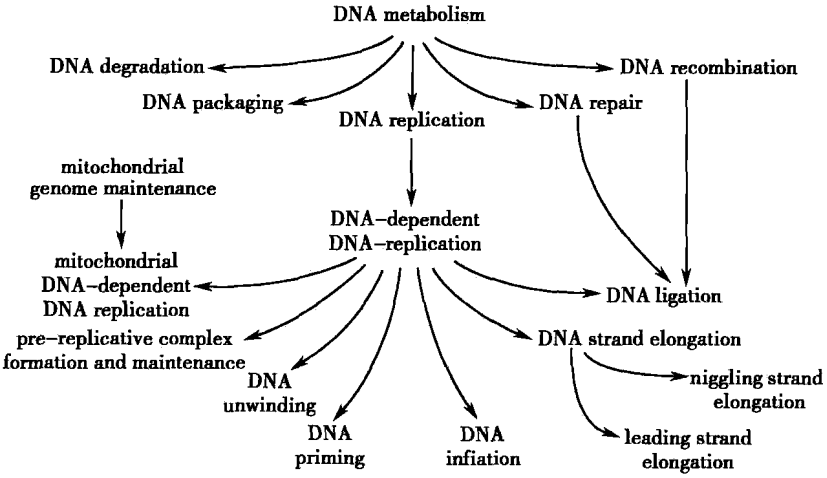


图 8-2 GO 中生物学过程的 DNA 代谢部分功能类示意图

(二) 使用 GO 数据库

1. 用关键词检索 GO 数据库 检索 GO 数据库通常先进入 AmiGO 的首页(图 8-3)。在 GO 数据库中,每条记录都有一个数据标识号 GO:XXXXXX 和对应的术语。因此检索时需要知道待查基

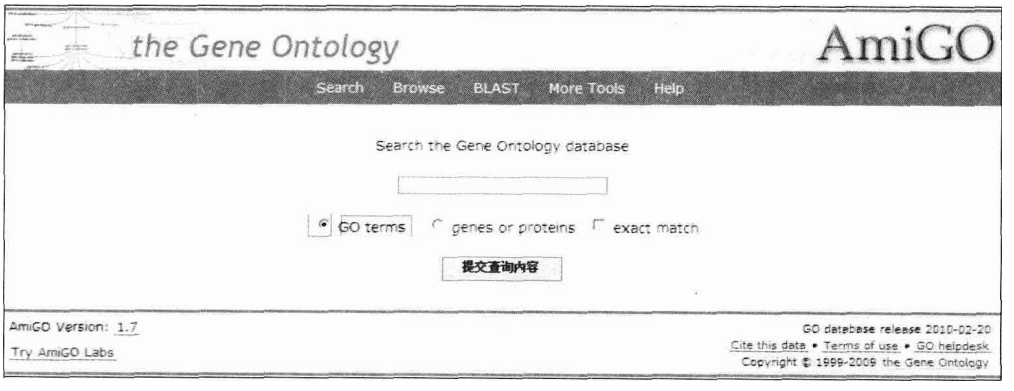


图 8-3 AmiGO 检索网页

因的数字标识号或术语，将它们直接输入框中检索即可。如果检索的基因或蛋白质存在别名，可在检索框下勾选“gene or proteins”，并在检索框中输入别名检索；“exact match”表示是否完全匹配，可供选择。

这里以检索神经源性分化因子 6(NEUROD6)为例。在检索框中输入“NEUROD6”并勾选“gene and proteins”和“exact match”，运行后所得基因产物检索结果如图 8-4 所示。

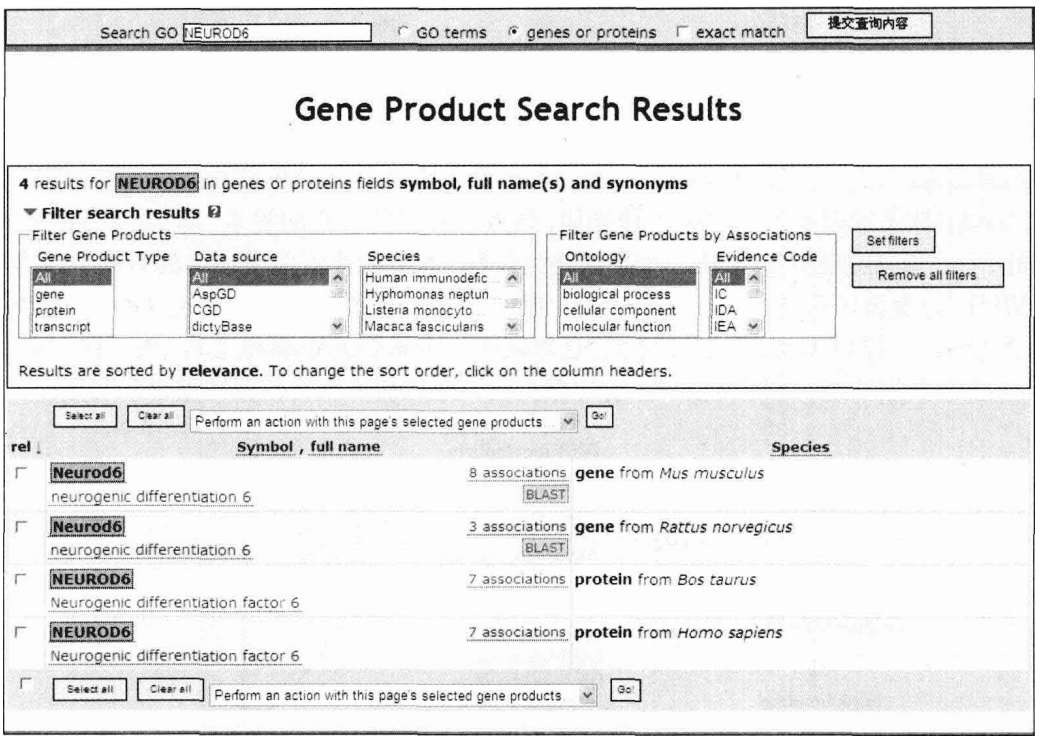


图 8-4 AmiGO 检索结果示例

检索得到的四个记录分别是不同物种中的神经源性分化因子 6，点击物种为人类的“NEUROD6”记录，得到结果如图 8-5 所示。

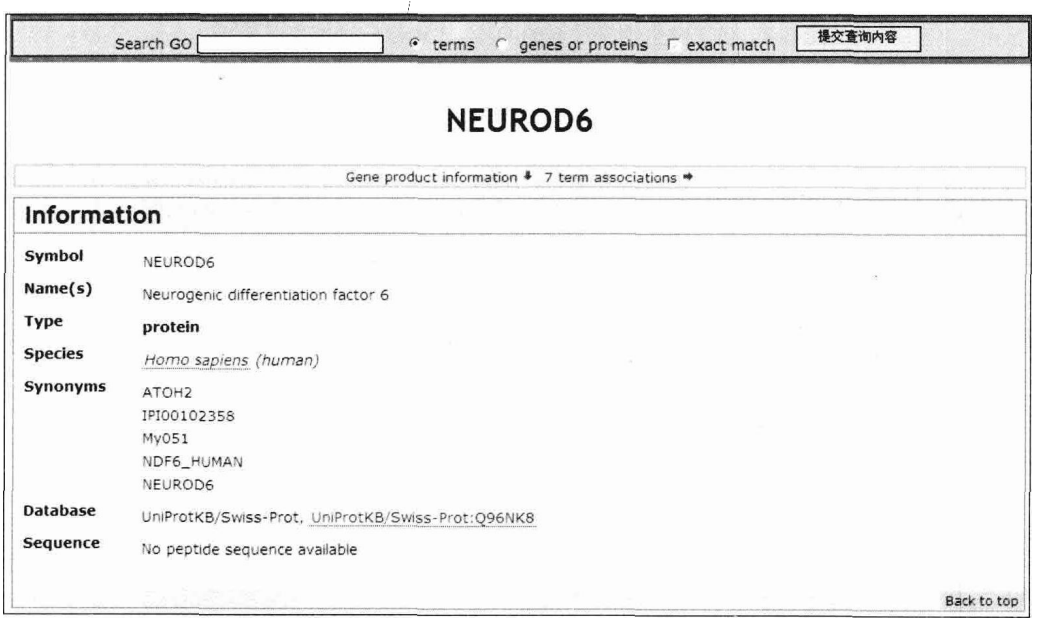


图 8-5 AmiGO 基因描述示例 1

图 8-5 显示了该基因产物的基本信息，包括类型、物种、别名来源和序列；图 8-6 显示了该基因产物的术语关联(term associations)图，图中记录名称 “Term” 是 GO 记录的名字，“Ontology” 是该基因产物的特性，如要查看其分子功能，可点击其中的一条记录如 “nervous system development”(图 8-7)。

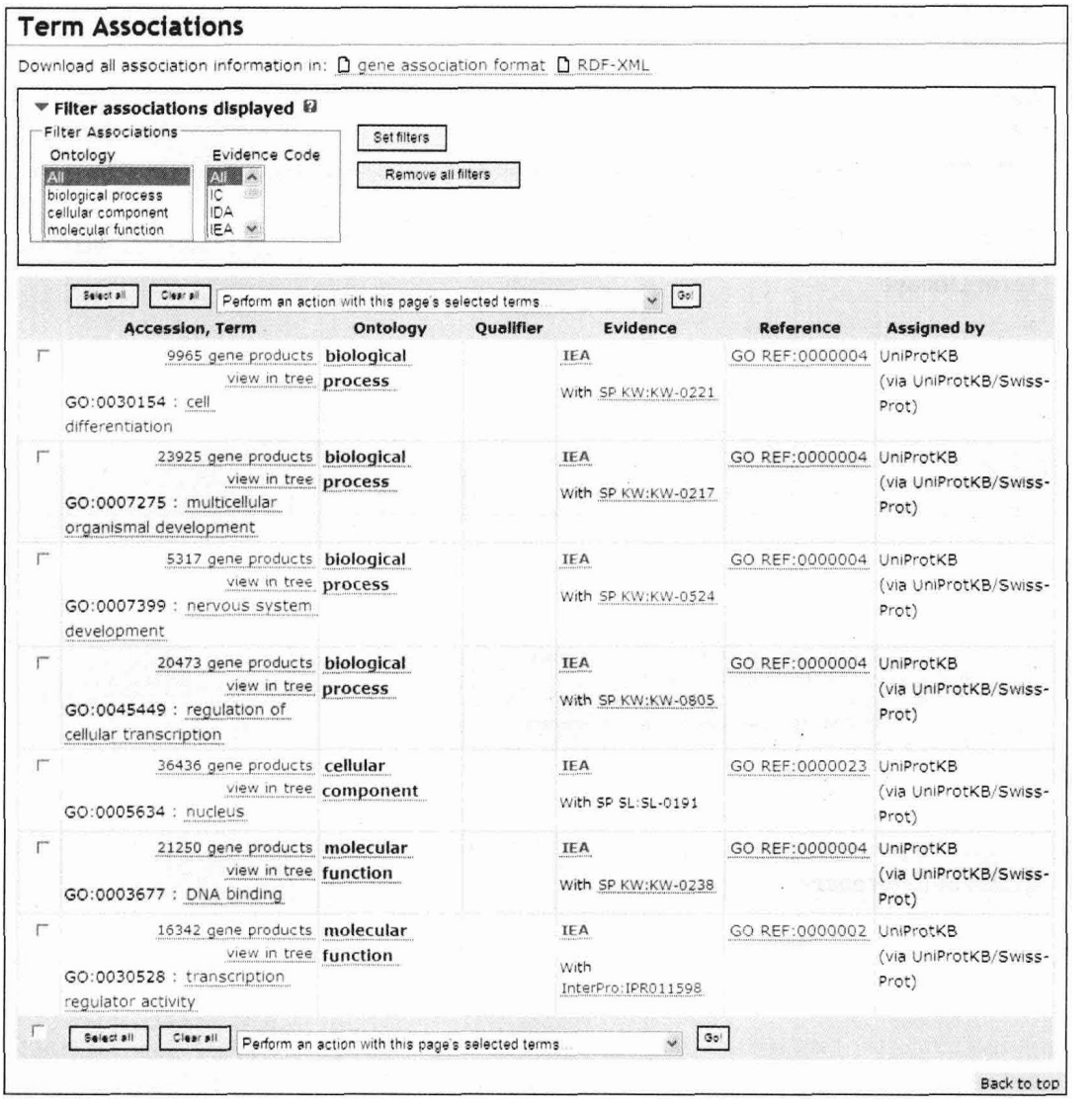


图 8-6 AmiGO 基因描述示例 2

图 8-7 上部先对神经源性分化因子 6 的相关信息做简单描述，中间术语系谱(term lineage)成阶梯状分布，记录了 GO 数据库中全部分子功能所处的位置和关系。下方 “External Reference” 提供了与外部相关数据的链接。点击右上方的可视化视图(graphical view)能更清晰地显示分子功能记录之间构成的复杂网状结构，既有上下隶属关系，也存在平行关系(图 8-8)。

2. 用序列检索 GO 数据库　对于未知基因名的序列，可以用序列直接检索 GO 数据库。点击 AmiGO 首页上方的 “BLAST”，进入检索界面。在检索框中输入氨基酸或核酸序列，检索工具能自动识别并相应地选择 BLASTP 或 BLASTX 来与数据库中的序列进行比对。这里以检索 RPIA 基因的序列为例，如图 8-9 所示。

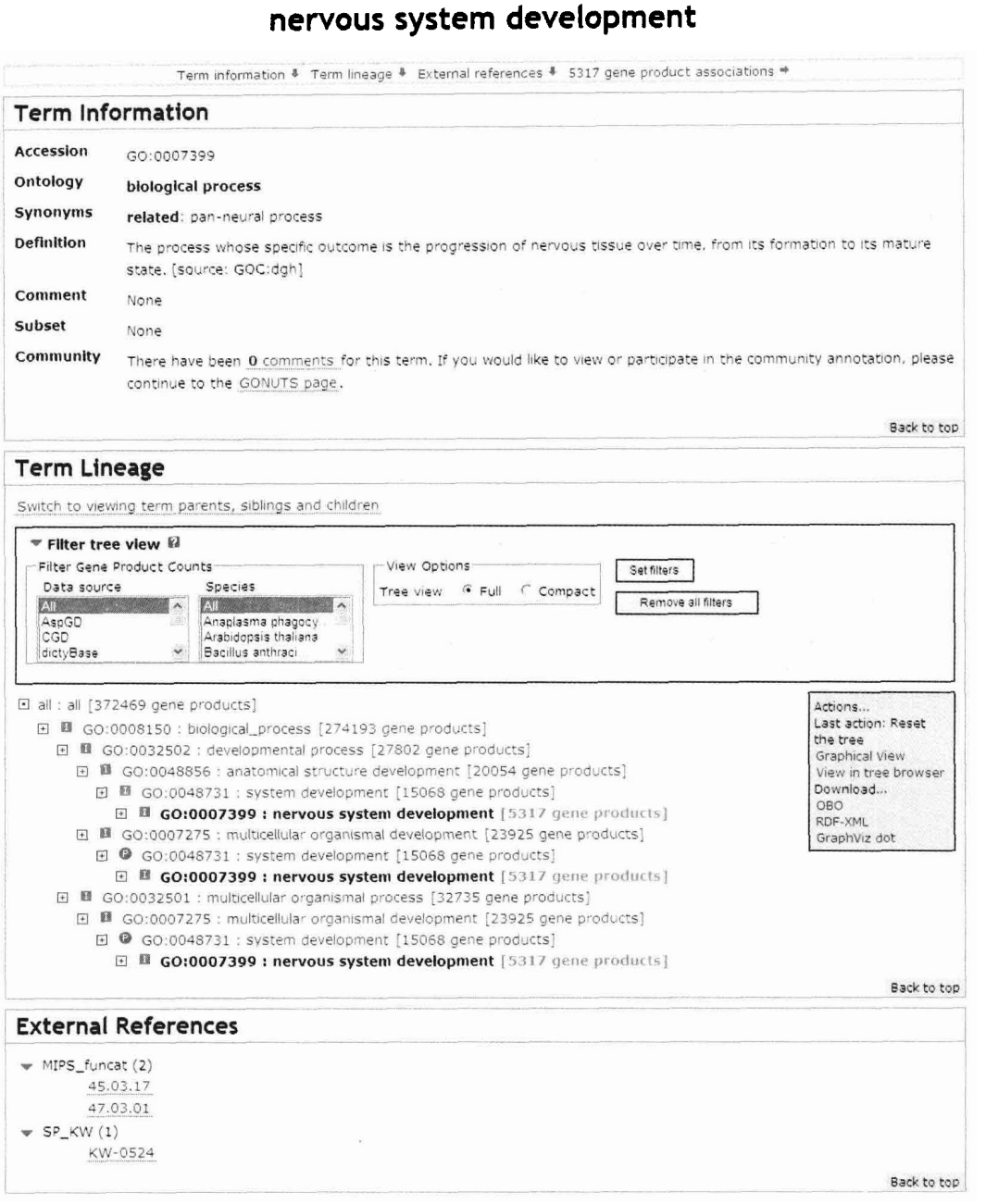


图 8-7 AmiGO 基因功能描述示例

二、京都基因与基因组百科全书数据库

京都基因与基因组百科全书(Kyoto encyclopedia of genes and genomes, KEGG)是系统分析基因功能、基因组信息的数据库，它整合了基因组学、生物化学以及系统功能组学的信息，有助于研究者把基因及表达信息作为一个整体进行研究。

KEGG 提供的整合代谢通路查询十分出色，包括碳水化合物、核苷酸、氨基酸等代谢及有机物的生物降解，不仅提供了所有可能的代谢通路，还对催化各步反应的酶进行了全面的注解，包含其氨基酸序列、到 PDB 数据库的链接等。此外，KEGG 还提供基于 Java 的图形工具访问基因组图谱、比较基因组图谱和操作表达图谱以及其他序列比较、图形比较和通路计算的工具。因此，KEGG 数据库是进行生物体内代谢分析、代谢网络分析等研究的强有力工具之一。

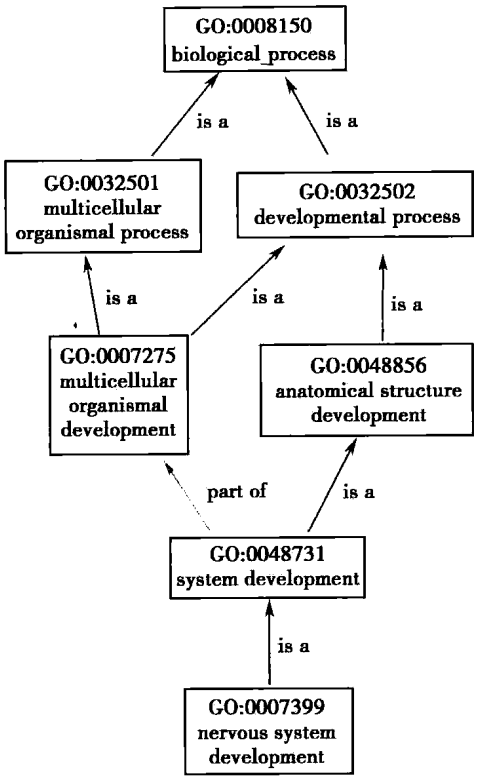


图 8-8 AmiGO 查询结果图形视图

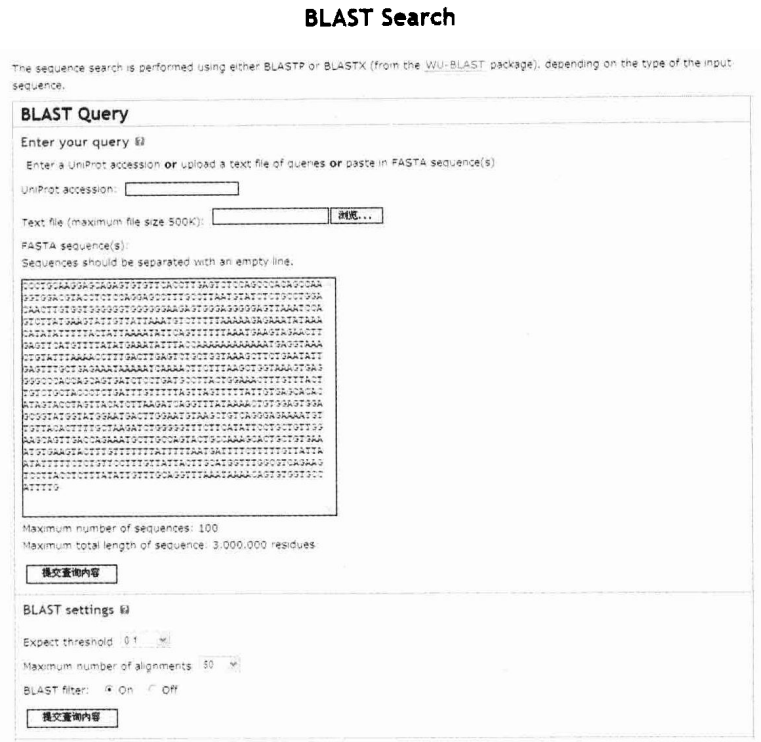


图 8-9 AmiGO BLAST 序列检索网页

（1）KEGG 存储内容：KEGG 目前共包含了 19 个子数据库。①基因组信息存储在 GENES 数据库里，包括全部完整的基因组序列和部分测序的基因组序列，并伴有实时更新的基因相关功能的注释，更高级的功能信息则存储在 PATHWAY 数据库里，包括图解的细胞生化过程如代谢、膜转运、信号传递、细胞周期和同系保守的子通路等信息；一些直系同源的基因数据作为 PATHWAY 数据库的补充，形成了 PATHWAY 数据库中一些保守的子通路(pathway motifs)，这些子通路通常有一些在染色体位置上邻近的基因编码，这对于基因功能的预测十分重要。②KEGG 中化学信息的 6 个数据库被称为 KEGG LIGAND 数据库，包含化学物质、酶分子、酶化反应等信息。KEGG BRITE 数据库是一个包含多个生物学对象的基于功能进行等级划分的本体论数据库，它包括分子、细胞、物种、疾病、药物以及它们之间的关系，该数据库将基因与外界环境影响联系起来。例如，可以通过 BRITE 数据库分析药物和靶点之间的关系。③一些小的通路模块被存储在 MODULE 数据库中，该数据库还存储了其他的一些相关功能的模块以及化合物信息。④KEGG DRUG 数据库存储了目前在日本所有非处方药和美国的大部分处方药品。⑤KEGG DISEASE 是一个存储疾病基因、通路、药物以及疾病诊断标记等信息的新型数据库。

（2）KEGG 数据库的注释与检索：KEGG 通常被看作是生物系统的计算机表示，它囊括了生物系统中的各个对象与对象之间的关系。在分子层面、细胞层面、组织层面都可以进行检索。每个数据库中的检索条目按照一定规律被赋予一个检索号，也就是 ID。表 8-2 中列出了 KEGG 的 13 个核心数据库的检索号，其中 GENOME 和 ENZYME 使用了这一领域通用的标准命名规则，GENOME 使用了 3~4 个字母作为名称来区分不同的物种。此外，其他的数据库 ID 命名均采用 5 个数字，并以一个大写英文字母(如 K、C 等)或者 2~4 个小写字母(如 map、br 等)为前缀。例如：C00047 代表赖氨酸，hsa05210 代表结肠癌通路。利用这些 ID 号在 KEGG 提供的搜索工具 DBGET 中进行检索，可以直接获得各个数据库的相应结果。另外，这些 ID 号也被当前比较流行的网络搜索引擎(如 Google、Yahoo 等)所接受，可以直接在 KEGG 相应的数据库中得到搜索结果。

表 8-2 KEGG 的 13 个核心数据库的检索号

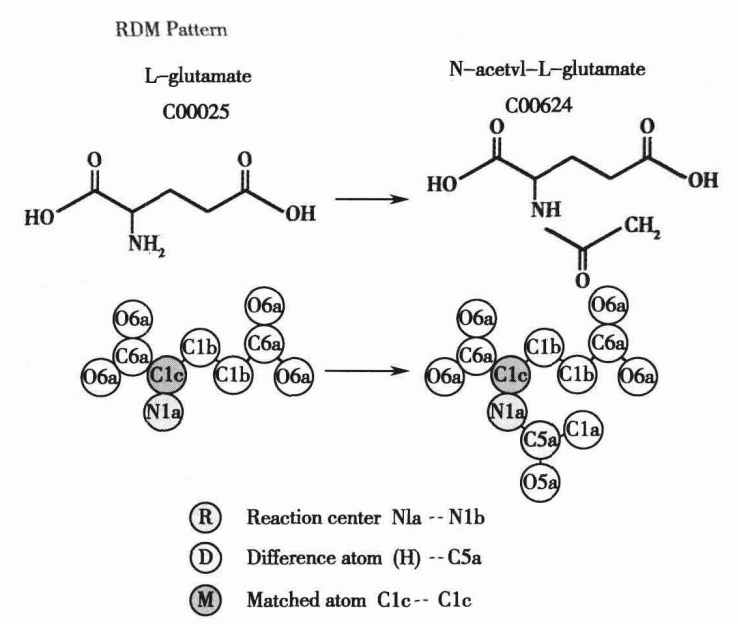
Release	Database	Object Identifier
1995	KEGG PATHWAY	map number
	KEGG GENES	locus_tag / GeneID
	KEGG ENZYME	EC number
	KEGG COMPOUND	C number
2000	KEGG GENOME	organism code / T number
2001	KEGG REACTION	R number
2002	KEGG ORTHOLOGY	K number
2003	KEGG GLYCAN	G number
2004	KEGG RPAIR	RP number
2005	KEGG BRITE	br number
	KEGG DRUG	D number
2007	KEGG MODULE	M number
2008	KEGG DISEASE	H number
2009	KEGG PLANT	
Future releases	KEGG MEDICUS	Integrate KEGG DISEASE, KEGG DRUG, and various aspects of human body systems

KEGG 通过 KO 标识(KEGG Orthology，也称为 KO 号)对基因进行注释，每个 KO 标识代表一个来自不同物种的直系同源基因组。在 KEGG 通路中，每个 KO 标识代表着通路图中一个网络结点(在通路图中以一个方盒子表示)。在 KEGG 对每个对象的功能及其他等级划分中，KO 标识则代表着底层的叶子结点。

KO 标识是基因组通过 KEGG 通路以及 KEGG 等级划分与生物学系统关联的基础。对于 KEGG 中的每个物种来说，物种特异性通路以及功能等级的划分是通过计算的方法自动实现的，在这一过程中 KO 标识是必不可少的。有了这些物种特异性通路以及功能等级划分，由基因芯片表达谱等高通量方法得到的基因便可以注释到相应的位置，以此来系统的分析该基因在细胞或组织中的功能。除了对基因或蛋白质的功能等级划分之外，KEGG BRITE 数据库还包含了化合物(C、D、G、R 标识)以及其作用关系的等级划分。

KO 标识还可以将基因的基因组信息以及转录组信息与通路总化合物分子的化学结构联系起来，因此，KO 分类系统还可以应用在化学信息注释上。这一过程的基本原理是每个 KO 下的基因所标识的酶不同，其对应化学底物也不同。例如：糖类的生物合成是通过一系列的生化反应来完成的，这些反应都是由糖基转移酶催化。在 KEGG PATHWAY 中，与糖类生物合成相关的通路图中各种糖类相关的化合物都是通过一条边与糖基转移酶的一组同源基因(KO group)直接相连，一旦在通路中确定了基因的注释位置，则与其相关的糖类化合物也能被找到。应用相似的方法可以对基因芯片表达谱数据进行糖类结构及其功能的预测，这一方法已被广泛使用。除了糖类化合物之外，在 KEGG 数据库中还存储了很多其他化合物(多聚不饱和脂肪酸、萜类化合物、聚酮化合物等)的结构和功能信息，通过以上方法可以对基因进行化学信息的注释。

另外一种化学注释的方法是以小分子化学结构的生物学意义为特征来实现的。和先前提到的一样，在 KEGG 数据库中，酶与酶之间的反应信息以及相关的化学结构信息分别存储在 KEGG REACTION 数据库和 KEGG REPAIR 数据库中。每个化合物的化学结构都被转化为 RDM(atom type changes at R:reaction center, D:diffevent atom, M:matched atom)模式(图 8-10)。大多数的 RDM 模式在 KEGG 数据库中都会被唯一存储，并且相对于其他存储的化合物会被优先找到。利用这一点可以预测代谢中较为重要的异生化合物。



(Example)RDM pattern for AO4458

图 8-10 KEGG 数据库存储的 RDM 模式

下面以人类编码葡萄糖磷酸变位酶的基因“PGM1”为例：首先进入 KEGG 首页，在首页顶端的输入框中输入类葡萄糖磷酸变位酶基因名称“PGM1”(图 8-11)。

点击搜索按钮“GO”进入查询结果页面(图 8-12)，该页面会列出针对基因“PGM1”在 KEGG 数据库中的搜索结果，除人类外，包含“PGM1”基因的物种条目也会被列出。

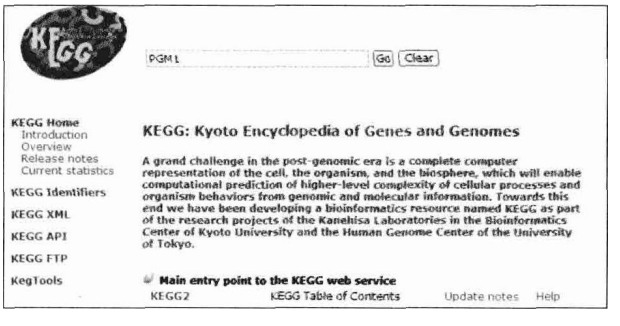


图 8-11 KEGG 查询首页

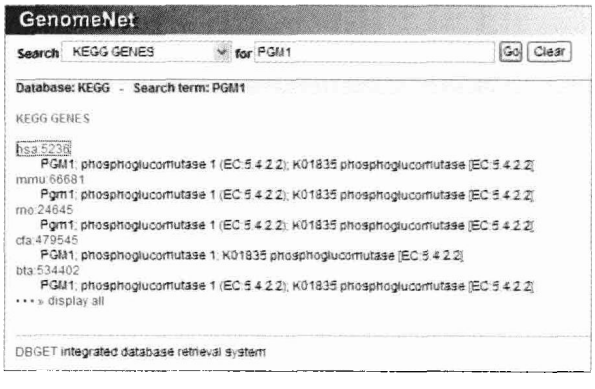


图 8-12 查询结果

其中排在第一行的是人类基因“PGM1”的相关信息，点击该条目进入到详细信息页面(图 8-13)。

该页面以表格的形式列出了该基因有关的详细信息，包括基因编号，基因的详细定义，所编码的酶的编号，基因所在通路，以及序列的编码信息。同时，在页面的右侧还提供了该基因在其他分子生物学数据库的链接，如 OMIM、NCBI、GenBank 等。

通过点击相应的链接，可以进入该基因相应信息的页面。在 pathway 这一栏中列出了该基因所在的生物学通路，点击编号为 hsa00010(糖酵解或糖异生通路)的通路，进入到该通路的相应页面(图 8-14)。

编号为 hsa00010 的通路页面以简单的几何图形显示出了糖酵解 / 糖异生相关生物过程。图中红色的方框即为基因“PGM1”所编码的酶，可以通过该酶所在位置以及通路的拓扑结构来综合分析基因。

KEGG Homo sapiens (human): 5236

Help

Entry	5236 CDS H.sapiens
Gene name	PGM1
Definition	phosphoglucomutase 1 (EC:5.4.2.2)
Orthology	K01835 phosphoglucomutase [EC:5.4.2.2]
Pathway	hsa00010 Glycolysis / Gluconeogenesis hsa00030 Pentose phosphate pathway hsa00052 Galactose metabolism hsa00500 Starch and sucrose metabolism hsa00520 Amino sugar and nucleotide sugar metabolism hsa01100 Metabolic pathways
Class	Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis [PATH:hsa00010] Metabolism; Carbohydrate Metabolism; Pentose phosphate pathway [PATH:hsa00030] Metabolism; Carbohydrate Metabolism; Galactose metabolism [PATH:hsa00052] Metabolism; Carbohydrate Metabolism; Starch and sucrose metabolism [PATH:hsa00500] Metabolism; Carbohydrate Metabolism; Amino sugar and nucleotide sugar metabolism [PATH:hsa00520] BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster GFIT
Motif	Pfam: PGM_PMM_I PGM_PMM_III PGM_PMM_II PGM_PMM_IV PROSITE: PGM_PMM Motif
Other DBs	NCBI-GI: 213611621 NCBI-GeneID: 5236 OMIM: 171900 HGNC: 8905 HPRD: 01959 Ensembl: ENSG00000079739 UniProt: P36871
Position	1p31
AA seq	562 aa AA seq DB search MVKIVTVKTQAYQDQKPGTSGLRKRVKVFQSSANYAENFIQSIISTVEPAQRQEATLVVG GDGRFYMKEAIQLIARIAAANGIGRLVIGQNGILSTPAVSCIIRKIKAIGGIILTASHNP GGPNGDFGIKFNISNGGPAPEAITDKIFQISKTIEEYAVCPDLKVDLGVLGKQQFDLENK FKPFTVEIVDSVEAYADMLRSIFDFSALKELLSGFNRLKIRIDAMHGVTGPYVKKILCEE LGAPANSAVNCVPLEDFGGHHPDPNLTYAADLVETMKSGEHDFGAAFDGDGDRNMILGKH GFFVNPSDSVAVIAANIFSIPYFQQTGVRGFARSMPTSGALDRVASAIKIALYETPTGWK FFGNLMDASKLSLCGEESFGTGSDHIREKDGLWAVLAWLSILATRKQSVEDILKDHWQKY GRNFFTRYDYEEVEAEGANKMMKDLEALMFDRSFVGKQFSANDKVYTVEKADNFEYSDPV DGSISRNQGLRLIFTDGSRIVFRLSGTGSAGATIRLYIDSYEKDVAKINQDPQVMLAPLI SIALKVSQLQERTGRTAPTVIT
NT seq	1689 nt NT seq atggtgaagatcgtgacagttaagacccaggcgtaccaggaccagaagccgggcacgagc gggctgcggaagcgggtgaaggtgttccagagcagcgccaactacgcggagaactttcatc cagagtatcatctccaccgtggagccggcgcagcggcaggaggccacgctggtggtgggc ggggacggccggttctacatgaaggaggccatccagctcatcgctcgcatcgctgccgcc aacgggatcggtcgcttggttatcggacagaatggaatcctctccacccctgtgtatcc tgcatcattagaaaaatcaaagccattggtgggatcattctgacagccagtcacaaccca gggggcccaatggagattttggaatcaaattcaatatttctaatggaggtcctgctcca gaagcaataactgataaaatttccaaatcagcaagacaattgaagaatatgcagttgc cctgaccgaaagtagaccttggtgttctgggaaagcagcagtttgacttggaaaataag ttcaaaccattcacagtggaaattgtggattcggtagaagcttatgctacaatgctgaga agcatctttgatttcagtgcactgaaagaactactttctgggccaaacgcagtgaagatc cgtattgatgctatgcatggagttgtgggaccgtatgtaaagaagatcctctgtgaagaa ctcggtgcccctgcgaactcggcagttaactgcgttcctctggaggacttttggaggccac cacctgaccccaacctcacctatgcagctgacctggtggagaccatgaagtcaggagag catgattttgggggctgcctttgatggagatgggatcgaaacatgattctgggcaagcat gggttctttgtgaaccttcagacttctggtgtcattgctgccaacatctttcagcatt ccgtatttccagcagactgggggtccgcggctttgccagggagcatgcccacgagtggtgct ctggaccgggtggctagtgctacaaagattgctttgtatgagaccccaactggctggaag ttttttgggaatttgatggacgcgagcaaactgtcctttgtgggaggagagcttcggg acggttctgaccacatccgtgagaaagatggactgtgggctgtccttgcctggctctcc atcctagccaccgcaagcagagtgtggaggacattctcaaagatcattggcaaaagtat ggccggaatttcttcaccaggtatgattacgaggaggtggaagctgagggcgcaaacaaa atgatgaaggacttggaggccctgatgtttgatgcgtcctttgtggggaagcagttctca gcaaatgacaaagtttacactgtggagaaggccgataactttgaatacagcgacccagtg gatggaagcatttcaagaaatcagggcttgggcctcattttcacagatggttctgaatc gtctttccgactgagcggcactgggagtgccgggggccaccattcggctgtacatcgatagc tatgagaaggacgttgccaagattaaccaggaccccaggtcatgttggccccctttatt tccattgctctgaaagtgtcccagctgcaggagaggacgggacgcactgcaccccactgtc atcacctaa

All links

Pathway (6)
KEGG PATHWAY (6)
Disease (1)
OMIM (1)
Chemical reaction (1)
KEGG ENZYME (1)
Genome (1)
KEGG GENOME (1)
Gene (14)
KEGG ORTHOLOGY (1)
NCBI-Gene (1)
NCBI-GI (3)
UniGene (1)
HGNC (1)
HPRD (1)
ENSEMBL-REA (1)
Protein sequence (4)
UniProt (1)
RefSeq(pep) (1)
IPI (2)
DNA sequence (11)
RefSeq(mRNA) (1)
GenBank (5)
EMBL (5)
Protein domain (5)
Pfam (4)
PROSITE (1)
All databases (43)

DBGET integrated database retrieval system

图 8-13 详细信息页面

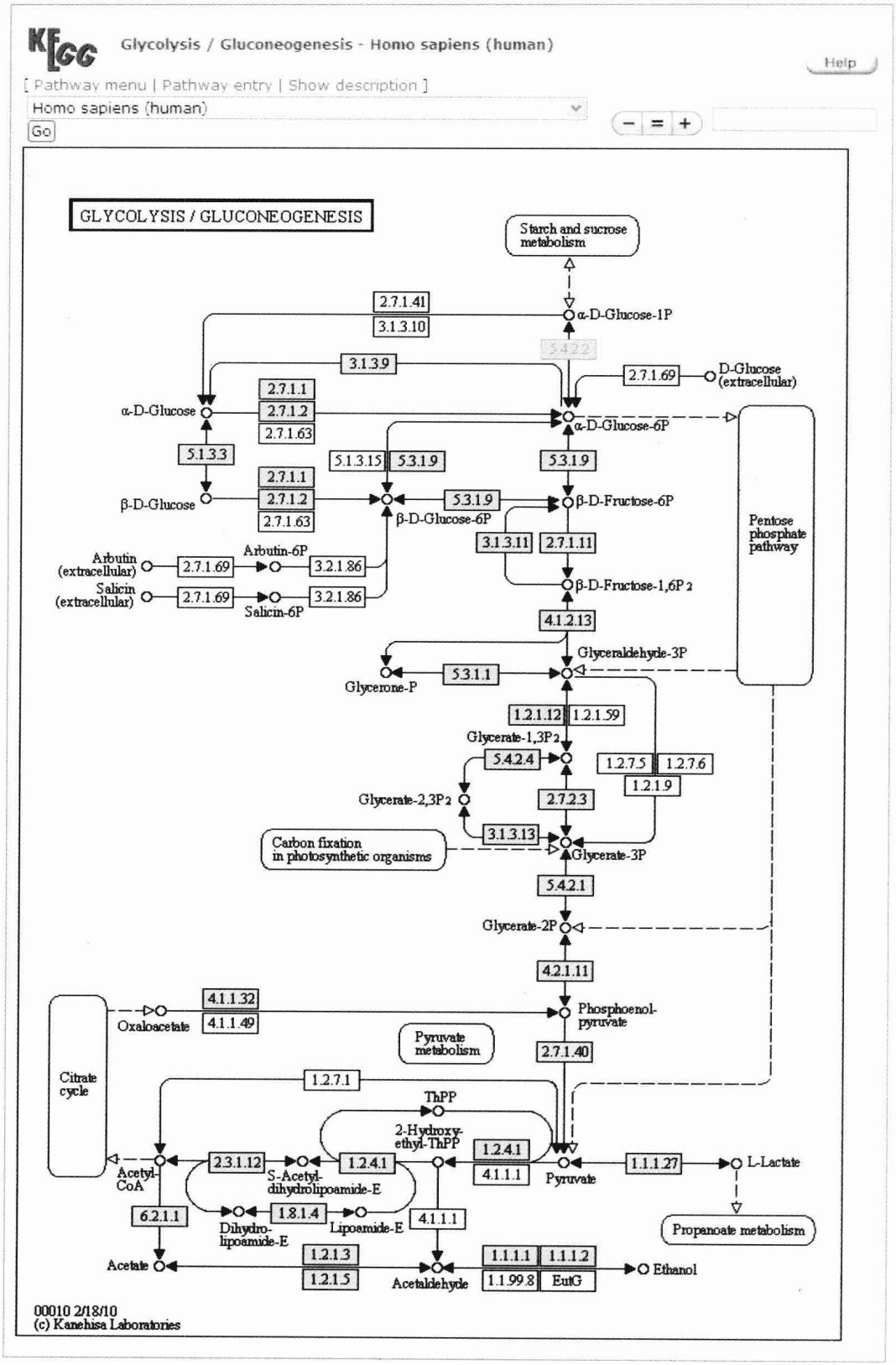

KEGG
Glycolysis / Gluconeogenesis - Homo sapiens (human)
Help
[Pathway menu | Pathway entry | Show description]
Homo sapiens (human)
Go
GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
2.7.1.41
3.1.3.10
α-D-Glucose-1P
5.4.2.2
3.1.3.9
2.7.1.69
D-Glucose (extracellular)
2.7.1.1
2.7.1.2
2.7.1.63
α-D-Glucose
α-D-Glucose-6P
5.1.3.3
5.1.3.15
5.3.1.9
5.3.1.9
2.7.1.1
2.7.1.2
2.7.1.63
β-D-Glucose
5.3.1.9
β-D-Fructose-6P
β-D-Glucose-6P
Pentose phosphate pathway
3.1.3.11
2.7.1.11
Arbutin (extracellular)
2.7.1.69
Arbutin-6P
3.2.1.86
Salicin (extracellular)
2.7.1.69
Salicin-6P
3.2.1.86
β-D-Fructose-1,6P2
4.1.2.13
Glyceraldehyde-3P
Glycerone-P
5.3.1.1
1.2.1.12
1.2.1.59
Glycerate-1,3P2
5.4.2.4
1.2.7.5
1.2.7.6
1.2.1.9
Glycerate-2,3P2
2.7.2.3
3.1.3.13
Carbon fixation in photosynthetic organisms
Glycerate-3P
5.4.2.1
Glycerate-2P
4.2.1.11
4.1.1.32
4.1.1.49
Oxaloacetate
Phosphoenol-pyruvate
Pyruvate metabolism
2.7.1.40
Citrate cycle
1.2.7.1
ThPP
2-Hydroxy-ethyl-ThPP
1.2.4.1
1.2.4.1
4.1.1.1
Pyruvate
1.1.1.27
L-Lactate
2.3.1.12
1.2.4.1
Acetyl-CoA
S-Acetyl-dihydrolipoamide-E
6.2.1.1
1.8.1.4
Dihydro-lipoamide-E
Lipoamide-E
4.1.1.1
Propanoate metabolism
1.2.1.3
1.2.1.5
1.1.1.1
1.1.1.2
Acetate
Acetaldehyde
1.1.99.8
EutG
Ethanol
00010 2/18/10
(c) Kanehisa Laboratories

图 8-14 通路图

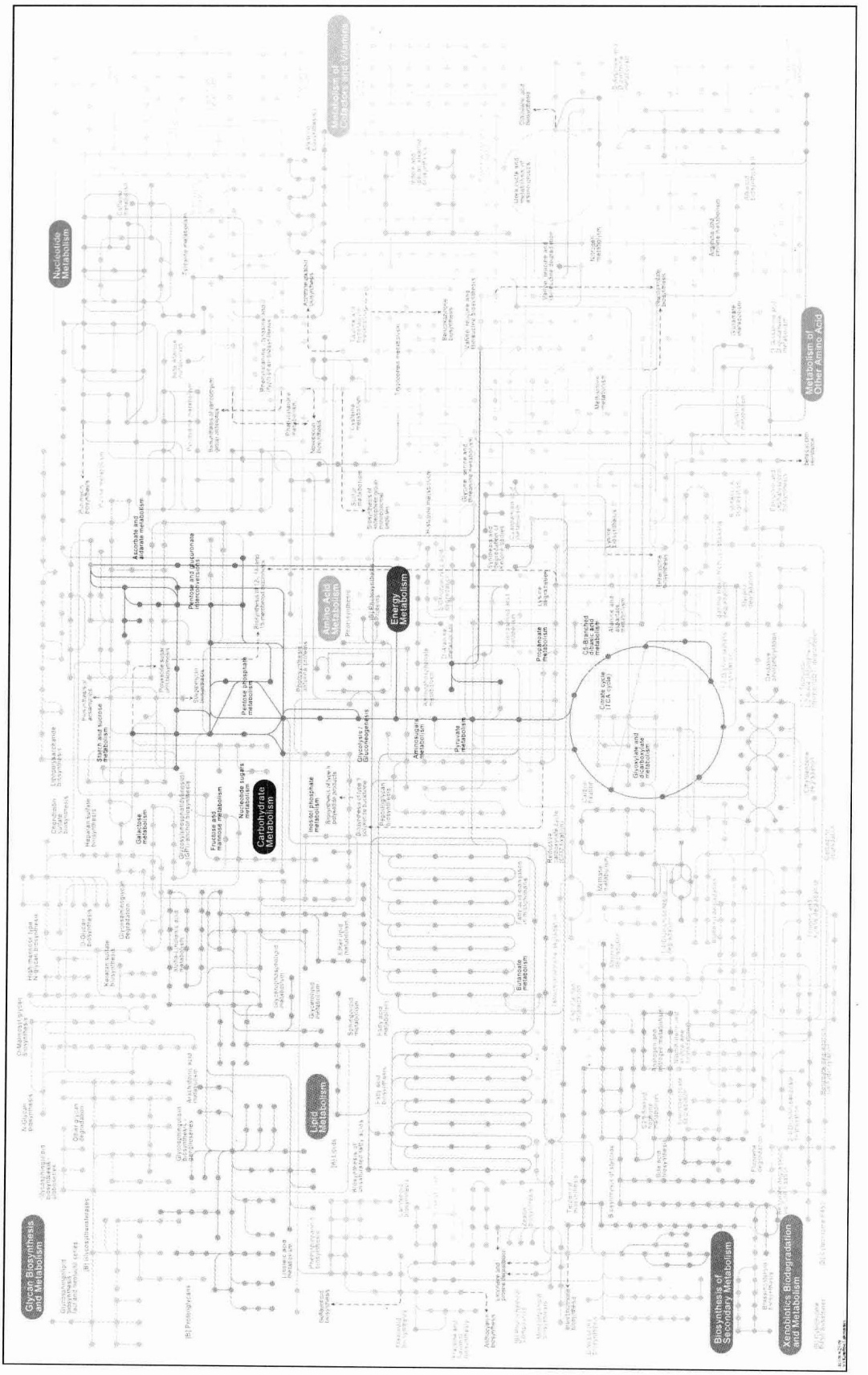


图 8-15 KEGG 全局通路图

此外，可以通过页面顶部的下拉列表框来选择该通路在其他物种中的信息，也可以通过该列表框的选择来查看相关的基因、酶、反应、化合物等相关通路信息。

（3）KEGG 数据库在医疗和药物研究中的应用：KEGG PATHWAY 还存储了一些人类疾病通路数据，这些疾病通路被分为六个子类：癌症、免疫系统疾病、神经退行性疾病、循环系统疾病、代谢障碍、传染病循环系统疾病。尽管这些疾病通路数据在快速地增加，但大多数的数据片段还很零散，还没有组建出完整的疾病通路。

KEGG DRUG 数据库也在不断地完善，其中的药物数据几乎涵盖了日本的所有非处方药和美国的大部分处方药品。DRUG 是一个以存储结构为基础的数据库，每条记录都包含唯一的化学结构以及该药物的标准名称，以及药物的药效、靶点信息、类别信息等。药物的靶点通过 KEGG PATHWAY 查询，药物的分类信息是 KEGG BRITE 数据库的一部分，通过药物的标准名称可以找到该药物的商品名，还可以找到药物销售的标签信息。此外，DRUG 还包括一些天然的药物和中药的信息，有些药物被日本药典所收录。

（4）KEGG 数据库的改进与更新：为了满足日益增长的科学研究需求，KEGG 数据库在最近几年里不断扩充，新增加的 50 多个通路使 KEGG PATHWAY 数据库更加完善。这 50 多个新增加的通路包括信号传导通路、细胞生物过程通路和人类疾病通路等。KEGG 对通路数据新增了两个补充内容：第一个补充的是一张全局通路图（图 8-15），这张全局通路图是通过手工拼接 KEGG 的 120 多个现存通路图生成的，存储为 SVG 文件。在全局通路图中每个结点（在图中以圆圈表示）代表一种化合物，两个结点的连线（包括直线以及曲线）代表若干个连续的生化反应。这张全局通路图为研究人员提供了整个代谢通路的分布情况，可以通过这张图对若干个代谢通路进行比较。同时，这张通路对应的 XML 文件也便于操作。另一个补充内容是 KEGG MODULE 数据库，这是一个收集了通路模块以及其他一些功能单元的新型数据库，功能模块是在 KEGG 子通路中被定义为一些小的片段，通常包括几个连续的反应步骤、操纵子、调控单元，以及通过基因组比对得到的系统发生单元和分子的复合物等。

第三节 基因集功能富集分析

Section 3 Gene Set Enrichment Analysis

已建立的基因及其产物注释数据库包含了丰富的知识和复杂的结构，促使研究人员开展以注释数据库为知识基础的基因功能研究，以便更好地利用注释系统。

一组基因直接注释的结果是得到大量的功能结点。这些功能具有概念上的交叠现象，导致分析结果冗余，不利于进一步的精细分析，所以研究人员希望对得到的功能结点加以过滤和筛选，以便获得更有意义的功能信息。目前最常用的方法是基于 GO 或 KEGG 的富集分析。人们通过多种方法获得大量的感兴趣基因，如差异表达基因集、共表达基因模块、蛋白质复合物基因簇等，然后寻找这些感兴趣基因集显著富集的 GO 结点或 KEGG 通路，这有助于指导进一步深入细致的实验研究。

一、富集分析算法

一个生物过程通常是由一组基因共同参与，而不是单个基因单独完成。富集分析的主要依据是，如果一个生物学过程在已知的研究中发生异常，则共同发挥功能的基因极可能被选择出来作为一个与这一过程相关的基因集合。因此，富集分析方法通常是分析一组基因在某个功能结点上是否过出现（over-presentation）。这个原理可以由单个基因的注释分析发展到大基因集合的成组分析。由于分析的结论是基于一组相关的基因，而不是根据单个基因，所以富集分析方法增加了研究的可靠性，同时也能够识别出与生物现象最相关的生物过程。富集分析中常用的统计方法有累计超几何分布、Fisher 精确检验等。

累计超几何分布公式：

$$P(X>q)=1-\sum_{x=1}^{q}\frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}} \qquad \text{式 8-1}$$

其中 N 为注释系统中基因总数，n 为将要考察的结点或通路本身注释的基因数，m 为感兴趣的基因集大小，x 为基因集与结点或通路的交集数目。

Fisher 精确检验公式：

$$P=\frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \qquad \text{式 8-2}$$

n 为系统中基因总数，a 为感兴趣的基因集中的基因数目，b 为将要考察的结点或通路本身所注释的基因数目，c 为去除感兴趣基因以外的基因数目，d 为待考察结点基因去除与感兴趣基因重合的数目。

还有多种统计方法可以应用于富集分析，如 Z-score，Kolmogorov-Smirnov-like statistic 等，这里不再详细介绍。此外，由于在进行富集分析时通常需要同时进行大量检验(多重检验)，所以需要采用多重检验校正的方法对检验结果进行校正。这些方法主要包括邦弗朗尼校正(Bonferroni correction)，邦弗朗尼递减校正(Bonferroni step down correction)，本杰明假阳性率校正(Benjamini false discovery rate correction)等。

二、常用富集分析软件

利用富集分析方法，对基因注释数据库做生物信息学研究产生了很多富集分析工具。这些工具对促进基因功能分析以及研究高通量的生物学数据起到了重要作用。表 8-3 列举一些常用富集分析工具。基于不同的算法原理，可以将目前常用的富集分析工具分为三类：单一富集分析(singular enrichment analysis)、基因集富集分析(gene set enrichment analysis)、模块富集分析(modular enrichment analysis)。

第一类富集分析方法利用预先选定的注释基因计算每个 GO 结点的显著性，之后显著富集的结点被列出，这一方法是最传统的算法，也是最常用的方法。

第二类基因集富集分析方法的特点是：无需预先选择感兴趣的基因集；实验值整合成 P 值计算。

第三类模块富集分析方法继承了单一富集分析的主要思想，但是在计算 P 值是考虑了结点间或基因间关系。这一方法的优点是考虑了结点间或基因间关系的生物学意义，而这些生物意义无法由单个基因体现。这种模块化的分析更接近生物数据结构的本质。常用富集分析工具的分类结果列在表 8-3 中。

表 8-3 常用富集分析工具集

Enrichment tool name	Year of release	Key statistical method	Category
Onto-express	2002	Fisher's exact;hypergeometic; binomial; chi-square	Class I
GOArray	2004	Hypergeometric; Z-score; permutation	Class I
FACT	2005	Adopt GeneMerge and GO: TermFinder statistical modules	Class I
BayGO	2006	Bayesian; Goodman and Kruskal's gamma factor	Class I
Gene Class Expression	2006	Z-statistics	Class I
GOALIE	2006	Hidden Kripke model	Class I

续表

Enrichment tool name	Year of release	Key statistical method	Category
JProGO	2006	Fisher's exact; Kolmogorov-Smirnov test; student's t-test;	Class I
ProbCD	2007	Yule's Q; Goodman-Kruskal's gamma; Cramer's T	Class I
GO-Mapper	2004	Gaussian distribution; EQ-score	Class II
iGA	2004	Permutations; hypergeometric; t-test; Z-score	Class II
GSEA	2005	Kolmogorov-Smirnov-like statistic	Class II
GAzer	2007	Z-statistics; permutation	Class II
POSOC	2004	POSET (a discrete math: finite partially ordered set)	Class III
GENECODIS	2007	Hypergeometric; chi-square	Class III
GOSim	2007	Resnik's similarity	Class III
PalS	2008	Percent	Class III
ProfCom	2008	Greedy heuristics	Class III
ermineJ	2005	Permutations; Wilcoxon rank-sum test	Class I,II
DAVID	2003	Fisher's Exact (modified as EASE score)	Class I,III
GOToolBox	2004	Hypergeometric; Fisher's exact; Binomial	Class I,III
ADGO	2006	Z-statistic	Class II,III

三、富集分析应用实例

上面介绍了多种富集分析工具，这里以目前应用较为广泛的 DAVID 为例对基因集进行具体分析。DAVID 是一个综合工具，不但提供基因富集分析，还提供基因间 ID 的转换、基因功能的分类等（图 8-16）。

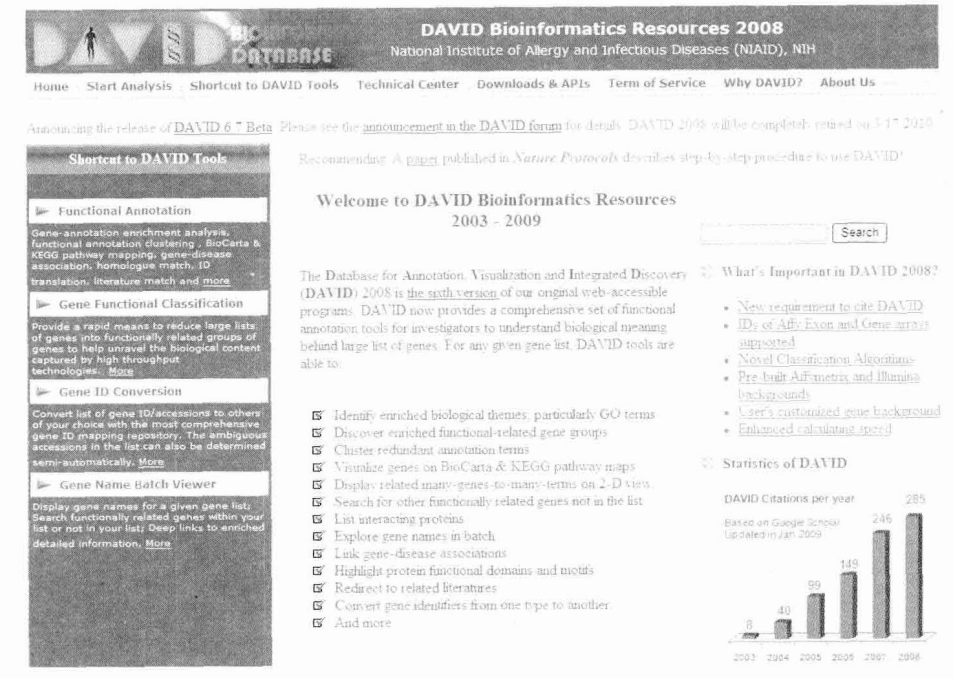


图 8-16 DAVID 工具应用首页

点击“Start Analysis”后，第一步为提交基因集，选择基因标识名和基因集类型；第二步得到注释结果摘要，包括多种注释数据；然后选择感兴趣的注释内容得到富集分析结果，见图 8-17。

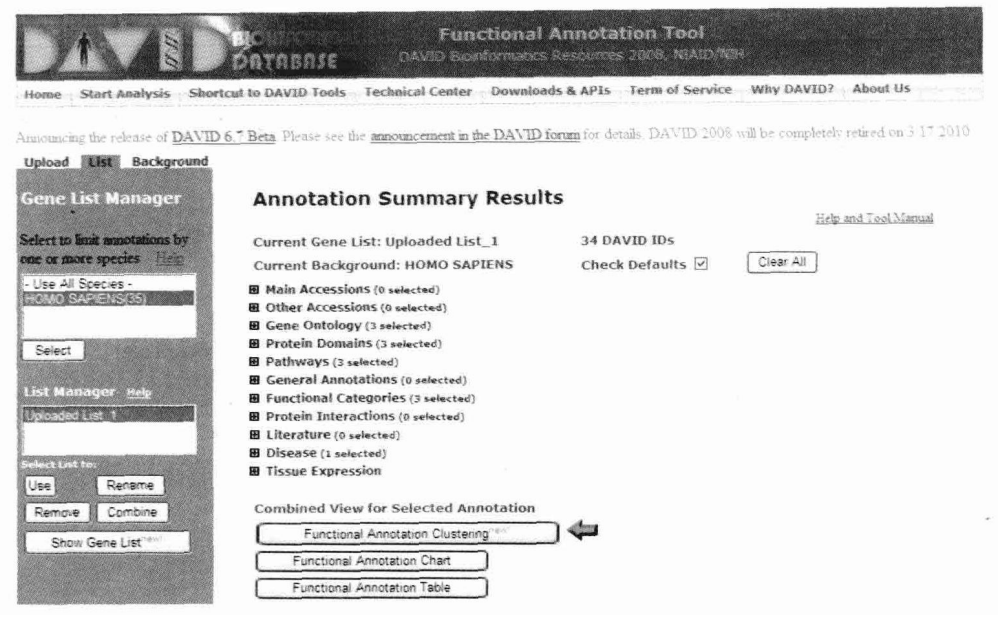


图 8-17 DAVID 富集分析注释结果摘要

这里以 KEGG 通路的富集分析为例。提交之后的结果如图 8-18，可以看到，对提交的基因集做富集分析，找到 5 个具有显著性的通路。这里的“P-Value”是通过 Fisher 精确检验得到的 P 值，“Benjamini”指的是本杰明假阳性率校正方法。

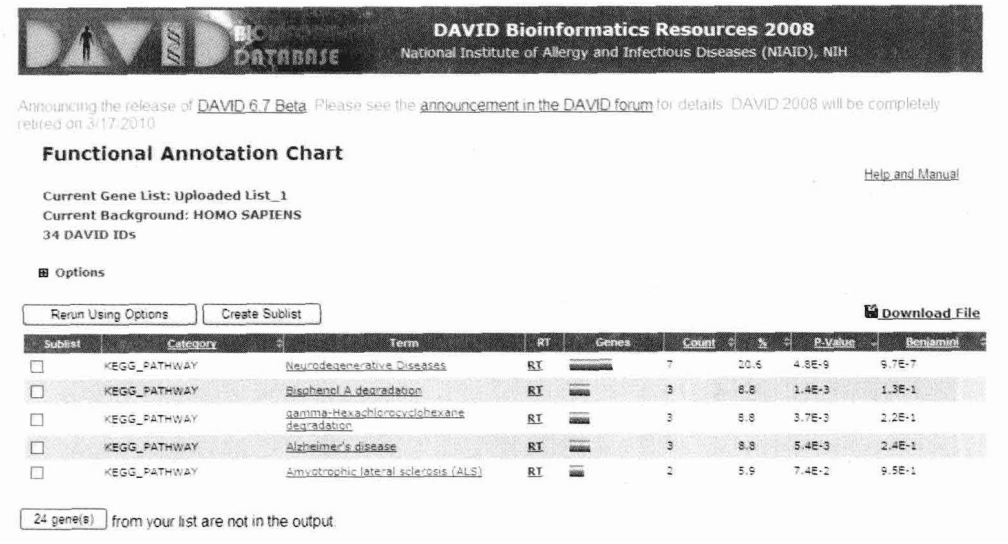


图 8-18 DAVID 在 KEGG 上富集结果实例

第四节 基因功能预测

Section 4 Gene Function Prediction

一、基因功能预测算法

目前，大量参与重要生命活动的基因功能仍然未知。因此，生物信息学的重要任务之一是在全基因组范围内对基因功能进行预测。传统的基因功能预测方法主要依赖于序列的同源性，而近来已经发展了很多基于 GO 数据库或 KEGG 数据库的方法，利用高通量的基因表达和蛋白质互作数

据进行功能预测，其中一些新开发的方法试图整合多种数据类型，通过构建功能相关网络的方式预测基因功能(图 8-19)。GO 数据库包含了基因参与的生物过程，所处的细胞位置及具有的分子功能三方面功能信息，通过 GO 中的注释信息，可以对基因的功能进行预测。KEGG 是系统分析基因功能、联系基因组信息和功能信息的知识库，KEGG 的 PATHWAY 数据库提供了基因编码的生物学大分子酶或者蛋白质在生命体内相互联系相互影响的情况。同一生物学通路内的基因大多参与了此代谢通路所揭示的生命过程。根据功能相似的基因可能导致相似的表型这一依据，可以通过网络拓扑性质对基因的功能进行预测，并利用 GO 和 KEGG 功能富集分析方法进行进一步的预测。当前基于 GO 或 KEGG 的基因功能预测策略一般为：首先，从总体上宏观地概括抽取信息，如不同样本间、不同时间点间全部差异基因；其次，通过 GO 或 KEGG 分析，即从 GO 分类结果找到实验涉及的显著功能类别或将差异基因映射到通路中，根据基因在通路中的位置及表达水平的变化算出受影响显著的通路，从而预测未知的基因功能等。

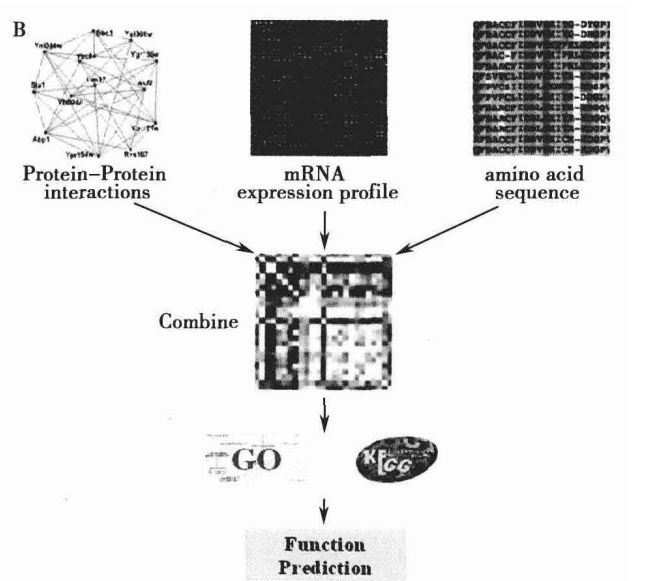


图 8-19 整合蛋白质互作数据、表达谱和序列数据的功能预测

基于 GO 或 KEGG 的基因功能预测通常需要定义基因集，基因集的定义基于统一的先验生物学知识，如已发表的有关基因共表达、生物通路等。一个基因集是基因芯片上一组具有相同生物学功能或位于同一生物通道的基因，产生基因集的数据包括基因表达谱数据和蛋白质互作数据。

(一) 基于 GO 的基因功能预测

1. 对差异表达基因进行功能预测 GO 应用的一个重要方面就是用来指导基于基因表达谱数据的基因功能预测。在基因芯片的数据分析中，研究者可以找出哪些差异表达基因属于一个共同的 GO 功能分支，并用统计学方法检验结果是否具有统计学意义，从而得出差异表达基因主要参与了哪些生物功能。

目前，大量的基因功能预测方法利用 GO 作为功能分类的来源或结果证实。在已知的大多数相关研究中，研究者首先将感兴趣基因注释到 GO 上，然后筛选出显著性富集的 GO 结点作为功能标签，考察这组基因是否共同注释到同一个功能结点上，或注释的结点是同一个结点的直接子结点，并认为这样的基因具有相似的功能，这项工作实现了对未知基因功能预测，是 GO 结构信息的进一步发掘。这是直接利用 GO 注释的方法进行基因功能预测。

目前许多已知功能的基因只注释到了描述很不具体的功能类，称之为已知部分功能的蛋白质。显然寻找这些基因的精细功能对于了解这些基因和提供必要的数据来学习其他基因的功能都具有重要意义。为了寻找已知部分功能的基因更精细的功能，目前有一种深层预测算法：该算法利用蛋白质互作数据，将基因从其已注释到的功能类向下预测一层或多层，发现其更精细的功能。由于已知部分功能的基因参与一个子功能类的先验概率增大，预测的可靠性可能会提高，因此使用注释到同一个功能类中的基因可以过滤部分假阳性互作。

具体做法为：首先，选定一个 GO 结点作为深层预测的目标结点，定义它的任何一个祖先结点为预测空间，按照 GO 注释的提示，将注释到预测空间而没有注释到它的任何一个子结点的基因定义为已知部分功能的基因，即预测对象；然后，通过连接注释在预测空间中互作的基因构建一个功能特异的子网，孤立的基因被排除在外，在互作子网中，注释到目标结点的蛋白质被当作阳性样本，而除

预测对象外的其他蛋白质被当作阴性样本。

通常一个蛋白质被赋予与其直接相互作用的邻居蛋白质中出现频率最高的几个功能。尽管一个蛋白质可以执行多个功能，这里选择只为蛋白质赋予一个可信度最高的子功能。因为目标结点中阳性样本要和预测空间中所有其他子结点的阴性样本竞争，因此对于预测一个阳性结果来说是保守的。可以采用留一法来评价分类器的预测效果。每一个训练样本都要被轮流留出来作为测试样本。计算真阳性(TP)、真阴性(TN)、假阳性(FP)和假阴性(FN)，再计算精确度、覆盖率和 F 指标。基于蛋白质互作数据和深层预测方法，以高于 90% 的精确率，为几千个已知部分功能的酵母和人类蛋白质预测了精细的功能。预测的精细功能对于指导随后的实验和提供必要的功能知识来学习其他蛋白质的功能都具有重要的意义。

2. 蛋白质互作网络用于基因功能预测　传统的基因功能注释及预测方法是根据基因相关的一些统计特征集，利用机器学习方法来得出功能注释的规则用于预测。基因功能实现的复杂性以及功能定义的模糊性，使得传统的利用特征预测的方法很难准确地进行预测。而蛋白质相互作用网络能够利用蛋白质之间的相关性，对未知功能的基因进行注释。目前，利用相互作用网络进行功能注释主要有两种方法，即直接注释方法(direct annotation schemes)和基于模块的方法(module assisted schemes)。

(1) 直接注释方法：直接注释方法根据网络中某个蛋白质的连接情况直接推测该蛋白质的功能。这类方法基于的假设是：在蛋白质相互作用网络中，距离相近的两个蛋白质更加倾向于拥有相似的功能。而通过两蛋白质在网络中的距离来计算并判断这两个蛋白质功能相似性有许多的方法：①邻居结点计算法(neighborhood counting)：这种方法是最简便也是相对较早出现的方法。它根据网络中某个蛋白质直接相关的邻居已知蛋白质的功能来确定该未知蛋白质的功能注释。这种方法假设某未知蛋白质的邻居中有超过 n 个蛋白质具有一样的功能，就将这种功能赋予该蛋白质。这种方法虽然简单并且有时候非常有效，然而它在功能注释过程中不能为这种关联性提供非常有显著意义的解释，并且它也没有考虑到网络的全局拓扑结构。②图论方法(graph theoretic method)：图论方法不同于邻居结点计算法，它可以考虑网络的全局拓扑结构。基本思路是：对一个未知功能蛋白质赋予某种功能，要使得注释为相同功能的蛋白质(未注释或者已注释)的连接数目最多。③马可夫随机场方法：注释方法中有许多基于概率的方法，它们均基于马可夫假设，即蛋白质的功能独立于与其直接相邻的邻居之外的所有蛋白质。根据这个假设，人们也提出了马可夫随机场模型用于蛋白质功能的注释。

(2) 基于模块的方法：基于模块的方法首先将网络中相关的蛋白质组成不同的模块，然后根据该模块中成员的功能来得到整个模块所共有的可能的功能，从而用来预测其中未知成员的功能。一个功能模块指其中的蛋白质所处的细胞位置以及相互作用使得它们可以实现一个特定的功能。而基于功能模块的蛋白质功能注释方法也不再单独的预测单个蛋白质的功能，而是试图发现模块中所有蛋白质的共同内在的功能。一旦模块确定，那么可以通过一些简单的方法来预测其功能，比如该模块中如果大部分的蛋白质都具有某种功能，那么这种功能就将赋予该模块。对蛋白质相互作用网络进行模块划分的常用方法有以下几种：①分级聚类方法(hierarchical clustering based methods)：聚类就是将相似功能的蛋白质归为同一类(模块)。分级聚类的关键问题是如何评判蛋白质对之间的相似性，最简单的方法是以两个蛋白质之间的距离作为基准。但是在分级聚类中，大量蛋白质对之间的距离都是相同的，通常认为同一个模块中的蛋白质成员更加可能拥有最短的路径距离谱(path distance profiles)。根据这个假设，所有短路径的蛋白质对聚成一类。这个方法实施比较复杂，很难在整个基因组水平上的网络上进行分析，但在一些子网络中它已经得到很好的应用，比如对酿酒酵母的核蛋白的相互作用网络分析。②图形聚类方法(graph clustering methods)：大量的图形聚类方法也用于图形化描述二元相互作用。早期的图形聚类方法用于相互作用网络模块的构建主要有两类，一类是基于 SPC 聚类(super paramagnetic clustering)方法，另一类为基于蒙特卡洛算法(monte

carlo algorithm)。其中 SPC 算法在决定那些内部密度很高但松散的连接于其他部分的模块效果非常好。在最近，又不断发展出许多新的图形聚类算法，如高连通子图算法(highly connected sub graphs，HCS)、有限邻居搜索聚类算法(restricted neighborhood search clustering，RNSC)以及马可夫聚类算法(Markov clustering，MCL)等。

3. 利用 GO 体系结构比较基因功能　此外，还有一些基于信息理论的相似性概念比较基因间的功能相似性，从而对基因功能进行预测。通常认为如果两个基因产物的功能相似，那么它们的表达也就相近，同时它们在 GO 中注解的结点就相似，所以只要能找出 GO 中结点对的相似度，就可以近似估计两基因表达的相似度，从而判断两基因产物的功能的相似度。被人们广泛了解的是 Resnik 在 1995 年提出的对分类系统中的每个类定义的语义相似性算法，计算两个类的语义相似性，后有多位科学家经过改进等提供了多种类相似性的计算测度。在 2002 年 Lord 第一次提出把语义相似性理论应用到 GO 分类系统中，计算两个结点之间的相似性，从而可以利用不同的方法计算基因间的功能相似性，最后可以根据功能相似性得分预测未知基因的功能。

在分类系统中，利用 GO 结构信息和基因注释信息，首先设一个函数，计算得到每个结点的信息含量值：$p(c)=\frac{freq(c)}{N}$，$freq(c)$表示结点及它的子结点上注释的所有基因数，$p(c)$是结点 c 的概率，并且随着结点 c 在层级结构中的升级，概率 p 是单调递增的，top 结点概率是 1。越往上层，概率越大，信息含量越小。即如果 $c1$ 是 $c2$ 的下属，则 $p(c1)\leqslant p(c2)$。则结点 c 的信息含量值为：$IC=-\log(p(c))$。得到每个结点的信息含量值后，计算任意两个结点的相似性方法有多种，Resnik 最早提出语义相似性概念，它的定义为两个结点的公共祖先中最近距离的祖先结点的 IC 值即为它们的相似性值，即：

$$sim(c1,c2)=\max_{c\in S(c1,c2)}[-\log p(c)] \qquad \text{式 8-3}$$

$$sim(c1,c2)=\frac{2IC_{ms}(c1,c2)}{IC(c1)+IC(c2)} \qquad \text{式 8-4}$$

在 GO 系统中，可以计算得到任意两个结点的相似性值，则可根据基因注释在哪些结点上而计算两个基因之间的功能相似性。最简单的方法是取两个基因所注释的结点对的最大值或平均值，来作为两个基因的功能相似性，还有最优分配法，目前已经有一些比较基因间的关联程度的算法和工具，利用语义相似性原理计算基因间的功能相似性的工具已经有 GOSim、csbl.go、G-SESAME 等。

(二) 基于 KEGG 通路分析的基因功能预测

通路分析是现在经常被使用的芯片数据基因功能分析法。与 GO 分类法(应用单个基因的 GO 分类信息)不同，通路分析法利用的资源是许多已经研究清楚的基因之间的相互作用，即生物学通路。研究者可以把表达发生变化的基因集导入通路分析软件中，进而得到变化的基因都存在于哪些已知通路中，并通过统计学方法计算哪些通路与基因表达的变化最为相关。现在已经有丰富的数据库资源帮助研究人员了解及检索生物学通路，对芯片的结果进行分析。主要的生物学通路数据库有以下两个：①KEGG 数据库：迄今为止，KEGG 数据库(Kyoto encyclopedia of genes and genomes)是向公众开放的最为著名的生物学通路方面的资源网站。在这个网站中，每一种生物学通路都有专门的图示说明；②BioCarta 数据库：它在其公共网站上提供了用于绘制生物学通路的模板。研究者可以把符合标准的生物学通路提供给 BioCarta 数据库。BioCarta 数据库不会检验这些生物学通路的质量，因此其中的资源质量参差不齐，并且有许多相互重复。然而 BioCarta 数据库数据量巨大，且不同于 KEGG 数据库，包含了大量代谢通路之外的生物学通路，所以也得到广泛的应用。如图 8-20。

芯片数据通路分析的第一步是差异基因的通路定位，一些商业软件如 Genespring 可以做到，基于 EASE 算法的开放在线程序 DAVID 也可以实现定位。目前的通路分析方法还存在很多局限性，例如，只注意到基因集合定位到了哪个通路而忽略了其在通路中的位置，如果一个通路由某个基因

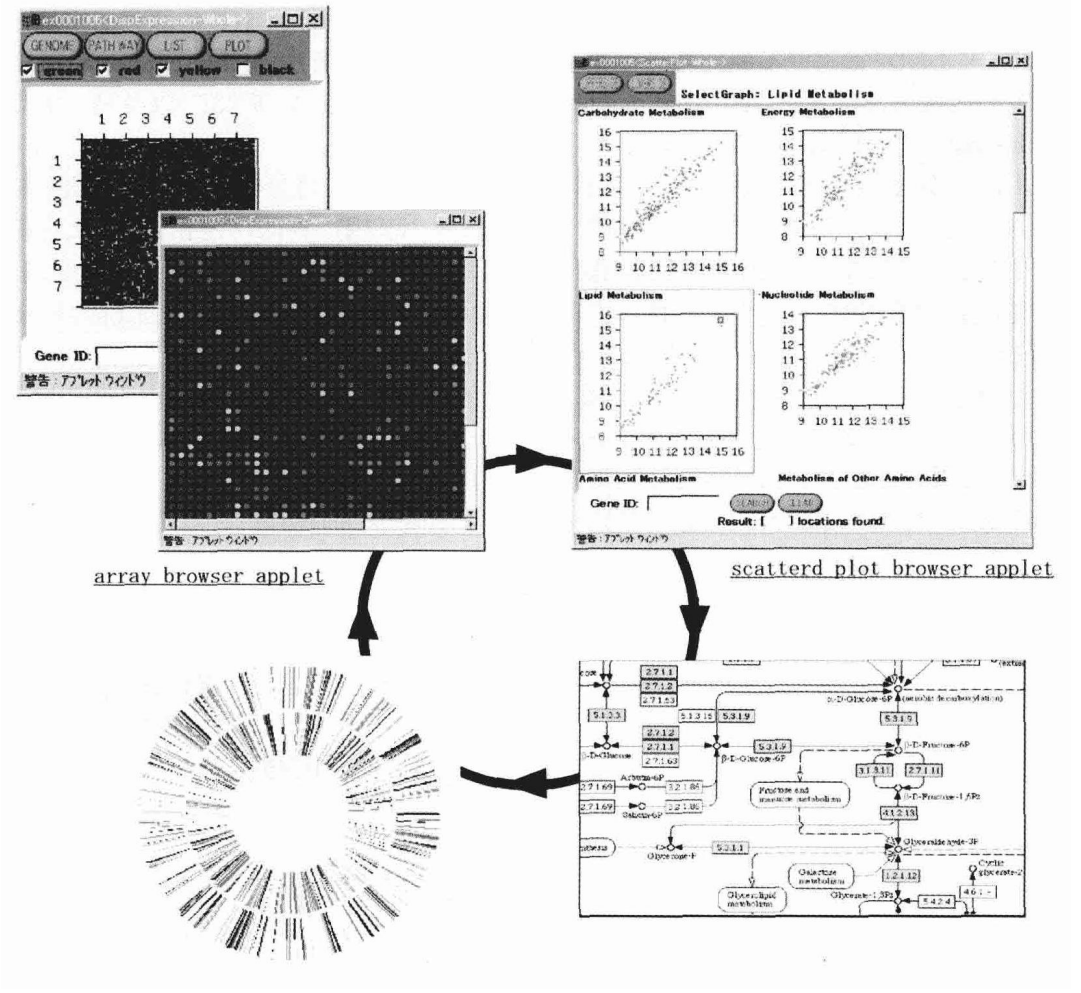


图 8-20 通过表达谱数据进行通路定位

产物触发或被单个受体激活，并且特定的蛋白质没有表达，这个通路就会受到严重影响甚至关闭；相反，如果多个基因与某个通路相关但都只出现在通路的下游，那么其表达水平的变化就可能不会对通路造成很大影响。另外，一些基因往往有多个功能分布于不同的通路发挥不同的作用，要得到相对准确的结果还必须考虑通路的拓扑结构。目前很少有能将基因差异表达值变化应用于通路分析的方法，Pathwayexpress 提出了一种基于 IF(impact factor)的通路分析方法，综合了差异基因的标化的差异表达值、通路中基因的统计学显著性以及信号通路的拓扑学三方面内容。Pathwayexpress 主要基于 KEGG 库，结果输出中自动把差异基因以不同颜色定位于通路中，红色为上调，蓝色为下调，这些定位着上调和下调基因的通路图可以在 Java 控制台中找到绝对路径，在浏览器中打开或保存，也可以 GML 格式导出，然后直接导入 Cytoscape，用 merge 结点功能把多个相关 pathway 连接起来，显示互作网络，并分别以红蓝色显示显著性通路中上调下调的基因(结点)，以及这些基因与其他基因间的相互作用(边)，可以从不同视角观察其位置，不断放大就可以看到结点的基因名称。其他的可视化工具还有 pathwaystudio、genmapp、arrayxpath、osprey 等。Biolayout 也是一款分子作用网络展示工具，所不同的是结果为三维图形界面。

二、常用基因功能预测软件

(一) 基于 GO 的基因功能分析软件

EASE(expressing analysis systematic explorer)是比较早的用于芯片功能分析的网络平台。由美国国立卫生研究院(NIH)的研究人员开发。研究者可以用多种不同的格式将芯片中得到的基因导入

EASE进行分析，EASE会找出这一系列的基因都存在于哪些GO分类中。其最主要特点是提供了一些统计学选项以判断得到的GO分类是否符合统计学标准。EASE能进行的统计学检验主要包括Fisher精确概率检验，或是对Fisher精确概率检验进行了修饰的EASE得分(EASE score)。

由于进行统计学检验的GO分类的数量很多，所以EASE采取了一系列方法对“多重检验”的结果进行校正。这些方法包括Bonferroni校正法、Benjamini falsediscovery rate和bootstraping。同年出现的基于GO分类的芯片基因功能分析平台还有底特律韦恩大学开发的Onto-Express。2002年，挪威大学和乌普萨拉大学联合推出的Rosetta系统将GO分类与基因表达数据相联系，引入了“最小决定法则”(minimal decision rules)的概念。它的基本思想是在对多张芯片结果进行聚类分析之后，与表达模式不相近的基因相比，相近的基因更有可能参与相同的生物学功能的实现。比较著名的基于GO分类法的芯片数据分析网络平台还有很多，这里列举了其中的一部分(表8-4)。

表8-4 用GO分类法进行芯片功能分析的网络平台

Name	Internet Site
Onto-Tools	http://vortex.cs.wayne.edu/projects.htm
ROSETTA	http://rosetta.lcb.uu.se/general/
GOToolBox	http://burgundy.cmmt.ubc.ca/GOToolBox/
GOstat	http://gostat.wehi.edu.au/
GFINDer	http://www.medinfopoli.polimi.it/GFINDer/
FatiGO	http://www.fatigo.org/
EASE	http://david.abcc.ncifcrf.gov/ease/ease.jsp

(二) 基于KEGG的基因功能分析软件

最先出现的通路分析软件之一是GenMAPP(gene microarray pathway profiler)，它可以免费使用，其最新版本为Gen-MAPP2。在这个软件中，使用者可以用几种灵活的文件格式输入自己的表达谱数据，GenMAPP的基因数据库包含许多从常用的资源中得到的物种特异性的基因注释和识别符(ID)。这些ID可以将使用者输入的基因与不同的生物学通路的基因联系起来。这些生物学通路存在于GenMAPP的MAPP文件中。MAPP文件需要时常下载更新。它包含有许多KEGG生物学通路，一些GenMAPP自己的生物学通路和许多GO分类的MAPP文件，全部操作简单明了。而且依靠其自带的MAPPBuilder和MAPPFinder两个软件，使用者可以自己绘制生物学通路和对MAPP文件进行检索。由于使用者可以自己绘制生物学通路保存为MAPP格式，这个文件很小，易于在网络上传播，所以GenMAPP数据库更有利于研究者之间的及时交流。由于上述特点，GenMAPP数据库及软件仍是现今免费平台里应用比较广泛的。

2004年发表的Pathway Miner也是应用较为广泛的免费通路分析网络平台，由美国亚利桑那大学癌症中心建立维护，其最突出的特点就是信息全面，操作简便。使用者可以在这个网站中获得单个基因的序列、功能注释，以及有关它们编码的蛋白质结构功能，组织分布，OMIM等信息。对于通路分析部分，使用者给出基因集及它们的表达变化值，网站可以根据三大公用的通路数据库：KEGG、GenMAPP和BioCarta，生成变化基因参与的通路，并用Fisher精确概率检验。PathwayMiner自动把得到的通路分成两大类：代谢通路和细胞调节通路。方便使用者根据不同的研究目的选择需要查看的结果。2006年国内也开发了用于通路分析的网络平台，即KOBAS(KO-based annotation system)，其基于KEGG数据库建立，由北京大学生命科学院开发和维护。其特点是可直接采用基因或蛋白质的序列录入基因，并对录入的基因集进行KO注释。对于结果的可靠性检验提供了四种统计方法。使用者可以在网站进行注册，网站会为使用者保存输入的数据，方便日后直接调用。最近推出的软件Eu.Gene整合了来自KEGG、Gen-MAPP以及Reactome的通路数据，并采用Fisher精确

概率检验及基因集富集分析(gene set enrichment analysis, GSEA)来检验结果是否具有统计学意义。这里列举了部分通路分析的网络平台及它们的网址(表 8-5)。

表 8-5 通路分析网络平台

Name	Internet Site
GenMAPP	http://www.genmapp.org/
PathwayMiner	http://www.biorag.org/pathway.html
KOBAS	http://kobas.cbi.pku.edu.cn
GEPAT	http://gepat.bioapps.biozentrum.uni-wuerzburg.de/GEPAT/index.faces
VitaPad	http://bioinformatics.med.yale.edu/group
KEGGanim	http://biit.cs.ut.ee/kegganim/
WholePathwayScope	http://www.abcc.ncifcrf.gov/wps/wps_index.php
VisANT 3.0	http://visant.bu.edu/
Eu.Gene	http://www.ducciocavalieri.org/bio/Eugene.htm

(三) 利用 Onto-Express 预测基因功能

Onto-Express 是 Wayne State University 开发的 Onto-Tools 软件包中的一个表达谱数据分析工具，利用 Gene Ontology 中的数据信息对基因的功能进行分析，可以在 http://vortex.cs.wayne.edu:8080/index.jsp 上免费下载该软件。

1. 数据输入　下面通过提供的测试数据阐述 Onto-Express 的使用方法，该芯片的测试数据可在 http://www.ebi.ac.uk/~jane/TestData/ 下载，输入数据为 total 和 under.over，输入数据为文本格式，包含 accession numbers，cluster identifiers 或 probe identifiers。进入 Onto-Express 的输入窗口，如图 8-21 所示。

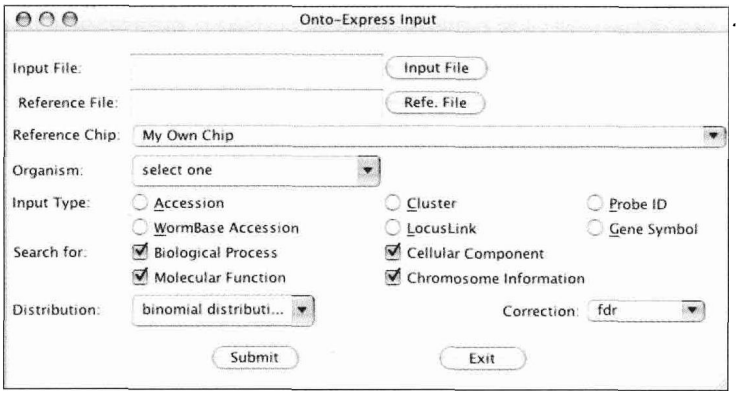


图 8-21　Onto-Express 输入窗口

点击“Input File”按钮输入 under.over 文件，这个文件包含一列实验产生的上调或下调基因集合。点击“Reference File”按钮输入 total 文件，这个文件包含了实验中的所有基因，并选择“My Own Chip”，这里注意到：如果实验中使用的是商业芯片，可以不用上传芯片的基因，选择芯片的类型即可。

对于物种选项，这里选择 homo sapien，“Input Type”选择 Gene Symbol。其他的选项默认即可，然后点击“Submit”结果将在几分钟后生成。

2. 结果页面　返回的结果页面见图 8-22。选择“Tree View”，将显示 GO 的树状图，可以单击收缩或展开显著 term 的信息。GO term 上的黑体字是输入的上调或下调基因集合注释到该 term 上的数目。P 值是该结点含有上调或下调基因的数目大于随机期望的概率。

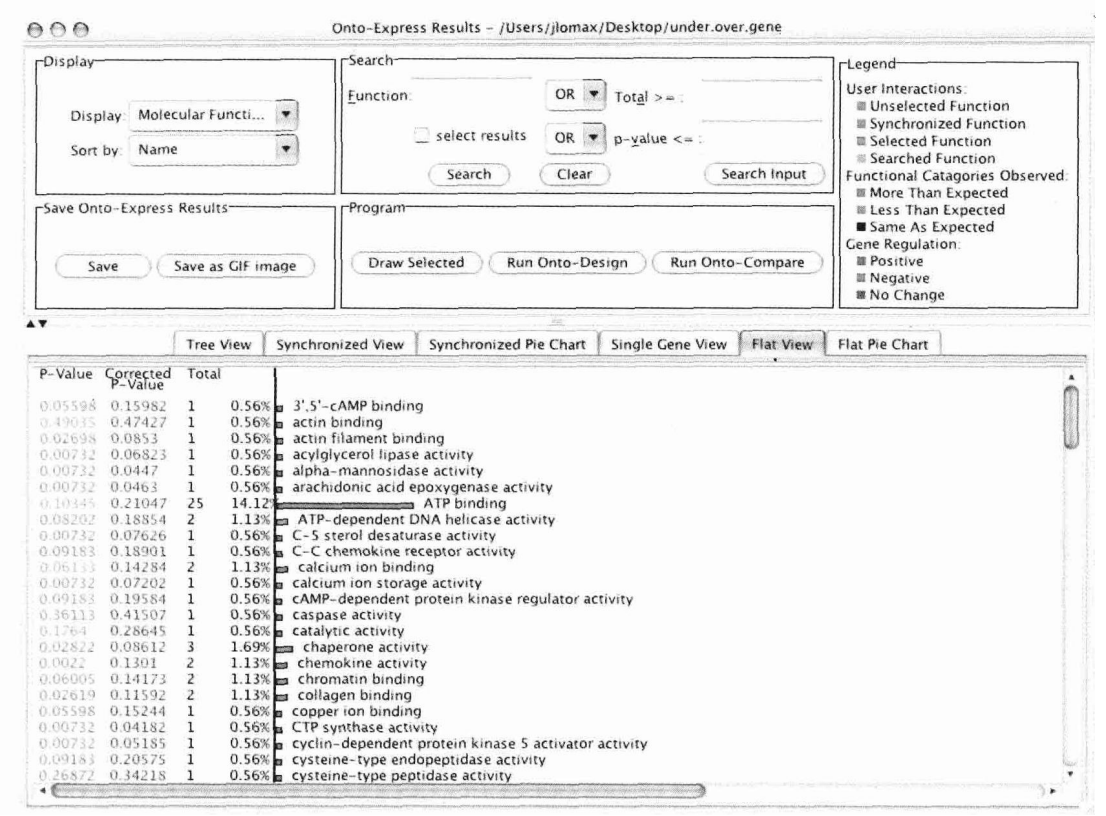


图 8-22 Onto-Express 结果窗口

小 结

基因注释与功能分类是功能基因组学和计算系统生物学的重要基础。本章重点介绍了 Gene Ontology(GO)数据库和 Kyoto Encyclopedia of Genes and Genomes(KEGG)数据库。分别从基因功能注释和通路注释两个层面阐述功能注释与分类。

随着功能基因组学在人类复杂疾病研究中应用的逐步深入，基因功能注释的尺度也逐步从单基因注释发展到多基因注释和通路(或特定功能的基因集合)注释。基于 GO 和 KEGG 发展起来的 David、GOEAST、GOSim、KEGGSpider、KEGGArray、PathwayMiner 等软件从不同角度实现注释、富集分析和功能预测，方便临床医学工作人员对感兴趣的基因或基因组进行研究。

Summary

Gene annotation and functional classification are important basis for functional genomics and computational system biology. In this chapter, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) are introduced while functional annotation and classification are summarized in terms of gene functional annotation and pathway annotation. As functional genomics are widely used in human complex disease, gene functional annotation is also improved from single gene annotation to gene set annotation and pathway annotation. Softwares such as David、GOEAST、GOSim、KEGGSpider、KEGGArray、PathwayMiner can perform annotation, enrichment analysis and functional prediction which quicken the study of gene and gene product for clinical researchers.

(李亦学 汪强虎 李 霞)

习 题

1. 富集分析方法的目的是什么?
2. 简述多重检验校正的作用。
3. 常用富集分析软件可以分为几类,请简述各类特征。
4. 应用 DAVID 找出一组基因在 GO 中 BP 分支上的显著结点。
5. 简述利用 GO 和 KEGG 进行基因功能预测的基本步骤。
6. 列举常用的基因功能预测软件和分析平台。
7. 简述基于 GO 的软件 EASE 预测基因功能的基本步骤。
8. 简述 GenMAPP 软件预测基因功能的基本步骤。

主要参考文献

1. Huang D. W., Sherman B. T., Lempicki R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 2009, 37(1): 1-13.
2. Huang D. W., Sherman B. T., Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nat. Protoc., 2009, 4: 44-57.
3. Ashburner, M., Ball, C. A., Blake, J. A., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet., 2000, 25: 25-29.
4. Hu, P., Bader, G., Wigle, D. A., et al. Computational prediction of cancer-gene function. Nature Reviews Cancer, 2007, 7(1): 23-34.
5. Murali, T. M., Wu, C. J., Kasif, S. The art of gene function prediction. Nature Biotechnology, 2006, 24(12): 1474-1476.
6. Sharan, R., Ulitsky, I., Shamir, R. Network-based prediction of protein function. Molecular System Biology, 2007, 3(88): 1-13.
7. Zhou, Y., Young, J. A., Santrosyan, A., et al. In silico gene function prediction using ontology-based pattern identification. Bioinformatics, 2005, 21(7): 1237-1245.

第九章　蛋白质分析与蛋白质组学
CHAPTER 9　PROTEIN ANALYSIS AND PROTEOMICS

第一节　引　　言
Section 1　Introduction

一、发展概述

人类基因组测序的完成标志着一个新的生物学研究时代——后基因组时代(post-genomic era)的来临，而传统的对单个蛋白质进行研究的方式已无法满足后基因组时代的要求。这是因为：①生命现象的发生往往是多因素、多水平影响的，必然涉及多个蛋白质；②多个蛋白质的参与是交织成网络的，或平行发生，或呈级联因果；③在执行生理功能时蛋白质的表现是动态的、多样的和可调控的。要全面、深入认识生命的复杂活动，必然要在整体、动态水平上对蛋白质进行系统研究。20 世纪 90 年代中期，一门新兴学科——蛋白质组学(proteomics)应运而生。蛋白质组(proteome)一词由澳大利亚学者 Williams 和 Wilkins 于 1994 年首先提出，源于蛋白质(protein)与基因组(genome)两个字的结合，意指“一种基因组所表达的全套蛋白质”，即包括一种细胞乃至一种生物所表达的全部蛋白质。因此，蛋白质组的内涵是一个细胞、一类组织或一个生物的基因组所表达的全部蛋白质。蛋白质组学，是以细胞内全部蛋白质的存在及其活动方式作为研究对象，注重研究参与特定生理或病理状态的所有蛋白质种类及其与周围环境(分子)的关系。其研究不仅能为生命活动规律提供物质基础，也能为众多种疾病的机制阐明及防治提供理论根据和解决途径。蛋白质组学已逐步成为联系基因组序列与细胞行为研究的学科。

但是，蛋白质组处于新陈代谢的动态变化过程中，蛋白质的合成受诸多因素调控，即使是同一种细胞，在不同时空、不同条件下(如正常生理状态和病理状态)其蛋白质组也会发生改变。正因为蛋白质组具有这种时空性和可调控性，因此目前尚不可能获得细胞内存在的所有蛋白质。为此，科学家又提出“以细胞在某一特定时间所表达或与某个功能相关的蛋白质集合为研究对象的功能蛋白质组学(functional proteomics)”这一全新学科概念。功能蛋白质组学研究是从生命大分子(基因、蛋白质)水平到细胞水平研究的重要桥梁环节，已成为后基因组学的重要组成部分。

生物功能主要由蛋白质体现，而蛋白质有其自身特有的活动规律，仅从基因的角度或水平进行研究已远远不够。只清楚基因 DNA 序列，尚不能全面解决基因的表达时空、表达量、表达调控、蛋白质翻译后修饰等基因行为，这些在基因组学中不能解决的问题，功能蛋白质组学的研究可为其找到答案，因为功能蛋白质组学能够在细胞和生命有机体的整体水平上阐明生命现象的本质和活动规律。此外，功能蛋白质组学的研究还可为食品改造、疫苗开发和生物制药等提供重要依据。

二、研究对策、范围和内容

1. 研究对策　主要策略有两种：一种称为“竭泽法”，即采用高通量的蛋白质组研究技术分析生物体内尽可能多乃至接近所有的蛋白质。这种观点从大规模、系统性的角度来看待蛋白质组学，也

更符合蛋白质组学的本质。但是，由于蛋白质表达随时间和空间不断变化，要分析生物体内所有的蛋白质是一个难以实现的目标。另一种策略称为“功能法”，即研究不同时期细胞蛋白质组成的变化，如蛋白质在不同环境下的差异表达，以发现有差异的蛋白质种类为主要目标。这种观点更倾向于把蛋白质组学作为研究生命现象的手段和方法。

2. 研究范围　根据以上两种研究策略，蛋白质组学的研究范围也相应地分两种：一种是“完全”蛋白质组学或表达蛋白质组学(expression proteomics)，主要分析构成蛋白质组蛋白质的种类和数量，并以此来探讨细胞、组织、个体或特定状态的特征；另一种是“差异”蛋白质组学或功能蛋白质组学，主要筛选和鉴定不同种类或状态下各样品间蛋白质组的区别与变化，通过分析蛋白质组中构成蛋白质间相互作用及细胞内的功能单位，解析蛋白质组与细胞功能之间的相关性。蛋白质组学研究试图比较细胞在不同生理或病理条件下蛋白质表达的异同，对相关蛋白质进行分类和鉴定，并分析蛋白质间的相互作用和功能。

3. 研究内容　早期蛋白质组学主要是研究蛋白质组的组成成分，即蛋白质组的表达模式(expression profile)。随着学科的发展，蛋白质组学的研究范围不断完善和扩充，蛋白质组功能模式的研究不断深入，蛋白质翻译后修饰和蛋白质 - 蛋白质相互作用的研究也已成为蛋白质组学研究的重要组成部分。目前，以上三部分研究构成了蛋白质组学的主要研究领域。

第二节　蛋白质分析方法

Section 2　Protein Analysis Methods

一、蛋白质的指纹特征

(一) 蛋白质的指纹即肽质量指纹谱具有特征性

由于每种蛋白质的氨基酸序列(一级结构)都不同，蛋白质被识别特异酶切位点的蛋白质酶水解后，产生的肽片段序列也各不相同，其肽混合物质量数亦具有特征性，称为肽质量指纹谱(peptide mass fingerprint, PMF)，即蛋白质的指纹特征(图 9-1)。

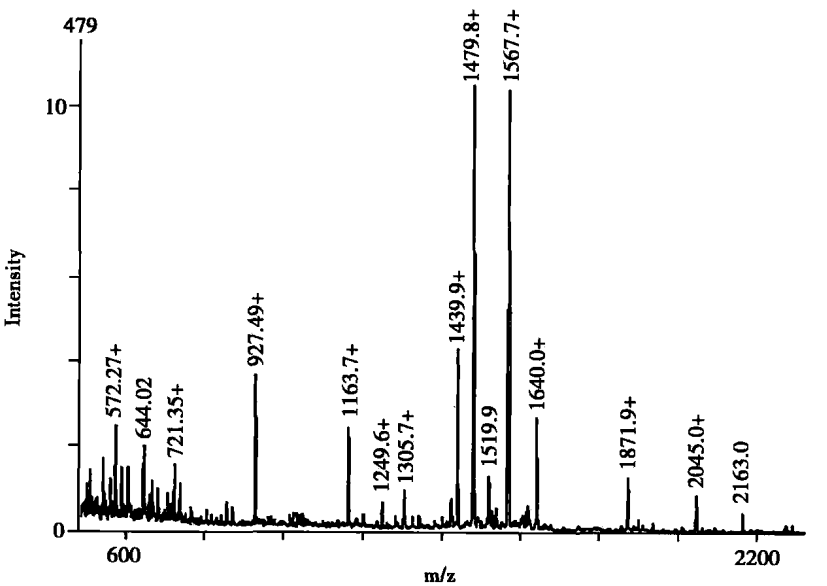


图 9-1　BSA 的 MALDI-TOF 质谱图谱

横坐标代表质量数，纵坐标代表峰强度，是该质量离子多寡的表示

(图片来源于：Rune Matthiesen. Mass Spectrometry Data Analysis. Humana Press, 2003.)

（二）肽质量指纹谱是鉴定蛋白质的常用方法

PMF是目前蛋白质组研究较为常用的鉴定方法，质谱是最有效的测定分析肽混合物的仪器。一般来说质谱指纹分析需要经过蛋白质原位酶解（包括蛋白质凝胶的脱色、还原和烷基化、酶解、萃取及合并萃取液冻干后进行质谱分析）、MALDI-TOF肽质量指纹图测定（包括蛋白样品脱盐及制备、质谱仪进行肽质量指纹图测定）及蛋白质鉴定数据库搜寻等三个步骤。

（三）肽质量指纹谱可用作其他参数测定

除了用来测定蛋白质的一级结构、表征氨基酸序列，以鉴定蛋白质外，指纹分析还可测定一些位点和翻译后修饰位和氨基酸变异位、缺失位。利用不同物种的保留特性可以进行功能分析，还可用于基因诊断。

1. 测定不同物种间的保留特性，从而推断分子的功能。由生物多样性和进化上远离引起的氨基酸残基取代，显示了蛋白质中的特征功能区。

2. 在一个蛋白消解物中，用来检测在化学或酶处理前后的“非匹配”（即和预测片段不符）的肽，从而表征蛋白质的修饰。例如，用特殊的内外糖苷酶分解糖蛋白，然后用质量指纹谱检测相对于假定的糖基化肽的质量位移，可以测出聚糖结构的位置和形式。

二、蛋白质的定位、修饰

蛋白质功能模式的研究是蛋白质组研究的最终目标，其主要研究目标是要揭示蛋白质组成员间的相互作用的关系，并深入了解蛋白质的结构与功能的相互联系，以及基因结构与蛋白质结构功能的关系。蛋白质定位、蛋白质翻译后修饰及后面章节提到的蛋白质-蛋白质相互作用都是目前蛋白质组学研究的重要内容。

（一）蛋白质的定位

成熟蛋白质必须在特定的细胞部位才能发挥其生物学功能，蛋白质在细胞内不同组分中的定位对其生理功能有着直接的影响。处于合适的亚细胞定位的蛋白质才能行使其功能。

1. 十二类亚细胞结构　生物体细胞是一个高度有序的结构，胞内根据空间分布和功能不同，可以分成不同亚细胞器或细胞区域。目前，对于蛋白质在细胞内的定位主要数据基于12类亚细胞：细胞膜、细胞质基质、内质网、高尔基体、溶酶体、过氧化物酶体、线粒体、叶绿体、细胞骨架、液泡、细胞核、细胞外基质。

2. 转运与定位机制　蛋白质在核糖体中合成后经蛋白质分选信号引导后被转运到特定的细胞器中，部分蛋白质则被分泌到细胞外或留在细胞质中。只有转运到正确的部位才能参与细胞的各种生命活动，每一个蛋白质都要在膜（细胞膜、细胞器膜）或水相腔室（细胞质、细胞器基质或内膜腔）中正确定位，如受体蛋白、离子通道蛋白和转运蛋白需要嵌在质膜内；DNA、RNA聚合酶需要送到核内；蛋白酶和过氧化氢酶应分别转运至溶酶体和过氧化物酶体；其他如细胞外间质和激素则需要分泌到细胞外等。

真核细胞中的蛋白质在细胞内的转运机制主要有两种类型：翻译后转运机制及共翻译转运机制。

（1）翻译后转运（post-translational translocation，PTT）：即蛋白质转运定位发生在蛋白质翻译后，游离型核糖体上合成的蛋白质必须等蛋白质完全合成并释放到胞质溶胶后才能被转运。其机制主要与蛋白质序列中与进入不同细胞器相关的特异前导肽（leading peptide）有关。翻译后蛋白转运入不同的细胞器有各自相应的前导肽信号，如蛋白质进入线粒体的前导肽序列、蛋白质进入细胞核的细胞核定位序列（nuclear location sequence，NLS）和输入蛋白（importin）序列及蛋白质进入过氧化物酶体的过氧化物酶体定位信号等。此外，此类蛋白质合成后还有相当一部分直接存在于细胞质溶液中，包括细胞骨架蛋白、各种反应体系的酶或蛋白质等。

（2）共翻译转运（co-translational translocation）：即蛋白质翻译与转运定位同时进行，其机制是膜结合型核糖体上合成的蛋白质，在它们进行翻译的同时就开始了转运，主要通过定位信号即信号肽，

一边翻译一边进入内质网，并在内质网内滞留进行加工即翻译后修饰；部分蛋白质翻译修饰后经分拣穿过内质网，经高尔基体，接着跨过细胞膜，进行定位与分泌。在膜结合核糖体上合成的蛋白质通过信号肽(signal peptide)，经过连续的膜系统转运分选才能到达最终的目的地，这一过程又称为蛋白质分选或蛋白质运输(protein trafficking)。

3. 预测亚细胞定位　除传统的亚细胞分离技术外，融合绿色荧光蛋白、质谱和同位素亲和标签等实验技术提供了一些比较精确的亚细胞定位数据。但单纯通过生物学实验方法来进行蛋白质亚细胞定位是十分费时费力和费钱的。开展广泛的生物信息学研究，构建可靠的数据库，对蛋白质的亚细胞定位进行分析与预测，能加速蛋白质亚细胞定位的研究。通过数据库分析和亚细胞定位相关的蛋白质序列特征是亚细胞定位预测的基础，从而寻找到亚细胞定位的生物学规律并确定蛋白质功能。

4. 亚细胞定位的数据库　研究蛋白质亚细胞定位的数据基本来自 SwissProt 数据库。除了一些综合数据库(SwissProt、MIPS 等)和模式生物数据库收录有部分蛋白质亚细胞信息外，目前还出现了一些专门的亚细胞定位数据库，如针对某个物种的、某个单个细胞器的、或收录预测数据、或同时收录经过实验验证和计算预测的数据。这些数据库的构建主要基于计算机预测、大规模实验和文献挖掘技术，如附表 9-1 所示。

5. 蛋白质信息的提取　蛋白质信息的提取是亚细胞定位预测的最基本步骤，体现了亚细胞定位的生物学内涵。蛋白质在合成过程中被分选到特定的亚细胞器中发挥生物学功能，很大程度上是由蛋白质的特征所决定，包括分选信号、序列、结构域特征和残基的理化性质等等，目前所采用的特征参数或特性的提取基本上都是基于某一特征或几个特征的综合。根据各类方法中抽取特征信息的不同，大致分为以下七种。

(1) 蛋白质分选信号：合成的蛋白质必须要定向地转运到特定细胞器中，一个重要的原因就是蛋白质中包含了各种不同的分选信号，这些信号指导新合成的蛋白质分选到特定的亚细胞器中，一种信号序列决定了特定蛋白的转运方向，可以被细胞器上的分选受体特异性识别。N 端分选信号包括信号肽、线粒体引导肽、叶绿体运输肽、核定位信号、内囊体腔转移肽和过氧化物酶体定位信号等。ChloroP 用以预测叶绿体运输肽；SignalP 专门用来识别信号肽；此外尚有 PSORT 及 TargetP 都是利用蛋白质的 N 端序列来进行亚细胞定位的预测。

(2) 蛋白质序列的氨基酸组分：氨基酸组成是一种最基本的序列特征，也是亚细胞定位预测中使用最为普遍的一种蛋白质特征信息。蛋白质一般由 20 种氨基酸组成，氨基酸组成将 20 种氨基酸在蛋白质序列中出现的频率抽取出来作为一个 20 维的向量。使用氨基酸组成来预测蛋白质亚细胞定位的方法主要有 Nakashima 预测法、ProtLock 预测法、Reinhardt 预测法等。

(3) 蛋白质的功能域信息：蛋白质序列在长期的进化过程中，某些特定位点上的氨基酸残基具有高度的保守性。这些特定的位点联系着蛋白质特定的生物学功能，被称为蛋白质的功能模体(motif)。这些功能模体具有特异性，可有效地刻画蛋白质序列。Scott 等基于蛋白质的模体信息发展出了 PSLT 预测法。

(4) 序列比对信息：蛋白质的比对就是找出两条序列之间的一种最佳匹配。这种匹配对应一个数值型输出结果，用来度量序列之间的相似程度，这种相似程度的度量可以直接用于预测蛋白质的亚细胞定位。BLAST 是一个常用的序列比对算法，可用于蛋白质序列相似性计算。

(5) GO 注释信息：由于蛋白质必须在特定的亚细胞中通过与其他的蛋白质进行相互作用才能执行特定的生物学功能，所以蛋白质的功能与它所处的亚细胞位置密切相关。如果知道了蛋白质的功能信息，就可以知道它所处的亚细胞位置。GO 是一个公认的基因功能注释标准化项目(详见 http://www.geneontology.org/)，包括了分子功能、生物学过程和细胞组件 3 种基本的信息。与功能域信息一样，这种信息的有效性取决于 GO 数据库的完善程度。

(6) 氨基酸物理化学性质：蛋白质序列是由氨基酸残基构成的，序列中氨基酸残基的物理化学

性质从根本上决定了蛋白质序列的整体物理化学性质，因此氨基酸残基的物理化学性质是描述蛋白质序列的一种重要信息。

（7）混合性特征参数：由于不同的特征从不同的角度刻画蛋白质序列，目前尚无一种特征能够很好地刻画蛋白质的亚细胞定位特征，因此将多种特征参数结合起来已经成为亚细胞定位预测中最为普遍的一种方法。这种混合提取特征信息建模的方法使得信息的输入更加完备，显著地提高了预测能力，并且使得人们更好地理解蛋白质的亚细胞定位与其序列、结构、物化性质和功能之间的关系。

（二）蛋白质的修饰

蛋白质翻译后修饰(post-translational modification，PTM)，指蛋白质在翻译中或翻译后会在个别氨基酸链上共价结合各种非肽类基团，形成翻译后修饰，包括磷酸化(如核糖体蛋白的 Ser，Tyr 和 Trp 残基常被磷酸化)、糖基化(如各种糖蛋白)、甲基化(如组蛋白，肌蛋白)、乙基化(如组蛋白)、羟基化(如胶原蛋白)、二硫键的配对、焦谷氨酸化、蛋白质降解泛素化、S- 硝酸化以及 ADP 核糖基化等二十多种，是蛋白质行使正常生理功能所必需的。

1. 常见的蛋白质翻译后修饰类型及其功能

（1）磷酸化：磷酸化是通过蛋白质磷酸化激酶将 ATP 上的磷酸基团转移到蛋白质的特定位点上的过程。磷酸化的作用位点为蛋白质肽链上的 Ser、Thr、Tyr 残基，在磷酸化调节过程中，细胞的形态和功能都发生改变。可逆的磷酸化过程几乎涉及所有的生理及病理过程，如细胞信号转导、肿瘤发生、新陈代谢、神经活动、肌肉收缩以及细胞的增殖、发育和分化等。

（2）糖基化：蛋白质的糖基化是低聚糖以糖苷的形式与蛋白质上特定的氨基酸残基共价结合的过程。根据氨基酸和糖连接方式的不同，糖基化可以分为四类：O 位糖基化(寡糖连接在 Ser、Thr 或羟基 -lys 的羟基上)、N 位糖基化(由寡糖连接在 Asp 的氨基形成)、C 位甘露糖化以及聚糖磷脂酰肌醇锚(GPI-anchor)糖基化。蛋白质的糖基化在多种生物过程中起着重要作用，如免疫保护、病毒复制、细胞生长、细胞与细胞间的黏附、炎症的产生等。

（3）甲基化：蛋白质的甲基化修饰主要是指是在甲基转移酶的催化下，在赖氨酸或精氨酸侧链氨基上进行的甲基化。甲基化增加了立体阻力，取代了氨基的氢，影响氢键的形成。因此，甲基化可以调控分子间和分子与目标蛋白间的相互作用，其中组蛋白上的甲基化在真核细胞染色体的遗传外修饰中占有中心地位，与细胞分化、发育、基因表达、基因组稳定性及癌变等有关联。

（4）泛素化：泛素由 76 个氨基酸组成，高度保守，普遍存在于真核细胞内。共价结合泛素的蛋白质能被蛋白酶识别并降解，这是细胞内短寿命蛋白和一些异常蛋白降解的普遍途径。泛素 - 蛋白酶系统是一个对真核细胞非常重要的调节系统，泛素化对于细胞分化、细胞器的生物合成、细胞凋亡、DNA 修复、新蛋白生成、调控细胞增殖、蛋白质运输、免疫应答和应激反应等生理过程都起到重要的作用。

（5）脂基化：为长脂肪链通过 O 或者 S 原子与蛋白质缀合得到蛋白缀合物的过程，通常是蛋白质分子中半胱氨酸残基的 S 键被棕榈酰基乙酰化，或者被法呢基烷基化。这两种脂肪链通常共同修饰同一个蛋白质分子，通过脂肪链与生物磷脂膜良好的相溶性，将蛋白质固定在细胞膜上。脂基化对于生物体内的信号转导过程起着非常关键的作用，脂基化蛋白相当于细胞信号转导的开关。

2. 蛋白质翻译后修饰分析方法　目前，蛋白质翻译后修饰的解析主要采用电喷雾(ESI)和基质辅助激光解吸电离(MALDI)两种质谱技术，可通过质谱对特征离子监测确定磷酸化肽，通过串联质谱确定磷酸化位点来鉴定磷酸化修饰；可通过质谱、蛋白酶解和糖苷酶解相结合的方法寻找糖肽，鉴定糖基化位点，再依靠串联质谱(MSn)分析糖链组成、结构甚至分支情况等。

（1）质谱分析：主要通过质量偏移(mass shift)来识别翻译后修饰蛋白。一种特定的翻译后修饰通常会作用于一定的氨基酸，经修饰后的氨基酸会增加相应的分子量，如磷酸化肽段因加入磷酸化基团而产生 +80 的质量偏移。翻译后修饰引起的肽段质量偏移表现为修饰蛋白在二维电泳和质谱

峰中相位的偏移，质谱通过测定多肽离子片段的质量鉴定肽段，可检测出翻译后修饰导致的质量偏移，进而识别发生翻译后修饰的蛋白。几种重要的翻译后修饰引起的质量偏移如附表 9-2 所示。

（2）质谱数据分析：由质谱数据鉴定翻译后修饰肽段的主要方法分为两大类：数据库搜索法（Database Searching）和从头测序法（DeNovo Sequencing）。

1）数据库搜索：首先将数据库中的蛋白质理论上酶解，产生在一定误差范围内和母离子质量匹配的候选肽段，然后比较实验质谱和候选肽段的理论质谱，为实验质谱指派肽段，并用合适的打分函数给出分值。从头测序法试图仅利用实验质谱的信息重建其对应的肽段序列，从而可能发现数据库中不存在的肽段。常用的数据库搜索法有：SEQUEST、Mascot、X!Tandem、MS-Align 等。

SEQUEST（http://fields.scripps.edu/sequest/）：最早被广泛使用的数据库搜索法。它计算理论质谱和实验质谱的相关性，并评价每个质谱对应的最好肽段与次好肽段之间的差异，相关性越高，差异越大，鉴定的可靠性就越高。

Mascot（http://www.matrixscience.com/search_form_select.html）：由 MOWSE 发展而来。MOWSE 是最早用肽指纹图谱鉴定蛋白质的算法，搜索速度快，但对新蛋白酶酶切得到的质谱或蛋白质翻译后修饰的鉴定非常困难。Mascot 改进了 MOWSE 的打分函数，给鉴定到的肽段指派概率，提高了鉴定的准确性，也提高了鉴定修饰肽段的能力。

X!Tandem（http://www.thegpm.org/TANDEM）：是用于蛋白质鉴定的开源软件，可用于搜索翻译后修饰、点突变、半酶切肽段等。X!Tandem 也需比较质谱和数据库中所有可能的候选肽段，用实验质谱和理论质谱的点积作为初始分值，然后考虑匹配上的 b/y 离子个数，改进打分函数，最后用类似 blast 算法给出质谱对理论肽段的期望值，表征鉴定结果的可靠性。

MS-Align：基于动态规划的质谱比对法，其改进了打分函数，考虑多个相关参数（如片段离子的信号强度、类型，肽段长度等），提高了候选肽段的可靠性，并用动态规划算法，显著加快了运行速度，从而可以搜索各种各样的翻译后修饰，方便以后扩充新的翻译后修饰类型，并可以根据一批质谱数据的结果统计表发现新的翻译后修饰。

2）从头测序法：常用的从头测序法主要是 PEAKS 法，由 Bin 等于 2003 年提出，可独立于蛋白质数据库鉴定肽段和翻译后修饰。PEAKS 首先根据母离子的质量，用不同的氨基酸组合得到 10 000 个候选肽段。对每个肽段序列，计算序列产生的理论离子与实验质谱谱峰的匹配程度。高信号强度的谱峰被匹配上的越多，肽段序列就越可能是真实的肽段。

3. 生物信息学预测和鉴定方法　目前，蛋白质翻译后修饰的研究仍面临着很多困难，如检测低拷贝的修饰肽方法灵敏度不过高，对修饰后蛋白质的实时定量分析较困难，修饰状态的稳定不易维持，及鉴定所有可能修饰蛋白质序列的测定工作量过于巨大等。所以，尽管二维凝胶电泳和质谱等蛋白质组技术不断完善，但是从整体上了解蛋白质的翻译后修饰仍面临着巨大的挑战。

鉴于蛋白质翻译后修饰鉴定的精确性、高通量化、定量分析还不很成熟，识别未指定修饰类型时的翻译后修饰，盲搜尤其困难，随着生物信息方法的介入，能够从序列和质谱两个角度帮助大规模鉴定翻译后修饰，有助于翻译后修饰蛋白质组学研究的迅速发展。

（1）常用蛋白质翻译后修饰相关数据库（表 9-1）：目前，涉及与功能密切相关的翻译后修饰的数据仍很少，多集中在磷酸化和糖基化方面，其中 SwissProt 是高质量的非冗余蛋白质数据库，包含有多种翻译后修饰的注释信息；PhosphoELM 和 Phosphosite 详细地收录了实验验证的磷酸化数据。

表 9-1　常用蛋白质翻译后修饰相关数据库

数据库	数据类型	网址链接
SwissProt	实验验证的各种 PTM	http://www.expasy.ch/sprot/userman.html
PhosphoELM	实验验证的磷酸化蛋白	http://phospho.elm.eu. org/about.html
Phosphosite	文献中确证的蛋白磷酸化位点	http://www.phosphosite.org/homeAction.do

续表

数据库	数据类型	网址链接
PHOSPHONET	人的蛋白磷酸化位点	http://phosphonet.ca/Default.aspx?AspxAutoDetectCookieSupport=1
Phosida	质谱鉴定的磷酸化位点	http://www.phosida.com/
PhosphoPep	4个物种的质谱磷酸化(果蝇/人/线虫/酵母)	http://www.phosphopep.org/
O-glycbase	实验验证的糖基化位点	http://www.cbs.dtu.dk/databases/OGLYCBASE
Resid	各种PTM包括氨基酸末端修饰	http://www.ebi.ac.uk/RESID/
Dbptm	预测的翻译后修饰	http://dbptm.mbc.nctu.edu.tw/

(2) 基于序列法预测蛋白质翻译后修饰：蛋白质的翻译后修饰需要相应酶的催化作用，发生在特定氨基酸或者多肽的特殊位置上。如甲基化修饰是在甲基转移酶的催化作用下专一性地发生在赖氨酸或精氨酸残基上；磷酸化是在磷酸酶的催化作用下发生在丝氨酸、苏氨酸或酪氨酸残基上；泛素化调节蛋白质的降解则需要泛素激活酶、泛素结合酶和泛素连接酶的参与。因此，同类翻译后修饰位点周围的片段往往都具有很强的序列保守性，通过对常见翻译后修饰数据的收集，及对发生同类修饰的蛋白序列特征的研究(如保守模体)，使得基于序列预测翻译后修饰成为可能(表9-2)。

表9-2 常用的基于序列预测翻译后修饰的方法及预测工具

预测工具	内容	网址链接
NetPhos	磷酸化位点	http://www.cbs.dtu.dk/services/NetPhos
Predikin	丝氨酸/苏氨酸蛋白激酶底物	http://florey.biosci.uq.edu.qu/kinsub/predikin.htm
Scansite	磷酸化位点	http://scansite.mit.edu/
NetphosK	真核蛋白磷酸化位点	http://www.cbs.dtu.dk/services/NetPhosK/
Disphos	磷酸化位点	http://core.ist.temple.edu/pred/pred.html
Predphospho2	激酶特异性磷酸化位点	http://www.nih.go.kr/phosphovariant/html/seq_input_predphospho2.htm
Netoglyc	真核蛋白糖基化位点	http://www.cbs.dtu.dk/services/NetOGlyc/
ELM	真核蛋白功能位点	http://www.elm-tech.com/
Prosite	蛋白功能域及功能位点	http://www.expasy.ch/prosite
Netacet	乙酰化位点	http://www.cbs.dtu.dk/services/NetAcet/
Sulfinator	酪氨酸硫酸化位点	http://www.expasy.ch/tools/sulfinator/
Gpi-som	磷脂酰肌醇锚定信号	http://gpi.unibe.ch/
Kinasephos	激酶磷酸化位点	http://kinasephos.mbc.nctu.edu.tw/

第三节 蛋白质组学数据的获取与分析

Section 3 Proteomics Data Acquisition and Analysis

蛋白质组数据库(proteome database)包含所有鉴定的蛋白质信息，如蛋白质的序列、核苷酸顺序、2-D PAGE、3-D结构、翻译后的修饰、基因组及代谢数据库等。蛋白质组数据的获取和分析可采用二维凝胶电泳分析技术、蛋白质芯片分析技术、酵母双杂交系统、Rosentta Stone方法等。

一、二维凝胶电泳分析技术

二维凝胶电泳(two-dimensional electrophoresis，2-DE)是蛋白质组研究的常用技术之一。

(一) 定义及特点

2-DE广义的定义是将样品进行电泳后在它的直角方向再进行一次电泳，又称双向电泳。第一向

是等电聚焦(isoelectric focusing, IEF),蛋白质沿 pH 梯度分离至各自的等电点。第二向是 SDS 聚丙烯酰胺凝胶电泳(SDS-PAGE),蛋白质进行分子量的分离。样品经过电荷和质量两次分离后,可获得样品分子等电点(isoelectric point, pI)和分子量等信息,分离的结果不是带,而是点。这是目前所有电泳技术中分辨率最高、信息最多的技术。

对 2-DE 而言,目前主要有三种方法分离蛋白:①ISO-DALT(isoelectric focus)以 O'Farrell's 技术为基础。第一向应用载体两性电解质(carrier ampholyte, CA),在管胶内建立 pH 梯度。尽管该系统有很高分辨率,但其第一向电泳存在很多问题,如因阴极漂移而丢失碱性蛋白;载体两性电解质 pH 梯度不稳定、受电场和时间的影响大、重复性不好等。②NEPHGE(non-equilibrium pH gradient electrophoresis)用于分离碱性蛋白(pH>7.0)。如果聚焦达到平衡状态,碱性蛋白会离开凝胶基质而丢失。因此,在等电区域的迁移须在平衡状态之前完成,但很难控制。③IPG-DALT 发展于 80 年代早期。由于固相 pH 梯度(Immobilized pH gradient, IPG)的出现解决了 pH 梯度不稳的问题。IPG 通过固化电解质(immobiline)共价偶联于丙烯酰胺产生固定的 pH 梯度,克服了 IEF 的缺点,从而达到高度的重复性,是目前最常用的方法。

(二)固相 pH 梯度 -SDS 双向凝胶电泳(IPG-DALT 电泳)

作为目前分辨率最高的电泳方法,固相 pH 梯度二维凝胶电泳的操作原理及技术流程主要有以下六个步骤:

1. 样品制备　样品制备的目的是从成分复杂的细胞、组织等中取得高纯度、尽可能完整的蛋白质组分。蛋白质提取质量的好坏,直接关系到获取蛋白质组信息的完整性,因此样品制备是双向电泳实验首要的关键环节。

样品制备的方法应尽量简单、避免蛋白质丢失或降解;在实验的整个过程中要保持蛋白质的充分可溶性,选择合适的电泳缓冲体系。通常可采用细胞或组织中的全蛋白质组分进行蛋白质组分析,也可以进行样品预分级,即采用各种方法将细胞或组织中的全体蛋白质分成几部分,分别进行蛋白质组研究。

2. 蛋白质定量　进行凝胶间的差异比较以及不同长度、pH 梯度胶条和不同检测方法的选择,都要求对蛋白质进行定量。常用的蛋白质定量方法有 BCA 法、Bradford 法、UV280 法等,但由于这些定量方法都是基于吸光度的测定,而样品溶液中往往含有高浓度的尿素等溶剂可能影响吸光度的准确测定,故推荐使用专门用于双向电泳蛋白质定量的试剂盒进行定量。

3. 一向电泳　一向电泳等电聚焦(isoelectric focusing, IEF),是根据不同蛋白质的 pI 值不同,在电场力的作用下将其分离。pH 值梯度的存在对等电聚焦技术相当重要。在 pH 梯度胶内,不同 pI 的蛋白质分子在电场作用下,将移动到胶条上不同 pH 值梯度位置。一向电泳不仅能将蛋白质在其等电点上浓缩,还能根据不同蛋白质所带电荷的微小差异将不同蛋白质分离。

4. 一向胶条的平衡　进行第二向电泳前,需要对 IPG 胶条进行平衡(Equilibration),平衡过程是将 IPG 胶条浸没在第二向电泳所必需的 SDS 缓冲体系中,以便于被分离蛋白质与 SDS 完全结合并顺利转移入二向电泳的凝胶中。平衡后应立即进行第二向电泳。

5. 二向电泳　第二向电泳,即十二烷基磺酸钠 - 聚丙烯酰胺凝胶电泳(Sodium Dodelyl Sulfate-PolyAcrylamide Gel Electrophoresis, SDS-PAGE)是根据蛋白质分子量的大小不同,在电场中的泳动速率也不同的原理而分离蛋白质的方法。

6. 凝胶检测　分离后的斑点检测(spot detection)对于 2-DE 至关重要,尤其对于"差异蛋白质组"研究。适用于 SDS 凝胶中蛋白质检测的方法都可用于双向电泳凝胶检测。

银染和考马斯亮蓝(R250、G250)染色是蛋白质组研究中最为广泛使用的两种染色方法,考马斯亮蓝染色方法因简单易行、价格便宜且与后续质谱鉴定的兼容性较好,但是灵敏度较低。银染色的灵敏度较考马斯亮蓝染色高约 100 倍,可检测少到 2~5ng 的蛋白,适合于含量低的蛋白样品分析,但其敏化过程中戊二醛对蛋白质的可能修饰及银离子对后续鉴定过程中质谱的峰值有一定影响,

降低了鉴定率。Cydye、Deep purple、SYPRO Ruby 和 Pro-Q Diamond 等荧光染料灵敏度高，线性范围宽，且与质谱分析的匹配性好，是非常理想的检测方法，但是价格较昂贵并需要借助紫外光或激光扫描。

二、蛋白质组质谱分析技术

质谱(mass spectrometry，MS)是按照物质的质量与电荷的比值(质荷比)顺序排列成的图谱。质谱分析法是按照离子的质荷比(mass-to-charge ratio，m/z)大小对离子进行分离和测定从而对样品进行定性和定量分析的一种方法。自 20 世纪初该技术产生时起，质谱已成为连接蛋白质与基因的重要技术，是蛋白质组研究中发展最快，也最具活力和潜力的技术。

(一) 质谱仪

质谱仪(mass spectrometer)是利用电磁学原理使离子按照质荷比进行分离，从而测定物质的质量与含量的科学实验仪器，一般由进样器、离子化源、质量分析器、离子检测器、控制电脑及数据分析系统组成，其中样品入机的离子化源和测量被介入离子分子量的质量分析器是两个关键的部件。进样器把分析样品送进离子源；离子化源把样品中的原子、分子电离成离子；质量分析器使离子按照质荷比的大小在空间或时间上分离开来；检测器用以测量、记录离子流强度而得出质谱图。

基质辅助激光解吸/电离(matrix assisted laser desorption/ ionization，MALDI)和电喷雾(electrpspray ionization，ESI)是蛋白质组学质谱分析中最常用的两种电离技术。MALDI 利用激光脉冲将与基质结晶混合的蛋白质样品升华并电离出来。ESI 将分析物从溶液中电离出来，可以方便地与液相色谱(liquid-chromatography，LC)联用。MALDI-MS 通常用来分析成分相对简单的肽混合物，而集成液相色谱的 ESI-MS 系统(LC-MS)是分析较复杂样品的首选。

质量分析器是质谱仪的核心，在蛋白质组学分析中，它的技术指标是灵敏度、分辨率、质量精确度和从肽碎片中得出信息丰富的质量谱的能力。有四种质量分析器常用于蛋白质组学研究：飞行时间(time-of-flight，TOF)、离子阱(ion trap，IT)、四极杆(quadrupole，Q)、傅立叶变换离子回旋共振(Fourier transform ion cyclotron resonance，FT-ICR)分析器。这些质量分析器既可单独使用，也可发挥它们的各自长处串联起来构成串联质谱仪(Tandem-MS)。

(二) 质谱的应用

1. 分子量测定　分子量是蛋白质、多肽最基本的物理参数之一，是蛋白质、多肽识别与鉴定中首先需要测定的参数。生物质谱可测定的生物大分子分子量高达 40 万，准确度高达 0.1%~0.001%，远高于目前常规使用的 SDS 电泳与高效液相色谱技术。

2. 肽谱测定　肽谱是基因工程重组蛋白结构确认的重要指标，也是蛋白质组研究中大规模蛋白质识别和发现新蛋白质的重要手段。生物质谱通过与特异性蛋白酶解相结合，可测定肽质量指纹谱并给出全部肽段的准确分子量，结合蛋白质数据库检索就可实现蛋白质的快速鉴别和高通量筛选。

3. 肽序列测定　串联质谱技术可直接用于肽段的测序，从一级质谱产生的肽段中选择母离子进入二级质谱，经惰性气体碰撞后，肽段沿肽链断裂，由所得各肽段质量数差值推定肽段序列，并用于数据库查询，称为肽序列标签技术(peptide sequence tag，PST)，目前广泛应用于蛋白质组研究中的大规模筛选。

4. 巯基和二硫键定位　利用生物质谱的准确分子量测定特性，同时结合碘乙酰胺、4- 乙烯吡啶等化学试剂对蛋白质进行烷基化和还原烷基化以及蛋白质酶切、肽谱技术等，可实现对二硫键和自由巯基的快速定位。

5. 蛋白质翻译后修饰　目前已有将生物质谱技术应用于蛋白质翻译后修饰的识别与鉴定研究的报道，如用 MALDI-TOF-MS 对双向电泳分离蛋白质磷酸化位点进行定位、MALDI-TOF-MS 结合不同酶解方式确定糖基化位点等。

目前，以质谱为基础鉴定和注释蛋白质主要通过两种路线：一种是通过 PMF 和数据库搜寻匹配的路线，进行这种分析的质谱仪首选 MALDI-TOF 质谱仪；另一种是通过测出样品中部分肽段二级质谱信息或氨基酸序列标签和数据库搜寻匹配的路线，主要应用于 ESI 串联质谱仪。

（三）基质辅助激光解吸电离飞行时间质谱（MALDI-TOF-MS）分析技术

目前用于蛋白质鉴定的质谱仪主要有两种：基质辅助激光解吸电离飞行时间质谱（MALDI-TOF-MS），用飞行时间作为质量分析器；电喷雾（四极杆）质谱（ESI-MS），采用四极杆质量分析器。基质辅助激光解吸电离飞行时间质谱（MALDI-TOF MS）为脉冲式的离子化技术，从固相标本中产生离子，并在飞行管中测其分子量，其特点是对盐和添加物的耐受力高，样品测量速度快，且操作简单。

1. MALDI-TOF 质谱测定肽质量指纹图　肽质量指纹谱（peptide mass fingerprint，PMF）分析为目前双向凝胶电泳分离的蛋白质进行微量鉴定时使用最广泛的方法。将质谱分析获得的肽段分子质量与蛋白质数据库中理论肽段的分子质量进行比较（理论肽是由实验所用的酶来“断裂”蛋白所产生的），通过软件分析就可获得蛋白质信息，根据匹配情况判断出所鉴定分析的蛋白质是已知的还是未知的。这一技术能够完成的肽质量可精确到 0.1 个分子量单位，是大规模蛋白质鉴定的重要手段。

互联网上有很多专门服务于生命科学研究领域的网站（表 9-3），其中 ExPASY（蛋白质分析专家系统）是常用于蛋白质肽质量指纹图鉴定的网站。

表 9-3　用于蛋白质 PMF 鉴定的数据库搜索软件及地址

软件名称	地址
MultiIden	http://expasy.hcuge.ch/sprot/multiident.html
Peptide Search	http://www.narrador.emblheidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html
MS-Fit	http://falcon.1udwig.ucl.ac.UK/vcsfhtml/msfit.html
ProFound	http://Prowl.rockefeller.edu/PROWL/Pro-id-main.html
MOWSE	http://www.d1.ac.uk/SEQNET/mowse/html/
Mass Search	http://cbrg.inf.ethz.ch/subsection3-1-3.html

2. MALDI-TOF 质谱技术用于蛋白质 C- 端序列分析　肽质量指纹术对其自身而言并不能揭示所衍生的肽片段或蛋白质，为进一步鉴定蛋白质，出现了一系列的质谱方法用来描述肽片段（peptide fragment）。在质谱仪内，应用源后衰变（post-source decay，PSD）和碰撞诱导解离（collision-induced dissociation，CID）可产生包含有仅异于一个氨基酸残基质量的一系列肽峰的质谱。此外，用酶或化学方法从 N- 或 C- 末端按顺序除去不同数目的氨基酸，亦可形成大小不同的一系列的梯形肽片段，所得的一定数目的肽质量由 MALDI-TOF-MS 测量。

（四）电喷雾质谱分析

电喷雾电离质谱（ESI-MS）采用连续离子化的方法，从液相中产生离子，能快速、准确地解决从小分子到生物大分子或不稳定有机分子质量的测定问题，常作为与其他分析技术联用的首选仪器，联合四级质谱或在飞行时间检测器中测其分子量，可以和液相色谱（liquid-chromatography，LC）、毛细管电泳等现代化的分离手段联用。

1. 电喷雾电离质谱测定蛋白质和多肽分子质量　蛋白质和多肽分子经电喷雾电离时，会吸附一个或多个质子，形成一系列带电荷状态不同的分子离子，在质谱中形成荷质比不同的谱峰。一般可根据谱峰的同位素离子峰分布情况以及利用相邻两峰的荷质比和电荷数关系计算求得离子的分子质量。

2. 液相色谱 - 电喷雾质谱法鉴定双向凝胶电泳蛋白质　对双向凝胶电泳分离的蛋白质点经酶解后的多肽混合物进行液相色谱 - 电喷雾质谱联用（LC-ESI MS）鉴定分析，同样可以得到肽质量指纹图（PMF）。

（五）串联质谱（MS/MS）

单纯用 PMF 不能明确鉴定时，就要用其他信息来鉴定蛋白质。串联质谱的使用能够对基于

PMF 的结果进行再分析或对未赋值的质谱峰信号进行研究。对于初始用 PMF 法鉴定的蛋白，可选择其中部分肽段峰进行 MS/MS 分析，能得到肽段的序列。PMF 和肽序列联合检索有助于进一步确定鉴定结果，提高蛋白质鉴定的成功率。而对于那些未匹配到的肽段峰进行 MS/MS 分析，则有可能发现感兴趣的修饰位点。

串联质谱法可选择性分析混合物中的某个组分而不必将各组分分开。第一级质量分析器获得所有组分的分子离子峰(母离子)，经质量过滤器挑出需进一步分析的母离子，此母离子在碰撞室内经高流速惰性气体碰撞诱导离解(collisionally induced dissociation，CID)产生碎片离子(子离子)，子离子进入第二级质量分析器获得母离子的碎片峰。通过研究母离子和子离子的裂解关系，可以获得多肽和蛋白质的结构信息。

最近面市的最新型串联生物质谱仪液相色谱 - 电喷雾 - 四极杆飞行时间串联质谱仪(LC-ESI-QTOF)在传统电喷雾质谱仪的基础上采用飞行时间质量分析器代替四极杆质量分析器，其一级质谱是电喷雾电离源、四极杆质量分析器，第二级质谱改用飞行时间质量分析器，大大提高了仪器的分辨率和灵敏度，可对微量样品进行序列分析。

带有串联质谱功能的 MALDI-TOF 质谱仪，其商品名有 Q-STAR，在质谱中加入了源后降解(post-source decay，PSD)模式或碰撞诱导解离模式，MALDI 方式产生的离子在飞行管道飞行时会产生失去中性碎片的亚稳离子，通过分析亚稳离子与其母离子的裂解规律而得到母离子的结构信息，从而使生物大分子的测序成为可能。

三、蛋白质芯片分析技术

蛋白质芯片(protein chips)技术又称蛋白质微阵列(protein microarrays)，是一种高通量的、小型化的、平行性的生物检测技术。

(一) 基本原理与特点

蛋白质芯片是将已知蛋白点印在固定于不同种类的支持介质上，制成由高密度的蛋白质或多肽分子的微阵列组成蛋白微阵列，阵列中固定分子的位置及组成是已知的，用未经标记或标记(荧光物质、酶或化学发光物质等标记)的生物分子与芯片上的探针进行反应，然后通过特定扫描装置如激光扫描系统(laser scanner basessystem)或电荷偶联照像系统(charge coupleddevicecamera，CCD)对信号强度进行检测，进一步对杂交结果进行量化分析，检测蛋白质的存在情况。广义上，蛋白质芯片包括固定有保持天然活性的蛋白质、能与蛋白质特异性结合的 DNA 和 RNA、糖类、合成多肽及其他能够从复杂蛋白质混合物中特异性地捕获目的蛋白的小分子物质的微阵列。

蛋白质芯片具有以下特点：①特异性强，这是由抗原抗体之间、蛋白与配体之间的特异性结合决定的；②敏感性高，可以检测出样品中微量蛋白的存在，检测水平已达 ng 级；③通量高，在一次实验中对上千种目标蛋白同时进行检测，效率极高；④重复性好，不同次实验间相同两点之间差异很小；⑤应用性强，样品的前处理简单，只需对少量实际样本进行沉降分离和标记后，即可加于芯片上进行分析和检测；⑥适用范围广，适用于包括组织、细胞系、体液在内的多种生物样品。

(二) 分类

1. 根据功能　可分为功能研究型芯片(functional protein microarrays)和分析检测型芯片(analytical protein microarrays)。功能研究型芯片多为高密度芯片，载体上固定的是天然蛋白质或融合蛋白，该种芯片主要用于蛋白质活性以及蛋白质组学的相关研究。分析检测型芯片的密度相对较低，固定的是抗原抗体等，主要用于生物分子的大量、快速检测。

2. 依蛋白质种类　可将蛋白质芯片划分为抗体芯片和抗原芯片。抗体芯片由特异性的抗体或抗体类似物固定在载体上构成，检测标本中抗原是否存在及浓度如何，如蛋白质表达谱研究、测定血液中某种疾病特异性蛋白的浓度、测定细胞因子等。第二类是抗原芯片，检测自身免疫性疾病中的特异性抗体、过敏性疾病的过敏原和受微生物感染的宿主体内的抗体等。

3. 根据芯片表面的不同化学成分　分为化学表面芯片和生物表面芯片。化学表面芯片的构想来源于经典色谱法(反相层析、离子交换层析、金属螯合层析等),分为疏水、亲水、阳离子、阴离子和金属螯合芯片。其原理为铺有相关介质的芯片可以通过介质的疏水力、静电力、共价键等结合样品中的蛋白质,用洗脱液去除杂质蛋白质而保留感兴趣的蛋白质,用于检测未知蛋白质,并获取指纹图谱。生物型蛋白质芯片则是将生物活性分子结合到芯片表面,用于捕获靶蛋白,分为抗体 - 抗原、受体 - 配体、DNA- 蛋白质芯片等。

4. 按点样蛋白质有无活性功能　分为无活性和有活性芯片。无活性的芯片是将已经合成好的蛋白质点在芯片上,其制作方式主要分为原位合成、点合成、光蚀刻术 3 类;有活性的芯片是指点在芯片上的样品是活的生物体(如细菌),在芯片上原位表达蛋白质。相对于无活性的芯片,有活性的芯片可以提供模拟的机体内环境,对于蛋白质功能分析更为有利。

5. 按载体的不同　可分为普通玻璃载体芯片(plain-glass slide)、多孔凝胶覆盖芯片(porous gel pad chip)、微孔芯片(microwell chip)等。

(三)蛋白质芯片的制备及分析

1. 被测物质的准备　蛋白质芯片的检测对象包括蛋白质、酶的底物或其他小分子。以蛋白质为例进行介绍,首先将被测蛋白质进行标记,标记物既可以是荧光剂如 Cy3、Cy5、Bodipy-FL,又可以是酶,如辣根过氧化物酶、碱性磷酸酶、β-D- 葡萄糖醛酸酶等。也可根据实验需要,在被测蛋白质与芯片反应后进行特异性标记,如在免疫反应中,利用酶标二抗间接标记。

2. 反应过程　先将蛋白质芯片与被测样品溶液在适宜温度下孵育一定时间,然后用 PBST 洗去未反应的分子。再根据标记物的不同或直接检测(如荧光标记)或显色后检测(如酶标记)。

3. 芯片的检测　生物芯片技术的核心是生物芯片的制备及反应信号的检测。对于荧光标记的芯片,用激光共聚焦显微镜进行扫描,然后通过计算机分析出每个点的平均荧光密度;对于酶标记的芯片,显色后用 CCD 照相机进行拍摄,将信号通过计算机处理得到每个点的灰度。

4. 结果的分析　在每个芯片的制作过程中应设计有阴阳性对照反应,或已在多次实验中找到一个判断阴阳性结果的界值,作为判断结果的根据。将每个点的荧光密度或灰度除去背景干扰后与相对界值进行比较,根据信号的有无、多少进行定性或定量分析。在结果分析过程中需要对大量数据进行复杂处理,这些都需要通过特定的计算机软件来完成。

(四)应用领域

1. 基因表达的筛选　将 cDNA 文库表达的蛋白制备成蛋白质芯片,可用来筛选特异性的基因表达产物。

2. 特异性抗原抗体的检测　蛋白质芯片上的抗原抗体反应具有很好的特异性,结合使用特异性抗体,可用来度量整个细胞或组织中的蛋白质丰富程度和修饰程度。

3. 蛋白质组学研究　蛋白质芯片在蛋白质组学应用研究中主要用于定量检测组织和细胞中蛋白质的表达水平比较健康、病理和药物治疗等不同条件下蛋白质谱的表达差异,为大量蛋白质表达水平的差异分析提供了一种高通量的分析技术平台,弥补了传统的二维凝胶电泳和质谱分析技术的不足。

4. 蛋白质相互作用的研究　蛋白质点在芯片上后仍能保持其生物活性,在芯片上高密度固定大量感兴趣的蛋白质,就可以平行而高通量地研究蛋白质 - 蛋白质、蛋白质 - 小分子以及酶 - 底物之间的相互作用,以筛选新的蛋白质。此外,还能够用于其他方法不能检测到的,如蛋白质 - 药物、蛋白质 - 脂质之间的相互作用,以及蛋白质与小分子物质的作用,如蛋白质与 DNA、RNA 分子等,可用于高通量、大规模地筛选药物靶分子。

四、酵母双杂交系统

酵母双杂交系统(yeast two-hybrid system)是一种直接于酵母细胞内检测蛋白质 - 蛋白质相互作用而且灵敏度很高的分子生物学方法。该系统在 1989 年由 Fields 和 Song 等在研究真核基因转录调

控中首次建立并得到广泛的应用。此项技术的出现使以往因耗时、繁琐的物理测定和基因筛选限制而无法实现的许多设想，都可能逐一实现。酵母双杂交系统是鉴定及分析蛋白质-蛋白质间相互作用的最常用及最有效的工具之一，此技术现已逐渐推广到了如信号传导、细胞周期调控及基因表达调控等多个研究领域。

（一）酵母双杂交系统原理

真核生物中有一种特殊的上游激活序列(upstream activating sequence，UAS)，可以在转录水平上进行基因表达的调控，它的作用就是与激活蛋白结合从而大大增加启动子的转录速度。酵母双杂交系统是建立在人们对酵母转录因子 GAL4 蛋白的认识基础上。酵母中存在着转录因子 GAL4 蛋白，该蛋白能激活转录主要是因为它有两个结构上可以分开的、功能上相互独立的结构域，即位于氨基(N)端的 DNA 结合结构域(DNA-Binding domain，DNA-BD)以及位于羧基(C)端的转录激活结构域(transcriptional activation domain，TAD)。这两个结构域在它们分开时仍分别具有功能，即 DNA 结合结构域仍具有结合 DNA 能力，但因缺少转录激活结构域而不能激活转录；而转录激活结构域虽然有激活转录功能，但因无法正确定位于 DNA 上也不能激活转录。即 DNA-BD 和 AD 单独分别作用并不能激活转录反应，只有将这两部分通过适当的途径连在一起后才恢复激活转录的活性，激活 UAS 下游启动子，使启动子下游基因得到转录。导致结合和激活发生的关键是 DAN-BD 和 AD 两个结构域通过共价或非共价连接建立的空间联系。

根据这一特性，可以构建两种重组质粒，分别表达 GAL4 蛋白 N 端的 1~147 个氨基酸(DNA-BD)和羧基端的 768~881 个氨基酸(AD)。若在 DNA-BD 上再接上一个“诱饵”蛋白 X，在 AD 上接上一个“猎物”蛋白 Y，再将这两个质粒共同转化至酵母体内。如果 X、Y 蛋白在酵母核内发生交互作用，则相当于将 GAL4 的 DNA-BD 和 AD 又连在一起，从而激活 UAS 下游启动子调节的报告基因的表达，使转化子由于报告基因的表达而可以在特定的缺陷培养基上生长，同时因激活转录下游 GAL1-LacZ 和/或 MEL1 基因的表达，从而在 X-Gal 存在下显蓝色。这样就可以利用报告基因的转录指示诱饵蛋白 X 与猎物蛋白 Y 之间是否反应。最后，将阳性克隆子在二缺平板上划线 3 次，提取酵母质粒，转化 *E.coli*，提取质粒，进行 PCR 或双酶切去除含有相同长度 cDNA 的克隆子，并对插入片段测序，将获得的序列通 NCBI 数据库进行相似性搜索。

由于该系统将一个蛋白 X 与 GAL4 的 DNA-BD 杂交，再将第二个蛋白 Y 与 GAL4 的 AD 杂交，因此称该系统为双杂交系统。目前双杂交系统大多都应用于酵母中，故称为酵母双杂交系统。酵母是真核细胞，表达的蛋白质可以在细胞核中产生交互作用。而且酵母菌有很多营养缺陷型标记，有利于对质粒的选择，筛选比较方便，如 MEL1、LacZ 报告基因，通过颜色反应即可验证蛋白 X 与蛋白 Y 是否发生交互作用。所有的报告基因的启动子都经过了特殊改造，插入了所采用的 DNA 结合蛋白的特异结合序列，经改造的启动子和报告基因能稳定地整合到酵母细胞的基因组。酵母细胞本身并不能表达能激活报告基因的转录因子，报告基因的表达只依赖于体系中 X、Y 蛋白的同时存在和相互作用。由于酵母双杂交系统是通过利用两蛋白相互作用后，观察能否激活报告基因，而决定两蛋白间是否有相互作用的，因而在酵母体内表达的诱饵蛋白或猎物蛋白两者单独均不能激活报告基因的转录因子。

（二）酵母双杂交系统特点与优点

1. 双杂交系统具有独特优势　首先，它实现了真正的克隆基因目标化。利用该系统不但可找到相互作用的蛋白质，而且能直接有效地克隆到这些蛋白质的基因。其次，使用该系统分析蛋白质的反应时，是在细胞内进行并完成的，故而检测分析不受外界的限制，且反映了蛋白在体内的交互作用。另外，该系统还可以高灵敏度地检测蛋白与蛋白之间的交互作用。利用其来检测两个蛋白之间是否存在交互作用，只要将编码两个蛋白质的 DNA 序列分别克隆到双杂交系统的载体中，即可通过检测报告基因是否表达从而论证这两个蛋白质是否发生了交互作用。它之所以具有高度敏感性，主要是因为杂合蛋白是由高拷贝质粒的强启动子过量表达的蛋白，信号测定是在自然平衡浓度条件

下进行；另外，酵母表型、X-Gal 及 HIS3 等蛋白表达等检测方法均很敏感。双杂交系统是在真核模式生物酵母细胞内进行的，对蛋白质间微弱的及瞬间的作用也能够通过报告基因的表达产物检测得到，是一种具有高灵敏性的检测蛋白质间关系的技术。它不仅可以用于研究哺乳动物基因组编码的蛋白质之间的相互作用，还可以用来研究植物等基因组编码的蛋白质之间的相互作用。

2. 应用方便快捷　随着这项技术的不断改进，它的适用领域不断拓宽。这一系统的最主要应用是可以快速直接地分析已知蛋白质 - 蛋白质间的相互作用，另外可以筛选 cDNA 文库以分离新的与已知蛋白作用的配体及其基因序列。酵母双杂交技术已经成为发现新基因的一个主要途径，同时也是研究蛋白和蛋白间交互作用最有力的工具之一，并且还具有很大的发展潜力，在许多研究领域中有着广泛的应用。

（三）酵母双杂交系统的局限性及优化

1. 局限性　如在双杂交过程中，要经过转化，但是酵母细胞的转化效率相当低，要比细菌低约 4 个数量级，转化步骤往往成为该技术的一个瓶颈。

由于双杂交过程是在细胞核内完成的，然而许多蛋白质的相互作用不仅仅局限于核内，因而对于研究膜蛋白、分泌蛋白、膜受体及胞质蛋白有很大的局限性。分离的泛素系统（split-ubiquitin system）是一种不依赖转录激活机制的酵母双杂交系统，更适合于发生在胞膜或胞质中的蛋白质相互作用研究。SOS 蛋白介导的双杂交系统（SOS recruitment system）也可用于分析膜受体及细胞外分泌蛋白相互作用。

2. 假阳性　可以针对表达型载体进行改进，比如融合蛋白可以被置于一些可诱导的启动子调控之下，使蛋白质在酵母的表达量能依需要进行调控。其次，发展哺乳动物细胞双杂交系统可以较好地研究此类蛋白质间的相互作用。

此外，由酵母双杂交衍生出的单杂交系统（one-hybrid system）、三杂交系统（three-hybrid system）、反向双杂交系统（reverse two-hybrid system）等对传统的双杂交系统作出了重要补充和扩展。

酵母双杂交技术已经在蛋白质间的相互作用研究、筛选新的蛋白质、研究蛋白质的结构与功能等诸方面发挥着重要作用，随着酵母杂交体系技术的不断完善与提高，以及在应用方面经验的不断积累，基于酵母双杂交原理而建立的一系列方法将为生命科学的进展起到积极的推动作用。

五、Rosetta Stone 方法

随着测序技术的飞速发展，越来越多物种的基因组序列得以完成，这使得应用高通量的实验技术及计算方法构建基因组范围内的蛋白质相互作用网络成为可能。应用蛋白质相互作用网络，不仅可以预测蛋白质的功能，另一方面还可能发现新的细胞系统，甚至可以研究网络结构对蛋白质进化速率的影响。Rosetta Stone 方法即是近年来开发出的基于基因组上下文的方法来预测蛋白质相互作用，并用于研究非同源基因之间功能关联的方法。

（一）Rosetta Stone 方法的来源

罗塞塔石碑（Rosetta Stone）是 1799 年在埃及罗塞塔发掘出的刻有古埃及文字及希腊文的石碑。在古埃及建筑及石碑上面，铭刻着许多古埃及文字，由于古埃及象形文字久已失传，后人难以看懂。人们试图破解古埃及文字，但都未成功。罗塞塔石碑的出土则为破译这种象形文字提供了关键的线索。碑文上段与中段分别是草体埃及文字及古埃及象形文字，下段是希腊文字。石碑上用的是三种文字书写的是同一段话，是对同一内容文字的互译。而希腊文及草体埃及文字是当时已经了解的文字。所以这三种文字相互对照，就可以知道一个个古埃及象形文字所代表的含义。这样借助于罗塞塔石碑，由此而完全破译了古埃及象形文字。

罗塞塔石碑方法（Rosetta Stone method）与上述相似，此法又称为基因融合法（gene fusion method）。比如一个物种中的基因 C 包含的两个部分分别与同一物种或另一物种中的基因 A 及 B 同源，那么就可以认为基因 A 与基因 B 存在功能上的相关性，基因 C 在此能够将另外两种没有同源性的基因

A 及 B 联系在一起。借助于基因 C 的作用，就可以找到本无同源性的基因 A 和 B 之间的关联。基因 C 在此可称为罗塞塔石碑基因，其对应的表达蛋白可称为罗塞塔石碑蛋白。由于罗塞塔石碑蛋白 C 的存在，据此可预测蛋白质 A 与蛋白质 B 之间可能存在相互作用。

如酵母内的一个单一的蛋白质 DNA 拓扑异构酶Ⅱ（DNA topoisomeraseⅡ），它包括的两个结构域（domain）分别与大肠杆菌内的两个不同蛋白亚基 DNA 解旋酶 A（gyrase A）及解旋酶 B（gyrase B）高度同源。在此，酵母 DNA 拓朴异构酶Ⅱ可称为大肠杆菌 DNA 解旋酶 A（gyrase A）及解旋酶 B（gyrase B）的罗塞塔蛋白。还有在一些寄生原虫如疟原虫内存在一种双功能融合蛋白二氢叶酸还原酶 - 胸苷酸合成酶（dihydrofolate reductase-thymidylate synthase，DHFR-TS），该蛋白质具有两种不同酶的活性，其 N 端有二氢叶酸还原酶活性，而 C 端有胸苷酸合成酶的活性。此酶对于维持疟原虫四氢叶酸水平极为重要，同时也是疟原虫脱氧胸苷酸生物合成通路中必不可少的酶。而在人体内，二氢叶酸还原酶及胸苷酸合成酶是两种单独的蛋白酶。

（二）Rosetta Stone 方法的应用

利用 Rosetta Stone 方法，在大肠杆菌中已经发现一共有 6809 种潜在的蛋白质间相互作用，而在酵母中共有 45 502 种。值得指出的是，由这种计算机预测得到的蛋白质相互作用网络，必须要对其进行进一步的实验分析以提高精确性。利用迄今已发展的包括如噬菌体展示技术、酵母双杂交系统、免疫共沉淀法、X 射线结晶学以及表面等离子共振技术等多种有效的研究蛋白质间相互作用的高通量的实验技术，这些都为蛋白质组学的发展奠定了坚实的基础。尽管由计算机预测方法得到的蛋白质相互作用的可信度相比由实验技术鉴定出来蛋白质相互作用要低一些，但这些数据有助于对扩大实验设计的范围提供方向，并可从整体的角度观察细胞内所有的蛋白质相互作用。如果对预测得到的蛋白质相互作用网络进行详细分析，并与已有的生物学资源结合起来，将会对研究基因的功能更有价值，并能够为在蛋白质组范围内研究细胞活动机制提供更多的信息。

六、蛋白质组学分析软件与数据库

数据库技术和相关的分析软件，是蛋白质组学研究不可缺少的信息学手段，可用于这些数据的存储、管理与分析（表 9-4）。

表 9-4 常用蛋白质组学分析工具及网址链接

分析工具	网址链接
Base Peak	http://base-peak.wiley.com
EMBL	http://www.narrador.embl-heidelberg.de
ProteinProspector	http://prospector.ucsf.edu
PeptideSearch	http://www.narrador.emblheidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html
Mascot	http://www.matrixscience.com
Mowse	http://www.seqnet.dl.ac.uk/bioinformatics/webapp/mowse
ProFound	http://prowl.rockefeller.edu/cgi-bin/ProFound
Lutefisk97	http://www.lsbc.com:70/lutefisk97.html
M-scan	http://www.m-scan.com
Protana	http://www.protana.com
Expasy proteomics tools	http://www.expasy.ch/tools
NCBI homepage	http://www.ncbi.nlm.nih.gov/
dbest database	http://www.ncbi.nlm.nih.gov/dbest/index.html
Saccharomyces cerevisiae genome database	http://www.genome-stanford.edu/saccharomyces/
SEQUEST	http://fields.scripps.edu/sequest/start.html

（一）常用的蛋白质组分析工具

目前常用的蛋白质组分析工具可分为两大类：

1. 蛋白质表达分布图数据库　这种数据库通常与 GenBank 等核酸序列和表达序列标签(EST)相连接，如日内瓦大学的 ExPASy 系统已成为现阶段蛋白质组学分析中最基础也是最重要的信息来源。

2. 蛋白质组图谱自动识别软件包　如肽图(peptide mapping)包含一个蛋白质全部的质谱(MS)信息，肽段(peptide fragment)则包含蛋白质多个片段的质谱信息(类似于 EST)。这两种策略都需将实测的蛋白质谱与数据库中每种蛋白的理论质谱数据进行比较及统计分析。

此外，其他常用方法(如多序列对位排列和结构排列等)也用于蛋白质结构预测，以提高蛋白质组分析的效率和准确性。

（二）相关分析软件

大规模蛋白质组分析过程主要包括：样品制备、图像分析、蛋白质成分的分析与鉴定，其分析的关键技术包括图像分析、微量测序、肽段氨基酸组分分析和质谱分析等。

1. 图像分析　分析蛋白质组双向凝胶电泳图谱的工作现在都是由计算机图像分析系统来完成(包括斑点检测、背景消减、斑点匹配和图像数据库构建等)。图像采集系统一般采用电荷耦合器件(CCD)相机、激光密度仪和荧光扫描仪等对图像进行数字化，然后进行图像加工和斑点检测。相关软件多是以控制斑点的重心或最高峰来进行边缘检测和邻近分析，一旦图像上的斑点被检测出来后就可以进行不同凝胶间(即样品间)的蛋白质组分的匹配。目前，各种凝胶图像分析软件多是由各生物公司开发并与公司相应的电泳系统进行匹配使用。

2. 微量测序(microsequencing)　经凝胶分离的蛋白质先印迹在 PVDF 膜或玻璃纤维膜上，染色、切割，然后直接置于测序仪中进行序列测定。

3. 质谱技术　质谱数据的计算机处理和蛋白质的数据库搜寻鉴定：获得的数据经计算机处理后，可以用三种方式通过网上数据库搜寻来“鉴定”该蛋白质：①利用 MS 数据搜寻，即 PMF 法；②利用“原始”MS/MS 数据搜寻法，即不对获得的串联质谱数据进行解析而直接用于网上搜寻；③先对串联质谱数据进行解析，获得部分多肽片段氨基酸序列后对蛋白质进行序列查询法鉴定。

4. 肽质量指纹图谱与肽段部分测序　PMF 采用酶(胰酶)对由双向凝胶电泳分离的蛋白在胶上或在膜上于精氨酸或赖氨酸的 C- 末端处进行断裂，所产生的精确的肽相对分子质量通过质谱来测量。所有的肽质量最后与数据库中肽段质量相配比(理论肽由实验所用的酶来“断裂”蛋白所产生)。

5. 氨基酸组分分析　利用氨基酸组分的异质性，可用双向凝胶电泳图谱鉴定蛋白质。通过放射标记的氨基酸来测定蛋白质的组分，或将蛋白质印迹到 PVDF 膜上并进行酸性水解，每一样品的氨基酸自动衍生并由色谱分离。根据代表两组分间数目差异的分数，可以对数据库中的蛋白质进行排列，挑选出最可能的蛋白质。

（三）蛋白质组分析数据库

蛋白质组数据库包含各种已鉴定的蛋白质信息，如蛋白质序列、核苷酸序列、双向凝胶电泳图谱、蛋白质三维结构、翻译后修饰、基因组及代谢数据库等。为了有效地实现信息资源的交换和共享，很多国际知名的实验室和研究机构均建立了各种蛋白质相关数据库。例如，SWISS-2DPAGE 数据库是瑞士日内瓦大学 ExPASy 分子生物学服务器的一部分，包括人类、细菌、细胞等丰富的蛋白质信息。

在蛋白质序列数据库的基础上，研究人员进一步开发了基于数据库的检索程序。这些蛋白质组数据库与蛋白质序列数据库等链接，根据输入检索程序的某些特征参数检索数据库，由匹配结果确定凝胶分离蛋白质的归属。这些特征参数包括：肽质量指纹谱、N- 末端序列、肽序列标签、氨基酸组成、分子量和等电点等。此外，根据蛋白质分离过程中是否进行还原烷基化处理来设定蛋白质的修饰方式，如碘乙酰胺、碘醋酸、4- 乙烯吡啶等。一般网站将分子量、等电点范围与其余某参数结合检

索蛋白质，也有多个参数结合进行检索，以增加结果可靠性。

1. 综合性蛋白质 2DE 数据库 SWISS 2D-PAGE 数据库、Argonne 2D-PAGE 数据库、Max Planck 感染生物学研究所(MPIB)创建的蛋白质 2D-PAGE 数据库、SIENA 2D-PAGE 数据库、捷克军事医学科学院 PMMA 创建的 2D-PAGE 数据库、美国 LSBC 公司研发的 2D-PAGE 数据库、上海生物科学研究所创建的 SBS 2D-PAGE 数据库、中国国家生物医学分析中心的双向电泳 - 质谱蛋白质数据查询平台(NCBA-2D)等。

2. 哺乳类 2DE 数据库 丹麦 Aarhus 大学人类基因组研究中心创建的 2D-PAGE 数据库、英国心脏科学中心 Harefield 医院维护的心脏内皮细胞 HSC-2D PAGE 数据库、德国柏林心脏研究所的人类心肌 2D-PAGE 数据库、美国国立癌症研究中心建立的有关 60 种癌症细胞的 2DE 蛋白质表达数据库、澳大利亚 JPSL 网站组建的有关乳腺癌细胞等细胞组织的蛋白质表达数据库、英国 Durham 大学造血干细胞研究小组创建的造血干细胞分化相关的蛋白质表达数据库、法国 paris13 大学生物化学蛋白质技术实验室创建和维护的有关白血病细胞系的 2DE 数据库、日本东京首都老年医学研究小组创建的年龄相关蛋白质组数据库 TMIG 2D-PAGE、美国华盛顿大学内耳蛋白质数据库(Wash U. Inner Ear PDB)等。此外，尚有各种微生物、真菌类 2DE 数据库，以及植物类 2DE 数据库。

3. Bioknowledge Library Bioknowledge Library(http://www.proteome.com/)是一个有关模式生物蛋白质组的数据库系统。每个数据库中收录了关于单一蛋白质已发表信息，如其生化功能、在细胞和整个生物体中的作用、定位、调节、域和基序及与其他蛋白质的交互作用等。

(四) 质谱信息查询和蛋白质鉴定软件

依肽段质谱信息进行数据库查询和蛋白质鉴定的常用软件有：

1. PepSea(http://pepsea.protana.com/) 一种简单、快捷的肽图分析软件，检索前必须先获得肽序列标签(PST)。在检索较大蛋白时积分较高，随机匹配的可能性也较大。

2. Sequest(http://fields.scripps.edu/sequest/) 与 PepSea 相比较为费时，但可以使用未分离的肽段质谱进行数据库检索。

3. PeptIdent/MultiIdent(http://www.expasy.ch/tools/) 基于遗传算法的蛋白质鉴定软件，训练集来自数据库中已有的蛋白质谱数据。在检索较大蛋白质时随机匹配的可能性也很大。

4. MS-Fit(http://prospecter.ucsf.edu/) 也是利用肽图进行蛋白质鉴定。

5. MOWSE(http://srs.hgmp.mrc.ac.uk/、http://prospecter.ucsf.edu/) 利用数据库中蛋白质的平均组成来提高鉴定的灵敏度。

6. ProFound(http://prowl.rockefeller.edu/cgi_bin/profound) 基于 Bayesian 判别法对可能的蛋白质进行分类，可以同时考虑单一蛋白质或多种蛋白质的信息。

(五) PMF 质谱分析概述

基于凝胶分离的蛋白质组学通常有两种通过质谱来鉴定蛋白质的策略，一种是通过肽指纹图谱(PMF)数据，另一种是通过串联质谱(MS/MS，tandem mass spectrumeter)数据。现主要介绍通过 PMF 数据来鉴定蛋白质的方法。

基于 PMF 的蛋白质鉴定策略的实验数据主要来自于由 MALDI-TOF 对酶切消化蛋白进行质谱分析而得出的质量数列表。胰酶(trypsin)是最常用的蛋白酶，将其酶切消化后的各蛋白质肽段峰质量数数据在 PMF 搜索程序中作为查询内容来搜索模拟数据库，该模拟数据库为将给定的蛋白质序列数据库中各蛋白质进行计算机模拟酶切而得到的理论肽段质量数数据库，最后得到与蛋白质序列数据库中每种蛋白质进行匹配的评分值(score)，该评分值取决于实验肽指纹图谱质量数数据与已知序列蛋白质理论酶切后所得肽质量数数据相匹配的程度。评分越高的匹配蛋白质则很有可能是样品中所鉴定的蛋白质。

由于有众多因素能够影响实验肽图谱质量数数据与理论肽图谱质量数数据的匹配，使得 PMF 法并不是非常可靠的鉴定方法。有些酶切后肽段自身不易被离子化并能干扰其他肽片段的电离，从

而导致某些肽段峰图无法在图谱上观察到，原因尚不清除。另一个问题是，只有质量数处于已记录质量范围的肽段能被观察到，通常介于700~3500Da之间。此外，大量靠计算机模拟酶切无法预测出的肽段峰信号也能在谱图中观察到，这些肽段可能是被修饰过的肽段，在进行PMF搜索时必须考虑到肽段的特定修饰才可能匹配到结果。同时，样品中的来自头发、灰尘的角蛋白等污染也能影响谱图中肽段峰的分析。

多种软件工具能用于PMF搜索，搜索结果通常为与样品中蛋白质能够最相匹配的蛋白质信息列表。许多搜索工具的计算法都考虑了诸如数据库的大小、给定蛋白大小范围内特定肽段质量数的分布频率以及质量数精确性的分布等因素。其他参数，如允许漏切的酶切位点数量、预计的修饰位点、拟搜索的蛋白质序列数据库以及搜索要求的质量数准确度等由操作者进行设置。这些搜索参数决定了搜索的特异性，确定了搜索的参数值，也就限制了可供搜索考虑的蛋白质数量，但在限制范围内使得正确的蛋白质信息不被剔除。

（六）PMF质谱分析基本步骤

1. 核对获得的谱图，扣除本底等因素引起的失真，进行峰值校正及分析范围的选择。

在分析质谱图谱前，需明确几个质谱图相关概念：

（1）相对丰度：以质谱中最强峰为100%(称基峰)，其他碎片峰与之相比的百分数。

（2）总离子流(TIC)：即一次扫描得到的所有离子强度之和，若某一质谱图总离子流很低，说明电离不充分，不能作为一张标准质谱图。

（3）动态范围：即最强峰与最弱峰高之比。若太窄，会造成有多个强峰出头，都成为基峰，而该要的(常为分子峰)却记录不出来。这样的图也是不标准的，检索、解析起来都很困难。因此，要确定好扫描的分子量范围。

（4）本底：未进样时，扫描得到的质谱图，空气成分、仪器泵油、底物、缓冲液及吸附在离子源中其他样品等所导致的背景峰。

获得图谱后，需检查仪器检测质量数是否准确，扫描条件是否合适等，以得到一张标准而可靠的质谱图。

①检查质谱图总离子流、动态范围及本底等值，要控制进样量及放大器放大倍数，还要扣除本底，以得到一张标准的质谱图；②以内标肽段峰质量数为基准对谱图肽段质量数进行校正。内标是指在质谱点样时在样品中加入已知理论质量的肽段，这样可以通过它来校准精确度，从而提高搜库时的可信度。有时，Matrix和Trypsin的自酶解峰就可以作为内标。但通常加入一个分子量3000左右的肽段作为内标，这样大部分酶解的肽段在两个内标之间，校准的精度就更高。一般认为用分子量大于800的肽段去搜库的可信度较好。图9-2是牛血清白蛋白(bovine serum albumin，BSA)PMF图谱的一个例子，图9-3是其500~900区域的放大图。

2. 确定肽指纹谱峰值数据集，剔除与所鉴定蛋白无关的质量峰　来源于其他杂质的肽段峰，如胰酶自消化酶解峰、角蛋白肽段峰及基质峰等会影响PMF的搜库结果，增加假阳性结果的可能，甚至能导致搜库的失败。因此在核对初始谱图时，需要把基质峰、酶自解峰、角蛋白峰、及知道的污染峰等给剔除掉，将剩下的峰值列表进行PMF搜索。

经剔除基质峰、酶自解峰等信号，图9-2中BSA的肽指纹质量数数据集为721.355、927.490、1163.654、1249.633、1305.694、1439.850、1479.815、1567.733、1639.953、1871.888、2044.991。

3. 数据库搜索及参数设置　在将PMF结果进行数据库搜索之前，需要根据样品的准备方法对搜索参数进行设置，如样品来源于何种组织、是纯化还是经过凝胶分离所得、有无经过特殊的化学处理如还原和烷基化。特殊的化学处理有可能对蛋白质的氨基酸残基进行修饰，从而影响数据库的搜索结果。

（1）选择允许的化学修饰：通常根据样品在制备过程中的化学处理方法选择肽段的固定修饰(fixed modification)，根据样品的组织来源可选择肽段的可变/选择性修饰(variable modification)。对

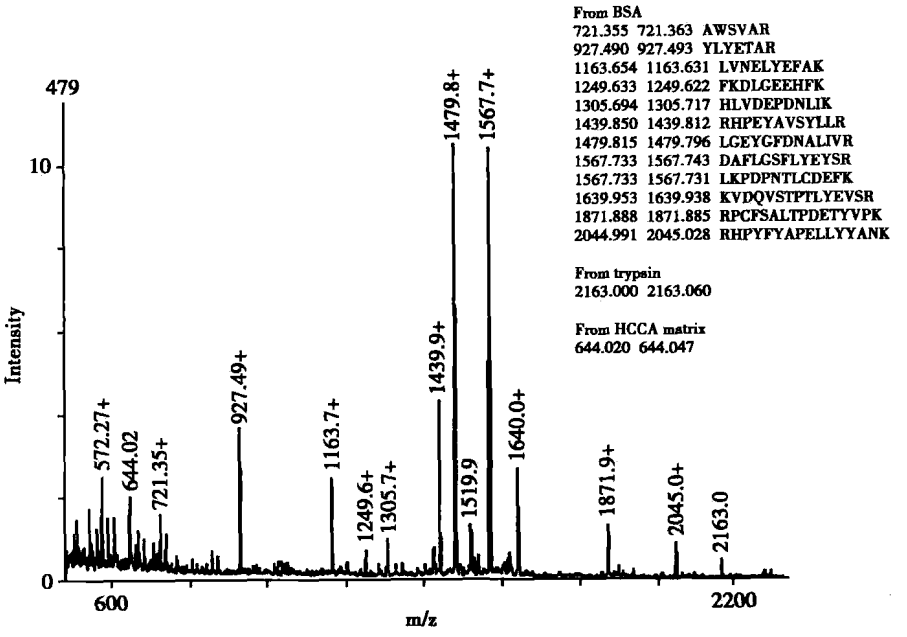


图 9-2 BSA 的 MALDI-TOF 质谱图谱

以牛血清白蛋白(bovine serum albumin, BSA)PMF 图谱为例 右上角显示质谱分析数据，数据分析由 VEMSmaldi v2.0 完成。第一列表示实验肽段质量数，第二列表示理论酶切后肽段质量数，第三列表示 BSA 酶切后各肽段序列 质谱图中各肽段峰上数字表示各峰相对应的质荷比(m/z)值，(+)表示该实验峰质荷比值与理论酶切后肽段峰质荷比值相匹配

(图片来源于：Rune Matthiesen. Mass Spectrometry Data Analysis. Humana Press, 2003.)

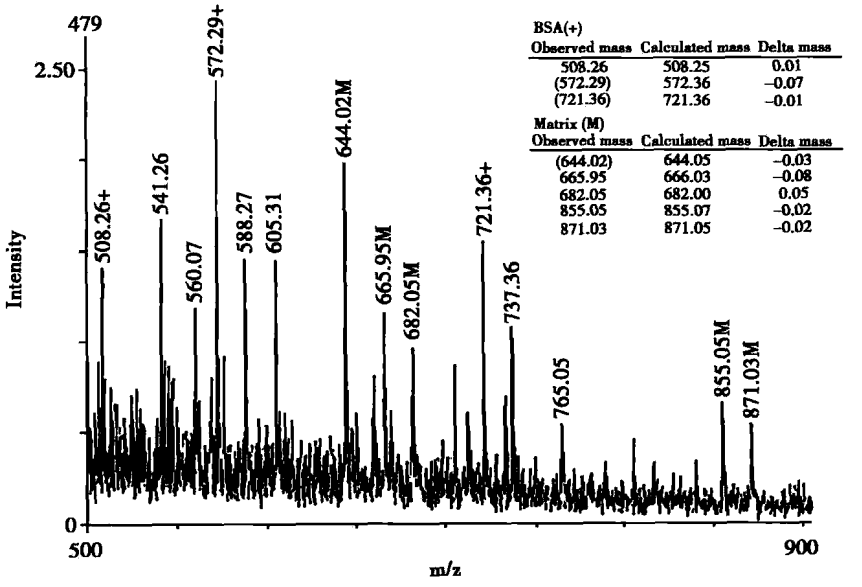


图 9-3 BSA 的 MALDI-TOF 质谱图谱 500~900 区域放大图

各标记肽段峰上，(+)表示 BSA 酶切后肽段峰，(M)表示基质峰 数据分析由 VEMSmaldi v2.0 完成

(图片来源于：Rune Matthiesen. Mass Spectrometry Data Analysis. Humana Press, 2003.)

于真核组织蛋白质，常选择肽段 N 末端的乙酰化(acetylation)及甲硫氨酸的氧化，oxidation(M)，作为可变 / 选择性修饰；半胱氨酸(cysteins)常被还原及烷基化，据烷化剂的不同其固定修饰亦不同，如用碘乙酰胺(iodoacetamide)进行烷基化，应在固定修饰项中选择酰胺甲基化，carbamidomethylation(C)。

(2) 确定可耐受的质量数精确度(mass tolerance)：可耐受的质量数精确度通常指对于这次搜

索可以耐受的精确程度。这一参数描述实验所得肽段峰质量数与搜索理论数据库中质量数的匹配程度，用以反映实验所得数据的质量数准确性，单位可用道尔顿(Da)或百万分率(parts per million，ppm)。一般选择 50~200ppm 或 0.1~0.3Da，可根据需要变化。

(3) 确定酶切所用蛋白酶：最常用来进行 PMF 分析的蛋白酶为胰蛋白酶(trypsin)。该酶酶切位点为 /K/R-\P，其中 /K/R 表明是在赖氨酸(lysine，Lys，K)和精氨酸(arginine，Arg，R)的前面切，-\P 表示若 K、R 后面紧跟着 P(脯氨酸)，则无法酶切。

(4) 确定允许漏切的酶切位点个数：多数蛋白酶消化都可能产生含有酶切位点但未能被酶切的肽段。根据各种蛋白酶的酶切效率，在搜索时最好考虑此类可能被漏掉的酶切位点个数。对于胰蛋白酶，通常允许在每个肽段内有一个漏切的酶切位点。如果搜索失败，即无法得出显著和可靠的匹配结果，可以将此参数设置为 0 或 2 进行再搜索。如果设置为 2，则在搜索后需要对匹配的结果进行人工分析和评价。

(5) 确定肽段质量数值(mass values)及计算模式：肽段质量数值的选择是指确定该质谱图为哪种离子化法的谱图，一般 MALDI-TOF 所电离的肽段都是 M+H 质子化的。质量数值计算模式有单同位素模式(monoisotopic)和平均值模式(average)两种。质谱仪测定的不是酶切后肽段本身的质量，而是质量同电荷的比值(m/z)。在 MALDI-TOF 高分辨率质谱仪检测中，产生的多为单电荷离子，肽段峰通常都是带有一个质子的峰，即单同位素峰。

(6) 根据搜索蛋白的匹配对象选择合适的数据库及物种(taxonomy)限定：至少要选择包含有目标蛋白或其同源蛋白的数据库。通常选择 NCBI 数据库，该数据库涵盖了目前已公开的所有蛋白质序列，最大可能的包含了目标蛋白质的信息。数据库的大小也可通过限制其物种分类来缩小，如此不仅可以减少搜索时间，也可减少跨物种蛋白质鉴定结果的可能性。但这种基于同源蛋白质的跨物种间搜索有助于鉴定那些全基因组序列尚不清楚的物种相关蛋白质。

物种限定不能设的太窄，至今对每个物种的基因还不完全清楚，不同物种可以相互补充。

(7) 确定估计等电点(isoelectric point，pI)及分子量(molecular weight，MW)数值：用来 PMF 搜索的蛋白质样品制备通常都是通过凝胶分离，目标蛋白质的等电点及分子量等值可通过凝胶分析进行估计，并可作为附加参数用以限制搜索结果。但不是所有的搜索程序都会考虑此类参数，通常 pI 和 MW 都不用限制，根据检索结果和已知数据确定。

蛋白质鉴定的软件和算法 MASCOT 和 PEPTIDENT 的参数设置如表 9-5 和表 9-6 所示。

现以牛血清白蛋白(bovine serum albumin，BSA)为例，采用 MASCOT 搜索工具进行 PMF 分析鉴定(图 9-4、图 9-5、图 9-6、图 9-7)。

表 9-5 MASCOT 参数设置

参数	参数设置
Database searched	SWISS-PROT
Taxonomy	Mus musculus
Enzyme	Trypsin
Max missed cleavages	1 or 2
Fixed modifcation	Carbamidomethyl
Protein mass	None
Peptide mass tolerance	±0.5Da
Mass values are	Monoisotopic
Report top	20

注：mascot 的优点在于引入了显著性检验，即在多少分以上实验获得肽段与理论上的酶切肽段之间的匹配是由于蛋白质相同而不是由于随机原因造成的。

表 9-6 PEPTIDENT 参数设置

参数	参数设置
Database searched	SWISS-PROT
Protein mass range	40-60KDa
PI range	4-6
Species searched	Mus musculus
Digested used	Trypsin
Peptide mass accurance	±0.5Da
Cysteines are treated with	Idoacetamide
Peptide masses are	Monoisotopic
Number of peptides requried for match	4
Number of missed cleavage sites	1
Number of matching proteins display	20

注：PepIdent 没有从统计学上来保证数据库的查询质量，其可信度有限。

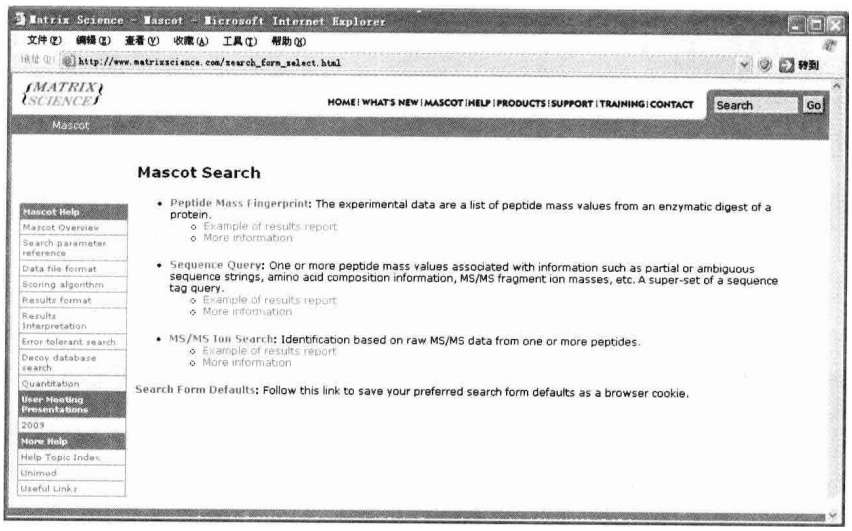


图 9-4 MASCOT 搜索主界面

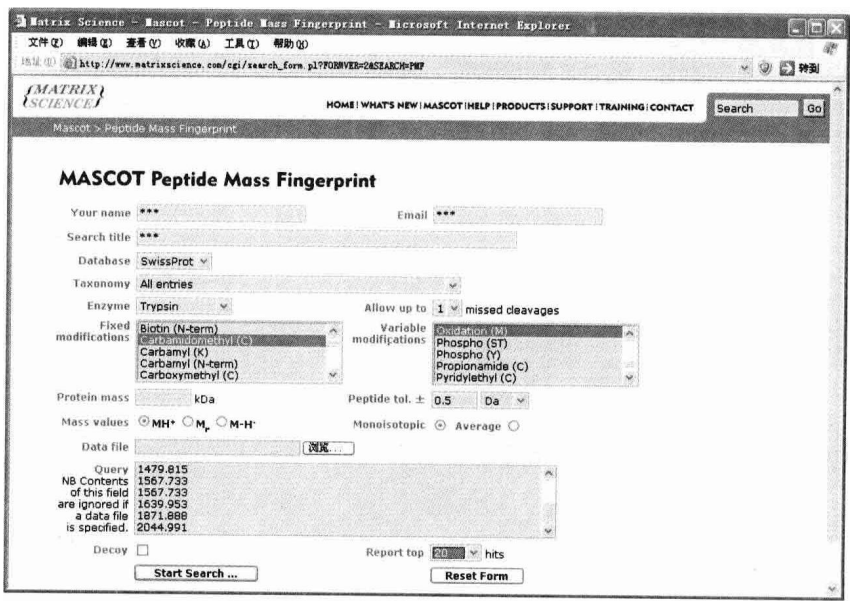


图 9-5 选择 MASCOT Peptide Mass Fingerprint 程序

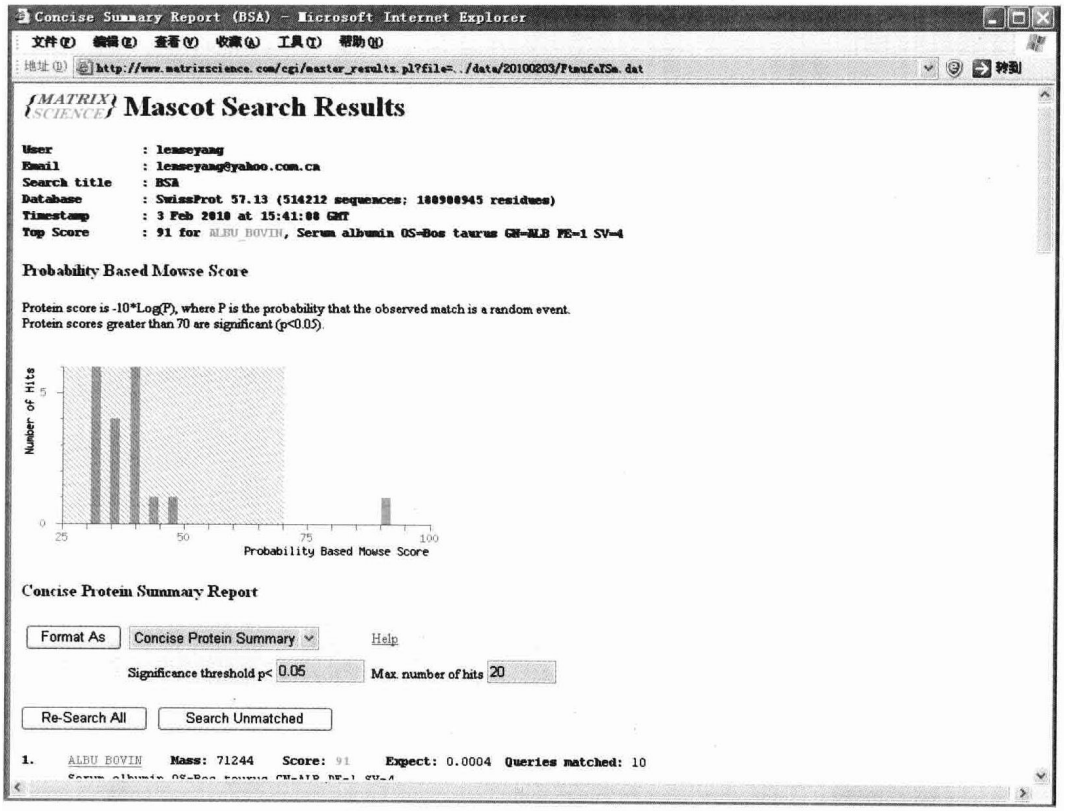


图 9-6 MASCOT PMF 搜索结果界面

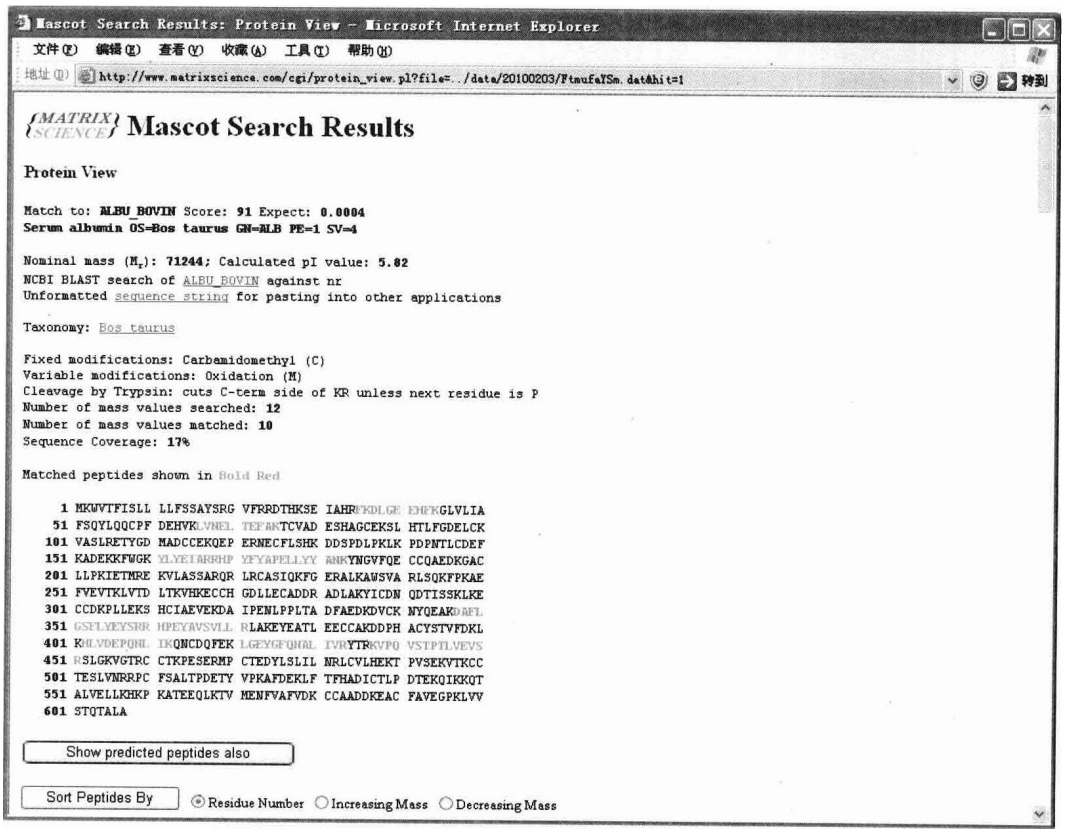


图 9-7 搜索结果蛋白详细信息

小 结

功能蛋白质组学研究是生命科学进入后基因组时代的标志之一。蛋白质组学以蛋白质组为研究对象。蛋白质组最初是指“由一个细胞或一个组织的基因组所表达的全部相应的蛋白质”。功能蛋白质组是指一种细胞、组织或完整生物体在特定时空上所拥有的全套蛋白质。蛋白质分离、蛋白质鉴定、蛋白质相互作用分析及生物信息学数据处理为蛋白质组研究的基本支撑。质谱是目前发展最快、应用最广的蛋白质鉴定技术。研究蛋白质相互作用的技术包括从原子水平、分子水平和细胞水平进行研究。大规模高通量的研究蛋白质相互作用研究方法主要是酵母双杂交、蛋白质芯片技术等。生物信息学在蛋白质组学应用主要包括：构建与分析双向凝胶电泳图谱；蛋白质数据库的建立和搜索；蛋白质结构与功能预测；蛋白质预测分析软件的开发与应用等。常用的蛋白质数据库有UniPro、TreMBL、EXPASY、NCBInr、PDB、PIR 等。这些数据库在国际互联网上的共享，为功能蛋白质组学研究提供了平台。蛋白质组学在人类疾病中的大规模运用，为发掘新的疾病相关生物标志物和揭示疾病发生、转移及耐药机制，为疾病的早期诊断、靶向治疗、新药开发和疫苗研制提供了重要的理论基础。

Summary

Functional proteomics, as a new frontier in life science, has become one of the landmarks during the post-genomics period. The term “proteome” was first coined to describe “the set of proteins encoded by the genome from a given cell or tissue”. Functional proteome refers to the set of all protein subunits and modifications, the interactions between them, the structural description of proteins and their higher-order complexes etc. Protein separation, protein identification, protein interaction analysis and bioinformatics data processing are four basic supporting techniques in proteomic study. Mass spectrometry technology is the fastest developing and the most extensive application technology among various protein identification methods currently. The methods of study on the interaction between proteins include atomic, molecular and cellular levels. Yeast two-hybrid and protein array are large-scale and high throughout methods for investigation on interaction between proteins. Application of bioinformatics on proteomics includes construction and analysis of 2-DE map, construction and search of protein database, structure and function prediction of protein, development and application of related software etc. UniPro, TreMBL, EXPASY, NCBInr, PDB and PIR are the familiar protein databases. The sharing of these databases online will build a platform for proteomics, and the latter will pave the way for mining of new biomarkers of diseases, for understanding mechanisms of occurrence, migration and drug-resistance of diseases and for early-stage diagnosis, targeted treatment, development of drugs and vaccines.

（杨琳琳 李学荣 吴忠道）

习 题

1. 简述蛋白质、蛋白质组、蛋白质组学和功能蛋白质组学的定义。
2. 名词解释：MALDI TOF MS；肽质量指纹谱；蛋白质翻译后修饰。
3. 鉴定蛋白质翻译后修饰的重要方法是：

A. 双向电泳
B. 质谱
C. 蛋白质芯片技术

4. 蛋白质组学的主要研究领域和目标是什么?
5. 简述双向电泳的基本原理及操作步骤。
6. 试述蛋白质芯片技术的原理及其应用前景。
7. 试述蛋白质功能预测的思路及一般流程。
8. 常用蛋白质组学数据库获取与分析的技术有哪些?
9. 试述蛋白质相互作用的研究方法。
10. 试述蛋白质组学理论与技术对临床医学的意义。

主要参考文献

1. 李伯良. 功能蛋白质组学. 生命的化学. 1998, 18: 1-4.
2. 邱宗荫, 尹一兵. 临床蛋白质组学. 北京: 科学出版社; 2008.
3. Bock J. R., Gough D. A. Whole-proteome interaction mining. Bioinformatics, 2003, 19(1): 125-134.
4. Chautard E., Thierry-Mieg N., Ricard-Blum S. Interaction networks: from protein functions to drug discovery. A review. Pathol Biol (Paris), 2009, 57(4): 324-333.
5. Gorg A., Postel W., Gunther S. The current state of two-dimensinal electrophoresis with immobilized pH gradients. Electrophoresis, 1988, 9: 541-546.
6. Jonathan P. 生物信息学与功能基因组学. 孙之荣, 主译. 北京: 化学工业出版社; 2008.
7. Larkin S. E., Zeidan B., Taylor MG., et al. Proteomics in prostate cancer biomarker discovery. Expert. Rev. Proteomics, 2010, 7(1): 93-102.
8. Nguyen T. P., Ho T. B. An integrative domain-based approach to predicting protein-protein interactions. J. Bioinform. Comput. Biol., 2008, 6(6): 1115-1132.
9. Pandey A., Mann M. Proteomics to study genes and genomes. Nature, 2000, 405: 837-847.
10. Richard J., S. 蛋白质与蛋白质组学实验指南. 何大澄, 主译. 北京: 化学工业出版社; 2006.
11. Skrabanek L., Saini H. K., Bader G. D., et al. Computational prediction of protein-protein interactions. Mol. Biotechnol., 2008, 38(1): 1-17.
12. Wasinger V. C., Cordwell S. J., Cerpa-pojak A., et al. Prosgress with gene, product mapping of the Mollicutes: mycoplasma genitalium. Electrophoresis, 1995, 16: 1090-1094.
13. Woychik R. P., Klebg M. L., Justice M. J., et al. Functional genomics in the post-genome era. Mut Res. 1998, 400(1-2): 3-14.
14. Yates J. R. Mass spectrometry from genomics to proteomics. Trends Genet, 2000, 16: 5-8.

第十章 蛋白质结构分析
CHAPTER 10 ANALYSIS OF PROTEIN STRUCTURE

第一节 引 言
Section 1 Introduction

1958年解析和确定了第一个球蛋白(肌红蛋白)的三维结构，随后展开了对蛋白质结构与功能关系的广泛研究，发现不同蛋白质有独特而复杂的结构特征。Sanger证明每种蛋白质的氨基酸残基序列都是独特的。1961年Anfinsen研究了核糖核酸酶的变性和重折叠，提出了蛋白质一级结构(氨基酸序列)决定其三维结构的原则。蛋白质只有折叠成特定的空间构象才能具有相应的活性和生物学功能。分子生物学的中心法则确定了DNA与蛋白质氨基酸序列间的关系，称为第一套遗传密码子；确定蛋白质氨基酸序列与三维结构间的关系，被称之为“第二套遗传密码子”。因此，发掘蛋白质序列中决定高级结构的信息和高级结构中决定功能的信息，是深入解析中心法则的关键。

蛋白质在生命活动过程中有复杂而精细的生物学功能。参与生命活动的蛋白质高级结构信息包含二级结构(secondary structure)、三级结构(tertiary structure)和四级结构(quaternary structure)等。蛋白质能发挥各种各样的精细功能都是以其高级结构及基于此高级结构与其对应分子发生高度特异的相互作用为基础。通常蛋白质的某种精细的局部结构可用于实现某种局部的生物化学功能。蛋白质高级结构的形成和变化都遵循基本的物理化学原理。发掘蛋白质高级结构的特征信息是理解蛋白质行使其生物功能的机制、认识蛋白质与蛋白质(或其他分子)间相互作用的基础，对于生物医药整个领域的研究也是非常重要的。

可用实验方法解析蛋白质的高级结构，但实验测定过程非常复杂，技术难度大且成本较高。因此，至今已测定高级结构的蛋白质数量比已知氨基酸序列的蛋白质数量要少得多。另一方面，不依赖于测定高级结构的实验技术，而根据结构已知的蛋白质的结构特征和形成规律，并在此基础上发展理论方法来预测未知蛋白质的高级结构，正逐步成为结构生物信息学研究领域的重点。蛋白质序列数据的快速积累和高通量结构分析技术的应用大大加快了结构生物信息学的发展，从而促进了蛋白质结构的预测、基于蛋白质结构的新功能预测、以蛋白质为靶点的配体类药物设计等基础与应用研究的快速发展。

在结构生物信息学领域，蛋白质结构分析的主要目标是建立研究蛋白质结构信息发掘与预测的方法；利用这些方法和技术，研究参与生命活动过程中蛋白质的物理性质、空间架构(结构)、功能片段和相互作用，进而探索基于蛋白质结构表征的生物学意义和得到新的预测性知识。

第二节 蛋白质的高级结构
Section 2 Advanced Structures of Protein

一、蛋白质的高级结构特征

蛋白质高级结构即蛋白质构象(conformation)，有其自身形成规律；活性蛋白质有最适构象但此

构象有动态性(dynamicity)。蛋白质构象源自二级结构的精巧组装，提取蛋白质高级结构中的二级结构信息成为发掘高级结构同其序列及生物学功能之间联系的必然途径。

(一) 二级结构的主要类型和特征

蛋白质的二级结构是指多肽链主链骨架盘绕折叠而形成的构象，借氢键维系。其主要分为α螺旋、β折叠、β转角及无规卷曲等。

1. α螺旋　α螺旋(α-helix)的结构特征为：①主链骨架围绕中心轴盘绕形成右手螺旋；②螺旋每上升一圈是3.6个氨基酸残基，螺距为0.54nm；③相邻螺旋圈之间形成许多氢键；④侧链基团位于螺旋的外侧。通常，球状蛋白质直径有限，其所含α螺旋长度有限，一般不超过11个残基，长度约17Å(图10-1)，而纤维蛋白α螺旋长度可达数十纳米。脯氨酸残基含α亚氨基，使其形成的二面角φ受到限制不能参与α螺旋，这是识别α螺旋起点和终点的重要标志之一，见图10-1(E)和(F)。

如图10-1(D)所示，α螺旋形成后，氨基酸残基的侧链全部位于外侧，多肽链的特殊序列形成α螺旋时可将不同性质的侧链归集到不同侧面，有利于促进α螺旋组装和与环境相容；只要单个α螺旋的外部侧链足够疏水，此α螺旋就可独立存在于膜脂中，如某些典型跨膜蛋白。不同氨基酸的侧链性质差异使其出现在α螺旋中的概率不同。除了脯氨酸外，Gly、Tyr和Ser都不利于α螺旋形成，但Ala、Glu、Leu和Met都促进α螺旋形成，这些信息对识别和预测α螺旋也有辅助价值。

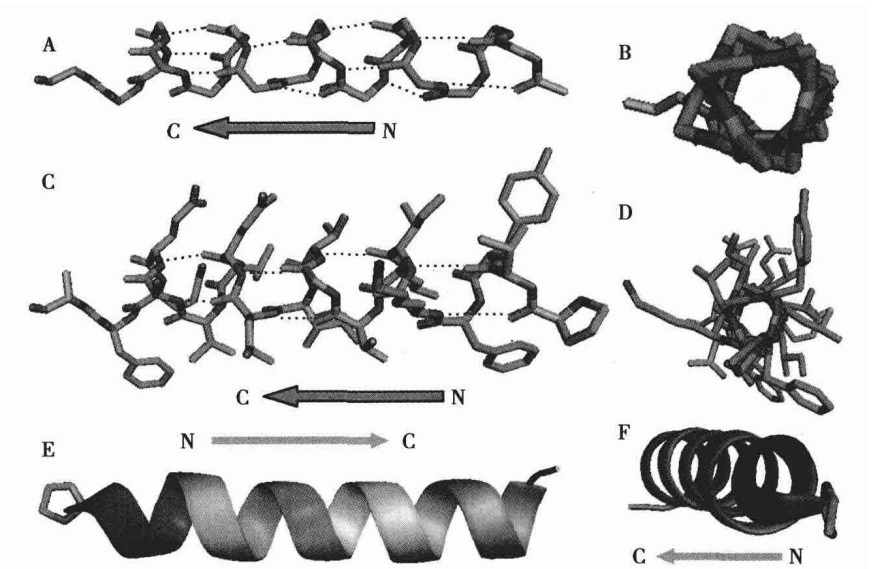


图10-1　人细胞珠蛋白(2DC3.pdb)的第121到140位残基对应的α螺旋侧面和顶部(N端)视图

A. 隐藏侧链显示的树枝状主链的侧面视图；B. 隐藏侧链的树枝状主链的顶部视图；C. 显示树枝状侧链的侧面视图；D. 显示树枝状侧链的顶部视图；E. 用飘带显示的侧面视图；F. 用飘带显示的顶部视图　图A和C中虚线为α螺旋中每四个氨基酸残基的肽键基团之间形成氢键的示例　其中绿色表示C原子，蓝色表示N原子，红色表示氧原子　图E和F中显示肽链的N端第121位对应的脯氨酸残基，其侧链没有进入此段α螺旋区域　图形均用Pymol Version 0.99输出；※如未另作说明，本章以后图片中原子标示颜色与此同

2. β折叠　β折叠(β-sheets)是描述主链的另一种特殊走向模式，由位于不同区段的多个连续主链片段(β折叠链)相互靠近，连续片段肽键基团间形成氢键并侧向聚集构成(图10-2)。其结构特征为：①若干条肽链或肽段平行或反平行排列成片；②所有肽键的C=O和N−H形成链间氢键；③侧链基团分别交替位于片层的上、下方。通常，球蛋白直径限制每股β折叠链的长度在3到10个残基之间。球蛋白中一条主链形成β折叠结构必然需要与其他二级结构连接。

多肽链主链的方向性(从N端到C端)决定了片层中任两股的方向可同或不同，故存在平行和反平行的β折叠，见图10-2(C)。另外，脯氨酸的特殊β折叠氨基结构使得其也不利于形成β折叠，见图10-2(C)，其他不同氨基酸形成β折叠的能力也各不相同。不同性质侧链可聚集在β折叠的上层

或下层，使得β折叠具有两亲性。但β折叠对外部的接触面积较大，分布在膜蛋白表面与疏水膜酯类接触需要较大体积且较多的疏水侧链，故其出现在膜结合蛋白与膜脂接触面的频率比较α螺旋低。如β折叠链扭曲形成封闭的β折叠桶则可稳定存在于膜脂中，此组装模式常见于离子通道蛋白。

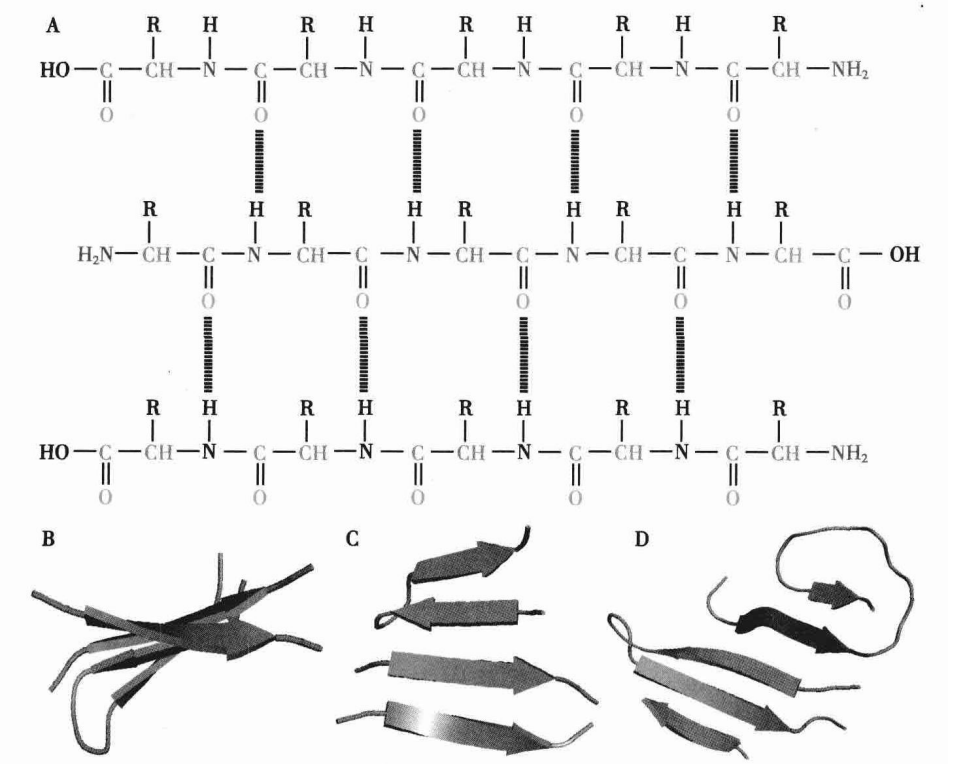


图 10-2 β-折叠示意图

A. 反平行和平行的多个β折叠链形成一个完整β折叠结构的氢键示意图；B. 来自人 pi 型谷胱甘肽 -S- 转硫酶中单个亚基中连续主链的部分β折叠结构(2DGQ.pdb)侧面视图，可见转角(turn)；C. 来自人 pi 型谷胱甘肽 -S- 转硫酶一个亚基中连续主链的部分β折叠结构顶部视图，可见转角(turn)；D. 来自人信号传递蛋白 SMAD4(1DD1.pdb)的一个亚基中部分β折叠结构顶部视图，可见到大的环区(loop)

3. β转角 β转角是连接相同主链上α螺旋和β折叠等二级结构的关键结构(图 10-3)。β转角常出现在水溶性球状蛋白的表面，也是与其他生物分子相互作用的常见位点。其结构特征为：多肽链 180°回折部分，通常由四个氨基酸残基构成，借 1、4 残基之间形成氢键维系。在球蛋白中β转角因其特殊结构作用有较高保守性。Asp、Asn、Ser、Thr 和 Gln 等常出现在β转角中，这些残基侧链亲水性强，有利于β转角出现在水溶性球状蛋白的表面(图 10-3)。另外，不利于形成α螺旋和β折叠的 Gly 和 Pro 也经常出现在β转角。同时，β转角及其附近也是蛋白质发生化学修饰最常见的位点。

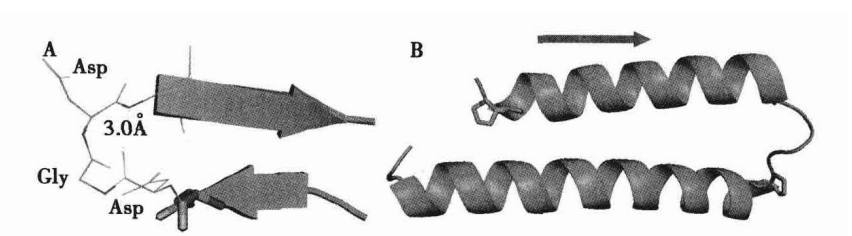


图 10-3 β-转角及其连接的β-折叠链和α-螺旋

A. 人谷胱甘肽 -S- 转硫酶 pi 第 56 到 59 位残基的β转角连接了来自相同主链的两段β折叠链，片层末端残基显示为粗枝状，β转角中 Gly 和 Asp 显示为细线，转角区域内第一个 Asp 的α羰基氧与其后第三位α氨基成氢键(3DGQ.pdb)；B. 来自人细胞珠蛋白(2DC3.pdb)的两段α螺旋由β转角连接，用粗树枝状显示了两段螺旋末端的脯氨酸

β转角的空间定位和对环境的敏感性，是含β转角蛋白质功能精细调节的结构基础之一。

另外，涉及更多氨基酸残基且外形类似于β转角的结构也是连接蛋白质中其他二级结构的过渡结构，称为Ω环(Ωloop)，图10-2(D)。此类二级结构涉及8个左右的残基，一般不超过16个残基；其主要作用是改变主链的走向，使得主链出现折叠、回环等，形成β折叠片所常见的连接结构。Ω环与β转角区分的主要标志之一是起点与终点间的氨基酸残基数量和形成氢键的模式。

4. 无规卷曲　无规卷曲是规律性较低而难以描述的特殊类型二级结构，其所涉及的残基数量差异大，整体外形变化大，可采取多种折叠且不同构象间的能量差异小而容易相互转变，故其结构的规律性很低。无规卷曲实际是描述此类二级结构的习惯用语。无规卷曲在球状蛋白质表面出现较多，也是连接其他规则二级结构的结构模式。无规卷曲各种构象状态仍遵循物理化学原理，但构象波动较大，对温度变化敏感；实验测定结构时往往缺失其坐标，即使有坐标其温度因子也较高。无规卷曲同Ω环的区分主要是其长度和形状的波动性。

（二）超二级结构和结构域的主要类型和特征

超二级结构(supersecondary structure)指位于同一主链的多个二级结构组装形成的特定组装体，可直接作为三级结构或结构域的组成单元，是从蛋白质二级结构形成三级结构的一个过渡结构形式，也称为立体结构形成的模体。

超二级结构主要有如下类型：①β转角或Ω环等连接连续四个α螺旋形成的四α螺旋捆；②中部固定位置含有亮氨酸及其他疏水侧链氨基酸残基、在螺旋两端含有强亲水侧链氨基酸的α螺旋组成的亮氨酸拉链(Leucine zipper)；③一条主链中相邻七个两亲α螺旋通过过渡结构形成的七次穿膜螺旋组；④连续主链中两段α螺旋连接三段β-折叠链形成的Rossmann折叠；⑤β-转角连接α螺旋构成的α螺旋-β转角-α螺旋；⑥Ω环连接α螺旋构成的α螺旋-Ω环-α螺旋等。⑦β-折叠都为超二级结构。超二级结构通常并不对应生物化学功能，但其结构模式是解析蛋白质组装机制的关键信息之一。

结构域(domain)也是蛋白质构象中二级结构与三级结构之间的一个层次，它是在较大的蛋白质分子中，由于多肽链上相邻的超二级结构紧密联系，形成在空间上可以与蛋白质亚基结构明显区别的结构形态。一般每个结构域约由100~200个氨基酸残基组成，各有独特的空间构象，可承担特定的生物化学功能。

（三）三级结构的主要类型和特征

蛋白质三级结构即蛋白质分子中所有共价相连原子的空间相对位置，由多肽链在二级结构的基础上进一步盘绕和折叠形成；蛋白质如有特殊的必须辅基，其三级结构也包括来自这类辅基的原子的空间位置。稳定蛋白质三级结构主要靠氨基酸侧链之间的疏水相互作用、氢键、二硫键、范德华力和静电作用等。不同类型的蛋白质局部结构分解后可具有很高的相似性，但在三级结构层面不同蛋白质所体现的各自整体结构特征通常不同。

1. 水溶性蛋白质三级结构的基本特征　稳定的活性水溶性蛋白质构象有如下特征：①各个二级结构恰当组装以使疏水性侧链的氨基酸主要集中在内部，同时把尽可能多的亲水侧链氨基酸残基分布在表面；其内部需要尽量压缩以减小体积，从而容易维持稳定的动态高级结构并获得稳定内核作为结构骨架(图10-4)；内部还同时生成氢键、离子键等以减弱来自主链和被包埋极性侧链的不利偶极相互作用，在多数蛋白质中主链肽键基团有90%左右参与形成氢键；在蛋白质表面通过氢键结合水分子生成水化膜以稳定蛋白质。水溶性蛋白质主要是球形的，表面亲水而内部也有氢键或离子键等相互作用。但表面有过多或局部暴露区域有较多疏水残基的构象在水溶液中是不稳定的，容易聚集而不溶。②通过表面对应区域的形状和理化性质互补和多位点协同实现与相应分子的特异性相互作用。表面应有裂穴状结构以与小分子多位点相互作用提高特异性，即使同其他蛋白质等大分子相互作用也需粗糙表面以促进多位点接触而提高特异性。因而蛋白质的表面应粗糙而非光滑。③球状蛋白质结构有柔韧性，且总处于亚稳定状态，不仅结构动态可变，且在结构变化到显著改变其活性之

前还可逆恢复构象并恢复活性；不同结构区域可相对独立发挥作用，但依赖于整体的结构；当整体构象发生显著改变后（即变性），局部结构将难以维持，局部功能也难以维持。这些特征是识别水溶性蛋白质活性位点、相互作用位点的重要依据。

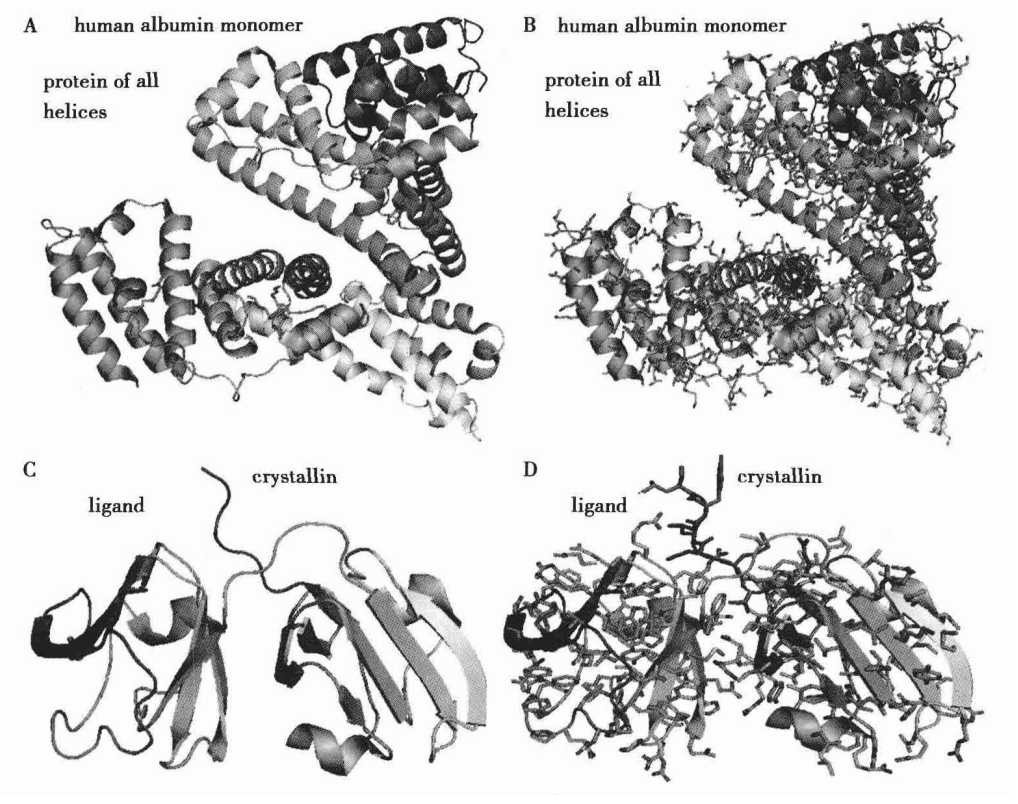


图 10-4 代表性蛋白质的三级结构

A. 飘带显示全α螺旋人血清白蛋白单体三级结构，结构略微松散（2T2Z.pdb）；B. 飘带显示全α螺旋人血清白蛋白单体三级结构，树枝状显示氨基酸侧链，结构明显紧密；C. 飘带显示全β折叠人晶状体蛋白三级结构，结构略微松散（2JDF.pdb），全蓝色的树枝状结构为配体；D. 飘带显示全β折叠人晶状体蛋白的三级结构，树枝状显示氨基酸侧链，结构非常紧密，全蓝色的树枝状结构为配体

2. 膜蛋白三级结构的基本特征 膜蛋白难结晶，因此，已解析结构的膜蛋白较少。膜蛋白构象形成规律性依然遵从上述基本原理。膜蛋白中最常见的二级结构是α螺旋。在膜中α螺旋可独立稳定存在而在水溶液中却很难；α螺旋中主链形成的氢键封闭在内部而侧链全部暴露在外侧，只要这些暴露的侧链足够疏水则此α螺旋可稳定存在。β折叠作为二级结构在膜蛋白中主要位于三级结构内部；全β折叠膜蛋白必须采用特殊的组装模式，即形成封闭的β折叠链桶，以避免β折叠最外边的两条折叠链暴露于膜脂，而且β折叠链需要适度扭曲以使得能用最少的β折叠链形成桶状结构，这在离子通道蛋白中常见。在膜蛋白内部可形成很强氢键以消除内部极性基团的不利偶极作用。细菌视紫红质、离子通道、膜表面标志物和膜受体等是膜蛋白质的代表（图 10-5）。细菌视紫红质结晶时含有结合的脂类物质，图 10-5（A~C）。

3. 蛋白质三级结构中二级结构的折叠和组装 通常按二级结构组装模式对蛋白质进行分类。蛋白质二级结构组装模式主要是全α螺旋、全β折叠、α螺旋/β折叠，还有少量α螺旋+β折叠类。全α螺旋蛋白质几乎不含β折叠，主要由α螺旋加上连接结构组成，典型代表如人血清白蛋白，见图 10-4（A、B）和细菌视紫红质，见图 10-5（A~C）；全β折叠蛋白质几乎不含α螺旋，由β折叠和连接结构组成，代表如人晶状体蛋白，见图 10-4（C、D）和某些离子通道，见图 10-5（F）；α螺旋/β折叠蛋白质指中心为β折叠外围为α螺旋的蛋白质，如细胞表面标志蛋白 CD98，见图 10-5（D）及糖酵解的绝大多数酶蛋白，见图 10-6（A）；α螺旋+β折叠类蛋白质指含有α螺旋和β折叠，且中间有β转角等其

他结构进行分隔，其规律性较弱且较少见，典型代表如 TATA- 盒结合蛋白，见图 10-6(B)。

目前，已有数据库(如 SCOP 和 CATH 等)基于蛋白质折叠等结构信息对蛋白质进行分类，详见本章第三节。

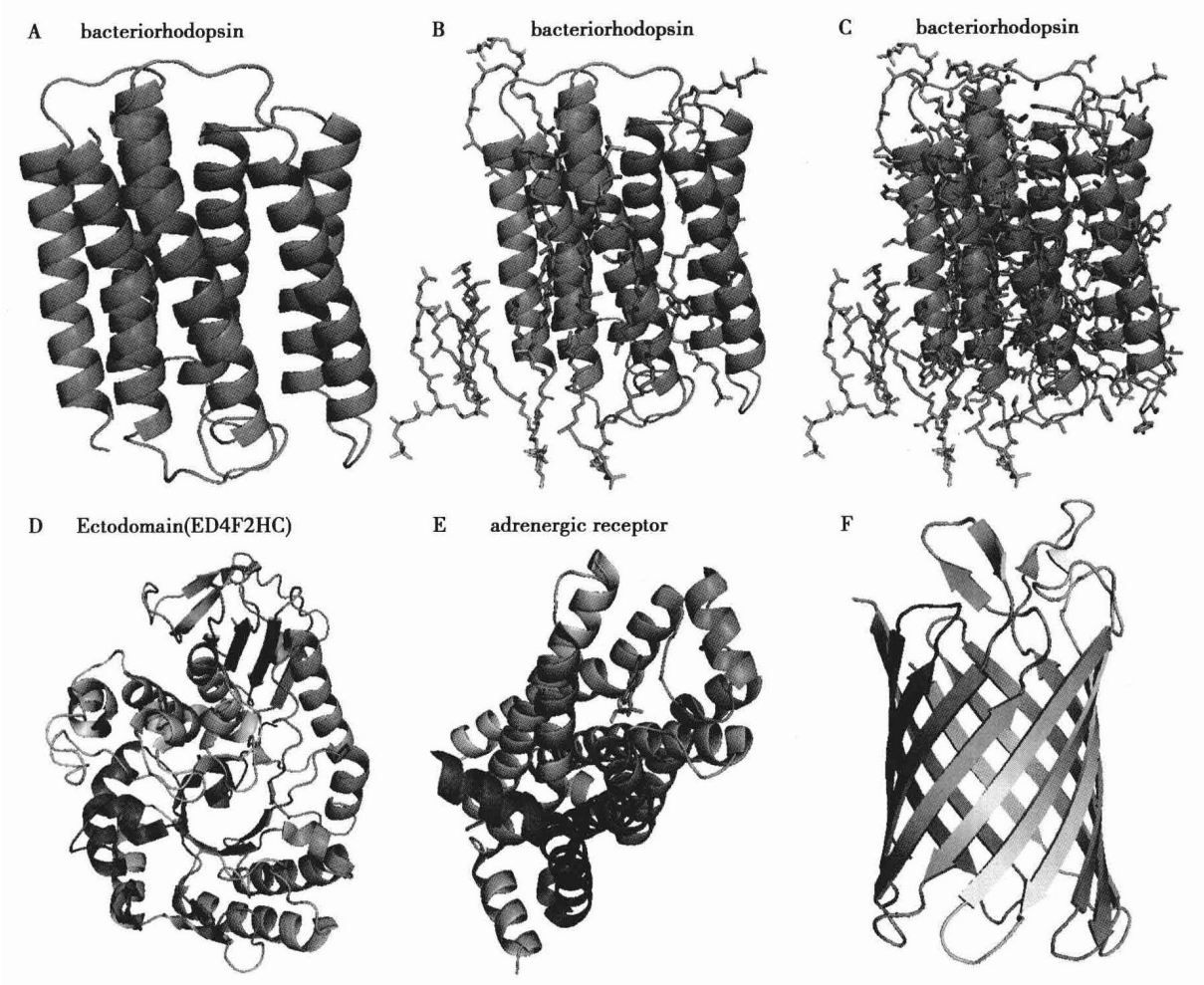


图 10-5 代表性膜蛋白

A~C. 细菌视紫红质蛋白，结晶时结合了大量脂类(2BRD.pdb)；D. 人淋巴细胞激活抗原 CD98(2DH2.pdb)；E. 鸡 β1- 肾上腺素受体，七螺旋跨膜蛋白(2VT4.pdb)并结合有其配体；F. 大肠杆菌 NANC 离子通道蛋白(2WJR.pdb)

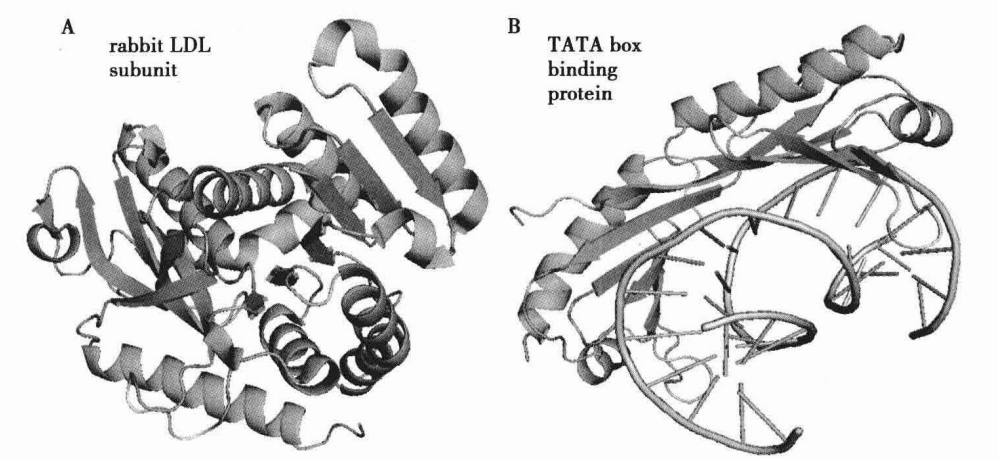


图 10-6 代表性的 α/β 和 α+β 组装模式蛋白

A. 兔子肌肉 LDH 亚基中的构象(3H3F.pdb)；B. 人 TBP 与双螺旋 DNA 复合物(1CDW.pdb)

（四）四级结构的主要类型和特征

有独立三级结构的单元通过非共价键聚集成的非共价复合物称为四级结构，其所含独立三级结构单位为亚基(subunit)。

四级结构是独立三级结构形成的复合物，其中每个独立三级结构为亚基，也称为单体(monomer)。含两个亚基蛋白质称二聚体(dimer)；含三个亚基则称三聚体(trimer)；含四个亚基则称四聚体(tetramer)；含五个亚基则称五聚体(pentamer)；含六亚基则称六聚体(hexamer)。水溶性蛋白质四级结构主要含偶数个亚基且数量大多在六个以下，含奇数个亚基或超过六个亚基的蛋白质都较少。

形成四级结构全部依靠非共价键相互作用，且来自不同亚基的二级结构间可发生强的相互作用以稳定四级结构，如生成跨亚基的更大β折叠结构或α螺旋聚集体；其中，氢键、疏水相互作用和静电作用是主要维持力。为了形成稳定的四级结构，必然要求相互作用的任两个蛋白质间在空间外形互补以增加接触面且理化性质互补。这些特征也是预测蛋白质间相互作用时有用的辅助依据。

图 10-7 PBO-1 蛋白质呈现的对称结构

通常四级结构能可逆解离或聚合并多数伴随活性的显著改变，这是蛋白质序列改变后蛋白质构象改变并改变蛋白质聚集状态从而引起疾病的重要机制(详见本章第六节)。在外观形状上，偶数亚基形成的四级结构具有较高的对称性，包括中心对称和二面体对称等方式(图 10-7)。水溶性四级结构蛋白质主要是球形，而胶原和鞭毛等蛋白质具有纤维状四级结构，其水溶性较差。

具有四级结构的蛋白质通常有多个相同或不同的活性位点，比单纯的三级结构蛋白质具有更复杂的功能和调节机制。特别是形成同亚基的四级结构，对生物体是一种简约化的行为，因为可用短的基因片段编码一个很大的蛋白质，这无疑比直接编码一个大蛋白质节约且可靠。

二、蛋白质二级结构的测定与指认

二级结构识别是蛋白质高级结构识别的基础，目前主要有如下识别策略。

1. 蛋白质二级结构词典(dictionary of secondary structures of proteins，DSSP) 来自模式识别技术，其仅依据主链肽键基团的坐标判断主链肽键基团间是否形成氢键，计算氢键能量低于 −0.5kcal/mol 则有氢键形成，用于搜索α螺旋和β片层结构是否存在。α螺旋定义为，主链的第 i 个氨基酸的α羧基中羰基氧原子与第 i+4 个肽键的 -NH- 形成氢键系统($i \rightarrow i+4$ 主链氢键)的存在；出现α螺旋和β片层交替时，α螺旋优先；一个最小长度的α螺旋为有两个主链氢键；但此定义系统没有约定螺旋的起点和终点。DSSP 中的β片层结构被定义为两段主链间含有氢键，或被β片层中的两个以上氢键所围绕。这种定义模式包含了经典的平行、反平行及氢键围绕等模式，最短的β折叠链包含两个残基。

2. 原子坐标搜索二级结构(search secondary structures in proteins from atomic coordinates，STRIDE) STRIDE 程序用特殊方法判定主链肽键之间的氢键是否存在并用二面角参数辅助识别指认二级结构。此方法所需二面角参数来自已知二级结构参考蛋白质数据的平均和优化。但是不同蛋白质的测量精度不同，所选的参考蛋白不同都会带来识别结果的差异。这也是不同自动化识别系统差异的主要原因之一。与 DSSP 类似，STRIDE 也定义最短的α螺旋为两个连续的 $i \rightarrow i+4$ 主链氢键，但如果二面角限制，则缩短螺旋甚至判断氢键不合理。最小的片层结构也为两个残基组成。

3. 模板 另一种二级结构识别方法，将待测主链的α碳原子的坐标与作为参考的二级结构模型进行叠合，测定α碳原子坐标与参考模型的偏离，只要偏离在允许的范围内，就定义属于对应的二级

结构类型。这种方法从模式识别的角度最容易理解，但显然参考模型的选择和距离偏离矩阵的界限决定了结果的可靠性。其他类似的测定方法也在建立中。

目前还没有满意的二级结构指认方法。现有的二级结构指认方法中，DSSP 比较常用。对二级结构的识别和基于序列对二级结构的预测，是预测蛋白质三级结构的重要手段，这一直是结构生物信息学的研究热点。

三、蛋白质结构域与家族分类

蛋白质的复杂结构和功能依赖于多个结构域的协同；蛋白质缺失某个结构域则其必然缺失对应的生物化学功能。据蛋白质序列相似度或生物化学功能与结构的相似度可将蛋白质分类为家族(family)；同一家族蛋白质有某种类似的生物化学功能或者类似的高级结构。因此，了解蛋白质结构域及家族分类信息，对于蛋白质结构分析有着很重要的意义，并且很多蛋白质家族分类数据库是基于结构分类基础之上构建的。

（一）蛋白质结构域

结构域是构成蛋白质亚基的紧密球状区域，为介于二级与三级结构之间的一种结构层次；是蛋白质中可以具有相对独立三级结构的部分，通常由一个基因外显子编码，并可具有特定的功能。在较大的蛋白质中结构域之间通过较短的多肽柔性区互相连接；蛋白质的结构域有时还可分为一些次级结构，称为组件(module)。组件是在稳定的蛋白质功能域中常见的一种进化上保守而又独立的折叠单位，也是在进化压力下发生外显子迁移的基本单位，它还参与新基因的产生。结构域可以作为蛋白质三级结构的组件，通常不具有完整的生物学功能但有特殊的生物化学作用，这也是结构域与三级结构的关键区别。

在一级结构中的氨基酸序列的某些区域相邻的氨基酸残基形成有规则的二级结构(如 α 螺旋、β 折叠、β 转角和无规卷曲等)，然后再把相邻的二级结构片段集装在一起，形成超二级结构，在此基础上，多肽链再进一步折叠，或为近乎球状的三级结构就可成为一个结构域。最常见的结构域约含有 100~200 个氨基酸残基，一般至少有 40 个、多的可达 400 个以上；对于较小的蛋白质分子或亚基，其结构和功能都较简单，难以区分出独立的不同结构域，这类蛋白质是属于单结构域分子(如卵溶菌酶等)。对于一个较大球状蛋白质分子来说，一条很长多肽链往往由两个或两个以上相对独立的三维实体缔合而成三维结构体。从功能角度看，很多蛋白质属于多结构域的蛋白，其功能位点基本都位于结构域之间，这是由于：①通过结构域容易构建具有特定三维排布的功能中心；②结构域之间常只有一段肽链相连，使域间容易发生相对运动，这将有利于功能位点与对应成分相互作用或施加应力，有利于产生别构效应而对蛋白质的功能实现精细调节。

（二）蛋白质结构比对与算法

对蛋白质高级结构的比对(alignment)是探索远源蛋白质的有效办法，也是从高级结构预测蛋白质功能、基于结构对蛋白质进行分类等的重要基础。但结构比较算法很复杂且至今还没有统一，目前的代表性方法有 VAST、CE、FRACT 等，大多有对应的免费软件可用。

现有方法对两个蛋白质进行结构比较时大多涉及如下步骤：①用不依赖于坐标的方式描述两种蛋白质结构；②比较两种蛋白质的结构；③优化两种蛋白质间的比较；④与随机情况比，计算相似性的统计量度。结构比较要求给出最优化比对并对已有蛋白质与新目标的结构相似性给出排序或量化，这造成在搜索数据库时计算工作量过于庞大。为了提高计算效率，可排除序列相似性高(>30%)和折叠类型差别大的蛋白质；只选择同种折叠类型中的代表蛋白质主链与目标蛋白质进行结构比较，但此代表性蛋白质主链的选择标准很复杂。在 CE 算法中按照如下标准选择代表性蛋白质主链：①两条链之间的 rmsd 小于 0.2nm；②两条链间长度差异小于 10%；③两条序列对比时缺失序列要少于比对序列的 20%；④在类型划分时，要求同类型的非代表性链也至少要有 66% 残基能与所选代表性链获得比对。其他算法的比较标准不同，如选择结构域进行比对(表 10-1)。

表 10-1 常用结构比对算法和其网上资源

方法	结构特征	搜索策略	网址
DaliLite	C 原子间的距离矩阵	分枝界限法	http://www.ebi.ac.uk/DaliLite/
SSM	二级结构节点图	C 原子三维比对，同型亚结构图	http://www.ebi.ac.uk/msd-srv/ssm/
SSAP	C~C 向量和其他特征	二级动态规划算法，打分排序	http://www.cathdb.info/cgi-bin/cath/SsapServer.pl
VAST	二级结构节点	同型亚结构图	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
CE	任意八个残基的 C 原子	启发式方法，八肽片段比对	http://cl.sdsc.edu/ce.html
DEJAVU	基于二级结构向量的距离和角度矩阵	递归搜索，分枝界限法	http://portray.bmc.uu.se/cgi-bin/dejavu/scripts/dejavu.pl
FATCAT	C 原子	类似于 CE 但考虑结构的柔性	http://fatcat.burnham.org/
MATRAS	二级结构向量	分枝界限法，动态规划优化配对	http://biunit.aist-nara.ac.jp/matras/

通过结构相似蛋白质的功能可类推目标蛋白质的功能再实验验证，这是从高级结构预测蛋白质功能的主要思路。多结构比较的算法更复杂，可参考相关专著和文献。

（三）蛋白质家族分类

蛋白质结构域对于了解蛋白质的结构和功能意义重大。目前建立在结构域基础上的蛋白质家族数据库有 PROSITE、PRINTS、Pfam、SMART、SWISS、PROT、ProDom 和 BLOCKS 等。因为每个数据库都有各自的分类原则和积分标准，将它们结合起来可以更准确地归类蛋白质家族和描绘结构域。随之出现了 InterPro 数据库，它是将蛋白质的结构域和功能位点加以统一而建立的数据库资源。InterPro 联合 PROSITE、PRINTS、Pfam 和 ProDom 四个独立完整的蛋白质结构域数据库组成站点，截止到 2009 年 12 月 16 日，共收录了 18 349 个条目，再现了 5149 个结构域、11 082 个蛋白质家族等信息。此外，PDB、SCOP、CATH、HOMSTRAD、CAMPASS 等蛋白质结构数据库运用不同的原理来识别结构相似的蛋白质超家族。蛋白质的结构域在进化过程中比序列保守，一些通过核苷酸序列识别不到的蛋白质超家族在这些数据库中可以被用户检索查询得到(表 10-2)。

表 10-2 常用的蛋白质结构域查询网址

数据库	网址
PROSITE	http://www.expasy.ch/prosite/
BLOCKS	http://blocks.fhcrc.org/
Pfam	http://pfam.sanger.ac.uk/
ProDOM	http://prodom.prabi.fr/
SMART	http://smart.embl-heidelberg.de/
InterPro	http://www.ebi.ac.uk/interpro/
SBASE	http://www.icgeb.trieste.it/sbase
PRINT	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html

四、蛋白质高级结构的解析方法

目前蛋白质结构分析主要有蛋白质晶体 X- 衍射分析、磁共振波谱和冷冻电镜技术等。此外，高通量的酵母双杂交系统和免疫共沉淀技术可用于分析蛋白质间的相互作用，这也是结构生物信息学中分析蛋白质四级结构状态的有力工具，相关的信息和细节可参考对应文献。

（一）蛋白质晶体结构 X- 衍射分析

X- 射线晶体分析法(X-ray diffraction crystallography)是解析生物大分子结构的基本方法，也是

目前分辨率最高的方法，已用于解析了大量蛋白质的三维结构。该法需要将待分析的蛋白质形成晶体，所用蛋白质样品量很大，故常将该蛋白质的基因克隆到表达载体在特定宿主细胞（如大肠杆菌）中诱导表达，纯化后优化条件结晶；然后将晶体进行 X 射线衍射，收集并整合相应的衍射图谱，通过复杂的计算和数据解析过程得到蛋白质中的原子坐标信息。目前，X 射线衍射测定蛋白质结构需要人工处理大量复杂的数据，在成本和效率方面还有待改进。

优化蛋白质结晶条件和快速处理晶体衍射数据是目前蛋白质晶体结构分析的两大难题；发展高通量的蛋白结晶技术和高可靠性的结构解析技术，是当前结构生物学的重要任务。高通量晶体结构解析主要涉及数据处理与分析、重原子的定位、密度修饰、分子替换、图形整合、模型加工和确认等环节。近年来，随着自动化和高通量分析技术的应用，提高了新结构被解析的速度。尽管如此，晶体衍射数据处理和分析的算法仍需进一步完善。晶体衍射数据分析的常用软件有 SOLVE 和 RESOLVE 等。随着晶体结构的运算法则和计算机科学的发展，相信新一代的自动化分析软件将进一步解决高通量结构分析的技术问题，并将适时处理各种衍射数据和加快图形整合过程。

（二）核磁共振波谱分析

自从 1985 年第一个蛋白质的空间结构由核磁共振方法确定以来，已经有几千种蛋白质的空间结构通过核磁共振（nuclear magnetic resonance，NMR）测定，至今核磁共振技术解析的蛋白质结构已达到 PDB 数据库中蛋白质总量的 13%。NMR 主要利用氢原子之间的相互作用所产生的特殊信号来分析蛋白质等生物大分子的结构，通常为一系列的可能结构。该分析过程可在溶液状态进行而得到蛋白质分子在溶液中的构象，条件更接近于蛋白质的生理状态，是研究蛋白质的折叠和构象稳定性对生理环境温度、盐浓度和 pH 值等环境条件变化敏感性的重要工具。在溶液环境中，可以观察到整个结构表面的一些松散肽链的运动性，而蛋白质的功能部位往往是在整个结构的表面。因此，NMR 是研究蛋白质与蛋白质、蛋白质与小分子配体间相互作用的动力学特征和性质的有效手段；随着 NMR 技术的发展，其在结构基因组学中的应用也将愈加广泛。

与 X- 衍射晶体分析技术相比较，NMR 技术尽管在蛋白质结构测定中限制较大，但其无需制备晶体，故 NMR 法常用于解析无法获得晶体的蛋白质或膜蛋白的结构。目前，NMR 技术主要用于解析分子量在 20kD 以下且水溶性很好但培养晶体困难的蛋白质结构。用 NMR 测定蛋白质结构的数据处理涉及许多复杂的算法。此过程中首先是将核磁共振的信号经过傅立叶变换转换为不同的峰值，然后采集各种不同的峰组成图谱，并筛选出具有特定结构特征的图谱。这些过程常用 NMRPipe 和 SPARKY 软件（http://www.cgl.ucsf.edu/home/sparky/）处理，也使用 XEASY，DYANA 和 GARANT 等软件分析侧链或骨架结构。随着 NMR 所用磁场强度的增强、计算资源的提升和分析软件的进一步发展完善，核磁共振技术在蛋白质结构解析领域的应用会越来越广。

（三）冷冻电子显微镜技术

冷冻电子显微镜（cryoelectron microscopy）技术是从 20 世纪 70 年代提出的，在 80 年代趋于成熟已成为研究生物大分子结构与功能的强有力手段。该技术大致包括样品制备、数据采集和图像处理和三维重构等环节，其确定三维结构的方法主要有电子晶体学方法、单粒子重构法和电子断层成像技术。这种方法采用高压快速液氮冷冻方法使样品包埋在玻璃态的水环境中，这种环境也接近于生理状态，减少了样品在制备过程中的结构破坏，以便观察生物大分子在天然状态下的结构；同时冷冻的速度极快，这就有可能把细胞在其生理活动（例如，肌肉收缩）的某些特定时刻固定下来，并进而显示此时的结构特点，进而可通过不同功能状态的瞬时构象变化来研究生物分子的功能。故冷冻电镜获得的是处于天然状态下未经染色的分子的二维投影像。将样品进行不同角度的倾斜所获得的数据进行综合分析，并依据样品的不同特性使用不同的重构技术获得分子的结构，在此基础上观察多种成分的图像变化，可追踪生物大分子的装配及其动力学过程。

冷冻电子显微镜技术主要用于蛋白质及其复合物的外部形貌观察，可用不同的方法对均一的（如膜蛋白的二维晶体、二十面体对称的病毒等对称结构）和不均一的（如核糖体等）样品进行三维

结构重构，同时可应用的蛋白质分子大小范围很宽。冷冻电子显微镜技术观察生物大分子的空间构象需要借助生物信息学方法、模式识别(pattern recognition)、数据库分析和同源建模(homology modeling)等技术的整合。由冷冻电镜技术所获得的蛋白质三维结构与X射线晶体技术非常相似，而且其信噪比非常低，并适合于膜蛋白的分析。此技术目前应用面并不太广，也没有形成相应的数据库。各种相关技术的发展和整合将能提供研究生命现象与本质的强有力的技术手段。

五、蛋白质结构的可视化

目前已有蛋白质高级结构数据存储的通用格式和数据库，可通过软件将蛋白质高级结构可视化，这些资源是蛋白质高级结构信息分析的关键基础之一。可视化分析蛋白质的高级结构有利于从原子间相互作用的层次理解生命活动过程的信息控制机制，理解蛋白质分子结构和各种微观性质与宏观性质之间的关系。

(一) 常用蛋白质分子图形系统

目前，蛋白质分子图形学软件已很普及。蛋白质结构数据可从蛋白质数据库中直接获得(详见第三节)，只要安装蛋白质分子图形学软件，配以免费的小分子图形设计系统(如ACD FREE)或商业软件，就可开展结构生物信息学的探索性工作。

这里着重介绍蛋白质三维图形相关的免费软件Pymol的基本应用。

1. 软件安装、启动和教程　Pymol可在http://www.pymol.org/寻找链接下载，该软件目前持续更新和升级，其主程序是免费的。下载后Pymol的安装与其他Windows系统下软件的安装相同。

Pymol启动后显示双界面，对分子操作的常用命令及按钮都集成在一个图形显示界面，但文件读入、背景设置、操作转变、图像输出和特征分析等功能主要集中在另一个不显示分子图形且使用下拉菜单的界面，并带有命令行操作模式，关闭任意窗口则程序关闭。图形界面左上侧列出主要的可操作对象并分成几个层次，包括所选对象、蛋白质和整体等；每个层次的对象有五种主要操作：动作(A：action)、显示(S：Show)、隐藏(H：hide)、标记(L：Label)和上色(C：Color)。Dispaly下拉菜单中可设置背景(论文中这类图一般用白色背景，而报告中常用黑色背景增加视觉效果)，Wizard中有对分子常用性质测定模块，包括距离、电荷等以及尝试进行蛋白质分子改造的功能。需要仔细阅读每个下拉菜单包含的功能才有利于发挥该软件的作用。可先读教程文件进行学习(图10-8)。

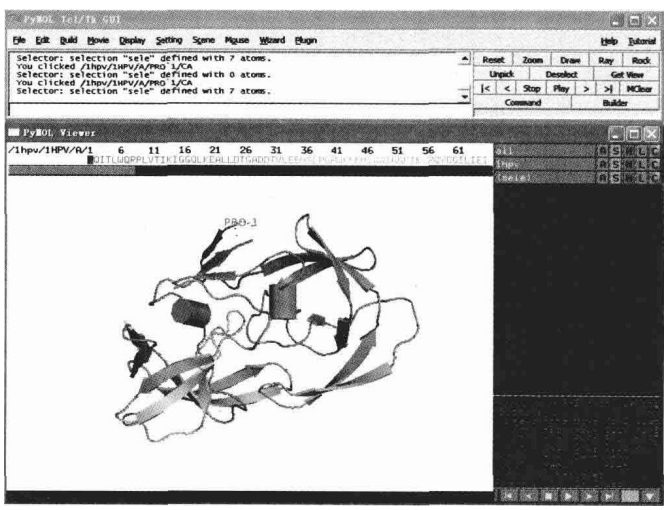


图10-8　Pymol启动后的两个操作界面(上下两个窗口)，随后读入教程所用结构

2. 主要的分子图形操作和性质测定　鼠标是主要的图形操作工具，左键旋转，右键调整大小，也可在另一个窗口的下拉菜单中选择放大缩小；可设置鼠标的模式(两键与三键鼠标等，见Mouse下拉菜单)。可显示蛋白质中每条肽链的序列和非蛋白质成分(单击图形界面右下角字母S或在Display下拉菜单中选择)；鼠标左键单击序列选中特殊待操作的残基可同时显示对象所在位置(图10-8)。在Wizard中有多种性质测定功能可灵活使用。

Pymol是强大的分子图形显示和基本特征测定系统，在带有专业显卡的计算机上输出图形更绚丽。但Pymol对非英文文件名和长文件名支持不够。Pymol自带二级结构定义词典，但对α螺旋的定义不严格，有时会给出一些不尽合理的α螺旋；不过有些商业软件不能识别同源建模所得蛋白质

中的二级结构而 Pymol 可以识别这些二级结构，这是 Pymol 的一个优势。

（二）集成的分子模拟与分析的图形学系统

集成结构生物信息学、分子操作绝大部分功能和 MD 模拟轨迹分析等功能的商业软件已面市，如 Insight II、Discover Studio 和 Sybyl 等；这些商业图形操作界面系统价格不菲，集成在图形界面里进行分子模拟、分子对接和分子改造等操作，并有各种高质量的图形显示，对应用研究人员可事半功倍。

（三）其他的蛋白质可视化软件介绍

还有很多界面友好的蛋白质结构可视化软件和在线服务器，如 RasMol 和 Jmol 等，已与 PDB 数据库链接；及 Cn3D、Mage、KiNG 等可视化软件（表 10-3）。

表 10-3 目前常用的蛋白质可视化软件

软件名称	主要功能	下载地址
RasMol	直观再现生物分子 3D 微观立体结构；提供可以旋转等多个模式效果图；提供多种结果图片存储形式；提供命令行操作，源代码开放用户可自行维护	http://www.bernstein-plus-sons.com/software/rasmol/
Jmol	以 3D 形式查看蛋白质等生物大分子化学结构，提供命令行操作，提供结构查询工具，基于网络界面可通过网址或本地文件读取结构，无需安装（Jmol 提供的功能适用于小分子，晶体，材料和生物分子）	http://jmol.sourceforge.net/
Cn3D	生物分子三维结构、序列以及序列比对结果的可视化工具；读取输入数据格式为 MMDB 格式文件，不能读取 PDB 格式文件；可紧密联系结构与序列信息，可根据基于结构的序列比较显示分子结构之间的关系；可自定义标签特征，输出结果格式多样，并可对结果进行文献注释；通过网络浏览器来作为 NCBI 的 Entrez 系统的一个辅助工具，也可作为一个独立的程序使用	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
QuickPDB	用 JAVA 编译的显示结构和序列的工具；网络浏览可直接显示序列信息，可以控制设置残基属性等；支持多种文件格式输入、可以不同形式显示三维结构	http://www.sdsc.edu/pb/Software.html
Mage	广泛应用于教学与研究中，输入为 *.kinemage 文件格式，该文件内含有蛋白质结构的各种信息与相关命令；可实时旋转效果图、并对效果图进行蛋白质结构的三维动画演示，部分图像可隐藏和显示；输出格式为 .kinemage，也可以多种其他格式输出	http://kinemage.biochem.duke.edu/software/mage.php
VMD	主要处理目标是分子动力学数据，可对生物分子进行结构分析和表征；提供多用途交互式图形界面操作；开源并提供强大脚本语言，可用于程序扩展	http://www.ks.uiuc.edu/Research/vmd/
KiNG	KiNG 即 Kinemage，是在 Mage，JavaMage 和 Kinemage 软件基础上发展起来的三维分子显示软件，可展示生物大分子结构	http://kinemage.biochem.duke.edu/software/king.php
Spdbv	即 Swiss-Pdb Viewer 或 DeepViewer。可同时分析几个蛋白质的 PDB 文件并分析结构相似性、比较活性位点或其他有关位点；可以很容易获得氢键、角度、原子距离、氨基酸突变等数据；可直接从软件链接到 Swiss-Model 服务器对蛋白质理论立体结构进行构建，并调用 POV-Ray 软件生成高质量的结构图像	http://mac.softpedia.com/get/Math-Scientific/SPDBV.shtml
WebMol	用 JAVA 语言编译的结构呈现程序，网络浏览，可从 URL 上载结构	http://www.cmpharm.ucsf.edu/~walther/webmol/download.html
Raster3D	可显示蛋白质三维结构并生成蛋白质结构的分子艺术图片（TIFF 格式与 JPG 格式）	http://www.fyxm.net/Raster3D-93918.html

第三节 蛋白质结构数据库

Section 3 Protein Structure Databases

蛋白质三维结构数据库是一类重要的生物分子信息数据库，是结构生物信息学的关键组成。随着 X-射线晶体衍射技术、NMR 和冷冻电子显微镜技术等的发展，已测定了很多蛋白质的结构；随着蛋白质结构分类研究的深入，出现了蛋白质家族、折叠模式、结构域和回环等数据库。总体而言，目前常用的蛋白质结构数据库主要是存储蛋白质结构的 PDB 数据库(protein data bank，PDB)、进行蛋白质结构比较的 SCOP 和 CATH，及存储次级结构的 targetDB、FSSP 和 DSSP 等。

一、蛋白质三维结构数据库

PDB 是用于保存生物大分子结构数据的常用档案库，由美国 Brookhaven 国家实验室于 1971 年创建的。1998 年 10 月为适应结构基因组和生物信息学研究的需要，由美国国家科学基金委员会、能源部和卫生研究院资助成立了结构生物学合作研究协会(Research Collaboratory for Structural Bioinformat-ics，RCSB)。之后，PDB 数据库的维护主要是由该组织负责，目前主要成员为拉特格斯大学(Rutgers University)、圣地亚哥超级计算中心(San Diego Supercomputer Center，SDSC)和国家标准化研究所(National Institutes of Standards and Technology，NIST)。PDB 数据库网站主页见图 10-9。

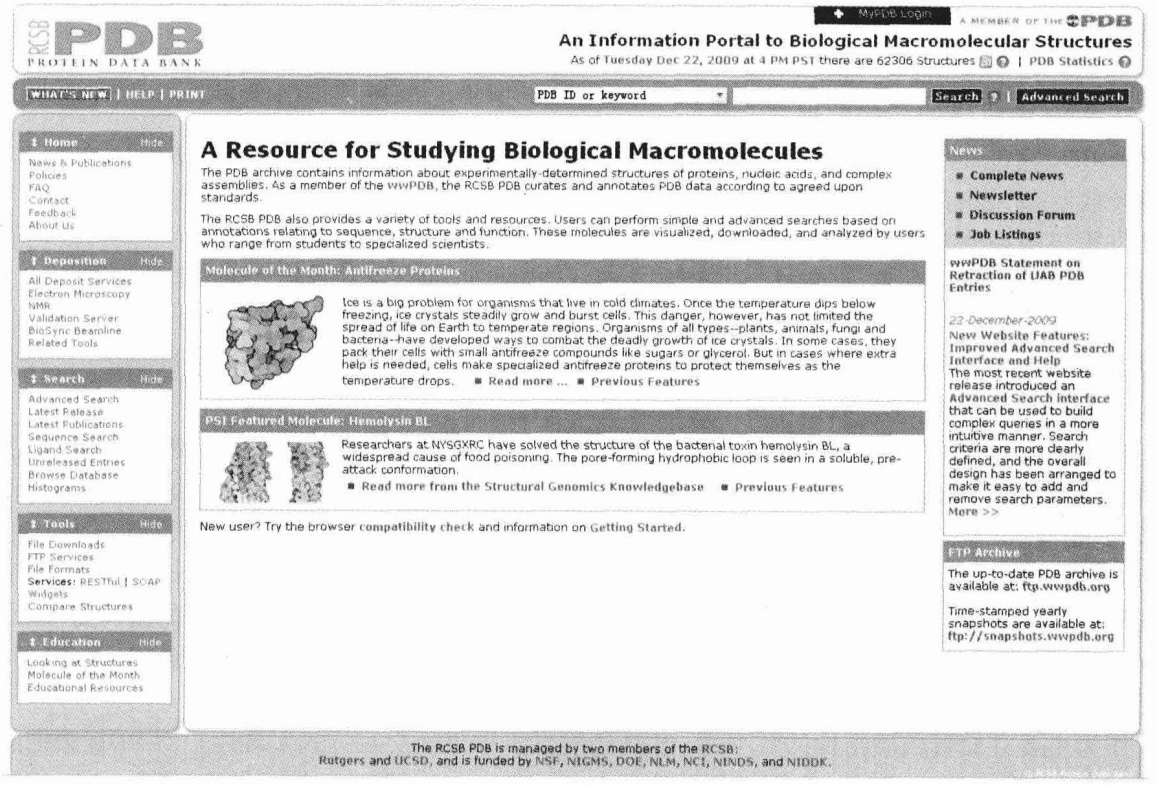


图 10-9 PDB(Protein Data Bank)数据库网站主页

PDB 中包含了通过 X 射线单晶衍射、磁共振和电子衍射等实验手段确定的蛋白质、多糖和核酸等生物大分子的三维结构数据。最初 PDB 中只含有七个生物大分子的结构，之后随着结构测定方法的成熟以及人们对数据共享观点的改变，PDB 库中的数据量迅速增加(图 10-10)。目前 PDB 库的信息是每周进行更新，截止到 2009 年 9 月 8 日，PDB 总共收录了 60 046 条结构数据，其中，收录蛋白质结构为 55 519 条，收录核酸 2058 条。详细数据见表 10-4。

表 10-4 PDB 数据库收录条目一览表

实验方法	分子类型				总数
	蛋白质	核酸	蛋白质 / 核酸复合物	其他	
X- 射线衍射	48 225	1168	2216	17	51 626
NMR	6993	869	150	6	8018
电镜	171	16	65	0	252
其他	130	5	5	10	150
总数	55 519	2058	2436	33	60 046

PDB 数据库以文本文件的方式存放数据，每个分子各用一个独立的文件。除了原子坐标外，还包括物种来源、化合物名称、结构以及有关文献等基本注释信息。此外，还给出分辨率、结构因子、温度系数、蛋白质主链数目、配体分子式、金属离子、二级结构信息、二硫键位置等和结构有关的数据。除了能用文本编辑的方式查看这些数据外，还可以利用一些图形软件直观观察蛋白质的三维结构，例如 VMD、Jmol、Swiss-PDBviewer 及 RasMol 等。

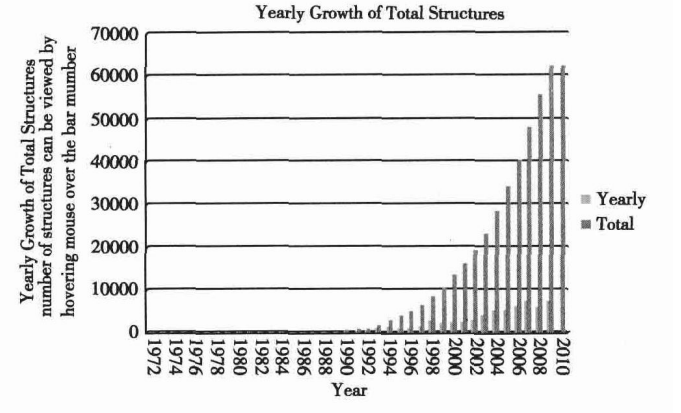


图 10-10 PDB 数据库规模在过去 28 年中的变化

在 PDB 中收集的结构数据都有唯一的 PDB-ID，它包含 4 个字符，由大写字母和数字组成(如血红蛋白的 PDB-ID 为 4HHB)。PDB-ID 编码系统较复杂，没有特征明显的顺序，但相关结构数据记录的 PDB-ID 仍然有明显联系。PDB 数据库允许用户用各种方式以及布尔逻辑组合(AND、OR 和 NOT)进行检索，可检索的字段包括功能类别、PDB 代码、名称、作者、空间群、分辨率、来源、入库时间、分子式、参考文献和生物来源等项。用户不仅可以得到生物大分子的各种注释、坐标和三维图形，并能链接到一系列与 PDB 相关的数据库，包括 SCOP、CATH、Medline、ENZYME 和 SWISS-3DIMAGE 等。下面以人类泪液载脂蛋白为例(图 10-11)，具体介绍其在 PDB 数据库中结构检索和可视化过程。

利用搜索关键字 "HUMAN TEAR LIPOCALIN" 在 PDB 数据库主页搜索框内进行搜索；点击查询结果页面的一个检索条目 1XKI，打开其链接页面；在结果页面右侧列表信息中查看生物结构信息面板 "Biological Assembly" 部分；点击 "Biological Assembly" 面板查看 1XKI 结构图(需要 JavaScript 插件)；在结果页面中，可查看提供该蛋白质结构的作者信息(Deposition Summary)及实验细节信息(Experimental Details，包括分辨率 resolution、空间群 space group 和近体的单位晶胞尺度 unit cell dimension 等)。另外，还可以链接到其他一些浏览结构信息的可视化工具如 Jmol 和 Kiosk 等进行精细结构的观察和分析。

作为主要存储蛋白质结构的数据库，PDB 还提供多种界面交互方式实现用户对 PDB 数据的浏览，可通过三种查询方式对其主要服务器站点 SDSC、Rutgers、NIST 和其镜像网站(表 10-5)进行查询，也可进行相应数据的下载操作。数据库的查询方式见表 10-6：①1999 年 2 月建立的 SearchLite 是一个关键词检索工具，在该界面的对话框内键入与生物大分子相关的关键词，点 "Search" 或者回车键即可，如键入 "protein kinase"，则可以查询所有包含蛋白激酶的结构。PDB 中所有原文资料、存储和发布日期以及一些实验数据可以通过简单的浏览或结构浏览得到；②SearchFields 是 1999 年 5 月建立的一个惯用浏览方式，可以用化合物、作者引用、序列(通过 FASTA 搜索)、存储日期或发布日期来查询。当用 SearchLite 或 SearchFields 浏览时，在 "Query Result Brower" 的界面可得到一些综

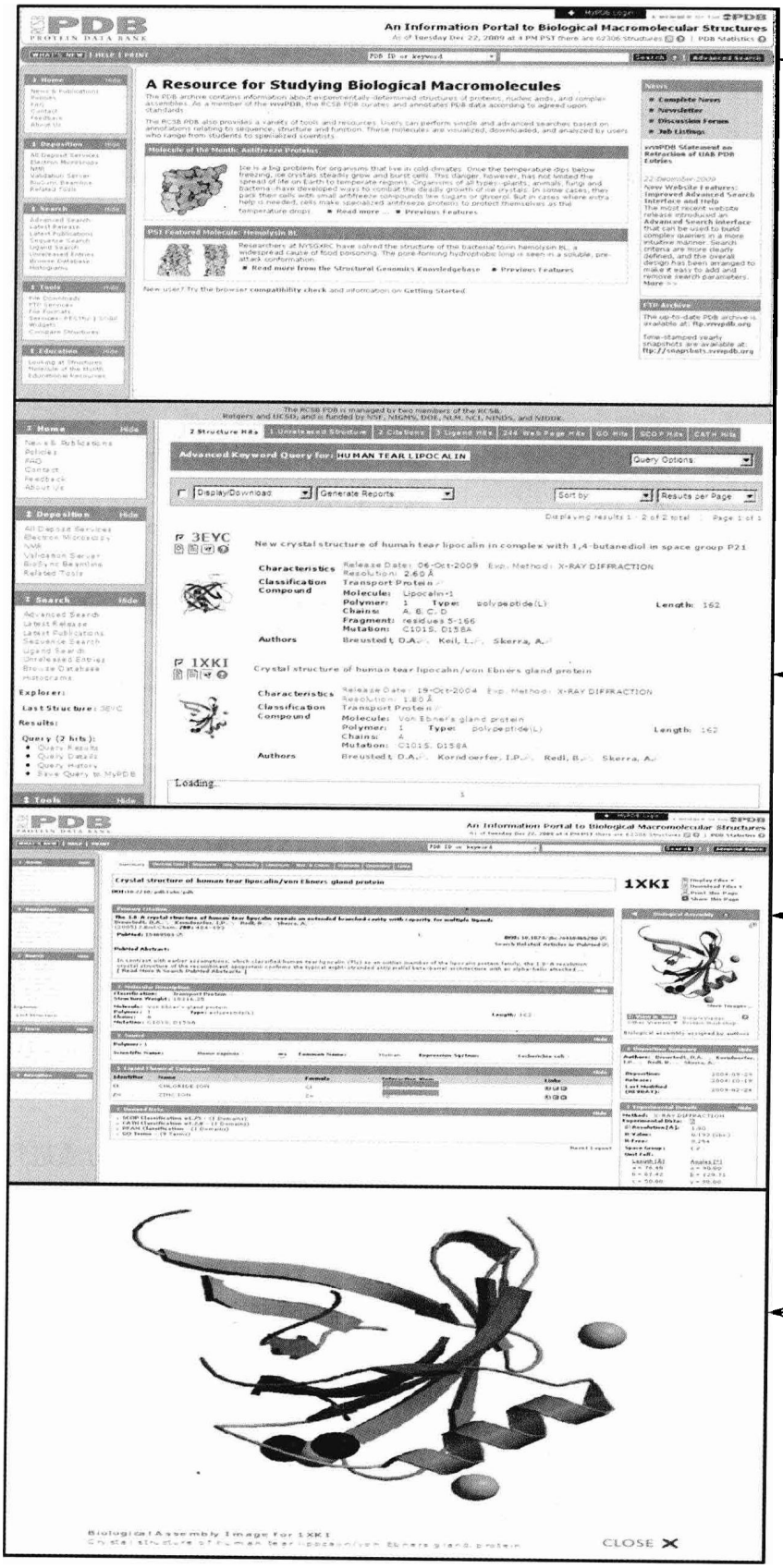


图 10-11 在 PDB 数据库中搜索人类泪液载脂蛋白结构的结果

合信息及图表中的详细信息，并可下载 PDB 中系列数据文件，下载的数据以纯文本格式或压缩文件的形式保存。“Struture Explorer”界面提供每个蛋白质结构的信息以及与许多大分子结构数据库的交叉链接。

表 10-5 PDB 的主要镜像点

PDB 的主要镜像点	镜像所在国家	网址及 ftp 地址
RCSB 成员站点		
圣地亚哥超级计算中心	美国	http://www.pdb.org/ ftp://ftp.rcsb.org/
拉特格斯大学	美国	http://rutgers.rcsb.org/
国家标准化研究所	美国	http://nist.rcsb.org/
其他 RCSB 镜像点		
剑桥晶体数据中心	英国	http://pdb.ccdc.cam.ac.uk/ ftp://pdb.ccdc.cam.ac.uk/rcsb/
新加坡国立大学	新加坡	http://pdb.bic.nus.edu.sg/ ftp://pdb.bic.nus.edu.sg/pub.pdb/
大阪大学	日本	http://pdb.protein.osaka-u.ac.jp/ ftp://pdb.protein.osaka-u.ac.jp/
米纳斯联邦大学	巴西	http://www.pdb.ufmg.br/ ftp://vega.cenapad.ufmg.br/pub/pdb/

表 10-6 数据库查询方式

查询方式	使用方法
SearchLite	PDB 所包含的任意词或词组
SearchFields	1. 一般信息：PDB 编码，作者已用，链型（蛋白质、DNA 等），PDB HEADER，试验方法，存储或发布日期，复合物资料，BC 数字或上下文检索 2. 序列或二级结构：链长，FASTA 检索，短序列方式和二级结构内容检索 3. 晶体试验信息：溶剂，空间基团，单体相关参数
Status	PDB 编码，存储信息作者，题目，存储日期或发布日期

二、蛋白质结构分类数据库

蛋白质结构分类数据库（structural classification of protein，SCOP）是对已知结构蛋白分质进行分类的数据库（图 10-12），其根据不同蛋白质的氨基酸组成及三级结构的相似性，详细描述已知结构蛋白间的功能及进化关系。SCOP 数据库的构建除了使用计算机程序外，主要依赖于人工验证。SCOP 数据库建立于 1994 年，数据库中信息主要由 Alexdi G Murzin 和其同事每年更新。

目前 SCOP 数据库的最新版本是 2009 年 2 月 23 日发布的 1.75 版，在该版本中共含有 38 221 个已有结构的蛋白质以及 110 800 个蛋白质结构域，表 10-7 为 SCOP 数据库最新版本中详细的信息统计。

在 SCOP 数据库中，按照从简单到复杂的顺序对蛋白质进行分类，分类基于四个层次，位于分类层次顶部的是类（Class），之后依次为家族（Family），超家族（supper family）、折叠子（Fold）、蛋白质结构域（protein domain）、单个 PDB 蛋白质结构记录。SCOP 数据库可以通过其分级结构导航进行浏览，用关键字、PDB 标志码查询，或通过一个蛋白质序列进行同源搜索。在各个分类层次中，家族用来描述相近的蛋白质进化关系；超家族用来描述远源的进化关系；折叠子用来描述空间的几何关系。在 SCOP 数据库中结构域又被分为以下几类：全 α 螺旋，全 β 折叠，α 螺旋和 β 折叠，α 螺旋 +β 折叠以及复合结构域。除此之外，SCOP 提供一个非冗余的 ASTRAIL 序列库，这个库通常被用来评估各

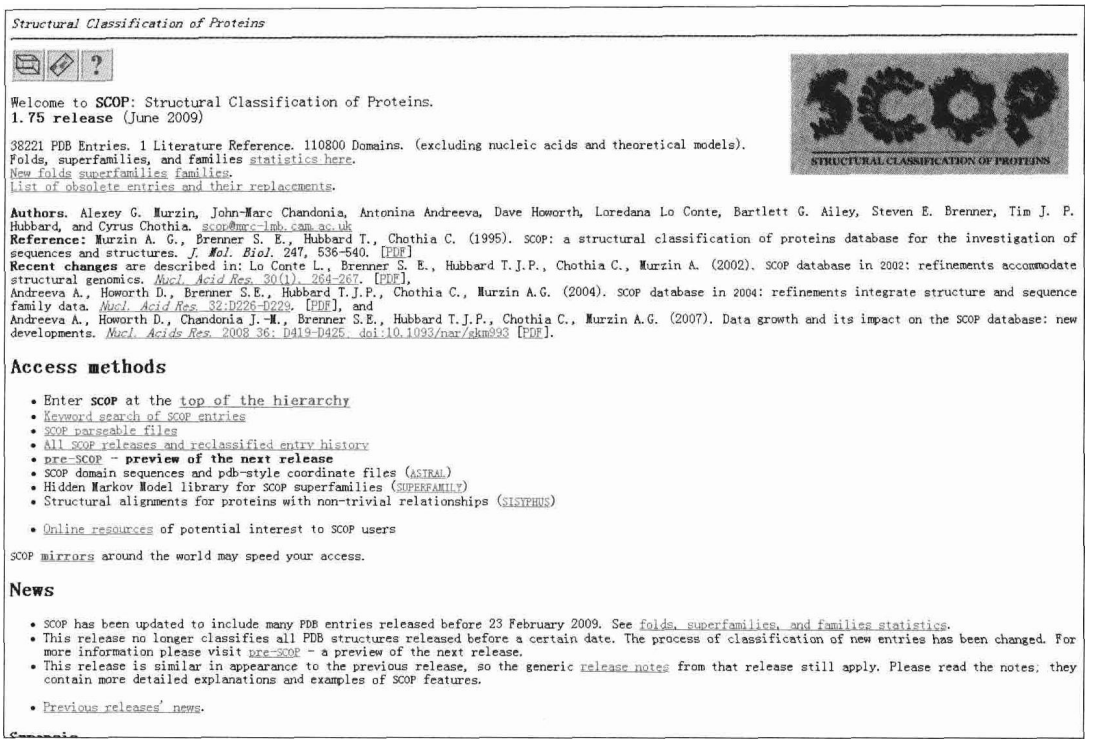


图 10-12 SCOP 数据库主页

表 10-7 SCOP 数据库中 1.75 版本中详细信息

蛋白质种类(Class)	折叠子数目(Folds)	超家族数目(Superfamilies)	家族数目(Families)
全 α 螺旋蛋白	284	507	871
全 β 折叠蛋白	174	354	742
α 螺旋和 β 折叠	147	244	803
α 螺旋 +β 折叠	376	552	1055
复合结构域蛋白	66	66	89
膜蛋白	58	110	123
小蛋白	90	129	219
总和	1195	1962	3902

种序列比对算法；同时 SCOP 还提供一个 PDB-ISL 中介序列库，通过与这个库中序列的两两比对，可找到与未知结构序列远源的已知结构序列。除了显示蛋白质结构与进化的信息外，SCOP 数据库通常可以链接到 PDB、SP3D 和 NCBI Entrez 等数据库来显示原子坐标，蛋白质序列及同源蛋白信息。SCOP 对多方用户都具有广泛的用途，全世界不同地区具有其相应的镜像站点，见表 10-8。探究所研究的蛋白质相近的结构空间区域时，蛋白质的分类层次有助于对蛋白质进行定位，而且数据库提供的交叉链接，方便对预测结果进行生物学解释。

表 10-8 全世界不同地区的 SCOP 镜像站点

地区	位置	URL
UK	SCOP home server in Cambridge	http://scop.mrc-lmb.cam.ac.uk/scop/
China	Peking University	http://mal.ipc.pku.edu.cn/scop/
Taiwan	National Tsing Hua University	http://scop.life.nthu.edu.tw/

续表

地区	位置	URL
USA	University of California, Berkeley	http://scop.berkeley.edu/
Russia	Institute of Protein Research	http://scop.protres.ru/
Japan	Biomolecular Engineering Research Institute (BERI)	http://beri.co.jp/scop
Israel	Wizmann Institute	http://pdb.weizmann.ac.il/scop/
Italy	Center of Biomedical Engineering, Politecnico of Turin	http://loki.polito.it/scop/
Australia	Walter and Eliza Hall Institute Australian National Genomics Information Service (ANGIS)	http://scop.wehi.edu.au/scop
Singapore	National University of Singapore Madurai Kamaraj University Indian Institute of Science Center for DNA Finger Printing and Diagnostics (CDFD) University of Pune	http://gene.tn.nic.in/scop/ http://scop.physics.iisc.ernet.in/scop/ http://www.cdfd.org.in:5555/scop/

三、蛋白质分类数据库

另一个代表性蛋白质结构分类数据库是由伦敦大学于 1993 年开发和维护的 CATH(图 10-13)。该数据库的名称 CATH 分别是数据库中四种分类类别的首字母，即蛋白质的种类(class，C)，蛋白质中二级结构的构架(architecture，A)；蛋白质的拓扑结构(topology，T)和蛋白质同源超家族(homologous superfamily，H)。SCOP 注重从蛋白质进化角度进行分类，而 CATH 偏重于从结构角度对蛋白质分类，同时数据库对蛋白质进行分类时既使用计算机程序，也进行人工检查。

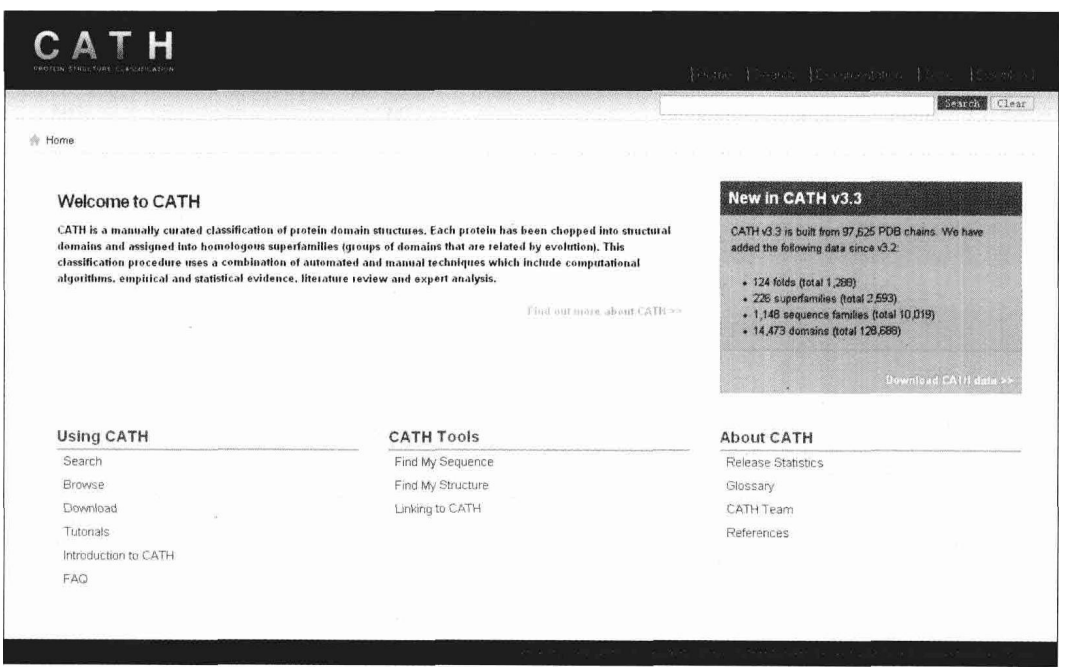


图 10-13　CATH 数据库主页

目前 CATH 数据库最新版本是 2008 年发布的 3.2 版，该版本中含有 114215 个蛋白质结构域，40 个二级结构构架，1110 个拓扑结构以及 2178 个同源蛋白质超家族。同 PDB 蛋白质结构数据库相似，每一个蛋白质都会有一个不重复的标号，在 CATH 数据库中表现为不同分类层次都将有 CATH 号，并且不同水平的 CATH 号的标准不同。例如位于 CATH 数据库最底层的分类的蛋白种类类别，

它的CATH号的范围为1~4，每一类之间的间隔量为1；而其他类别之间的间隔量为10。图10-14是CATH数据库中各个类别的层次划分结构。与SCOP不同的是，CATH把蛋白质分为四类，即全α、全β、α-β(α/β型和α+β型)和低二级结构类。低二级结构类是指二级结构成分含量很低的蛋白质分子。CATH数据库的第二个层次为由α螺旋和β折叠形成的超二级结构排列方式，而不考虑它们之间的连接关系。形象地说来，就是蛋白质分子的构架，如同建筑物的立柱和横梁等主要部件，这一层次的分类主要依靠人工方法。第三个层次为拓扑结构，即二级结构的形状和二级结构间的联系。第四个层次为结构的同源性，它是先通过序列比较然后再用结构比较来确定的。除了以上提到的四种分类外，CATH数据库还有另外一种分类层次为序列层次，在这一层次上，只要结构域中的序列同源性大于35%，就被认为具有高度的结构和功能的相似性，从而被划分为在同一序列家族(Sequence family)中。

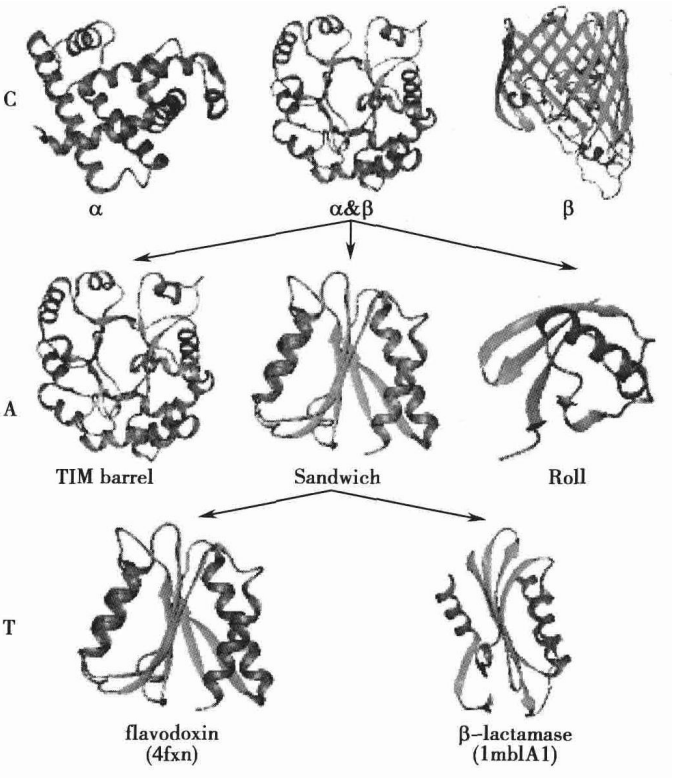


图10-14 CATH数据库中不同蛋白分类的结构图

主要包括：α，α和β，β，“TIM折叠桶”结构，“三明治”结构，“肉冻卷”结构，黄素蛋白、β-内酰胺酶结构

CATH数据库可以通过英国伦敦大学(UCL)的生物分子结构和模拟实验室的网络服务器来实现用户数据的查询和分析。在CATH首页右上角搜索框内输入待查询关键字，点击“Quick Search”查询。CATH给用户提供了满足不同需求而进行的相应的数据查询方式，具体包括：

(1) 搜索一个特定结构域信息，需要链接至“PDB code/Domain ID search”。用户搜索条目可以为CATH domain ID，CATH Chain ID或者PDB code，输入搜索条目关键字，点击首页右上角的“Quick Search”或者转到“Search CATH by ID/sequence/text”页面，利用“Search by ID/Keywords”模块进行搜索。

以蛋白质1ucr为例，搜索结果见图10-15。

通过搜索，可见该蛋白质1ucr包括两个结构域‘1ucrA00’和‘1ucrB00’，这两个结构域属于同一同源家族1.10.10.10。通过点击“CATH code”的超链接，可以获取该同源家族中其他结构域成员信息。同时，结果显示1ucr为二聚物，它的每条链都有自己特异的链标识(如1ucrA和1ucrB)。另外，用户可以通过查询结果的Pdbs表获得该查询蛋白质的PDB code、图像和功能信息。

通过进一步查询，可以获得特定蛋白质结构域更详细的信息，见图10-16。如点击上述查询结构页面domain ID为1ucrA00的超链接，CATH数据库将列出该结构域相关的序列家族、结构、序列和数据更新历史记录等结果；同时，点击查询结果页面的二维图像下面的图标，可用RasMol软件显示该结构域。

(2) 搜索与用户给定结构或功能关键字相关的信息，需要链接至“Text Search”实现文本搜索查询。用户输入的搜索关键字可以是描述功能起源的“chaperone”或结构相关的“helix”。将搜索关键字输入到搜索框，点击首页右上角的“Quick Search”按钮进行查询或者转到“Search CATH by ID/sequence/text”页面，利用“Search by ID/Keywords”模块进行搜索。以搜索关键字“lysine”为例，搜索结果见图10-17：

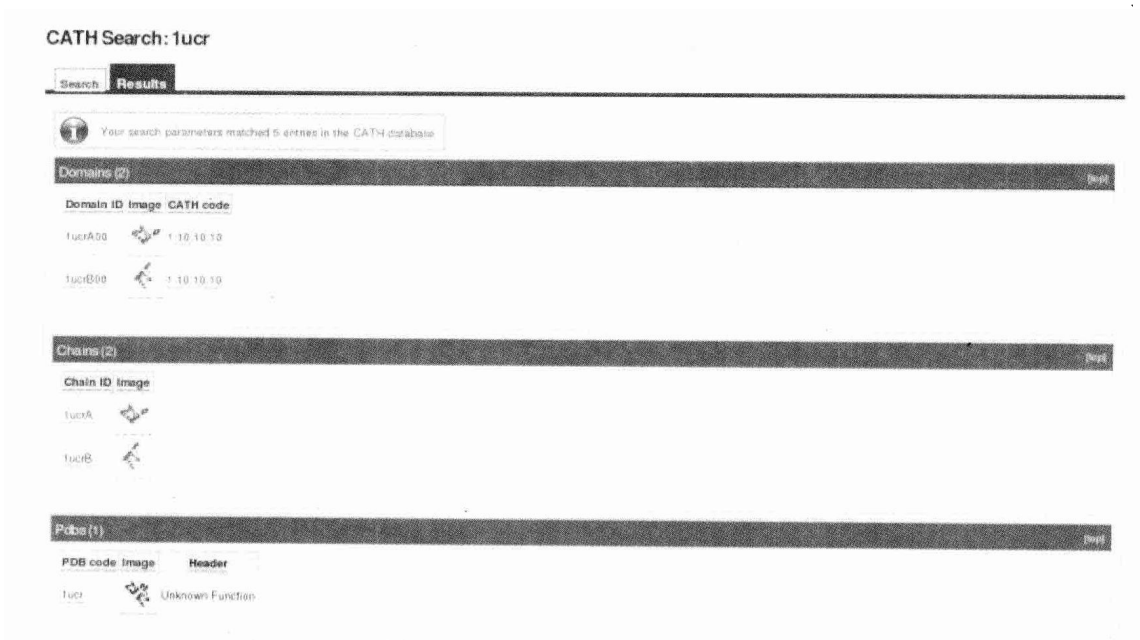


图 10-15　在 CATH 数据库中通过蛋白质 PDB 标识 1ucr 搜索其特定结构域的结果

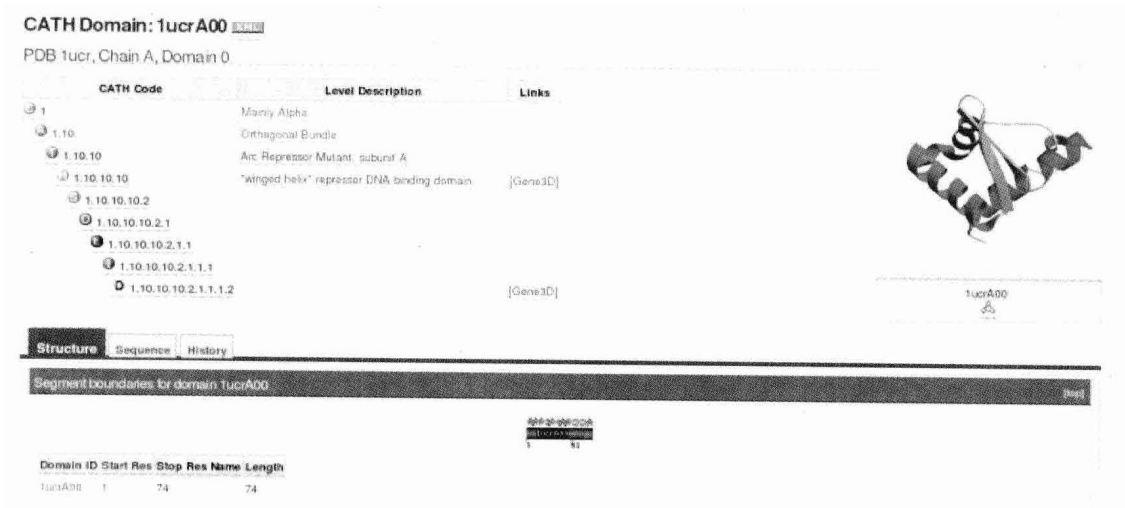


图 10-16　蛋白质 1ucr 的特定结构域 1ucrA00 的详细信息

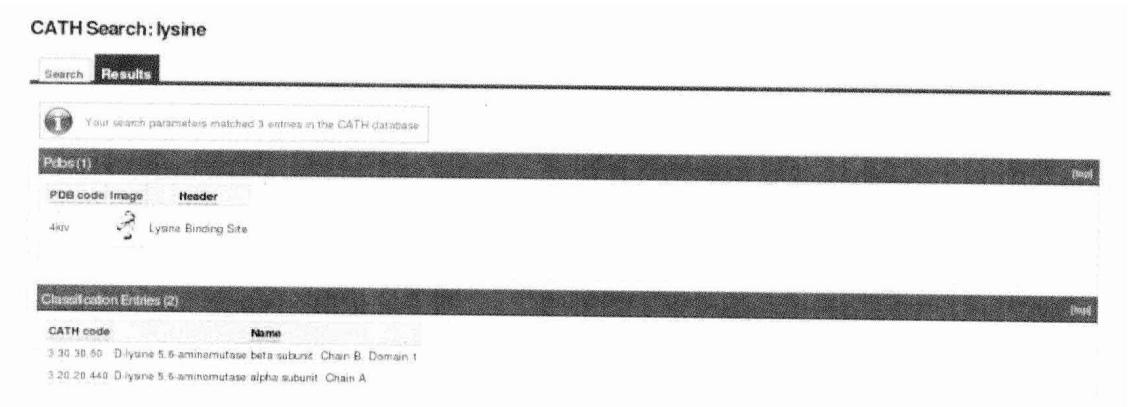


图 10-17　在 CATH 数据库中通过文本 lysine 进行数据库搜索的结果

通过搜索，用户可以检索到与查询关键字相关的 PDB 信息栏和分类信息。

（3）搜索 CATH 不同层次结构相关的信息，需要链接至“Browse the CATH hierarchy”，可查看数据库数据分类信息。也可通过“Search CATH by ID/sequence/text”页面，点击“Browse”按钮链接至“CATH hierarchy”，见图 10-18。CATH 数据库分为四个层次：全 α、全 β、α 与 β 混合以及少量的二级结构。通过选择各个类别，用户可以得到相应类别下的 CATH 编码(CATH code)、CATH 子类名称、结构域数目和相应的结构等信息，见图 10-19。

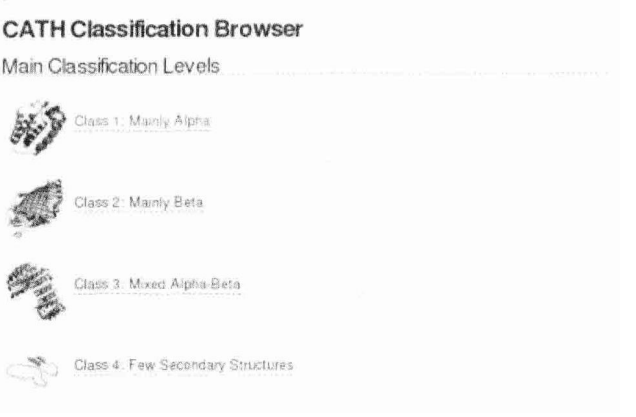


图 10-18 CATH 数据库分类查询主页

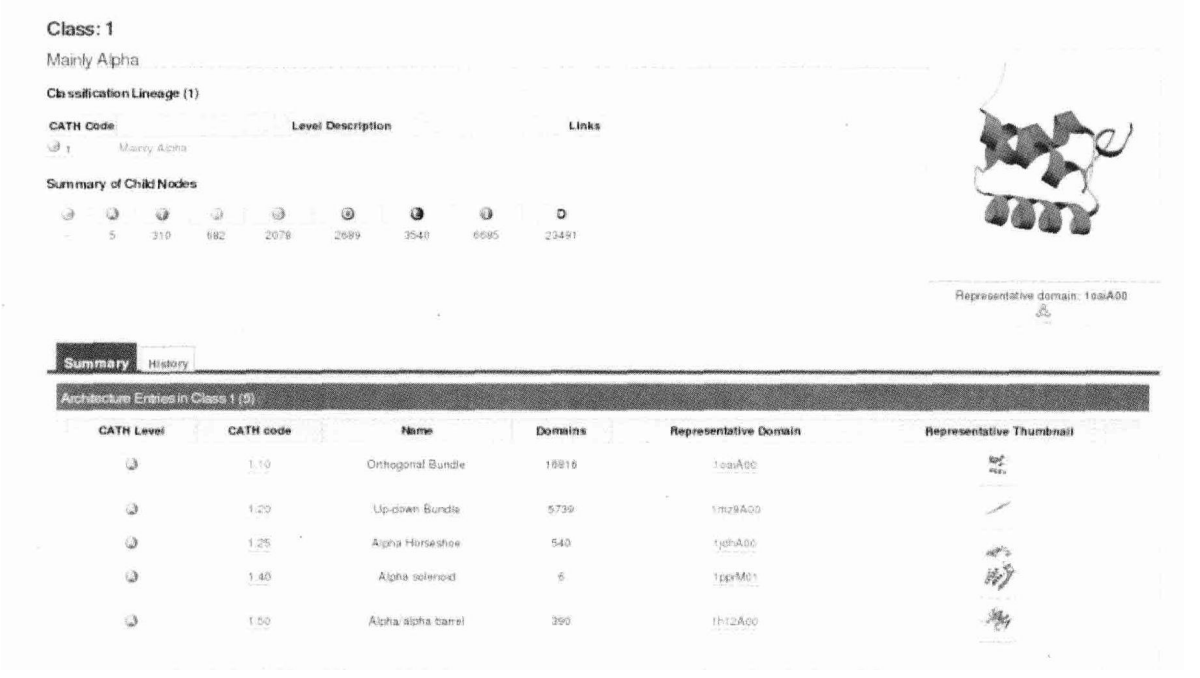


图 10-19 CATH 数据库分类等级中的全 α 类别数据情况一览图

其次，CATH 数据库还提供了分析模块，可以提交感兴趣的查询条目，CATH 数据库将提供该查询条目相关的详细的结构和相应的功能信息。

在 CATH 数据库主页上，选择“Tools”，进入“CATHEDRAL Server”分析服务器，允许用户根据 PDB ID 标识或 CATH code 编码，进行相应的结构和功能分析。如用户可以从 CATH 数据库中获悉给定蛋白质 FtsA(pdbid: 1e4f)在不同物种中的进化相关性以及与之密切相关的生物学功能信息。首先，检索 CATHEDRAL Server 服务器获取该蛋白质上所有的结构域、结构域家族和蛋白质超家族信息。需要指出的是，对于一个序列已知结构未知的蛋白质，CATH 数据库可以根据其结构比较算法将感兴趣的蛋白质与 CATH 数据库中的背景蛋白质进行相似结构搜索，最终确定出该蛋白质的结构和相应的结构域信息。然后，根据蛋白质 FtaA 结构域所在的超家族信息，用户可以获悉该蛋白质处于核酸转移酶结构域家族(CATH code: 3.30.420.40)中。其次，利用 CATH 和 Gene3D，检索 CATH 超家族 3.30.420.40 所包含的所有结构域信息。在 CATH 主页上输入结构域编码(1e4f+结构域标识 / 链

标识），检索获取相应的结构域信息；在 H-level 同源超家族层，可以找到 3.30.420.40，点击进入可以查看该同源超家族中其他已知结构域的结构信息；通过结构域链接，可以获取其结构信息和特定的分子细胞功能信息等。具体内容可登录到数据库站点查看，这里不作赘述。

通过 UCL 生物分子结构和模拟实验室的网络服务器还可以查询 PDB 数据库 PDBsum。PDBsum 数据库提供对 PDB 数据库中所有结构信息的总结和分析。每个总结给出了与 PDB 库中条目相关的简要信息，如分辨率，R 因子，蛋白质主链数目，配体，金属离子，二级结构，折叠图和配体相互作用等，还提供了获取一维序列、二维序列模体和三维结构信息的整合界面。

四、其他蛋白质结构数据库

（一）SWISS-MODEL 数据库

SWISS-MODEL 数据库收录的蛋白质结构都是使用 SWISS-MODEL 同源建模对 Swiss-Prot 蛋白质序列数据库或其他蛋白质序列进行自动同源建模所得到的结构数据，该数据库保持定期更新。建立该数据库的主要目的在于提供最新的蛋白质 3D 结构注释信息。

至 2008 年 9 月，SWISS-MODEL 数据库共收录数据 340 万条，覆盖了 UniProt 数据库中 270 万个不同的蛋白质序列。SWISS-MODEL 数据库允许用户对数据库中的模型质量进行评价，允许用户搜索另外一种可变模板结构（alternative template structures），用户还可以使用 SWISS-MODEL 工作平台（http://swissmodel.expasy.org/workspace/）构建蛋白质的三维模型。最后对结构模型的注释信息即包括功能信息可通过与其他数据库进行交叉链接，通过这些链接，用户就可以在蛋白质序列数据库和结构数据库之间自由切换。

（二）生物磁共振数据库（BMRB）

生物磁共振数据库由美国威斯康星大学麦迪逊分校组织构建的专门用于存放蛋白质、多肽、核酸等物质磁共振 NMR 波谱数据，以及对应的分子研究的源数据、研究所使用的实验条件和设备、与研究相关的重要出版物等信息。

随着测序技术和预测方法不断发展，涌现了很多蛋白质结构相关的数据库。这些数据库存储蛋白质序列、分类、家族、二级或三级结构、膜蛋白、结构域以及结构修饰等信息（表 10-9）。

表 10-9 常用蛋白序列和结构数据库

数据库	说明	网址链接
PDB	蛋白质三维结构	http://www.rcsb.org/pdb
READ	蛋白质结构修饰数据库	http://pir.georgetown.edu/cgi-bin/resid
中国蛋白质结构数据库	中国蛋白质结构数据库	http://lifecenter.sgst.cn/cnpdb/cn/pdbHome.do
BMRB	生物磁共振数据库	http://www.bmrb.wisc.edu/
SWISS-PROT	蛋白质序列数据库	http://kr.expasy.org/sprot/
PIR	蛋白质序列数据库	http://pir.georgetown.edu/
OWL	非冗余蛋白质序列	http://www.bioinf.man.ac.uk/dbbrowser/OWL/
EMBL	核酸序列数据库	http://www.embl-heidelberg.de/
TrEMBL	EMBL 的翻译数据库	http://kr.expasy.org/sprot/
GenBANK	核酸序列数据库	http://www.ncbi.nih.gov/Genbank/
PROSITE	蛋白质功能位点	http://kr.expasy.org/prosite/
SWISS-MODEL	从序列模建结构	http//www.expasy.org/swissmod/SWISS-MODEL.html
SWISS-3DIMAGE	三维结构图示	http://us.expasy.org/sw3d/
DSSP	蛋白质二级结构参数	http://www.cmbi.kun.nl/gv/dssp/
FSSP	已知空间结构的蛋白质家族	http://www.bioinfo.biocenter.helsinki.fi

续表

数据库	说明	网址链接
SCOP	蛋白质分类数据库	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH	蛋白质分类数据库	http://www.biochem.ucl.ac.uk/bsm/cath/
Pfam	蛋白质家族和结构域	http://pfam.wustl.edu/
tmbase	跨膜蛋白数据库	ftp://ulrec3.unil.ch(/pub/tmbase)

第四节 蛋白质结构的预测

Section 4 Prediction of Protein Structure

蛋白质结构模型的确定和分析有重要的应用，如点突变作用分析、酶促反应、蛋白质复合物和活性位点的界面相互作用分析，还可以用于晶体衍射数据的定相(phasing)以及相近家族的归类、配体类药物设计和虚拟筛选等。目前，蛋白质序列数据库中多肽链序列记录的数量与结构数据库中结构记录的数量相差很大，需要发展计算机预测蛋白质结构的方法。从序列预测和指认蛋白质主链的折叠和组装模式是最有希望的蛋白质高级结构预测策略。蛋白质折叠和蛋白质结构域预测是蛋白质结构分析中非常重要的两个问题，蛋白质折叠研究的是蛋白质动态变化过程，蛋白质结构域预测研究的是蛋白质相对静态的结构，它们的本质是统一的，都是探讨蛋白质一级结构序列如何影响和决定蛋白质高级结构。

一、蛋白质二级结构预测方法及软件

(一) 蛋白质二级结构的预测方法

蛋白质二级结构是一级结构与三级结构之间的桥梁。每一段相邻的氨基酸残基都具有形成一定二级结构的倾向。蛋白质二级结构由 α 螺旋、β 片层或折叠和卷曲组成。根据 α 螺旋和 β 片层的百分比，蛋白二级结构可分为：全 α、全 β、α/β 和 α+β 四类。所有蛋白质中约 85% 的氨基酸残基处于 α 螺旋、β 折叠和转角等三种基本二级结构状态，并且各种二级结构非均匀地分布在蛋白质中。其中 α 螺旋和 β 片层被认为是规则的二级结构，卷曲是低规律性的二级结构。因此，进行二级结构预测需要通过统计和分析发现这些倾向或规律。

1. DPM(双重预测方法) 该方法先预测蛋白质的结构分类再预测序列的二级结构，分为四个步骤：①从氨基酸组分预测蛋白质结构分类；②从简单算法初步预测二级结构；③将两个独立的预测进行比对；④优化参数得到其二级结构。

2. DSC 该算法将二级结构预测分为两步，首先预测基本概念，然后利用简单且线性的统计方法结合概念预测二级结构，这种预测方法准确性较高。

3. PHDsec 是一种借助神经网络系统来预测二级结构的方法，每个结果都据局部序列间关系和整体蛋白质性质(蛋白质长度、氨基酸频率等)来预测残基的二级结构，其致力于研究氢键。该算法首先将提交的靶序列进行 BLASTP 查询 SWISS-PROT 数据库得到同源序列，将查询结果过滤后再进行 CLUSTALW 多序列比对，最后将多序列比对的结果作为神经网络计算的输入值进行计算。PHD 被认为是二级结构预测的标准。

4. SOMPA SOMPA 在蛋白质二级结构的预测方面做出了很大的改进，它从同一家族序列比对的结果中获取信息。该方法能够对氨基酸序列进行螺旋、折叠和卷曲等三种形式描述，准确性达到 69.5%。位于法国里昂的 CNRS(Centre National dela Recherche Scientifique)用五种相互独立的方法进行预测，并将结果汇集整理成一个“一致预测结果”。这 5 种方法包括 Garnier-Gibrat-Robson (GGR)、Levin 同源预测方法、DPM、PHD 和 CNRS 的 SOPMA 方法。

5. MLRC 该算法集 GOR4、SIMPA96 和 SOMPA 为一体。它首先处理了蛋白质二级结构预测的结果并形成了分类的后概率事件的估计。

6. Jpred 该方法由 Barton Group 创建于 1998 年。通过提交单一蛋白质序列或多重蛋白质序列并运行就可预测出蛋白质序列的 α 螺旋，β 折叠或无规则卷曲等三种二级结构。Jpred 运用了 Jnet 神经网络算法，准确率可达到 76.4%。

（二）蛋白质结构域识别方法

结构域对应着独立折叠的一段连续氨基酸序列，它是蛋白质结构的一个过渡层次，是蛋白质工程化设计的基本单位。目前对结构域的识别方法主要包括根据蛋白质空间结构信息利用机器学习方法获取结构域信息的方法 SSEP-Domain、通过对具有代表性三级结构的蛋白质建立隐马尔可夫模型方法、分析蛋白质序列构象熵值判定结构域边界的方法、运用神经网络从蛋白质序列获取结构域边界方法和基于经验的人工划分方法等。

下面简要介绍结构域识别的几个主要方法：

1. 通过蛋白质空间结构信息获取结构域信息 该方法是通过分析氨基酸 C-C 键的距离，将每一套蛋白质三维结构里的结构域进行测量。再通过结构域的稳定性，与折叠方面来确认蛋白质结构域的子结构。但该方法的缺点是并没有考虑到三维结构最佳分割的一般问题。

2. 运用图论法 将蛋白质看作是互相作用的残基的三维图形，这里不涉及任何共价结构，确定结构域的问题这里就变成将这个图分割成几批残基，使这几批残基之间的相互作用最小。在不知道限度与大小的情况下，先将所描述的分割程序重复 k 次，k 代表结构阈的尺寸的相关值，范围是 1 至 $n/2$，n 为蛋白质中的残基总数，经过对 k 的确认，每个最小接触区域也被记录下来，并根据 k 来设定这个区域，后通过结构域尺寸标化，寻找球状最小值。一旦球状最小值在最小接触密度图里被确认，则将这个值与给定的极限值相比较，若低于极限值则接受，不低于极限值则拒绝分割。最后一旦某一部分可以接受，对每一个产生的子结构进行这个过程的重复递归处理，直至没有缺失出现为止。

同样类似的方法是将蛋白质的三维结构看作一幅图像，以氨基酸残基为节点，把残基间的相互作用作为弧线，用网络流程策略寻找最优的分割方式，并对其进行评估。但该方法不进行判别式的分析。

3. 其他方法 另外两种方法的主要规则比较简单。首先，链上的每个氨基酸残基都被进行了数字标记（沿着序列方向的顺序氨基酸残基编号）。如果其中的一个氨基酸残基被其他的临近的残基所包围（通过使用距离约束来定义），并且这些残基的平均编号更高的话，这个残基的标记数字增加，反之下降。而后这个过程被反复应用于氨基酸链上的所有残基，并加入一些修正部分，以解决一些技术上的问题并避免非正常状态的出现：例如结构域之间链片段的过度拟合等。

（三）蛋白质二级结构预测相关软件

目前较为常用的二级结构预测软件 PSIPRED、Jpred、PREDATOR、PSA、SOMPA 等都有在线服务器；进入这些软件的主页，输入 Fasta 格式的目的蛋白序列，在网页上直接选取适合的蛋白质结构预测算法，点 submit 运行即可。

人基质金属蛋白酶 MMP14（Matrix metalloproteinase，MMP14）氨基酸序列的 fasta 形式可从 NCBI 的蛋白质数据库获得（>gi|4826834|ref|NP_004986.1|matrix metalloproteinase 14 preproprotein [Homo sapiens]）。

这里介绍 Jpred 和 SOMPA 二级结构预测的使用方法。

1. Jpred 预测二级结构

进入 Jpred 首页（http://www.compbio.dundee.ac.uk/~www-jpred/），见图 10-20；在“Sequence”下的空白处直接输入序列，点击“Make Prediction”；也可以选择“Advanced”高级模式，选择 Email 提交方式或留空为网页结果显示，输入蛋白质序列或从本地文件夹中获取，再点击“Make Prediction”；在电子邮箱中找到结果地址，在弹出的结果显示界面选择进行简单结果浏览、图形化输出等操作。

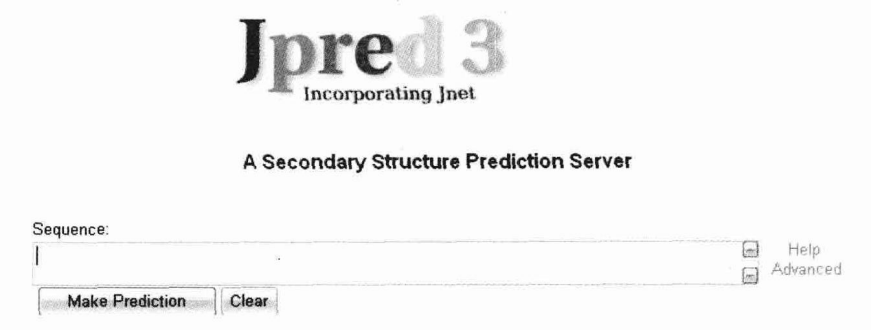


图 10-20 Jpred 首页

分析结果见图 10-21，Jpred 方法预测的 MMP14 二级结构有 8 个 α- 螺旋区(H)和 23 个 β- 折叠区(E)，其他区域均为无规则卷曲区(-)。

2. SOPMA 二级结构预测

进入 SOPMA 主页：

(http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html)；如图 10-22 所示，在“Paste a protein sequence below”下的空白处提交蛋白质序列，设置拟定的参数，点击“SUBMIT”按钮进行分析。

图 10-21 Jpred 二级结构预测

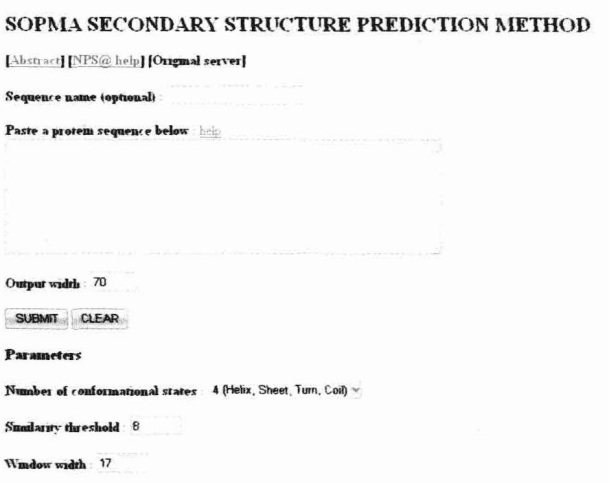


图 10-22 SOPMA 首页

结果如图 10-23，SOPMA 方法预测的二级结构主要含有 α 螺旋(h)，占 25.77%；延伸链(e)，占 18.90%；β 折叠(t)，占 8.93%；无规则卷曲(c)，占 46.39%。

二、蛋白质三维结构预测方法及软件

目前，蛋白质的结构预测主要有比较建模(comparative modeling，CM)、穿线(threading)及自由建模(free modeling)等三类。

应用比较建模时，可通过 BLAST 或 PSI-BLAST 来获得同源的模板并进行目标序列和模板序列间的比对。当缺乏与目标序列的序列相似性高的已知结构模板时，就需用复杂方法获得合适的模板并产生更确切的比对，这种过程被称为远缘同源建模(distant homology Modeling)、折叠识别(fold recognition)或穿线。不直接用已知模板的方法称为自由建模或从头建模(ab initio)。

(一) 穿线法预测蛋白质高级结构

穿线法是用于检测进化相关的序列和相似的折叠，接受与模板蛋白非常相似的结构。该方法是将目标序列与模板蛋白已解析的三维结构直接匹配，即使序列无明显的同源性，仍能识别相似的

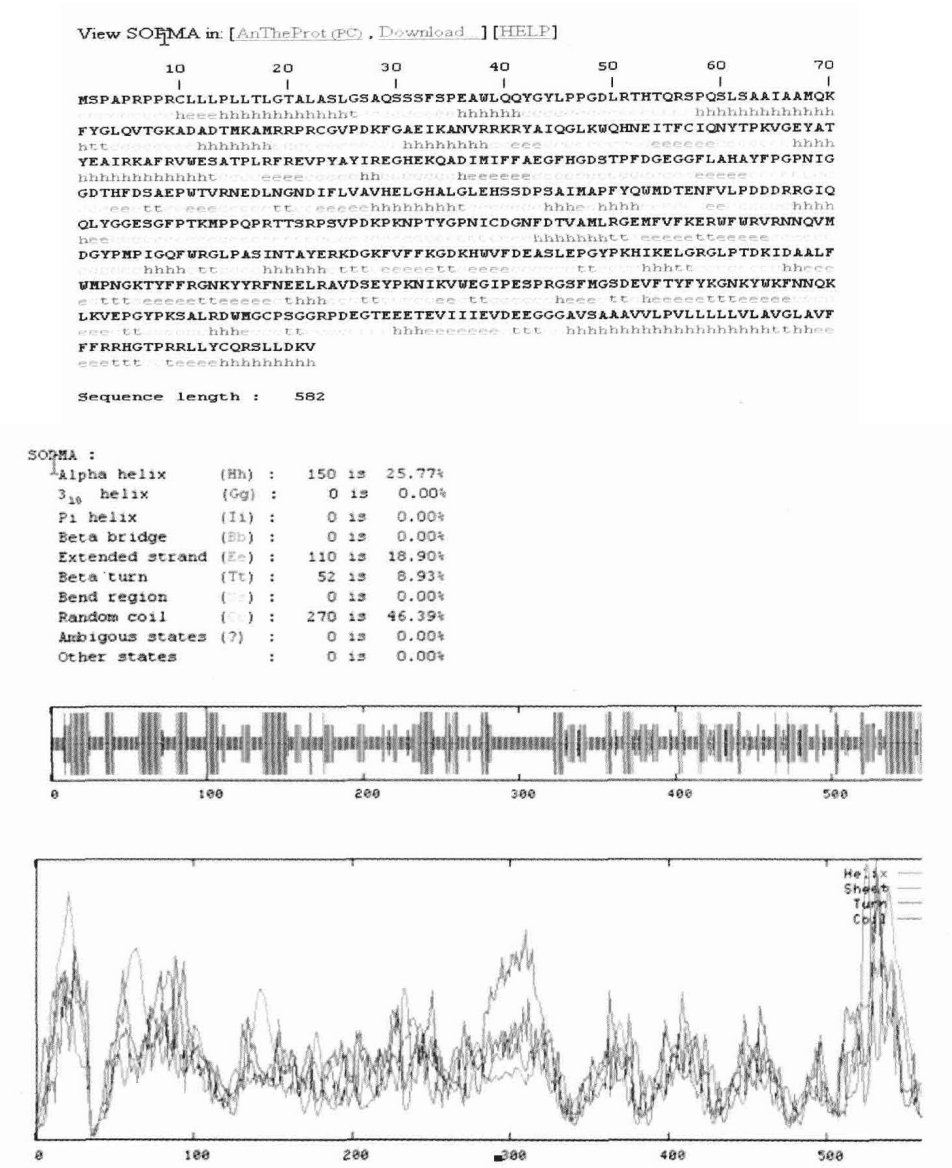


图 10-23 SOPMA 预测结果

蛋白质折叠。它是结构类似物检测的有效工具，该方法已相对成熟。通过结构的优化有利于使结构模型更加接近其天然结构。穿线法是蛋白质结构预测最活跃的领域之一，有大量的算法包括序列 profile-profile alignments(PPA)、structural profile alignments、隐马尔可夫模型和其他机器学习算法等。序列 PPA 可能是最常用和有效的穿线方法，与靶 - 模板单序列比对不同，PPA 将靶多重序列比对(multiple sequence alignment，MSA)与模板 MSA 一起比对。在 Live-Bench-8 实验中，表现较好的方法是 BASD/MASP/MBAS、SFST/STMP、FFAS03 和 ORF2/ORFS 等，基于序列 PPA 算法。

(二) 比较建模法预测蛋白质高级结构

比较建模方法：利用序列相似的已知结构来预测未知序列三级结构的主要方法，这是目前蛋白质结构预测最为成功的方法。最常见的就是利用网络直接搜索 PDB 数据库和 SWISS-MODEL 服务器。

1. 比较建模的原理　比较建模法又称为同源建模法(Homology modeling)，其原理较简单，它是基于进化相关的序列具有相似的三维结构，且进化过程中三维结构比序列保守而利用进化相关的结构模板信息建模。

2. 比较建模的基本步骤　比较建模包括以下几个步骤且常需不断重复才能获得满意的结构模型：①将目标序列作为查询序列来搜索 PDB 和 SWISS-PROT 等已知蛋白质结构数据库，确定和识

别一个同源(潜在远缘)或来自已知结构的同源物作为结构模型的模板;②将目标序列和模板序列进行比对,利用多种比对方法或手工校正以改进和优化靶序列和模板结构的比对,比对中可以加入空格;③以模板结构骨架作为模型,建立目标蛋白质骨架模型;④对侧链建模,包括构建环区(loops)和侧链,优化侧链位置,并从模板结构到目标蛋白精练整个模型;⑤优化和评估产生的模型,使用能量最小化和已知的优化知识来优化结构,如利用分子动力学、模拟退火等进行结构优化。

3. 比较建模法的局限性 传统的比较建模是通过 PSI-BLAST 找到已知结构的相关蛋白。最近如进行 profile-profile 比较和有效利用结构信息的更加复杂的方法已显著增加了比对的质量以及远程同源(remote homologue)检测的能力。因此,比较建模和折叠识别在基于模板的建模方法中的区别现已十分模糊,但在检测远程同源模板和预测新蛋白结构的技术方面至今缺乏实质性的进展。

比较建模最大的挑战是对模板链进行空隙和插入的建模。目标蛋白与模板结构保守性的程度及序列比对的正确性严重影响预测模型的准确性。与模板超过 50% 序列一致性的模型通常可靠,其 Ca 原子位置与实验结构的平均偏差约 1Å;蛋白质序列一致性在 30%~50% 时,至少可共有 80% 的结构,在该范围的最好的 CASP 模型与其本身结构 Ca 原子位置平均偏差 <4Å(典型为 2~3Å),且其误差主要在环区;当序列一致性为 20%~30% 或甚至低于 20% 时,结构保守性能低至 55%。因此,比较建模主要在大于 30% 序列一致性的序列间进行。

4. 常用比较建模服务器和软件简介

(1) SWISS-MODEL 服务器:SWISS-MODEL 是蛋白质 3D 结构自动比较建模服务器。它与 ExPASy 网站和 DeepView 程序紧密相联。可通过直接输入蛋白质序列来建模,是目前最广泛使用的基于网络的免费蛋白质 3-D 自动建模服务器(http://www.expasy.org/swissmod/SWISS-MODEL.html)。

(2) Predict Protein 服务器:Predict Protein 提供同源建模的服务,全球大约有 20 多个镜像,可直接输入蛋白质序列查询并建模。其网址:(http://cubic.bioc.columbia.edu.cn)。

(3) Accelrys Discovery Studio 软件:此系统可基于 Windows 和 PC 机进行分子建模和模拟。通过 Blast 搜索蛋白质 PDB 结构数据库,得到目标蛋白质的多个模板三维结构,然后对已知蛋白质进行同源建模和三维结构的重建,结合其他程序进行结构精修和分析,获得高质量的三维结构,并进行相关结构域和活性位点的分析。

(4) FAMS:一种新的自动同源建模软件,由数据库搜索和模拟退火两部分组成。该方法预测的三级结构类似于 X- 射线晶体衍射得到的结构,已经通过 Koji Ogata 等的实验得以证实,对保守结构区域主链和侧链原子结构的预测较好。其网址(http://physchem.pharm.kitasato-u.ac.jp/FAMS/fams.html)。

5. 用 SWISS-MODEL 预测三级结构实例 SWISS-MODEL 主要包括 Modelling、Tools、Repository 和 Documentation 四个模块。Modelling 有自动模式(Automated Mode)、比对模式(Aligment Mode)和文件模式(Project Mode)等 3 个工作模式。①自动模式:提交 FASTA 格式的目标序列,服务器自动运用 Fasta、BLAST、SIM 和 CompAli 等对结构数据库进行序列搜索和比对;如果可行,ProMod 程序将建立分子的结构模型。②比对模式(Aligment Mode):提交目标序列和模板序列比对结果(比对结果可为 Fasta、CluatalW、DeepView 等格式),将一致性比对文件(alignment file)及命令文件(command file)提交给 ProMod,可调整 SWISS-MODEL 的序阵或模板结构数。此建模过程以用户限定的目标 - 模板序列比对为基础。③文件模式(Project Mode):直接以文件提交 Fasta 格式目标蛋白的氨基酸序列。该模式采用 DeepView(Swiss-PdbViewer)集成的序列 - 结构工作站(workbench)。SWISS-MODEL 所得模型通过 email 发回,可选择在结果中提供蛋白质校验工具 WhatCheck(http://www.cmbi.kun.nl/gv/whatcheck)分析报告和原子平均力势能(atomic mean force potential)ANOLEA(http://swissmodel.expasy.org/anolea)评估报告。下面用自动方式以人的 MMP14 序列为例说明建模过程。

第一步:进入 SWISS-MODEL 三级结构预测服务器主页(图 10-24)。

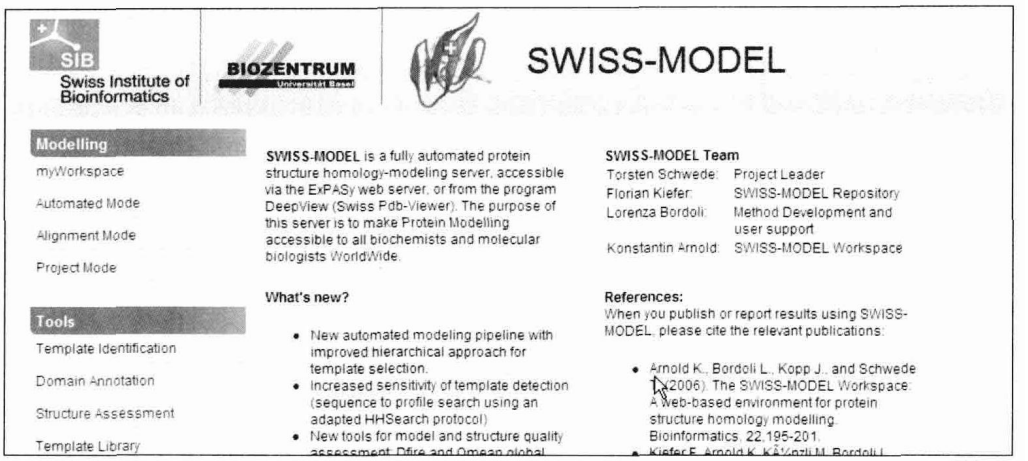


图 10-24 SWISS_MODEL 预测服务器主页

第二步：选择“Automated Mode”→粘入 MMP14 蛋白质序列；在这里可以填写 E-mail 地址，将结果发送至电子邮箱，也可以在新的网页上直接展示。

第三步：点击“Submit Modeling Request”即可。

第四步：直接在页面上查看 MMP14 蛋白质的三级结构信息(图 10-25)。

第五步：结果分析发现通过查询 Expdb 数据库，共得到 218 个可选项。选用其中相似性最高的两个模型，分别是 1bqqM 和 1su3B，从图 10-25 中可看出其三级结构含有 α 螺旋和平行 β 折叠链。从模板信息里可以得到所模拟目标蛋白的残基范围、所用模板、序列相似性及 E 值。另外，可通过展示模型获得模板的具体信息或者通过下载模板保存其三级结构的 PDB 格式。

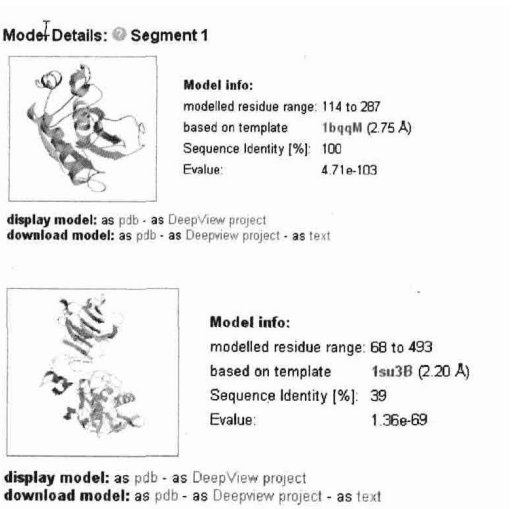


图 10-25 借助 SWISSA-MODEL 预测 MMP14 三级结构

6. 用 Discover Studio 建模简介 以人的 MMP14 序列加以实例说明 Accelrys Discovery Studio 软件的实用。

第一步：用 MMP14 序列和 Blast 搜索蛋白质结构数据库(PDB)，见图 10-26。选取序列同源性高的三维结构作为候选模板，分别是 1BQQ 的 M 链(Model-1)，1SU3 的 A 链(Model-2)和 1RM8 的 A 链，这些候选模板的三维结构如图 10-27。

Sequences producing significant alignments:		Score (Bits)	E Value
pdb|1BQQ|M	Chain M, Crystal Structure Of The Mt1-Mmp--Timp-2 ...	368	3e-102
pdb|1SU3|A	Chain A, X-Ray Structure Of Human Prommp-1: New In...	267	7e-72
pdb|1RM8|A	Chain A, Crystal Structure Of The Catalytic Domain...	251	6e-67
pdb|1FBL|A	Chain A, Structure Of Full-Length Porcine Synovial...	226	2e-59
pdb|2CLT|A	Chain A, Crystal Structure Of The Active Form (Ful...	221	5e-58
pdb|1SLM|A	Chain A, Crystal Structure Of Fibroblast Stromelys...	194	6e-50
pdb|1CK7|A	Chain A, Gelatinase A (Full-Length) >pdb|1GXD|A Ch...	189	2e-48
pdb|1RMZ|A	Chain A, Crystal Structure Of The Catalytic Domain...	169	2e-42
pdb|1Y93|A	Chain A, Crystal Structure Of The Catalytic Domain...	167	7e-42
pdb|1OS2|A	Chain A, Ternary Enzyme-Product-Inhibitor Complexe...	167	1e-41

图 10-26 在 PDB 数据库 Blast 搜索 MMP14 的结果

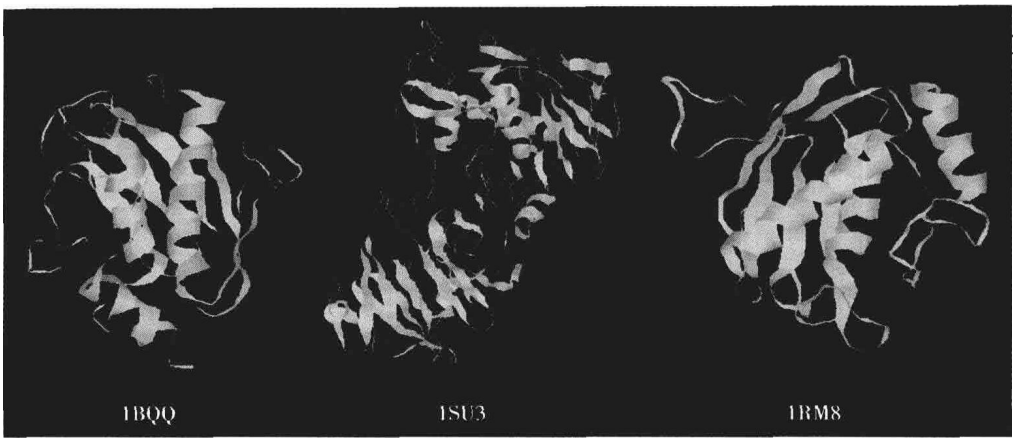


图 10-27 蛋白质 MMP14 序列相似性蛋白质(1BQQ/1SU3/1RM8)的三维结构模型

第二步：把 3 个三维结构导入到“Discovery Studio”的主界面中，调节角度使得 3 个三维结构展示在一个界面里。在主菜单中选择“Structure”|“Superimpose”|“Molecular overlap”，在弹出的对话框中点击“Yes”。按下组合键“Ctrl+D”，弹出“Display Style”对话框，在“Atom”一栏中选“None”，在“Protein”一栏中选“Solid ribbon”，折叠结果如图 10-28。

第三步：基于结构的序列比对。在“Protocol Explorer”中，选择“Protein Modeling”下的“Align Structure”(MODELER)。在打开的“Parameter Explorer”中，选择“Input Sequence Alignment”为 model-1，点开“+”号，确保在“Input Protein Molecules”中包含以上三个蛋白分子，将“Gap Extension Penalty”一栏中的参数改为 3.0，其余参数均不变。最后得到的比对结果所示的序列相似性分别如下：identify：24.4%，simility：28.2%(图 10-29)。由此可以看出该模型的整合效果还是可以的。

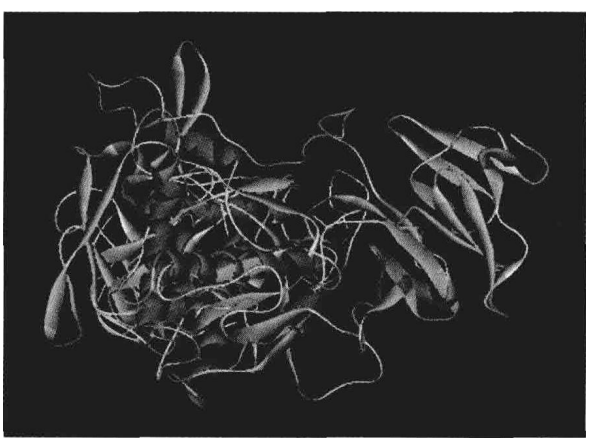


图 10-28 三个蛋白质分子的叠合图

Parameter Name	Parameter Value
⊟ Input Sequence Alignment	Model_1
Input Protein Molecules	Model_1, Model_2, 1rm8A
Gap Open Penalty	0.0
Gap Extension Penalty	3.0
Fast Alignment	False
Initialize Alignment	True

图 10-29 比对参数设置(基于结构)

第四步：靶序列与模板序列的比对。在“Protocol Explorer”中，选择“Protein Modeling”下的“Align Sequence with Structure protocol”，鼠标双击。在“Parameter Explorer”中，将“Gap Open Penalty”一栏中的参数改为 −450，“Gap Extension Penalty”设为 −25，其余参数均不变(图 10-30)。运行之后序列 Identify 为 18.1%，similarity 达到 20.9%。

第五步：同源模型的建立。在“Protocol Explorer”中，选择“Protein Modeling”下的“Build Homology Models Protocol”，鼠标双击。该模块通过使用 Modeler，从序列比对结果出发构建蛋白的三维结构模型。在“Parameter Explorer”中，“Input Sequence Alignment”栏选择 model-1，点开“+”号，“Input Model Sequence”栏选择 MMP14，“Input Template Structure”栏选择 model-1，model-2 和 1rm8，将“Cut Overhangs”栏改为 False，其余参数均保留默认值(图 10-31)。

Parameter Name	Parameter Value
⊟ Input Structure Alignment	Model_1
Protein Structures	Model_1, Model_2, 1rm8A
Input Sequence Alignment	mp142
Scoring Matrix	as1
Gap Open Penalty	-450
Gap Extension Penalty	-25
⊞ 2D Gap Weights	

图 10-30 比对参数设置(目标序列和模板序列)

Parameter Name	Parameter Value
⊞ Input Sequence Alignment	Model_1
Cut Overhangs	False
Disulfide Bridges	
Cis-Prolines	
Additional Restraints	
Copy Ligands	
⊞ Number of Models	1

图 10-31 参数设置(同源模型建立)

第六步：经过结构比对、目标序列与模板序列的比对等步骤，最后获得一个比较好的人 MMP14 三维模型(图 10-32)。窗口中的蛋白模型以飘带的形式显示，不同部位采用不同的颜色和宽度。颜色和宽度依 Verify score 而定，score 越高则蛋白结构越理想，从蓝色到红色 score 值依次降低，蓝色表示 score 值很高，白色表示 score 中等，而红色则表示 score 值较低。飘带的宽度与 score 值成反比，蛋白的结构越不理想，则该处的飘带越宽。

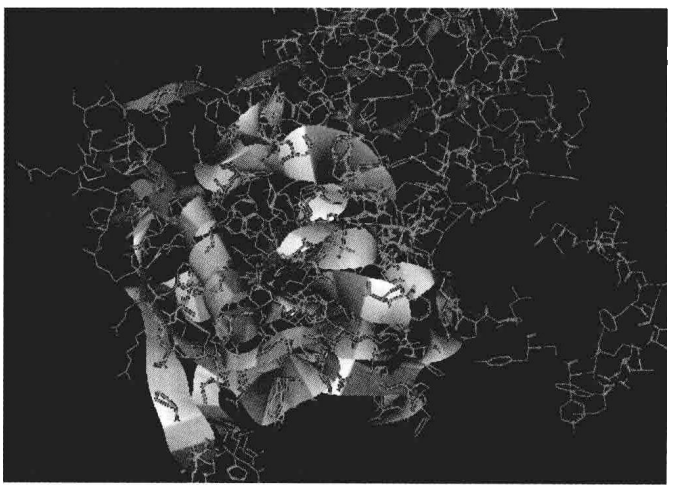


图 10-32 预测的 MMP14 三维飘带模型

第七步：模型验证。图 10-32 不能直观的看出模型建立的好坏。在 MMP14.msv 所在的 3D-Windows 中，按组合键“Ctrl+T”调出“Data Table”，选择“Amino Acid”。找到“PDF Total”一栏，鼠标单击将此列选中，然后选择主面板上的“Chart”|“Line Plot”，对此列作图(图 10-33)可见在 PDF 值较低的部位，蛋白的结构是比较合理的，而在 PDF 值较高的部位蛋白结构能量较高，可能还需要进一步的优化。例如：第 125 位和第 250 位残基附近的高峰区。

第八步：采用图形直观地展示模型的好坏；选择主面板上的“Chart”|“Ramachandran Plot”，对整个蛋白作图。如图 10-34 所示，位于蓝色区域以内的残基结构合理，处于蓝色区域以外紫色区域以内的残基结构比较合理，位于这两个区域以外的残基结构则合理性较差。

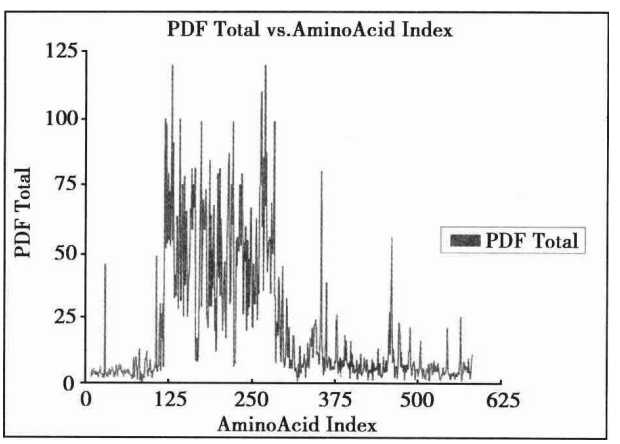


图 10-33 模型验证

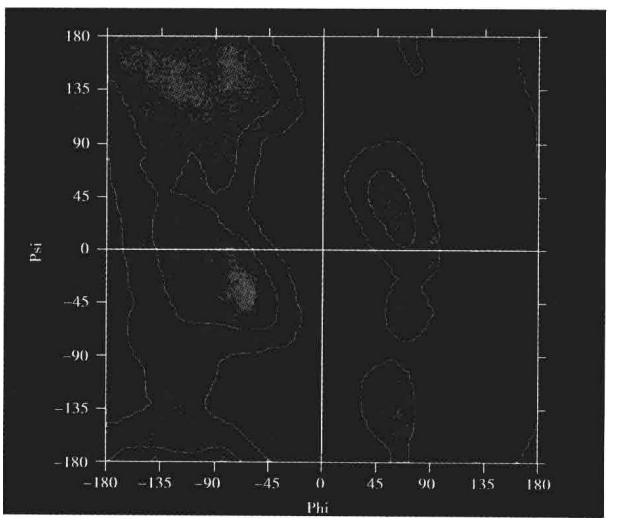


图 10-34 采用图形进行模型验证

（三）蛋白质三维结构的从头预测方法

如果目标蛋白质缺乏已知结构的同源蛋白质，则可采用从头预测方法（Ab initio，de novo design）或自由建模获得目标序列的三维模型，具体有两个主要策略：①根据已经预测的二级结构，把可信度较高的二级结构进一步组装，得到最后的蛋白质结，但是由于该方法较多地依赖于前面二级结构的预测，在很大程度上受到了一定的限制；②不依赖二级结构预测的结果，直接预测三维结构。

至2002年，ab initio所建模型仍是粗糙的，但ab initio与序列比对和穿线法一起，可望为弱同源模板结构的蛋白质折叠识别发挥作用。与其他的结构预测方法相比，从头预测方法的发展是缓慢而艰难的。ab initio预测蛋白质高级结构的理论依据是Anfinsen在1973年提出的热力学假说和量子力学，即在给定条件下蛋白质的天然结构对应其自由能最低的状态，单纯的从物理方法出发对蛋白质结构预测仍然是个假设。成功的ab initio方法依赖于以下因素的有效性：①蛋白质结构的最优代表具有充分结构可靠性和计算可控性；②具有近天然结构的力场；③一个方案可以在一定量的CPU时间内有效地寻找构象空间的重要区域；④一种方法可用于从模拟分析的伪品中正确鉴别出近天然结构，此方法还在发展中。

（四）蛋白质高级结构的其他预测方法

折叠识别法是近年来发展起来的一种比较新的三维结构预测方法，旨在应用到同源模建不能达到可靠预测的时候，即同源性小于30%的蛋白质通常采用该方法。适用于当未知蛋白找不到已知结构的蛋白质作为模板时的情况，且不需预测二级结构，即直接预测三级结构，从而可以绕过现阶段二级结构预测准确性比较低的限度，是一种有潜力的预测方法。

折叠识别法包括两步：第一步就是将目的蛋白序列和已知的折叠结构进行匹配，在已知的结构中找到一个或几个匹配最好的结构模型，作为目的蛋白的预测结构。第二步基于已有的知识找到最好的模型，毫无疑问，目前影响评价模型好坏的主要因素还是目标序列和模板结构序列的比对质量。

结构在进化上的保守性要高于序列。目前PDB中已经包含了足够的结构来覆盖小的蛋白质结构，这也为折叠识别法的发展奠定了一定的基础。但这种方法局限性在于已有的蛋白质折叠类型还是有限的，序列相似的蛋白也可能具有明显不同的折叠模式，这是结构预测需要关注的问题。

三、对结构预测结果的评价

面对多种的模型和预测方法，有一些多种公共范围的实验评估方法，主要有LB、CASP、CAFASP和EVA等。

1. LiveBench（LB）实验方法　该实验方法由Rychlewski和Fischer创建。每周收集新公布的蛋白质结构，利用这些相对大量的预测靶，LB不断地对各自动服务器进行能力评估，约半年评估这些预测方法一次。

2. CASP和CAFASP实验方法　即蛋白质结构预测技术评估规范（Critical Assessment of Techniques for Protein Structure Prediction，CASP）和CAFASP实验，该方法每两年举行一次，用于检测现行研发模建方法的能力和局限、确定研发的进展并阐明特殊的瓶颈，是蛋白质结构预测领域的一个重要里程碑。CASP1始于1994年，至今已完成CASP8。但CASP方法仍是很有争议的。最新的CASP实验显示，在比较建模领域，最成功的方法是利用同义策略（consensus strategies），其是基于多重模板或蛋白片段的重组构建最终模型。对自动结构预测方法的评估是通过CAFASP（Critical Assessment of Fully Automated Structure Prediction）试验来实现的。其与CASP平行操作，针对相同的靶。

3. EVA实验方法　EVA（http://cubic.bioc.columbia.edu/eva）是另一种评估方法。

结构比较模建和无模板建模（template-free modeling），后者尽管也很接近，但主要是获得一些小蛋白质的接近正确的模型。目前主要的挑战是比较建模的完善，从而与实验精度更匹配，获得精确序列的比对以用于基于进化远距离关系的建模，以及拓展无模板建模方法从而产生更多精确的模

型。以及针对大的结构预测。

开发更复杂和自动的计算机建模方法将极大地增加结构基因组的建模蛋白的范围。在该领域关键的问题和努力包括：①对于在 PDB 库中相似的序列(尤其是那些与靶蛋白弱或远距离同源的)如何确定正确的模板和如何优化模板使其与天然构象相近；②若序列无合适的模板，如何从头开始进行正确拓扑的建模。

随着人们对蛋白质序列、结构和功能相互关系的更深入的了解、技术的不断进步以及新算法、新方法的呈现，基于实验和预测方法将会有越来越多的蛋白质结构被精确解析和获得。

第五节 基于结构预测蛋白质功能

Section 5 Prediction of Protein Function Based on Structures

蛋白质间序列相似性高于 40% 时该蛋白质同其相似序列蛋白可能有某些相同生物化学作用；蛋白质间序列保守性低于 40% 时，可从高级结构预测功能。蛋白质有多个功能域可对应该蛋白质的某些精细功能。从高级结构预测功能实际上是预测蛋白质的某些局部的基本生物化学作用而不是全部生物化学功能。按蛋白质功能分类的数据库如 SPIN-PP 和 MIPS 等，为新蛋白功能预测提供了很多有用信息，详见本章第三节。

一、基于结构分类的蛋白质功能预测

知 识 拓 展

蛋白质构象的动态性是其复杂而精细功能的基础；分子动力学模拟可观察某些非常快或很小而重要的局部构象动态的特征，这也是探索蛋白结构功能关系和基于结构预测蛋白质功能的重要基础。常用分子动力学模拟方法以牛顿力学原理描述体系内原子间的相互作用及对体系能量的贡献，促进体系朝能量最小化方向演变，观察演变过程中原子间相互作用的变化，用统计力学原理将微观性质与宏观性质相关联。分子动力学模拟计算量大，已有 CHARMM、AMBER 和 Gromacs 等并行计算软件可用。详细介绍参见本书所附光盘。

蛋白质在进化中保守的高级结构通常对应某些保守的精细生物化学功能，故结构相似的蛋白质会有某些相似的精细生物化学功能。对已知结构的蛋白质进行分类，搜寻同类蛋白质的功能是预测目标蛋白功能的有效手段。

最早基于结构进行蛋白质功能注释的方法是搜索与目标蛋白质结构相似的蛋白质，并将其功能转移给目标蛋白质。此过程中需要进行蛋白质的结构比对和判断结构相似程度。可将这种相似性估值转化为序列比对问题，利用序列比对经典算法来解决结构比对问题，如 DaliLite、SSM、STRUCTAL、MultiProt 和 3DCoffee 等。基于“具有相似功能的蛋白质定位于结构空间图中相邻近的位置”，Hou 等(2005)使用多维度标度技术(multi-dimensional scaling，MDS)构建了一个蛋白质结构空间图(SSM)，根据 DaliLite 结构比对方法进行相似性打分，最终在构建的结构空间中按照距离阈值将一个新的蛋白质归类到某个功能类别中。值得注意的是蛋白质存在趋同进化，也存在趋异进化。如两个蛋白质结构相同但是序列不同，他们的结构相似性可能是趋同进化的结果，对应的功能保守程度不一定很高，即两个结构相似的蛋白质可能有不同的功能。

还有一些方法试图将结构相似性方法与其他方法结合进行功能决策。例如，考虑一个系统发育上下文中的结构相似性，会增加功能注释精确性。综合不同方法在特定生物学背景下解决结构比对问题，有助于提高结构预测功能精确性。

二、基于结构预测蛋白质间相互作用

细胞内存在与细胞生命活动密切相关的蛋白质相互作用网络。目标蛋白质同其他蛋白质的相互作用是其重要的功能之一，预测蛋白质间的相互作用是预测这类功能的有效策略。预测蛋白质间相互作用涉及预测可相互作用蛋白质和相互作用位点。目前主要有如下策略用蛋白质的高级结构信息预测蛋白质相互作用。

（一）基于结构的物理对接

理论上可用物理对接预测两个蛋白质间的相互作用位点，但对体积很大的蛋白质分子，相互作用的可能界面太多而计算工作量很大。目前有些软件将目标蛋白质作为刚性球体，通过评价候选作用蛋白质与其在表面形状和理化性质的互补性，并基于分子动力学等技术进行优化，尝试识别结合位点和探索复合物的结构信息。应用中此技术如考虑目标蛋白质的柔性则计算工作量更大。

（二）识别相互作用界面序列特性模式进行预测

发掘蛋白质相互作用界面的序列特性信息可用于预测相互作用且更易于实践，目前已有如下的应用方式。

1. 关联性突变法　能保持相互作用的蛋白质，相互作用界面的残基突变存在关联性。这种策略不需要目标蛋白的高级结构而只需要序列信息，且计算量比基于结构的物理对接小得多。

2. 联用方法　联用高级结构和序列信息可提高预测可靠性。用关联性突变法预测结合区域；通过刚性对接模式获得候选蛋白与其复合物的结构；再用关联性突变法为距离限制标准筛选真正的候选蛋白质复合物。用这种方式已成功的预测了血红蛋白亚基之间的相互作用。

3. 人工神经网络学习法　利用高级结构信息和序列特征进行训练，可建立蛋白质间相互作用界面的预测方法。这种方法中利用高级结构信息来定义临近残基的界面区域，使用多序列比对获取这些界面的序列特征；用已知相互作用的蛋白质复合物进行训练；利用建立的相互作用界面区域信息预测已知结构的未知蛋白质的可能相互作用界面是否存在。这种方法的预测准确度可达到 70%。

三、其他蛋白质功能预测方法

还有其他基于结构预测功能方法，具有代表性的如下：

（一）基于基序的方法

基于基序的方法(motif-based approaches)通过识别功能相关的蛋白质中保守的三维基序，并建立这些保守的基序和保守的蛋白质功能间的映射关系用于预测目标蛋白质的某些生物化学功能。酶进化过程中其催化残基通常最保守。相同或相似功能的酶进化后序列差异可能很大，但围绕催化位点的结构信息可能具有很好的保守性。这种基于结构比对的保守性分析策略是预测未知功能酶蛋白的有效手段。这种策略有多种成功应用尝试，已有下列具体方法和软件。

1. SITE 程序和数据库储存了酶活性位点保守基序信息　此数据库用位点匹配程序寻找关键的功能位点残基作为保守残基。但是这些数据分析发现：即使高度同源的蛋白质，有些程序认证的功能位点残基也不保守，即不属于结构基序。因此，需要仔细分析这些信息以寻找新蛋白质的未知功能。

2. TESS 程序　采用了几何散列算法，通过模板研究和重叠，从蛋白质的高级结构中寻找保守的必须残基。通过匹配一个模板蛋白和未知功能的新蛋白，考察预期的保守残基是否存在来认证有无相似性。但 TESS 采用功能残基的侧链坐标进行匹配，对高级结构测定精度要求较高。

3. 模糊功能形态(FFF)　从三维信息角度认证与生物学功能相关位点的保守性。其用主链 α 碳原子坐标进行匹配。通常高级结构中主链 α 碳原子坐标解析精度较高，故 FFF 的适用性强于 TESS。此方法预测对硫氧化还原蛋白的功能预测比较成功，延伸到人肿瘤抑制基因产物 N33 也获得成功。

4. SPASM　同时用主链 α 碳原子和侧链基团作为分析对象，寻找并列的保守残基，并用于搜寻

结构数据库中能匹配的已知功能蛋白，此程序容易使用。

5. 分子识别策略分析　是基于已知功能域四周原子的重叠认证保守性预测蛋白质功能。这种策略认为当已知功能蛋白和未知功能蛋白在预期的功能域四周都具有很高的原子重叠时，它们结构相似性较高就可能有类似功能。此方法对蛋白间的序列相似性要求很低。用于识别腺嘌呤结合域四周的原子重叠，发现即使序列相似度极低的蛋白激酶、依赖 cAMP 的蛋白激酶和酪氨酸蛋白激酶等需要结合腺嘌呤的蛋白质符合很好。用于磷酸盐结合位点的分析也得到预期结果。

蛋白质侧链的保守模式分析预测功能类似于前述的 TESS、FFF 和 SPASM，分析重复出现的氨基酸侧链的保守性。这种方法只需新蛋白质的结构数据和与之关联的多重序列排列。保守性的主要约束机制是氨基酸残基和距离，但是不考虑与活性位点无关的残基的性质。但此方法判断标准为残基的偏离均方差根，并用统计学意义进行评估，本质上属于物理学方法。

（二）基于表面的方法

基于表面的方法(surface-based approaches)对给定蛋白质进行表面模型化，利用与结构相关联的蛋白质表面模型，识别蛋白质表面上的结构(如空间特性或穴等)特征，进而利用这些特征来推断蛋白质功能。与基序分析相似，蛋白质结构模型可以通过蛋白质分子表面的互补性来展示在氨基酸或原子水平下由分子内相互作用所体现的特定的生化功能。例如，疏水性表面经常作为相互作用分子之间的接口，静电分子表面也常被用来解释蛋白功能。这些方法必须基于两个蛋白表面之间的匹配模型，常用图论技术解决结构匹配问题。

SURFACE 数据库提供对输入蛋白质进行局部表面特征模式(local surface patterns，clefts)识别，进而进行蛋白质功能注释的系统。该系统首先使用 SURFNET 算法搜寻蛋白质表面的 clefts；根据 PROSITE 数据库利用 GO 功能进行表面 clefts 模式功能注释；再用 RMSD 和 PAM 相似矩阵方法，综合考虑了结构和残基相似性预测相互作用。这种匹配算法以多对多方式对 PDB_SELECT 数据组中结构所构建的结构模式进行估值，精确性一般能达到 90% 左右，但计算量很大。

（三）基于学习的方法

基于学习的方法(learning-based approaches)是利用有效的分类方法，从最相关的结构特征中识别最合适的功能类别，如 SVM 和 KNN 等分类方法。基于学习的方法以蛋白质结构特征作为分类依据，功能分类作为样本标签，通过数据对象之间的相似性矩阵对训练蛋白质进行结构与功能关系的评估。某些代表性的方法比较两个蛋白质结构的核函数 kernel 包含：①定义两个亚结构间相似性 KPattern_Sim(S，T)；②基于蛋白质基序 CxxC 定义巯基化合物 / 二硫化物和氧化还原酶蛋白的功能相似性 KRedox_Func(S，T)；③将蛋白质看成由多个氨基酸构成的具有给定半径的一组球体，定义两个蛋白质的结构相似性为 K3Dball(P1，P2)。利用这些 kernel 函数，可构建 KNN 和 SVM 分类器，针对两个独立的实验数据集(一个来自 SCOP 的 10 个蛋白质超家族，一个来自 PDB 的 21 个巯基化合物 / 二硫化物、氧化还原蛋白)进行训练，最终发现 K-NN 方法获得了比 SVM 方法更好的基于蛋白质结构的功能预测效果。更多的蛋白质结构解析将为这类方法提供更加完备的特征训练库，使其有望超越其他现有的方法。

四、蛋白质结构与功能关系数据库

蛋白质结构与功能关系数据是进行蛋白质功能预测及蛋白质设计的基础。目前已有一些蛋白质结构与功能关系的数据库，如 PIR、Pfam 和 InterPro 等。

（一）Pfam 数据库

Pfam(the Protein Families database)是通过自动比对构建的蛋白质结构域家族数据库，它收集了大量的蛋白质多重序列排布以及 HMMs 文件(profile hidden Markov models)的数据，将具有结构相似性的序列归为一类，可用类的名称查询到原始序列比对信息。它可广泛用于通过序列比对推测蛋白质结构域排布形式及其功能。最新的 Pfam 24.0 版本涵盖了 11912 个蛋白质家族，这些 Pfam 家族

是基于 SWISS-PROT 以及 TrEMBL 中的蛋白质数据的。应用 Wise 2 软件包可以用基因组 DNA 对 Pfam 文库进行直接搜索，在 http://pfam.sanger.ac.uk/ 搜索的结果如图 10-35。有多个网站支持这类数据库和搜索。

图 10-35 Pfam 数据库主页

Pfam 数据库包含 Pfam-A.seed 和 Pfam-A.full 等文件，这些是以 Stockholm 格式注释的“seed”和“full”排布。PfamFrag 是为搜索相匹配的蛋白片段而特别设计的 HMMs 文件文库。PfamB 是以 Stockholm 格式注释的 Pfam-B 家族数据文件；Diff 是用来对 Pfam 来源数据进行更新的文件；Pfamseq 是以 fasta 格式注释的序列数据。Pfam 数据库包括文本搜索、蛋白质 HMM 搜索、DNA HMM 搜索、浏览 PFAM、NIFAS 和结构域查询等几个部分。

Pfam 包括功能注释、参考文献，以及与每个家族相链接的数据库。每个 Pfam 家族包括“seed alignment”(由家族中具有代表性的成员构成)和“full alignment”(所有家族成员构成)两部分。所有排布都采用来源于 Pfamseq 的数据。在“seed alignment”基础上，应用 HMMER(http://hmmer.wustl.edu)建立了 HMM 文件对 Pfamseq 序列数据库进行搜索。Pfam 的重要功能包括将蛋白质快速自动划分入不同的结构域家族。当前主要运用 HMMer 软件对蛋白质翻译进行注释，或应用 Gene Wise 2 软件直接预测基因并注释基因组 DNA。GeneWise 的检测结果表明，在同源区域内它预测基因的准确性可达到 98%。结构域边界选择错误可能造成家族分类重叠或遗漏。但随着 Pfam 数据库的不断完善，其功能将日趋完善。

(二) PIR 蛋白质功能预测数据库

PIR 全称 The Protein Information Resource，是集成了蛋白质功能预测数据的公用数据库。PIR 与 MIPS(the Munich Information Center for Protein Sequences)、JIPID(the Japan International Protein Information Database)合作，共同构成了 PIR- 国际蛋白质序列数据库(PSD)，目前版本 15.13，包括了 250 000 个蛋白。为了提高蛋白质功能预测可靠性，PIR 建立了一套系统用于递交、分类和提取文献信息。PIR 在超家族、域和模体水平上对蛋白质分类，同时提供蛋白质的结构和功能信息，并给出了与其他 40 个数据库之间的相互参考。PIR 自身提供非冗余的蛋白质数据库，包括从 PIR-PSD、SWISS-PROT、TrEMBL、GenPept、RefSeq 和 PDB 收集来的约 800 000 条序列，对每条序列给出了一个符合的名称和相关文献。PIR 采用开放的数据库框架，利用 XML 技术进行数据发布。

除了蛋白质序列数据以外，PIR 还包含以下信息：①蛋白质名称、蛋白质的分类和蛋白质的来源；②关于原始数据的参考文献；③蛋白质功能和蛋白质的一般特征，包括基因表达、翻译后处理、活化等；④序列中相关的位点、功能区域。PIR 提供三种类型的检索服务：①基于文本的交互式查

询，用户通过关键字进行数据查询；②标准的序列相似性搜索，包括 BLAST、FASTA 等；③结合序列相似性、注释信息和蛋白质家族信息的高级搜索，包括按注释分类的相似性搜索和结构域搜索等。

（三）InterPro 数据库

整合蛋白质结构域和功能位点资源数据库(Integrated Resources of Proteins Domains and Functional Sites，InterPro)是集成了蛋白质结构域和功能位点的数据库，其主页地址为：http://www.ebi.ac.uk/interpro/。InterPro 包含关于蛋白质家族、域和作用位点的整合的数据资源，它最初是作为一种对 PROSITE、PRINTS、Pfam 和 ProDom 数据库工程的一种补充手段而建立的。InterPro 已经成为 Oracle 中的一个相关数据库，用户可以直接利用 JAVA 服务器进入数据库，它以 XML 形式文件分发存储数据。同时，InterPro 数据库提供了一个位于常用的署名数据库之上的整合层，能提供友好的人机界面和基于文本的搜索和序列扫描(表 10-10)。Interpro 包含很多来自不同数据库的诊断签名的人工助理文件，形成了对一个给定的蛋白质家族、结构域和功能位点的独特描述。

表 10-10 InterPro 子成员数据库信息

子数据库名称	版本号	子数据库蛋白质条目	整合到 InterPro 数据库的条目
HAMAP	280509	1633	517
PANTHER	6.1	30 128	2234
Pfam	24.0	11 912	9151
PIRSF	2.70	2742	2691
PRINTS	39.0	1950	1926
ProDom	2006.1	1894	990
PROSITE patterns	20.52	1308	1304
PROSITE profiles	20.52	860	836
SMART	6.0	809	804
TIGRFAMs	8.0	3603	3580
GENE3D	3.0.0	2147	1025
SUPERFAMILY	1.69	1538	1093

InterPro 蛋白质数据库，当前版本为 24.0，发布日期为 2009 年 12 月 16 日，包含了 18 349 个与蛋白质相关的条目信息，它们包括 83 个活性位点、59 个绑定位点、551 个保守位点(保守 motif)、5149 个结构域、11 082 个蛋白质家族、23 个后转录修饰、1189 个蛋白质区域和 213 个重复区域等信息。

（四）常用数据库

目前常用的蛋白质结构和功能数据库见表 10-11。

表 10-11 蛋白质结构和功能关系数据库

数据库	结构信息	网址	功能信息
SignalP	信号肽	http://www.cbs.dtu.dk/services/SignalP	蛋白质信号肽信息
ScanProsite	结合位点	http://us.expasy.org/tools/scanprosite	检索 Prosite 数据库的快捷方式，提供结合位点描述信息
Pfam	结构域	http://pfam.sanger.ac.uk/	结构域常用数据库，提供结构域功能描述
SMART	结构域	http://smart.embl-heidelberg.de	结构域常用数据库，提供结构域功能描述
InterPro	结构域	http://www.ebi.ac.uk/interpro/scan.html	结构域常用数据库，提供结构域功能描述
MATA	拓扑结构	http://cubic.bioc.columbia.edu/predictprotein/submit_met.html	可自动链接到不同的拓扑结构分析程序，如 PROFphd、PSIPRED、SSpro2 等，通过 Email 方式反馈结果
TMHMM	跨膜结构	http://www.cbs.dtu.dk/services/TMHMM-2.0	常用的跨膜结构预测平台

续表

数据库	结构信息	网址	功能信息
PSORT	细胞定位	http://psort.nibb.ac.jp/form2.html	查找细胞定位信号或基序
PDB	3D 结构	http://www.pdb.org	新发现蛋白质通常为阴性结构，但可与同源蛋白质进行结构比较
MIPS	物理结构互作	http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction	收集酵母中蛋白质相互作用
COG	同源性家族	http://www.ncbi.nlm.nih.gov/COG	存储多物种同源蛋白质信息，蛋白质家族信息

第六节 蛋白质结构异常与疾病

Section 6 Protein Structure and Diseases

当蛋白质保守位点发生性质截然相反的突变(如：亲水性氨基酸被疏水性氨基酸替代)时，蛋白质的高级结构可能被显著改变而影响其功能。另外，蛋白质序列不变而高级结构发生显著改变，例如变性(denaturation)或错误折叠(misfolding)，也会造成蛋白质功能的显著改变，特殊情况下就造成病理生理现象。

知 识 拓 展

基于大规模基因组测序和高通量蛋白质三维结构预测，可以获得全基因组中各个基因的结构分布信息。将 SNP 分配到三维蛋白质结构中可为蛋白质结构和功能相关性研究提供极大的便利，非同义 SNP 由于其可能影响蛋白质的序列、结构和功能，所以有可能与人类疾病的进程相关。TopoSNP 正是一个将非同义 SNP 与三维结构结合起来，以助于了解 SNP 在疾病进程中的作用的数据库。

一、蛋白质序列变化引发的疾病

氨基酸序列决定了蛋白质的三维结构，有时即使是一个残基的变化也会引起结构的显著改变。Ratjen 和 Doring 研究表明囊性纤维化病的病因是在编码囊性纤维化跨膜调控蛋白(CFTR)的基因内发生了变异。比较普遍的变异是△F508 导致了 CFTR508 位苯丙氨酸的缺失。这种缺失的后果是改变了该蛋白质中 α 螺旋的含量。CFTR 正常情况下位于肺上皮细胞质膜，而这种结构变化在一定程度上阻碍了 CFTR 通过旁分泌到达此位点的过程。

与疾病关联的蛋白质序列变化并不一定导致蛋白质结构上的巨大变化，这样的一个例子是血红蛋白基因突变和镰刀形细胞贫血症。血红蛋白 β 基因第 6 位氨基酸由谷氨酸突变成缬氨酸所造成的镰状细胞贫血症，是蛋白质序列改变引起高级结构和功能显著改变继而引起疾病的典型代表。

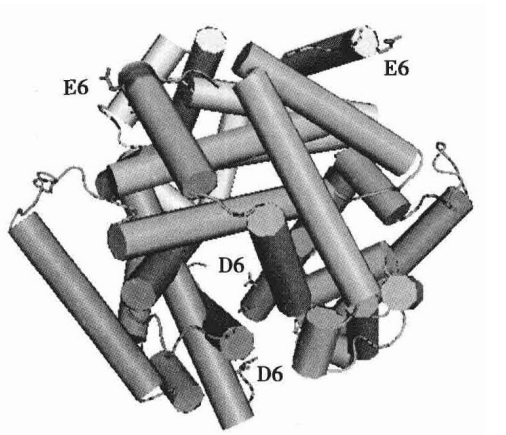


图 10-36 血红蛋白结构示意图(1VWT.pdb)
图中标示了 α 亚基上第 6 位的天冬氨酸(D6)和 α 亚基上第 6 位的谷氨酸(E6，红色)

血红蛋白为四聚体氧载体，含 α 和 β 亚基各两个，每个亚基都含有血红素辅基(图 10-36)。目前发现的血红蛋白基因突变类型多达 1000 种，约有 40% 左右的突变伴随

临床症状。其中，β链的第6位氨基酸由谷氨酸突变成缬氨酸造成镰状细胞贫血，其对应血红蛋白为HbS。Hb高级结构中第6位谷氨酸位于N端第一个α螺旋起点；第6位突变为缬氨酸后，HbS表面的静电分布，尤其是对应螺旋两端的静电分布发生了显著的改变，这可能会影响第一个α螺旋同其他螺旋的偶极相互作用和疏水相互作用，使得HbS结合氧前后的构象差异比正常Hb更大。当HbS与氧结合后其构象显著不同于未结合氧的HbS，其不形成聚集的纤维丝，对红细胞行为无影响。红细胞中Hb浓度太高；HbS在未与氧结合时其特殊的构象和表面性质诱发聚集生成HbS纤维，并与细胞膜接触，降低了细胞膜的变形性，使得红细胞通过毛细血管末端释放氧后容易破裂，造成贫血。

二、蛋白质折叠错误引发的疾病

诱发神经系统退行性病变的淀粉样蛋白(amyloid-protein，A)、突触核蛋白(synulcein)是蛋白质序列相同但四级结构不同而诱发疾病的典型代表，朊蛋白也是这种作用机制的致病蛋白之一。

在阿尔茨海默病(AD)发生过程中出现淀粉样蛋白。A是由特殊水解酶对其前体蛋白的水解作用产生的。A有两种构象，一种为螺旋且可溶而存在于健康个体脑组织，此类A为单体没有四级结构；另一种为片层且是多个A聚集形成的链间片层，此类A不溶且出现在AD患者脑组织(图10-37)。诱发A从可溶螺旋转变成不溶片层聚集体的机制不清，但已广泛证实这种构象转变是AD的重要诱因。由图10-37(D)可见其每条链N端和C端首尾靠近区域全是强疏水残基，片层外侧也主要是疏水或中性残基。按此前描述水溶性蛋白质特征，这种特殊形状和表面结构的蛋白质水溶性肯定低。针对这种难溶的错误折叠蛋白质，通过设计能稳定其结构的小肽，使其复合物组装成模式且可溶，图10-37(E)，这是治疗AD的药物研发方向之一。

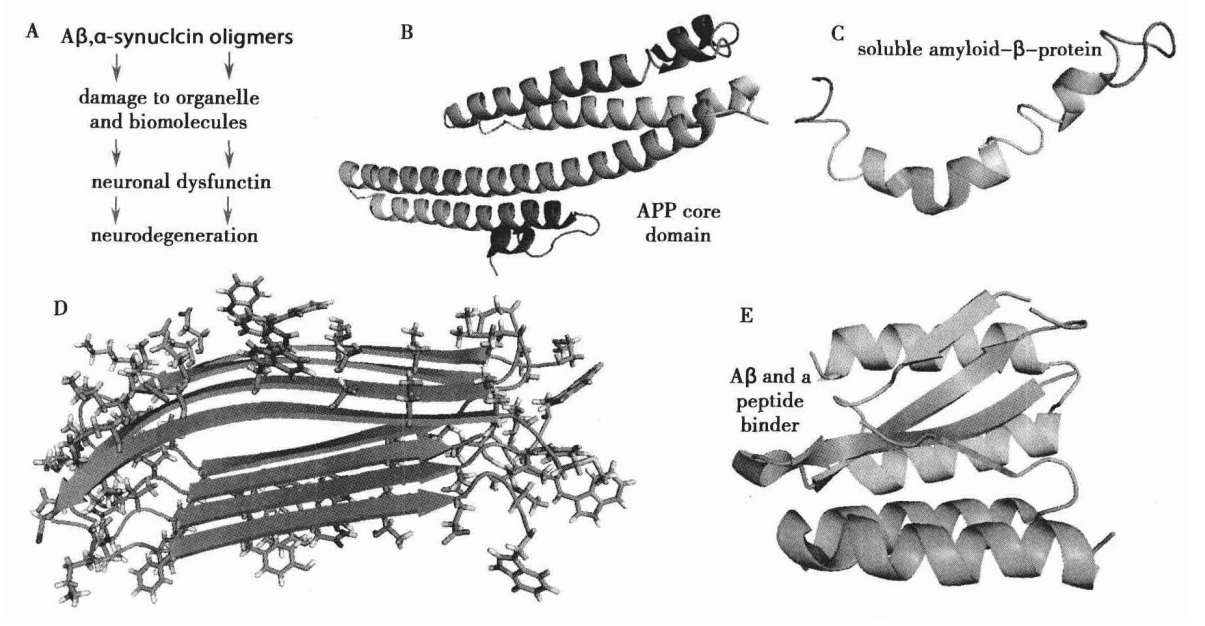


图10-37 退行性神经病变和Aβ构象

A. NDD的假设；B. Aβ核心结构域(1RW6.pdb)；

C. 可溶Aβ构象(1ZOQ.pdb)；D. 聚集的纤维状Aβ构象(2NNT.pdb)；

E. Aβ和设计的结合蛋白的复合物构象(2OTK.pdb)

另一个代表性例子是朊蛋白，最早发现于疯牛病；此蛋白质也可诱发退行性神经病变，现证实对多数动物都有这类同源蛋白质(图10-38)。这类蛋白质和A一样也都是错误折叠而聚集不溶的蛋白质。但与A不同的是朊蛋白有传染性，即错误折叠的朊蛋白可诱发本来可溶的朊蛋白转变成不溶的

聚集体。至今只明确朊蛋白致病和传染的病理作用同其聚集生成不溶性聚集物有关，并初步发掘了其可能的聚集位点，但这些聚集物如何诱发疾病仍不清楚。

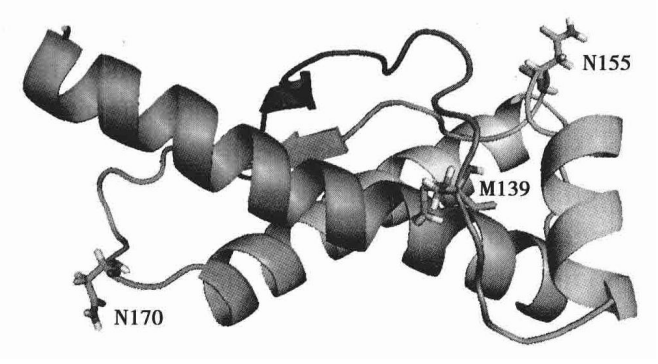


图 10-38 仓鼠朊蛋白的结构示意图
M139、N155 和 N170 是初步推测的聚集接触位点

三、疾病过程中蛋白质的相互作用

蛋白质结构互作引发疾病的过程，往往伴随着蛋白质序列变异或错误折叠等结构突变，从而导致正常的蛋白质结构互作缺失，引起特定的功能或表型异常。几乎所有的急性骨髓样白血病(AML)都是由基因突变引起的。在这些基因中，有一部分基因的编码产物组成了一种转录因子复合物——核心结合因子(CBF)。CBF 由两部分组成，其一是 CBFα，它直接和 DNA 发生结合；其二是 CBFβ，它能够帮助 CBFα，结合 DNA。所有的 CBFα 亚基都含有一段保守序列——Runt 结构域，CBFα 正是通过该结构域结合 DNA 并和 CBFβ 起反应的。人体内共有三种编码 CBFα 亚基的基因，这些基因的突变可导致人类疾病的发生。2009 年 4 月《Nature》报道，英国剑桥医学研究协会的科学家 Allan Warren 教授及其合作者，运用 X 射线成功破译了 CBFα 亚基和 DNA 结合复合物的三维结构。他们发现，正是由于 CBFα 亚基 Runt 结构域的变异，导致了 CBF 失去其结合 DNA 的能力，使得 CBFα 和 CBFβ 之间不能正常反应，进而引发了各种 AML。

蛋白质的高级结构决定了其在生物体内的功能，多个蛋白质发挥作用时常需要与其他蛋白质协同作用，不同蛋白质之间形成复合体(complex)。每个蛋白质可以看成复合体的一个亚基(subunit)，亚基间相互作用，形成紧密的复合体结构或共同组成复合体的活性中心。这种相互作用往往涉及蛋白质的结构域，当结构域中的氨基酸因突变被替换后，会导致复合体去稳定化(destablization)，破坏了原有的蛋白质间的结构互作关系，复合体形成障碍而被降解。随着蛋白质精细结构的逐步解析，从蛋白质结构互作的角度来研究和探索复杂疾病的潜在发生机制，将会成为结构基因组学十分有意义的研究方向。

小 结

本章解析蛋白质的三维结构特征是解释生命活动机制的重要部分。蛋白质结构数据库的日益完善和结构生物信息学技术的发展，大大加速了蛋白质的结构解析过程及与其紧密相关的蛋白质的功能研究进程。根据蛋白质高级结构特征，可预测蛋白与小分子配体的作用、蛋白质间的作用及蛋白质与核酸的相互作用，并可以据其结构同类家族的功能特征预测新蛋白的功能。随结构生物信息学研究的进一步深入，基于生物信息学分析方法对单个蛋白质的结构、蛋白质多聚体和复合体的结构，及它们在相互作用过程中蛋白质结构的动态变化过程进行研究，必将有助于从结构角度揭示生命过程和复杂疾病的致病机制。

Summary

The exploration on protein conformation contributes significantly to the understanding of biological mechanisms during life processes. With the accumulation of protein structure databases and the development of bioinformatics techniques, more and more researchers dedicate their focus onto interpretating structures and functions of proteins effectively. Based on the structure of a protein, it becomes feasible to predict complicated protein associations, which include protein-ligand interactions, protein-protein interactions and protein-nucleic acid interactions, and then to predict potential functions of the protein. The advancement of bioinformatics will enable researchers to in-depth understand the underlying importance of protein structures, protein polymerides and protein complexes, which could provide additional insights into life process and the pathogenensis of complex diseases from the protein structure perspective.

（廖 飞 茆灿泉 陈丽娜 魏冬青）

习 题

1. 蛋白质高级结构的分类及特征有哪些？
2. 简述实验解析蛋白质高级结构方法的特点。
3. 蛋白质三大结构数据库 PDB、SCOP 和 CATH 有何异同？
4. 采用哪些理论计算的方法可预测蛋白质高级结构？
5. 基于结构如何对蛋白质进行分类？如何预测蛋白质的功能？
6. 安装一款免费的分子图形学软件到电脑上。
7. 网上获取 HIV 反转录酶和 Bcr/Abl 激酶等蛋白质的晶体结构数据。
8. 图形显示上述蛋白质的二级结构和三级结构的特征。
9. 以 Bcr/Abl 激酶为未知功能蛋白，检索结构分类数据库预测其功能。
10. 选择 HIV 反转录酶中某些氨基酸序列进行改变，再用 Swiss-Model 进行同源建模，用图形学系统分析突变位置，据本章所述原理预测该蛋白质的结构和功能的可能改变。

主要参考文献

1. Aloy P., Pichaud M., Russell R. B.,. Protein complexes: structure prediction challenges for the 21st century. Curr. Opin. Struct. Biol., 2005, 15(1): 15-22.
2. Aung Z., Tan K. L. Rapid retrieval of protein structures from databases. Drug Discov. Today, 2007, 12(17-18): 732-739.
3. Gherardini P. F., Helmer-Citterich M. Structure-based function prediction: approaches and applications. Brief Funct Genomic Proteomic,2008, 7: 291-302.
4. Hasegawa H., Holm L. Advances and pitfalls of protein structural alignment. Curr. Opin. Struct. Biol., 2009, 19(3): 341-348.
5. Jellinger K. A. Recent advances in our understanding of neurodegeneration. J. Neural. Transm., 2009, 116: 1111-1162.
6. Tseng Y. Y., Dundas J., Liang J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J. Mol. Biol., 2009, 387(2): 451-464.
7. Moore R. A., Taubner L. M., Priola S. Prion protein misfolding and disease. Curr. Opin. Struct. Biol., 2009, 19(1): 14-22.

8. Sacan A., Toroslu I. H., Ferhatosmanogl H. Integrated search and alignment of protein structures. Bioinformatics, 2008, 24(24): 2872-2879.
9. Soto C., Estrada L., Castilla J. Amyloids, prions and the inherent infectious nature of misfolded protein aggregates. Trends Biochem. Sci., 2006, 31(3): 150-155.
10. Uversky V. N. Intrinsic disorder in proteins associated with neurodegenerative diseases. Front Bio. Sci., 2009, 14: 5188-5238.
11. Watson J. D., Laskowski R. A., Thornton J. M. Predicting protein function from sequence and structural data. Curr. Opin. Struct. Biol., 2005, 15(3): 275-284.
12. Wolfson H. J., Shatsky M., Schneidman D. From structure to function: methods and applications. Curr. Protein. Pept. Sci., 2005, 6: 171-183.
13. P. E. 波恩, H. 魏西希. 结构生物信息学. 刘振明, 刘海燕译. 北京: 化学工业出版社; 2009.
14. G. A. 佩特斯科, D. 格林. 蛋白质结构与功能入门. 葛晓春译. 北京: 科学出版社; 2009.
15. 卡尔·布兰登, 约翰·图兹. 蛋白质结构导论. 王克夷, 龚祖埙译. 上海: 上海科技出版社; 2007.

第十一章 转录调控的信息学分析
CHAPTER 11 BIOINFORMATICS ANALYSIS OF TRANSCRIPTIONAL REGULATION

第一节 引 言
Section 1 Introduction

基因表达是指基因在生物体内的转录、剪接、翻译以及转变成具有生物活性的蛋白质分子之前的所有加工过程。人类基因组大约有两万多个基因，但是在单个细胞中，同时表达的基因往往只有几千甚至几百个，而且很多基因只在特定组织或发育阶段表达。从一套基本不变的基因组中产生出多元化的细胞类型是由调控基因活性的各种信号途径所控制。作为基因表达的第一步——转录是调控机制的中心。转录调控因子(transcription factors，TF)，也称之为反式作用因子(trans-acting factor)有序地结合在目标基因启动子(promoter)序列中的特殊位点，启动基因的转录和控制基因的转录效率(图 11-1)。这些位点被称为转录因子结合位点(transcription factor binding sites，TFBS)，又被称为顺式调控元件(*cis*-regulatory elements)，其长度从几个到十几个碱基对不等。每个转录因子的结合位点通常都有特定的模式，被称为模体(motif)。找到这些特定的序列片段对研究基因的转录调控有着重要意义。

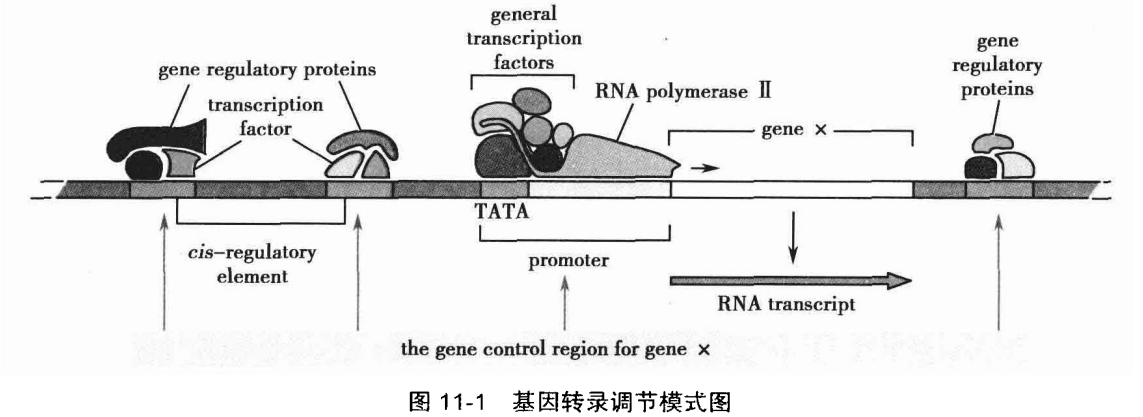


图 11-1 基因转录调节模式图

实验中常常选用荧光素酶报告基因(luciferase report gene)、凝胶迁移(electrophoreticmobility shift assays)、染色质免疫沉淀(chromosome immunopreciation，ChIP)或 DNase 足迹法(DNase footprinting)等方法来确定转录因子结合位点。尽管这些方法比较准确，但是尚不能够实现大规模、高通量的分析。近年来，随着基因芯片和高通量测序等数据的出现，计算方法在转录因子结合位点的分析中得到了广泛的应用。并且，利用微阵列芯片的海量数据和日益完善的生物信息学分析工具，对基因转录调控区进行详细分析已成为实验手段的重要补充。

第二节 转录调控的高通量实验测定

Section 2 High-throughput Techniques in Transcriptional Regulation Analysis

真核生物的基因组 DNA 以染色质形式存在。因此，研究蛋白质与 DNA 在染色质环境下的相互作用是阐明真核生物基因表达调控机制的基本途径。ChIP 是目前研究 DNA 与蛋白质相互作用最有效的方法之一。近些年来，随着基因芯片和新一代测序技术的迅猛发展，对哺乳动物在全基因组水平上进行高分辨率的 DNA 结合蛋白研究成为可能，由此也就诞生了在 ChIP 技术基础上建立的高通量分析蛋白质与 DNA 相互作用的技术平台 ChIP-chip 和 ChIP-seq。

一、ChIP 技术

ChIP 技术是由 Alexander Varshavsky 及其所在团队于 20 世纪 80 年代末创立的。其基本原理是先用甲醛(formaldehyde)处理活细胞，使 DNA 碱基上的氨基或亚氨基和蛋白质上的 α- 氨基及赖氨酸、精氨酸、组氨酸、色氨酸的侧链氨基与其他 DNA 或蛋白质上的氨基或亚氨基交联在一起，在几分钟内形成生物复合体，即稳定的蛋白质 -DNA 复合物。随后裂解细胞提取基因组 DNA，并用超声处理(sonication)使之断裂为长度为 0.2~1kb 的片段。这些附着有蛋白质的 DNA 片段随后用特异的抗体进行免疫沉淀(immunoprecipitation)。最后，经过去交联反应(reverse cross-link)，使相互作用的蛋白质和 DNA 片段分离，并纯化目的蛋白结合的 DNA 片段(图 11-2)。通过 PCR 对目的片段的检测(ChIP-PCR)，最终获得蛋白质与 DNA 相互作用的信息。其优势在针对特定候选蛋白质，如某一转录因子，是否特异性结合于所调节的靶基因(target gene)的某一预定区域内(如启动子区)进行检测。更重要的是，对同一 DNA 底物，可以运用多种不同的抗体，分别进行免疫共沉淀，以确定多种抗原，即多种结合蛋白在同一染色质片段上的结合。如图 11-2 所示，分别用 Sp1、c-Jun 或 c-Fos 抗体进行

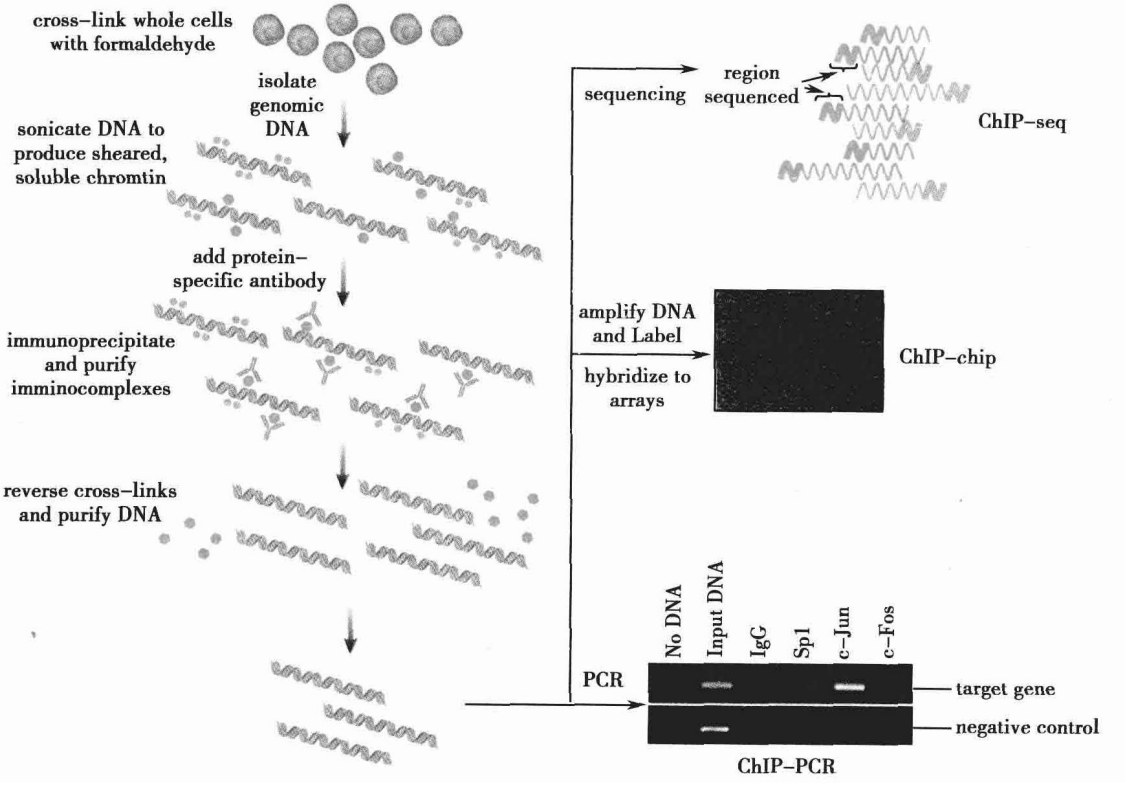


图 11-2 ChIP、ChIP-chip 和 ChIP-seq 工作流程

免疫共沉淀，然后对调节的靶基因启动子区进行 PCR 检测，发现仅 Sp1 和 c-Jun 能同时结合到靶基因的 DNA 序列上，提示此靶基因的转录可能被 Sp1 和 c-Jun 两种转录因子同时调控。这种方法为 DNA 结合的蛋白复合体的研究提供了新的思路，使研究者能更清晰地描绘蛋白的结合次序和相互协作关系。值得注意的是，为了确证实验结果的可靠性和特异性，在实验操作中需设立多种对照：①免疫沉淀时，与加入特异性抗体相对应，需取出另一份 DNA 样本，加入与特异性抗体同一种属的免疫球蛋白(IgG)作背景对照；②进行靶基因特异性 PCR 时，除了空白对照外(无 DNA)，应用免疫沉淀前的 DNA 片段作阳性对照；③除了用靶基因特异性引物进行扩增外，还需用无关序列引物进行 PCR。只有上述对照成立，才能判定某种蛋白与相应的 DNA 序列存在相互作用关系。

二、ChIP-chip 技术

ChIP-chip 技术是由 Richard A. Young 及其所在团队于 2000 年创立的。ChIP-chip 是一种全基因组范围内的定位分析技术，建立于 ChIP 和芯片技术的联合运用之上，用于分析细胞中 DNA 结合蛋白对特异结合位点，包括启动子、增强子、抑制子、沉默子、绝缘子、边界元件，以及 DNA 复制的调控序列的鉴定，整体研究生物体发育和病变过程中的复杂信息网络，是一个绘制基因组功能元件作用网络的技术平台。与单一 ChIP 相比，ChIP-chip 实现了鉴定任何一个特定 DNA 结合蛋白如转录因子或组蛋白等的靶基因群的高通量分析，研究者不需要预先考虑蛋白质的可能结合位点。

ChIP-chip 作为一种新技术，也有一些缺陷需要改进。①成本较高，一般公开发表的运用 ChIP 方法的研究中都至少重复三次以保证其可靠性，其高昂的成本是一些实验室运用此方法的限制因素；对芯片实验得到的大量数据进行统计和分析，规范实验程序，制定确切的有生物学意义的实验标准等方面也尚待完善；②另一个限制因素是所能得到的 DNA 片段的大小，一般的超声裂解只能得到约 200bp 的片段，但为了得到高分辨率的结果，需要得到更小的甚至单核苷酸的片段；③与 ChIP 一样，ChIP-chip 也要求特异性很高的抗体，要在自由溶液和固定液中都能识别其抗原决定簇(epitope)。需特别注意的是，由于基因芯片是一个“封闭系统”，它只能检测已知序列的特征，所以 ChIP-chip 并不能获得一个特定 DNA 结合蛋白的全部可能结合位点。

三、ChIP-seq 技术

知识拓展

DNA 测序技术经历了至少三次革命性的更新换代。如果说从最初的放射性标记法到荧光标记法算作第一次革命的话，那么第二次革命就是从平板电泳仪到毛细管电泳仪的变迁。而第三次革命才刚刚开始，这是一个崭新的阶段，它以样品的微量化、操作的规模化和平台的多样化为标志，称之为新一代测序(next-generation sequencing)。此测序技术可以一次性测定 100Mb 以上的核苷酸，是最好的毛细管测序仪通量的 20~100 倍。2007 年，新一代测序技术因其在众多领域中的成功应用而得到广泛认可。

ChIP-seq 技术是伴随着新一代高通量测序技术的发展应运而生的。高通量测序技术是指一次可对几十万到几百万条 DNA 分子进行序列测定，是对传统测序的一次革命性改变。因此，在一些文献中称其为下一代测序技术(next -generation sequencing)。同时高通量测序使得对一个物种的转录组和基因组进行细致全面的分析成为可能，所以又被称为深度测序(deep sequencing)。ChIP-seq 就是在染色质免疫沉淀以后的 DNA 直接进行高通量测序(图 11-2)，对比基准基因(reference sequence)可以直接获得蛋白与 DNA 结合的位点信息。

ChIP-seq 技术是由 Steven J.M. Jones 及其所在团队于 2007 年率先提出的。他们使用 ChIP-seq

在人 HeLa S3 细胞中首次高通量鉴定了 γ- 干扰素刺激前后的 STAT1 转录因子结合位点，并就部分位点与 ChIP-PCR 和 ChIP-chip 方法进行了比较。结果显示，应用 ChIP-seq，共获得了刺激前后的序列数分别为 129 万和 151 万条，潜在的 STAT1 结合位点分别是 11 004 和 41 582 个。对于目前已知的 34 个 STAT1 结合位点，ChIP-seq 鉴定出 24 个，阳性率为 71%；与 ChIP-PCR 和 ChIP-chip 相比，ChIP-seq 的灵敏度在 70%~90% 之间，特异性至少为 95%。

与 ChIP-chip 相比，ChIP-seq 的优势主要在于：①它是一个“开放系统”，其寻找新的信息能力，从本质上高于芯片技术。可以检测更小的结合区段、未知的结合位点、结合位点内的突变情况和蛋白亲和力较低的区段；②成本低，周期短，收获的 DNA 可直接测序，省去了标记和杂交等步骤，并且无需多次重复实验，极大提高了工作效率；③分辨率可提高至 30~50bp。作为两个高通量的基因组学研究技术，在应用的某些方面存在重叠和竞争，但是在更多方面是优势互补，两种方法联合使用，将解决以前的单种技术难以解决的问题。有学者同时用 ChIP-chip 和 ChIP-seq 对 STAT1 的结合位点进行了检测，结果非常有趣，两种技术对于强阳性的区段具有非常好的相关性，但是对于一些弱的结合位点，ChIP-chip 和 ChIP-seq 都会丢失部分信息，其中一种方法丢失的信息又恰好能被另一种方法所检出，所以完整的数据是来自两部分的整合。

第三节 转录因子结合位点的信息学预测方法

Section 3 Prediction of Transcriptional Factor Binding sites

大量的实验证据表明，转录因子结合位点的长度一般在 6~12bp 之间。然而，ChIP-chip 技术的分辨率在 200~800bp 左右，远大于转录因子结合位点的长度，所以需用计算方法来确定转录因子结合位点的确切位置。与之相比，ChIP-seq 技术的分辨率可以达到 100bp 甚至更高。因此随着基因芯片和深度测序等高通量数据的出现，计算方法在转录因子结合位点的分析中得到了广泛的应用。对转录因子结合位点的计算研究可分为两类问题：第一类问题是通过收集可能被同一转录因子调控的基因启动子序列，在其中寻找具有统计显著性的短片段作为转录因子可能的结合位点，称之为转录因子结合位点的识别(identification of transcription factor binding site)。第二类问题是根据若干已知的转录因子结合位点的模体，在所研究基因的启动子区域内搜索相应转录因子可能的结合位点，称之为转录因子结合位点的定位(location of transcription factor binding site)。

一、转录因子结合位点的表示方法

1. 共性序列 转录因子结合位点最简单的表示方法是共性序列。不同基因的启动子区域中，同一转录因子的结合位点并不完全相同。可能与同一个转录因子结合的所有 DNA 片段按照对应位置进行排列(图 11-3)，在每个位置上选择最可能出现的碱基，组成了该转录因子结合位点的共性序列。共性序列中用 A、C、G、T 之外的字母来表示结合位点中各个位置上可能出现的碱基组合(表 11-1)，这些字母称为 IUPAC 简并码(IUPAC degenerate codes)。共性序列的表示方法简明易懂，却不能够反映每个位置上不同碱基出现的概率。

表 11-1 IUPAC 简并码

IUPAC code	Nucleotide	IUPAC code	Nucleotide
W	A or T	B	C, G or T
R	A or G	D	A, G or T
K	G or T	H	A, C or T
S	C or G	V	A, C or G
Y	C or T	N	A, C, G or T
M	A or C		

2. 位置频率矩阵(position frequency matrix, PFM) 相对于共有序列表示方法，位置频率矩阵可以反映出每个位置上不同碱基出现的概率。该模型的一个前提假设是各个位置上碱基出现的概率相互独立。矩阵每一列表示模体相应位置上四种碱基出现的概率。对于长度为 n 的模体，碱基 i(i={A, C, G, T})在模体第 j 个位置上出现的频率为 $q_{i,j}$，则整个模体用矩阵 M 表示如下：

$$M=\begin{Bmatrix} q_{A,1}\, q_{A,2}\cdots q_{A,n} \\ q_{C,1}\, q_{C,2}\cdots q_{C,n} \\ q_{G,1}\, q_{G,2}\cdots q_{G,n} \\ q_{T,1}\, q_{T,2}\cdots q_{T,n} \end{Bmatrix}$$

目前位置频率矩阵是在转录因子结合位点研究中应用最广泛的模型。

3. 序列标识图(sequence logo) 为了更加直观地区分结合位点中不同位置上的碱基倾向性及其在结合过程中的作用，人们提出了序列标识图(sequence logo)的概念(图 11-3)。序列标识图依次绘出模体中各个位置上出现的碱基，每个位置上所有碱基的累积可反映出该位置上碱基的一致性，每个碱基字母的大小与碱基在该位置上出现的频率成正比。这种表示方法直观地给出模体各个位置上碱基出现的倾向性和整个模体的序列的一致性，应用非常广泛。

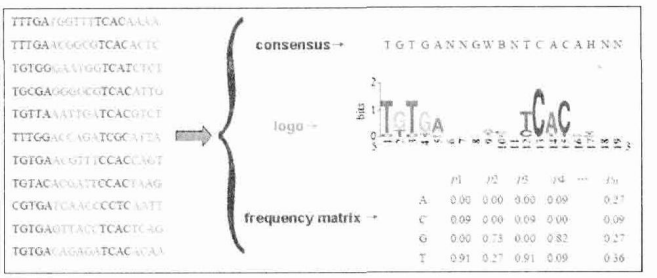


图 11-3 转录因子结合位点表示法

二、转录因子结合位点的识别

要鉴定出同一转录因子调控多个靶基因表达的共有结合位点，收集可能被同一转录因子调控的多基因序列是首要步骤。然后通过多种计算方法从不同角度或不同层面去进行计算、评估和分析，尽可能地屏蔽掉冗余序列和噪音序列，寻找出具有统计显著性的短片段，作为转录因子可能的结合位点；最后到相关转录因子数据库中查询以确定是什么转录因子。图 11-4 所示为转录结合位点识

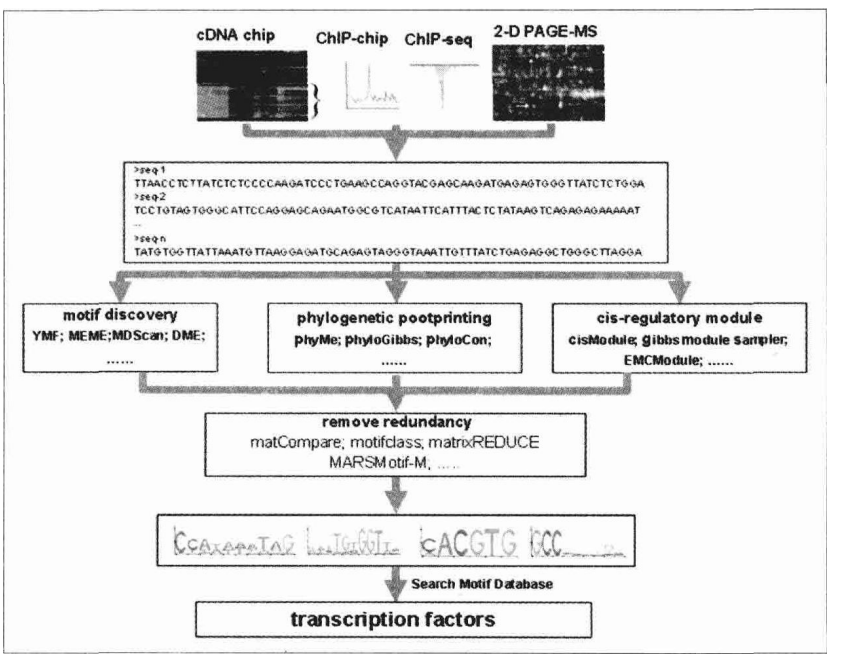


图 11-4 转录因子结合位点识别的工作流程

别的全部流程。表 11-2 是识别转录因子结合位点可利用的资源网站。

表 11-2 转录因子结合位点分析可利用网络资源

Category	Program	URL
Single motif discovery	MobyDick	http://genome.ucsf.edu/mobydick/
	YMF	http://bio.cs.washington.edu/software.html
	Consensus	http://ural.wustl.edu/software.html
	MEME	http://meme.sdsc.edu/meme/intro.html
	Gibbs Sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html
	MDScan	http://ai.stanford.edu/~xsliu/MDscan/
	DME	http://rulai.cshl.edu/software/index1.htm
	SISSR	http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/
Comparative genomics	PhyMe	http://edsc.rockefeller.edu/cgi-bin/phyme/download.pl
	PhyloGibbs	http://www.imsc.res.in/~rsidd/phylogibbs/
Cis-module analysis	CisModule	http://www.stat.ucla.edu/~zhou/CisModule/
	EMCModule	http://www.bios.unc.edu/~gupta/emcmodule.html
Regression methods	REDUCE	http://bussemaker.bio.columbia.edu:8080/reduce/
	MatrixREDUCE	http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/
	MotifRegressor	http://www.math.umass.edu/~conlon/mr.html
	MarsMotif-M	http://rulai.cshl.edu/software/index1.htm
Motif search Database	TRANSFAC	http://www.gene-regulation.com/
	Jaspar	http://jaspar.cgb.ki.se/
	DBTBS	http://dbtbs.hgc.jp/
	TRED	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home

（一）获取靶向序列

随着基因组学和蛋白质组学的迅速发展，获得被同一转录因子调控的多基因序列主要来自以下几个方面：

1. 从基因差异表达谱芯片数据出发获取启动子序列　一张基因芯片（microarray）可以同时检测数万个基因在某个组织样本中的表达值，对在不同条件下获得的基因芯片数据进行聚类分析（详见第七章第五节），可以得到一组或几组有相似表达模式的基因，提示这些基因很可能受到共同转录因子的调控。相同的转录因子在这些基因启动子区域上的结合位点应当是相同或者相似的。通过计算方法寻找这些相似的转录因子结合位点（模体），就可以完成转录因子结合位点的识别。找到一组共调控的基因之后，就可以利用 NCBI 上相关核酸数据库（详见第一章第二节）确定基因的启动子区。一般认为，转录因子结合位点主要在转录起始位点（transcription start site，TSS）附近出现，但还有一些转录因子结合在基因上游很远的区域（被称为远程作用）。根据研究问题的不同，启动子序列的长度可以取几百到几千个碱基不等，通常选取转录起始位点附近 1000~2000bp 的长度作为启动子区，例如，转录起始位点上游 1000bp 和下游 200bp。序列太短会丢失部分结合位点。如果序列取的过长，在包含了少量真实结合位点的同时，却引入了大量的背景噪声，使真正的转录因子结合位点淹没在噪声中无法区分。

2. 从差异表达蛋白质数据出发获取启动子序列　一个基因的转录表达，最终是通过翻译成蛋白质行使其功能的。配对样品（实验组和对照组）的双向电泳结合生物质谱分析，可以同时获得数十到数百个差异表达的蛋白。通过蛋白质组的功能分析（详见第九章），就可以得到一组或几组由同一信号通路调控的蛋白质，表明它们的转录可能被同一转录因子调控。基于此，利用所获得的蛋白质信息，就能从 SWISS-PROT 和 NCBI 等数据库中获得编码基因的启动子区。

3. 从 ChIP-chip 和 ChIP-seq 数据出发获得结合位点序列　与由基因芯片和功能相关获得的包含共同转录因子结合位点的启动子序列相比，ChIP-chip 和 ChIP-seq 确定的包含共同结合位点的区域更加准确，也更广泛；不仅在启动子区域，也会出现在内含子和 3' UTR，甚至间隔区。

（二）转录因子结合位点识别的计算方法

1. 单个模体预测算法　得到一组候选启动子序列后，可以直接利用计算方法寻找其中具有统计显著性的短片段作为转录因子可能的结合位点。根据不同的表示方法，转录因子结合位点识别算法总体上可以分为基于共有序列和基于位置频率矩阵两类。第一类是基于共有序列的识别方法，通过穷举所有可能的序列组合得到具有统计显著性的转录因子结合位点，例如 MobyDick 和 YMF 算法等。第二类是基于位置频率矩阵的识别方法，利用贪婪算法、期望最大化(expectation maximization，EM)或 Gibbs 采样方法(Gibbs sampling method)等得到转录因子结合位点对应的位置频率矩阵，这类方法包括 MEME 和 Gibbs Motif Sampler 等。MEME 基于 EM 算法，它的优点是具有较高的敏感度，但计算复杂度高，计算时间较长。Gibbs Motif Sampler 计算速度快，但需要多次重复实验才能得到稳定的结果。

这些单个模体预测算法在线虫、酵母等低等生物中得到了很好的结果，但是在高等生物中的应用却存在假阳性高的问题。这主要是由于真核生物的转录过程更加复杂，比如受染色质结构的影响，转录因子同基因的远程作用，以及转录因子之间的相互作用等。

2. 比较基因组学　由于转录因子结合位点功能上的重要性，结合位点所在区域的突变速度会慢于无功能序列。因此，比较基因组学在转录因子结合位点的识别中可以起到重要作用。随着多个真核生物全基因组测序的完成，人们可以通过比较基因组学的方法得到启动子序列在多个物种间的保守性，并将保守性信息同传统的方法相结合进行识别。比较基因组学在转录因子结合位点分析中的应用可以分为两类：一类先利用传统的方法进行模体识别，然后再检测得到的模体在不同物种中的保守性，筛除不保守的模体；另外一类是以候选启动子区及其在不同物种中的直系同源序列为输入，在识别过程中考虑不同物种间的保守性和模体的信号强度这两种因素。

遗传系谱印记法(phylogenetic footprinting)近几年在转录因子结合位点的分析中应用非常广泛，PhyME、PhyloGibbs 和 PhyloCon 都属于这一类方法。这类方法的基本假设是功能元件在进化过程中存在选择压力，因此它们进化速率要慢于周围的非功能序列。在识别过程可以利用遗传系谱印记法基于不同的物种间的进化关系寻找保守片段，然后利用传统方法进行模体识别。应用保守性时，物种间的进化距离不能太远。在漫长的进化过程中，DNA 序列可能发生大段的插入或删除，这种情况下转录因子结合位点不可能保守。另一方面，如果物种间的进化距离太近，大部分序列没有发生过变异，就无法把转录因子结合位点序列与无功能序列分开。例如，识别人类基因的转录因子结合位点时，选择小鼠的直系同源基因比较合适。

利用不同物种间的保守性信息来筛选真正的转录因子结合位点是很有效的方法。但是转录因子结合位点通常都是很短的序列，即使它们有重要的功能，在进化过程中还是有可能发生变异，研究中也发现了一些物种特异的转录因子结合位点。因此比较基因组学的应用存在着一定的局限性。

3. 顺式调控模块识别方法　许多真核生物基因的表达是由多个转录因子结合位点的组合所调控的，这些转录因子结合位点的特定组合被称为顺式调控模块(cis-regulatory module，CRM)。在顺式调控模块中，不同转录因子结合位点出现的个数和顺序以及它们之间相对位置都存在一定的规律，研究者不仅关心有哪些模体存在，还关心这些模体在结构上是如何组织的。因此，开发了从共调控基因的启动子序列识别调控模块的方法。目前的顺式调控模块识别方法可以分为两类：一类从已知的模体集合出发，看哪些模体组合在启动子序列中的出现频率明显高于其他组合。这类方法还可以细分为寻找只包含两个模体的调控模块和寻找包含多个(两个或两个以上)模体的调控模块。由于目前大部分转录调控因子的结合体还未知，使得上述基于已知模体集合的方法受到了很大的限制。另外一类直接从共调控基因的启动子序列识别调控模块的方法逐渐出现，如 CisModule、Gibbs

Module Sampler 和 EMCModule，这三种方法都可以寻找包含多个模体的顺式调控模块。

4. 基于启动子区重要性差异的识别算法 有时收集到的候选启动子序列并不是同样重要的，有些序列包含目标转录因子结合位点的可能性更大。为了充分利用这些可信度更高的序列，提出了对重要性不同的序列区别对待的方法(discriminative method)，包括 MDScan 和 DME(Distributed mutual exclusion)等。根据包含目标结合位点的可能性，MDScan 方法对候选启动子序列进行排序，将其分为“最可能”和“次可能”两组。MDScan 在“最可能”组内寻找富集的模体，再用“次可能”组对其进行校正。ChIP-chip 数据中，覆盖芯片上探针的信号强度反映了探针对应序列与转录因子结合的可能性。信号越强，这个区域与转录因子结合的可能性越大，信号强度自然成为可信度的标准。MDScan 方法对候选启动子序列区别对待，首先从高信号强度的序列中寻找模体，这在数据存在大量噪声的情况下增加了找到目标结合位点的可能性。MDScan 方法的计算速度快，但最后识别的模体中包含较多的简单重复序列，所以在应用该方法前要对候选启动子序列中的简单重复序列进行屏蔽。

DME 方法不区分候选集合内各条序列包含目标结合位点可能性的高低，而是另外构造一个由无关序列组成的背景集合，搜索候选集合相对于背景集合显著富集的模体。背景集合的构造可以在全基因组范围内随机选取启动子序列，或者根据已有知识选择一些不太可能包含目标转录因子结合位点的序列。在 ChIP-chip 数据的应用中，DME 方法以低结合强度序列为背景，搜索在高结合强度的序列中显著富集的模体。由于应用了背景序列，DME 方法可以避免识别的模体中出现简单重复序列的问题，但是该方法的缺点是计算复杂度较高，计算时间长。另外选择不同的背景集合对分析结果的影响较大，使用中可以尝试多种背景集合，找到其中较稳定的结果。

5. SISSR 算法 从上一节的论述中已经知道，目前为止，ChIP-seq 技术是鉴定转录因子结合位点分辨率最高的研究方法。通过 ChIP-Seq 读出的一段短小序列通常都能够与基准基因相对应。但也存在许多特殊情况，使得这样的一种对应关系出现错误，从而不能够精确地定位结合位点。美国国家健康中心的学者们最近提出了一种新颖的运算法则——短序列读数位点验证(Site Identification from Short Sequence Reads，SISSR)，很好地解决了上述问题。SISSR 首先计算出阅读序列的平均长度，然后结合片段长度、阅读方向、背景模型和其他一些控制参数，将结合位点的定位缩小到数十个碱基范围内，从而大大提高了基因定位的准确性。通过测试数个大型基因组的结果证实了该算法的灵敏度与准确性，该算法对 ChIP-Seq 技术的推广使用来说意义重大。

(三) 处理识别结果

1. 去除冗余及质量控制 利用转录因子结合位点的识别方法，可以得到一组在候选启动子序列中具有统计学显著性的模体。识别结果中通常会有很多相似的模体，所以需要对结果去除冗余。利用模体比对的方法可以将重复出现的相似模体进行归并。很多计算方法可以实现模体的比对。对基于共有序列表示的模体，可以直接比较共有序列是否匹配；对基于矩阵表示的模体，比对方法包括计算皮尔森相关系数(Pearson correlation coefficient)和平均对数似然比(average log-likelihood ratio)统计量；对几乎在所有启动子序列中都富集的模体，如 TATA box 或高 GC 含量的重复序列，这样的模体并非感兴趣的功能模体，应该去除。Motifclass 方法能够区分在所有启动子序列都频繁出现和只在特定基因集中富集的模体。将特定基因集的启动子序列作为前景集合，并构造一组与之对应的背景序列集合。背景集合的构造可以在全基因组范围内随机选取启动子序列，或根据已有知识选定一组与前景集合不同的序列。给定前景、背景和需要区分的模体集合，Motifclass 能够找到那些在前景集合内富集而在背景集合内相对较少的模体。被区分的模体可以是识别到的新模体，也可以是数据库中收录的已知模体。

得到非冗余的模体后，可以将这些新发现的模体与数据库中已知的模体进行比对，判断它们之间可能的对应关系。这些对应关系可以反映出选定的从共调控基因受哪些已知的转录因子调控。另外，它们也可以作为衡量识别结果的标准之一。

2. 通过回归分析寻找特定条件下起作用的模体　转录调控是一个时空特异过程。虽然在启动子区域上转录因子结合位点时刻存在，但只有在特定组织或发育阶段，转录因子与位点结合，才能促进或者抑制下游基因的转录。通过识别方法可以得到在候选启动子区显著富集的转录因子结合位点，但是却无法判断它们在什么情况下起作用。为了找到在特定条件下起作用的模体，近年来出现了一些基于回归的算法。这类算法以转录因子结合位点的出现情况作为自变量，以基因芯片上的基因表达值作为因变量来进行回归。它们的基本假设是：基因的表达值都是由其启动子区的转录信号决定的，其中一种信号就是转录调控因子与其对应位点的结合。某一特定条件下，基因芯片记录的基因表达值反映了在该条件下转录因子与其结合位点间的作用，可以用于发现转录因子结合位点起作用的时空条件。目前所用的回归算法主要有 REDUCE 算法、MatrixREDUCE 算法和 MARSMotif-M 算法。REDUCE 算法以模体出现的次数作为自变量来进行简单线性回归，对比方法进行改进，提出了用位置频率矩阵的打分作为自变量进行回归的 MatrixREDUCE 方法。这两种方法都是针对一定长度，遍历所有可能的序列或位置频率矩阵，所以模体的长度和计算的速度都受到限制。上述算法应用在酵母数据中，取得了很好的结果。高等生物的调控更加复杂，简单的线性模型往往不能很好地反映其中的调控关系，所以又有人提出了基于多变量适应回归模型(multiple adaptive regression splines，MARS)的 MARSMotif-M，将该方法应用在人类基因表达数据中，取得了较好的结果。

三、转录因子结合位点的定位

上面主要介绍如何从可能被共同转录因子结合的基因启动子序列中寻找相似的转录因子结合位点的问题，而有些研究中人们只关心某一个或几个基因受哪些已知的转录因子调控，这种情况下可以搜索目标基因的启动子区中包含哪些已知的转录因子结合位点。这类根据已知的转录因子结合位点模体搜索其可能结合位点的问题，称为转录因子结合位点的定位。

（一）转录因子结合位点定位的计算方法

对任一长度为 n 的已知模体位置频率矩阵 M，转录因子结合位点定位就是判断某一长度为 n 的序列片段与 M 的匹配程度。考虑到 DNA 序列本身有可能存在碱基组成上的偏向性，通常把位置频率矩阵转换为位置权重矩阵(position weight matrix，PWM)，用位置权重矩阵的打分来衡量模体与任意给定序列的匹配程度。在位置权重矩阵中，引入碱基 $i(i=\{A, C, G, T\})$在背景序列中出现的频率记为 b_i 来消除 DNA 序列本身碱基组成偏向性的影响。位置权重矩阵中的每一个元素记为 $S_{i,j}$：

$$\boldsymbol{S}_{i,j}=\log\left(\frac{q_{i,j}}{b_i}\right) \qquad \text{式 11-1}$$

这里，$q_{i,j}$ 是碱基 i 在模体中第 j 个位置处出现的频率。

则 M 被转换为的位置权重矩阵 S 为：

$$\boldsymbol{S}=\begin{Bmatrix} S_{A,1}\, S_{A,2}\cdots S_{A,n} \\ S_{C,1}\, S_{C,2}\cdots S_{C,n} \\ S_{G,1}\, S_{G,2}\cdots S_{G,n} \\ S_{T,1}\, S_{T,2}\cdots S_{T,n} \end{Bmatrix}$$

对于长度为 n 的 DNA 序列片段，它作为模体 M 对应的转录因子结合位点的打分为：

$$\boldsymbol{S}_{i,j}=\sum_{j=1}^{n} St_j, j \qquad \text{式 11-2}$$

其中，t_j 表示相应序列第 j 个位置上出现的碱基。给定阈值 T，如果序列片段由上式给出的打分 $S \geq T$，则认为它有可能是相应转录因子的结合位点。在实际应用中，对长度为 L 的区域，用一个长度为 n

的窗口在序列上滑动，每次步长为 1，遍历所有长度为 n 的片段。选出对数似然比打分高于阈值 T 的那些片段，则为可能的结合位点(图 11-5)。

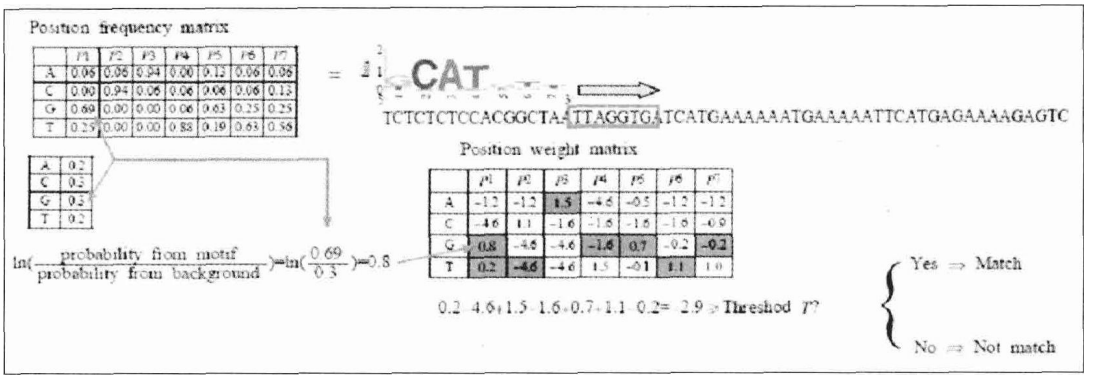


图 11-5 应用位置权重矩阵预测潜在结合位点

在应用中一个重要的问题是如何选择阈值 T。如果 T 很高，大部分片段都不符合打分高于阈值 T 的要求，这时可以避免高假阳性的出现，但同时也会丢掉很多真的转录因子结合位点；如果 T 很低，在包含更多真实转录因子结合位点的同时就会引入假阳性结果。实际应用中可以根据研究问题的需要，在高检出率和低假阳性率之间取一个折中。阈值 T 的选择可以根据打分 S 的统计显著性，其中最具代表性的是 Staden 方法。对于特定的模体 M 和打分 S，Staden 方法可以计算随机位点的打分大于等于 S 的概率，概率越低说明 S 的统计显著性越强。数据库 TRANSFAC 的 MATCH 算法提供了基于大规模采样的阈值选取规则。在其基础上，TRANSFAC 数据库还开发了一种结合序列搜索和对数似然比打分方法的 P-Match 算法。

在转录因子结合位点的定位问题中，一个值得注意的问题是分析结果的假阳性率较高。转录因子结合位点通常只有几个到十几个碱基长，这么短的序列片段在基因组上随机出现的概率很大，况且转录因子对其结合位点的识别并不要求完全匹配，这些因素都导致了转录因子结合位点定位问题中假阳性率较高。如何结合具体的生物问题降低转录因子结合位点定位的假阳性，是值得研究的重要问题。

(二) 转录因子结合位点定位的预测

目前，一些生物学数据库，收录了大量转录因子的结合位点及其位置频率矩阵，这些信息为定位问题提供了可能。TRANSFAC、JASPAR、TRED 和 DBTSS 等是其中最常用的数据库(详见本章第四节)，其收录的数据都是经过实验验证的。基于这些数据资源开发的预测转录因子结合位点定位的相应软件，方便了研究者使用。

1. TRANSFAC 众所周知，TRANSFAC 是世界上著名的转录因子数据库(详见本章第四节)，同时它也包括了多种转录因子及其结合位点的预测工具，如 AliBaba、P-Match、Patch 和 MatrixCatch 等。在这些工具的主界面上，都有对其计算方法和各项参数含义的详细描述，可随时查阅。进入 TRANSFAC 主页后(http://www.gene-regulation.com)，首先要进行注册登记(免费)，然后点击“programs”超链接进入工具栏，接着就可以根据自己的目的选择各种工具进行序列分析。现就这些工具的主要特点和使用流程简述如下：

(1) AliBaba2.1：AliBaba 2.1 是一个预测未知 DNA 序列中转录因子结合位点的商业性程序。2000 年由 Niels Grabe 借助于 TRANSFAC 4.0 数据库中所收集的结合位点编写而成。AliBaba 2.1 是目前预测结合位点最特异的工具之一，可以在线使用，但需预先注册(免费)。在其主界面有超链接“Documentation”，点击进入后，可了解 AliBaba 2.1 的详细使用信息，如各项参数的含义与设定等。图 11-6 为 AliBaba 2.1 的操作流程。

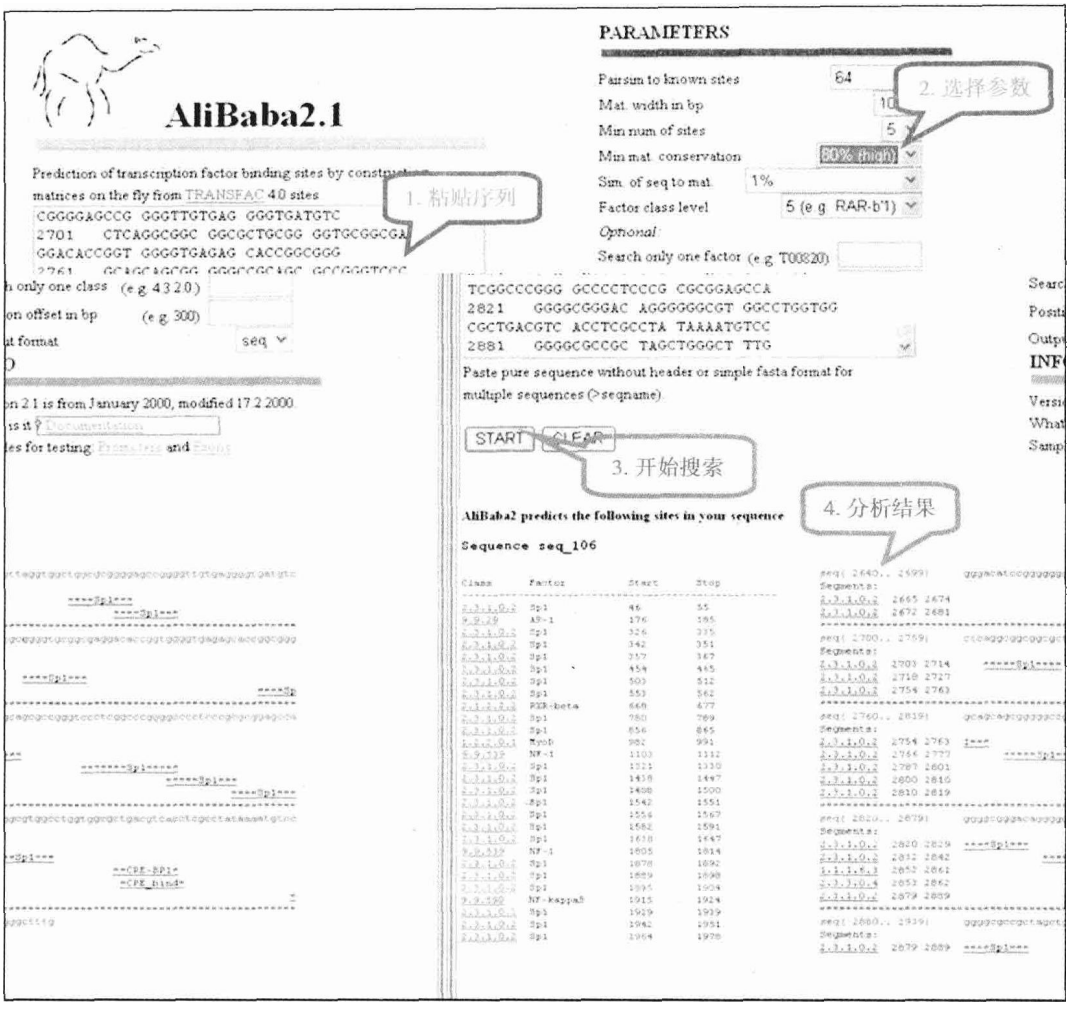


图 11-6 AliBaba 2.1 程序运行流程图

（2）P-Match：P-Match-Public 1.0 Public 是由 Dmitry Chekmenev、Carla Haid 和 Alexander Kel 三人联合建立的鉴定 DNA 序列中转录因子结合位点的新工具。P-Match 综合了模式匹配(pattern matching)和权重矩阵策略两种方法，大大提高了识别结合位点的准确性。P-Match 是使用来自 TRANSFAC 6.0 中收集的单核苷酸权重矩阵以及与这些矩阵相关的位置排列(site alignment)编写而成的。P-Match 不仅可以直接搜索转录因子结合位点，而且还可以针对特定的组织或器官(如肌肉组织和肝脏等)，对特异转录因子表达模式进行限定，即只搜索在特定组织或器官表达的转录因子，使输出的结果更集中。另外，P-Match 还可以建立、编辑和删除自己感兴趣的某种细胞、组织或器官的转录因子结合位点矩阵模式。P-Match 可以免费在线使用，并有“Help”菜单进行详细注释。图 11-7 为 P-Match- Public 1.0 Public 的操作流程。

（3）Patch：Patch 1.0 是利用模式匹配方法寻找感兴趣序列中潜在的转录因子结合位点的一种工具，是由 Jochen Striepe 和 Ellen Goessling 共同建立的。所采集的数据来自 TRANSFAC 专业数据库中的转录因子结合位点和 TRANSFAC 专业版的权重矩阵共有序列。Patch 1.0 可以免费在线使用。图 11-8 为 Patch 1.0 的操作流程。

（4）MatrixCatch：MatrixCatch 2.7 工具是 gor Deyneko 和 Alexander Kel 为了在 DNA 序列中寻找潜在的转录因子复合元件(composite elements，CE)而设计的。MatrixCatch 所使用的 CE 矩阵模型(CE matrix model)程序库是在 TRANSCOMPEL 数据库中收集的复合元件以及 TRANSFAC 6.0 公共数据库中的单核苷酸权重矩阵基础上建立的。MatrixCatch 2.7 可以免费在线使用，并有“Help”菜单

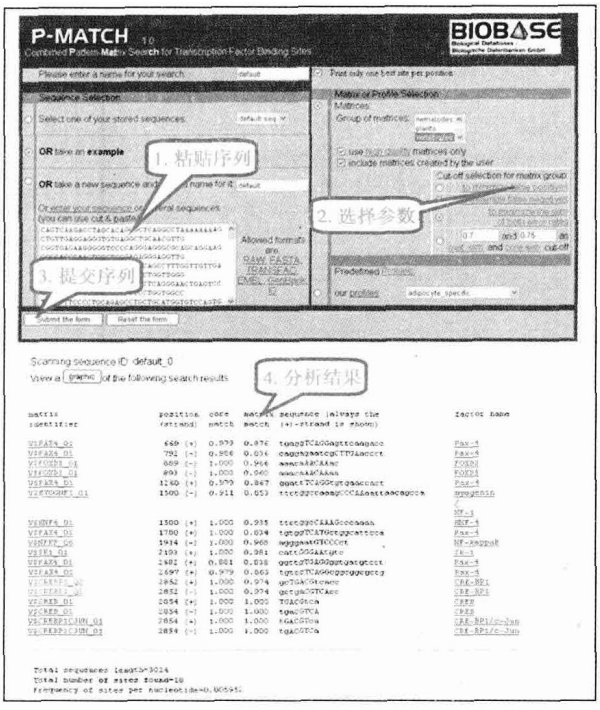


图 11-7 P-Match- Public 1.0 Public 程序运行流程图

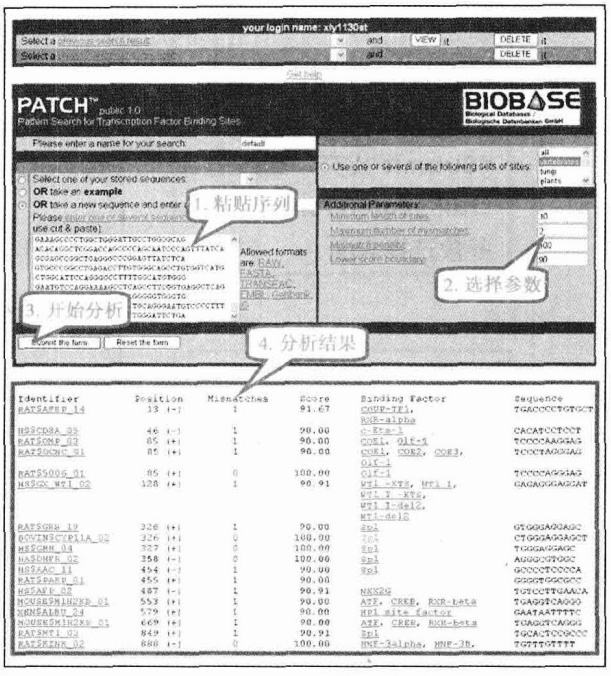


图 11-8 Patch 1.0 程序运行流程图

进行详细注释。图 11-9 为 MatrixCatch2.7 的操作流程。

2. TESS(transcription element search system)工具 TESS 是直接预测 DNA 序列中转录因子结合位点的在线工具(http://www.cbil.upenn.edu/cgi-bin/tess),它是由宾夕法尼亚大学医学院计算生物学与信息学实验室的 Jonathan Schug 和 G. Christian Overton 于 1997 年建立,并于 1998 年投入使用。运行十几年来,已更新多次,最近一次是 2007 年 2 月。TESS 工具的主要特点是:①使用 TRANSFAC、JASPAR、IMD 和 CBIL-GibbsMat 数据库中的位点或共有字符串和位置权重矩阵鉴定转录因子结合位点;②界面简单明了,可在主界面的对话框中直接粘贴序列,进行快速搜索,也可以通过设定各种参数后,再进行精确搜索;③既可以使用自己克隆的序列,也可以是感兴趣的基因组序列中获取;既可以一次输入一个序列(≤2000bp),也可以输入多个序列(总长度≤100 000bp)进行 TESS 搜索;④结果输出格式多样,可以直接把搜索结果发到使用者电子信箱中,也可以直接下载以表格形式输出。不足之处是运算速度慢,在线数据保留时间短(最长一个月);预测的转录因子结合位点太多,给后续的实验验证带来困难。TESS 的操作流程如图 11-10 所示。

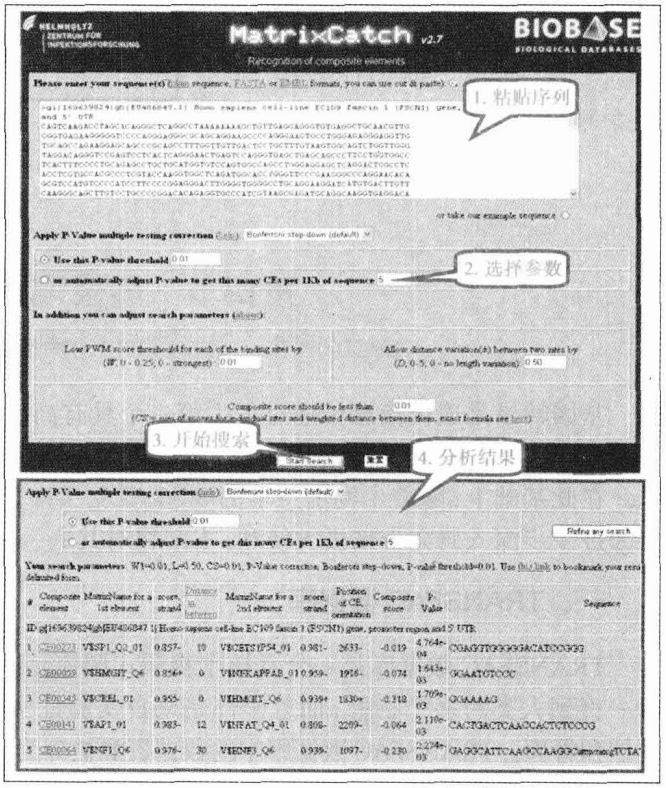


图 11-9 MatrixCatch 2.7 程序运行流程图

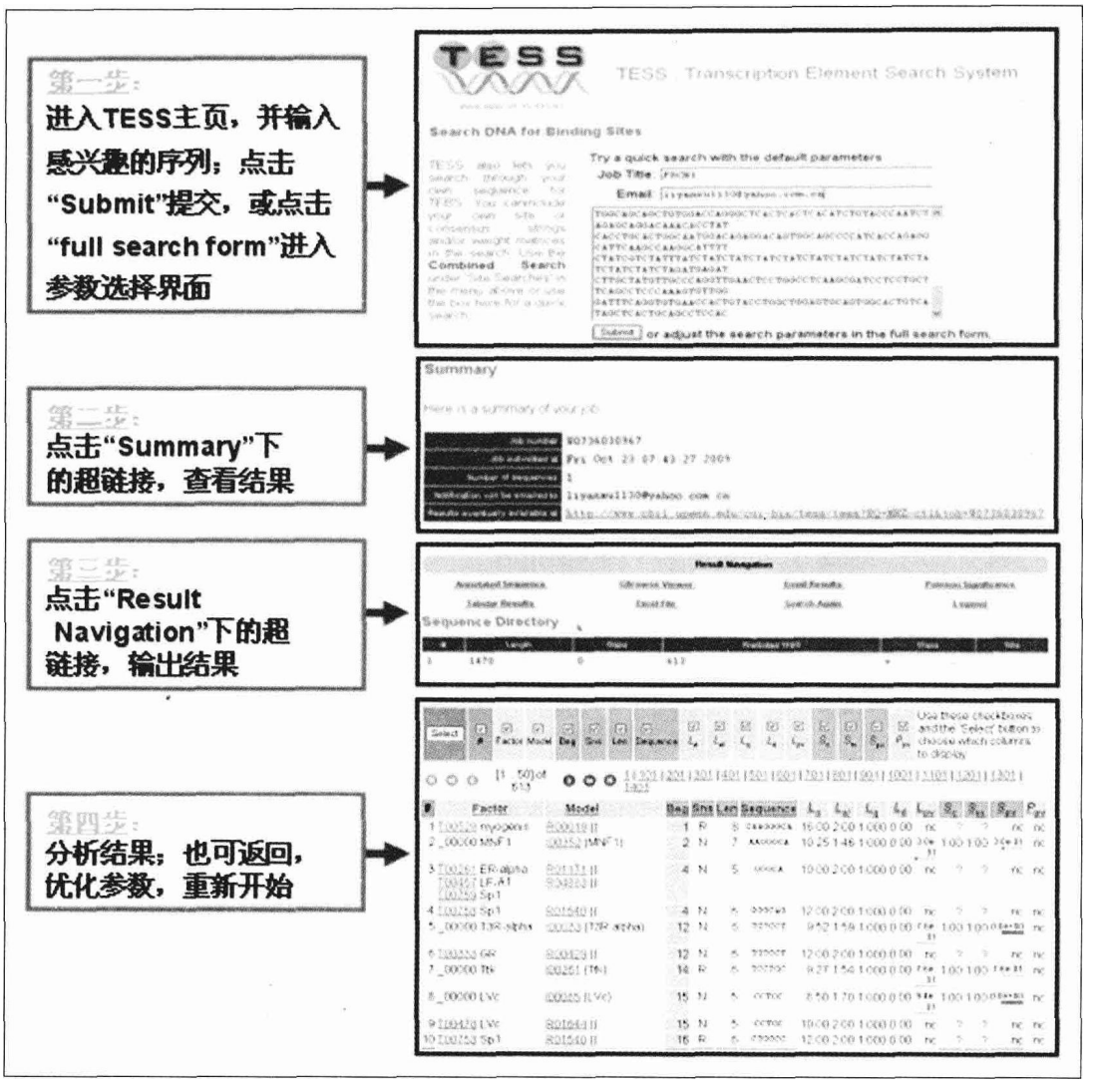


图 11-10 TESS 工具运行流程图

第四节 转录调控相关数据库

Section 4 Transcriptional Regulation Databases

基因转录是遗传信息传递过程中第一个具有高度选择性的环节，对基因转录调节的研究一直是基因分子生物学的研究中心和热点，发表了大量实验数据，也增强了人们对基因转录调控机制的认识。在此基础上，随着生物信息学的崛起，建立了许多很有价值的数据库资源，对这些数据库的了解将为进一步深入研究带来极大便利，本节将对其中一些数据库进行简要介绍。

一、TRANSFAC 数据库

TRANSFAC 数据库是一个真核生物顺式调控元件和反式作用因子数据库，数据搜集的对象从酵母一直到人类跨多个物种，而且全部是来自实验证实的数据。TRANSFAC 目前由 BIOBASE GmbH 维护和管理。TRANSFAC 数据库中的数据资源被分为六大数据表，可根据自己的研究需要进行使用和搜索。这六大数据表包括：①FACTOR 描述了结合到转录因子结合位点的蛋白信息；②SITE 给出了转录因子结合位点的相关信息；③MATRIX 给出了转录因子结合位点的核苷酸矩阵信息；④GENE

给出了转录因子的基因信息；⑤CLASS 包含了转录因子背景介绍及其所属类别；⑥CELL 给出了转录因子与其结合位点发生相互作用时的细胞信息。

TRANSFAC 数据库分为学术(免费)和专业(付费)两个版本。自 1996 年发布第一版以来已作了多次更新，目前学术版已更新至 TRANSFAC 7.0(2006 年)，所收集的数据信息详见表 11-3；专业版本已升至 TRANSFAC 2008.3(2008 年)。与学术版相比，专业版的信息量更多，不仅包含了全部学术版的内容，而且还增加了 ChIP-chip 的数据，2009 年 12 月还增加了 ChIP-seq 的数据。

表 11-3 TRANSFAC 7.0 数据库收集的数据

Table	TRANSFAC_7.0
FACTOR	6133
其中：	
Homo sapiens(人类)	1040
Mus musculus(小鼠)	765
D.melanogaster(黑腹果蝇)	233
A.thaliana(拟南芥)	1751
S.cerevisiae(啤酒酵母)	368
SITE	7915
MATRIX	398
GENE(all entries)	2397
其中：	
H.sapiens	608
M.musculus	417
D.melanogaster	145
A.thaliana	115
S.cerevisiae	195
GENE(entries with SITE links)	1504
CLASS	50
CELL	1307

TRANSFAC 学术版及其相关数据库可以免费检索和查询(http://www.gene-regulation.com/pub/databases.html)。使用时，如同本章第三节所述进入主页后，首先注册登记，然后点击进入“Database”主页，选择 TRANSFAC7.0 下设的超链接“Search”即可进入 TRANSFAC7.0 主界面。值得注意的是，上述六个数据表是相互独立的(除部分 GENE 和 SITE 有交叉)，所以存在冗余现象，应予以筛选，与此同时，同一个转录因子在不同数据表中又可以获得不同的信息描述，起到相互补充的作用。图 11-11 显示的是用同一个转录因子“Sp1”在六个数据表中搜索的部分结果，从中也可以看出各数据表中的 ID 号等设置完全不同。

此外，还有几个与 TRANSFAC 密切相关的扩展库：①TRANSPATH 数据库提供了参与信号转导通路的信号分子数据和它们所介导的反应，以及最终形成的相互作用组分之间的复杂网络关系。其着重强调的是在一定的细胞环境下信号转导级联放大所导致的转录因子活性及其下游基因表达谱的变化。②TRANSCompel 数据库收集了影响真核基因转录的复合调控元件。③PathoDB 数据库收集了与病理现象有关的转录因子及其结合位点的突变形式，它主要由已确证能引起病理障碍的缺陷型转录因子或突变的转录因子结合位点组成。④S/MARt DB 数据库呈现了真核基因组上的核骨架或核基质结合区以及与之相结合的蛋白质相关信息。⑤Cytomer 数据库显示了人类和小鼠转录因子在各个器官、细胞类型、生理系统和发育时期的表达状况。

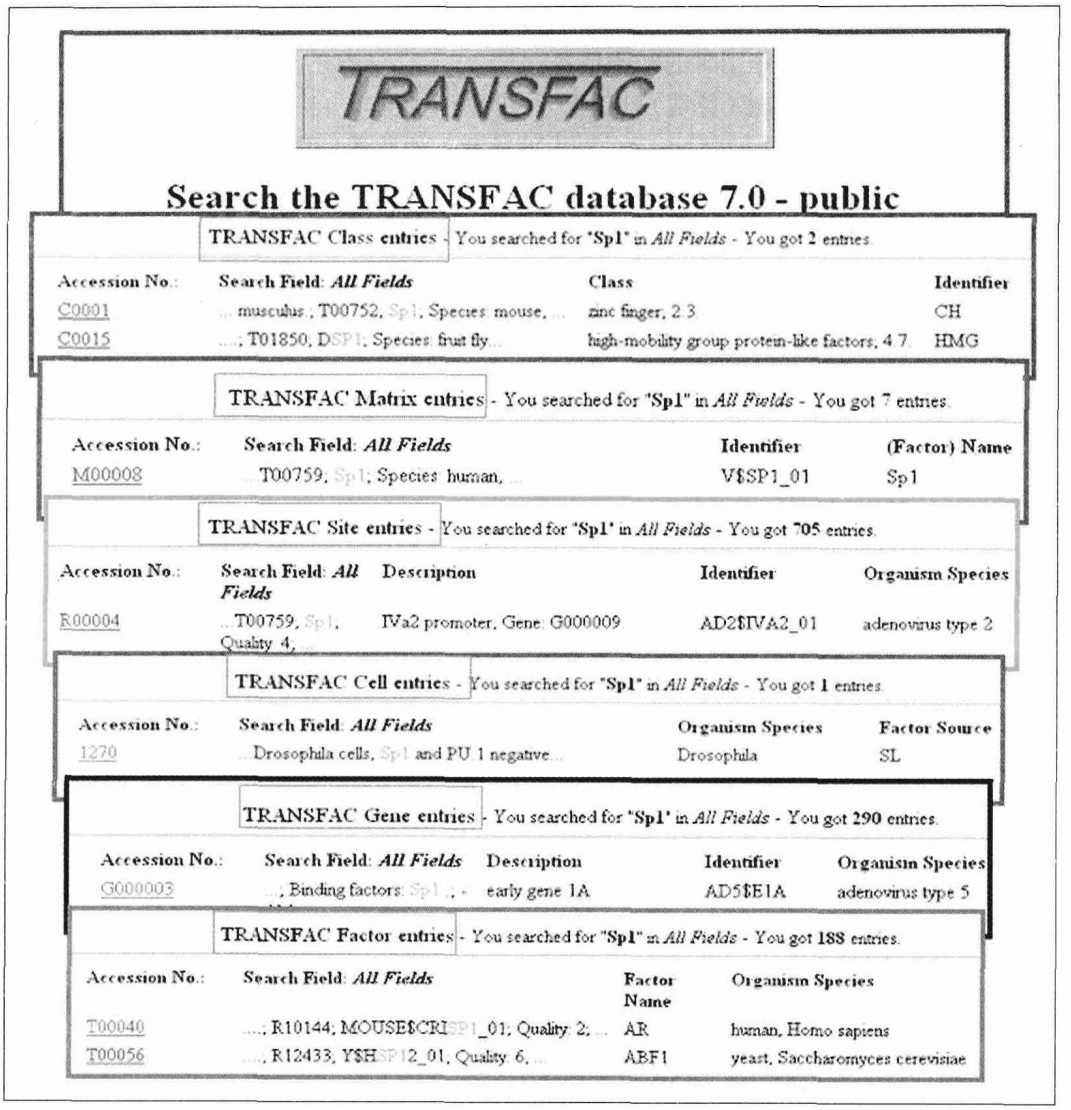


图 11-11 TRANSFAC 数据库的应用

二、JASPAR 数据库

JASPAR 是收集有注释的、高质量的多细胞真核生物转录因子结合部位的开放数据库(表 11-4)。所有序列均来源于通过实验方法证实能结合转录因子，而且通过严格筛选的序列，再通过模体(motif)识别软件 ANN-Spec 进行联配。ANN-Spec 是利用人工神经网络和吉布斯(Gibbs)取样算法寻找特征序列模式的软件(频数矩阵)。联配后的序列再利用生物学知识进行注释。

表 11-4 JASPAR 数据库的特点

数据库名称	特点
JASPAR CORE	高质量，非冗余的转录因子数据库，收录了 460 个序列模式，用于寻找特异转录因子模型或其结构类型
JASPAR FAM	包含 11 种转录因子结构类型的模型，用于搜索未知基因组序列某一转录因子家族的共有模式和鉴定新模式的分类
JASPAR PHYLOFACTS	由 174 种系统发育中保守的基因上游调控元件组成，用于分析启动子的组织特异性
JASPAR POLII	保存了 13 种与 RNA 聚合酶 II 核心启动子连接的 DNA 模型，用于分析潜在的核心启动子

续表

数据库名称	特点
JASPAR CNE	收集了 233 个人类保守的非编码元件，但是其生化和生物学功能尚不清楚，用于分析潜在的增强子
JASPAR SPLICE	包含有 6 种人类高度可靠的经典和非经典剪切位点的矩阵模式。用于分析剪切位点和选择性剪切
JASPAR PBM	保存有 104 种小鼠转录因子矩阵模式
JASPAR PBM HOMEO	保存有 176 种小鼠同源结构域矩阵模式
JASPAR PBM HLH	保存有 19 种线虫碱性螺旋环螺旋(bHLH)转录因子模型

2004 年创建以来，JASPAR 已多次进行更新。除了包括最常用的 JASPAR CORE 数据库外，还包括一些与转录调控相关的扩展数据库。这些数据库的特点比较见表 11-3。从中可以看出，尽管各数据库容量不大，但是各有特色。与相似领域数据库相比，JASPAR CORE 数据库具有很明显的优势：①它是一个非冗余的可靠的转录因子结合部位序列模式数据库；②数据的获取不受限制；③功能强大且有相关的软件工具使用。JASPAR 与 TRANSFAC 有较明显的差异，后者收录的数据更广泛，但包含不少冗余信息且序列模式的质量参差不齐，是商业数据库，只有一部分可以免费使用。

JASPAR 数据库所有内容可到主页下载(http://jaspar.cgb.ki.se)。通过主页界面(图 11-12)，用户可进行下列操作：①通过用转录因子 ID 号、物种(species)、转录因子家族(class)或种群(Taxonomic group，仅指脊椎动物，昆虫，植物和脊索动物)浏览转录因子结合的序列模式；②通过矩阵(matrix)或 IUPAC 字符串(IUPAC string)以及转录因子名称(name)等搜索序列模式；③将用户提交的序列模式与数据库中的进行比较；④利用选定的转录因子搜索特定的核苷酸序列。用户可到 ConSite 服务器(http://www.phylofoot.org/ consite)进行更复杂的查询。JASPAR 其他数据库的使用与此相类似。

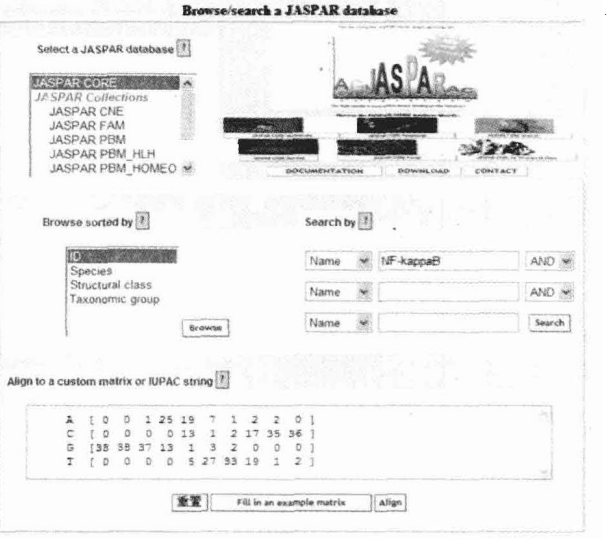


图 11-12 JASPAR 数据库主页

三、TRED 数据库

TRED(transcriptional regulatory element database)数据库是美国冷泉港实验室于 2007 年建立的。TRED 收集了哺乳动物的转录调控元件的数据库，对人、小鼠、大鼠等物种的启动子区域有相对完整的注释。其启动子数据主要来自某些数据库如 GenBank，EPD 和 DBTSS 等中的已知数据，并通过使用启动子发现程序 FirstEF 和 mRNA/EST 的信息以及物种交叉比较对这些数据进行了测评。TRED 数据库还采用人工操作进一步确认了所收集数据的准确性，每一个启动子的注释都有可靠的证据支持。TRED 的网址是 http://rulai.cshl.edu/TRED。通过主页界面，用户可以进行下列操作：①浏览数据库的全部内容及所涉及的人、小鼠和大鼠物种转录因子及其靶基因、启动子结合模体的总数(表 11-5)；②搜寻启动子，检索转录因子靶基因及其结合模体，特别是针对与肿瘤发生和发展相关的 36 个转录因子有详细注释(表 11-6)；③检索 36 个肿瘤相关转录因子的调控网络。图 11-13 显示了转录因子雌激素受体(ER)分别在人、小鼠和大鼠中的调控网络。此外，TRED 还提供了多个与其他相关网站的链接，如上节提到 MEME 和 Gibbs Samper 算法。

表 11-5 TRED 数据库统计表

相关数据	人类	小鼠	大鼠
版本	hg15: UCSC Human GoldenPath Apr. 03	mm3: UCSC Mouse GoldenPath Feb. 03	rn2: UCSC Rat GoldenPath Jan. 03
基因数	30 981	31 683	26 064
启动子数	58 229	50 764	30 386
转录因子有效靶点	3409 个基因，9085 个启动子，1249 个结合模体	1126 个基因，3089 个启动子，366 个结合模体	461 个基因，1132 个启动子，150 个结合模体
同源组数(两种或三种)		23 471	

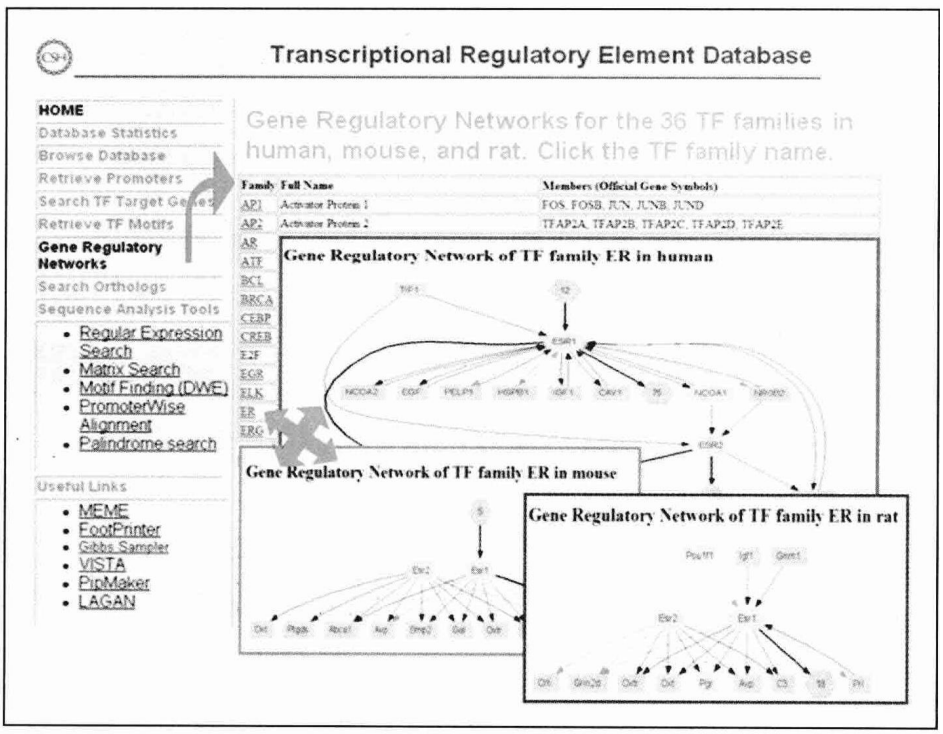


图 11-13 36 个转录因子家族的基因转录调控网络典型页面

表 11-6 与肿瘤相关的 36 个转录因子家族成员所靶向的启动子 / 基因数

转录因子家族	人类	小鼠	大鼠
AP1(Activator Protein 1)	432/383	217/190	157/143
AP2(Activator Protein 2)	338/318	123/123	90/86
AR(Androgen Receptor)	69/49	19/19	24/15
ATF(Activating Transcription Factor)	189/173	59/59	26/26
BCL(B-cell CLL/lymphoma)	21/19	15/15	0/0
BRCA(breast cancer susceptibility protein)	20/20	4/4	0/0
CEBP(CCAAT/enhancer binding protein)	335/325	152/134	241/179
CREB(cAMP responsive element binding protein)	224/220	138/133	95/93
E2F(E2F transcription factor)	1593/1329	141/127	11/11
EGR(early growth response protein)	120/111	67/55	33/26

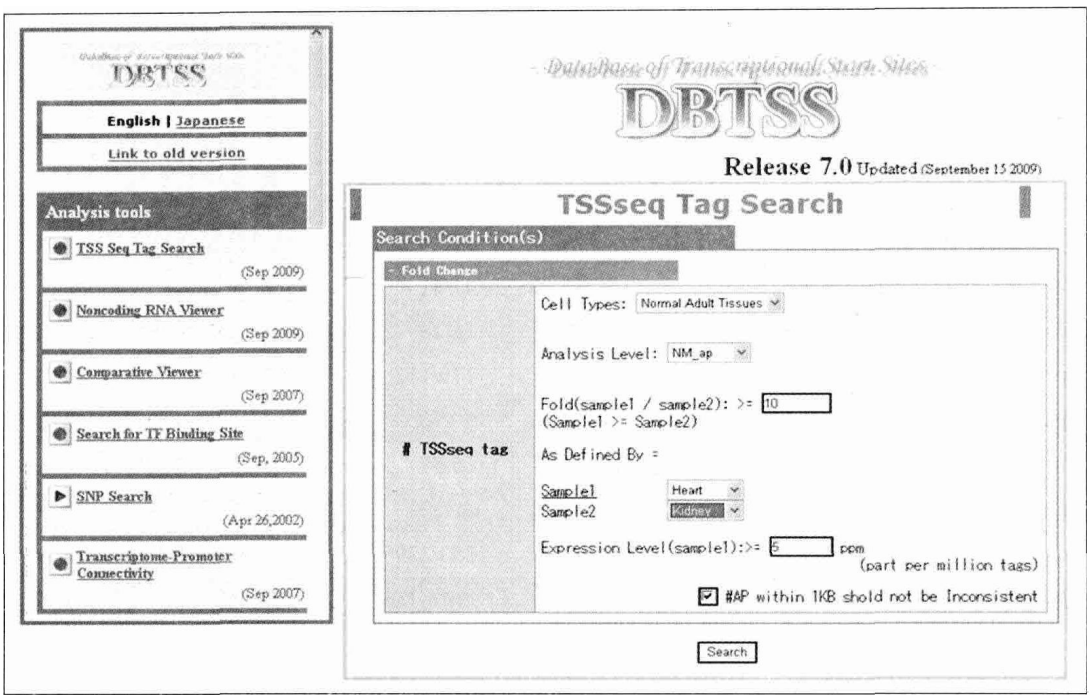
续表

转录因子家族	人类	小鼠	大鼠
ELK(member of ETS oncogene family)	47/41	15/13	6/6
ER(Estrogen Receptor)	169/152	40/39	32/31
ERG(ets-related gene)	21/21	5/5	0/0
ETS(ETS-domain transcription factor)	445/412	207/196	51/51
FLI1(friend leukemia integration site1)	41/41	17/16	0/0
GLI(glioma-associated oncogene homolog)	16/16	8/8	0/0
HIF(Hypoxia-inducible factor)	119/112	63/60	29/29
HLF(hepatic leukemia factor)	10/10	5/5	2/2
HOX(homeobox gene)	65/57	93/81	5/5
LEF(lymphoid enhancing factor)	40/33	26/23	5/5
MYB(myeloblastosis oncogene)	253/239	40/40	6/6
MYC(myelocytomatosis viral oncogene homolog)	2676/785	108/38	128/62
NFI(nuclear factor I; CCAAT-binding transcription factor)	136/127	75/62	73/65
NFKB(Nuclear factor kappa B, reticuloendotheliosis oncogene)	445/396	202/181	87/87
OCT(Octamer binding proteins)	232/195	123/108	34/34
p53(P53 family)	337/313	135/130	32/30
PAX(paired box gene)	52/47	76/61	13/11
PPAR(Peroxisome proliferator-activated receptor)	149/149	125/124	88/84
PR(Progesterone Receptor)	31/27	14/14	10/10
RAR(retinoic acid receptor)	233/218	71/71	40/40
SMAD(Mothers Against Decapentaplegic homolog)	139/130	76/75	17/17
SP(sequence-specific transcription factor)	655/515	296/263	235/220
STAT(signal transducer and activator of transcription)	245/218	111/106	48/46
TAL1(T-cell acute lymphocytic leukemia-1 protein)	15/14	9/6	0/0
USF(upstream stimulatory factor)	235/215	94/91	72/62
WT1(Wilms tumor 1, zinc finger protein)	78/49	16/16	8/8

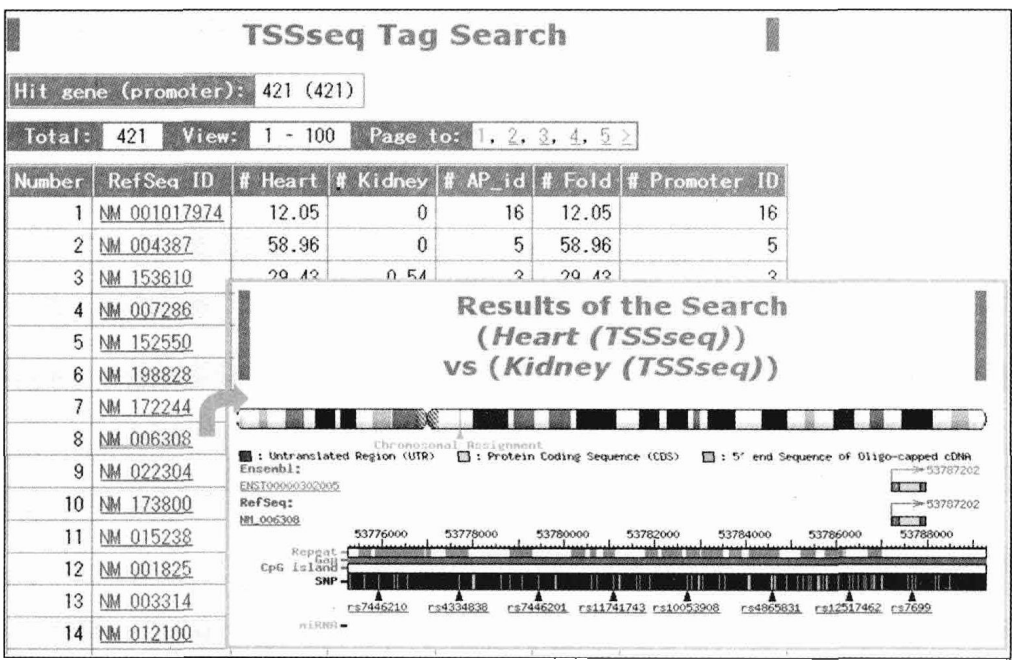
四、DBTSS 数据库

DBTSS(dataBase of transcriptional start sites)由东京大学人类基因组中心维护，网址是 http://dbtss.hgc.jp。最初该数据库收集用实验方法得到的人类基因的转录起始位点(transcriptional start site，TSS)数据。对 TSS 的确切了解具有非常重要的意义，能够更准确的预测翻译起始位点，可用于搜索决定 TSS 的核苷酸序列，而且可更精确地分析上游调控区域(如启动子)，见图 11-14。

自 2002 年发布第一版以来已作了多次更新，2009 年 9 月已是 DBTSS 7.0 版本。目前包含的 TSS 标签已达到 3.28×10^8 个，涵盖了人和小鼠的全长 cDNA。这些 TSS 标签是从 33 种不同的细胞类型或培养条件下获得的。在其主页面用户可以使用下列分析工具(这里仅列举与转录调控相关的工具)：①从来源于同一组织或细胞的一对样本中搜寻差异变化的 TSS 标签；②查看非编码 RNA 序列，目前已收集 2311 个非编码 RNA 序列，并且以图的形式显示某一非编码 RNA 在基因组 DNA 上的 TSS；③通过与 TRANSFAC 数据库链接搜索转录因子结合位点；④对 MCF7(乳腺癌细胞)或 HEK293(人胚肾上皮细胞)细胞不同药物处理后基因转录与启动子活性关联性的比较。



A



B

图 11-14 DBTSS 主页(A)及 TSS 搜索结果(B)

五、TRRD 数据库

TRRD(The Transcription Regulatory Regions Database)由西伯利亚分校细胞与遗传学研究所于1993 年创建，网址是 http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/。其数据来源于已发表的科学论文，每一个 TRRD 的条目里包含了特定基因的各种结构 - 功能特性，包括转录因子结合位点、启动子、增强子、沉默子的位置以及基因表达调控模式等。目前 TRRD 最新的版本为 7.0(2005 年)，综合了7609 篇科学文献中的 2344 个基因，10 135 个转录因子结合位点，3490 个调控区域(包括启动子、增

强子和沉默子等），14个座位控制区（Locus control regions）和14 407个表达模式。在TRRD数据库中，所有信息被分列于五个相关的数据表中（图 11-15）：①TRRDGENES，包含所有 TRRD 收录的基因及其调节单元的基本信息；②TRRDSITES，包含转录因子结合位点的详细信息；③TRRDFACTORS，包含 TRRD 收录的转录因子的详细描述；④TRRDEXP，包含对基因表达谱的详细描述；⑤TRRDBIB，包含 TRRD 所有注释涉及的参考文献。TRRD 收录的基因根据种属特异性、基因编码的蛋白质的类型以及基因的功能等进行分类。TRRD 的主页提供了对这几个数据表的检索服务。除此之外，数据库还提供了另外两个工具：①序列获得系统 SRS，用于搜索 TRRD 及与外部信息和软件资源进行整合；②TRRD viewer，以基因图谱的形式提供相关信息的描述。

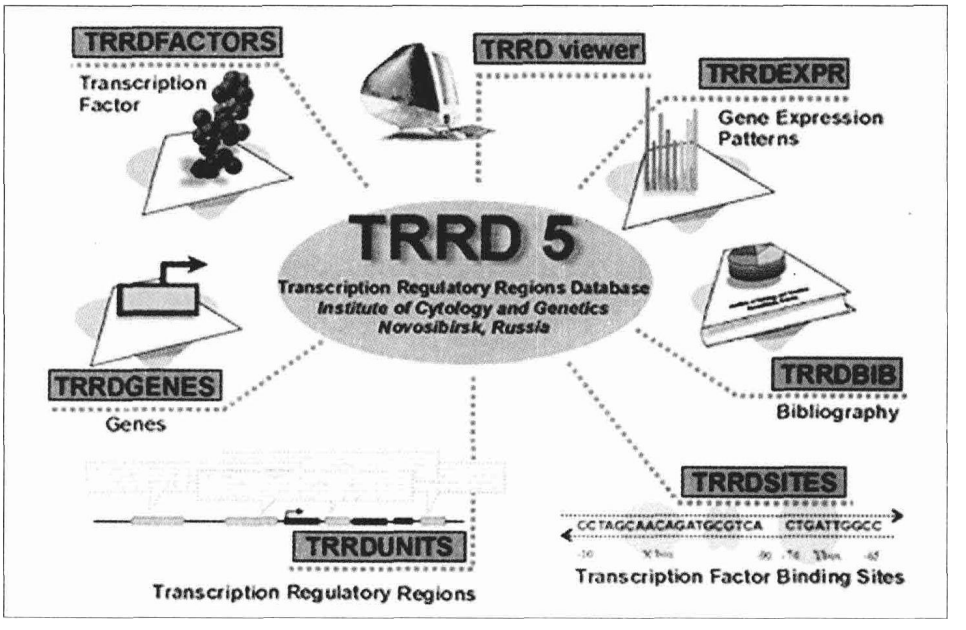


图 11-15 TRRD 数据库的组成

六、其他转录调控相关数据库

除了上面介绍的数据库以外，还有许多有关转录调节位点和转录因子的数据库，如：①EPD 是关于真核 RNA 聚合酶Ⅱ型启动子的非冗余数据库（http://www.epd.isb-sib.ch）；②DBTBS 是针对枯草杆菌转录调控的数据库，包括枯草杆菌的启动子、操纵子和终止子等（http://dbtbs.hgc.jp.）；③DPInteract 是关于大肠杆菌（*E.coli*）的 DNA 结合蛋白及其结合位点的数据库（http://arep.med.harvard.edu/dpinteract）；④PLACE 是关于植物顺式调控 DNA 元件的数据库（http://www.dna.affrc.go.jp/PLACE）；⑤HvrBase 是注释灵长类线粒体 DNA 调控区序列的数据库（http://www.hvrbase.org/）。

对真核生物转录调控区进行计算机预测和鉴定是具有挑战性的研究工作。到目前为止，尽管相关数据库和软件资源得到了很大的丰富和发展，但仍存在着明显不足，如：①大多数数据库对于数据的创新、精确性和准确性缺少权威评价，数据过多、重复、分类较粗等；②公共数据库中，针对人类只有极少数被实验证实的顺式作用元件，绝大多数基因的转录调控区或启动子仍然未知；③采用人类基因组信息来预测植物、真菌等远缘物种的基因结构时，数据准确性不高，但目前针对植物、真菌等的生物信息学数据库远没有人类的全面和完善；④真核生物的顺式作用元件比原核生物的复杂，需要考虑多种因素；⑤基因的转录不仅具有时空性和组织特异性，还呈现网络化，基因转录调控网络的预测方法还较少。因此高效的实验方法和设计良好的预测软件仍是生物学家面临的严峻课题。随着分子生物学、遗传学和生物信息学的高速发展，更多的真核生物启动子序列将得到分析，各顺式作用元件的功能也会逐渐明确，启动子的计算机预测研究工作也将有更广阔的发展空间。

小 结

基因的转录是通过转录因子结合到靶基因的特异位点(转录因子结合位点)来完成的。近年来随着基因芯片和下一代测序技术的发展,高通量鉴定转录因子及其结合位点的实验方法 ChIP-chip 和 ChIP-seq 开始应用。两者的共同特点是数据多,信息量大,为生物信息学分析提供了重要条件。在信息学分析中常常采用共有序列、序列标识图、位置频率矩阵和位置权重矩阵等来表示转录因子结合位点。基因转录调控的信息学分析包括三方面的研究内容:一是从众多序列中鉴定出同一转录因子的共有序列,称之为转录因子的识别。其主要步骤是首先筛选出可能被同一转录因子调控的多基因序列;其次分别应用单个模体测算法、遗传系谱印记法、顺式调控模块识别法、SISSR 算法等多种方法进行评估和分析,找出具有统计显著性的短片段;然后采用 Motifclass 方法或回归模型进一步去除冗余序列;最后通过搜索相关转录调控数据库确定可能与之结合的转录因子。二是根据已知转录因子结合位点模体,预测目的基因调控区序列中转录因子结合位点,称为转录因子结合位点的定位。其主要方法是应用一些软件或程序进行打分,给出预测结果。其数据的多少与所设定的阈值和相关参数密切相关。目前比较常用的程序有 AliBaba、TESS 和 MatrixCatch 等。三是对海量实验数据进行整理和挖掘,建立转录调控相关数据库。目前比较公认的数据库有 TRANSFAC、JASPAR 和 TRED 等。这些数据库各有所长,它们为转录因子结合位点的识别和定位研究提供了重要数据资源。

Summary

Gene transcription is performed through transcription factors binding to the specific sites (transcription factor binding sites) of target gene. Recently, ChIP-chip and ChIP-seq approaches have been applied to identify transcription factors and their binding sites with the development of gene chip and next-generation sequencing technologies. Since there is a lot of data in both ChIP-chip and ChIP-seq, bioinformatics has displayed an important role. In the bioinformatics analysis, the transcription factor binding site is often represented by consensus sequence, sequence logo, position frequency matrix (PFM) and position weight matrix (PWM). The bioinformatics analysis of the transcriptional regulation includes three directions of study content. One is called identification of transcription factors to identity the consensus sequence for each transcription factor from multitude sequences, the first archae-procedure of which is to screen possible sequences regulated by each transcription factor. The second is to evaluate and analyze them with a series of methods: motif discovery; phylogenetic footprinting; *cis*-regulatory module (CRM) and site identification from short sequence reads (SISSR) and find out the short fragment with statistical significance. Then, remove the redundancy sequences with Motifclass system or regression model. At last, define the transcription factor possibly binding to the site through retrieving the correlated transcriptional regulation database. Another one is called the location of transcription factor binding site that is to compare with the known motif of transcription factor site using some programs or software and to predict the transcription factor binding site in the sequences of target gene regulation region. The export data depends on selected threshold value and parameters. Some programs or software, such as AliBaba, TESS and MatrixCatch are frequently used. The third one is to collect and integrate the great experimental data and to construct transcriptional regulation databases. TRANSFAC, JASPAR and TRED are relatively received databases, in which each has a unique, providing significant data resource to study the identification and location of transcription factor binding site.

(许丽艳)

习 题

1. ChIP-chip 的分辨率是：

A. 6~20bp B. 30~50bp C. ~200bp

2. ChIP-seq 的分辨率是：

A. 6~20bp B. 30~50bp C. ~200bp

3. 从多个基因启动子序列，找出一个或几个转录因子共有结合位点的研究，称之为：

A. 转录因子结合位点的识别

B. 转录因子结合位点的定位

4. 根据已知的转录因子结合位点模体在感兴趣靶基因启动子区域内搜索相应转录因子可能结合位点的研究，称之为：

A. 转录因子结合位点的识别

B. 转录因子结合位点的定位

5. 简述共有序列、位置频率矩阵和位置权重矩阵的概念。

6. 试述转录因子结合位点识别的详细操作流程。

7. 简述 P-Match、Patch、MatrixCatch 和 TESS 等程序或软件的主要特点。

8. 比较 TRANSFAC、JASPAR、TRED、DBTSS 和 TRRD 等数据库的优劣势。

9. 试述转录调控数据库的现状和存在的不足。

10. 应用 NCBI 等核酸数据库的基因信息，尝试进行转录因子结合位点的识别和定位分析。

主要参考文献

1. Robertson G., Hirst M., Bainbridge M., et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods, 2007, 4(8): 651-657.
2. Euskirchen G. M., Rozowsky J. S., Wei C. L., et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: com-parison of array- and sequencing-based technologies. Genome Res., 2007, 17(6): 898-909.
3. 李婷婷，蒋博，汪小我等. 转录因子结合位点的计算分析方法. 生物物理学报，2008，24(5): 334-347.
4. Jothi R., Cuddapah S., Barski A., et al. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res., 2008, 36(16): 5221-5231.
5. Wingender E., Dietze P., Karas H., et al. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res., 1996, 24(1): 238-241.
6. Sandelin A., Alkema W., Engström P., et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res., 2004, 32(Database issue): D91-94.
7. Jiang C., Xuan Z., Zhao F., et al. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res., 2007, 35(Database issue): D137-140.
8. Suzuki Y., Yamashita R., Nakai K., et al. DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. Nucleic Acids Res., 2002, 30(1): 328-331.
9. Kolchanov N. A., Ignatieva E. V., Ananko E. A., et al. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res., 2002, 30(1): 312-317.
10. Sierro N., Makita Y., de Hoon M., et al. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. Nucleic Acids Res., 2008, 36 (Database issue): D93-99.

第十二章 生物分子网络
CHAPTER 12 BIOLOGY MOLECULAR NETWORK

第一节 引 言
Section 1 Introduction

网络是人们很熟悉的概念，例如近年来我国修建的高速公路将众多城市连接为一个巨大的网络，城市作为网络中的节点通过公路与其他城市连接在一起，部分临近的城市由直通的公路连接，而更多的城市则可以通过其他城市的中转而被连接起来。给世界带来巨大变化的国际互联网本身也是一个巨大的网络，网络服务器、个人计算机和其他计算设备被通讯线路连接在一起，通过网络中节点之间的连接，实现了全球计算机间的高速通讯与信息资源共享。在人们的周围还有一些网络是无形的，例如人群间的人际关系网络等社会学网络，人或者群体作为节点被多种多样的关系关联起来。可以说，网络是复杂系统存在的普遍形式。而通过已有的经验和知识重构知识网络，并以其为工具进一步分析复杂系统的内在规律是研究复杂系统的有效和重要途径。

“人类基因组计划”引发的生命科学革命带来了一个全新的“后基因组时代”。生命活动本身的复杂性和迅速增加的海量数据资源要求生命现象必须要在成千上万个生物分子组成的复杂系统层面上予以认识。因此，系统全面地研究各生物大分子及其间存在的相互作用成为“后基因组”生物学研究的关键目标。为揭示数量巨大的生物大分子及其间的相互作用如何在复杂的生存环境中行使生物学功能，需要研究者采用不同于传统生物学研究手段的新技术。本章将介绍网络分析在系统生物学中的应用。

第二节 生物分子网络概述
Section 2 Description of Biology Molecular Network

一、生物分子网络的基本概念

近年来，复杂网络理论和技术发展迅速，大量复杂的技术网络和社会学网络被发掘和分析。在生物系统中同样包含很多不同层面和不同组织形式的网络。目前，基因转录调控网络、生物代谢与信号传导网络、蛋白质相互作用网络是最常见的生物分子网络。这些网络通常由许多不同的参与生物过程的分子元件组成，其中最重要的元件是基因和蛋白质。但对“系统”而言，关键不是元件本身，而是元件之间的关系。从生物分子的角度来看，关系可以是分子与分子之间的相互作用，也可以是某种化学反应。而为了能够清晰地重构与分析这些网络，必须先明确网络的基本概念。

（一）网络的定义

通常可以用图 $G=(V,E)$ 表示网络(network)，其中 V 是网络的节点集合，每个节点代表一个生物分子，或者一个环境刺激；E 是边的集合，每条边代表节点之间的相互关系。当 V 中的两个节点 v_1 与 v_2 之间存在一条属于 E 的边 e_1 时，称边 e_1 连接 v_1 与 v_2，或者称 v_1 连接于 v_2，也称作 v_2 是 v_1 的邻居。

（二）有向网络与无向网络

根据网络中的边是否具有方向性或者说连接一条边的两个节点是否存在顺序，网络可以分为有向网络与无向网络，边存在方向性为有向网络(directed network)，否则为无向网络(undirected network)，如图 12-1 所示。生物分子网络的方向性取决于其所代表的关系，如调控关系中转录因子与被调控基因之间是存在顺序关系的，因此转录调控网络是有向网络，而基因表达相关网络中的边代表的是两个基因在多个实验条件下表达的高相关性，因此是无向的。

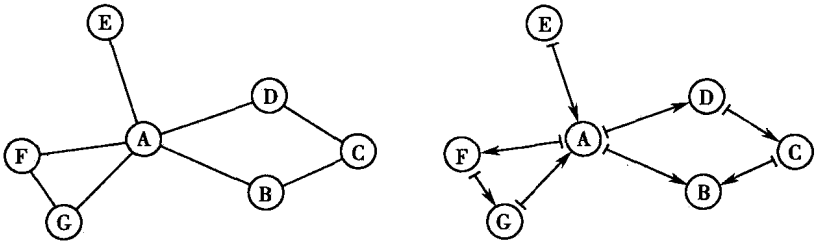


图 12-1 A 无向网络；B 有向网络

（三）加权网络与等权网络

网络中的边在网络中具有不同意义或在某个属性上有不同的价值是网络中普遍存在的一种现象。比如交通网中，连接两个城市(节点)的道路(边)一般具有不同的长度，而在互联网中两台直接相连的计算设备间通讯的速度也不尽相同。

如果网络中的每条边都被赋予相应的数值，这个网络就称为加权网络(weighted network)，所赋予的数值称为边的权重。权重可以用来描述节点间的距离、相关程度、稳定程度、容量等各种信息，具体含义依赖于网络和边本身所代表的意义。

如果网络中各边之间没有区别，可以认为各边的权重相等，称为等权网络或无权网络(unweighted network)。

（四）二分网络

如果网络中的节点可分为两个互不相交的集合，而所有的边都建立在来自不同集合的节点之间，则称这样的网络为二分网络(bipartite network)。例如，药物分子与其靶蛋白的结合关系即可以用二分网络的形式来表示。

（五）网络中的路径与距离

网络中的路径是指一系列节点，其中每个节点都有一条边连接到紧随其后的节点。对包含节点数目有限的路径来说，第一个结点称为起点，最后一个节点称为终点，二者均可称为路径的端点，其余的节点则称为路径的内点或中继点。这样的路径也称为连接起点与终点的路径。例如，图 12-1(A)中节点 G 到节点 C 的路径有 $l_1=\{G, A, B, C\}$，$l_2=\{G, A, D, C\}$，$l_3=\{G, F, A, B, C\}$ 和 $l_4=\{G, F, A, D, C\}$。对无向网络来说，只要将路径的顺序颠倒就可以得到从原来的终点指向起点的路径，但是在有向网络中，起点与终点是不可逆的，如图 12-1(B)所示网络中节点由 A 出发到节点 C 间存在路径 $l=\{A, D, C\}$，但 C 不能找到路径回到 A。

网络中如果两个节点间能够由一条路径连接，则称这两个节点是连通的。所有能够彼此连通的节点和它们之间的边构成了一个连通分量。

路径中所经过边的权重之和称为路径的权重，也称为路径的长度。对于等权网络而言，路径的长度即为路径中所经过边的数目，上述图 12-1(A)中从节点 G 到节点 C 的路径中，l_1 和 l_2 的长度为 3，l_3 和 l_4 的长度为 4。

在连接两个节点的所有路径中，长度最短的路径称为最短路径，最短路径的长度称为从起点到终点的距离，上述图 12-1(A)中从节点 G 到节点 C 的距离为 3。

二、基因调控网络

所有生物在生长发育和分化的过程中，以及在对外部环境的反应中，各种相关基因有条不紊的表达起着至关重要的作用。与原核生物相比，真核生物基因表达的调控更为复杂，真核生物基因表达的调控主要是指编码蛋白质的 mRNA 产生和行使生物功能过程中的调节与控制。从理论上讲，基因表达调控可以发生在遗传信息传递过程的各个水平上，其中转录调控是基因表达调控中最重要、最复杂的一个环节，也是当前研究的重点。

（一）基因调控检测

20 世纪 90 年代开发的微阵列技术是检测基因表达水平的有力工具，相关的数据和分析方法比较完善和成熟。其中，最重要的是染色质免疫沉淀技术（Chromatin Immunoprecipitation，ChIP）和在此基础上发展起来的 ChIP-chip 芯片等技术。

1. 染色质免疫沉淀技术　ChIP 是一种在体内研究 DNA 与蛋白质相互作用的方法。ChIP 不仅可以检测体内转录因子与 DNA 的动态作用，还可以用来研究组蛋白的各种共价修饰与基因表达的关系。近年来，这种技术得到不断的发展和完善。ChIP 与体内足迹法相结合，用于寻找转录因子的体内结合位点；RNA-ChIP 用于研究 RNA 在基因表达调控中的作用。它与 DNA 芯片和分子克隆技术相结合，可用于高通量地筛选已知蛋白的未知 DNA 靶点和研究反式作用因子在整个基因组上的分布情况。

2. ChIP-chip 芯片技术　ChIP-chip 芯片的基本原理是在生理状态下把细胞内的蛋白质和 DNA 交联在一起，超声波将其打碎为一定长度范围内的染色质小片段，然后通过所要研究的目的蛋白特异性抗体沉淀此复合体，特异性地富集目的蛋白结合的 DNA 片段，通过对目的片段的纯化与检测，获得蛋白质与 DNA 相互作用的信息。ChIP-chip 技术的发展为分析活细胞或组织中 DNA 与蛋白质的相互关系提供了极为有力的工具。

（二）基因转录调控数据库

各种搜集整理了转录调控信息的生物学数据库为研究工作提供了有力的帮助。常用的基因转录调控数据库包括 TRANSFAC、TRRD、RegulonDB 和 COMPEL 数据库。

1. TRANSFAC 数据库　TRANSFAC 数据库是关于转录因子、它们在基因组上的结合位点的数据库。由 SITE、GENE、FACTOR、CLASS、MATRIX、CELLS、METHOD 和 REFERENCE 等结构组成。此外，还有几个与 TRANSFAC 密切相关的扩展库：PATHODB 库收集了可能导致病态突变的转录因子和结合位点；S/MARTDB 库收集了与染色体结构变化相关的蛋白因子和位点的信息；TRANSPATH 库用于描述与转录因子调控相关的信号传递网络；CYTOMER 库表现了人类转录因子在各个器官、细胞类型、生理系统和发育时期的表达状况。该数据库网址为：http://www.gene-regulation.com/pub/databases.html。

2. TRRD 数据库　TRRD 数据库是在不断积累的真核生物基因调控区结构 - 功能特性信息基础上构建的。每一个 TRRD 的条目里包含特定基因各种结构 - 功能特性：转录因子结合位点、启动子、增强子、沉默子以及基因表达调控模式等。TRRD 包括五个相关的数据表：TRRDGENES 包含所有 TRRD 库基因的基本信息和调控单元信息；TRRDSITES 包括调控因子结合位点的具体信息；TRRDFACTORS 包括 TRRD 中与各个位点结合的调控因子的具体信息；TRRDEXP 包括对基因表达模式的具体描述；TRRDBIB 包括所有参考文献。TRRD 主页提供了对这几个数据表的检索服务。该数据库网址为：http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/。

3. RegulonDB 数据库　RegulonDB 数据库是一个提供转录起始和调控网络信息的数据库。其中存储着与转录调控相关的操纵子、转录单元、启动子、结合位点以及转录因子和受控基因等信息。该数据库的网址为：http://regulondb.ccg.unam.mx/。

4. COMPEL 数据库　COMPEL 数据库提供了许多复合转录元件。这些复合转录元件是组合转

录调控的最小单位，包括两个属于不同转录因子在位置关系上紧密相连的结合位点。它们强调高级生物复杂微妙的表达是通过转录因子的特定组合调节实现的。构成复合元件的两个结合部位之间的距离和先后顺序很重要，它们的排列必须支持转录因子的三维空间结构，使两者能够正确地结合。该数据库网址为：http://compel.bionet.nsc.ru/。

此外，SCPD(http://rulai.cshl.edu/SCPD/)，JASPAR(http://jaspar.genereg.net)和 DBD(http://www.transcriptionfactor.org)等数据库也提供了丰富的有关转录调控的信息。

关于转录调控的检测技术和数据库信息，可以参考本书第十一章。

(三) 基因转录调控网络

通过基因转录调控数据可以构建基因转录调控网络。基因转录调控网络是以转录因子和受调控基因作为节点，以调控关系作为边的有向网络，见图 12-2(A)。有的时候，根据转录因子是促进还是抑制受控基因的表达，调控网络中的边可以分为正调控和负调控。

三、蛋白质互作网络

蛋白质是构成生物体的重要物质，也是行使生物功能的重要生物大分子。单独蛋白通过彼此之间的相互作用构成蛋白质相互作用网络来参与生物信号传递、基因表达调节、能量和物质代谢及细胞周期调控等生命过程的各个环节。系统分析大量蛋白在生物系统中的相互作用关系，对于了解生物系统中蛋白质的工作原理，了解疾病等特殊生理状态下生物信号和能量物质代谢的反应机制，以及了解蛋白质间的功能联系都有重要意义。

蛋白质互作通常可以分为物理互作和遗传互作。物理互作是指蛋白质间通过空间构象或化学键彼此发生的结合或化学反应，是蛋白质互作的主要研究对象。而遗传互作则是指在特殊环境下，蛋白质或其编码基因受到其他蛋白质或基因影响，常常表现为表型变化之间的相互关系。

(一) 蛋白质互作检测技术

早期的蛋白质互作检测工作主要基于免疫共沉淀(co-immunoprecipitation)技术。近些年来，一些高通量的检测技术应用于检测蛋白质间的相互作用关系(蛋白质互作)。其中较为常用的技术有酵母双杂交(yeast two hybrid, Y2H)技术和串联亲和纯化 - 质谱分析(Tandem Affinity Purification-Mass Spectrometry, TAP-MS)技术。

1. 免疫共沉淀技术　免疫共沉淀技术的原理是在非变性条件下分解细胞，使用特异性的抗体将目标蛋白 A 沉淀提取，在自然状态下与蛋白 A 通过相互作用结合的另一个蛋白也随之沉淀下来，然后通过抗体抗原反应(western 印迹法)检测与目标蛋白 A 一同沉淀下的与目标蛋白发生互作的蛋白，由此判定互作关系的存在。免疫共沉淀方法是目前最为可靠的蛋白质互作检测技术，其优点是能够检测自然状态下蛋白间的相互作用，可靠性比较高。缺点是无法检测短时不稳定的蛋白质互作关系，同时需要预先选定待检测的互作关系以准备相应的抗体，因此检测效率较低，不能并行大规模检测互作关系。

2. 酵母双杂交技术　酵母双杂交技术是研究活细胞体内的蛋白质与蛋白质之间相互作用的技术。该系统的建立是基于对真核生物调控转录起始过程的认识。细胞起始基因转录需要有反式转录激活因子的参与，而酵母的某些转录因子(如 GAL4 转录因子)包含两个相对独立的功能区域，即特异性 DNA 结合结构域(binding domain, BD)和转录激活结构域(activating domain, AD)。只有当这两个结构域在空间上接近时，才能表现出完整的转录因子活性。根据这一特性，首先将编码已知诱饵蛋白(bait)的基因与报告基因的 DNA 结合区域基因融合，再将待测 cDNA 文库基因(prey)与转录因子激活结构域基因融合，将两组融合基因通过载体转染酵母细胞，在酵母中表达出不同的融合蛋白。从而可以根据报告基因(被转录因子所激活的基因)是否表达，判断待检测的蛋白质之间是否存在互作关系。大量的研究文献表明，酵母双杂交技术既可以用来研究哺乳动物基因组编码的蛋白质之间的互作，也可以用来研究高等植物基因组编码的蛋白质之间的互作，因此，它在许多研究领域中都有

着广泛的应用。

3. 串联亲和纯化 - 质谱分析技术　串联亲和纯化 - 质谱分析技术包括串联亲和纯化(TAP)以及高通量质谱分析蛋白质复合物检测技术(High-throughput Mass-Spectrometric Protein Complex Identification, HMS-PCI)。这种方法首先通过免疫共沉淀反应(co-immunoprecipitation)或串联亲和纯化反应获得含有目的蛋白质的蛋白质复合物，分离提纯后用质谱分析或者蛋白质测序来鉴定复合体的各个组分。串联亲和纯化技术是一种能够高通量检测蛋白间互作的技术，其检测可靠性高于酵母双杂交技术。与酵母双杂交技术检测蛋白质之间存在的物理互作不同，对于已知诱饵蛋白(bait)，这种技术将检测出与其同属于至少一个复合物的蛋白质。与免疫共沉淀类似，它适用于检测稳定互作，不适用于检测瞬时互作。

酵母双杂交技术和串联亲和纯化 - 质谱分析技术等高通量的蛋白质互作检测技术的广泛应用，使得大规模地分析相互作用的蛋白质成为可能。其主要的应用有：①发现新的蛋白质和蛋白质的新功能；②在细胞体内研究抗原和抗体的相互作用；③筛选药物的作用位点以及药物对蛋白质之间相互作用的影响；④建立基因组蛋白连锁图，在蛋白质组尺度上进行研究等。

4. 蛋白质互作预测技术　近年来，随着生物信息学的发展，开发了一系列的蛋白质互作预测技术。这些技术的应用对指导实验检测、降低实验成本以及弥补实验检测数据的不足起到了重要的作用。

基于同源性的预测技术是一类最普遍的蛋白质互作预测技术。生物学家认为彼此互作的一对蛋白更可能共同进化。基于这个观点，可以通过检查一对蛋白的系统发生距离来推断其间存在的互作关系。还可以利用进化的保守性，检测一对蛋白是否有同源序列参与已知的互作结构，由此对这对蛋白质的互作关系作出预测。此外，还可以通过与已知的蛋白质结构模式进行比对，预测分别包含互作结构域的一对蛋白质的互作关系。

基于多重数据源和机器学习算法的蛋白质互作预测技术不同于同源性预测技术。其基本原理是运用 Bayes 网络等机器学习技术整合多种数据源的信息，构建蛋白质互作网络。

5. 遗传互作检测技术　遗传互作的常见检测方法包括剂量增长补足(dosage growth defect)和联合致死(synthetic lethality)等。其中剂量增长补足是指如果一个基因突变或者被敲除时另一个基因的表达量明显增加。联合致死则是表示在基因敲除实验中，只敲除任何一个都不会造成细胞死亡，而两个基因一同被敲除时细胞就会死亡。

(二) 蛋白质互作数据库

目前，已经有大量蛋白质互作数据存储在公共数据库中，提供了大量的蛋白质相互作用信息，其中包括 BIND 数据库、DIP 数据库、MIPS 数据库和 BioGRID 数据库等。从这些数据库中，可以得到不同物种的蛋白质互作信息及其实验证据。

1. BIND 数据库　BIND 数据库是生物分子对象网络数据库(Biomolecular Object Network Databank)中最重要的组成部分之一。主要记录蛋白质互作等生物分子间的相互作用信息，并将其中的信息分为经过人工检查的可信信息和高通量数据信息。用户可以通过网络工具查询互作信息也可以将互作信息下载到本地进行处理。该数据库的网址为：http://bind.ca/。

2. DIP 数据库　DIP 数据库是专门存储蛋白质相互作用信息的数据库。该数据库中也包含人工检查的可靠信息和由自动计算方法所获取的高通量数据。该数据库可以按照不同的物种选择下载不同格式的蛋白质互作信息。用户可以通过网络工具查询互作信息也可以将互作信息下载到本地进行处理。该数据库的网址为：http://dip.doe-mbi.ucla.edu/。

3. MIPS 数据库　MIPS 数据库是一个跨物种的综合性数据库，包含多种数据库信息。其中的 CYGD 数据库提供了比较完整的酵母蛋白质互作信息。而 MIPS 哺乳动物数据库 MPPI 则提供了经过人工检查的哺乳动物蛋白质互作信息。用户可以通过网络工具查询互作信息也可以将互作信息下载到本地进行处理。该数据库的网址为：http://www.helmholtz-muenchen.de/en/mips/。

4. BioGrid 数据库 BioGrid 数据库是一个包含多物种蛋白质互作信息的数据库。数据库中包含来自多个物种的互作信息，其中既包括物理互作信息也包括遗传互作信息。用户可以通过网络工具查询互作信息也可以将互作信息下载到本地进行处理。该数据库的网址为：http://www.thebiogrid.org/。

（三）蛋白质互作网络

蛋白质互作网络是系统显示蛋白质互作信息的基本方法。将蛋白作为节点，相互作用关系作为边，将蛋白质组整体连接到一个系统网络当中，见图 12-2(B)。一般情况下，蛋白质互作网络是一个规模较大的无向网络。目前蛋白质互作网络是被研究最充分的生物分子网络之一，蛋白质互作网络也往往是规模最大的生物分子网络，常常包含数千甚至上万个节点以及为数更多的边。

四、代谢网络和信号传导网络

在生物化学领域，代谢通路是指细胞中代谢物在酶的作用下转化为新的代谢物过程中所发生的一系列生物化学反应。而代谢网络则是指由代谢反应以及调节这些反应的调控机制所组成的描述细胞内代谢和生理过程的网络。

生物中的信号传导(signal transduction)则是指细胞将一种类型的生物信号或刺激转换为其他生物信号最终激活细胞反应的过程。同代谢通路一样，信号传导的过程中多个生物分子在酶作用下按照一定顺序发生一系列生理化学反应，由此得到了信号传导通路。信号传导网络即是指参与信号传导通路的分子和酶以及其间所发生的生化反应所构成的网络。

这些网络是研究和分析代谢过程和信号传导过程的重要工具，随着许多物种基因组测序的逐步完成以及新的生物检测技术的开发，对生物细胞内生化反应的知识也正以极快的速度增加，这就使构建人类等物种完整的生物代谢网络和信号传导网络成为可能。目前代谢和信号传导通路信息被收集和整理到一些重要的通路数据库当中，这些信息是构建代谢网络与信号传导网络的基础。

（一）通路数据库

1. KEGG 数据库 KEGG 数据库是关于基因、蛋白质、生化反应以及通路的综合生物信息数据库。它由多个子库构成，其中的 KEGG PATHWAY 数据库中包含有大量物种的代谢与生物信号传导通路信息。该数据库的网址为：http://www.genome.jp/kegg/。

2. ERGO 数据库 ERGO 数据库是关于多个物种基因组信息的综合数据库。其中包含有关于代谢通路和非代谢通路的综合信息。该数据库的网址为：http://ergo.integratedgenomics.com/。

3. BioCyc 数据库 BioCyc 数据库是为不同物种单独构建的代谢通路数据库的合集，目前已包括超过 500 个不同的数据库。该数据库根据数据的可信程度，分为三个层级，即第一层细致确认数据库，其中最重要的是针对大肠杆菌(Escherichia Coli)的 EcoCyc 数据库和针对多物种的 MetaCyc 数据库，第二层初步确认计算通路数据库和第三层未经确认计算通路数据库。该数据库的网址为：http://www.biocyc.org/。

4. GeneDB 数据库 GeneDB 是关于多物种基因信息的综合数据库，当用户输入相关基因查询时，GeneDB 会提供该基因所参与的通路信息。该数据库的网址为：http://www.genedb.org/。

（二）代谢网络和信号传导网络

代谢网络和信号传导网络中包括大量不同的通路，而每条通路也包含不同的生物分子之间的多种生理和化学反应，因此代谢网络具有不同于其他生物分子网络的复杂性。根据研究的目的常常需要构建不同层次的代谢网络。其中包括：

1. 完全网络 最完整的保存代谢通路中各个反应，以及每个反应中的底物、产物和酶，如果同一个酶参与不同反应则在网络中应以不同的节点表示，见图 12-2(C、D)。

2. 多反应物网络 包含参与生物通路的代谢物即底物、产物和酶的有向网络，其中每种代谢物只由一个节点表示，边由底物指向产物，酶与底物、产物之间的边则可以由双向边来表示，也可以作为边的属性。

3. 主要反应物网络　在部分研究中，研究者不关心代谢反应中的酶和其他一些如提供能量与磷酸键的 ATP 等的共反应因子，由此就得到了只包含主要代谢底物指向主要产物的网络。

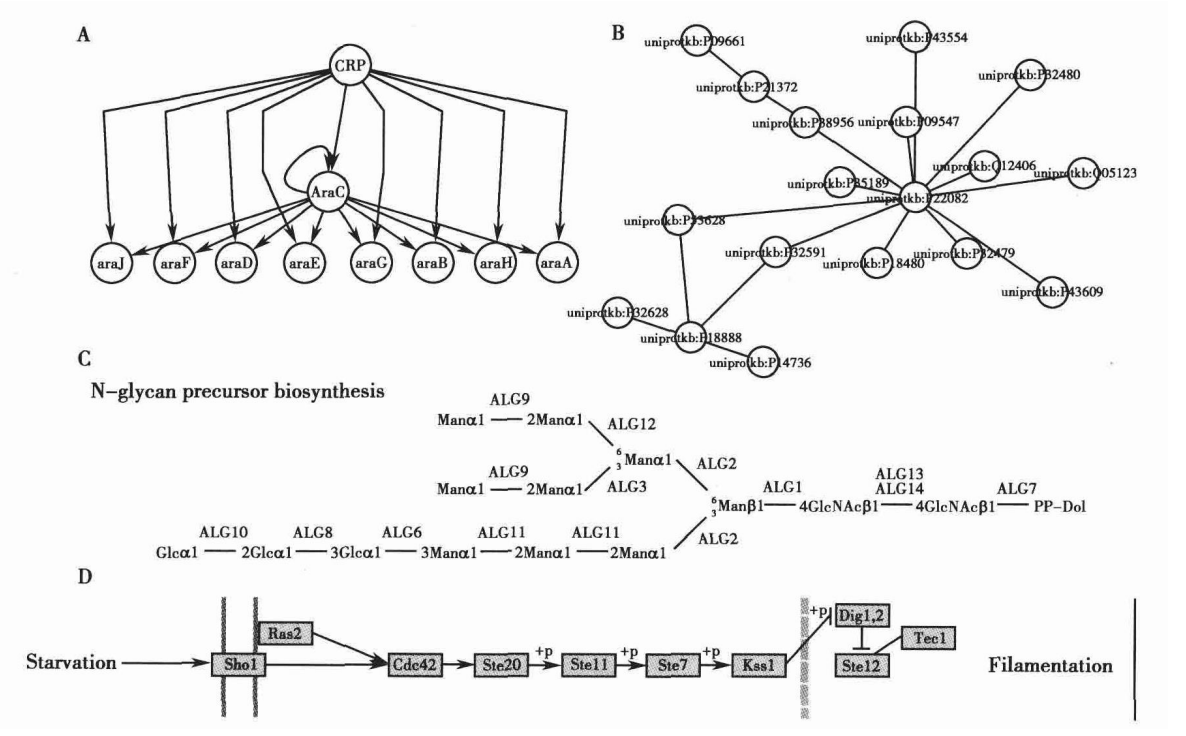


图 12-2　生物分子网络样例

A 为基因转录调控网络；B 为蛋白质互作网络；C 为完全代谢网络；D 为完全信号传导网络

（三）细胞间通讯网络

生物信号的传递不仅发生在细胞内部，同样也出现在不同细胞之间。细胞间通讯是生物体调节细胞基本活性、协调细胞活动的信息传递机制。细胞间通讯广泛存在于各种生物中，单细胞生物之间、多细胞生物同组织细胞之间和多细胞生物不同组织细胞之间都存在着各种形态的细胞间通讯。通过正确的接收和应答细胞间传递的信息，细胞能够协作完成自身发育、组织修复、免疫保护并维护正常组织的体内平衡。而细胞间信息传递过程中的错误则可能引发多种疾病，如癌症、自身免疫性疾病和糖尿病等。理解细胞间通讯过程对进一步研究生命活动的原理以及疾病的治疗都具有重要意义。

与传统的孤立研究不同，系统生物学将细胞间的各种通讯作为网络来进行研究。研究内容包括细胞间通讯网络的基本结构及其在信息传递过程中的作用原理。细胞间存在多种通讯方式，根据细胞间的距离可以分为直接接触（邻分泌通讯）、近距离通讯（旁分泌通讯）和远距离通讯（内分泌通讯），此外还包括仅针对同类型细胞的自分泌通讯。环境中的信息通过一类被称为受体的蛋白接收并传递到细胞内，并与细胞内的信号传递网络相连接，从而促使细胞在必要时做出相应反应。

第三节　生物分子网络分析

Section 3　Analysis of Biology Molecule Network

一、网络的拓扑属性

网络的拓扑属性是描述网络本身及其内部节点或边结构特征的测度。这些测度对进一步分析网络结构和探索关键节点有重要的意义。

（一）连通度

连通度(degree)是描述单一节点的最基本的拓扑性质。节点 v 的连通度是指网络中直接与 v 相连的边的数目，例如在图 12-3(A)中节点 A 的连通度为 3。对于有向网络往往还要区分边的方向，由节点 v 发出的边的数目称为节点 v 的出度，指向节点 v 的边数则称为节点 v 的入度。在本章中，符号 k 表示连通度，k_{out} 表示出度，k_{in} 表示入度。在图 12-3(B)中，节点 A 的入度为 1，出度为 2。

连通度描述了网络中某个节点的连接数量，整个网络的连通性可以使用其平均值来表示。对于由 N 个节点和 L 条边组成的无向网络，其平均连通度为 $2L/N$。

连通度是一种简单而十分重要的拓扑属性。在研究中，连通度较大的节点称为中心节点(hub)，它们很自然地成为目前研究的重点。研究显示，在蛋白质互作网络等生物分子网络中，支持生命基本活动的必需基因或其翻译产物在中心节点中出现的频率显著高于一般节点。同时，人类蛋白质互作网络的研究表明，中心节点显著富集着与癌症等遗传性疾病相关的基因。

（二）聚类系数

在很多网络中，如果节点 v_1 连接于节点 v_2，节点 v_2 连接于节点 v_3，那么节点 v_3 很可能与 v_1 相连接。这种现象体现了部分节点间存在的密集连接性质，可以用聚类系数(clustering coefficient) CC 来表示，在无向网络中，聚类系数定义为：

$$CC_v = \frac{n}{C_k^2} = \frac{2n}{k(k-1)} \qquad \text{式 12-1}$$

其中 n 表示在节点 v 的所有 k 个邻居间边的数目。在无向网络中由于 n 的最大数目可以由邻居节点的两两组合数 $C_k^2 = k(k-1)/2$ 来确定，所以 CC 值位于[0, 1]区间。当节点 v 的所有邻居都彼此连接时，v 的聚类系数 $CC_v = 1$；相反，当 v 的邻居间不存在任何连接时，$CC_v = 0$。在图 12-3(A)中，节点 A 有三个邻居 {B, C, D}，其间只有一条边连接，所以节点 A 的聚类系数 $CC_A = \frac{2\times 1}{3\times(3-1)} = \frac{1}{3}$。

在有向网络中，由于两个节点间可以存在两条方向相反的边，则标准化的聚类系数被定义为：

$$CC_v = \frac{n}{P_k^2} = \frac{n}{k_{out}(k_{out}-1)} \qquad \text{式 12-2}$$

其中 k_{out} 指 v 的出度，n 指所有 v 所连接的节点彼此之间存在的边数。在图 12-3(B)中，节点 A 连接 2 个节点 B 和 C，其间只有 1 条边 {C→B}，则节点 A 的聚类系数为 $CC_A = \frac{1}{2\times(2-1)} = \frac{1}{2}$。

（三）介数

一个节点的介数(betweenness)是衡量这个节点出现在其他节点间最短路径中的比例。节点 v 的介数 B_v 定义如下：

$$B_v = \sum_{i\neq j\neq v\in V} \frac{\sigma_{ivj}}{\sigma_{ij}} \qquad \text{式 12-3}$$

其中，σ_{ij} 表示节点 i 到节点 j 的最短路径的条数，σ_{ivj} 表示其中通过节点 v 的路径条数。

介数也可以用标准化至[0, 1]区间的形式表示：

$$B_v = \frac{1}{(n-1)(n-2)} \sum_{i\neq j\neq v\in V} \frac{\sigma_{ivj}}{\sigma_{ij}} \qquad \text{式 12-4}$$

介数表明了一个节点在其他节点彼此连接中所起的作用。介数越高，意味着在保持网络紧密连接性中节点越重要。

如在图 12-3(A)中，A 以外的节点有 4 个，彼此间存在 $C_4^2 = 6$ 对节点关系，每对关系都只能找到 1 条最短路径，则所有的 $|\sigma_{ij}| = 1$，而只有 {B, A, D}，{C, A, D}，{D, A, C, E} 以及它们的逆序路径共 6 条最短路径通过节点 A，所以，节点 A 的介数为 6。

而在图 12-3(B)中，由于存在方向性，节点 A 以外 4 个节点间彼此间可能存在的连通路径按排

列数计算有$P_4^2=12$条，但真正连通的路径只有{C, B}, {D, A, B}, {D, A, C}, {D, A, C, B}, {E, C}, {E, C, B}。其中经过节点 A 的路径有 2 条，则节点 A 的介数为 2。

（四）紧密度

紧密度(closeness)是描述一个节点到网络中其他所有节点平均距离的指标。节点 v 的紧密度 C_v 定义如下：

$$C_v=\frac{1}{n-1}\sum_{j\neq v\in V}d_{vj} \qquad \text{式 12-5}$$

其中 d_{vj} 表示节点 v 到节点 j 的距离。紧密度测度衡量节点接近网络“中心”的程度，紧密度测度越小，节点越接近中心。

图 12-3(A)中，节点 A 到 B、C、D、E 的距离分别为 1、1、1、2。则节点 A 的紧密度为 1.25。

（五）拓扑系数

类似于聚类系数，拓扑系数(topology coefficient)是反映互作节点间共享连接比例的测度，节点 v 的拓扑系数 T_v 可以定义为：

$$T_v=\frac{1}{|M_v|}\sum_{t\in M_v}C_{v,t}/\min\{k_v,k_t\} \qquad \text{式 12-6}$$

其中，$C_{v,t}$ 表示与节点 v 和节点 t 都连接的节点数。M_v 为所有与节点 v 分享邻居的节点集合。拓扑系数反映了节点的邻居间被其他节点连接在一起的比例。

如图 12-3(A)，与 A 节点共享邻居的节点共有 3 个，则 $M_A=\{B, C, E\}$ 其连通度分别为 $k_B=2$，$k_C=3$，$k_E=1$。则节点 A 的拓扑系数 $T_A=\frac{1}{3}\left(\frac{1}{2}+\frac{1}{3}+1\right)=\frac{11}{18}$。

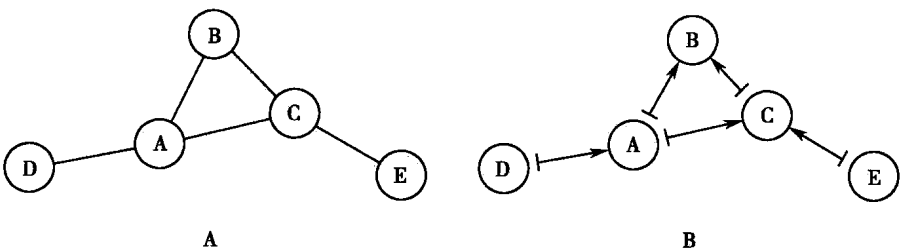


图 12-3 有向网络与无向网络

（六）直径

直径(diameter)是描述网络总体性质的一个属性。网络的直径是指网络中任意两个连通节点间距离的最大值。网络的直径代表了网络中节点连接可能出现的最远距离，标志着网络紧密的程度。

（七）平均距离

网络的平均距离也是描述网络总体性质的一个属性。网络的平均距离是指网络中任意两个连通节点距离的平均值，也是衡量网络紧密程度的重要指标。

（八）连通度的分布函数和聚类系数函数

通过统计不同连通度的节点占全部节点的比例，能够得到一种重要的描述网络连通性的属性：连通度的分布函数 $P(k)$, $k=1, 2, \ldots\ldots$。而类似的还可以建立起随连通度变化的聚类系数函数 $C(k)$，当自变量等于 k 时，$C(k)$即为连通度为 k 的节点聚类系数的平均值。与连通度分布函数 $P(k)$类似，$C(k)$也广泛应用于描述网络结构的基本性质。相比于拓扑性质指标的平均数，由于连通度的分布函数以及依赖于连通度的聚类系数函数包含更多的信息，对分布函数的分析往往可以揭示更为深刻的网络性质。

二、无标度网络

无标度(scale free)网络是1999年首次提出的，近年来，人们在互联网和人际关系网络等社会学网络的研究中都发现了这一特性。无标度网络中，大部分节点通过少数中心节点连接到一起，这就意味着节点在网络中的地位是不平等的，中心节点在连接网络完整性方面起更加重要的作用。

(一) 无标度网络定义

无标度网络，是指网络中连通度的分布符合幂率分布，即 $P(k)\sim k^{-r}$ 的网络，如图12-4(B)所示。这种分布说明，在无标度网络中大部分节点的连通度较低，但存在少数连通度非常高的节点使网络连接在一起。在这种网络中，平均连通度等标度已经不足以描述网络的规模和结构。

如果网络中节点间的连接完全是随机的，那么连通度的分布应该符合泊松分布或者在大尺度的情况下近似认为符合正态分布，即度的分布比较均匀，大部分节点的连通度都与平均连通度相差不多，只有极少数节点具有很低或很高的连通度，如图12-4(A)所示。

随机网络中直径或网络平均距离与节点的数目的对数成正比，即 $l\sim\log(N)$。对于包含大量节点的网络，其直径相对要小得多，任意两个节点间只需要较少的转接即可以连接在一起。一方面网络中包含有大量节点和边，表现出“大世界”的景象，另一方面，连接任意节点间的距离却相对较小，呈现“小世界”的特征。这种“小世界”网络是复杂系统互作网络的共同特性，因此成为目前网络研究分析的一个热点问题。

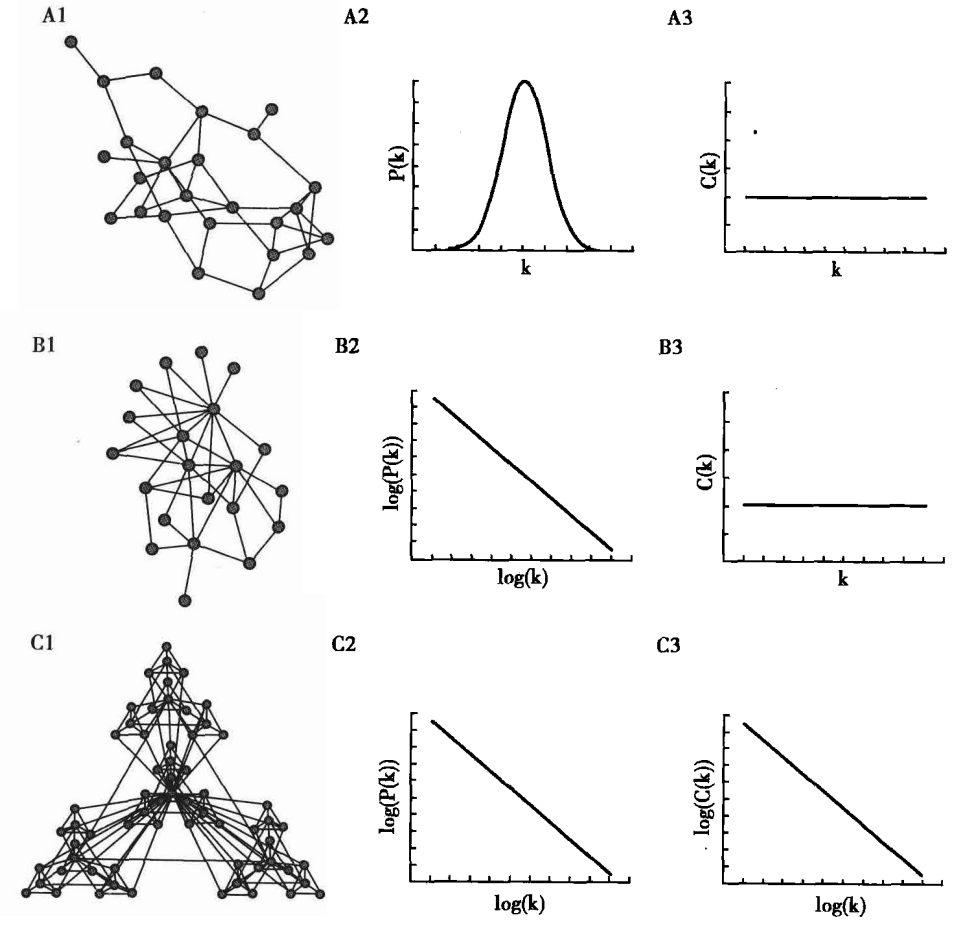


图12-4 随机网络，无标度网络和层次网络及其连通度分布和聚类系数函数趋势图

A为随机网络，其连通度分布符合泊松分布，在大尺度情况下近似服从正态分布

B为无标度网络，其连通度分布符合幂率分布，平均聚类系数函数曲线水平

C为层次网络，其连通度分布符合幂率分布，平均聚类系数与连通度的倒数成正比

无标度网络对网络的另一个重要影响是使网络的直径相对较小，一般来说无标度网络直径的大小正比于网络中节点数目的对数值的对数值，即 $l \sim \log[\log(N)]$。由此可以发现无标度网络比一般小世界网络直径更小，联系更紧密。

（二）无标度网络形成的生物模型

为了解释无标度网络为何会成为包括大部分生物分子网络在内的复杂系统网络模型，Barabási 和 Albert 提出了形成无标度网络的 Barabási-Albert 模型。

该模型首先从一个包含 m_0 个节点的网络开始，其中 $m_0 \geq 2$，初始网络中每个节点的连通度都应大于零，否则在后续过程中将无法与网络连接。而后，通过一个循环过程扩大网络，在每次循环中只增加一个节点，并依次按照概率 $\pi_i = \frac{k_i}{\sum_{v \in V} k_v}$ 决定是否与原网络中节点 i 建立连接，其中 k_i 是节点 i 的连通度。因此，原有连通度较高的节点将更有机会与新加入节点连接，从而获得更高的连通度。按照这种机制构建起的网络即可以得到无标度网络。

例如，在互联网形成的初期，网络中的连接呈现随机特性，而当一个新的节点加入网络时，人们会倾向于访问已经具有一定知名度的网站，也就更有可能与这样的网页建立连接。这样随着越来越多的节点引入网络，网络连接便呈现出无标度特性。这个模型很好地解释了网页连接网络中少数权威网站存在的现象，也为生物分子网络中无标度特性的形成原因提供了很好的启示。

根据这一模型，有学者提出蛋白质网络中出现无标度特性的原因在于基因复制，即在细胞分裂过程中复制产生的基因产物会与相同的蛋白发生相互作用。因此，与发生复制的蛋白连接的蛋白节点将会获得新的连接。高度连接的节点更有可能与发生复制的基因产物发生互作，从而获得额外的连接。因此在生物进化的过程中，就出现了蛋白网络的无标度特性。

同时，还有另一种不同的看法存在，即目前蛋白网络中呈现的无标度特性是来自于目前的诱饵 - 猎物模式的蛋白质相互作用检测方式和目前还远未完善的数据资源。在来自不同结构的随机网络背景中按照诱饵 - 猎物模式抽取部分网络，结果发现来自其他模型的数据也可能随机抽选出无标度的子网。

三、生物分子网络的模块性

细胞功能经常以模块化的形式展现出来。模块是指彼此协同工作从而执行一致功能并在物理上或者功能上紧密联系的一组生物分子（节点）。事实上，在复杂系统中通常包含很多模块，例如，人类通过结成不同层次的各种团体，联系成为整个复杂的人类社会；计算机互联网中相关内容的网页通过页面间的链接密集连接组成一个个独特的模块；近似领域的科学文献间互相引用的频率较高等。在人类的工业化生产中，也往往有意识地采用模块化设计，从而提高工程效率和稳定性。这种模块化的属性已经应用在小到移动电话、个人电脑，大到大型客机、航天器械的设计和制造当中。

生物系统同样包含有大量的模块化现象。例如，蛋白质往往结合成为相对稳定的复合物来行使生物学功能，而蛋白质与核酸分子所组成的复合物在从核酸合成到蛋白质降解的生物基本功能中都发挥了重要的作用。在生物应激反应过程中，共同调控的生物分子也协同完成了使生物体适应内外环境变化的生物功能。总之，细胞中的大多数生物分子或者参与到多分子复合物中行使功能，或者在某个时刻与受到同样调控机制的其他分子协同参与某个生物过程。也就是说，生物分子行使功能的机制中往往会包含有模块化的特性，而网络中这种由许多分子相互结合形成的，有着稳定结构和功能的复合体，称为网络“模块”（module）。

网络的模块性指网络间的节点存在着内部彼此高度连接的子节点集合。由此，模块化的网络连通更为紧密。与同样规模的随机网络相比，虽然拥有相同的节点数与边数，模块化网络的连接却更为密集，这一现象可以由聚类系数 CC 的提高表现出来。同时，模块化的网络往往也同时具有无标

度的特性，即存在一些连通度较大的中心节点连接起不同的模块。连通度的分布 $P(k)$符合 k 的幂率分布，如图 12-4(C2)所示。

此外，聚类系数依赖于连通度的函数 $C(k)$在网络的模块性判别中也起到了重要的指示作用。模块化的性质说明大尺度的网络是由内部密集互作的小模块通过少数中心节点连接在一起的。这就意味着，大型模块化网络中连通度较低的节点往往具有较高的聚类系数，而另一方面，连通度较高的节点连接了不同的模块，从而使其聚类系数比较低。

考虑到很多真实网络当中同时具备模块性、无标度性以及局部高连接性的特征，学者提出节点集整合成为网络的过程类似一个循环迭代的过程，从而使网络成为一个层次性网络，见图 12-4(A1)。在此类网络中，聚类系数函数 $C(k)$正比于 k 的倒数，如图 12-4(C3)所示。

研究显示，不同的生物分子网络往往表现出相似的性质。大部分的真实生物网络如代谢网络、蛋白质互作网络、蛋白质结构域网络等都是无标度网络，其网络平均聚类系数都比具有同样大小和连通度分布的随机网络更高，且聚类系数均值正比于连通度的倒数，从而表明层次化是生物网络的一项基本性质。

四、网络模体

网络模体(network motif)是指网络中出现次数远超过随机期望的子网模式。这里子网模式是指一组节点按照特定的顺序连接而成的结构。针对不同网络的研究显示，在真实的网络中不同的子网模式出现的频率并不一致，有些模式在网络中频繁出现，远远超过随机网络中期望出现的次数。在某些网络中，特定出现的模体甚至是整个网络的基本组织形式，网络可以被看作是这些网络模体的组合。在生物学网络中，无论是有向网络还是无向网络，都包含有这些特殊的网络模体。在生物网络中搜索特殊模体有助于深入理解生物网络执行生物功能的基本形式，也有助于进一步从网络中发现节点间的功能联系。

(一) 有向网络模体

研究者从基因调控网络等有向网络中发现了一些特殊的模体，比较重要的有自调控环(auto-regulator loop，ARL)、前馈环(feed-forward loop，FFL)和单输入模体等(single input motif，SIM)。

自调控环模体包括正向自调控环，见图 12-5(A)，和负向自调控环，见图 12-5(B)，即转录因子促进或抑制自身转录的机制。在大肠杆菌(*E. Coli*)基因调控网络中存在较多的自调控模体。

前馈环模体则是在很多物种中常见的一种调控机制，即转录因子 A 调控转录因子 B 和基因 C，而同时转录因子 B 也调控基因 C，见图 12-5(C)。事实上，由于调控机制本身可以为正向和负向，前馈环还可以分为 8 种不同的类型。出现频率较多的有两种，一种是全部正向调控的一致前馈环，另一种是 A 正向调控 B 和 C，但 B 负向调控 C 的不一致前馈环。

单输入模体由同一个转录因子同时调控多个基因的表达，转录因子通常是自调控的，而所有调控符号(正、负向)都相同，且受控基因都不再受其他因子调控。这种模块在随机网络中并不多见，但在针对大肠杆菌(*E. Coli*)基因调控网络的分析中发现该模块经常出现在与蛋白质组装和代谢通路相关的基因调控中。在此类问题中，由一个转录因子控制参与生物过程基因的表达，能够有效地保持受控基因的比例，提高效率，见图 12-5(D)。

除上述模体外，研究者在调控网络中还

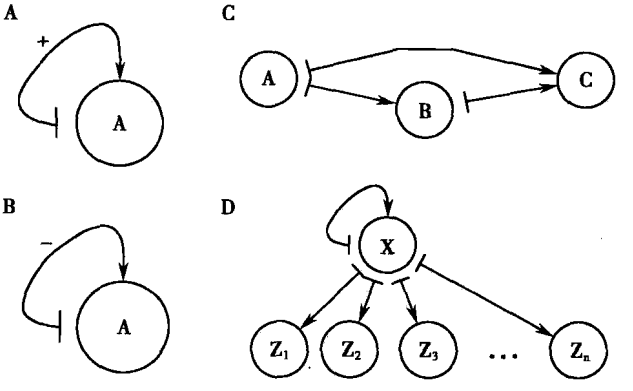


图 12-5 有向网络模体

A 为正向自调控环；B 为负向自调控环；C 为前馈环；D 为单输入模体

发现了其他一些模体，如密集重叠调控(Dense Overlapping Regulation，DOR)、多输入模体(Multi Input Motif，MIF)和调控链(Regulator Chain，RC)等。这些不同的网络模体结构代表了不同的转录调控机制，对这些模体的研究将极大地帮助人们了解生物过程的控制机制。

（二）无向网络模体

在无向网络中也可能存在一些特殊的模体，在生物网络中出现的频率远超过随机的情形，其中比较重要的是全连接集(clique)。

全连接集是指任意两点都被边连接在一起的子网。如果全连接集中包含 n 个节点，则称这个全连接集为 n- 全连接集，见图 12-6。

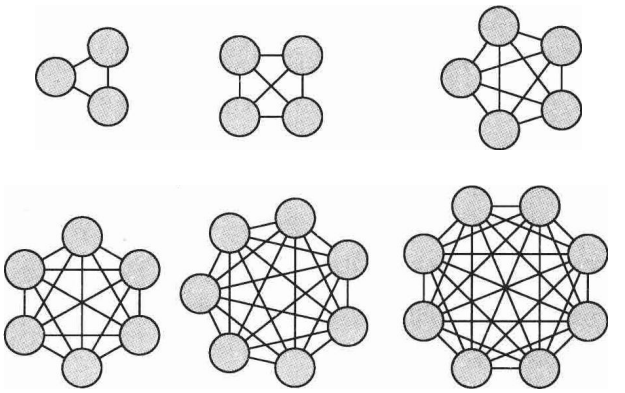


图 12-6 全连接集

（三）网络模体搜索算法

网络模体搜索算法指在网络中寻找与模体同构的子网的过程。要从一个包含 N 个节点的网络搜索模体的过程包括：①定义包含 k 个节点的子网模式；②在网络中搜索全部 C_N^k 个包含 k 个节点的节点子集，并检查其中结构与所搜寻的模式相符的个数；③将各个模式在真实网络中出现的频率和在大量随机网络背景中所出现的频数进行比较，从而发现网络模体。

这一过程在算法上是 NP 完全问题，即解决这一问题的计算时间可能需要花费多项式时间的计算问题。因此对包含节点数目更多的子网模式来说，网络中存在的组合数目比较多，待比较的类型也会更多，从而导致搜索的复杂度更高。因此，目前网络模体的搜索往往只针对一些较小的子网模式来进行分析。

理论上，模体的搜索也可以从边出发。在现实中，生物分子网络大多是比较稀疏的，也就是大部分的节点间不存在边的连接，因此，基于边的搜索会比基于节点的算法更快。

五、生物分子网络的动态性

生物分子网络并不是静态不变的。生物分子间发生相互关系需要特定的时间和空间条件。例如在富氧和缺氧状态下，葡萄糖的代谢途径并不相同；在应激反应中，生物体针对不同的外界刺激开启不同的信号通路予以应对；分子组装和能量代谢发生在特定的细胞器上。在不同的时间和空间，生物体执行着不同的生物过程。要揭示生命活动的真正过程，必须要考虑到生物分子网络的动态特性。

（一）含有时空信息的生物分子网络

基因芯片技术等针对特定实验条件的检测技术，提供了在特定时间和空间上生命活动的重要信息。通过对这些信息整理和分析，能够得到实验条件特异的生物分子网络。例如，利用一组在不同时间点获得的基因表达谱信息，可以构建表达相关网络，获取基因组中共同行使功能的基因集合，也可以构建基因调控网络，分析细胞循环过程中内在的调控机制等。

（二）整合时空信息的生物分子网络

生物分子网络的时空特异性是一项普遍存在的性质，即便是主要由一些非实验条件相关的检测技术所检测得到蛋白质相互作用信息，同样存在着时空特异性。蛋白质间相互作用的发生并非是静态而一成不变的，部分相互作用是稳定而持久的，还有一些相互作用则是在特定的时间与空间场合才会发生。

受检测技术的限制，蛋白质互作网络等生物分子网络的时空检测标准还不存在，但可以通过结合包含有明确时间或空间信息的其他实验技术所测的结果来为这些网络补充时空信息。例如，基因表达相关性可以为转录调控和蛋白质互作在相应条件下是否存在提供旁证。即在特定的实验条件下，转录因子编码基因及其靶基因的表达水平显示了表达调控的开放状态，一对互作蛋白质的表达

水平可以表明是否存在互作关系。由此可以构建特定实验条件下的转录调控网络和蛋白质互作网络。

（三）生物分子网络的动力学分析

生命过程是一个动态的过程，生物分子网络也不可避免地具有动态性的特征。通过结合带有时空性质的实验信息，挖掘在特定时间、空间和环境条件下的生物分子网络，从而更加准确地理解生物分子网络行使功能的方式，为进一步地科学分析提供更准确的研究基础。

基于生物分子网络的动态性质，既可以类似普通静态网络对网络属性进行统计分析，也可以针对网络进行仿真计算以分析网络的动力学问题。如在基因转录调控，信号传导和代谢等生物过程中，信息的传递和生物反应是一系列在时间和空间上连续的过程，这个过程也就可以被设定为网络节点状态和拓扑结构的一系列变化。通过结合基因表达等动态信息，利用线性模型、微分模型、随机过程等算法，可以构建出随时间、空间和环境条件等变化的动态生物分子网络，从而更为准确地描述、解释和预测生物过程。

六、生物分子网络分析软件

目前有很多软件被用于生物分子网络可视化展示和网络分析。其中包括一些可以依据 GNU 协议免费应用的软件，也包括一些商业软件。

（一）CytoScape 软件

CytoScape 是一款图形化显示网络并进行分析和编辑的软件，见图 12-7，它支持多种网络描述格式，也可以用以 Tab 制表符分隔的文本文档或 Microsoft Excel 文件作为输入，或者利用软件本身的编辑器模块直接构建网络。CytoScape 还能够为网络添加丰富的注释信息，并且可以利用自身以及第三方开发的大量功能插件，针对网络问题进行深入分析。

CytoScape 对非盈利性客户免费，下载网址：http://www.cytoscape.org。

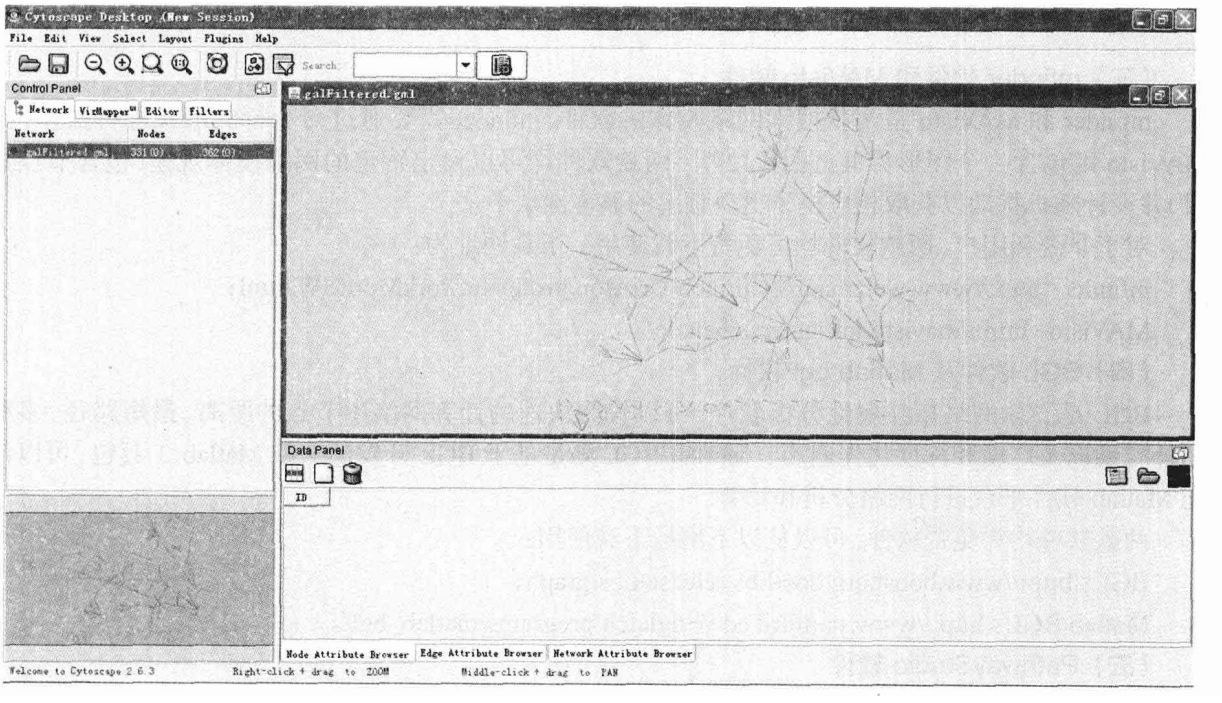


图 12-7 Cytoscape 工作界面

（二）CFinder 软件

CFinder 是一种基于全连接集搜索方法（the Clique Percolation Method，CPM）的网络密集集团模块搜索和可视化分析软件，见图 12-8。它能够在网络中寻找指定大小的全连接集，并通过全连接集

中共享的节点和边构建更大的节点集团。软件中可以使用以制表符分割的文本文件作为输入。算法主要针对无向网络，但也包含对有向网络的一些处理功能。

CFinder 允许非盈利性用户免费使用，并可以在 http://cfinder.org 免费下载和获取帮助。

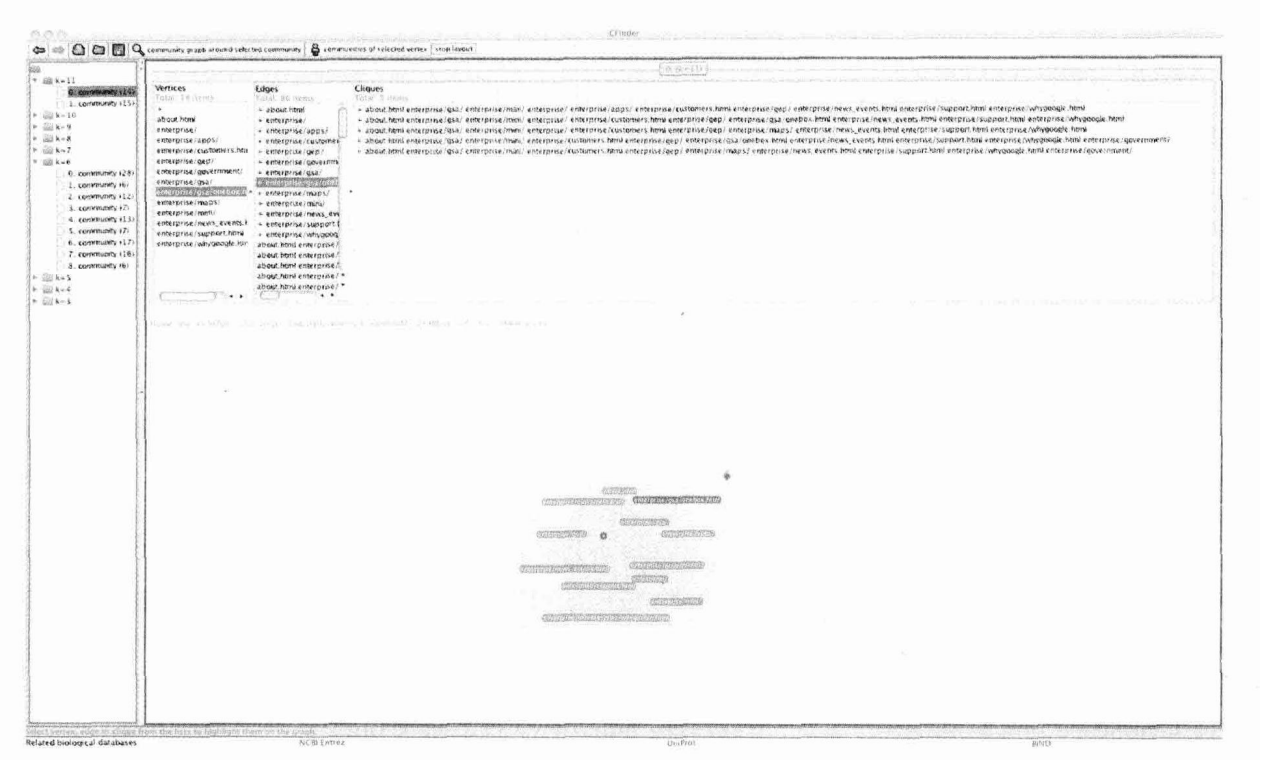


图 12-8 CFinder 工作界面

（三）mfinder 软件和 MAVisto 软件

mfinder 和 MAVisto 是两款搜索网络模体的软件，mfinder 需要通过命令行的形式进行操作，而 MAVisto 则包含一个图形界面，见图 12-9。两款软件均可以设定特定的网络模体规模（包含节点数目）并设计随机扰动以获取相应模体出现频率的显著性水平。

对于非盈利用户，两款软件均可免费下载使用。下载网址为：

mfinder：http://www.weizmann.ac.il/mcb/UriAlon/groupNetworkMotifSW.html；

MAVisto：http://mavisto.ipk-gatersleben.de/。

（四）BGL 软件及 Matlab bgl 软件

BGL 是一款网络拓扑属性分析软件，可以较为快速的计算网络中节点的距离、最短路径、多种拓扑属性以及广度和深度优先遍历。Matlab BGL 则是基于 BGL 开发的一款 Matlab 工具包，可以依托 Matlab 软件平台进行网络分析和计算。

两款软件均为免费软件，可以从以下网址下载使用：

BGL：http://www.boost.org/doc/libs/release/libs/graph；

Matlab BGL：http://www.stanford.edu/~dgleich/programs/matlab_bgl/。

（五）PathwayStudio 软件

PathwayStudio 生物通路可视化分析软件，是一款商业化生物信息学软件，见图 12-10，它能够以不同的模式绘制和分析生物通路，并且可以利用随带的 MedScan 软件通过公开发表的文献构建基于知识的生物通路网络。

（六）GeneGO 软件及数据库

GeneGO 是为系统生物学中的数据挖掘应用提供化学信息学和生物信息学软件解决方案的供应

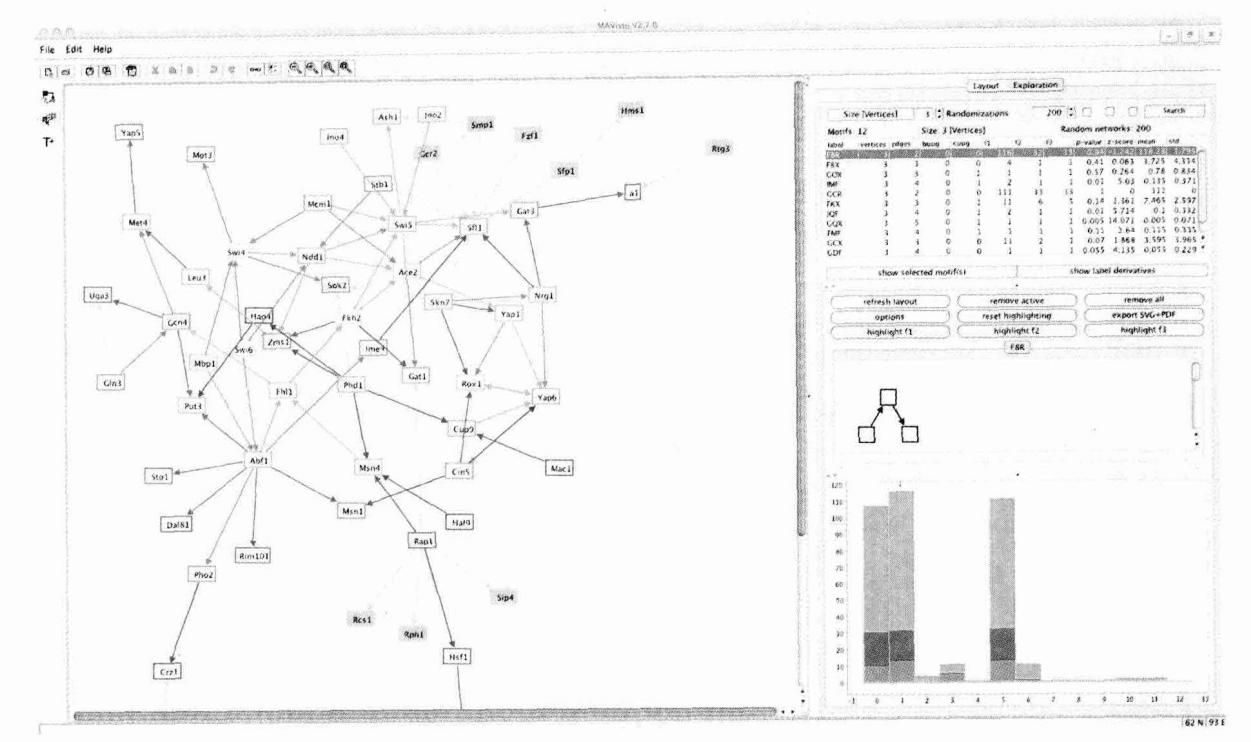


图 12-9 MAVisto 工作界面

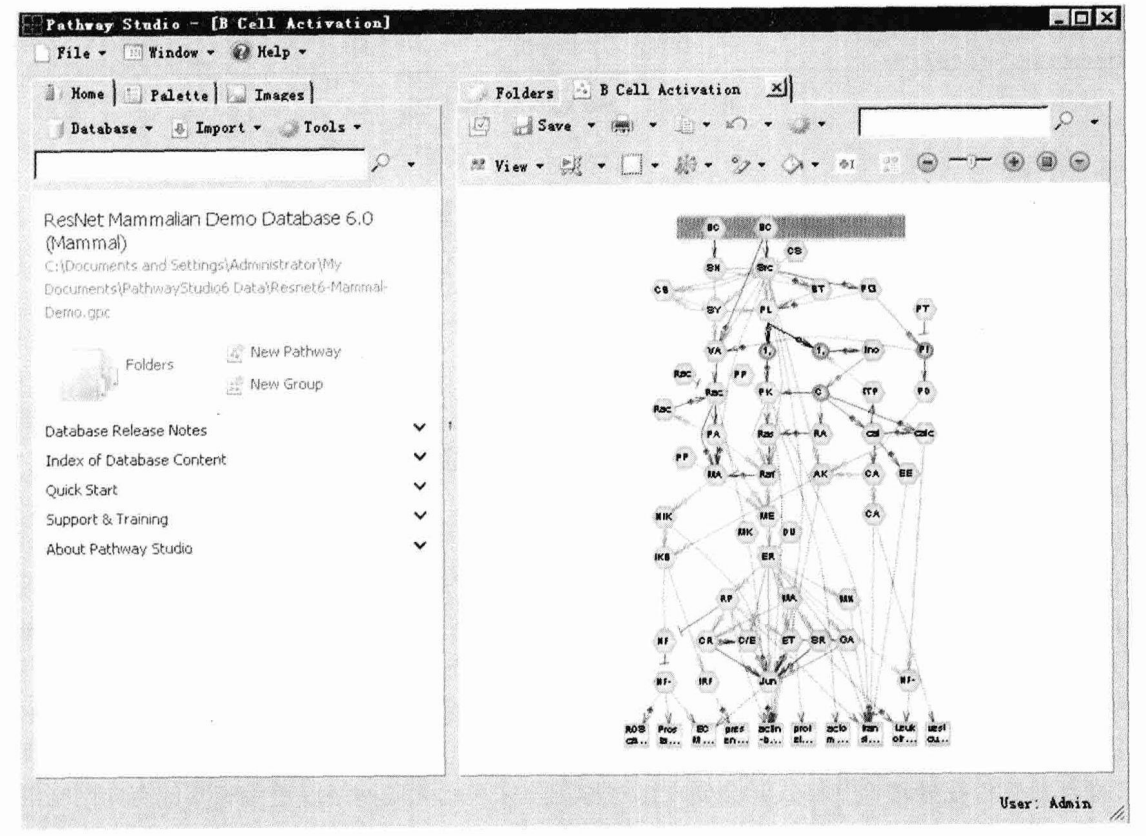


图 12-10 PathwayStudio 工作界面

商。其主要产品包括 MetaBase、MetaCore、MetaDrug 等。其中 MetaBase 是 GeneGO 专业研制的哺乳动物生物学与药物化学数据库。MetaCore(图 12-11)主要针对系统生物学研究中的通路分析和生物标记物发现提供了数据挖掘工具套件。MetaDrug 为 GeneGO 开发的系统药理学平台。

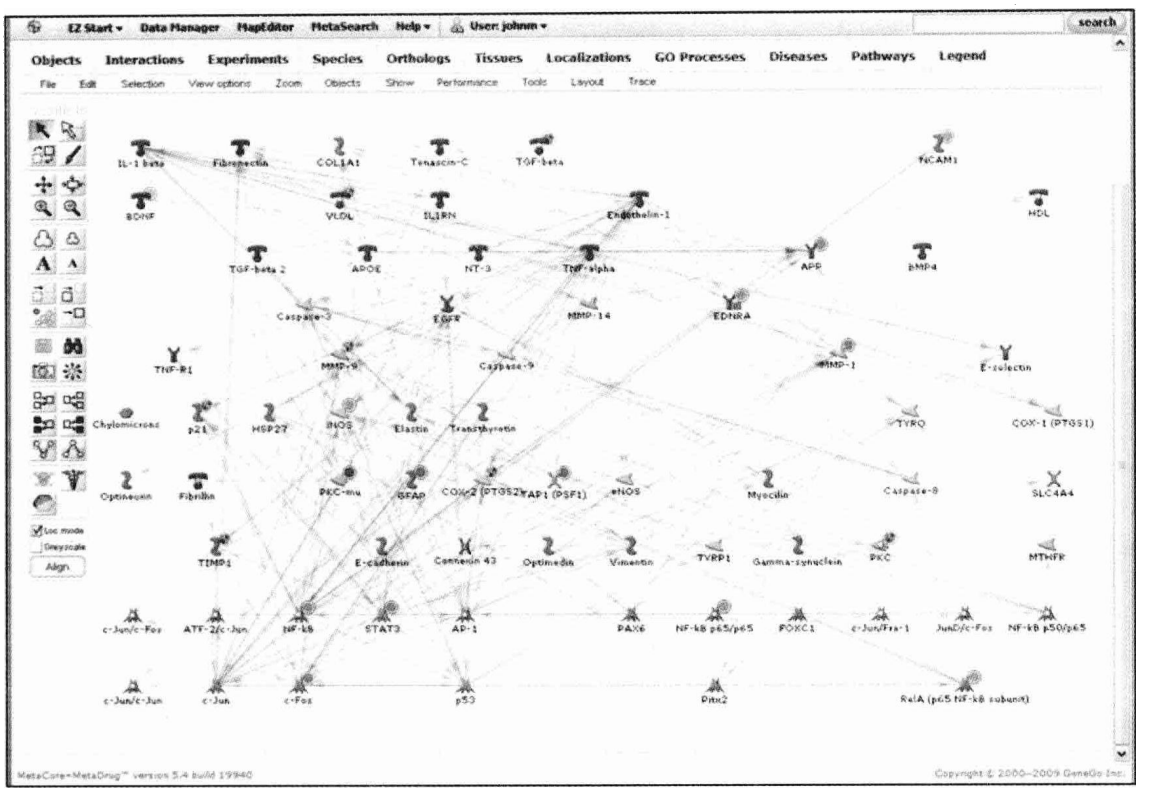


图 12-11 MetaCore 工作界面

第四节 生物分子网络的重构和应用

Section 4 Reconstruction and Application of Biology Molecule Network

一、生物分子网络重构的一般方法

高通量实验方法的出现和广泛应用为生物学分析提供了海量的数据资源，这些资源组织形式多样，包含信息丰富。基于这些数据信息，利用计算机技术重新构建网络，有助于综合分析数据，利用网络计算方法挖掘相关信息，从系统上分析生物分子网络。

(一) 网络的数据结构

在计算机中，存储网络的数据结构有很多形式，其中最常用的是连接矩阵表示法和边列表表示法。

1. 连接矩阵表示法 连接矩阵是一种比较直观的网络表示方法，通过构建与网络节点数目相同的方矩阵来表示网络。矩阵的每行表示有向网络中的源节点，每列则表示有向网络中的目标节点。矩阵中的非零元素代表一条由行节点指向列节点的边，而该元素的值则代表这条边的权重，而无权网络中往往取为 1(图 12-12)。对无向网络而言，矩阵表示法中的上三角阵(或下三角阵)即可表示整个网络，而部分软件在处理这种格式时会要求以对称矩阵来表示无向网络以避免和其他有向网络混淆。

连接矩阵表示法的缺点是占用较大的存储空间，由于在大型网络中，边的数量相对于可能存在的全部边数而言较少，网络矩阵中大部分元素都为 0。此时，只记录存在的边将会大大减少存储所需的空间。

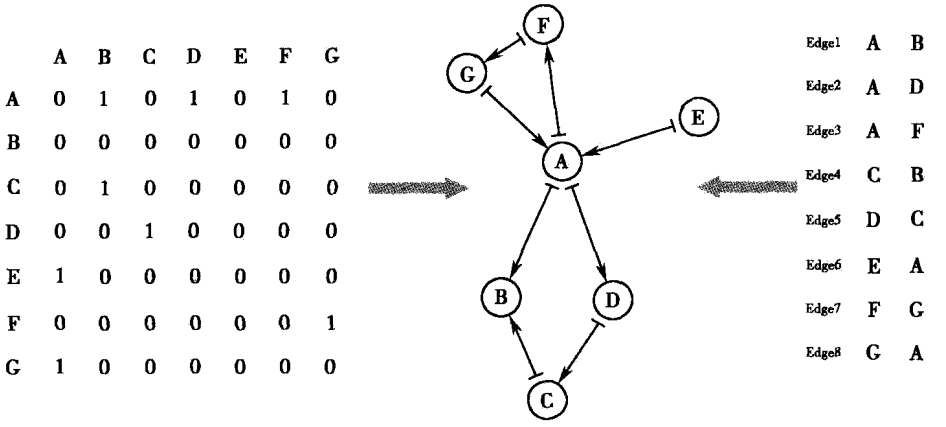


图 12-12 网络的数据结构

2. 边列表表示法 边列表表示法的记录方式一般包括两列数据，分别代表网络中的源节点和目标节点。每一行则代表一条由源节点指向目标节点的边(图 12-12)，还可以增加新的列表明边的类型、权重等信息。

3. 其他 用于存储网络的数据结构还有其他类型，如节点连接表形式，通过为每个节点保存一个可连接节点列表的方式记录网络；距离矩阵表示法，矩阵中每个元素记录其行和列所代表节点在网络中的距离等。

(二) 网络重构的方法

描述相互作用关系是生物网络最简单的功能，同时也是网络重构的简单方法。对于基本数据形式，即表示为两两关系的生物信息资源，如生物实验证实的转录因子与靶基因对应关系、蛋白质相互作用关系、药物分子与靶蛋白关系等，均可以将分子作为节点，分子间关系作为边，从而重构生物分子网络。

然而对于信号传导通路和代谢通路等形式复杂的功能网络，参与网络的生物分子类型多种多样，分子间关系复杂，以简单网络完整描述整个网络比较困难，因此需要对原信息进行过滤和抽象，在感兴趣的层次提取网络信息。例如，代谢网络中一种代谢物在酶的作用下转化为另一种代谢物并产生能量，在这个过程中，可以构建代谢物之间的转化网络，则酶与能量载体不以节点的形式出现在网络中；也可以将全部参与代谢通路的分子都作为节点，然后将其间发生的各种作用作为边，构建出更为复杂的网络。具体采用何种方式构建网络取决于构建网络的目的。

为了从实验数据中重构网络，需要通过数据统计或数据挖掘技术提取相应的作用关系。这种关系可能是简单的生物分子间是否存在连接，也可能是计算一系列定量指标，衡量分子间关系的紧密程度或可靠性等。

二、基因表达相关网络的重构和应用

DNA 微阵列、转录组测序等基因表达检测技术的广泛应用使研究者可以高通量并行研究大量基因在不同实验条件或细胞周期中的表达水平。为了完整系统地展示和分析基因间的共表达关系，可以构建基因表达相关网络。

基因表达相关网络可以以等权网络形式构建，构建步骤如下：

(1) 利用基因表达谱计算表达相关矩阵，得到任意两个基因间的表达相关性指标。其中表达相关性指标可以根据研究目的选用 pearson 相关系数、互信息或欧氏距离等。

(2) 选定阈值，获取显著相关的基因对。阈值的选定可以采取选定特定百分比、指标统计推断或者重排表达谱构建随机背景以获取显著性阈值等方法。

(3) 以相关性超过阈值的基因对作为边，基因作为节点，构建基因表达相关网络。

基因表达相关网络可以是等权的也可以以相关系数或由相关系数决定的函数作为权值，构建加权网络。通过对基因表达相关网络的分析，可以研究基因间的功能联系，进而获得在特定实验条件下的功能相关集合，也可以结合其他生物分子网络，构建实验条件特异的动态生物分子网络。

三、基因调控网络的重构和应用

基因调控网络中的节点包括转录因子和受控基因，如果受控基因的产物也是转录因子，往往会将受控基因及其产物视为同一个节点。由此，基因调控网络是一个有向网络，每条边由转录因子指向受控基因。从重构的方式来看，基因调控网络包括基于原始数据的网络和基于表达数据的网络。

（一）基于原始数据的基因调控网络

ChIP 等技术直接测得转录因子是否与 DNA 结合，因此可以比较简单的将转录因子作为源节点，受控基因作为目标节点，构建基于原始数据的有向基因调控网络。

例如，在 ChIP-chip 实验中，经过基因芯片处理后，每一个元件（基因或 DNA 区域）都对应一个强度值，反映了其经过特定感兴趣蛋白免疫共沉淀处理后的富集水平。对于双通道芯片，这个强度值常表现为处理组与控制组的强度比值或配对 t 统计量；而对单通道芯片，则可以表示为处理组与控制组的两样本 t 统计量。通过中值百分位数顺序法（median percentile rank），单芯片误差模型（the single-array error model）和移窗法（sliding-window approach）等数值和统计方法，就可以得到 DNA 区域与感兴趣蛋白之间发生结合互作的富集程度分值或概率分值。通过设定阈值的方式能够筛选出显著的蛋白质 -DNA 二元互作关系。由此即可得到由蛋白指向相应基因或 DNA 区域的边，整合这些互作关系，即可以重构基因调控网络，见图 12-13。

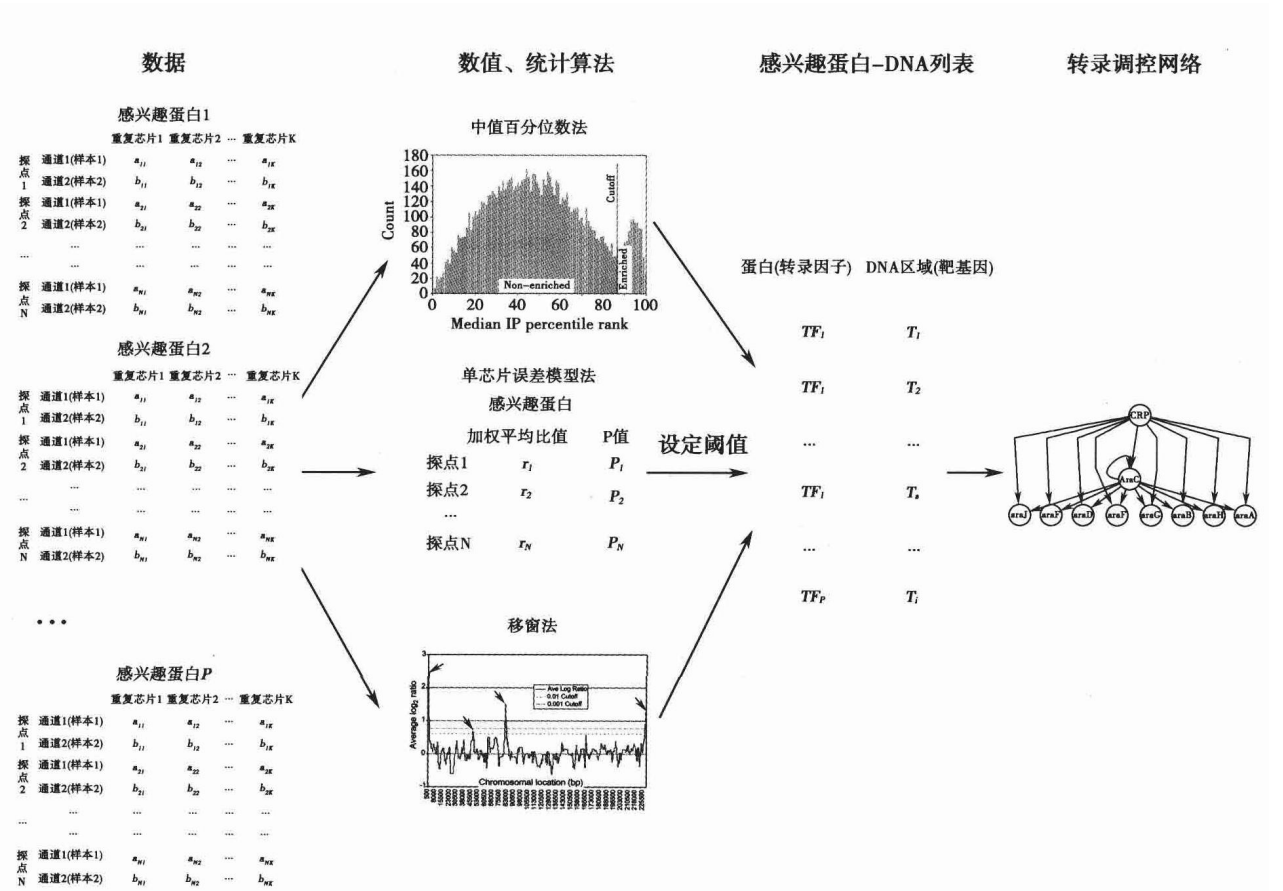


图 12-13 由 ChIP-chip 数据重构基因调控网络步骤

（二）基于表达数据的基因调控网络

基因转录在基因表达环节中起着非常重要的作用。例如受同一个转录因子调控的基因往往是共表达的，这些生物学原理可以用于指导基因调控路径的构建。因此为了弥补基因转录调控检测所得数据缺乏的缺陷，可以从反映了基因转录的调控机制的 DNA 微阵列基因表达谱数据出发挖掘基因转录调控关系。

利用基因表达信息等高通量数据挖掘基因调控关系，并以此重构基因调控网络对基因调控的研究有着重要的意义。最常用的构建基于表达数据的基因调控网络模型包括布尔网络模型、加权矩阵模型、线性组合模型等。

1. 布尔网络模型　基于表达数据重构的基因调控网络的一种最简单的模型就是布尔网络模型。在布尔网络中，每个节点代表一个基因，或者代表一个环境刺激。环境刺激可以是任何影响调控网络的生物、物理或化学因素，而不是基因或基因的产物。每条有向边代表基因之间的相互作用关系。当一个节点代表基因时，该节点与一个稳定的表达水平相联系，表示对应基因产物的数量。如果一个节点代表环境因素，则节点的值对应于环境刺激量。各节点的值或者是“1”，或者是“0”，分别表示“高水平”和“低水平”。

其中节点之间的相互作用关系可以由布尔表达式来表示，例如：

$$A \cap (\neg B) \rightarrow C \qquad \text{式 12-7}$$

读作“如果 A 基因表达，并且 B 基因不表达，则 C 基因表达”其中∩表示逻辑上的并且关系“and”，¬表示否定关系“not”。在网络上则可以表示为图 12-14(A)。布尔网络中的作用关系与上文所讨论的调控关系相比，增加了对多个因子综合作用（“并”，“或”，“与或”关系）的考虑，这种基于关系的信息输入称为连接。考虑网络中全部节点间的相互作用关系后就得到了如图 12-14(B)所示的布尔网络。当布尔网络中每个节点被赋予初值后，网络中的节点即能够自动的对下一个状态进行预测。这一过程可以被理解为布尔网络转化为一种接线图，见 12-14(C)。使用这种方法能够推导出下一步各节点的值，并通过迭代的方法获得以后各步运算的结果。经过迭代后网络出现了稳定状态，但由于初值的影响，稳定形态并不相同，如图 12-14(D)中，当选定初值为 A=0，B=0，C=0 时，一步迭代后网络各节点的值便稳定在 A=0，B=1，C=0 上。而当选定初值为 A=1，B=0，C=0 时，迭代结果则在第二次迭代时出现循环，结果始终在初值与 A=0，B=1，C=1 之间反复切换。

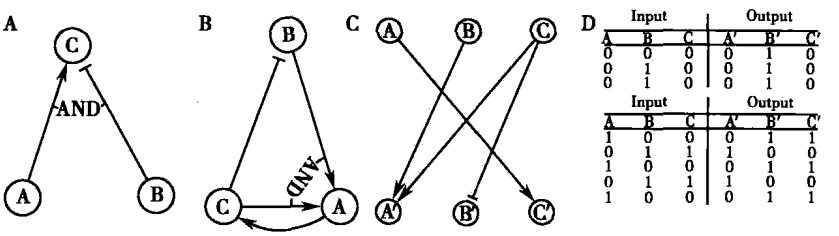


图 12-14　节点 C 的真值表

以上实验表明布尔网络能够模拟生物体的表达调控机制，布尔网络的迭代结果趋于稳定，类似于生物体在适应环境后的稳定状态。但在真实的问题中，布尔网络不是预先知道的。相反的，表达谱等数据信息能够为布尔网络模型提供前后不同时刻各个节点的取值。布尔网络模型分析的目的就是从多时刻的基因表达谱信息构建出特定环境下基因表达调控的网络结构。

基于表达数据使用机器学习或者其他智能训练的方法可以构建一个布尔网络。其基本原理是根据基因表达的实验数据建立待研究的基因之间的相互作用关系，确定每个基因的连接输入（或调控输入），并且为每个基因生成布尔表达式，或者形成网络系统的状态转换表。对于复杂的网络，在网络构造过程中，其搜索空间非常大，需要利用先验知识或合理的假设，以减小搜索空间，从而有效地构造布尔网络。

布尔网络模型简单，便于计算，但是由于它是一种离散的数学模型，不能很好地反映细胞中基因表达的实际情况，如，布尔网络不能反映各个基因表达的数值差异，不考虑各种基因作用大小的区别等。而在连续网络模型中，各个基因的表达数值是连续的，并且以具体的数值表示一个基因对其他基因的影响。

2. 线性组合模型　线性组合模型是一种连续网络模型，在这种模型中，一个基因的表达值是若干个其他基因表达值的加权和。基本表示形式为：

$$X_i(t+1)=\sum_j w_{ij}X_j(t) \qquad \text{式 12-8}$$

其中，$X_i(t+1)$是基因 i 在 $t+1$ 时刻的表达水平，$X_j(t)$是基因 j 在 t 时刻的表达水平，而 w_{ij} 代表基因 j 的表达水平对基因 i 的影响。在这种基因相互关系表示形式中，还可以增加一个常数项，反映一个基因在没有其他调控输入下的活化水平。

将上述表达式转换为线性差分方程，描述一个基因表达水平的变化趋势。这样，在给定一系列基因表达水平的实验数据之后，即给定每个基因的时间序列 $\{X_i(t)\}$，就可以利用最小二乘法或者多重分析法求解整个系统的差分方程组，从而确定方程中的所有参数，即确定 w_{ij}。最终，利用差分方程分析各个基因的表达行为。实验结果表明，该模型能够较好地拟合基因表达实验数据。

3. 加权矩阵模型　加权矩阵模型与线性组合模型相似，在该模型中，一个基因的表达值是其他基因表达值的函数。含有 n 个基因的基因表达状态用 n 维空间中的向量 $u(t)$表示，$u(t)$的每一个元素代表一个基因在时刻 t 的表达水平。以一个加权矩阵 W 表示基因之间的相互调控作用，W 的每一行代表一个基因的所有调控输入，w_{ij} 代表基因 j 的表达水平对基因 i 的影响。在时刻 t，基因 j 对基因 i 的净调控输入为 j 的表达水平(即 $u_j(t)$)乘以 j 对 i 的调控影响程度 w_{ij}。基因 i 的总调控输入 $r_i(t)$为：

$$r_i(t)=\sum_j w_{ij}u_j(t) \qquad \text{式 12-9}$$

这一形式与线性组合模型相似，若 w_{ij} 为正值，则基因 j 激活基因 i 的表达，而负值表示基因 j 抑制基因 i 的表达，0 表示基因 j 对基因 i 没有作用。与线性组合模型不同的是，基因 i 最终表达响应还需要经过一次非线性映射：

$$u_i(t+1)=\frac{1}{1+e^{-[\alpha_i r_i(t)+\beta_i]}} \qquad \text{式 12-10}$$

这种函数是神经网络中常用的 Sigmoid 函数，其中 α 和 β 是两个常数，规定非线性映射函数曲线的位置和曲度。通过上式计算出 $t+1$ 时刻基因 i 的表达水平。在最初阶段，加权矩阵的值是未知的。但是可以利用机器学习方法，根据基因表达数据估计加权矩阵中各个元素的值。

对于这样的模型，可以利用成熟的线性代数方法和神经网络方法进行分析。实验表明，该模型具有稳定的和周期稳定的基因表达水平，与实际生物系统相一致。在这种模型中还可以加入新的变量，模拟环境条件变化对基因表达水平的影响。

除上述模型外，还可以利用贝叶斯网络模型等方法由表达数据等信息重构基因调控网络，这些模型可以用于预测和验证基因间的转录调控关系，也为分析基因功能，研究生物信号与功能传递机制提供了重要的信息资源。

四、蛋白质互作网络的重构和应用

由于蛋白质相互作用数据本身可以提供蛋白质与蛋白质间相互作用关系信息，蛋白质互作网络的构建比较简单，只需将数据中的互作关系作为网络中的边，蛋白质作为网络中的节点，即可以重构蛋白质互作网络。

蛋白质互作网络是无向的无标度网络，少数连通度极大的节点将高度模块化的子网连接在一起。在对酵母等模式生物的分析中，几乎覆盖整个蛋白质组的蛋白质互作网络已经被重构出来，人类等高等物种的互作数据也在以极快的速度积累着。目前，酵母、小鼠、人类等物种的蛋白质互作网

络一般包含有数千个节点和数千到数万条边。在这样庞大的网络上，需要采用多种多样的计算方法对蛋白质网络进行分析。

（一）蛋白质网络的可靠性分析

目前，高通量的蛋白质互作检测技术和生物信息学预测方法极大地丰富了蛋白质互作数据资源，为进一步网络分析提供了数据基础，同时高通量检验结果中包含着大量不确定的结果，存在着严重的假阳性问题，因此确定数据的可靠性成为蛋白质互作网络分析前一项重要的工作。

一般认为，小规模生物实验所检测出的互作信息更为可靠。免疫共沉淀的阳性检测结果一般可以被作为互作存在的金标准。而当互作实验证据来自高通量实验时，往往用同一条互作信息在不同的高通量实验所证明的次数来反映互作信息的可靠程度。此外，还可以通过结合表达相关性等与互作关系密切相关的其他数据信息来检验互作信息的可靠性。

（二）基于蛋白质网络的蛋白功能预测方法

蛋白质通过彼此的连接来行使生物学功能，因此，存在一个很自然的假设，即彼此互作的蛋白具有相同或相近的功能。基于这一假设，开发了一系列基于蛋白质网络的蛋白功能预测方法。

这些方法中，邻居计数法是一种最简单的方法，即一个待测功能的蛋白质应同与其连接的大部分蛋白的功能一致，通过统计它的邻居中属于不同功能的蛋白数目，将计数最多的功能作为对待测蛋白的预测。

在邻居计数法的基础上，研究者又开发出了包括考虑功能类别本身规模影响的卡方法、结合全局信息的网络分割算法、基于不同概率模型的全局预测算法等。虽然这些方法普遍存在着预测准确率不高，预测范围有限等缺点，然而由于不需要利用同源信息，且其预测效率能够随互作信息和功能信息的完善而不断提高，因此，这类算法具有重要的意义。

（三）模体的搜索和分析

蛋白质互作网络中包含有大量的密集互作的子网模式，模体的出现暗示了生物分子行使生物功能的基本模式，挖掘这些模式对了解蛋白质如何行使功能，探索蛋白质间的功能联系以及寻找新的功能通路都有着重要的意义。

全连接集是蛋白质网络中普遍存在的一类模体，无论是各种全连接集出现的频率，还是最大全连接集的规模，真实的蛋白质网络都远远超过随机网络，从而说明，组成高度连接的蛋白质复合物是蛋白质行使生物功能的一种基本形式。研究发现，部分全连接集之间存在重叠，将这些存在重叠的全连接集合并在一起，可以获得密集互作的蛋白质子网。结果显示这些子网中往往存在功能上的关联性。另一方面，很多研究显示连通度最高的蛋白质往往没有出现在规模最大的全连接集中，表明此类高连通度的蛋白质节点更倾向位于连接不同模块的枢纽位置，而不是直接参与大型的蛋白质复合物。

此外，挖掘其他类型的蛋白质互作网络模体同样对于理解蛋白质行使功能的模式有着重要的意义。有研究通过整合基因表达相关性与蛋白质互作信息，以在互作网络中搜索高表达相关路径的方式预测潜在的生物通路等。

由多种互作关系所组成的复合生物分子网络存在特异性的模体，例如将蛋白质互作网络与表达调控网络结合来搜索其中的模体，可以获得在复合网络中显著富集的调控互作模块。

（四）基于拓扑属性的分析

利用拓扑属性分析网络中的节点是网络生物学中独特的方法。在蛋白质互作网络中，具有独特拓扑属性的节点蛋白往往具有独特的生物学意义。

研究显示，连通度较高的中心节点（在蛋白质互作网络中常以连通度大于5的节点作为中心节点）对网络的连通性起着特别重要的意义，其中显著富集着与生命基本活动相关的必需基因、疾病相关基因以及药物靶点基因等具有重要意义的基因的表达产物。中心节点的这种特点使得它们成为很多研究所关心的对象。介数和紧密度较高的节点往往也具有较高的连通度，这些节点在连接网络过

程中同样具有重要作用。

节点的拓扑属性是揭示节点在网络中意义的重要工具，通过对蛋白质互作网络中节点拓扑属性的分析，能够进一步理解其在生物网络中的重要作用。还可以利用模式识别方法对特定蛋白节点的功能进行预测。

五、代谢网络的重构和应用

生物代谢网络是一种较为复杂的网络，其原因在于其中包含的分子类型众多，反应类型多样。一个反应往往不是简单的两个生物分子的作用，而是以多个分子组成临时复合物的形式连接在一起。因而构建代谢网络的第一步重要工作是选择适当的水平来构建代谢网络。

（一）代谢网络的重构

根据选定代谢网络分子的类型可以将重构的代谢网络分为多反应物网络和主要反应物网络。多反应物网络是比较直观的一种构建代谢网络的方式。代谢通路中的反应的参与者主要是代谢底物和酶，此外还有一些其他的共反应因子。多反应代谢网络包含由主要代谢底物指向代谢产物的转化关系，也包含酶与底物以及不同底物分子之间的相互作用关系。参与反应的生物分子被设定为节点，转化关系和催化作用作为边，有时在反应中临时形成的媒介复合物也被作为网络的节点。由此构建的网络包括$(N+E+R)$个节点，其中N表示底物数量，E表示酶数量，R表示媒介复合物的数目。由于酶反应经常具有双向催化性，网络中边的方向需要在具体情况下分别考虑。

主要反应物网络，则忽略了参与反应的共反应因子，而直接以底物产物关系为网络的边，代谢反应中的底物和产物作为节点构建代谢网络。当考虑反应发生的主要方向时，网络可以设定为有向网络，否则也可将代谢网络作为无向网络。

此外，还可以从其他的角度着手构建代谢相关网络，如以酶作为节点，以在两个反应中分别作为底物和产物的代谢物作为关系将不同的酶联系起来构成酶关联网络，可以用于分析酶参与反应的相关程度和不同酶在代谢通路中所发挥的作用。

（二）代谢网络一般特征

代谢网络是最早发现无标度特性和层次化特性的生物分子网络。在针对数十个物种的分析中，代谢网络表现出了类似的无标度特性和层次化特性。作为有向网络，无论是代谢网络的出度还是入度都表现出了幂率分布$P(k)\sim k^{-r}$的特点。经分析在大肠杆菌(*E. coli*)中，无论是作为产物的分子的入度分布还是作为底物参与反应数目的出度分布都有$r=2.2$左右。

不同物种的代谢网络在大尺度特征上显示了高度的一致性，这说明代谢网络的特征不是随机的。高度模块化和无标度的属性的一个直接反应是代谢网络的直径远小于随机网络，这就使生物体对外界环境变化以及内部突变所做出的反应更为迅速和有效。

小　结

蛋白质、DNA、RNA以及生物小分子等细胞成员之间的相互作用是大多数生物功能发生的基本方式。系统研究活体细胞内生物分子及其间的相互作用是后基因组时代的一个重要目标。生物分子网络是研究和分析复杂生物分子系统的重要工具。

高通量的生物学检测技术产生了大量的信息资源，充实了各种生物信息学数据库。基于不同物种、不同类型的生物分子，出现了各种生物分子网络。其中最重要的是蛋白质互作网络，基因转录调控网络，代谢网络和信号传导网络。

无标度性是生物分子网络表现出的特殊网络性质之一。生物分子网络的连通度分布一般都服从幂率分布，与随机网络完全不同。生物分子网络的无标度特性是生物在进化过程中形成的特性，并有助于生物适应周围的环境。

生物分子网络的平均聚类系数远高于随机网络，网络中连通度高的节点往往具有较低的聚类系数，生物分子网络是高度层次化的。

网络模体是生物分子网络中出现频率显著高于随机网络的特定连接模式。通过对网络模体的搜索，有助于了解生物分子行使功能的基本方式。

生物分子网络的重构依赖于生物分子数据的组织形式，部分可以直接由原始数据构建，部分需要通过机器学习技术从生物数据中提取。

快速发展的网络生物学提供了研究生物学和疾病病理学的新视角，为解决复杂生物医学问题提供了新的途径。

Summary

Most biological characteristics arise from complex interactions between the cell's numerous constituents, such as protein, DNA, RNA and small molecules. To systematically study all molecules and their interactions within a living cell is one of the most important targets of postgenomic biomedical research. Biomolecular network is one of the major tools to analyze the complex biomolecular system.

Abundant bioinformatics sources generated by the high-through detected technology are stored in all kinds of bioinformatics dataset. Based on different organisms and different types of biology molecules, various kinds of biomolecular networks were reconstructed. The most important networks include protein-protein interaction network, gene regulatory network, metabolic network and signal transfer network.

"Scale free" is one of the universe signatures of biomolecular network. The degree of biomolecular network generally obeys the power distribution. Totally difference to the random network, "scale free" is one of the signatures during the biological evolutionary process and helps creature to adapt the surroundings.

The average clustering coefficient of biomolecular network is much higher than that of random network. The node with the higher degree is generally with the lower clustering coefficient. Those explain the high hierarchy of the biomolecular network.

Network motifs are the certain connected modes, which present significantly higher frequency in the biomolecular network than in the random network. Searching the network motifs will help to understand the essence manners of biomolecular function.

The reconstruction of the biomolecular network depends on the organism format of the biomolecular data. Part of them can be constructed by the original data, and part need to pick-up by biology data.

The development of the Network Biology provides the new view to study biology and pathology and the new approach for solving the complex biomedical problem.

（童隆正 高 磊 姜 伟）

习 题

1. 哪些生物分子网络通常是无向网络？
2. 如何通过网络拓扑属性分析节点在网络中的作用？

3. 请找出图 12-1(A)中任意两点间的最短路径，并思考在包含更多节点的网络中应如何寻找网络的最短路径。

4. 什么是中心节点(hub)？生物分子网络中的中心节点有什么特点？

5. 计算图 12-1(A)中各节点的连通度、聚类系数。

6. 什么是无标度网络？请举出一个例子。

7. 请用 Barabási-Albert 模型构建一个无标度网络，并验证它的各项性质。

8. 如何判断一个网络是层次网络？

9. 访问一个数据库，并获取数据，重构相应的生物分子网络。

主要参考文献

1. Barabasi A. L., Oltvai Z. N. Network biology: understanding the cell's functional organization. Nat. Rev. Genet., 2004, 5(2): 101-113.
2. Han J. D., Bertin N., Hao T., et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004, 430(6995): 88-93.
3. Jeong H., Tombor B., Albert R., et al. The large-scale organization of metabolic networks. Nature, 2000, 407(6804): 651-654.
4. Barabasi A. L. Scale-free networks: a decade and beyond. Science, 2009, 325(5939): 412-413.
5. Shen-Orr S. S., Milo R., Mangan S., et al. Network motifs in the transcriptional regulation network of Escherichia coli. Nat. Genet, 2002, 31(1): 64-68.
6. Palla G., Derenyi I., Farkas I., et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435(7043): 814-818.
7. Han J. D., Dupuy D., Bertin N., et al. Effect of sampling on topology predictions of protein-protein interaction networks. Nat. Biotechnol., 2005, 23(7): 839-844.
8. Shannon P., Markiel A., Ozier O., et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 2003, 13(11): 2498-2504.
9. Schreiber F., Schwobbermeyer H. MAVisto: a tool for the exploration of network motifs. Bioinformatics, 2005, 21(17): 3572-3574.
10. Uetz P., Giot L., Cagney G., et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 2000, 403(6770): 623-627.
11. Ho Y., Gruhler A., Heilbut A., et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature, 2002, 415(6868): 180-183.
12. Reguly T., Breitkreutz A., Boucher L., et al. Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J. Biol., 2006, 5(4): 11.
13. Xu J., Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 2006, 22(22): 2800-2805.
14. Yildirim M. A., Goh K. I., Cusick M. E., et al. Drug-target network. Nat Biotechnol, 2007, 25(10): 1119-1126.

第十三章 计算表观遗传学
CHAPTER 13 COMPUTATIONAL EPIGENETICS

第一节 引 言
Section 1 Introduction

表观遗传学是研究不涉及DNA序列改变的情况下，DNA甲基化谱、染色质结构状态和基因表达谱在细胞之间传递的遗传现象的一门科学。表观遗传信息模式在细胞分裂过程中稳定的遗传，使表观遗传调控成为决定细胞分化和细胞命运的关键机制。随机和环境诱导的表观遗传缺陷可引发衰老和癌症，也可能导致精神紊乱症和自身免疫性疾病等。高通量分子生物学实验技术的发展及其在表观遗传研究中的应用，如ChIP-chip、ChIP-Seq等，产生了许多高通量的表观遗传学数据，这些数据的处理与分析为生物信息学带来了巨大的挑战。基于计算方法对表观遗传事件的预测在解决表观遗传问题上起到了重要的作用。计算表观遗传学即是把生物信息学的研究策略和方法应用到表观遗传学的研究领域，具有快速、高通量、低成本的特点，可以为当前的表观遗传学的实验研究提供指导；同时，生物学实验可以用来验证运用计算表观遗传学方法推导的结论。结合实验方法和计算表观遗传学方法，是当前表观遗传学研究领域新兴的视角。计算表观遗传学所研究的内容主要是通过生物信息学技术储存管理大量的实验数据并开发适用于深度挖掘这些实验数据的生物信息学算法，有利于促进发育和疾病等的表观遗传调控机制的研究。

知 识 拓 展

1939年，Waddington CH首先在《现代遗传学导论》中提出了表观遗传学(epigenetics)，在表观遗传学命名方面做出了贡献。1942年，他把表观遗传学定义为“生物学的分支，研究基因与决定表型的基因产物之间的因果关系”。1979年，Holliday对表观遗传学进行了较为准确的描述，简单地说表观遗传是非DNA序列改变的核遗传。近几年，表观遗传学已经成为生命科学领域的研究热点之一，在“后基因组”时代，理解表观遗传学的运作机制可以帮助解决生物学和人类疾病中存在的疑问，例如肿瘤发生、再生及衰老等。

第二节 基因组的DNA甲基化
Section 2 Genome-wide DNA Methylation

一、CpG岛的DNA甲基化调控基因的表达

(一) DNA甲基化与CpG岛

DNA甲基化(DNA methylation)是一种发生在DNA序列上的化学修饰，可以在转录及细胞分裂

前后稳定地遗传。DNA 甲基化是重要的表观遗传修饰之一。

1. DNA 甲基化　在哺乳动物中，大约 60%~90% 的 CpG 二核苷酸是甲基化的，其中的 p 代表连接脱氧胞嘧啶核苷和脱氧鸟嘌呤核苷的磷酸基团。非甲基化(non-methylated)的 CpG 二核苷酸聚集成簇形成所谓的 CpG 岛(CpG islands，CGIs)。在哺乳动物细胞中，DNA 甲基化主要发生在 CpG 二核苷酸中胞嘧啶的第五位碳原子上，这样的胞嘧啶也叫做 5- 甲基 - 胞嘧啶(5mC)，其化学式如图 13-1 所示。在植物中，胞嘧啶可以在 CpG，CpNpG 和 CpNpN 环境中发生甲基化(这里 N 代表除鸟嘌呤外的其他碱基)。在真菌中，胞嘧啶的甲基化水平较低，大部分只有 0.1%~0.5%，有证据表明真菌的 DNA 甲基化可能控制状态特异的基因表达。由于甲基化最早是细菌用以识别自身的化学修饰，这种表观遗传代码可能是古细菌对其他生物感染后的持续遗留的产物。

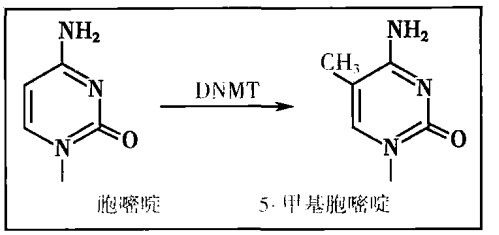


图 13-1　DNA 甲基化的机制

2. 催化 DNA 甲基化的生物酶　DNA 甲基化修饰发生时，甲基基团的添加主要有两类酶参与：甲基化维持酶和从头甲基化酶。在每个 DNA 复制周期中，DNA 甲基化的保持是维持甲基化活性所必需的。如果没有 DNA 甲基转移酶(DNA methyltransferase，DNMT)，复制结束后，子链甲基化修饰将会被动地缺失。DNMT1 是一种甲基化维持酶，它负责 DNA 复制过程中子链的甲基化模式的保持。DNMT3a 和 DNMT3b 都是从头甲基化酶，它们负责在发育早期建立 DNA 甲基化模式。DNMT3L 是类似于其他 DNMT3 的蛋白，虽然并没有催化活性，DNMT3L 通过与 DNA 的结合来辅助从头甲基转移酶。

3. CpG 岛与 DNA 甲基化的关系　CpG 二核苷酸倾向于聚集成簇，这样的区域称作 CpG 岛。CpG 岛的主要特点是 GC 的含量及 CpG 的含量非常高且大部分是处于非甲基化状态。CpG 岛覆盖了人类基因组大约 0.7% 的区域，但是却包含了所有 CpG 二核苷酸的 7%。CpG 岛主要分布在基因的 5' 非编码区、启动子和第一外显子区域，大约 60% 的基因的启动子含有 CpG 岛。这些区域的 CpG 二核苷酸的富集表明它们处于非甲基化状态(至少在生殖细胞中)，因此避免甲基化 CpG 带来高的突变率(这种突变是由于基因组的错配修复系统可以精确地识别并修正胞嘧啶碱基的脱氨基产物，而甲基化胞嘧啶的脱氨基产物则不被识别而发生缓慢的突变而转变为胸腺嘧啶，如果这种突变发生在生殖细胞中则是可遗传的)。尽管有研究认为 CpG 岛就应该是非甲基化状态，但是也有一些 CpG 岛在发育过程中被选择性地甲基化。哺乳动物基因组范围的研究表明大量 CpG 岛在终末分化细胞中是甲基化的。此外，大量的 CpG 二核苷酸处于重复元件中，但是它们在体细胞中被高度甲基化。

(二) DNA 甲基化对转录的调控

现已明确 DNA 甲基化的发生与转录沉默有关。DNA 甲基化参与的许多生物学过程都可以影响转录，其中一个公认的观点是 DNA 甲基化可以直接阻挡转录因子结合到 DNA 序列的靶点上而阻碍转录(图 13-2)。

1. DNA 甲基化阻碍转录因子的结合　许多转录因子倾向于结合包含 CpG 的序列，这些序列的 CpG 甲基化会阻止转录因子的结合(图 13-2)。c-Myc 是在细胞生长和分化过程中负责调控的转录因子，凝胶电泳实验表明 DNA 甲基化阻止 c-Myc 与它亲和的序列的结合。此外，在缺失染色质或甲基结合蛋白的情

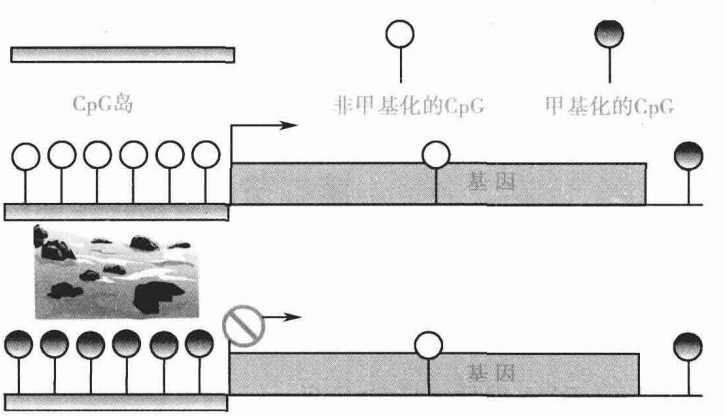


图 13-2　CpG 岛甲基化和转录的关系

况下，DNA 即使被甲基化也可以正常转录，这表明其他一些机制也可以导致基因沉默。

2. DNA 甲基化识别染色质标记 DNA 甲基化可以通过识别活性染色质标记而阻止转录的进行。通过 H3K4 甲基转移酶家族的催化，沉默的基因可以通过 H3K4 的甲基化而得以激活。一些 H3K4 甲基转移酶被认为靶向到包含 CpG 二核苷酸富集的区域。这些区域的甲基化通过阻止甲基转移酶的结合而阻碍基因的转录，这个机制使得 DNA 甲基化可以保持基因的沉默状态。

3. DNA 甲基化募集其他蛋白引起染色质沉默 除了直接抑制转录因子的结合外，DNA 甲基化还可以募集甲基化 CpG 结合蛋白(methyl-CpG binding proteins，MBPs)特异性地结合到甲基化的 CpG 位置上，它们在甲基化发生后的转录沉默过程中都扮演着重要的角色。MBPs 这类蛋白质家族包含五个成员，均包含一种同源的甲基 -CpG- 结合结构域(methyl-CpG-binding domian，MBD)。MBPs 可以结合沉默子以及组蛋白去乙酰酶，这是其导致染色质结构沉默的主要原因。

4. DNA 甲基化影响核小体定位 启动子区域内的 CpG 甲基化会通过影响基因转录起始位点附近的核小体定位(nucleosome occupancy)，进而影响这些基因的转录。核小体定位可阻碍转录因子和 RNA 聚合酶Ⅱ的结合，实验研究发现在转录起始位点附近，无核小体缠绕的 DNA 区域容易吸引转录激活因子和 RNA 聚合酶Ⅱ与 DNA 的结合。*MGMT* 和 *MLH1* 基因的启动子的研究表明 DNA 甲基化缺失影响体内无核小体区域的核小体定位。此外，一种 DNA 甲基化转移酶 DNMT3a 被发现和染色质重构物结合，表明染色质重构物可能直接和 DNA 甲基转移酶结合。

(三) DNA 甲基化的意义

1. DNA 甲基化与重复元件沉默 哺乳动物细胞必须保有使遗传元件沉默的机制才能使基因组达到长期稳定的目的，DNA 甲基化行使的即是此功能，而且在细胞分裂前后可以保持不变。哺乳动物基因组较低等生物基因组更复杂，主要是因为其不仅包含编码蛋白质的元素，还包含转座子和其他寄生元件，其中包括多种重复元件。许多重复元件包含长末端重复启动子，它可以使这些序列发生转录。由于这些序列的表达会导致基因组的寄生元件在基因组中的游动，因此通过 DNA 甲基化使其持久地沉默以保持基因组的完整性。

2. DNA 甲基化与染色体的选择性沉默 DNA 甲基化除了具有沉默重复元件的作用外，还在 X 染色体失活及基因印记的维持中发挥作用。X 染色体失活及基因印记均是非孟德尔遗传方式的一部分，从父本或母本得到的等位基因的一个等位发生甲基化而导致单等位表达。在胚胎形成过程中，两条 X 染色体中的一条发生失活也表现出单等位的表达，而在失活的 X 染色体上发现的 CpG 富集的启动子的甲基化使相应基因的抑制状态得到稳定。

3. DNA 甲基化与基因的组织特异表达 DNA 差异甲基化在发育过程中扮演着重要的角色。DNA 甲基化可以沉默生殖细胞特异的基因，而且大量的差异甲基化基因只是生殖细胞特异的基因，因此 DNA 甲基化通过抑制生殖细胞的关键基因而迫使细胞进入分化过程。在人类基因组中，CpG 岛的甲基化被认为对基因的沉默有直接的影响，然而大约 60% 的基因包含非 CpG 岛启动子。在发育和分化过程中，这些启动子的甲基化状态可能发生转变，从而介导基因的组织特异表达。

二、CpG 岛识别方法

对 CpG 岛的识别，大致有两种策略。一种是以生物信息学算法为基础开发的预测方法；一种是以限制性酶切法为代表的实验方法。CpG 岛最初是在对小鼠基因 DNA 使用甲基化 CpG 敏感的限制性酶 HpaⅡ进行酶切时发现的。

(一) CpG 岛识别的准则

1. 最初的 CpG 岛定义 CpG 岛的原始定义是 Gardiner-Garden 和 Frommer 于 1987 年提出的长度≥200bp，GC 含量≥50%，CpG O/E≥0.6 的一段序列。CpG 岛的这种定义方式看起来有些武断，许多启动子缺乏严格定义的 CpG 岛，但是却有组织特异的甲基化模式，与基因的转录活性有密切联系。例如 *Oct-4* 和 *Nanog* 启动子的甲基化状态和基因表达的相关性很高，尽管它们的启动子都没有

CpG 岛。

2. 改进的 CpG 岛定义　一直对 CpG 岛的定义主要是基于序列特征，目前有许多 CpG 岛定义的改变，包括在长度，GC 含量和 CpG O/E 比值的一个或全部阈值的变化。为了降低非 CpG 岛序列的错误引入，Takai 和 Jones 研究了增加最短长度、GC 含量和 CpG O/E 值分别到 500bp，55% 和 0.65% 对预测精度的影响。通过确定更加严格的阈值，最大程度地排除 Alu 重复元件，却排除了占原来数量 10% 的 CpG 岛，这表明一些真正的 CpG 岛可能也被排除。重复元件（例如“年轻”的 Alu 元件）的碱基组成和 CpG 岛的特点十分类似，显著地增加了鉴别 CpG 岛的假阳性率。大多数的多拷贝序列可以通过 Repbase 数据库中已知的重复类型得以剔除，在 Takai 和 Jones 的基础上应用重复元件筛选后剔除 1890 个非 CpG 岛，从而得到更加保守的 CpG 岛的数目估计是 27 000 个。

NCBI Mapview 有两套不同的参数组合方式用来分别提供宽松的和严格的识别 CpG 岛的标准，如表 13-1 所示。严格的标准预测了 24 163 个无重复的 CpG 岛，而宽松的标准识别了 307 193 个 CpG 岛。这种巨大的差异取决于以下因素：①长度、GC 含量和 CpG O/E 值的任意阈值的应用；②没有考虑到 CpG 岛的异质性；③基于 DNA 序列的预测方法忽略了 DNA 甲基化状态。

表 13-1　常见的 CpG 岛预测算法

预测方法	长度(bp)	GC 含量(%)	CpG O/E	重复元件屏蔽	备注
ENSEMBL	≥400	≥50%	≥0.6	否	严格的参数限制
NCBI 宽松	≥200	≥50%	≥0.6	否	总 CpG 岛数目 307 193
NCBI 严格	≥500	≥50%	≥0.6	否	总 CpG 岛数目 24 163
UCSC	>200	≥50%	>0.6	是	总 CpG 岛数目 28 226
EMBOSS	指定	指定	指定	否	参数可调
CpGProD	>500	>50%	>0.6	是	总 CpG 岛数目 76 793
CpGcluster	无限制	无限制	无限制	否	总 CpG 岛数目 197 727
CpG_MI	≥50	无限制	无限制	否	总 CpG 岛数目 40 926

3. 基于窗口滑动法的 CpG 岛预测算法　窗口滑动法是与最初 CpG 岛准则有很大不同的算法，它的一般步骤如图 13-3 所示。首先准备通过实验方法得到的候选 CpG 岛集合或全基因组序列，然后设定窗口宽度的大小。接着考察窗口内的序列片段是否满足 CpG 岛定义中的长度、GC 含量和 CpG O/E 值中的一个或几个阈值。一旦发现窗中的序列片段满足了 CpG 岛的定义，该片段就被选为候选 CpG 岛，同时扫描窗右移 1bp。如果扫描窗中的序列片段不满足 CpG 岛的定义，扫描窗右移一个窗口的长度。如果扫描得到的 CpG 岛区域有重叠，则将重叠部分合并。通过这一过程，得到了各种长度的 CpG 岛集合。然而，这种依赖于长度、GC 含量和 CpG O/E 值的一个或全部阈值的 CpG 岛识别算法有显而易见的缺陷：①由于这三个阈值的使用使得参数空间变得很大；②预测的 CpG 岛的长度和数目取决于窗口的长度和步长的预设值，存在主观任意性；③ CpG 岛的起始点一般不是 CpG 二核苷酸；④预测和筛选过程依赖于相同的参数；⑤方法经常需要针对特定物种

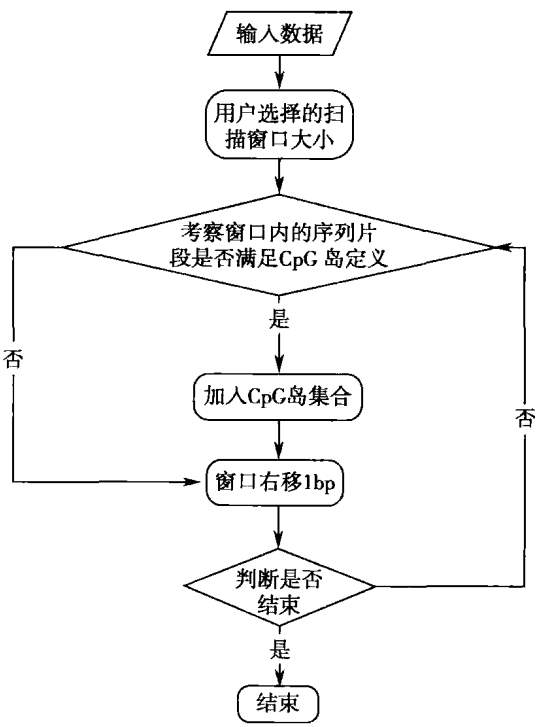


图 13-3　基于窗口滑动法的 CpG 岛预测算法框图

进行调整；⑥算法运行时间长，预测效率低。

4. 基于相邻 CpG 二核苷酸距离的 CpG 岛预测算法 预测的 CpG 岛总数是随着使用的序列参数而高度可变的。CpGcluster 是一种独特的方法，它并不依赖于任何 CpG 岛定义的阈值，并且由于只涉及算术运算，计算速度较快。它的工作原理是计算基因组范围的相邻 CpG 二核苷酸之间的距离。该算法利用几何分布估计出该距离的理论分布，从而计算出 CpG 二核苷酸进行汇聚的统计学阈值(40bp)。最终，该算法得到 197 727 个 CpG 岛。这些 CpG 岛的特点是短而多，但其中包含大量的重复元件。该算法具体的工作原理如下：①假设有如下一条序列：TTGCGGGTCCTAGAAGTCGCCTCCCCGCCTTGCCGGCCGCCCTTGCAGCCCCGAGCCGAGCAGC；②CpGcluster 首先找到所有的 CpG 二核苷酸的位置(斜体)TTG*CG*GGTCCTAGAAGT*CG*CCTCCCC*CG*CCTTGC*CG*GC*CG*CCCTTGCAGCCCC*CG*AGC*CG*AGCAGC；③然后得到 CpG 二核苷酸的位置 4，18，26，34，38，52，57；④通过公式 $d_i=x_{i+1}-x_i-1$ 计算相邻二核苷酸之间的算术距离：13，7，7，3，13，4；⑤考虑到假设：CpG 是伯努利实验的结果，这里设成功为 CpG，失败为 non-CpG。伯努利实验的概率 p 可以通过大量的序列算出。令序列的长度为 L，N 为 CpG 的数目，则 $p=N/(L-N)$。所以临近的 CpG 二核苷酸的距离服从几何分布，距离 d 等于失败的次数；⑥绘制长度(d)分布和几何分布的直方分布图(图 13-4)。从中可以发现观测值分布和理论分布差别很大，短距离出现的概率较大，中位数值恰好可以作为 CpG 二核苷酸富集的阈值。⑦为了计算之前步骤找到的 CpG 簇是 CpG 岛的概率，需要给出统计学 p 值，该 p 值可由负二项分布给出。基于 CpGcluster 的算法的原理，存在比随机出现 CpG 二核苷酸之间距离更短的 CpG 簇，通过合并重合的簇，最终得到的簇就被认为是 CpG 岛。

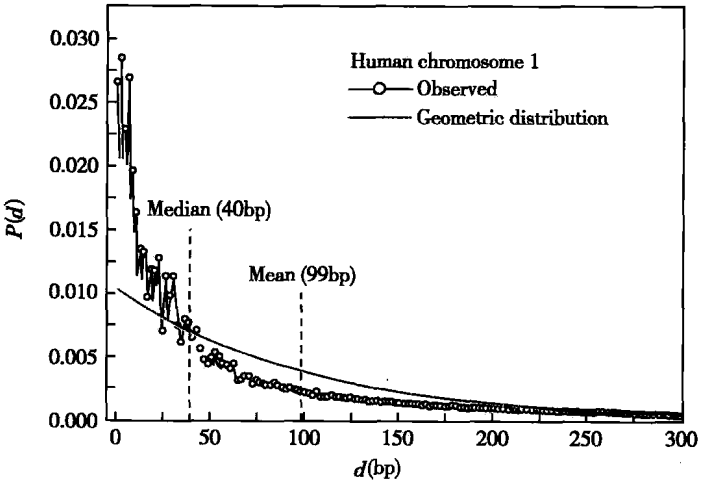


图 13-4 人类基因组 1 号染色体的邻接 CpG 二核苷酸之间距离的概率密度函数

观察值的分布以空心圆圈表示，而理论分布即几何分布则用实线表示 中位数值恰好和理论值吻合。距离小于中位数值的两个 CpG 二核苷酸则被纳入 CpG 岛的一部分。X 轴为距离 d，Y 轴为概率 p。Median 为中位数，Mean 为均值，带圈实线代表观测值的连线，实线代表几何分布的概率密度曲线

(来自于 CpGcluster: a distance-based algorithm for CpG-island detection)

另外，还有一种基于互信息的距离用于度量 CpG 二核苷酸距离预测 CpG 岛的方法。该方法通过计算累计互信息 $CMI(i)=\sum_{k=0}^{M} p_{(cg,cg)}^{(i)}(k)\log\frac{p_{(cg,cg)}^{(i)}(k)}{p_{cg}\times p_{cg}}$，$(i=1,2,\cdots n)$，刻画 DNA 序列上不同距离的相邻二核苷酸之间的互信息的累加。利用该度量可对哺乳动物基因组进行 CpG 岛预测，该方法可得到比之前的方法更理想的预测效果。主要的 CpG 岛预测方法的比较在表 13-1 中列出。

5. 结合功能基因组数据的 CpG 岛定位方法 大多数的预测算法和序列选择技术鉴别的 CpG 岛数目在 24 000 到 27 000 之间。尽管这些方法之间的差别不大，但是许多鉴别出来的 CpG 岛在不同

的预测结果中并不一致，可以通过结合包括 DNA 甲基化状态和染色质修饰在内的不同类型的信息添加到预测方法中减少预测结果的差别。在 CpG 岛预测算法中融合表观遗传信息和基因组属性可能有利于探测方法去除一些算法中定义的阈值。例如，Bock 等使用了 DNA 结构，组蛋白修饰，DNA 甲基化，转录因子结合谱，重复元件，进化保守，DNA 序列模式等信息定位人类基因组 CpG 岛，是目前较好的 CpG 岛定位方法。但该方法很难扩展到非人类的物种中，因为注释数据在其他物种并不全面。

（二）实验方法寻找 CpG 岛

CXXC 亲和纯化技术（CXXC affinity purification，CAP）通过提取非甲基化的 CpG 聚集的 DNA 片段识别 CpG 岛，克服了算法带来的假阳性等问题。该技术使用了半胱氨酸富集的对非甲基化的 CpG 位点有高亲和性的 CXXC 结构域。CXXC 结构域对只包含甲基化的 CpG 位点或缺乏 CpG 位点的 DNA 片段几乎没有亲和性。从小鼠 *Mbd1* 中得到的重组的 CXXC 结构域对非甲基化的 CpG 位点有高的结合特异性，并被用于从全基因组 DNA 中提取 CpG 岛。使用这种方法从人类血液中提取了超过 17 000 个 CpG 岛。

（三）CpG 岛的定位有助于发现新基因

CpG 岛是重要的转录调控元件，是基因起始的标志，可用于新基因的发现。同时，CpG 岛通常是不被甲基化的，可作为管家基因的重要标志之一。为了更好地认识和发现基因功能，需要开发定位 CpG 岛的新方法，以快速识别新测序物种的 CpG 岛，这有助于快速进行新基因组的注释。并不是所有 CpG 岛都在已知基因的转录起始位点附近，例如可以位于基因的 5' 区域、内含子以及基因间区（图 13-5）。然而，基因内的 CpG 岛可能表明这段区域存在未发现的新基因。

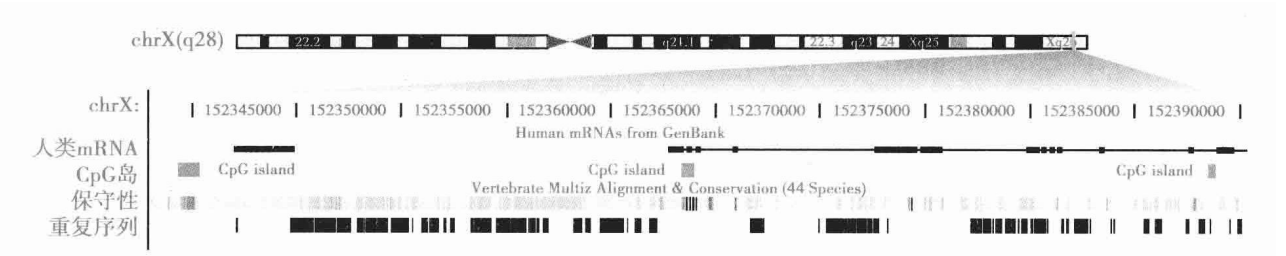


图 13-5 UCSC 数据库的截图（chrX：152，333，843-152，400，119）展示了三个 CpG 岛

三、DNA 甲基化状态的实验检测

（一）DNA 甲基化的检测方法

DNA 甲基化的检测比 DNA 碱基序列检测相对困难，这是因为甲基化的胞嘧啶并不影响 C：G 核苷酸的配对。目前常用的 DNA 甲基化检测方法是将待检序列中甲基化的胞嘧啶转化为其他碱基组成的变化。基本做法主要有两类：一种是依赖甲基化和非甲基化的胞嘧啶对特定的酶切处理的不同反应来实现；另一种则是用重亚硫酸钠（sodium bisulfite）转化 DNA 序列，利用甲基化和非甲基化的胞嘧啶的化学活性不同将其区分开来。此外，最新的检测方法通过利用基因微阵列（microarray）提高了 DNA 甲基化检测效率。下面对这些实验检测 DNA 甲基化的方法做简要介绍。

1. 限制性内切酶法　通过限制性内切酶特异性地将基因组 DNA 切割为甲基化和非甲基化的序列片段是发现功能 CpG 岛的重要手段。如图 13-6 所示，基于限制性酶的方法可以同时用于构建甲基化 DNA 文库和非甲基化 DNA 文库。最常用的限制性酶是 HpaⅡ和 MspⅠ，它们可以识别 CCGG 序列模式。HpaⅡ受甲基化的胞嘧啶的阻碍较大，任何胞嘧啶的甲基化均会阻碍它的切割，而 MspI 只会受到 CpG 环境外的胞嘧啶甲基化的阻碍。在基因组研究中，另外一种有用的酶是 McrBC，McrBC 能对 DNA 的甲基化位点进行识别而切割，可以将两个甲基化的胞嘧啶之间的片段剪切出来，从而得到非甲基化的序列片段。利用这些酶可以区分出甲基化和非甲基化的序列，得到感兴趣的 DNA 序列文库。

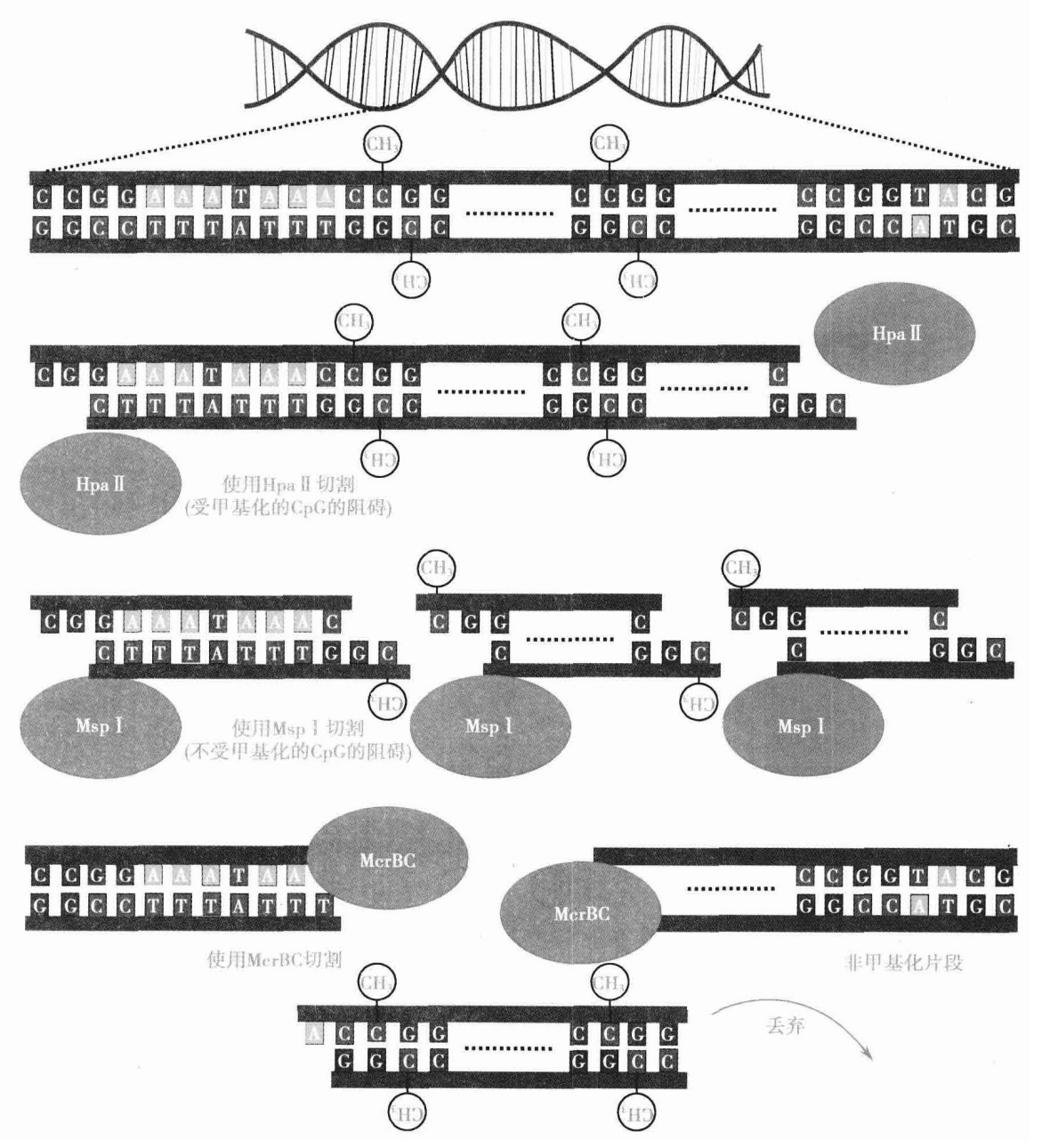


图 13-6 使用甲基化敏感的酶检测 DNA 甲基化

一段甲基化的基因组区域被 HpaⅡ、MspⅠ和 McrBC 切割。在最终的甲基化片段文库中，只包含 HpaⅡ处理的甲基化片段。基因组 DNA 被 McrBC 处理时，McrBC 切割甲基化的胞嘧啶位置(切点不确定)，小片段被丢弃，得到非甲基化片段文库

2. 重亚硫酸钠法 在基因组 DNA 序列上，甲基化的胞嘧啶和非甲基化的胞嘧啶在碱基配对特性上并无差异，即单纯利用测序技术无法确定甲基化水平。为了克服甲基化测定的难题，使用重亚硫酸盐将非甲基化的胞嘧啶转化为其他碱基，在合适的条件下，重亚硫酸盐会将非甲基化的胞嘧啶转化成尿嘧啶，而使甲基化的胞嘧啶保持不变(图 13-7)，转化的 DNA 在经过 PCR 扩增后得以转变为胸腺嘧啶。再对 PCR 产物进行桑格法测序、焦磷酸测序或质谱分析后，可以定量地考察每个胞嘧啶位点的甲基化程度。

重亚硫酸钠法检测 DNA 甲基化的基本原理是：重亚硫酸钠可以把非甲基化的胞嘧啶 C 转化为尿嘧啶 U，再经过 PCR 扩增克隆变成胸腺嘧啶 T 产生 T:A 配对，而对于甲基化的胞嘧啶重亚硫酸钠则没有作用。因此，通过重亚硫酸钠的处理，甲基化状态不同的 DNA 序列片段就转化为有碱基差

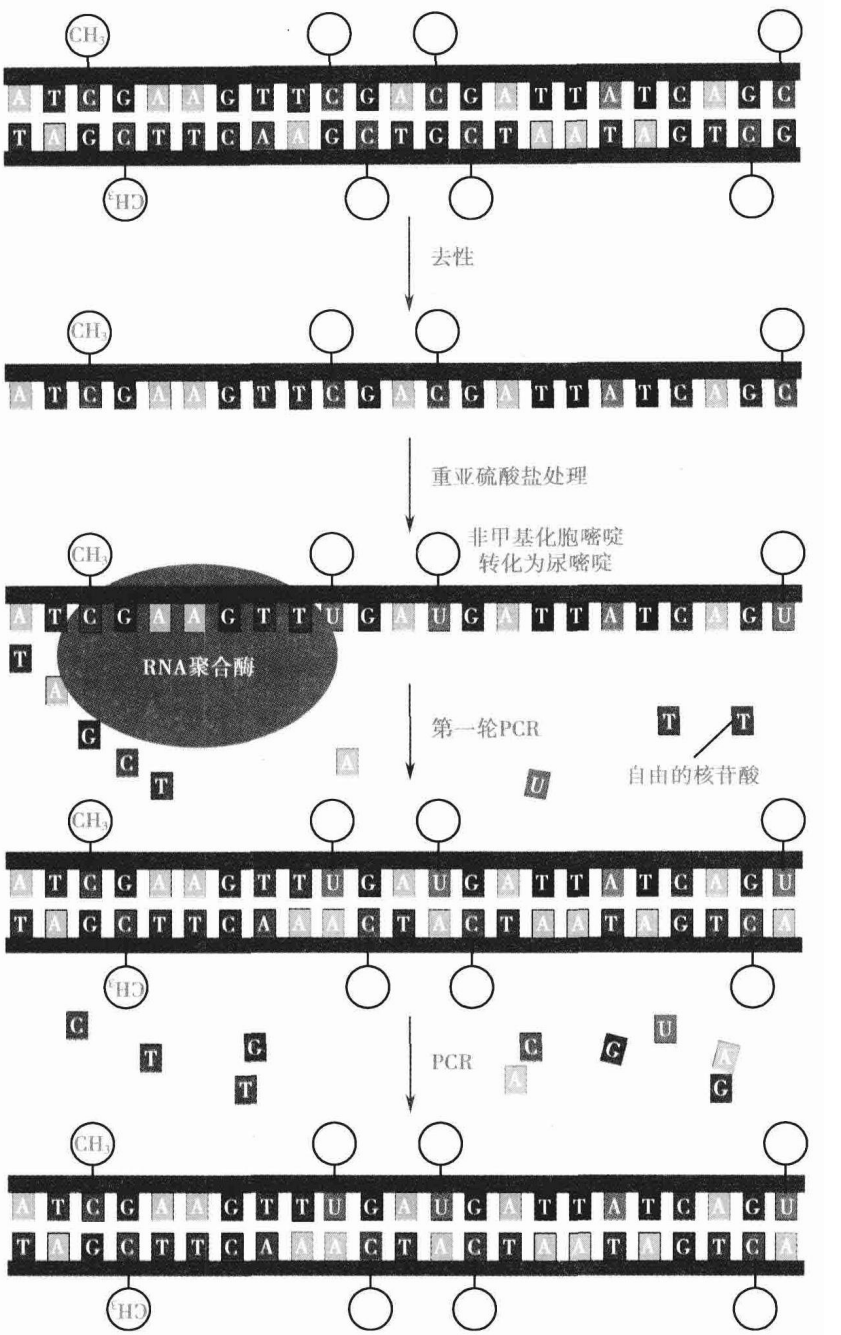


图 13-7 重亚硫酸盐转化实验原理

异的序列片段，这种差异可以进一步用 DNA 测序的方法进行测定(如图 13-7 所示)。

3. 亲和纯化 比较简单的建立甲基化 DNA 文库的方式是亲和纯化技术(图 13-8)。该方法发挥甲基化 CpG 结合结构域(methyl-binding domain，MBD)的优势，结合甲基化的 CpG 位点。首先需要从大肠杆菌中提取表达的 MBD 结构域，将之提纯，用于提取甲基化的 DNA 片段。相应地，也可以使用市售的特异识别甲基化胞嘧啶的单克隆抗体来提取甲基化的 DNA，这一过程被称作免疫共沉淀(co-immunoprecipitation)。对于哺乳动物研究而言，使用 MBD 方法的潜在优势是 MBD 只对 CpG 甲基化的 DNA 片段进行提纯，而对其他序列环境中的 DNA 甲基化不起作用。

(二) 基因组范围的 DNA 甲基化检测方法

人类表观基因组计划(Human Epigenome Project，HEP)通过重亚硫酸盐测序法对人类主要组织的部分 CpG 位点的甲基化状态进行测定。这是目前精度最高的方法，但是测定成本很高，耗费人力。

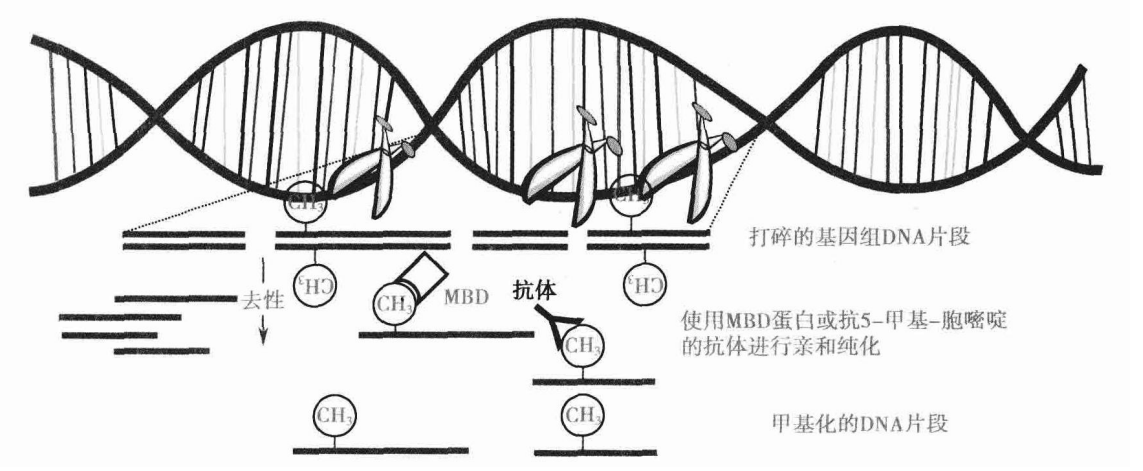


图 13-8 基因组 DNA 经过去性后使用抗体或 MBD 结构域而得以亲和纯化

目前，有许多方法支持大规模的 DNA 甲基化分析。早期的 DNA 甲基化微阵列的研究使用实验室或微阵列公司制作的斑点阵列。而现在，高质量的商业寡核苷酸阵列，包括 Illumina 推出的磁珠阵列，Affymetrix 和 NimbleGen 分别生产的平板阵列，以及 Agilent 推出的喷墨阵列都被广泛地使用。这些技术便利了近来的甲基化分析，例如 Illumina 阵列的设计便利了重亚硫酸盐转化的 DNA 的分析，而其他微阵列适合于限制性酶切和亲和纯化的实验(图 13-9)。各种高通量技术的优缺点和应用范围则在表 13-2 中列出。

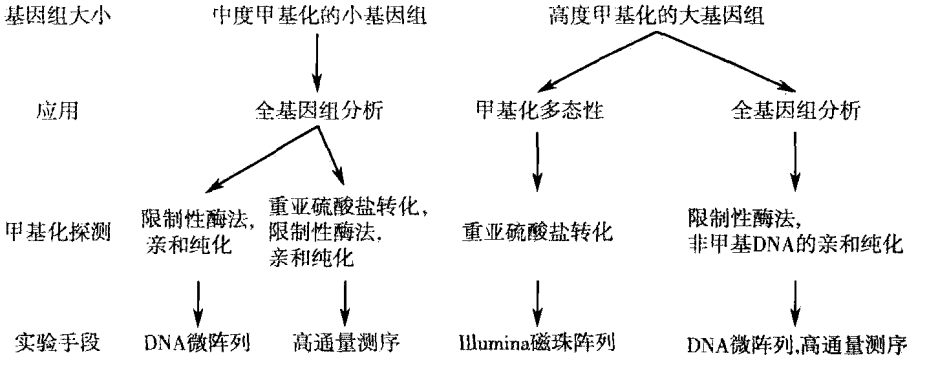


图 13-9 基因组尺度的 DNA 甲基化测定技术选择策略

表 13-2 DNA 甲基化大规模分析可用平台一览表

技术	应用	优势	局限
Illumina 磁珠阵列	甲基化多态性发现和分析	定量，多达 96 个样品的同时快速分析	需要设计引物文库，同时只能分析 1536 个位点
Affymetrix 芯片	全基因组甲基化测定	探针密度大，支持物种多，可定制，价格合理	短寡核苷酸噪声大，单通道杂交，定制芯片昂贵
NimbleGen 微阵列	全基因组甲基化测定	长寡核苷酸探针产生更纯净的数据，双通道杂交，定制芯片不昂贵，价格合理	较 Affymetrix 芯片的探针密度小
Agilent 微阵列	大规模甲基化测定	长寡核苷酸探针产生更纯净的数据，双通道杂交	较 Affymetrix 和 NimbleGen 芯片的探针密度小得多
Solexa 测序	全基因组甲基化测定，分析印记位点	定量化，无需杂交，并行的基因型信息	下一代技术，需要购买昂贵的仪器或服务

（三）基于高通量测序的 DNA 甲基化检测方法

高通量测序是最新发展起来的但却是最有前途的全基因组 DNA 甲基化分析方法。高通量测序技术的出现，使得产生大量序列信息的时间和成本均要低于桑格法。

高通量测序技术最初是作为寡核苷酸微阵列的替代品而引入 DNA 甲基化分析中。在微阵列实验中，无需标记和杂交，样品可以直接得以测序。当测得足够多的序列后，测序能达到和微阵列相媲美的信息量。相比其他方法，直接测序有许多优点。首先，短序列提供了甲基化丰度的定量表示，这要优于基于微阵列方法提供的相对度量。其次，除了测序中的一步需要扩增外，样品也无需额外进行扩增。单分子的测序方法目前还在不断开发中，将完全不需要扩增。再次，杂交所带来的偏差，也无需将待测片段设计成探针打印在玻璃板上。

目前，两种高通量的测序平台最为流行：一种是 454 生命科学公司开发的焦磷酸测序方法，另外一种是 Illumina 前身的 Solexa 开发的基于荧光核苷酸的系统。Solexa 系统可以产生较 454 系统数量多上百倍但长度在 25~35 碱基范围内的序列标签。但 30bp 对于大多数片段而言，可以基本保证匹配到参考基因组序列上。因此，Solexa 是目前此类分析首选的实验平台。

使用何种高通量检测技术取决于所要解决的问题及分析的基因组，而不是一味地推崇测序技术。对于一个重复序列较少的较小的基因组，例如拟南芥，对重亚硫酸盐转化的 DNA 实施直接测序是很好地选择。对于包含重复序列含量较多的较大基因组，例如人和小鼠的基因组，直接的重亚硫酸盐测序更加有挑战性，但是在小规模的前提下已经得以实施。对甲基化 DNA 的亲和纯化也是有挑战的，因为大多数基因组是甲基化的。最好的方法可能是先提取非甲基化的 DNA，然后辅以亲和提纯或使用限制性酶法（图 13-9）。

四、DNA 甲基化的预测算法

实验手段检测 DNA 甲基化状态的方法虽然比较可靠，但是其对于人力财力的需求以及检测技术方面的缺陷，使目前大规模的 DNA 甲基化的实验数据的产生还比较有限，开发计算方法预测 DNA 甲基化状态显得极为重要。另一方面，作为实验检测技术的补充，预测算法能够挖掘出数据中隐藏的重要特征，为进一步认识 DNA 甲基化机制提供重要依据。目前，DNA 甲基化的预测主要使用两种模型，一种是基于序列的判别模型，另一类是借助其他表观遗传修饰谱的整合模型。

（一）从 DNA 序列预测胞嘧啶甲基化

在生殖细胞中（除了印记区域外），CpG 岛通常被认为是非甲基化的。然而，许多研究试图给予 DNA 序列预测分化后的细胞中的胞嘧啶甲基化模式，发现人脑组织中甲基化和非甲基化的 DNA 序列片段的模体，尽管它们的功能意义尚不明确。

1. CpG 位点甲基化预测　Methylator（http://bio.dfci.harvard. edu/Methylator/）是最早的预测单个 CpG 二核苷酸甲基化状态的工具。该工具是基于 MethDB 数据库（http://www.methdb.de/）中人类的 2839 个 DNA 甲基化状态的数据。使用支持向量机（SVM）作为分类器，发现以 CpG 位点周围 39bp 窗口内的序列模式为特征时，该方法得到了 87% 的正确率，ROC 曲线下面积可以达到 0.82。

预测的原理如下：基于 n 个样本 $\{x_i, y_i\}$，$i=1, 2, \cdots, n$（其中 x_i 为 d 维特征构成的向量，y_i 取自 $\{-1, 1\}$ 而代表类别，-1 作为甲基化标记，1 作为非甲基化标记）作为训练数据，SVM 利用下面的判别函数进行训练和检验：$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i y_i K(x_i, x)+b\}$。其中，$\alpha_i$ 和 b 为待估参数，使得判别函数更好地拟合训练数据。当进一步增加窗宽时，SVM 的性能并没有显著地提高。该算法的流程在图 13-10 的左侧。由于单个 CpG 位点的甲基化状态一般不是一成不变的，因此 Methylator 不能满足组织特异分析的要求。实际上，目前尚缺乏有效的 CpG 位点的预测工具。

2. 基于序列特征的 CpG 岛甲基化预测　利用人脑组织全基因组范围的甲基化数据，得到 1948 个甲基化片段和 2386 个非甲基化片段，对甲基化和非甲基化的片段进行分类，并且在对这些片段进

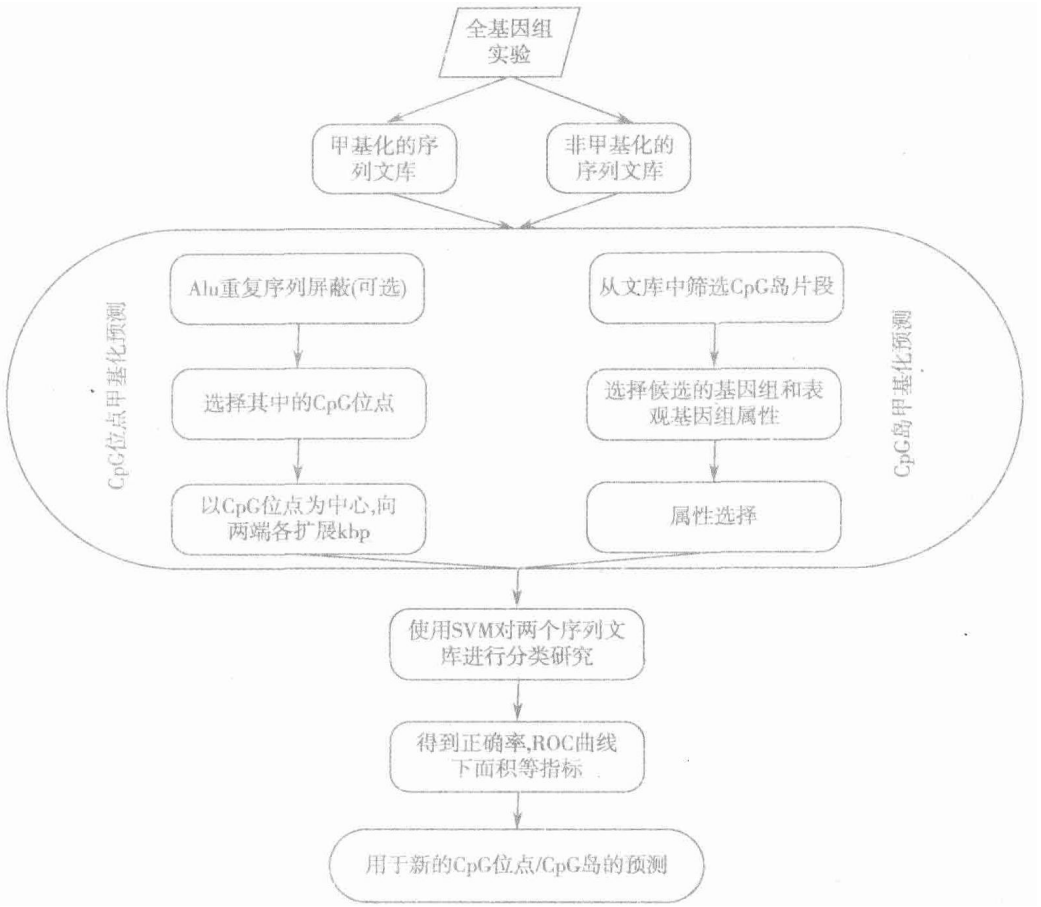


图 13-10 DNA 甲基化预测基本流程

行 Alu 重复序列屏蔽后，开始挖掘模体。具体而言，对每一个 CpG 二核苷酸两边各 250bp 范围内的序列进行 CpG 岛判别。如果它们之间有重叠的话，将找到的 CpG 岛进行合并。使用开发的位置权重矩阵枚举程序(判别矩阵枚举)找到对甲基化片段和非甲基化片段最有判别效力的 10 个模体。进行分析后，发现它们与已知的转录因子结合位点有关。利用开发的预测 CpG 岛片段甲基化状态的工具——HDMFinder，展示了预测的人类全基因组甲基化全景图谱。HDMFinder(http://rulai.cshl.edu/HDMFinder/methylation.htm)使用包括 Alu 在内的 17 个序列特征能够对长度为 800bp 的 CpG 岛片段以及 800bp 的普通 DNA 序列的甲基化状况进行准确预测，并且还对 22 个常染色体中的甲基化模式进行了预测。

MethCGI 则是另外一个专门针对 CpG 岛(包括 200bp、300bp 和 400bp)进行预测的在线工具，正确率达到 84%，并提供在线预测服务，http://bioinfo.au.tsinghua. edu.cn/MethCGI/MethCGI.html。该算法的流程在图 13-10 的右侧。该网络服务支持通过提交序列而预测的方式，结果输出提交的 CpG 岛序列和指定该 CpG 岛的甲基化状态(甲基化或非甲基化)。例如，用户输入一条长度在 200bp 以上的序列：gctggaggccactatcctaag...gtggtgtcagcg tccggggccgggggaggggtgtc(该序列可在网站找到)，输出的结果则通知用户该序列为非甲基化的 CpG 岛。

通过分析特定转录因子结合位点在甲基化和非甲基化的 CpG 岛片段中的分布情况，最终得到了 4 个在两类中差异最大的转录因子结合位点，这些结合位点可能为下一步研究 DNA 甲基化的内在机制提供重要依据。以非甲基化的 CpG 岛为中心，分别向两侧以 200bp 为单位进行划分，分析 CpG 岛内部和外侧的 TRANSFAC 数据库中注释的转录因子结合位点的个数，结果发现非甲基化的 CpG 岛外侧分布的转录因子结合位点的个数较 CpG 内部要显著。而 CpG 岛外侧同样分布着较多的隔离子

Sp1 和 CTCF 偏好的区域，它们都是公认的甲基化隔离子，能够保护 CpG 岛不受甲基化。同时，在非 CpG 岛边界分布富集的 70% 转录因子结合位点包含 C2H2 类型的锌指结构域。由此，推断锌指结构域可能也具有阻止 CpG 岛甲基化蔓延的机制。

3. 使用基因组特征有助于识别 CpG 甲基化　该方法利用人类淋巴细胞的第 21 号染色体已知甲基化状态的 132 个 CpG 岛构建判别模型。模型构建大致分为以下步骤：

步骤一：获取 CpG 岛数据集。首先去除人类基因组的 21 号染色体重复元件，使用传统的 CpG 岛识别算法预测潜在的 CpG 岛。然后对每个可能的 CpG 岛设计引物，并从外周血细胞中抽取相应的 DNA 片段。最后，通过甲基化特异的限制性内切酶(HpaII-McrBC)确定各个 CpG 岛的甲基化状态。对一些 CpG 岛进行重亚硫酸盐测序法进行确认。最终，取得了 149 个 CpG 岛的甲基化状态，并将它们分到如下各类之一：完全甲基化、完全非甲基化、不完全甲基化和差异甲基化。由于重亚硫酸盐测序法表明后两类可能受到实验误差的影响而产生，所以只选择前两类作为研究的对象，得到 21 号染色体的 29 个甲基化的 CpG 岛和 103 个非甲基化的 CpG 岛。

步骤二：属性计算。汇集 1184 个和序列直接或间接相关的 DNA 属性，包括 DNA 序列性质和模式、结构、保守元件、基因、SNP、转录因子结合位点、预测的 DNA 结构和重复元件等。大部分的属性以频率或数值分数的形式体现，并标准化。其中，这些属性中的一大类是序列模式，即对给定序列统计 4bp 序列模式出现的频率。除此外，大部分属性可以从 UCSC 数据库中下载到。对 132 个 CpG 岛计算这些属性的特征谱。此外，为了研究 CpG 岛周边的情形，并与 CpG 岛本身进行比较，将 CpG 岛本身进行扩展。使用 10 种序列扩展方式：−20kb~−10kb，−10kb~−5kb，−5kb~−2kb，−2kb~−1kb，−1kb~CpG 岛的左边界，CpG 岛的右边界~+1kb，+1kb~+2kb，+2kb~+5kb，+5kb~+10kb，+10kb~+20kb。经过属性计算，在 132 个 CpG 岛中的数值至少有 5 个甲基化的 CpG 岛和 5 个非甲基化的 CpG 岛不为 0 的特定属性被保留，共剩余 833 个属性。

步骤三：特征选择。统计学检验可以确定在两类 CpG 岛中显著差异的属性。首先使用非参数的 Wilcoxon 秩和检验比较所有属性在 CpG 岛的差异程度。然后，使用方差分析计算所有 CpG 岛的扩展区域的属性差异。显著性阈值使用 Bonferroni 多重检验方法进行调整，即整体的错误率在 1% 以下。

步骤四：使用机器学习构建预测模型。使用机器学习算法有两个目的，一是定量化 CpG 岛甲基化和 DNA 相关的属性之间的关联，另一个是预测新的 CpG 岛的甲基化状态。在第一个目的中，只是需要确认数据的关联性和预测的可能性。这里，使用一种分类器作为工具来定量化属性和 CpG 岛甲基化的关系。因为如果一个分类器使用一个特定的属性类别就能够成功而可靠地预测 CpG 岛甲基化，那么从功能上讲，这个属性类是和 CpG 岛甲基化相关，而且预测性能是衡量功能相关的一种尺度。在确认 CpG 岛甲基化和可能的属性的关联后，需要利用这些属性对未知的属性进行预测。在给定一些 CpG 岛的情况下，线性 SVM 被 90% 的数据重复地训练以评估对剩余的 10% 的预测效果。对这个过程重复 20 次。预测的结果使用检验集的真实值和预测值的相关系数进行评估。为了考察不同的分类器对评估的影响，选择多种分类器，如 AdaBoost M1、C4.5、线性 SVM 等。这些分类器已经被 Weka 和 R 等软件平台所实现，可以方便地调用。如果没有特定的要求，所有软件使用默认的参数即可。

步骤五：实验验证。通过重亚硫酸盐测序法对 12 个新预测的 CpG 岛进行确认。结果，同样在外周血细胞中，只有一个 CpG 岛的预测结果与实验结果不符，这充分说明了计算方法对实验的很好的指导效果。

通过使用 SVM 对 CpG 岛和 DNA 属性之间进行相关性研究，发现特定的 DNA 序列、结构和重复元件可用于区分高甲基化和低甲基化的 CpG 岛。对 12 个甲基化状态未知的 CpG 岛进行实验测定，正确率超过 90%。此外，在使用人类表观基因组计划 HEP 的甲基化数据对 6 号染色体的 CpG 岛进一步地确认，取得很高的预测精度。

（二）借助其他表观遗传修饰谱预测 CpG 岛甲基化

结合其他表观遗传学信息可以提高 CpG 岛预测精度。有使用 SVM 并整合全基因组范围的 ChIP-Seq 数据进行 CpG 岛甲基化预测，此外，结合序列特征，重复元件及转录因子结合位点等信息进一步增强了 SVM 的分类性能。另外发现添加了组蛋白甲基化修饰之后，该预测模型得到了更高分类正确率，同时发现了四种显著影响 CpG 岛甲基化的组蛋白修饰（H3K4me1、H3K4me2、H3K4me3 和 H3K9me1）。

五、异常 DNA 甲基化与疾病的发生

DNA 甲基化在控制基因活性方面起到重要作用。在人类正常基因组中，3%~6% 的胞嘧啶是被甲基化的。CpG 二核苷酸在基因组中的分布不是随机的，CpG 岛一般和已知基因的 5' 非编码区，启动子和第一外显子重叠。CpG 岛在正常细胞中通常是非甲基化的，使包含它的基因在转录激活元件的辅助下得以转录。在癌症细胞中，通过 CpG 岛的异常高甲基化即超甲基化（hypermethylation），肿瘤抑制基因的转录受到抑制，这通常是癌症发生的前奏（图 13-11）。

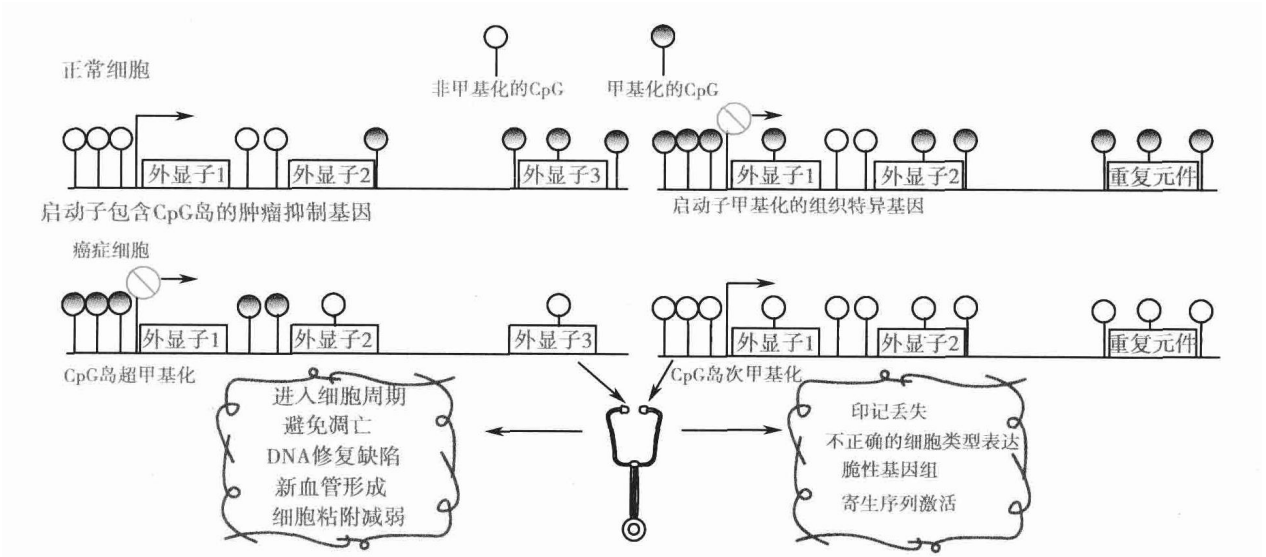


图 13-11 肿瘤中改变的 DNA 甲基化模式及分子生物学水平产生的效应

（一）基因组整体低甲基化

重复元件所在基因组区域在正常基因组中是甲基化的，这会保证基因组的稳定性，防止转座和基因断裂的发生。在癌症基因组中会发生全局性的基因组去甲基化，这一现象被称为次甲基化（hypomethylation）。次甲基化可以进一步导致癌症基因组的遗传性变异，这通常是肿瘤发生的特征。在 CpG 岛发生超甲基化的同时，癌症基因组经历了全局性的次甲基化。相比正常基因组，大约 20%~60% 的 5- 甲基 - 胞嘧啶的甲基基团脱落。癌症发展过程中经常伴随基因转录区域的整体性甲基化缺失以及重复元件 DNA 的去甲基化。

（二）印记丢失

DNA 甲基化还为生殖细胞特异基因和组织特异基因的表达提供表观遗传控制。基因组印记需要父本和母本等位中的一份发生超甲基化而建立单等位表达的模式。类似地，在正常情况下女性基因组中的一条 X 染色体发生异染色质化。在 DNA 甲基化的调控作用中，印记丢失（loss of imprinting）是许多癌症基因活化的一种机制。

（三）基因超甲基化是癌症的标志

表观基因组方法对促进癌症生物学理解的一个重要方面就是获得全面的基因异常甲基化信息。目前，从人类众多肿瘤中取得了众多的癌症超甲基化基因。超甲基化被认为是所有人类癌症的一般

标志，它几乎影响所有细胞通路。在过去几年内，癌症生物学中重要的一些基因，例如 DNA 修复基因 *MLH1* 以及 *BRCA1* 都被发现在癌症中经历了甲基化相关的沉默。许多癌症超甲基化基因本身就是肿瘤抑制基因，具有抗增殖作用的基因的转录被癌症细胞系的 CpG 岛超甲基化所抑制。使用表观基因组技术有助于鉴别出癌症异常甲基化基因，使用生物信息学技术可以进一步分析受累的通路。在不同的肿瘤类型中，CpG 岛超甲基化基因通常是不同的。每一种肿瘤亚型可能被一些超甲基化基因或表观遗传学标记所区分，这通常是癌症诊断十分重要的标志。因此，一个特定癌症的平均超甲基化基因数目变得十分重要，这会对理解遗传学和表观遗传学异常对癌症的协同作用有更深的理解。MeInfoText 和 PubMed 数据库汇总了癌症特异的异常甲基化信息。目前尚不清楚为什么一些基因在特定癌症中发生超甲基化，而另一些不表达的基因却保持非甲基化状态。

第三节 组蛋白修饰的表观基因组

Section 3 Epigenome of Histone Modifications

一、组蛋白密码是重要表观遗传标记之一

（一）核小体与组蛋白修饰

1. 核小体与组蛋白 组成染色质的基本单位是核小体(nucleosome)。每个核小体均由 5 种组蛋白共同构成。组蛋白是指所有真核生物的细胞核中，与 DNA 结合的碱性蛋白质的总称。在这些碱性蛋白质中，含精氨酸和赖氨酸等碱性氨基酸特别多，加起来为氨基酸残基总数的 1/4 左右。组蛋白与带负电荷的双螺旋 DNA 结合成复合物。组蛋白通常包括 H1，H2A，H2B，H3，H4 5 种。除 H1 外，其他 4 种组蛋白均分别以二聚体(共八聚体)的形式相结合，共同组成核小体的核心。DNA 完全缠绕在核小体的核心上。而 H1 则与核小体间的 DNA 结合。核小体间的 DNA 也叫连接区 DNA(linker DNA)。DNA 缠绕在组蛋白核心上。组蛋白缠绕 DNA 的松紧程度对基因表达乃至 DNA 损伤修复和 DNA 复制重组都有精确而动态的调节作用。鸟类、两栖类等含有细胞核的红细胞中，含有一种叫 H5 的特殊组蛋白。组蛋白可受到甲基化、乙酰化、磷酸化、聚 ADP 核糖酰化，以及泛素化等几种类型的修饰。组蛋白修饰扮演着十分重要的表观遗传调控作用。

2. 组蛋白修饰与转录 关于组蛋白修饰在转录中的作用，已经有许多模型如电中性模型、组蛋白密码以及信号通路模型被提出来。在电中性模型中，组蛋白乙酰化和磷酸化带的负电荷可以中和 DNA 的正电荷。根据电中性模型，组蛋白修饰可以使染色质纤维松弛。组蛋白密码假设指出多种组蛋白修饰可以协同地调节下游功能。信号通路模型假设组蛋白修饰便利酶和染色质结合并发挥功能，而不同的组蛋白修饰可以使生物学信号的传导更加鲁棒和特异。不同的组蛋白修饰类型的作用不尽相同。组蛋白乙酰化主要促使基因表达和 DNA 复制，使组蛋白乙酰化定位的基因得到动态的调控。组蛋白去乙酰化则使基因沉默。组蛋白的磷酸化可以改变组蛋白的电荷，对基因转录、DNA 修复和染色质凝聚等过程起调控作用。组蛋白的泛素化在细胞有丝分裂前后发生显著变化，被认为是信号传导的关键。

3. 组蛋白修饰命名法 一个组蛋白修饰的精确表示由三部分组成：组蛋白名称 + 组蛋白尾巴上的位点 + 修饰类型和个数。例如基因转录起始位点富集普遍存在 H3K4me3 修饰，它是组蛋白 H3 上，具体的位置为第四个位置即赖氨酸(Lysine，K)，该位置存在三个甲基基团。又如 H3K9ac，代表组蛋白 H3 上第九个位置即赖氨酸上发生的乙酰化修饰。当忽略组蛋白修饰的一部分时，例如 H3ac，则表示组蛋白 H3 上的乙酰化修饰，并没有指定位点信息；再如 H3K9me，则表示组蛋白 H3 上的第九位置上的甲基化修饰，但并没有指定甲基集团的数目，则泛指组蛋白甲基化修饰，这些模糊记法已被广泛地使用。目前，利用高通量实验技术广泛测量的组蛋白修饰类型如表 13-3 所示。

表 13-3 高通量实验测定的组蛋白修饰类型

组蛋白类型	组蛋白修饰
H2A	H2AK5ac, H2AK9ac, H2AZ
H2B	H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H2BK5me1, UbH2B*
H3	H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K27me1, H3K27me2, H3K27me3, H3K36ac, H3K36me1, H3K36me3, H3K4ac, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9ac, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H3ac*
H4	H4K12ac, H4K16ac, H4K20me1, H4K20me3, H4K5ac, H4K8ac, H4K91ac, H4Kac, H4R3me2, H4ac*

注：* 没有使用特异的抗体。

（二）激活性和抑制性的组蛋白修饰

根据对基因起到激活还是抑制作用，组蛋白修饰可以大致分为两类：激活性的组蛋白修饰和抑制性的组蛋白修饰。激活性的组蛋白修饰中最常见的就是 H3K4me。H3K4me 包括三种甲基化状态，且都是激活性的修饰。H3K4me 的三种修饰在基因组的分布差别较大，H3K4me3 的修饰主要在基因 5' 端的转录起始位点上下游附近。H3K4me2 和 H3K4me1 分别分布在 H3K4me3 的上下游外沿，强度比 H3K4me3 稍弱，沿着转录起始位点成对称状分布，且下游的强度较上游更强。除了定位活性基因外，H3K4me1 还被发现定位在基因的增强子。抑制性的组蛋白修饰中最常见的是 H3K27me。H3K27me 包括三种甲基化状态，但三种状态的组蛋白修饰都是抑制性的修饰。H3K27me 的分布强度较 H3K4me 要平坦许多，在活性基因中分布较少。H3K9me 的三种修饰和 H3K27me 具有类似的分布模式，对基因的调控功能也较为类似。

（三）组蛋白密码

1. 动态而又稳定的组蛋白密码　组蛋白的氨基酸残基可以接受许多种化学修饰，包括甲基化和乙酰化等修饰。质谱分析检测到组蛋白 H2A 有 13 个可以接受修饰的位点，H2B、H3 和 H4 则分别有 12 个、21 个和 14 个可以接受修饰的位点。每个氨基酸残基位点可以发生至少一种化学修饰。例如，一些赖氨酸残基可以发生甲基化和乙酰化修饰，对于甲基化而言，最多可以同时接受三个甲基基团的修饰。组蛋白修饰可能受到细胞的生理状态的改变和外界信号的刺激而发生瞬时的变化。在细胞周期的循环中，组蛋白修饰能够稳定地进行遗传。一个对人类肝脏组织的细胞周期过程中的组蛋白修饰模式的研究发现，*HNF-1*、*HNF-4* 和 *albumin* 基因的启动子的 H3K4me2/me3、H3K79me2、H3ac 和 H4ac 保持稳定。此外，在有丝分裂过程中这些活性组蛋白修饰并没有促使转录的发生，但染色质状态在细胞间得以稳定地保持，这说明组蛋白修饰在细胞分裂过程中的细胞间传递以表观遗传的方式进行。

2. 细胞分化过程中的组蛋白密码　组蛋白修饰的调控在许多生理过程中起到重要作用，其中包括细胞分化。研究发现组蛋白乙酰化对维持细胞的未分化和多能状态十分重要。使用组蛋白去乙酰酶抑制剂有助于维持干细胞的多能性(pluripotency)。相反，用去乙酰酶抑制剂刺激人类成熟细胞或癌症细胞会诱导分化的进行。因此，表观遗传调控对于细胞成熟至关重要。到底是什么类型组蛋白修饰或组蛋白修饰组合控制分化呢？如前所述，组蛋白乙酰化有助于保持细胞的多能性。此外如图 13-12 所示，H3K9me3 和 H4K20me3 也有类似

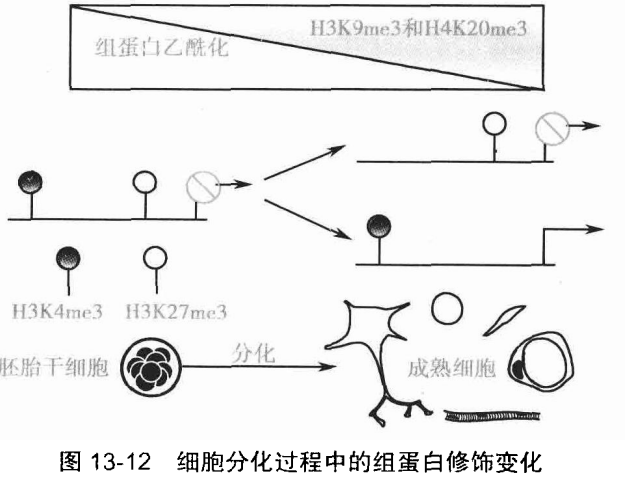


图 13-12　细胞分化过程中的组蛋白修饰变化

的作用。H4K20 甲基化和转录沉默有关，它可以控制 DNA 修复过程。然而，H4K20 甲基化被认为在细胞分化过程中高度变化。在从小鼠胚胎干细胞向多能神经前体细胞的分化过程中，H4K20me1 水平较高，而 H4K20me3 较低。随着分化的逐步进行，H4K20me3 的水平开始增加。在小鼠干细胞中，许多具有分化调控作用的基因都有二价结构域(bivalent domains)，包括具有转录抑制作用的 H3K27me3 和转录激活作用的 H3K4me3，拥有这样结构域的基因不会表达，看上去只是 H3K27me3 在发挥作用；在细胞分化过程中，二价结构域消失而只保留 H3K27me3 和 H3K4me3 中的一种修饰。在胚胎干细胞状态，基因都有 H3K4me3 标记，不管基因转录与否。因此，H3K4me3 是一个活性染色质修饰，但并不一定引起转录。拥有二价结构域的基因尽管受到 H3K27me3 的抑制不会转录，但 H3K4me3 和 H3K27me3 的平衡状态一旦被打破，基因就有可能倾向表达。

二、组蛋白修饰的分析方法

(一) 测定组蛋白修饰的高通量技术

从开始研究组蛋白修饰到现在，已经过去了几十年。但过去的几年却是发现组蛋白共价修饰的功能信息最多的几年，这得益于测定组蛋白修饰的高通量技术的不断成熟。这些高通量的实验技术提供了全面的表观遗传修饰图谱。基因组范围的数据的不断增多并结合计算表观遗传学的分析技术会增进对组蛋白修饰的理解。

1. ChIP-chip　在基因组范围上，检测组蛋白修饰的最流行技术就是染色质免疫共沉淀(Chromatin Immunoprecipitation，ChIP)与微阵列的结合即 ChIP-chip。简要地说来，染色质片段被特异性的抗体(例如针对 H3K4me3 的抗体)所沉淀，接着分离得到的片段，并进行扩增和荧光标记，最后使用 DNA 微阵列进行杂交检测。目前，该技术已被应用于啤酒酵母及哺乳动物等众多物种的组蛋白修饰测定中。

2. ChIP-SAGE　另外，针对 ChIP-chip 进行改进的技术正在日趋流行，其中之一是 ChIP 结合 SAGE(Serial Analysis of Gene Expression)的 ChIP-SAGE。也就是需要先进行 ChIP 实验，再进行 SAGE。从 SAGE 得到的测序文库中可以取得 21bp 的短序列标签，通过标签可以映射到基因组上。在某一基因组区域检测到的 tag 标签数据和该区域的修饰强度是成正相关关系。因为该改进技术没有探针杂交过程，该技术被认为比 ChIP-chip 更加定量化。

3. ChIP-Seq　近来快速兴起的一项新技术 ChIP-Seq 可以通过高通量并行的方式分析 ChIP DNA。简单地说，ChIP 得到的 DNA 片段的两头被加上 adaptor，并且进行有限次的扩增产生大量的 DNA。接下来，使加上 adapter 的序列在种有可与之共价互补结合的固相载体上杂交。通过桑格测序法确定结合到载体上的 DNA 片段的末端的 25~50bp 的碱基。因为 ChIP-Seq 不需要太多的 PCR 扩增循环，所以无需考虑探针杂交的效率问题，这使得 ChIP-Seq 标签是可以直接比较的，而 ChIP-chip 通常不能这么做。三种技术的比较见表 13-4。

表 13-4　ChIP-chip、ChIP-SAGE 和 ChIP-Seq 的比较

检测技术	ChIP-chip	ChIP-SAGE	ChIP-Seq
定量性	受杂交效率影响	定量	定量
分辨率的影响因素	染色质长度及探针密度	酶切效率	染色质长度，测序深度
全基因组范围实验花费	多	多	少
实验对于测定区域的局限性	局限于预设的基因组区域	受酶切位点的限制	可覆盖大部分基因组区域

(二) 分析基因组范围的组蛋白修饰数据

在进行全基因组范围的组蛋白修饰的 ChIP 实验后需要关注的一个问题就是如何从大规模的数据中抽取出有意义的生物学解释。通常，这些技术首要关注的是找出对应于特定基因组区域的信号

尖峰以及确定它的统计学水平。

1. 高通量组蛋白修饰分析工具 分析瓦式微阵列实验数据的分析工具中最有用的是 TileMap (http://biogibbs.stanford.edu/~jihk/TileMap/Index.htm)和基于模型的瓦式芯片分析算法(Model-based analysis of Tiling-array algorithm, MAT), http://chip.dfci.harvard.edu/~wli/MAT。这两个软件优于其他工具的地方在于它们支持多个样品的比较以及同一样品的重复测量的比较。序列标签分析和汇报工具(sequence tag analysis and reporting tool, START)是一个可以分析许多物种基因组的 ChIP-SAGE 产生的数据的工具。START 以 SAGE 文库的序列作为程序输入,运行后报告标签附近的基因,miRNA 和预测的转录因子结合位点的信息。

ChIP-Seq 数据的分析工具目前最实用的是 CisGenome(http://biogibbs.stanford.edu/~jihk/CisGenome/index.htm)。CisGenome 还支持 ChIP-chip 数据的分析。作为一个全面的整合分析平台,CisGenome 支持峰值探测以及下游的基因注释、从头模体发现、保守性分析以及基因组尺度的可视化。由于 Solexa 系统进行的 ChIP-Seq 实验测出的标签长度通常不超过 32bp,将这样的短片段对应到参考基因组上并且控制错配的碱基数小于 2 是比较困难的。ELAND(efficient large-scale alignment of nucleotide databases)程序可以对这样的数据进行处理,输出的标签对应到基因组上的精确位置。CisGenome 支持 ELAND 程序的输出文件的分析。除了 ELAND 格式,对于一种更为精炼的 BED 格式,CisGenome 同样支持。除了 CisGenome 外,MACS(model-based analysis of ChIP-Seq, http://liulab.dfci. Harvar d.edu/MACS/)也是不错的峰值探测工具。一旦得到一组精确的组蛋白修饰定位信息后,如何进行有效的显示也是较为困难的。目前,UCSC、CisGenome、整合基因组浏览器(integrated genome browser, IGB)均可以进行 ChIP-Seq 数据的可视化。这类可视化工具有助于从基因组角度解释表观遗传学修饰。

2. 组蛋白修饰峰值探测 与其他基于 ChIP 的高通量技术一致的是,从 ChIP-Seq 标签数据鉴别出可靠的组蛋白修饰谱,等价于寻找一段基因组区域内的统计学显著的组蛋白修饰标签的峰。一个最直接的想法是,对于一段长度一定的基因组区域来说,需要包含 R 个序列标签才可以从统计学水平支持这段区域被组蛋白修饰所定位。要固定这个数值,需要同时考虑几个参数的影响,即有效的基因组长度 gsize、期望的标签数 λ(为总标签数和与 gsize 的比值)、超声波降解得到的片段长度(bandwidth)、倍数富集(mfold)和标签偏移 d。通过构造泊松模型($1-\sum_{n=0}^{R-1} e^{-\lambda}\lambda^n/n!$),可以在一定统计学水平下(如 $p=0.01$)进行标签数估计,使得这个数值保证错误呼报(即窗口本不包含组蛋白修饰却被认为包含组蛋白修饰)的概率低于这个统计学水平。如果同时伴随 ChIP-Seq 实验数据还有实验控制数据的话,可以考虑使用实验控制数据对显著的标签数进行估计。即使没有实验控制数据,前述的两种分析程序也可以通过前面描述的统计模型构造背景模型。下面以 MACS 软件为例,解释影响 ChIP-Seq 峰值探测的一些因素。

①有效的基因组长度 gsize。由于人类基因组有些区域无法使用较短的序列进行唯一匹配,因此有效的基因组长度要小于期望值 3.2Gb。MACS 默认 gsize 为 2.7Gb。该值影响到基因组水平的 λ 计算,即 λ_{BG}。②标签偏移 d。由于 ChIP-Seq 的标签是对应于染色质片段的末端,标签对应到参考基因组的位置距离真正的组蛋白修饰的中心还有一定的偏差,所以需要对染色质标签进行位置调整以精确地反映组蛋白修饰的中心,通常的做法是将标签按给定测序方向的反方向移动 75bp,即差不多半个核小体 DNA 的长度。MACS 会随机挑选 1000 个高质量的峰,将沃森链和克里克链分开,分别计算两类链的标签的中点位置。如果沃森链在克里克链的左边,那么就将两类链向中心移动,形成混合的标签分布。如果沃森链和克里克链的标签分布的中心距离为 d 的话,那么两类链各自移动的距离为 $d/2$。③倍数富集(mfold)。给定一个 bandwidth 长度和 mfold 值,MACS 在基因组范围内生成 2*bandwidth 大小的众多窗口,以发现相对于背景分布的多于 mfold 倍数的富集标签的区域。MACS 默认设置 mfold 为 32。

MACS 与其他算法相比进行峰值探测有一些不同之处。首先去除冗余标签，有时相同的标签可能被重复地测序，这样的标签可能来自于 ChIP DNA 扩增和测序文库准备中所带来的偏差，这可能对最终的探测结果产生噪声。因此，MACS 去除了重复的标签。

目前，大多数的 ChIP-Seq 峰值探测算法均使用泊松分布进行建模。这个模型的优势在于只有一个参数：λ_{BG}，它既代表均值也代表方差。在 MACS 移动每个标签 $d/2$ 的距离后，就滑动 2*bandwidth 的长度寻找显著标签富集的可能的峰(泊松分布默认 p 值为 10^{-5})。交叠的峰值应当合并。最终的 tag 标签叠加后最高处即最可能为抗体的结合区域。

在控制样本中，经常观测到标签分布具有摆动和偏差的特点，这可能是局部染色质结构、DNA 扩增和序列偏好合力造成的结果。因此，使用 λ_{BG} 作为唯一的 λ 值是不合理的。MACS 因而使用一种动态的参数 λ_{local} 来估计局部的 λ。λ_{local} 被定义为 $\max(\lambda_{BG},[\lambda_{1k}],\lambda_{5k},\lambda_{10k})$。其中，$\lambda_{1k}$、$\lambda_{5k}$、$\lambda_{10k}$ 分别是从控制样本中的峰值位置为中心的 1kb、5kb 和 10kb 窗口范围估计的 λ。当没有控制样本时，λ_{1k} 不需要计算。λ_{local} 侧重于局部偏差的刻画，对于小区域的标签数目较少的情况亦有效。MACS 使用 λ_{local} 计算每个候选峰的 p 值，并去除由于局部偏差导致的潜在的错误的峰(即在 λ_{BG} 情况下满足，而使用 λ_{local} 则不满足)。p 值小于预设值(默认是 10^{-5})的候选峰则被认为是真实的峰，真实数据的标签数与 λ_{local} 的比值则为倍数富集值，在结果文件中随峰一起展示。

对于一个有对照的 ChIP-Seq 实验来说，MACS 可以为每个探测的峰估计错误发现率——FDR 的经验值。在每个经验 p 值下，MACS 使用相同的参数来发现相对于控制样本的 ChIP 峰和相对于 ChIP 峰的控制峰(交换)。经验 FDR 被定义为控制峰的数目与 ChIP 峰的数目的比值。MACS 也可以被应用于两个条件下的差异结合位点的情形，即将一个样本当作对照。因为任何一个样本均有生物学意义，所以不能通过使用交换计算 FDR。取而代之，需要选择一个真正的对照来评估每个样本的质量。

3. ChIP-Seq 实验的精简——特异性 ChIP-Seq　考虑到 ChIP-Seq 技术产生的标签集过大，而且如果考虑到特定的生物学需求的话，如只对某个基因座感兴趣的话，可以通过特异性的方法筛选感兴趣的基因组区域，之后进行 ChIP 后的测序即 ChIP-Seq，这样会大大减少标签数过于庞大，降低分析的难度和节约成本。原理见图 13-13。

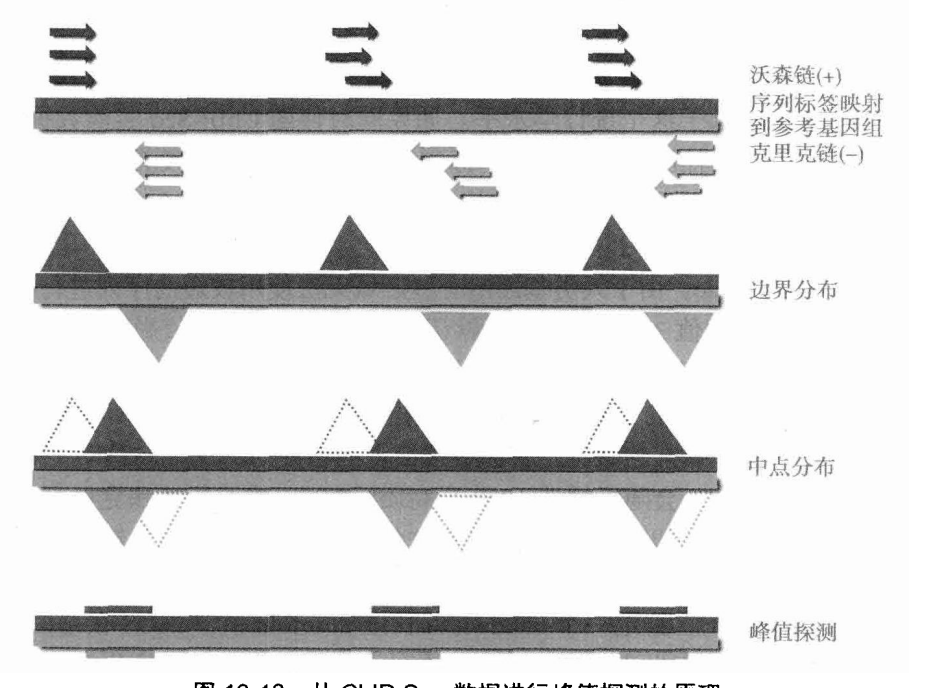


图 13-13　从 ChIP-Seq 数据进行峰值探测的原理

三、组蛋白修饰与其他表观遗传修饰存在协同调控关系

（一）DNA 甲基化和组蛋白修饰的相互作用

1. DNA 甲基化对组蛋白修饰的影响 DNA 甲基化对细胞分裂过程前后的组蛋白修饰模式的保持起到作用。在转录过程中，复制叉(replication fork)附近的染色质结构完全被破坏，因此在复制叉经过后，应该有一定的机制可以保持染色质状态得到很好的复原。DNA 甲基化模式应该是细胞分裂后重建染色质状态的主要标记。包含甲基化的 CpG 的区域在转录后重新组装为紧致的结构，而非甲基化的区域倾向于重新形成开放的染色质构象。使用 ChIP 技术发现，非甲基化的 DNA 倾向于和包含乙酰化修饰的组蛋白共处，而这种组蛋白正是开放染色质的标志；然而甲基化基团的出现和包含非乙酰化的组蛋白 H3 和 H4 的组装相关，这会导致紧致的染色质结构。DNA 甲基化和组蛋白修饰之间的关系可以部分受甲基化胞嘧啶结合蛋白例如 MECP2 或 MBD2 的调节。有证据表明 DNA 甲基化会抑制 H3K4 的甲基化。因此，发育过程中形成的 DNA 甲基化模式可能作为模板以维持许多代细胞分化的转录抑制模式，而无需识别 DNA 复制周围的序列或基因。

2. 组蛋白修饰对 DNA 甲基化的影响 研究表明在发育早期的 DNA 甲基化基础状态的建立可能受到组蛋白修饰的调节。根据这个模型，H3K4 甲基化的模式可能在 DNA 从头甲基化之前形成。H3K4 甲基化受 RNA 聚合酶Ⅱ的指导，因为 RNA 聚合酶Ⅱ募集特定的 H3K4 甲基转移酶。因为在早期胚胎基因组中，RNA 聚合酶Ⅱ大多结合在 CpG 岛，所以只有这些区域被 H3K4 甲基化标记，而其他基因组区域就不能被 H3K4 甲基化所标记。DNA 从头甲基化是 DNA 甲基转移酶 DNMT3a 和 DNMT3b 所行使的功能。由于 H3K4me 的干扰，胚胎中的从头甲基化只在基因组中的 CpG 位点发生，但是可能在 CpG 岛受到阻止。

（二）通过贝叶斯网络重构 DNA 甲基化和组蛋白修饰协同调控基因表达网络

DNA 甲基化和组蛋白修饰之间存在相互作用，且二者对基因表达都有直接的影响。贝叶斯网络是一种概率图形化的网络。目前，贝叶斯网络对于解决复杂多因素的关联研究已有广泛应用。首先计算全基因组范围的基因的组蛋白修饰和 DNA 甲基化的含量，然后，借助贝叶斯网络软件就可以进行表观遗传学网络的重构。在图 13-14 中，DNA 甲基化和 H3K4me3 之间存在密切的关系，而 H3K4me3 受到 PolⅡ间接的影响。此外，基因表达只受到 PolⅡ的直接影响。大部分属性的关系可以得到生物学证据的支持，另一部分新发现的关系为进一步的组蛋白代码的研究提供了很好的线索。

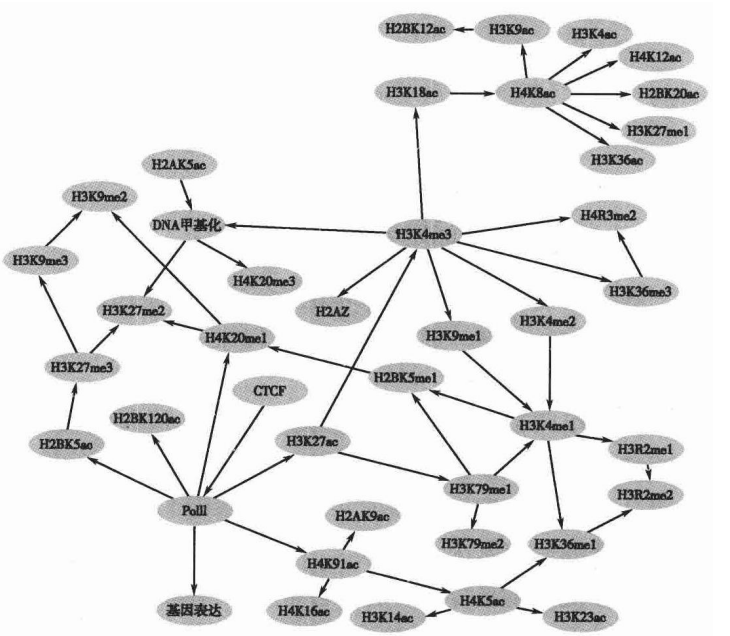


图 13-14 组蛋白修饰、DNA 甲基化和基因表达的贝叶斯网络

四、组蛋白修饰异常与疾病

尽管早在 20 世纪 80 年代中期就发现组蛋白修饰和染色质之间的联系，人们发现组蛋白修饰和癌症的关系已经比甲基化和癌症的关系要晚得多。许多研究表明 DNA 甲基转移酶以及其他辅助 DNA 甲基酶的效应蛋白都和组蛋白修饰酶有关联。因此，DNA 甲基化、组蛋白修饰模式和人类癌症有密切的关系。

（一）组蛋白修饰模式的改变被直接和癌症的发展相联系

许多组蛋白乙酰转移酶(histone acetyltransferase，HAT)基因在癌症中发生了改变，从而导致肿瘤抑制基因的H3乙酰化的丢失。负责维持平衡的组蛋白修饰状态如H3K4，H3K27和H3K79的甲基化的蛋白质的转座或过表达都会诱使癌症的发生。许多研究发现了组蛋白修饰和癌症的关系的证据，一个特异识别H3K4和H3K9的单甲基化和双甲基化形式的组蛋白去甲基转移酶LSD1的过表达和前列腺癌的发展有关。一个靶向H3K4me3的组蛋白去甲基转移酶PLU-1的上调和乳腺癌及睾丸癌相关。在食道癌中，一个编码JMJD2C结构域而特异识别H3K9me的组蛋白去甲基转移酶被频繁检测到。

（二）组蛋白修饰与其他疾病

组蛋白修饰除了与癌症密切相关外，与其他疾病包括神经退行性疾病，脑血管病，脱髓鞘性疾病和癫痫均有密切关系。在这些疾病中，均发现了组蛋白乙酰转移酶的升高，从而激活相关基因的乙酰化程度，促使机体保护机制相关的基因的表达。如果组蛋白乙酰转移酶发生突变或被组蛋白乙酰转移酶抑制剂抑制的话，将会导致疾病的发生。MeCP2可募集组蛋白去乙酰酶到甲基化的DNA区域而使组蛋白去乙酰化进而使染色质紧缩。MeCP2的异常可引起Rett综合征。

（三）食品营养与癌症表观遗传学

人们对营养和肿瘤的关系的研究已经有数十年的历史。目前普遍认为一定的饮食习性可以影响肿瘤的易患性。例如组蛋白甲基化由组蛋白甲基化酶所催化，组蛋白甲基化酶活性主要受到S-腺苷甲硫氨酸(S-Adenosyl methionine，SAM)的正性调节和其代谢产物S2腺苷同型半胱氨酸(S2-Adenosyl homocysteine，SAH)的负性调节。如果饮食中缺乏蛋氨酸和叶酸，则可引起SAM和SAH总量的降低，甲基化的能力降低。维生素B_{12}、叶酸、胆碱和维生素B_6可促进半胱氨酸的甲基化促使SAM和SAH的降解，增强组蛋白甲基化酶的活性进而抑制肿瘤的发生。这类研究有望成为今后表观遗传学的新的研究方向。

第四节 基因组印记

Section 4 Genomic Imprinting

一、基因组印记是表观遗传现象

哺乳动物的许多基因组印记特征都使基因组印记(genomic imprinting)成为后基因组学时代的有意义的生物学问题。基因印记一直以来都是一个谜，这部分是由于它们并不遵循传统的孟德尔遗传学规律。

基因组印记是在母本和父本之间产生功能性区别并在哺乳动物发育与生长中起重要作用的一种表观遗传学机制。在人类的基因组中有大量的基因易受基因组印记，主要表现为当亲本一方的基因被表达时另一方则被沉默。一个等位基因的沉默预先决定了所有与该基因有关的功能都只依赖于另一个活跃的等位基因。

在传统的遗传学中，子女会继承一个基因的两份拷贝，一份来自于父本，一份来自于母本，这两份拷贝的活性形式会影响子女的发育。但是当印记基因出现——这两个拷贝中一个会被来自母本或父本的分子调控机制所关闭，子女只会继承基因的一份拷贝的信息，从而承受更大的突变压力：如果一个功能拷贝受到损伤或遗失，那么就没有顶替的后备基因发挥作用。只拥有一个活跃等位的基因可能影响发育和导致人类疾病。印记基因对哺乳动物的发育非常重要，这类基因的突变会引起疾病。若维持印记机制的其他表观遗传标记发生异常也可能引起相同的疾病。

基因组印记是一种单等位基因表达的表观遗传现象。很多假设用于解释基因组印记在哺乳动物中的进化，但很少能解释是如何产生的。宿主防御假说认为印记产生于细胞内现存的机制对插入到

基因组中的外源 DNA 元件的沉默作用。此外，印记基因在基因组中通常是成簇出现，这些印记基因簇所对应的染色体区域被称作印记区(imprinted region)。

虽然发生、维持及进行印记的可能的分子机制正被识别，但还远未清楚这些单独起作用的印记基因的表达调控、功能及进化的分子机制。基于实验确定印记基因是具有挑战性的，因为一个特定印记基因的单等位表达可能只发生在特定的组织，或仅在特定的发育阶段。事实上，实验鉴定的人的印记基因的速度非常缓慢，通过预测发现潜在的印记基因被认为是扩充基因列表的一个有效方法。目前在人类基因组中预测的印记基因约有 200 多个，其中被实验验证具有印记表达的基因约有 50 多个。在小鼠中预测的印记基因约有 600 个左右，其中 70 多个经实验验证具有印记表达(http://www.geneimprint.com)，研究人员预计印记基因占人类所有基因的 1%，尚有一些基因有待发现。因此，利用机器学习技术识别印记基因仍是一种可行的方法，这将帮助研究人员揭示印记基因如何对人类健康的作用以及开展印记相关疾病的治疗。

二、机器学习是挖掘印记基因的有效方法

目前实验检测印记基因的主要方法是分析 DNA 甲基化和基因的差异表达。实验检测只是关注染色体的一小段区域。自从单等位和双等位基因不同的重复序列和 DNA 序列特性的被广泛关注，人们开始基于序列特征高通量的预测哺乳动物基因组的印记基因。目前主要预测印记基因的方法是用机器学习算法基于基因的序列特征预测全基因组印记基因。通过使用各种基因组特征和复杂的策略预测印记基因，对印记基因的发现有潜在价值。

1. 基于多元统计的方法　通过分析人类印记基因的 DNA 序列特征，发现序列组成对印记基因的识别有一定作用。在观察印记基因和非印记基因的编码区 DNA 水平的变量，通过适当的多元方法，主成分分析(PCA)和二次判别分析(QDA)，分析定量的基因组数据，得到对筛选印记基因有用的基因组属性。主成分分析(PCA)是一种多元统计方法。PCA 的主要思想是降低(代表大量相关变量)数据集的维度，同时保留尽可能多的数据集中的变量。二次判别分析(QDA)主要用于预测序列特征集中的成员。预测变量与二次判别相结合可以最好的预测预测组成员，使每一个基因基于它的序列特征可区分为印记基因和非印记基因。

通常使用的基因组特征分别是 [bp] %CpG 岛、[bp] % 串联重复序列、[bp] % 简单重复序列，它们能够区分人类印记基因的编码区，可以用于印记基因的预测。预测的流程图如图 13-15(左)所示。

应用这种序列分析方法，扫描人类全基因组基因，可以确定那些序列组成和印记基因相似的基因，对初步识别印记基因和后续的实验室工作指导具有潜在的价值。

预测模型如下：基于 n 个样本 $\{x_i, y_i\}$，$i=1,2,\cdots,n$(其中 x_i 为 d 维特征构成的向量，分为阳性样本集和阴性样本集，y_i 取自 $\{1,-1\}$ 而代表类别，使用 1 标记印记基因，-1 标记非印记基因)进行模型训练，利用主成分分析得到 p 个主要的基因组参数。主要原理：$Fp=a_{1m}Z_{X_1}+a_{2m}Z_{X_2}+\cdots a_{pm}Z_{X_p}$ 其中 $a_{1i}, a_{2i}, \cdots, a_{pi}(i=1,\cdots,m)$ 为 X 的协方差矩阵 Σ 的特征值对应的特征向量，$Z_{X_1}, Z_{X_2}, \cdots, Z_{X_p}$ 是经过标准化的原始变量值，因为在实际应用中，往往存在指标的量纲不同，所以在计算前须先消除量纲的影响，而将原始数据标准化。$A=(a_{ij})_{p\times m}=(a_1, a_2, \cdots a_m)$，$Ra_i=\lambda_i a_i$，$R$ 为相关系数矩阵，λ_i，a_i 是相应的特征值和单位特征向量，$\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_p\geqslant 0$。

根据软件的功能，利用内部和外部的验证方法对分类进行评估。

QDA 模型，使用一个内部验证方法称为交叉验证。此方法使用训练集检验模型。在这里，训练集分为 N 份。其中一份保留来进行验证结果，其余的用来建立模型。这一过程反复多次(训练集分为几部分)，即是 N 倍交叉验证。最后，所有的部分都被用来建立和验证模型。PCA 模型则使用外部验证检验集的方法。检验集的个数必须足够大(至少为训练集大小的 25%)，独立于训练集，而且检验集必须代表训练集。检验集的印记情况是已知的，用与建立模型不同的集合来评估 PCA 模型。

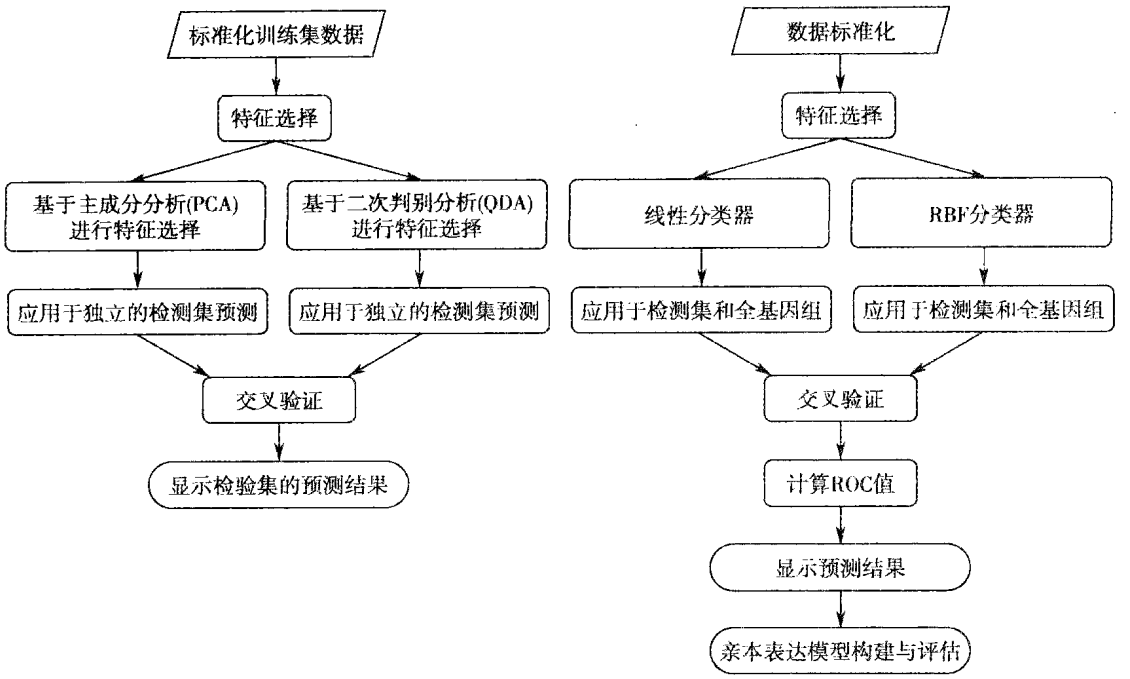


图 13-15 印记基因预测基本流程

2. 基于支持向量机(SVM)方法 目前有一种方法通过训练一个基于 DNA 序列特征的判别模型，不但能识别潜在的印记基因，还可以预测亲本的表达模式。在 23 788 个注释的常染色体小鼠基因中，模型预测出 600(2.5%)个可能的印记基因(64% 的印记基因为母本表达)。预测结果可以提供识别假定的候选基因，和亲本起源相关的复杂疾病的基因，可能涉及的疾病如老年痴呆症，孤独症，双相障碍，糖尿病，肥胖症，精神分裂症等。研究发现一个基因侧翼重复序列的数目类型及方向对预测一个基因是否印记很重要。基于 SVM 的预测流程如图 13-15(右)所示。

预测模型如下：基于 n 个样本 $\{x_i, y_i\}$，$i=1,2,\cdots,n$(其中 x_i 为 d 维特征构成的向量，y_i 取自 $\{-1,1\}$ 而代表类别，-1 作为印记基因，1 作为非印记基因)进行模型训练，SVM 利用下面的判别函数进行训练和检验：$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b\}$。其中，$\alpha_i$ 和 b 为待估参数，使得判别函数更好地拟合训练数据。具体的预测过程如下：

首先，收集供训练的阳性训练集印记基因共 44 个，阴性训练集非印记基因共 530 个。从 Ensembl 数据库中提取小鼠基因组注释及基因组区域包括：串联外显子的序列，串联内含子序列，基因上下游 100kb、10kb、5kb、2kb 和 1kb 范围内序列，CpG 岛含量及转录因子结合位点等。接下来，用 RepeatMasker 确定重复序列的出现率。计算不同重复序列类的统计量：每个窗口的总数，窗口内的属性重叠率。为了描述特征集之间的依赖性，用 t 检验计算所有训练集的互作对的两个特征，然后按 p 值排秩。在印记基因和非印记基因模型中，训练前 4000 个配对互作，把它们加到原始特征矩阵。因为所有这些互作的 p 值均满足 Bonferroni 控制的显著性阈值。在亲本表达预测模型中，前十个配对互作被保留($p<0.0003$)。最后，在一系列位点内和侧翼的 DNA 序列特征中对如重复元件，转录因子结合位点和 CpG 岛等进行定量化。根据这些特征，采用基于支持向量构建模型预测小鼠全基因组的候选印记基因和它们的亲本表达方式并通过交叉验证和独立的检验基因集评估预测模型。

三、基因组印记与表观遗传疾病有密切关系

哺乳动物的基因印记抑制基因表达，印记基因的异常表达会引发伴有复杂突变和表型缺陷的多种人类疾病。研究发现许多印记基因对胚胎和胎儿出生后的生长发育有重要的调节作用，对行为和

大脑的功能也有很大的影响，印记基因的异常同样可诱发癌症。如果抑癌基因有活性的等位基因失活便提高了发生癌症的概率，例如 *IGF2* 基因印记丢失将导致多种肿瘤，如 Wilm's 瘤。和印记丢失相关的疾病还有成神经细胞瘤，急性早幼粒细胞性白血病，横纹肌肉瘤和散发的骨肉瘤等。与基因组印记相关的疾病常常是由于印记丢失导致两个等位基因同时表达，或突变导致有活性的等位基因失活所致。调控印记基因簇的印记控制区发生突变将导致一系列基因表达失调，引发复杂综合征。基因组印记的本质为 DNA 修饰和蛋白修饰，所以和印记相关的蛋白发生突变也将导致表观遗传疾病。印记基因的缺陷可以导致一些严重的遗传疾病，如 Prader-Willi 综合征(PWS)、Angelman 综合征(AS)、Beckwith-Wiedemann 综合征(BWS)、假性甲状旁腺功能减退症(pseudohypoparathyroidism)、Russell-Silver 综合征等。

第五节 表观遗传学数据库及软件

Section 5 Databases and Softwares in Epigenetics

随着表观遗传学数据的不断增多，特别是高通量实验技术的不断涌现，在各基因组中产生了基因组范围的表观遗传遗传学修饰的图谱，对这些数据的存储和分析提出了挑战，应用于基因组研究的生物信息学解决了这一难题，研究者构建了专门的数据库用于存储表观遗传学实验测定的数据，并且开发了相应的算法对基因组范围内的数据进行分析。下面将分别介绍几个常用的表观遗传学数据库和分析软件。

一、表观遗传学常用数据库

表观遗传学数据库的构建和应用促进了表观遗传学的快速发展，有利于相关数据的重复利用。特别是近年来高通量技术的持续进步，数据量呈指数级增长，从原来几个基因的实验到现在的上百万的 CpG 位点甲基化的测定，数据量呈指数级上升，研究者在开发新的实验技术的同时也在不断开发新的数据库，用来存储高通量实验技术测定的数据。表观遗传学数据库主要是用来存储各种表观遗传学修饰(如 DNA 甲基化、组蛋白修饰等)的数据，例如人类表观基因组计划数据库(HEP)、人类组蛋白修饰数据库(HHMD)，以及表观遗传修饰修饰与某些生物学现象(如癌症等)的关系数据库，人类 DNA 甲基化与癌症数据库(MethyCancer)即是这样的数据库。下面将简单介绍这三个典型表观遗传学数据库的网站及其使用。

1. HEP 网址：http://www.epigenome.org/。HEP 的首页如图 13-16。人类表观基因组计划旨在确定、记录和解释人类所有基因在所有主要组织中的基因组范围内 DNA 甲基化模式。DNA 甲基化将基因组学、疾病和环境紧密的联系在一起，能够解释很多人类病理学现象。甲基化一般发生在 CpG 二核苷酸的胞嘧啶上，在不同的组织或疾病状态下，发生了截然不同的差异甲基化模式。这些甲基化可变性区域(MVP)是普遍的表观遗传学标记，与 SNP 一样，能够帮助认识和诊断人类疾病。

HEP 通过重亚硫酸盐技术测定 12 个人类正常组织中的 DNA 甲基化谱，目前已经完成人类 6 号、20 号和 22 号染色体上 DNA 甲基化谱的测定工作(图 13-16)，这些数据都可以从该数据库中免费获得。

2. HHMD 网址：http://bioinfo.hrbmu.edu.cn/hhmd，如图 13-17 所示。组蛋白修饰在染色质重塑、转录调控和人类疾病中起着重要的作用，当前基于 ChIP 技术的实验技术测定了人类基因组高通量的修饰位置信息，人类组蛋白修饰数据库是迄今为止收录各种实验测定的人类基因组组蛋白修饰最为全面的数据库，当前版本共涵盖了 43 种人类组蛋白修饰的大通量实验数据，并提供了通过文献得到的 9 种癌症相关的组蛋白修饰的信息。

该数据库包含用于可视化组蛋白修饰的工具 HisModView，用户可以通过 HisModView 在已有的基因组注释的背景下研究组蛋白修饰的分布、这些组蛋白修饰与 DNA 甲基化之间的关系，以及二者与相应基因功能元件的位置关系，据此来设计实验对感兴趣的区域进行湿实验研究，如图 13-18 所示。

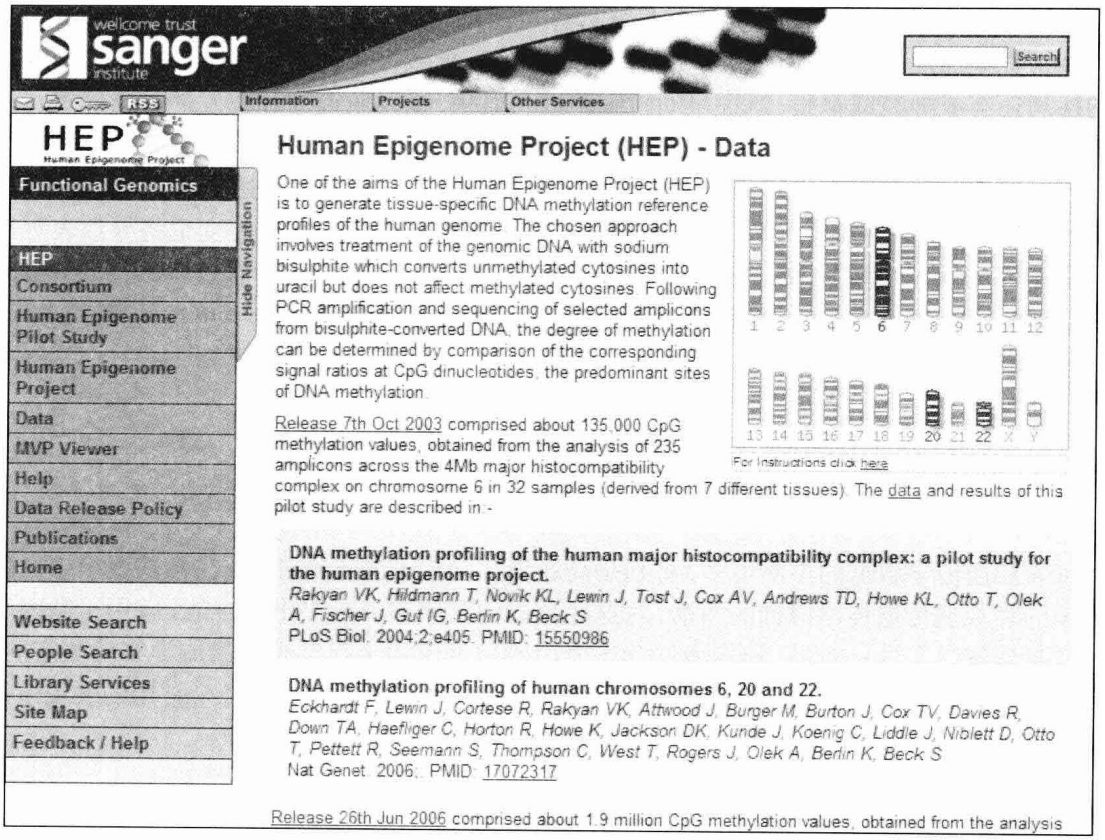


图 13-16 人类表观基因组计划数据库 HEP

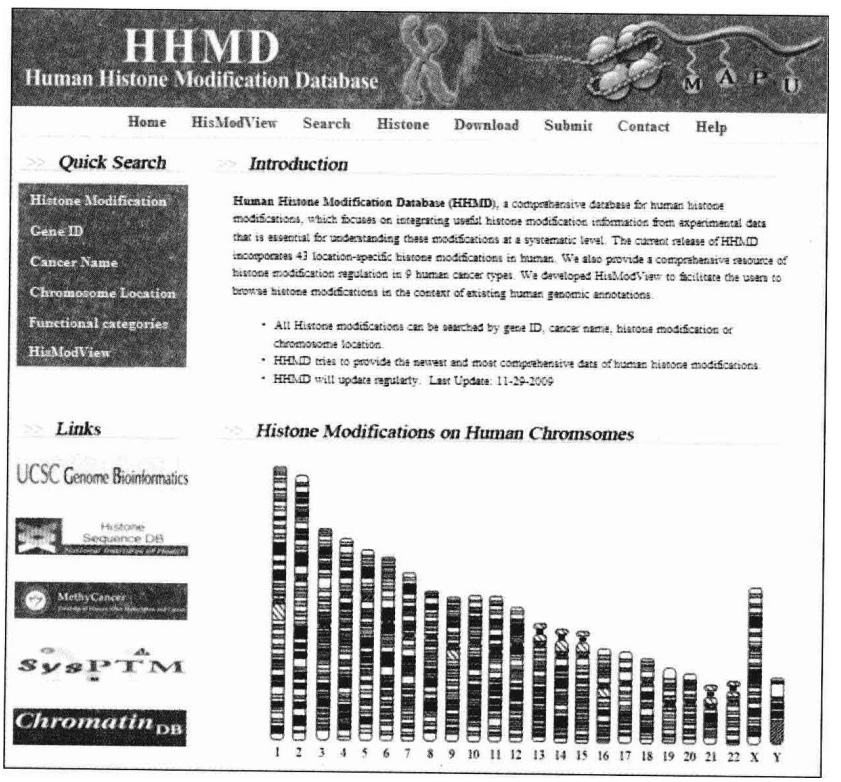


图 13-17 人类组蛋白修饰数据库 HHMD 首页

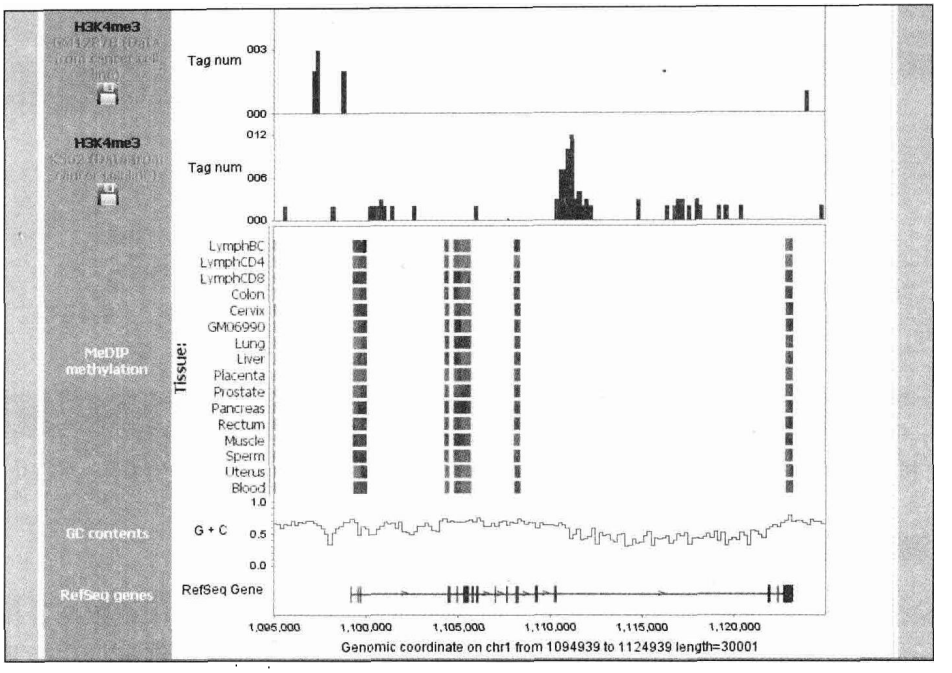


图 13-18 人类组蛋白修饰数据库 HHMD 的 HisModView 结果

用户可以通过搜索相应的基因组区域中的组蛋白修饰，该数据库提供了五种搜索组蛋白修饰的方式，分别是组蛋白修饰类型、基因 ID、功能注释、染色体定位、癌症类型(图 13-19)。并可以通过 HisModView 进行基因组水平可视化。整个数据库体现了很好的交互性操作，为研究组蛋白修饰与其他表观遗传调控元件如 DNA 甲基化之间的相互作用关系提供了一个很好的平台。

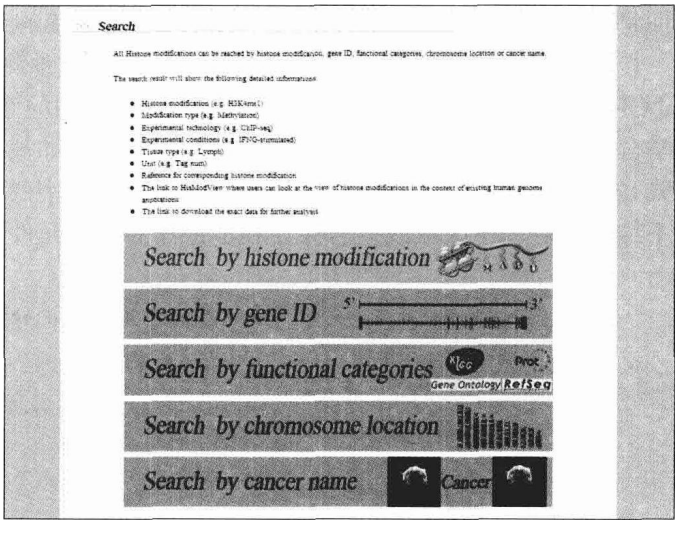


图 13-19 人类组蛋白修饰数据库 HHMD 的五个搜索入口

该数据库支持用户对搜索和可视化的结果进行下载，并且提供了处理基因组组蛋白修饰数据的 Java 程序。

3. MethyCancer 网址：http://methycancer.psych.ac.cn/。癌症被列为所有人类疾病的头号杀手之一，并继续威胁着人类的生命。目前的研究主要集中在加速对癌症发生的分子机制的阐明，并开发应用于癌症诊断、治疗及预后的有效手段。表观遗传学修饰的改变，特别是 DNA 甲基化的改变，通过调节癌基因激活、抑癌基因的沉默及染色质不稳定性，在肿瘤的发生中起着重要的作用。

MethyCancer 的开发旨在研究 DNA 甲基化、基因表达与癌症间的相互作用。该库包含 DNA 甲基化、癌症相关基因、突变、癌症信息和 CpG 岛等信息，对这些不同数据类型之间的互联互通进行了分析和讨论，并提供了搜索工具和可视化工具(MethyView)来帮助用户获取感兴趣的数据并在基因组的背景下查看 DNA 甲基化模式，如图 13-20 所示。

MethyCancer 的 DNA 甲基化搜索界面，可用来搜索用户感兴趣区域的甲基化模式。另外该数据库还提供了基因搜索界面、癌症搜索界面、序列搜索界面及重复序列搜索界面(图 13-21)。

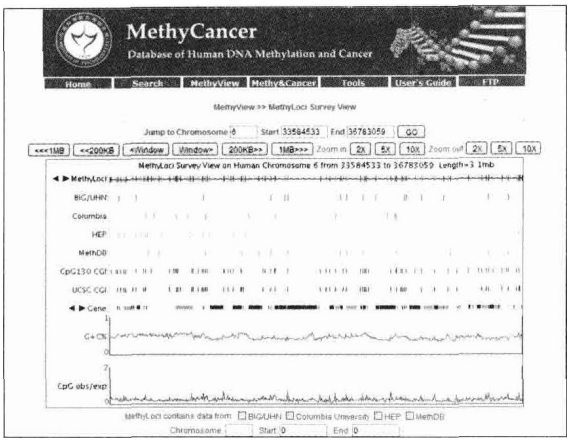


图 13-20 MethyCancer 的 MethyView 工具

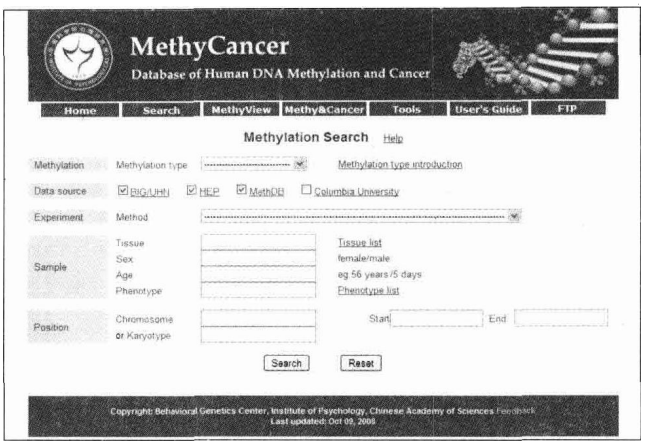


图 13-21 MethyCancer 的 DNA 甲基化搜索界面

二、表观遗传学常用软件

随着实验技术的不断进步及表观基因组数据不断增加，多种(表观)基因组数据分析软件也逐渐被开发。为了从基因组水平研究表观遗传学修饰，需要开发对表观遗传修饰进行功能基因组分析的软件。目前可用的软件中包括：界面友好的(表观)基因组分析和预测软件(EpiGraph)、基于支持向量机的 DNA 甲基化预测软件(Methylator)及基于互信息识别基因组功能 CpG 岛的软件(CpG_MI)。下面将对这三个计算表观遗传学软件的应用进行简单的介绍。

1. EpiGraph 界面友好的(表观)基因组分析和预测软件，网址：http://epigrap h.mpi-inf.mpg.de/WebGRAPH/，首页如图 13-22 所示。EpiGRAPH 可用于复杂的基因组和表观基因组数据集的生物信息学分析，这样的数据集经常包括共享特定属性(例如被一个特定的 TF 结合或表现进化保守的特定模式)的集合。通常，还需要一个包含落到相反的类中的基因构成的控制集，例如特定转录因子结合与不结合的启动子区域或显著保守和不保守的调控元件。即使当控制集没有给出，也很容易想到去

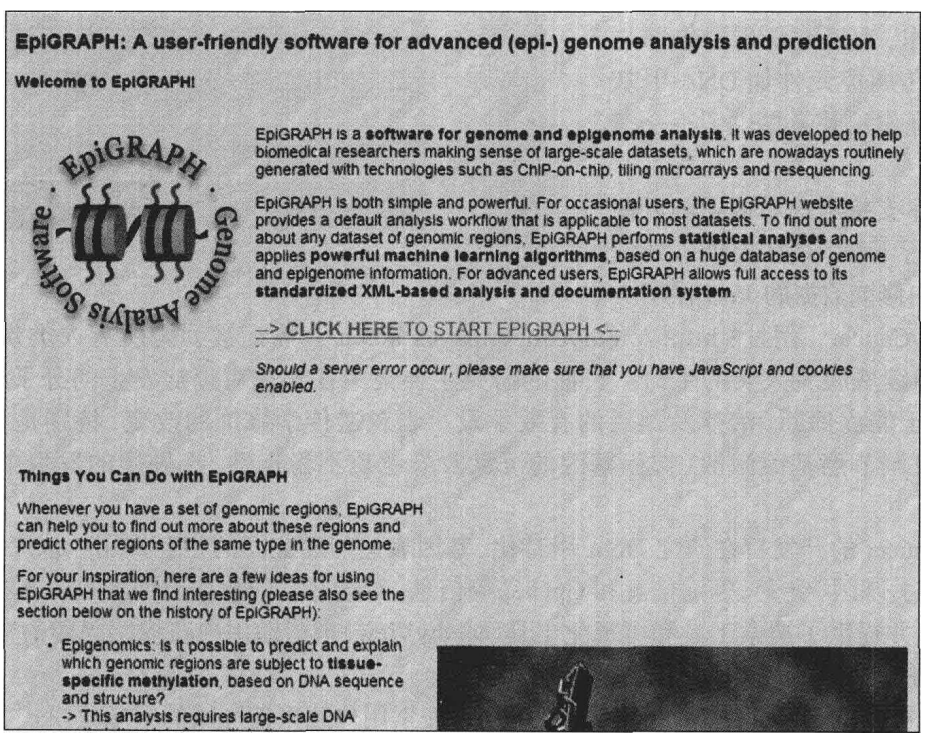


图 13-22 EpiGRAPH 的首页

手工建立一个控制区域的随机集合来作为一组给定基因组区域的补集。EpiGRAPH 因此将着重于以组为单位的涉及两类的基因组区域的分析。这个软件使得生物学家可以在脊椎动物基因组和表观基因组数据集中发现隐含的关联。

用户要通过 EpiGRAPH 进行研究首先应该在该网站上注册一个用户，之后分析时的数据和结果均存储在用户名下。登录网站用户可以根据网站的提示开始新的分析(图 13-23)，用户可以上传一组基因组区域，EpiGRAPH 将测试多种属性(包括 DNA 序列，染色质结构，表观遗传学修饰以及进化保守)是否在这些区域中富集或缺失。此外，EpiGRAPH 将会以预测的方式鉴别相似的基因组区域。该网站提供了详细的视频录像来指导用户使用该软件进行分析。

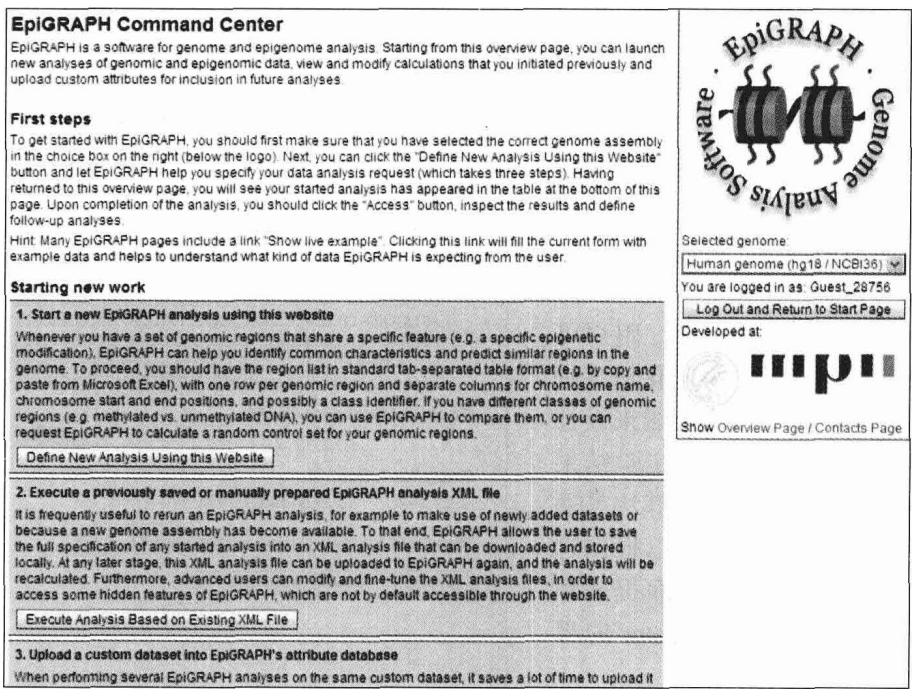


图 13-23 EpiGRAPH 执行一个新的分析任务的入口

通过以上各步骤，用户通过 EpiGRAPH 可得到的初步的结果(图 13-24)，还可通过 EpiGRAPH 进行进一步的分析。EpiGRAPH 为用户提供了统计分析的在线服务，用户只需要简单点击就可以完成对结果的分析，如网站相关的文献提供的单等位表达的例子中用于两组数据比较的箱式图。

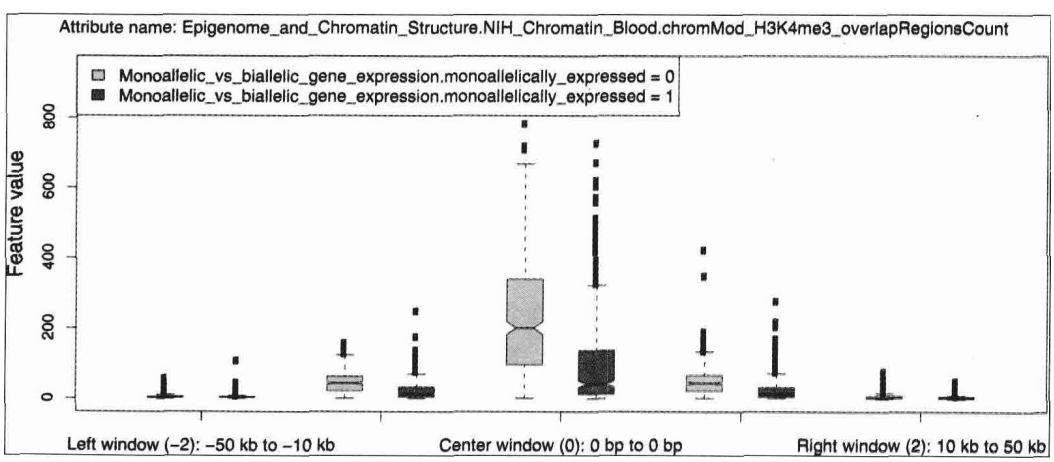


图 13-24 EpiGRAPH 的箱式图结果

(来自 EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data)

EpiGRAPH 解决了两个基因组生物学的普遍任务：一个是发现一组特定生物学作用的基因组区域(例如实验定位的增强子，表观遗传调控的热点或表现出疾病特异异常的位点)和从公共数据库中得到的大量的基因组注释数据的新的关联。另外一个是评价是否可能鉴别具有相似作用的额外的区域，而不必进行进一步的湿实验。

2. Methylator 基于支持向量机(SVM)的方法预测 CpG 二核苷酸中胞嘧啶的甲基化状态的软件，网址：http://bio.dfc i.harvard.edu/Methylator/。这种支持向量机模型在预测 DNA 甲基化方面比传统的机器学习方法(如神经网络、贝叶斯统计和决策树等)取得了更高的精度。该软件目前只提供在线服务。用户可以输入 GenBank、EMBL 或 GCG 数据库所要求的格式中任一 DNA 序列格式，也可以输入普通的序列格式。该在线平台提供两种上传数据的方法，一种是直接将序列粘贴到文本框中，另一种是直接上传序列文件。还可以更改甲基化分类的阈值，默认值为 −0.4(图 13-25)。

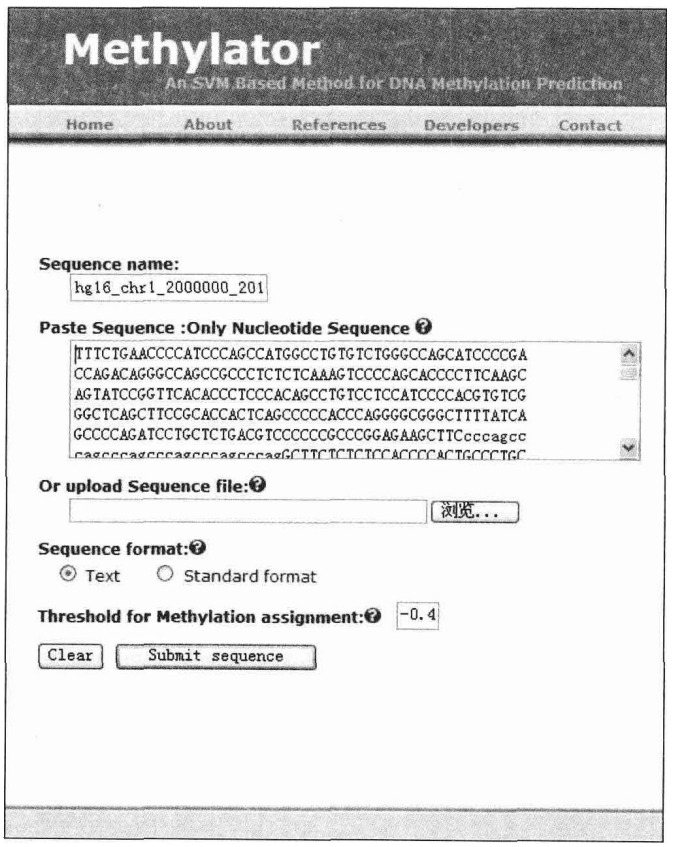


图 13-25 Methylator 的界面

通过 Methylator 的在线平台，用户可以对感兴趣的基因组区域的甲基化情况进行预测，从中挑选某些特殊的区域来设计实验，这样不仅节约实验成本，实验结果也有更好的可预见性。此外，该算法是通过 SVM 作为分类器，用户也可以根据这种思路，设计其他的分类器，从而提高预测 DNA 甲基化的精度。Methylator 的界面如图 13-25 所示。

提交序列后 Methylator 将在后台通过算法与甲基化的 CpG 位点进行预测，预测结果将在新的网页中显示(图 13-26)。甲基化的胞嘧啶将通过红色加粗的字来显示。目前在线服务每次只支持对一条 DNA 序列的甲基化预测，而且限制每次查询的序列长度限制在 50 000 个核苷酸以内，但这对基因组中的大部分功能区域应该是可用的。

3. CpG_MI 基于互信息识别基因组功能 CpG 岛的软件，网址：http://bioinfo.hrbmu.edu.cn/cpgmi/，界面如图 13-27 所示，提供在线平台和本地化软件。该方法不依赖于传统方法对 CpG 岛长度的限制，与之前用来识别 CpG 岛的方法相比，有着更高的精度，且识别出来的 CpG 岛大部分与组蛋白修饰区域相关。由于该算法只依赖于基因组 CpG 二核苷酸的分布，分析得到其他的脊椎动物基因组的 CpG 二核苷酸均服从相同的指数分布，因此可以将此算法推广到其他基因组中 CpG 岛的预测。

用户可以通过 CpG_MI 的在线服务来识别单个基因组区域中的 CpG 岛，用户有两种提交数据的方式：一种是提交基因组一段序列的位置信息，算法将在后台访问 UCSC 数据提取相应物种基因组区域的序列；一种是直接输入已经获取的 FASTA 格式的序列或上传 FASTA 格式的文件。如输入人类一号染色体 10 014 500 到 10 036 800 的基因组位置，结果页将列出这段区域中的每个 CpG 岛的位置、长度、CpG 数目、GC 含量和序列信息，将之下载到本地进行进一步的分析(图 13-28)。

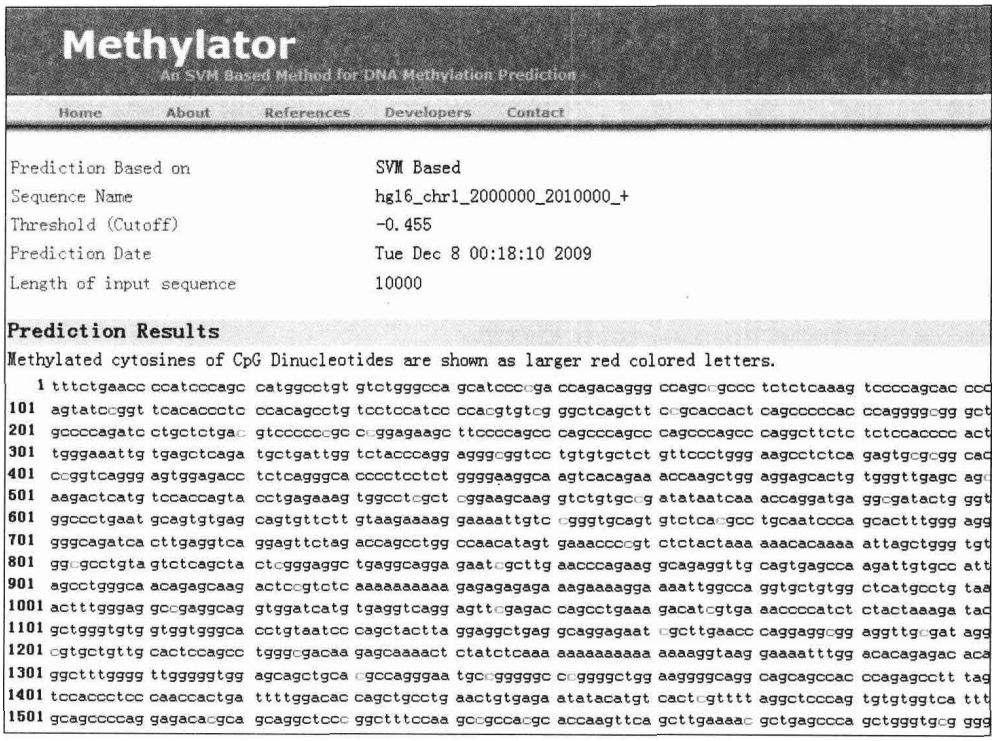


图 13-26 Methylator 的结果界面

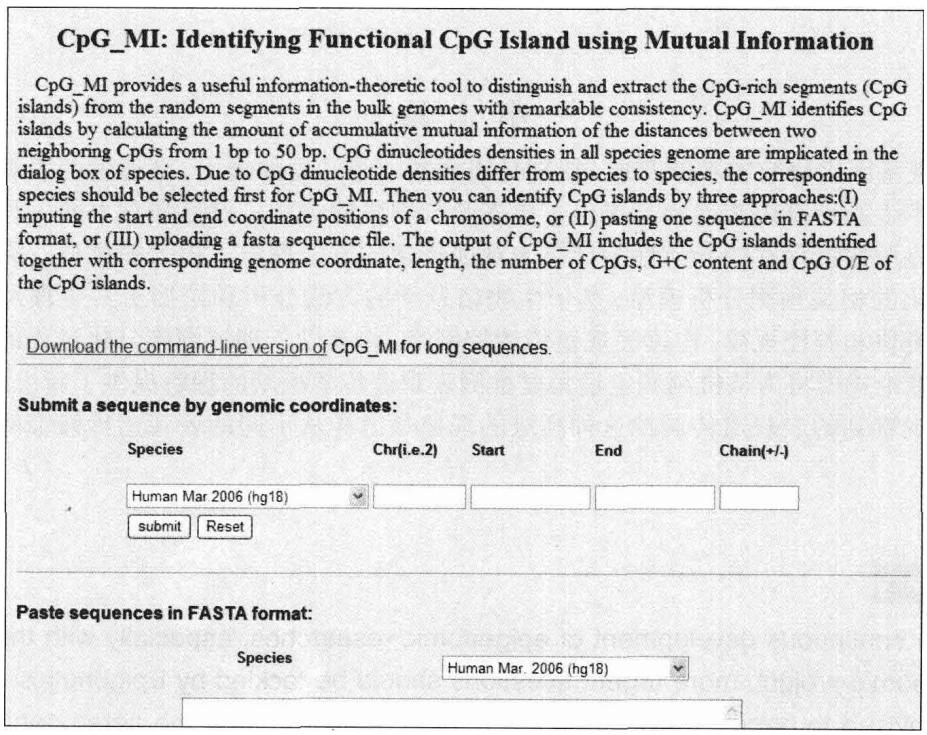
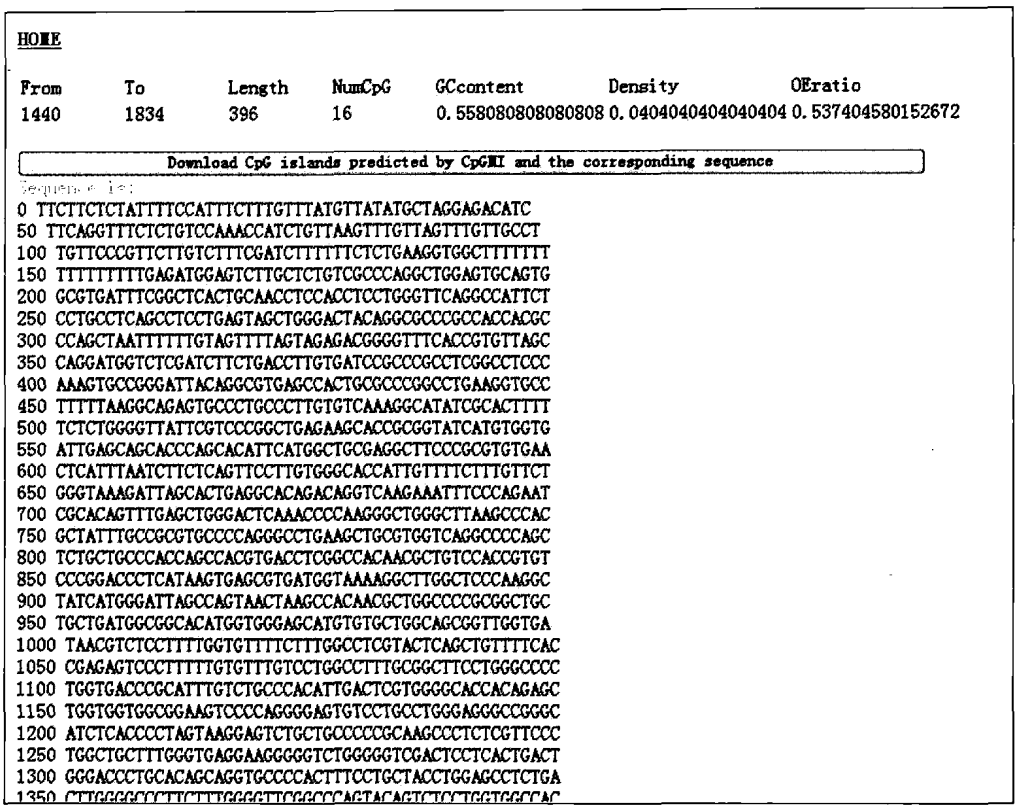


图 13-27 CpG_MI 的界面

通过CpG_MI，用户可以通过在线平台对其感兴趣的物种中的基因组区域中的CpG岛进行预测，也可以通过本地化软件对全基因组的 CpG 岛进行预测，对于 CpG_MI 所支持的物种，该软件对它们基因组的 CpG 岛已经进行了预测，用户可以直接下载使用。此外，通过 UCSC 可对这些 CpG 岛及其附近基因组区域进行可视化。


HOME

From	To	Length	NumCpG	GCcontent	Density	OEratio
1440	1834	396	16	0.558080808080808	0.0404040404040404	0.537404580152672

Download CpG islands predicted by CpGMI and the corresponding sequence

Sequence list:

0 TTCTTCTCTATTTTCCATTTCTTTGTTTATGTTATATGCTAGGAGACATC
50 TTCAGGTTTCTCTGTCCAAACCATCTGTTAAGTTTGTTAGTTTGTTGCCT
100 TGTTCCCGTTCTTGTCTTTCGATCTTTTTTCTCTGAAGGTGGCTTTTTTT
150 TTTTTTTTTGAGATGGAGTCTTGCTCTGTCGCCCAGGCTGGAGTGCAGTG
200 GCGTGATTTCGGCTCACTGCAACCTCCACCTCCTGGGTTCAGGCCATTCT
250 CCTGCCTCAGCCTCCTGAGTAGCTGGGACTACAGGCGCCCGCCACCACGC
300 CCAGCTAATTTTTTGTAGTTTTAGTAGAGACGGGGTTTCACCGTGTTAGC
350 CAGGATGGTCTCGATCTTCTGACCTTGTGATCCGCCCGCCTCGGCCTCCC
400 AAAGTGCCGGGATTACAGGCGTGAGCCACTGCGCCCGGCCTGAAGGTGCC
450 TTTTTAAGGCAGAGTGCCCTGCCCTTGTGTCAAAGGCATATCGCACTTTT
500 TCTCTGGGGTTATTCGTCCCGGCTGAGAAGCACCGCGGTATCATGTGGTG
550 ATTGAGCAGCACCCAGCACATTCATGGCTGCGAGGCTTCCCGCGTGTGAA
600 CTCATTTAATCTTCTCAGTTCCTTGTGGGCACCATTGTTTTCTTTGTTCT
650 GGGTAAAGATTAGCACTGAGGCACAGACAGGTCAAGAAATTTCCCAGAAT
700 CGCACAGTTTGAGCTGGGACTCAAACCCCAAGGGCTGGGCTTAAGCCCAC
750 GCTATTTGCCGCGTGCCCCAGGGCCTGAAGCTGCGTGGTCAGGCCCCAGC
800 TCTGCTGCCCACCAGCCACGTGACCTCGGCCACAACGCTGTCCACCGTGT
850 CCCGGACCCTCATAAGTGAGCGTGATGGTAAAAGGCTTGGCTCCCAAGGC
900 TATCATGGGATTAGCCAGTAACTAAGCCACAACGCTGGCCCCGCGGCTGC
950 TGCTGATGGCGGCACATGGTGGGAGCATGTGTGCTGGCAGCGGTTGGTGA
1000 TAACGTCTCCTTTTGGTGTTTCTTTGGCCTCGTACTCAGCTGTTTTCAC
1050 CGAGAGTCCCTTTTTGTGTTTGTCCTGGCCTTTGCGGCTTCCTGGGCCCC
1100 TGGTGACCCGCATTTGTCTGCCCACATTGACTCGTGGGGCACCACAGAGC
1150 TGGTGGTGGCGGAAGTCCCCAGGGGAGTGTCCTGCCTGGGAGGGCCGGGC
1200 ATCTCACCCCTAGTAAGGAGTCTGCTGCCCCCGCAAGCCCTCTCGTTCCC
1250 TGGCTGCTTTGGGTGAGGAAGGGGGTCTGGGGGTCGACTCCTCACTGACT
1300 GGGACCCTGCACAGCAGGTGCCCCACTTTCCTGCTACCTGGAGCCTCTGA
1350 CTTGGGGCCCTTCTTTGGGGTTCGGCCCAGTACAGTCTCCTGGTGGCCAC

图 13-28 CpG_MI 的结果界面

小 结

随着表观基因组研究的不断开展，特别是表观基因组计划的提出，高通量的表观基因组水平数据的不断产生给计算表观遗传学提出了更多亟待解决的问题。设计强有力的工具来处理海量的数据以及克服数据存在的异质性为计算表观遗传学的发展提出了很大的挑战。除了开发表观遗传修饰变化的相关预测分析模型，基于生物信息学的方法分析和挖掘正常生理及疾病状态下的表观调控模块和调控规律，构建表观遗传调控网络，将有助于理解细胞分化过程的染色质变化及其意义。理论建模将为从机制和定量角度推测表观遗传学机制的理解提供了新思路，建模研究有助于解释观察到的表观遗传调控这种高级的现象是怎样从不同的表观遗传修饰的动态互作中产生。

Summary

With the continuous development of epigenomic researches, especially with the proposition of the epigenomic project, more urgent questions should be tackled by Epigenetics. Robust tools should be devised to handle the vast amout of data and overcome the heterogenicity between data, which challenge the development of Computational Epigenetics. Besides the development of predictive models of variations in epigenetic modifications, analysis and mining of epigenetically regulatory modules and regulatory mechanisms in normal and aberrant physiological genomes and further the construction of Epigenetic networks by Bioinformatics would help understand the chromatin changes and the relevant significances during the cell differentiation. Theoretical

modeling is helpful in explanation of how the observed phenomenon is generated by the dynamic interplay among various epigenetic modifications.

（张 岩 苏建忠 王 芳）

习 题

1. 下列哪种现象不属于表观遗传异常：
   A. 印记丢失
   B. 全基因组整体去甲基化
   C. 抑癌基因沉默
   D. 基因印记
2. 下列说法正确的是：
   A. DNMT1 是一种从头甲基化酶
   B. 在体细胞中，甲基化的 CpG 可突变为 TpG，并可遗传
   C. Alu 元件是一种特殊的 CpG 岛
   D. DNA 甲基化阻碍转录因子的结合从而抑制转录
3. 以下因素对提升 CpG 岛甲基化预测准确性的贡献最小是：
   A. 是否整合表观基因组特征
   B. 是否特征筛选
   C. 是否使用最新的人类基因组版本的序列进行预测
   D. 是否只使用序列模式
4. 简单介绍 DNA 甲基化在转录调控中的作用。
5. 简单介绍 CpG 预测常用算法的基本原理。
6. 简单介绍 DNA 甲基化研究常用的实验方法和计算方法。
7. 简述 MACS 软件探测 ChIP-Seq 峰值的基本步骤。
8. 核小体的基本结构如何？
9. 常用核小体定位软件的基本原理及优缺点。
10. 请简述印记基因的预测方法。

主要参考文献

1. Benjamin L. Genes Ⅷ. United States: Benjamin Cummings; 2003.
2. 薛京伦. 表观遗传学——原理、技术与实践. 上海：上海科学技术出版社；2006 年.
3. Ferguson-Smith A. C., Greally J. M., Martienssen R. A. Epigenomics. Berlin: Springer; 2009.
4. Su J. Z., Zhang,Y., Lv, J., et al. CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. Nucleic Acids Res., 2010, 38(1):e6.
5. Zhang Y., Lv J., Liu H. B., et al. HHMD: the human histone modification database. Nucleic Acids Research, 2010: D149-154.
6. Esteller M. Epigenetics in cancer. N. Engl. J. Med., 2008, 358(11): 1148-1159.
7. Suzuki M.M., Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet., 2008, 9(6): 465-476.
8. Eckhardt F., Levin J., Cortese R., et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat. Genet.,

2006, 38(12): 1378-1385.

9. Segal E., Fondufe-Mittendorf Y., Chen, L. Y., et al. A genomic code for nucleosome positioning. Nature, 2006, 442(7104): 772-778.
10. Ioshikhes I. P., Albert I., Zanton S.J., et al. Nucleosome positions predicted through comparative genomics. Nat. Genet., 2006, 38(10): 1210-1215.
11. Laird P. W. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet., 2010, 11(3): 191-203.
12. Nowack M. K., Shirzadi R., Dissmeyer N., et al. Bypassing genomic imprinting allows seed development. Nature, 2007, 447: 312-315.

第三篇 生物信息学与人类复杂疾病

第十四章 人类复杂疾病与计算系统生物学

CHAPTER 14 HUMAN COMPLEX DISEASE AND COMPUTATIONAL SYSTEMS BIOLOGY

第一节 引 言

Section 1 Introduction

人类常见病包括肿瘤、心脑血管病、代谢系统疾病、神经系统疾病、精神和行为异常等，它们绝大多数都是复杂疾病。复杂疾病与单基因缺陷性遗传病不同，不符合孟德尔遗传定律，疾病的发生和发展涉及复杂的生物过程。如何诊断和治疗复杂疾病（研究复杂疾病机制）是21世纪生物医学重大的挑战之一。几百年来，由于科学技术的限制，虽然人们已经积累了大量的资料和数据，也取得众多研究成果，但对复杂疾病本质的认识还相差甚远。近30年来，伴随着生物科学、计算机技术的迅速发展，人们在生命活动的各个层面，尤其是在分子水平积累了大量的实验数据和研究成果，建立了多个疾病相关的数据库，对疾病有了更深刻的认识。生物“组”学（omics）和系统生物学（systems biology）方法的不断发展，为人们从多个层面系统地研究复杂疾病提供了有力工具。生物医学已经进入了系统和生物组学的新时代。应用复杂体系和整合医学（integrative medicine）的研究模式揭示复杂疾病的本质，认识其发病机制，寻找到正确的诊断和防治方法是当前乃至未来几十年复杂疾病研究的主要内容。本章主要介绍复杂疾病相关的基本概念和知识，重点介绍几个常用的复杂疾病数据库，最后探讨目前应用于复杂疾病研究的系统生物学方法。

第二节 复杂疾病概述

Section 2 Overview of Complex Disease

疾病是机体在一定病因的损害作用下，因机体自稳调节紊乱而发生的异常生命活动过程。各种致病因素包括遗传突变、DNA损伤和异常修复、表达调控紊乱、蛋白质功能异常等，往往与环境因素直接或间接地作用于易感个体，从而使机体发生一系列的功能、代谢和形态结构的变化，并由此产生各种症状和体征。现代医学认为复杂疾病是由内因和外因共同作用的结果。内因主要是遗传物质的变异，包括染色体异常、基因突变、单核苷酸多态的插入缺失变异、拷贝数变异、DNA修饰和核小体修饰等，这些遗传变异可能直接导致机体功能先天异常，或使机体对外界刺激的敏感性发生变化等。外因是诱发变异基因病变的多种外界因素，包括感染、损伤、环境、情绪和情感、教育和社会因素等。当具有某种遗传变异的人接触到相应不良外界因素的时候，疾病的发病率可能增加几倍、几十倍甚至上百倍。随着现代分子生物学和医学研究的不断发展，尤其是人类基因组计划（Human Genome Project，HGP）的完成和国际人类基因组单体型图计划（Human Haplotype Map，HapMap）的开展，积

累的大量的疾病分子水平的知识，使人们不仅可以认识疾病的本质，更可以利用这些知识探索和创造疾病诊疗的新方法和新技术。

一、孟德尔遗传疾病与复杂疾病

对于遗传因素作为主要发病原因的疾病，依据遗传因素可以将其区分成：单基因遗传疾病、染色体变异遗传疾病、多基因遗传疾病及线粒体变异引起的疾病。其中，单基因遗传疾病又称为孟德尔遗传疾病。孟德尔遗传疾病可分为：常染色体显性遗传，常染色体隐性遗传，X 连锁遗传，Y 连锁遗传等。

其他人类常见疾病如心脑血管疾病、肿瘤、代谢疾病、神经疾病等，往往不是由单基因或者单因素决定的，而是涉及多种基因、环境及遗传等方面因素。这些疾病的发生不符合孟德尔遗传定律，称为复杂疾病。

二、复杂疾病通常涉及多基因和蛋白质

复杂疾病(complex disease)，又称多基因病(ploygenic disorder)，是指由多个基因和环境因素共同作用的一类疾病，如高血压、哮喘、癌症等。这些疾病的发生通常是由于众多基因突变、表达调控紊乱等因素引起的蛋白质功能及互作关系紊乱、代谢和信号通路异常。与单基因病相比，复杂疾病具有遗传异质性、基因微效性、表型复杂性、种族差异性以及环境相关性等特点。随着基因表达检测技术、SNP 检测技术、蛋白质检测技术、表观遗传检测技术等高通量分子标记检测方法的迅速发展，人们已经开始在全基因组范围内系统地研究复杂疾病的发生过程。如 Ding 等利用基因芯片技术，试图从全基因组范围寻找肺癌的相关基因，最后识别出 188 种人类肺部癌变的 623 个已知或潜在相关的基因，并利用系统生物学方法从中筛选出和肺癌显著相关的 26 个基因。其中 *NF1*、*APC*、*RB1*、*ATM* 等基因的突变，*LRP1B* 基因序列的缺失以及 *PTPRD* 基因序列的改变都是诱导肺癌发生的原因。同时，Ding 等也发现了和肺癌相关的一些重要的信号通路如 MAPK、p53、WNT 和 mTOR(图 14-1)。

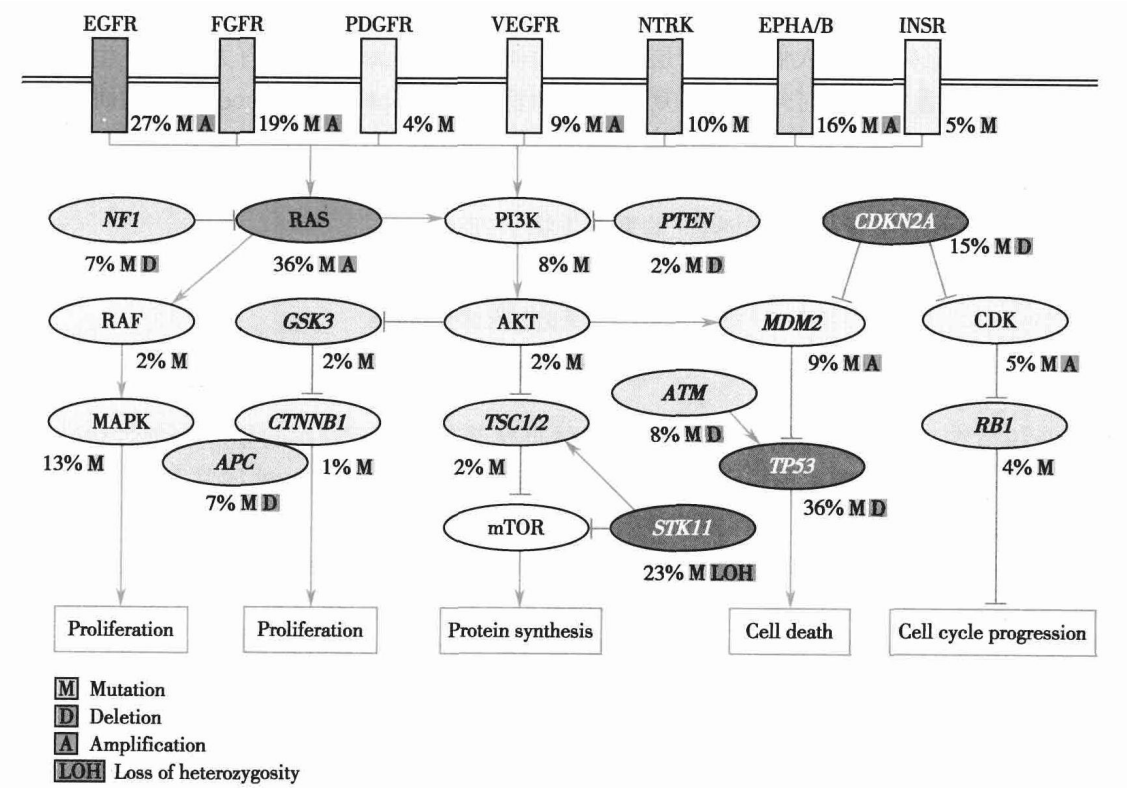


图 14-1 Ding 等研究的与肺癌相关的 26 个高频基因

同时 HapMap 的完成为研究者建立特定疾病(表型)与遗传多态性位点之间的联系提供了可能，进而为预防、诊断和治疗复杂疾病提供了新的方法。目前美国 NIH 全基因组关联分析数据库已收录了包括肿瘤(肺癌、结肠癌、直肠癌、前列腺癌等)，自身免疫病(糖尿病、类风湿性关节炎等)，心脑血管疾病(冠心病、心肌梗死、高血压等)，神经退行性疾病(阿尔茨海默病、帕金森病、肌萎缩性侧索硬化、多发性硬化症等)等 100 多种重要复杂疾病和性状易感基因。

三、复杂疾病受环境因素影响

复杂疾病不仅与遗传因素有关，与环境因素也有着非常密切的关系。据世界卫生组织报告，全球超过 20% 的疾病是由环境暴露造成的。而每年大约有 1300 万人死亡归因于的环境原因。在人类最不发达的地区，近三分之一疾病可以归因于环境原因。同时，报告还指出，在造成人类死亡率最高的几种疾病中(如心血管疾病、呼吸道感染、癌症、慢性阻塞性肺病等)，85% 以上的疾病受环境因素影响。可见，环境对于人类复杂疾病的影响之大。首先，环境因素会诱导基因的某些位点发生突变或表达变化最终引发疾病。比如癌基因在通常情况下处于抑制状态。如果细胞被紫外线照射或者受到环境因素的刺激，癌基因就可能从原来的抑制状态变成激活状态，进而使得正常细胞发生癌变转化为癌细胞。虽然每个人的基因组中都存在癌基因，但是一般都处于抑制状态。如果没有环境发生改变，就不会使癌基因从被抑制转为被激活。其次，基因的个体差异使环境对不同人的影响不同。个体基因组并不完全相同，对环境改变的敏感性也不同。所以当处于同样的环境下，有的人发病，有的人不发病。即使同卵双生的双胞胎在基因组序列上完全一致，对不同环境因素的响应也将导致表型差异。越来越多的实验证明基因与环境之间的相互作用在复杂疾病的发生或发展过程中起着关键作用，它们之间的相互作用是极其复杂和非线性的。一个基因在不同的环境中会产生不同甚至是完全相反的表型，因此单纯从遗传角度去研究疾病不足以全面了解复杂疾病的发生、发展过程。为了全面、系统地研究环境对疾病的影响，科学家们开展了环境基因组计划(Environment Genome Project, EGP)，并识别哪些人类基因能增加对环境相关疾病的个体易感性。该计划的目标主要是识别环境基因易感人群间的基因变化，并取得了关键性的进步。目前，研究人员已经完成了第一阶段工作，对 200 个具有环境响应性的基因进行了排序和编目。该计划为研究环境、人类基因和疾病发展之间的关系提供了一个平台。

四、疾病的分类

各种研究对疾病的分类方法不一。简单地，可以分为常见疾病和罕见疾病。罕见疾病(rare disease)是对于某一大群体年发病率少于一人的疾病。也可以根据疾病遗传因素的多少，分为孟德尔疾病和复杂疾病。

早在 20 世纪 60 年代，Victor 博士就开始带领一批学者和研究人员进行基因数据库 MIM 的创作，收集孟德尔疾病的相关信息并于 1966 年发布了第一个版本。后来形成了在线数据库 OMIM，并对收录的疾病进行了很大的扩充，不仅包括孟德尔疾病，对许多复杂疾病也进行了详细的记录，但是 OMIM 并未提供疾病的分类注释信息。

最早的疾病分类体系创建于 19 世纪 50 年代，并在 1893 年由国际统计研究所出版了《International List of Causes of Death》。世界卫生组织(WHO)于 1948 年开始负责 ICD(International Statistical Classification of Diseases and Related Health Problems)的编写任务，并首次加入了发病原因的信息。世界卫生大会(WHA)于 1967 年通过了世界卫生组织对疾病的命名规则，并要求其成员国使用 ICD 上疾病命名规则以及疾病死亡率和发病率的统计数据。ICD 疾病分类体系按照疾病的某些特征将疾病进行分门别类。现有版本(ICD-10)包含 15.5 万种编码。各个国家分别引进这种疾病分类体系并进行改进。如加拿大于 2000 年引进该版本并进行了改进成为“ICD-10-CA”等。中国根据“ICD-10”颁布了《第二次国家卫生服务调查疾病分类——编码表》对疾病进行了分类，共 19 类：①传染病；②寄

生虫病；③恶性肿瘤；④良性肿瘤；⑤内分泌疾病(营养和代谢疾病及免疫疾病)；⑥血液和造血器官疾病；⑦精神病；⑧神经系统疾病；⑨眼及附器疾病；⑩耳和乳突疾病；⑪循环系统疾病；⑫呼吸系统疾病；⑬消化系统疾病；⑭泌尿生殖系统疾病；⑮妊娠；⑯分娩病及产褥期并发症；⑰皮肤和皮下组织疾病；⑱肌肉、骨骼系统和结缔组织疾病；⑲损伤和中毒。

Bioinformatics Core Facility(BCF)联合 NuGene 工程创立了 Disease ontology(DO)，目的是为了将“ICD-9-CM”或者 SNOMED 上的疾病快速地映射到这个体系中。实际上，DO 的组织形式是一个有向无环图，它将疾病按照组织形式分成不同的类别并逐步细化。DO 最早的版本(version1.0)于 2003 年 8 月问世，现在支持的版本是 version2.1。

第三节 复杂疾病数据库

Section 3 Complex Disease Database

随着生物技术的不断发展，人们在疾病研究上取得了丰富的成果。不仅找到孟德尔疾病等简单疾病的遗传基因和染色体位点区域，而且在复杂疾病研究方面也取得了长足的进展，利用高通量分子标记检测技术和全基因组关联分析等方法定位了许多复杂疾病的染色体区域和候选基因座。研究人员分析整理了这些疾病相关的信息，形成了多个疾病数据库，为复杂疾病的进一步深入研究奠定了基础。这些数据库主要基于文献、关联分析及生物学实验的结果，记录了疾病表型、相关的染色体区域、候选基因等多方面的信息。其中比较著名的疾病数据库包括美国国家生物技术信息中心(National Center for Biotechnology Information，NCBI)的 OMIM 数据库，美国国立卫生研究院(National Institutes of Health，NIH)的 GAD 和美国癌症研究所(National Cancer Institute，NCI)建立的 CGAP。另外，世界卫生组织编纂的 ICD 疾病分类体系等疾病分类资源在本节中也将做简要的介绍。

一、人类孟德尔遗传在线

MIM(Mendelian Inheritance in Man)是一个将遗传病分类并链接到相关人类基因组中的基因数据库。它的在线版本是人类孟德尔遗传在线(Online Mendelian Inheritance in Man，OMIM)，可以通过网址 http://www.ncbi.nlm.nih.gov/omim/ 进行访问。OMIM 为临床医生和科研人员提供了权威可信的关于遗传疾病及相关疾病基因位点的详细信息。

OMIM 是目前权威的人类遗传疾病数据库，有着广泛的应用领域。例如，临床医生可以将病人的临床表型输入数据库中查找相关的疾病信息，还可以针对某些感兴趣的基因或者疾病进行搜索。在 OMIM 中搜索基因和疾病时，同时查询到基因和疾病相关的信息：如基因序列、染色体位置以及一些相关参考文献等。用户可以通过 MIM 号(ID)、疾病名、基因名或者疾病的一些表征进行搜索(图 14-2)。其中，Limits 选项提供给用户一些更严格的搜索条件，如范围、染色体、时间等。

在输入搜索关键词并运行后，网站会在搜索结果中列出与搜索记录最相近的 20 个记录，可依照个人习惯更改显示记录的数目。在 OMIM 数据库中，每一个记录都会有唯一的 6 位数编码，这种编码可以表示这种遗传病是常染色体显性(隐性)遗传、X 连锁还是 Y 连锁等，详见表 14-1。

同时在大部分 OMIM 号前面都会有一些特殊的符号，分别表示不同的含义。其中，“*”号代表着已知疾病基因的序列信息，没有加“*”号表示其遗传模式虽然已有推测，但没有被证实或者这个基因与其他记录所包含的基因位点的分离情况还不清楚；“#”号表示这种表型可以由两个或者多个基因中的一个发生突变而引起；“+”号表示这个记录包含基因的序列信息和表型；“%”表示记录中描述了一个已知的孟德尔表型，但是对其潜在的分子机制还不清楚；“^”表示该记录已不存在或者被其他记录所代替。此外，OMIM 数据库还提供下载功能，用户可以通过 FTP 方式下载

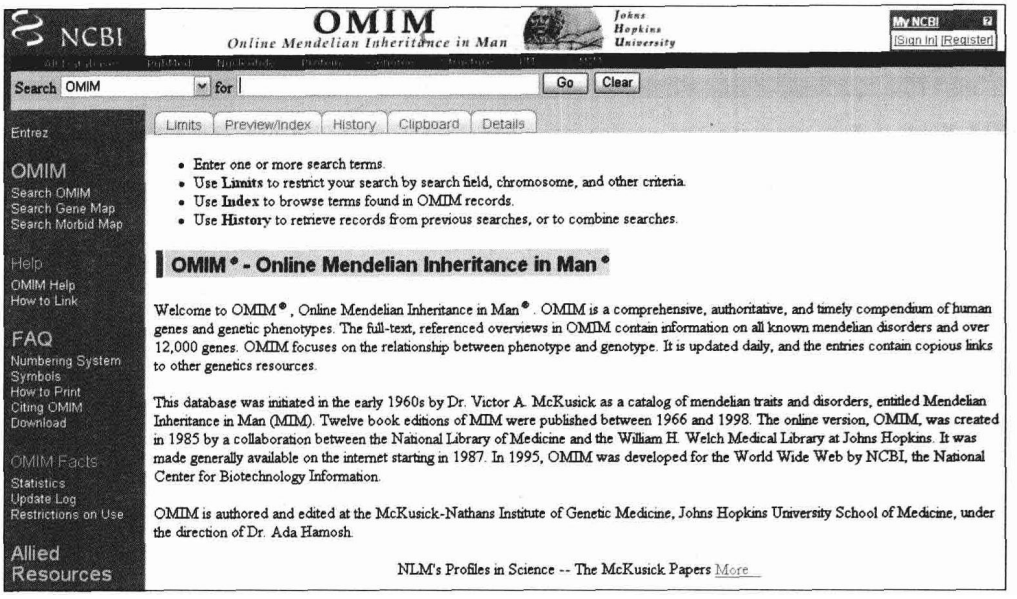


图 14-2 OMIM 主页面

表 14-1 OMIM 编号代表的遗传方式

MIM 编号范围	遗传方式
100000-199999	常染色体显性遗传或表型(于 1994 年 5 月 15 号创建)
200000-299999	常染色体隐性遗传或表型(于 1994 年 5 月 15 号创建)
300000-399999	X 连锁位点或表型
400000-499999	Y 连锁位点或表型
500000-599999	线粒体位点或表型
600000-	染色体位点或表型(于 1994 年 5 月 15 号创建)

(网址：ftp://ftp.ncbi.nih.gov/repository/OMIM/)，里面包含了全部的 OMIM 文件(omim.txt.Z)，基因文件(genemap)，解释文件(genemap.key)，以及疾病信息(morbidmap)。OMIM 还提供 genemap 和 morbidmap 的网络查询形式。如图 14-3，图 14-4。

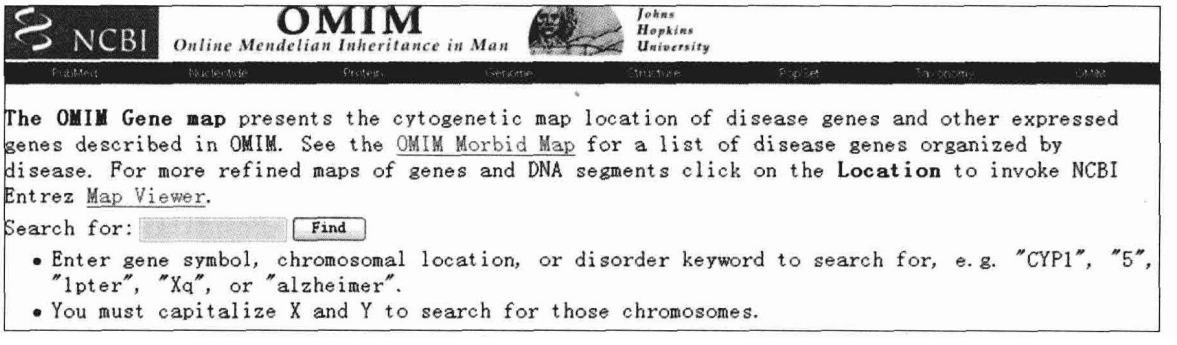


图 14-3 genemap 查询页面

下面以 Alzheimer's Disease(AD)为例，简单介绍一下 OMIM 数据库的使用。在 OMIM 查询框中输入“Alzheimer's Disease”就可以在 OMIM 上得到这种疾病相关的信息(图 14-5)。也可以输入该疾病的简写形式 AD，疾病表征(如 Senile Dementia)，或者与该疾病相关的基因名(APOE4)，查看检

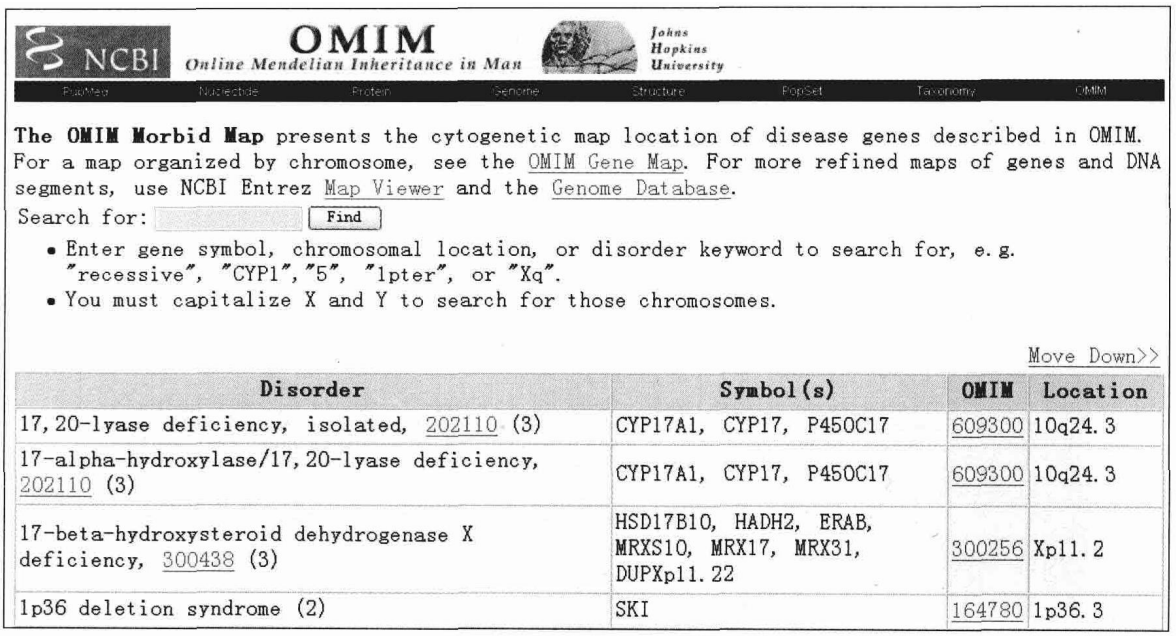


图 14-4 morbidmap 查询页面

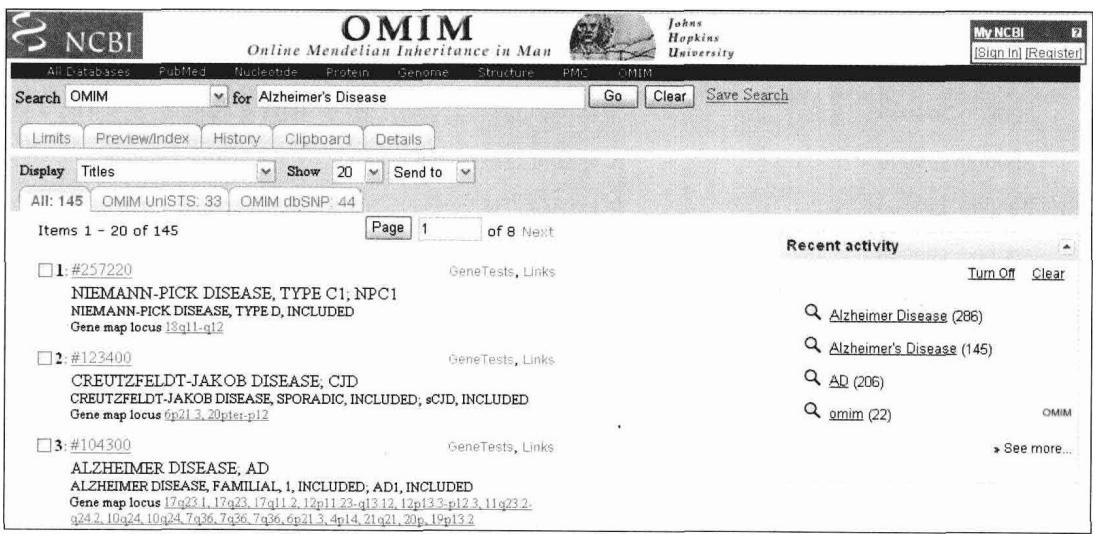


图 14-5 Alzheimer Disease 疾病 OMIM 查询结果

索结果(因关键词影响,结果略有差别)。

其中每一个记录表示在 OMIM 中与查询信息相关的内容。另外,可以在“Display”中选择查询结果的显示方式、条目数。选择任意一条记录都包含了如下信息:MIM 号(ID)、查询疾病的名称(别名),与疾病相关遗传信息的一般性描述,有文献支持的临床表征,生化特征,发病机制,遗传性及诊断,文献支持的基因信息,分子遗传学、群体遗传学等文献支持材料。最后,提供了大部分的研究参考文献(图 14-6)。

选择页面上的 Gene map locus 后面的基因区段,会显示出该区段在染色体图中的详细信息(图 14-7)。

主要的内容包括如下几方面的图信息:基因序列信息、表型信息(包括数量性状位点)、OMIM 疾病记录、细胞遗传上的基因分布等详细信息。其中部分数据可以下载或查看。

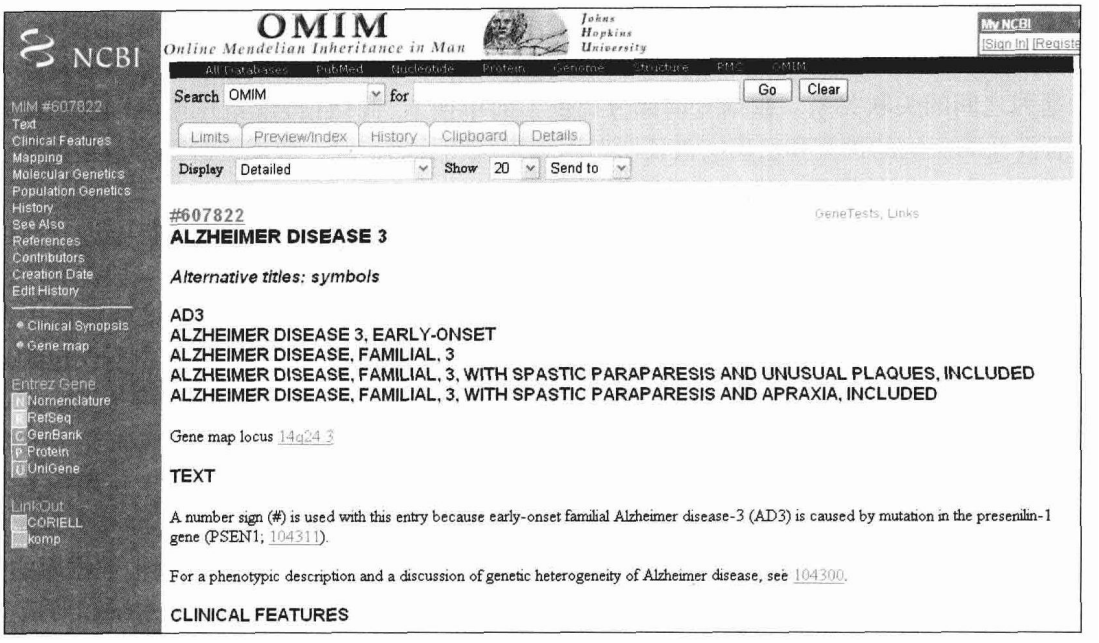


图 14-6 Alzheimer Disease 记录的详细信息

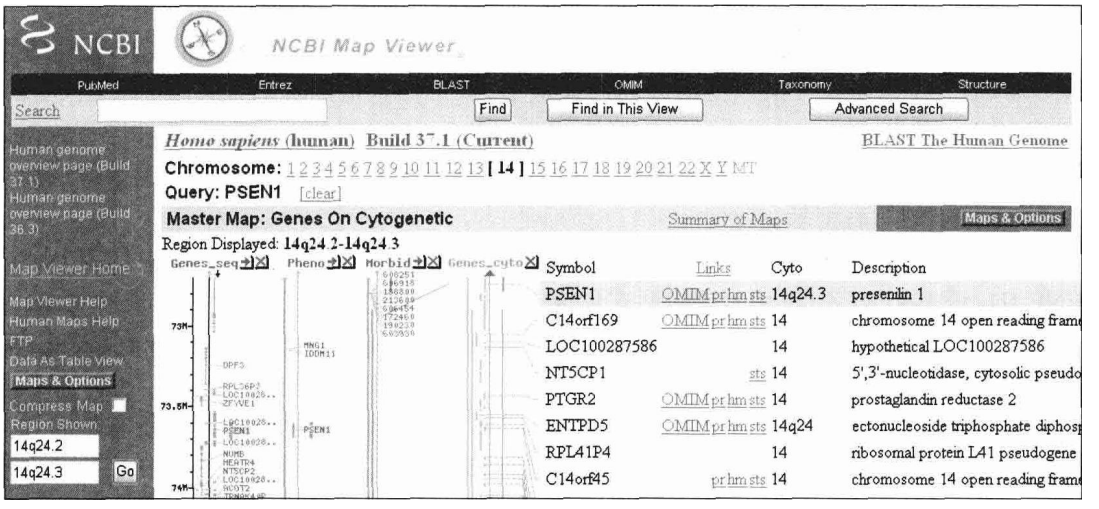


图 14-7 基因的主要图谱

二、遗传关联数据库

GAD 是由美国国立卫生研究院(National Institutes of Health, NIH)的 Kevin Becker 及其同伴们于 2004 年开发维护的遗传关联数据库(Genetic Association Database, GAD),该数据库中存储了大量的人类复杂疾病相关的基因及多态性信息,为研究人员从大量的多态性数据中快速地识别出疾病相关的多态提供了方便。数据库中的信息来源于对目前已有的关联分析结果的搜集和整理,这些信息是以基因为核心的,也就是说,数据库中的每条记录对应的是一个基因或者染色体位点,如果要研究某一特定疾病 6 个相关的基因,那么会在这个数据库中得到 6 条相应的记录。该数据库允许所有用户查看提交记录。

通过网址 http://geneticassociationdb.nih.gov/ 访问该数据库。用户可以在线查询某种特定遗传病相关的基因或某个基因相关的疾病的信息。也可以在免费注册后对整个数据库中的数据进行下载。截至目前,数据库中的记录数已经达到了 39 930 条。

GAD 数据库主要包含三部分功能(位于 GAD 主页左侧): 数据视图部分,数据查询部分,数据资源部分。数据视图部分主要是从疾病角度、基因角度、SNP 角度以及基因与环境互作的角度来查询疾病和基因之间的关联关系。数据查询部分提供了简单搜索、高级搜索、批量搜索以及通过基因来查看所有涉及的疾病的种类和已确实被证明基因和疾病相关的记录。数据资源部分包括提交疾病基因关联记录,对 GAD 数据库的意见以及数据下载等。

首先,可以通过数据库页面的左侧的相关链接选择不同的角度对数据表进行查询,GAD 会根据提交的查询从数据表中选择相应的字段返回结果页面,并且每条记录的第一个字段都有相应的详细的链接通过该链接,用户可以得到数据表中存储的与查询相关的全部信息。该数据库中存储的基因(多态)与疾病(表型)间的关系有一部分是通过关联分析得到的,因此数据表中不仅包含显著与疾病发生关联的基因的记录。同时也包含了关联关系不显著的记录,数据表中的字段 "Association? Y/N" 表明了具体的关系,该字段有三种取值: Y、N 和空,分别表示该记录的相应研究中的基因与疾病显著关联、不显著关联以及未明确是否关联。以疾病角度查询,可以得到特定疾病相关的基因 Symbol、染色体区段、基因组定位、对应的 OMIM ID、基因与疾病是否关联以及关联显著性水平的 p 值和相应参考文献的信息,另外 GAD 还给出了该疾病所属的疾病类信息以及与其他数据库的链接;以基因角度查询,可以得到基因相关的疾病表型描述、所属疾病类以及关联显著水平和对应参考文献的信息,同时还可以得到该基因在其他一级基因数据库中的 ID、名称、定位等基本信息以及与其他数据库的链接;另外,用户还可以从染色体角度出发,或者通过参考文献、环境因素等方面对数据表进行在线查询,当然,也可以选择 "All" 同时从多个角度对数据库中的相关信息进行查询。

其次,用户可以选择 "Simple Search",利用关键字实现对数据库中相关记录的简单查询。在 "Simple Search" 中,只需要提交以空格分隔的关键字,并选出查询内容的种类(Disease、Gene View、CH-SNP-HapMap 和 Reference)。还可以选择 "Advanced Search" 增加查询限定条件进行数据记录的高级搜索,包括更新时间、与疾病是否关联、疾病表型、疾病种类等。如果某些限定条件选择空白则会列出相关条件下的所有记录。

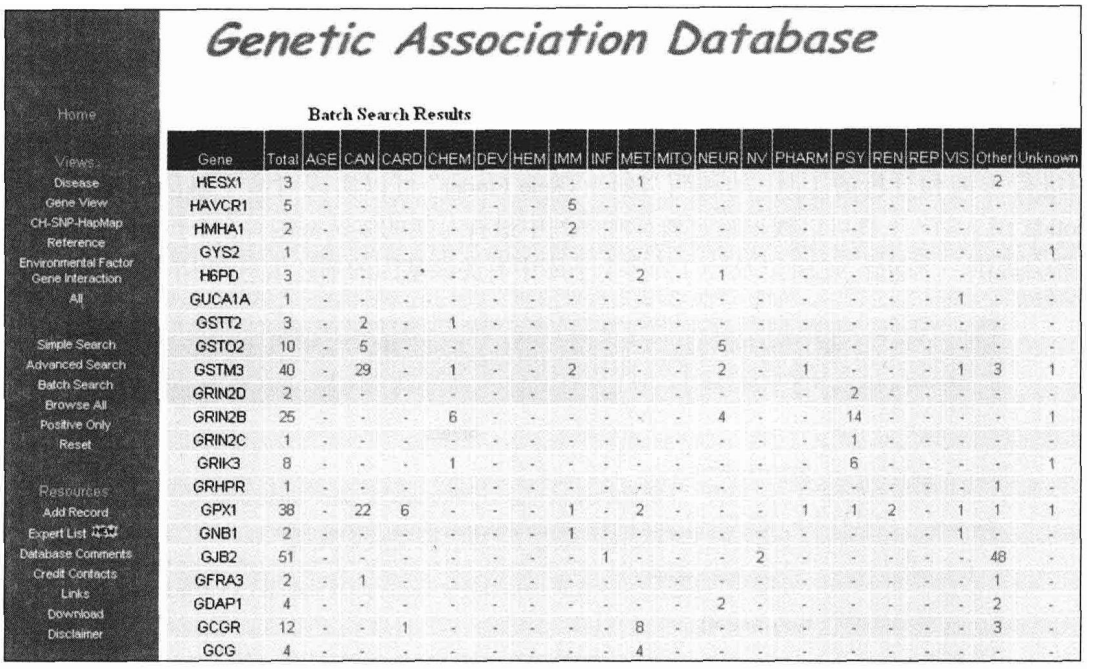
GAD 还支持对基因的批量查询,用户可以把小于 300 个基因以 HUGO 中的基因 Symbol, UNIGENE ID,或 ENTREZ GENE ID 的形式形成一个基因列表,并通过该列表实现对 GAD 中信息的批量查询。这样,GAD 就可以通过分析高通量实验(microarray、cDNA sequencing、SAGE 等)得到基因与人类疾病之间的关系。

选择 "Browser All" 链接可以得到图 14-8 中的结果,图中返回了数据库中的所有基因和与各类疾病间的关系,如第一条记录 HESX1 基因,它在数据库中共存在 3 条相关记录,其中与代谢类疾病(MET)相关的记录有 1 条,与其他类疾病相关的记录有 2 条。用户还可以选择 "Positive Only" 以筛选得到疾病与基因间存在显著关联的记录。另外,用户还可以通过 "Add Record" 页面实现向数据库中提交记录;通过 "Download" 页面实现对数据库中数据的下载。目前该数据库已经得到了研究人员的广泛应用。

三、癌症基因数据库

癌基因组解剖计划(Cancer Genome Anatomy Project, CGAP),是一项由美国癌症研究所(National Cancer Institute, NCI)于 1996 年发起并建立和主持的交叉学科计划。其目的在于产生用于解码肿瘤细胞的分子结构所需的信息,并创建一系列技术工具以挖掘与肿瘤相关的基因、蛋白质及其他的生物标记物,最终为癌症的研究提供信息资源和技术方法。CGAP 的总体目标是检测正常、癌前病变以及癌细胞的基因表达谱,使得研究人员可以借助于这些表达数据描述出肿瘤形成过程中的一系列细胞分子特征,最终改善对患者的检测、诊断和治疗。该计划通过与全世界范围内科学家的合作来增强其信息的科学性和完整性,为癌症相关科研人员提供方便。

CGAP 被分为五个部分,每一部分都有它自己的目的、信息学工具和资源。人类肿瘤基因索引



Gene	Total	AGE	CAN	CARD	CHEM	DEV	HEM	IMM	INF	MET	MITO	NEUR	NV	PHARM	PSY	REN	REP	VIS	Other	Unknown
HESX1	3	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	2	-
HAVCR1	5	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-	-
HMHA1	2	-	-	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-
GYS2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
H6PD	3	-	-	-	-	-	-	-	-	2	-	1	-	-	-	-	-	-	-	-
GUCA1A	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-
GSTT2	3	-	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GSTO2	10	-	5	-	-	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-
GSTM3	40	-	29	-	1	-	-	2	-	-	-	2	-	1	-	-	-	1	3	1
GRIN2D	2	-	-	-	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-
GRIN2B	25	-	-	-	6	-	-	-	-	-	-	4	-	-	14	-	-	-	-	1
GRIN2C	1	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
GRIK3	8	-	-	-	1	-	-	-	-	-	-	-	-	-	6	-	-	-	-	1
GRHPR	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
GPX1	38	-	22	6	-	-	-	1	-	2	-	-	-	1	1	2	-	1	1	1
GNB1	2	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	1	-	-
GJB2	51	-	-	-	-	-	-	-	1	-	-	-	2	-	-	-	-	-	48	-
GFRA3	2	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-
GDAP1	4	-	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	2	-
GCGR	12	-	-	1	-	-	-	-	-	8	-	-	-	-	-	-	-	-	3	-
GCG	4	-	-	-	-	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-

图 14-8 GAD 数据库中“Browser All”结果

(the human Tumor gene index, hTGI)指明了在人类肿瘤发生过程中的基因表达；分子表达谱(molecular profiling, MP)从分子水平分析人类组织样本的概念；癌症染色体变异计划(The Cancer Chromosome Aberration Project, CCAP)描述了同恶性转移相关的染色体变异；遗传注解索引(The Genetic Annotation Index, GAI)指明和描绘了同癌症相关的多态性；小鼠肿瘤基因索引(The Mouse Tumor Gene Index, mTGI)确定了在小鼠肿瘤发生过程中的基因表达。

用户可以通过网址 http://cgap.nci.nih.gov/ 对 CGAP 的网站进行访问(图 14-9)，并通过左侧导航栏 CGAP Info 中的相关链接了解更多有关该计划更为详细的信息。

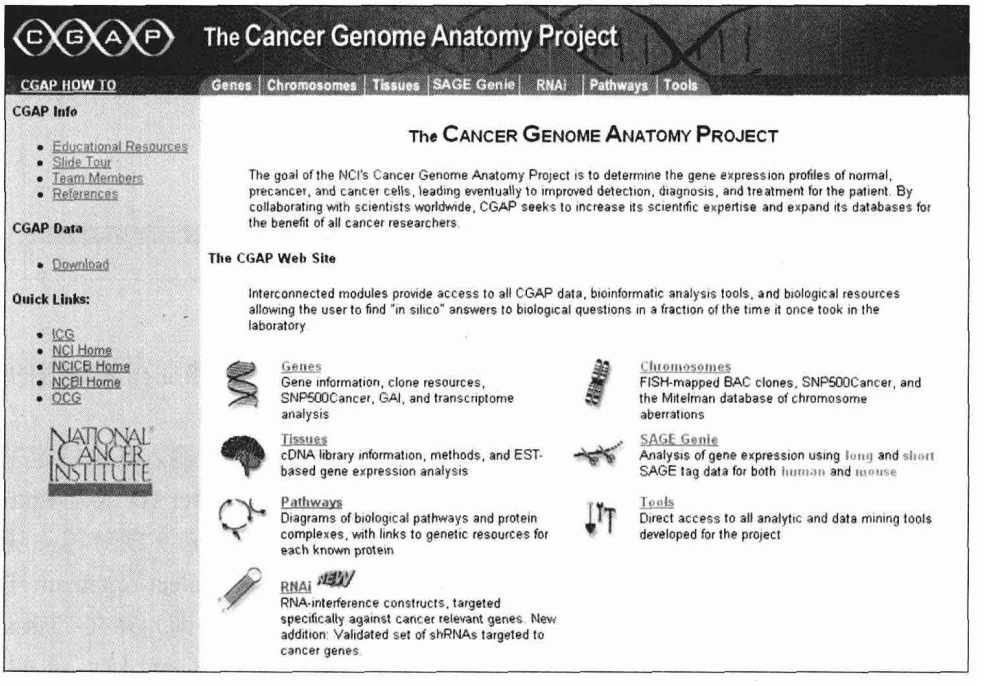


图 14-9 CGAP 数据库主页

该网站提供了七个相关模块用以对所有 CGAP 中包含的数据、生物信息学分析工具以及生物学相关资源的查询和获取，借助于这些模块用户可以实现对生物学问题的计算机模拟，从而快速地获得问题的解决方案。

进入“Genes”的标签页，可以得到图 14-10 所示的页面，该页面中提供了多种可用于对癌症相关基因进行查询和分析的工具，如利用“Batch Gene Finder”可以实现对多个基因的批量查询，利用“Nucleotide BLAST”工具可以找出给定核苷酸序列中最有可能的候选基因等，对于查询到的每个基因，CGAP都会提供一个包含NCBI以及NCI的多个子库中有关该基因的描述信息在内的“Gene Info”页面。

图 14-10 CGAP 数据库 Gene 标签页

下面以使用 Gene Finder 工具为例简要介绍如何在 CGAP 中实现对癌症相关基因的查询，并对查询结果进行简要解释。

Gene Finder 对应的标签页如图 14-11 所示，用户可以利用该工具通过输入某个特定基因的 Gene Symbol、GenBank 数据库中的 accession number、UniGene 数据库中的 cluster ID 或者 Entrez Gene ID 来查询基因的相关信息；也可以通过限定组织、功能、定位等方面的条件来实现对相关基因的查询。例如要查询与人类的结肠(colon)组织相关的基因，首先应在选择物种(Select organism)这一下拉列表中选择“Homo sapiens”(目前 CGAP 只支持对人类和小鼠两个物种的查询)，并在“Tissue Type”对应的下拉列表中选择“Colon”，提交查询后可返回一个包含所有结果的“Gene List”页面，对于感兴趣的基因，还可以通过页面中对应记录的最后一栏“Gene Info”链接去获取有关该基因的更为详细

的信息。对于结果列表中的第一个基因 *A1CF*，CGAP 中包含的有关该基因的全部信息如图 14-12 所示，其中包含 *A1CF* 在其他数据库中的 ID，名称，并提供其他数据库对该基因的描述链接，同时还包含了 *A1CF* 相关的序列、表达、细胞遗传学定位、染色体定位、对应蛋白、同源物以及相关的 GO 注释等多方面的信息(图 14-12)。

图 14-11 CGAP 数据库 Gene Finder 检索界面

CGAP 中还包含有染色体(chromosomes)、组织(tissues)、SAGE 精灵(SAGE Genie)、通路(pathways)、工具(tools)和 RNA 干扰(RNAi)六个模块。与 Genes 模块类似，每个模块都提供了很多相关的查询分析工具，可支持对 CGAP 中包含的染色体畸变、表达数据、蛋白质复合物、生物学通路等信息在内的多方面内容进行搜索，并可以根据查询得到的结果做进一步更深入的分析研究。特别是 RNAi 模块，其中收录了靶向癌症相关基因的 RNA 干扰，并包含有已经证实的靶向癌基因的短发卡 RNA (short hairpin RNA，shRNA)。

另外，CGAP 还允许用户对其中的数据资源进行下载，在 CGAP 主页左侧导航栏中包含有 CGAP Data 项，该项中的内容 Download 就是数据下载页面的链接，包含了人和小鼠两个物种的基因注释、基因表达以及相关的一些文库中的数据。

CGAP 计划还有另外一个目标，就是建立一套完整的基因及其变异目录，这些目录不仅有利于评价癌症的危险程度，而且可以根据遗传变异确定预防或治疗策略，最终根据分子特征达到治疗的目的。目前 CGAP 建立的注释基因索引包括利用表达序列标签(expressed sequence tags，EST)及基因注释等途径建立的人和小鼠的肿瘤基因索引和用于区分鉴定与肿瘤有关的基因的遗传变异的注释索引。CGAP 还建立了许多 cDNA 文库，不仅包括有全瘤组织文库，也包括癌症发展过程中不同阶段的细胞 cDNA 文库。同时 CGAP 也提供了诸多资源如克隆、BAC 及技术方法和检索工具等，为肿瘤研究提供了一个多学科的综合平台。

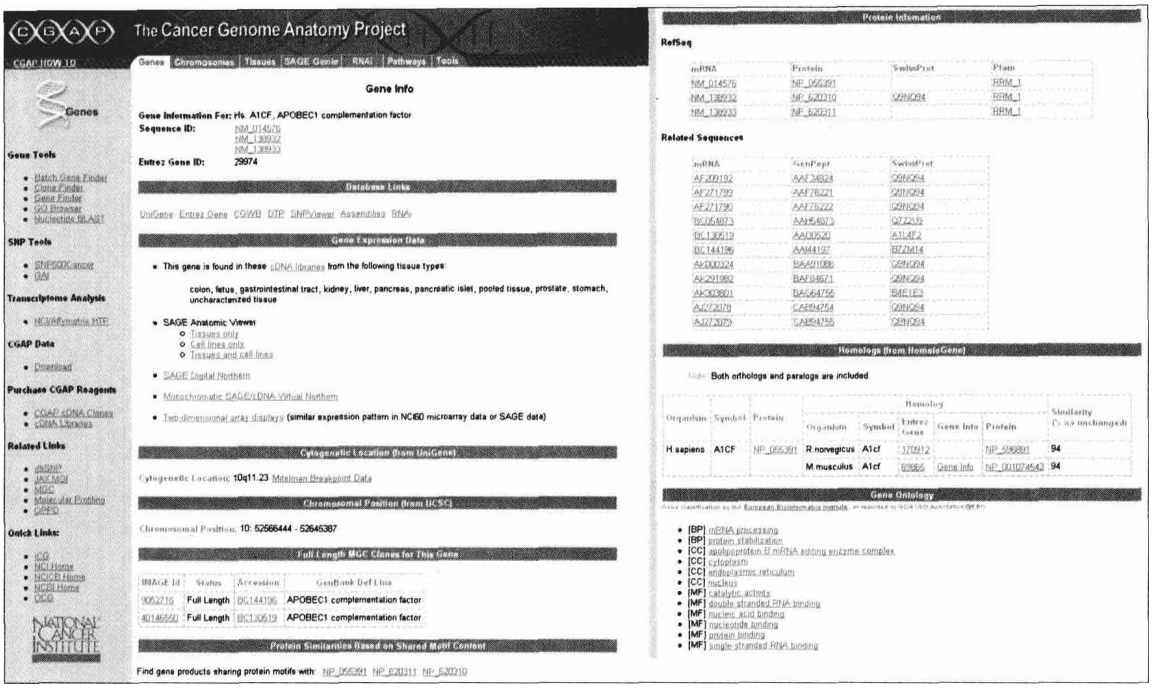


图 14-12 CGAP 数据库 Gene Finder 检索结果

CGAP 蕴涵了大量有用的信息，目前，已有许多科研工作者成功地利用这些信息实现了对肿瘤的研究，如 Loging 等用数据库及快速表达筛选方法，通过 CGAP 鉴定胶质瘤潜在的肿瘤标志和肿瘤抗原，获得了有意义的结果。

四、WHO 规范的疾病分类标准

国际疾病分类(international classification of diseases，ICD)是目前国际上共同使用的统一的疾病分类方法，是国际标准分类，它由世界卫生组织疾病分类合作中心负责进行国际疾病分类的修订、推广和应用工作。ICD 的目的是对不同国家或地区在不同时间收集到的死亡和疾病数据进行系统地记录、分析、解释和比较，其中包括对各人群组一般健康状况的分析，疾病发病和患病的监测以及与其有关的其他健康问题。ICD 把疾病诊断和其他健康问题的词句转换成字母数字编码，从而易于对数据进行贮存、检索和分析。国际疾病分类是国际标准，是各国进行卫生信息交流的基础，世界卫生组织每年出版一本《世界卫生年鉴》就是以它为标准收列了各国的死亡原因的统计资料。

ICD 是针对所有常见流行病的国际上标准的诊断学分类，被许多临床及卫生管理部门广为采用。该体系囊括了群体健康状况分析、疾病发病率及流行性监测等多方面内容，常被用于分类疾病及与其他许多类型的生命活动相关的健康问题。该体系除了可以用于存储和获取有关临床、流行病等的诊断信息外，还为疾病的发病率和死亡率统计提供了基础。

早在 18 世纪初，人们就试图系统地对疾病进行分类，并建立了多种分类体系，但更具实际意义的，如疾病的发病率和死亡率等统计学研究还是在 19 世纪才开始的。ICD 最初也正是为了这一目的而产生的，该分类的前身是 1893 年国际统计学会上被采纳的国际死因列表(International List of Causes of Death)，在经历了一系列修改后形成了现在的 WHO 规范的国际疾病分类体系。

ICD-10 是国际疾病分类体系的第十个版本，于 1990 年 5 月召开的第四十三届国际卫生会议上审核通过并从 1994 年开始在世界卫生组织会员国生效，是疾病分类体系的最新版本，ICD 在第六版中首次加入了发病原因的信息，并于 1948 年被当时刚刚建立的 WHO 接管和发布。

ICD-10 的书名由过去的《国际疾病分类》改变为《国际疾病与相关健康问题统计分类》，为保持分类的连续性，其简称仍使用 ICD。ICD-10 首次引用了字母编目，由原来的“纯数字编码”改为 26

个英文字母加数字编码形式的“字母和数字的混合编码”。编码的第一位是一个字母，每一个字母都与特定的章节有关，只有字母 D 和 H 除外，字母 D 同时用于第二章肿瘤和第三章血液及造血器官疾病和某些涉及免疫机制的疾患，而字母 H 同时用于第七章眼和附器疾病及第八章耳和乳突疾病，第二十章在编码时第一个位置有四个不同的字母 V、W、X、Y，未使用的“U”编码，WHO 建议 U00~U49 用于新发现疾病或病因不明疾病，U50~U99 用于特殊的临床研究。

国际疾病与相关健康问题统计分类(International Statistical Classification of Diseases and Related Health Problems)是世界卫生组织依据疾病的某些特征，按照规则将疾病分门别类，并用编码的方法来表示的系统。其中提供疾病的分类信息以及相关的症状、异常和引起疾病的社会环境因素等描述信息。国际疾病分类体系广泛应用于对发病率和死亡率的统计，为医药领域的诊断决策提供支持。

ICD-10 共分三卷，第一卷是疾病和有关健康问题的国际统计分类主要包括 ICD-10 全部 3 位数或 4 位数编码内容及其必要的注释和说明；第二卷是 ICD-10 指导手册，用于指导用户如何正确使用 ICD-10 的第一卷和第三卷，并对使用中需要遵循的各项规则和有关问题给予详细的介绍；第三卷是 ICD-10 字母索引，主要包括在查找疾病、损伤、中毒的临床表现和外部原因时详细的内容和编码。

ICD 在编制和使用中都以首先满足统计需要为前提，但为了适应各个医学领域对疾病分类的需求，ICD 采取了许多切实可行的措施，使其更加丰富和灵活，使各个医学领域在使用 ICD 的过程中都能找到适合的方法以解决本领域的特殊问题。

ICD 在分类轴心上强调“以病因为主、解剖部位和其他为辅”的原则，采用 3 或 4 位数的“字母数字编码”形式，即第 1 位为英文字母，第 2 至 4 位为阿拉伯数字，从“A00~Z99”对所有的疾病(包括损伤和中毒及其外部原因、与保健机构接触的理由)归成 21 大类疾病，再逐渐细分成小类、节、3 或 4 位数的详细内容。ICD-10 在分类结构上充分运用有限的位数来突出严重危害健康的疾病和情况，同时采用尾部编码开放式的技巧用于包括其他所有的疾病和情况(表 14-2)。ICD-10 在编码使用上只对前 4 位数有统一要求，对以后的扩展位数及编码排列没有限制，从而既可保证世界各地汇总资料的一致性，也允许各个领域或局部根据自己的实际情况自由发展。ICD-10 还继续沿用了 ICD-9 中各种形式的编码系统和特定意义的符号以满足卫生统计、预防医学、基础医学以及临床医教研各方面的需要。

表 14-2 ICD-10 疾病分类编码表

Chapter	Blocks	Title
Ⅰ	A00-B99	Certain infectious and parasitic diseases
Ⅱ	C00-D48	Neoplasms
Ⅲ	D50-D89	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
Ⅳ	E00-E90	Endocrine, nutritional and metabolic diseases
Ⅴ	F00-F99	Mental and behavioural disorders
Ⅵ	G00-G99	Diseases of the nervous system
Ⅶ	H00-H59	Diseases of the eye and adnexa
Ⅷ	H60-H95	Diseases of the ear and mastoid process
Ⅸ	I00-I99	Diseases of the circulatory system
Ⅹ	J00-J99	Diseases of the respiratory system
Ⅺ	K00-K93	Diseases of the digestive system
Ⅻ	L00-L99	Diseases of the skin and subcutaneous tissue
ⅩⅢ	M00-M99	Diseases of the musculoskeletal system and connective tissue
ⅩⅣ	N00-N99	Diseases of the genitourinary system
ⅩⅤ	O00-O99	Pregnancy, childbirth and the puerperium
ⅩⅥ	P00-P96	Certain conditions originating in the perinatal period

续表

Chapter	Blocks	Title
XVII	Q00-Q99	Congenital malformations, deformations and chromosomal abnormalities
XVIII	R00-R99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	S00-T98	Injury, poisoning and certain other consequences of external causes
XX	V01-Y98	External causes of morbidity and mortality
XXI	Z00-Z99	Factors influencing health status and contact with health services
XXII	U00-U99	Codes for special purposes

五、疾病本体论

疾病本体论(disease ontology, DO)最初是2003年在Northwestern大学启动的Nugene计划的一部分,其目标是提供一个与人类疾病相关的整合生物医学数据集的开源Ontology体系,以促进各种疾病及相关健康状况向特定医学代码的映射。DO期望构建成一个具有正确的Ontology体系结构并且在语义上可计算的结构形式,其中的每个节点都使用标准参考集给出明确的界定,每个节点都对应着诸如SNOMED、ICD-9和ICD-10、MeSH(Medical Subject Headings)以及UMLS (Universal Medical Language System)等目前常用的包含疾病及疾病相关概念的体系中的术语。DO是将不同数据库通过疾病概念整合到一起的开源疾病体系。这些数据库包括:MeSh、UMLS、ICD、Systematized Nomenclature of Human and Veterinary Medicine-Clinical Term,如图14-13。

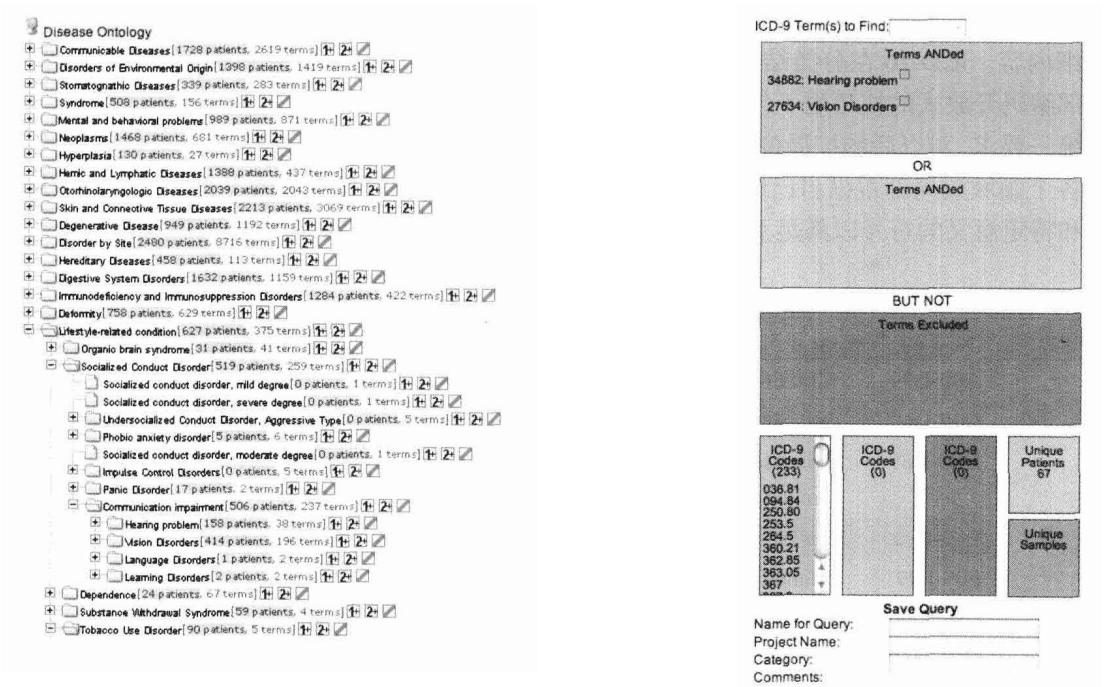


图14-13 Disease Ontology体系

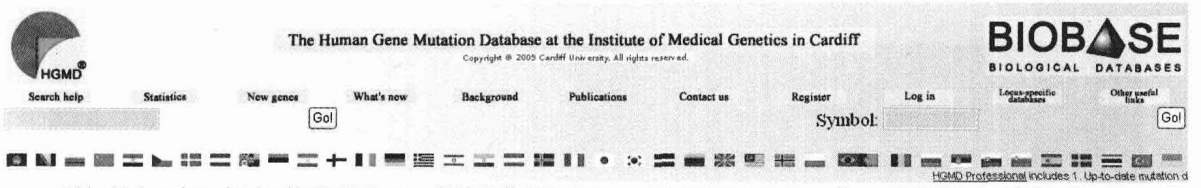
DO发展至今已经经历了3个版本,其中DO_V1是基于ICD编码的疾病体系;DO_V2的体系主要基于SNOMED, MeSH和UMLS;DO_V3在前一版本的基础上进行了修改,加入了更多的临床信息,如变异位点、环境、传染源和发病过程等等。DO是开源的,网址是:http://diseaseontology.sourceforge.net/。

六、其他疾病数据库

除了前几节介绍的数据库以外,还有许多复杂疾病相关的数据库,如人类基因突变数据库(Human Gene Mutation Database, HGMD), GeneCards等。

HGMD是由卡迪夫大学医学遗传研究所(Institute of Medical Genetics, Cardiff University)的Cooper等开发和维护的,该数据库存储了大量的人类遗传病相关的基因突变数据,截止到2009年10月15日,数据库中已经包含了发生在3526个基因上的93 347个不同的突变,并且记录数还在以每年超过9000条的速率快速增长。尽管其最初的建立目的仅仅是为了研究人类基因的突变机制,目前HGMD已经在众多领域得到了广泛应用,其用户范围已经遍及科研人员、临床医生等学术工作者以及生物制药、生物信息学等商业公司。目前,该数据库允许学术研究等非盈利性组织免费注册并从中获得数据,而商业用户则必须购买授权后才可以使用。

数据库中存储了多种类型的突变数据,具体的突变类型及相应类型的记录见图14-14。HGMD中不仅存储了疾病导致的突变,还存储了大量的与疾病表型相关联的多态性DNA序列变异以及相关的临床表型并不明确但具有明确功能的突变,这些突变对于研究不同个体间的疾病易感性差异是



This database is maintained by D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, K. Howells and M.E. Mort with the assistance of N.S.T. Thomas.

*Please note that this less up-to-date public version of our database is freely available only to registered users from academic institutions/non-profit organisations. All commercial users are required to purchase a license from BIOBASE, our commercial partner. A license to HGMD Professional is available to both commercial and academic/non-profit users wishing to access the most up-to-date version of the database (see example HGMD Professional entry). Read more about how HGMD is funded.

Table:	Description:	Public entries: This site. Academic/non-profit users only	Total entries: HGMD Professional 2009.3
	Mutation totals (as of 2009-11-26)	**67022**	**93347**
Gene symbol	The gene description, gene symbol (as recommended by the HUGO Nomenclature Committee) and chromosomal location is recorded for each gene. In cases where a gene symbol has not yet been made official, a provisional symbol has been adopted which is denoted by lower-case letters.	2473	3526
cDNA sequence	cDNA sequences are presented numbered by codon.		
Missense/nonsense	Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet.	38275	52591
Splicing	Mutations with consequences for mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions given as positive integers refer to a 3' (downstream) location, negative integers refer to a 5' (upstream) location.	6395	9015
Regulatory	Substitutions causing regulatory abnormalities are logged in with thirty nucleotides flanking the site of the mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiaton codon, polyadenylation site or termination codon is given.	912	1610
Small deletions	Micro-deletions (20 bp or less) are presented in terms of the deleted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	10935	14701
Small insertions	Micro-insertions (20 bp or less) are presented in terms of the inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	4403	6049
Small indels	Micro-indels (20 bp or less) are presented in terms of the deleted/inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	982	1371
Gross deletions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	3772	5705
Gross insertions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	679	1154
Complex rearrangements	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	511	870
Repeat variations	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	158	281

1499509 hits on this page since June 2005.

图14-14 HGMD的主页和内容

十分有价值的。HGMD 中存储的突变及多态性数据主要是通过整合现有突变数据库中所包含数据以及文本挖掘的方式获得的，这些数据经过了严格的筛选，是具有较高可信度的非冗余数据，但该数据库中只存储了与人类遗传病相关的核基因突变数据，并不包含体细胞突变和线粒体基因突变。

HGMD 可被应用于许多方面的研究，如通过对数据库中存储的突变数据的搜索来确定基因突变在以前的研究中是否被发现过；通过对数据库的查询来获得某个特定基因的所有已知突变谱；另外，该数据库还可以用于查询发生在特定位置上特定类型突变的例证，以作为研究发现的某一特定变异的病理学真实性的支持依据。

该数据库还以提供包括 OMIM、Entrez Gene 和 Human Gene Nomenclature Committee 在内的大量网络资源外部链接的形式，整合了表型、结构以及定位等多方面的信息。

GeneCards（图 14-15）建立于 1997 年，由以色列 Weizmann 科学研究所（Weizmann institute of science in israel）的 Crown 人类基因组中心（Crown human genome center）开发和维护，是一个整合了

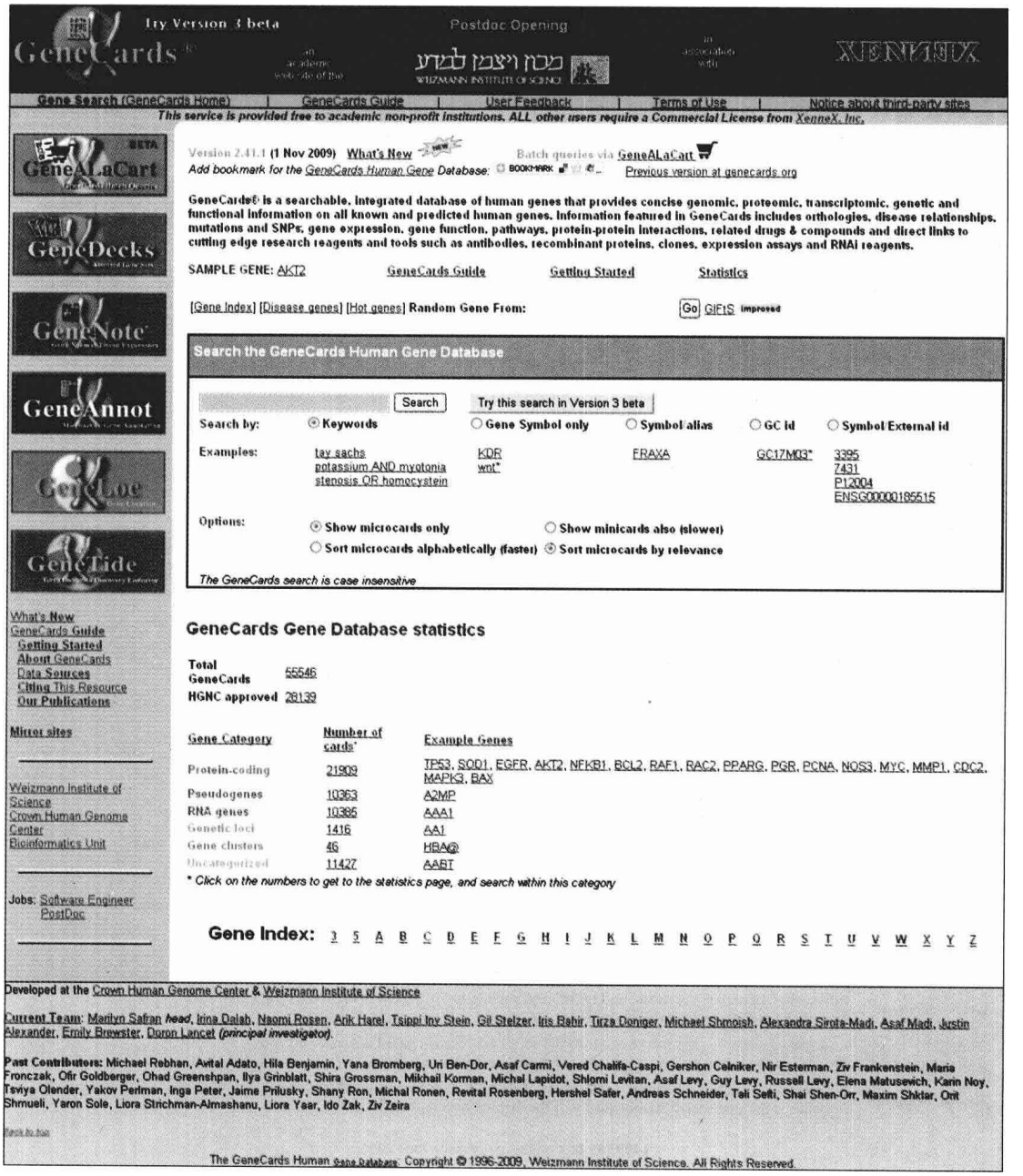


图 14-15 GeneCards 主页

人类已知及预测基因的基因组、蛋白质组、转录组、遗传及功能等多方面信息的综合性数据库。数据库中包含了同源、疾病、突变及SNP、基因表达、基因功能、通路、蛋白质互作、相关药物及化合物等丰富的内容。目前，该数据库已经发展到了2.41版本(更新时间为2009年7月19日)，记录数已经达到了55 546条(其中有28 139条可以在HGNC中找到相应的Symbol)，可以通过http://bioinfo.weizmann.ac.il/cards/对该数据库进行访问。

GeneCards搜集并整合了包括OMIM、GAD、CGAP、HGMD、GenBank、Ensembl、EntrezGene、HGNC、UniGene、SwissProt、dbSNP、GO在内的许多数据库中的数据。该数据库侧重于信息的全面性，因此相对于其他疾病数据库而言，GeneCards对人类疾病的具体描述相对较少，但它提供了更为全面的功能基因组数据及相关数据库的外部链接。另外，该数据库不仅支持简单搜索，还可以实现数据的模糊查询、由多个单词构成的字符串的查询以及高级搜索等功能，可以通过对GeneCards的查询来获取疾病相关基因的染色体定位、表达数据、同源基因、对应的蛋白质产物等众多信息。

其他疾病相关数据库还有：由癌症导致突变的基因、原癌基因、抑癌基因等相关信息的肿瘤基因数据库(The Tumor Gene Database，TGDB)，网址是http://www.tumor-gene.org/TGDB/tgdb.html；存储乳腺癌相关基因信息的乳腺癌基因数据库(The Breast Cancer Gene Database，BCGD)，访问网址为http://www.bcm.edu/test-bcgd/；包含群体发病率、基因与疾病的关联关系、基因与基因及基因与环境间互作等信息的人类基因组流行病学导航(The Human Genome Epidemiology Navigator，HuGE Navigator)；肿瘤及血液病相关的遗传学和细胞遗传学数据库(ATLAS of Genetics and Cytogenetics in Oncology and Haematology，http://atlasgeneticsoncology.org/)等。

第四节 疾病网络重构和计算系统生物学方法

Section 4 Complex Disease Network Reconstruction and Computational Systems Biology Methods

一、计算系统生物学概述

随着人类基因组计划(Human Genome Project，HGP)的完成，生物信息学的蓬勃发展，为从分子水平和系统观念来研究复杂疾病，以及研究模式从“序列→结构→功能”向“互作→网路→功能”的转变提供了契机。人类基因组计划的发起人、系统生物学创始人之一的美国科学家Leroy Hood提出：经典分子生物学是一种“垂直型”学科，既采用多种手段来研究个别基因和蛋白质，在DNA水平上寻找特定基因，通过基因突变、基因敲除的手段研究基因的功能，又在基因功能的基础上研究了蛋白质空间构象，蛋白质修饰以及蛋白质之间的互作等；而基因组学，蛋白质组学和其他各种“组学”都是“水平型”研究，系统生物学的特点就是要把“垂直型”研究和“水平型”研究相结合，形成一种“三维”立体式的研究。由于复杂疾病涉及了众多内在和外在的因素，所以从整体上考察疾病涉及的基因和蛋白质并结合转录调控、代谢通路等多层面信息就可能会揭示复杂疾病的发病规律。人体本身就是一个庞大复杂的网络体系，从DNA、RNA、蛋白质到信号转导、转录调控、细胞功能再到组织、疾病、表型，每一水平上都是一个复杂的生物网络，而不同水平上相互串联和交互构成了更加复杂的一个网络调控系统。每一个层面被称作一种组学(omics)，例如，蛋白质组、转录调控组、表型组等。而系统生物学是整合不同层次的组学信息来研究和理解生物系统是如何行使功能的。换句话说，系统生物学是对生物体整个生命过程做全面性的定量研究，并非以生物体的某一部分为研究对象。其目的是要建立模式并以实验来证实可预测的生物体的表现。

计算系统生物学的基本工作流程有四个步骤。首先是对选定的某一生物系统的所有组分进行了解和确定，描绘出该系统的结构，包括基因相互作用网络和代谢途径，以及细胞内和细胞间的作用机制，以此构造出一个初步的系统模型。第二步是系统地改变被研究对象的内部组成成分(如基因突

变)或外部生长条件,然后观测在这些情况下系统组分或结构所发生的相应变化,包括基因表达、蛋白质表达和相互作用、代谢途径等的变化,并把得到的有关信息进行整合。第三步是把通过实验得到的数据与根据模型预测的结果进行比较,并对初始模型进行修订。第四步是根据修正后的模型的预测或假设,设定和实施新的改变系统状态的实验,重复第二步和第三步不断地通过实验数据对模型进行修订和精练。系统生物学的目标就是要得到一个理想的模型,使其理论预测能够反映出生物系统的真实性。

在计算系统生物学中,信息(information)是计算系统生物学的基础。信息科学的研究方法是计算系统生物学的基础。首先生物学可以形象的用数字化的模式来表示,同时,无论是编码蛋白质的基因还是控制基因行为的调控网络,都可以用数字化的形式来表示。再者,生物学上的信息是有等级和方向次序的,一般而言都是以 DNA → RNA →蛋白质→蛋白质互作→细胞→器官→个体→群体这种方式进行信息的传递,而计算系统生物学的重要任务就是要尽可能多的获得每个层次的信息,并将其整合。

在计算系统生物学中,整合(integration)是计算系统生物学的灵魂。计算系统生物学与各种组学的不同之处在于它是一门整合型的大学科。首先,它要整合系统内部不同性质的构成要素(基因、mRNA、蛋白质、生物小分子及大分子等)。其次,计算系统生物学要实现从基因到细胞,到组织、再到个体各层面信息的整合。同时,计算系统生物学的研究又要实现研究思路和方法的整合,即将“水平型”研究和“垂直型”研究相结合,形成“三维”立体式研究。

整合多层面信息,构建二部网络甚至多部网络是目前利用计算系统生物学方法研究复杂疾病的重要方式。目前,大部分的研究工作结合两个层面的信息,如结合疾病和基因、疾病和通路、疾病和SNP、疾病和 miRNA、药物和靶蛋白、SNP 和基因表达等,构建整合两层面信息的二部网络,分析二部网络或重构网络的特点,从而对复杂疾病过程中的某些规律进行整体研究和分析。

二、Disease-Gene 网络重构分析

人们不仅能够从 OMIM、GAD 和 CGAP 等数据库中获取疾病相关的信息,这些数据库同时提供了疾病和相关的疾病基因之间的联系。这些关系有的是通过文献获得的,有的是根据关联分析结果推测的,有的是通过生物学实验证实的。复杂疾病的研究表明,某些疾病的病因或分子病理基础不同,同种疾病可能和不同的基因或基因集合有关;另外,由于基因多效性等因素,同一个或同一组基因在不同的组织或时期可能完成不同的生物学功能,它们的改变将导致或参与不同的疾病过程。基因和疾病之间存在复杂的多向对应关系,给人们研究复杂疾病的分子过程带来了巨大的挑战。OMIM、GAD 和 CGAP 等数据库为研究复杂疾病提供了重要的资源和线索,从这些数据资源中,人们不仅能够获得与某一种疾病有关的多个基因,同时也可以得到某一个基因参与的多个疾病。如果把这些疾病和基因连接起来,这就自然地形成了一个巨大的疾病 - 基因二部网络。Goh 等在 2007 年 PNAS 上发表了一篇题为“The human disease network”的研究报告,这项工作可以认为是通过疾病和基因关系研究复杂疾病的奠基文章,他们利用 OMIM 数据库存储的疾病和基因之间的关系,分别构建了疾病网络和疾病基因网络(图 14-16)。在疾病网络中,疾病通过是否共享相同的基因连接起来;同样,在疾病基因网络中,基因通过是否共享相同的疾病连接起来。网络中的疾病被划分为 22 个不同的疾病类,发现同一疾病类中的疾病在网络中倾向于聚集成团。这项研究从整体的角度利用网络分析方法研究了不同疾病之间在基因层面的相似性。随后《Genome Biology》、《Bioinformatics》、《Molecular Systems Biology》等杂志上也发表了多篇利用网络分析方法研究复杂疾病的文章,通过构建疾病和疾病基因之间的网络关系来识别复杂疾病潜在候选基因识别、预测疾病基因等。有些研究还结合了基因功能等层面的信息,对于疾病和基因的关系也不局限于数据库的信息,研究者利用信息整合的策略,从疾病数据库、实验、文献挖掘等多方面完善和丰富了疾病与基因之间的关系。

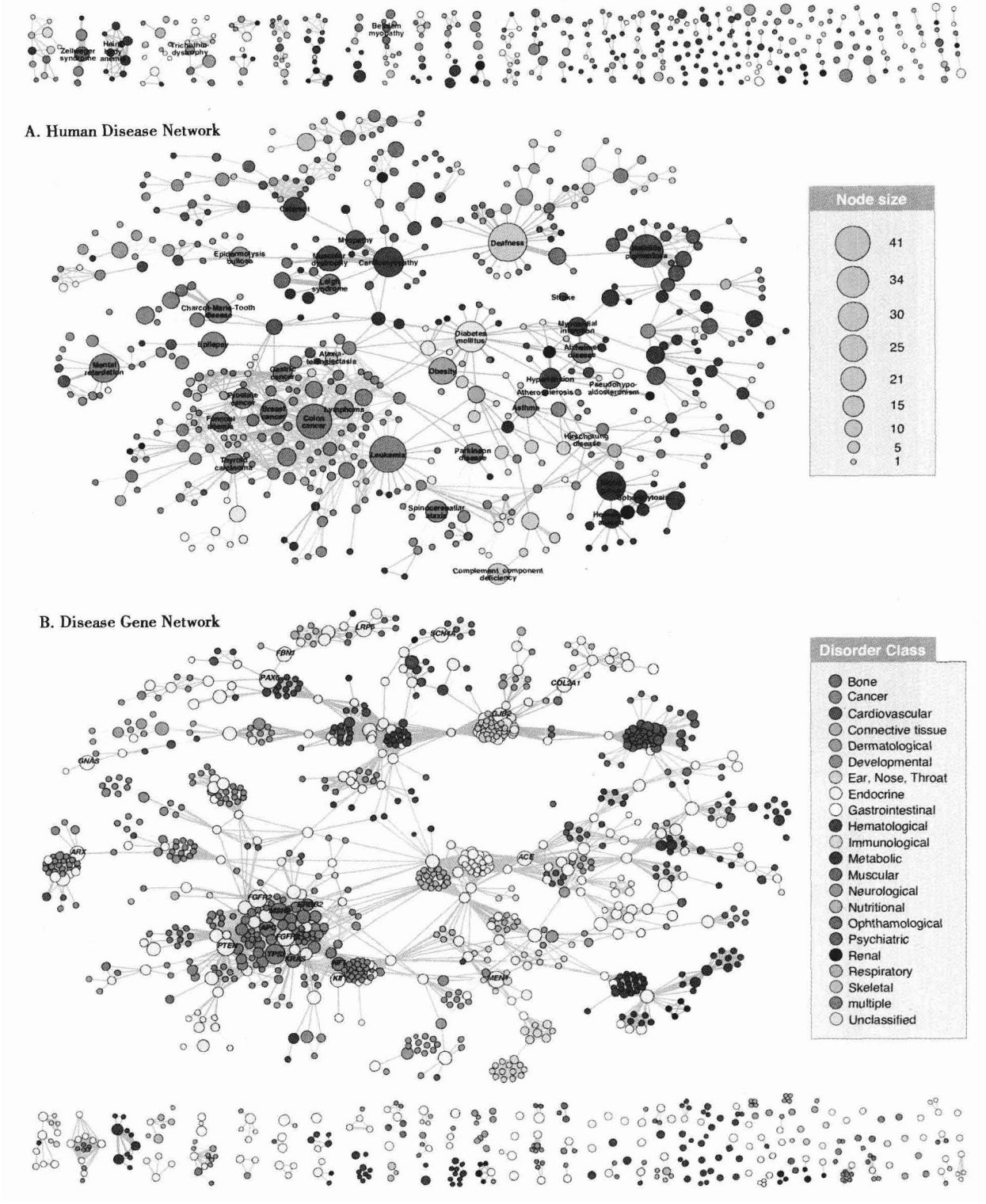


图 14-16 Human Disease Network

除了从数据库、文献等资源获得疾病和基因关系之外，随着高通量检测技术的发展，尤其是DNA 微阵列(DNA microarray)技术在各种疾病研究中的广泛应用，对于各种常见疾病，尤其是肿瘤研究，研究者检测了大量疾病的基因表达谱。这些数据资源存储在 GEO、SMD 等表达谱数据库中，且存储了大量疾病相关基因的信息，研究者为了寻找疾病相关的基因，开发了许多基于基因表达谱的疾病基因识别方法(见第七章)。利用这些数据和方法，研究人员可以方便地获得和某一种

疾病相关的基因集合。Xuebing 等利用他们提出的 CIPHER 算法构建包含 1126 种表型和 8919 个基因的网络，结合表型分类信息和基因功能聚类信息，从表型和基因的关联矩阵中(图 14-17)挖掘出了同一类疾病与功能基因簇之间的关系，为复杂疾病公共机制和公共通路研究提供了新的线索和思路。

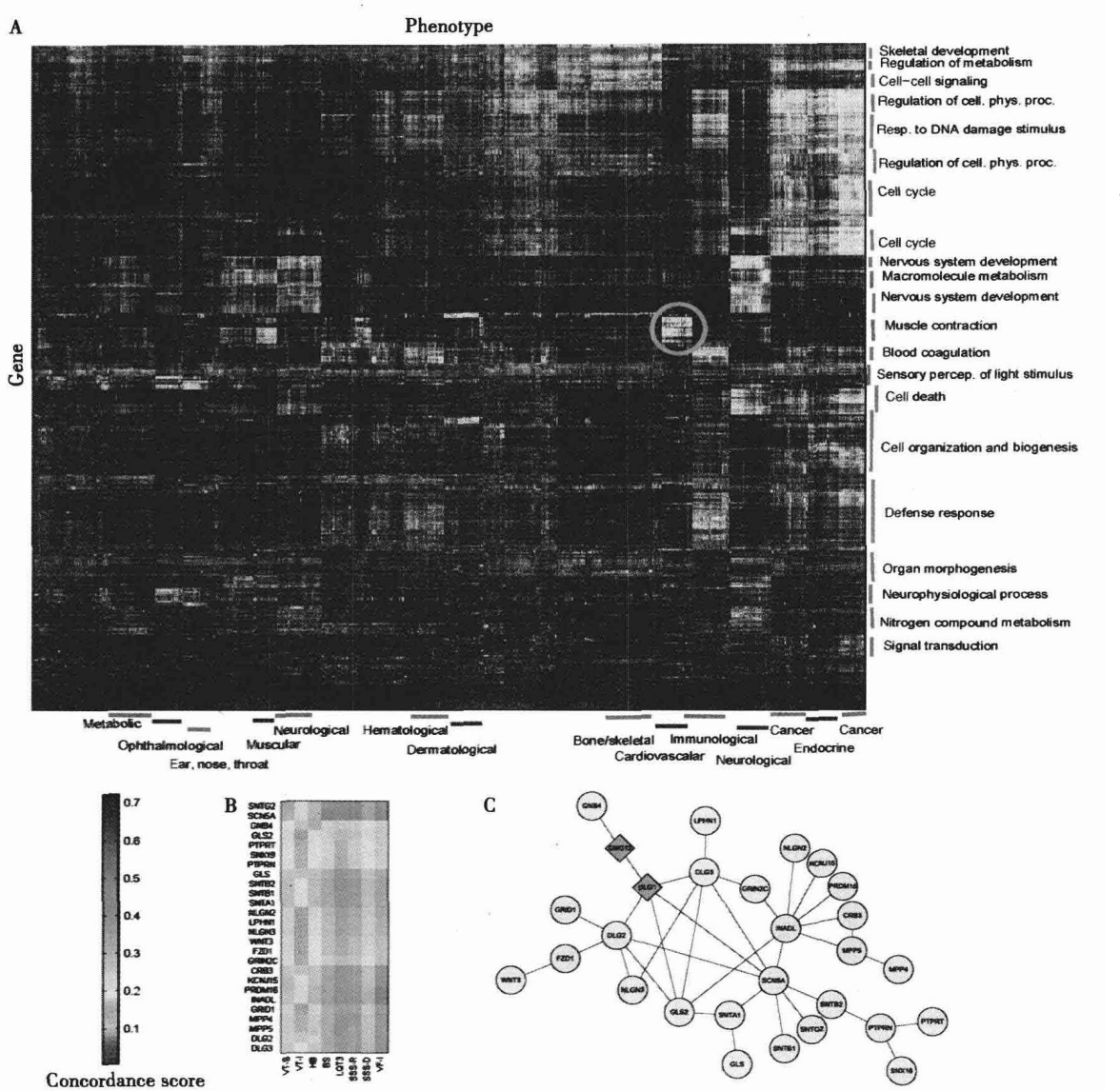


图 14-17 基因 - 表型聚类模块及基因 - 表型网络

三、Disease-Pathway 网络重构分析

代谢系统疾病是一类特殊的疾病，这类疾病主要是由参与各种代谢通路的酶或编码酶的基因的功能异常导致的。通常情况下，如果代谢通路中的一些重要酶发生异常，就会影响整个通路的功能。或者说，通路中的不同的酶或编码酶的基因的异常变化，都会导致整条通路异常，从而引发该种代谢疾病。因此，仅仅根据疾病相关的基因相同来分析几种复杂疾病分子基础的方法会忽略掉这些信息。Lee 等结合 KEGG 和 BIGG 通路数据库的信息，根据基因在同一代谢通路中的邻近关系重新构造了人类代谢相关疾病网络，同时还利用了 1990 年到 1993 年的美国人群流行病学调查数据，研究了网络中聚集成团的疾病在流行性、致死性以及发生率等方面的特点(图 14-18)。

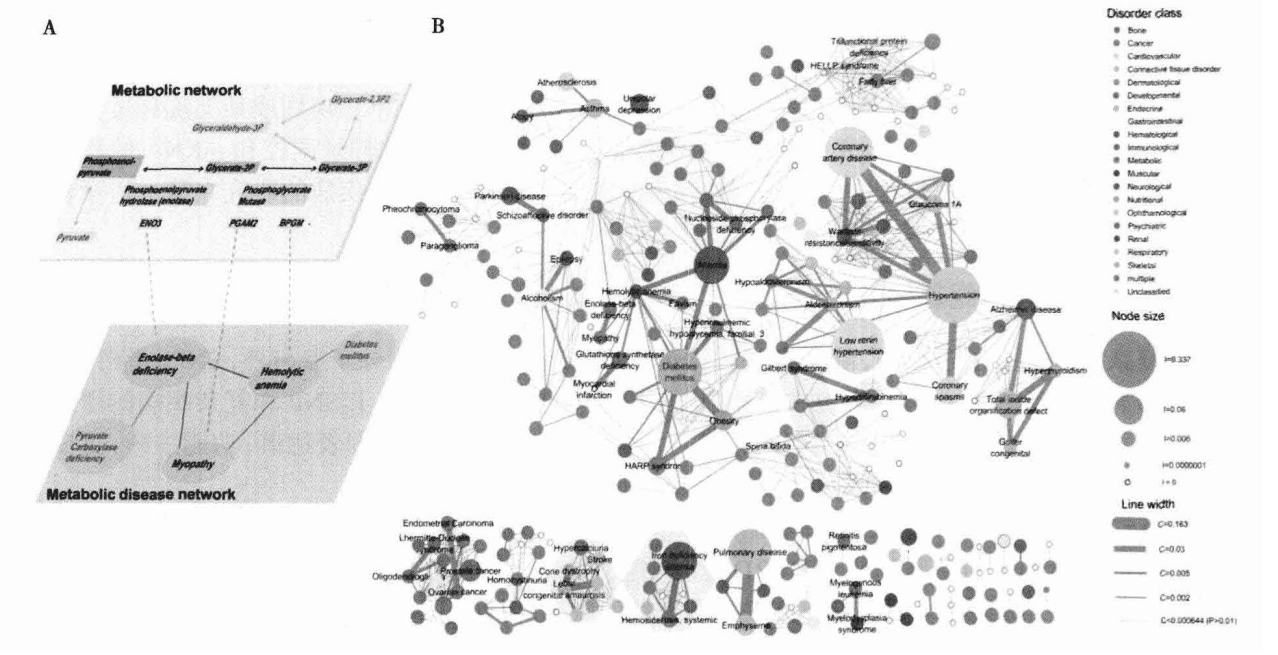


图 14-18 代谢疾病网络

把代谢通路数据库的信息引入到复杂疾病研究中的这一思想为人们提供了一个新的研究思路。不仅仅是代谢系统疾病，内分泌系统疾病、神经系统疾病等也可以采取这一策略，在疾病研究过程中引入相关的信息，如神经体液调节通路、信号转导通路等。多层面相关信息的引入，不仅使疾病研究更具体、信息更全面，也有助于对疾病的认识的不断深入和系统化。当然，如何合理有效地使用丰富的信息资源也给计算系统生物学带来了新的机遇和挑战。

四、Disease-miRNA 网络重构分析

对于 microRNAs(miRNA)的研究源自 20 世纪末第一次在线虫(C·elegans)中的发现(见第十六章)。这类约 22nt 长的小分子一开始并没有引起足够的重视。直到 2000 年第 2 个 miRNA——let-7 及其人类和果蝇中同源物的发现改变了人们的看法，miRNA 可能是一类进化上保守的、在生命中起着重要调控作用的分子。miRNA 能有效地抑制相关蛋白质的合成，导致靶 mRNA 的降解，或者其他形式的调节机制来抑制靶基因的表达。近年来发现 miRNA 可能在基因表达调控领域中起着超乎想象的重要作用，miRNA 序列、结构、丰度和表达方式的多样性，使其可能作为蛋白质编码 mRNA 的强有力的调节子。miRNA 的发现丰富了人们对蛋白质合成控制的认识，补充了在 RNA 水平对靶 mRNA 分子进行更迅速和有效的调节，展现了细胞内基因表达调控全方位多层次的网络系统。

miRNA 和疾病关系的研究开始于肿瘤研究。随着 miRNA 的不断发现，他们在肿瘤中扮演的角色也日趋重要。miRNA 不仅参与细胞的增殖、分化，在肿瘤形成早期也有重要的作用。利用 miRNA 表达谱可以比基因表达谱更好区分肿瘤类型和亚型。在肿瘤形成早期，miRNA 的表达变化也成为了一种重要的分子标记用来识别癌前和癌的早期形成。关于 miRNA 在肿瘤中的研究为攻克这一人类历史上最为重大的难题之一奠定了基础和提供了重要的线索。

miRNA 不仅参与肿瘤的疾病过程，它在心脑血管疾病、神经系统疾病、免疫系统疾病等众多复杂疾病中的作用也逐渐被科学研究所揭示。每年都有几百篇科研论文发表了关于 miRNA 和复杂疾病关系的新研究。生物信息学家也开始着手收集这些文献资料，利用文本挖掘等方法从中提取 miRNA 和复杂疾病的信息，也形成了几个小型数据库，如 HHMD、miR2Disease 等。这些数据库的出现是对疾病知识的重要补充，为人们深入研究疾病过程中复杂的分子调控机制提供了必要的信息来源。

疾病和基因网络分析方法当然也可以应用于从 miRNA 调控的层面研究复杂疾病，但是由于疾病和 miRNA 的关系研究还不完备，已有的数据库信息量相对较少，研究偏好和数据资料的不足将显著地影响研究结果。因此，基于网络方法的研究还主要致力于研究单一疾病或某类疾病过程中的 miRNA 调控关系。也有部分研究结合了转录调控信息，从基因的启动转录调控和 mRNA 的翻译调控两个层面共同研究疾病过程中的表达调控改变。关于 miRNA 和复杂疾病的研究在第十六章将详细叙述。

五、其他类型网络重构分析

在复杂疾病的分子过程中，除了有基因、miRNA、转录因子等生物分子参与外，SNP、蛋白质、代谢和信号转导过程中的小分子等也在疾病过程中发挥重要的作用。关于 SNP 和复杂疾病的研究将在第十五章详细叙述。除此之外，各个层面的生物学数据和知识：如基因表达谱、功能注释、蛋白质互作、代谢组学等信息也在复杂疾病研究中得到了应用。

Wang 等结合 OMIM 和 GenAge 数据库中存储的疾病和衰老与基因的关系，将疾病和衰老通过关联的基因映射到蛋白质互作网络 HPRD 上，构建了疾病 - 衰老网络 DAN(图 14-19)。通过对 DAN 网络性质的分析，研究者认为疾病和衰老具有密切的联系，并可根据与衰老的联系划分为两类疾病，同时识别出了联系疾病和衰老的基因集合。

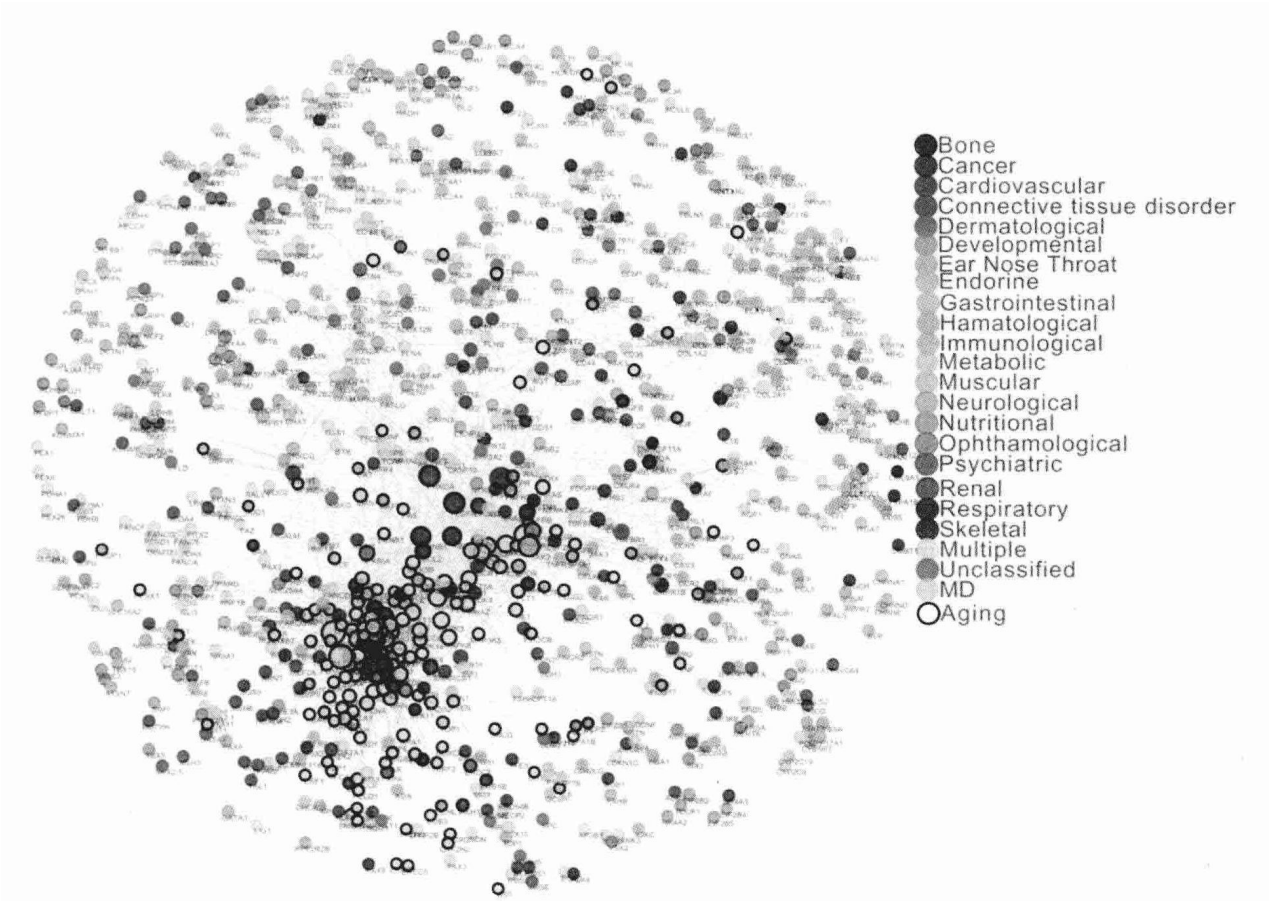


图 14-19 衰老网络

Hu 等利用从 GEO 数据库中获得的疾病和药物相关的基因表达谱构建了疾病 - 药物网络，从中发现了某些疾病在治疗药物影响上的相似性及疾病潜在的治疗药物和药物的适用范围等。

开发新的整合多层次信息的计算系统生物学研究方法和策略，尤其是结合基因表达调控、基因

组变异、基因功能、代谢和信号通路、蛋白质互作等从序列到功能的纵向多层膜信息整合研究，是复杂疾病分子基础研究的新机遇和挑战。

小 结

本章介绍了疾病的概念、特点、影响因素和分类方法。重点介绍了 OMIM、GAD、CGAP 等重要疾病数据库存储的主要信息和基本使用方法并给出了使用实例。这些数据库从实验、文献挖掘、关联分析等不同角度收录了疾病相关的基因、染色体风险位点、相关文献支持、实验证据等多方面信息。掌握这些数据库的使用方法，合理使用数据库中存储的疾病信息是复杂疾病的分子基础研究的重要保障。本章简要介绍了 ICD 和 DO 两个疾病分类体系。这两个疾病分类体系是针对所有常见流行病的国际上标准的诊断学分类，其中，ICD 被许多临床及卫生管理部门广为采用。本章还展望了计算系统生物学方法在复杂疾病研究过程中的应用，并结合科研实例探讨了疾病 - 基因网络、疾病 - 通路网络、疾病 miRNA 网络等前沿进展。

Summary

This chapter describes the concepts, characteristics, influencing factors and classification of disease. A brief introduction and usage of OMIM, GAD, CGAP and other important diseasesis presented with useful examples. These databases contain information from experiments, literatures and association analyses, including disease-related genes, the risk loci, the relevant documentary evidence supports, and other aspects. Information stored in these databases is very useful in complex disease study. This chapter briefly describes the ICD and the DO disease classification system. These two disease classification systems are international standard diagnostic classification systems, in which, ICD has been used by a number of clinical and health institudes. The chapter also views some computational systems biology methods in complex disease study. Combined with examples from journal papers, the disease-gene network, disease-pathway network and disease-miRNA network are discussed.

（李 霞 宫滨生 李 想）

习 题

1. 什么是复杂疾病？复杂疾病有哪些特点？
2. 什么是孟德尔疾病？复杂疾病和孟德尔疾病有何异同？
3. 简要说明复杂疾病与基因和环境的关系。
4. OMIM数据库包括哪些主要内容？请以白血病为例在OMIM数据库中获取白血病相关的信息。
5. GAD 数据库包括哪些主要内容？与 OMIM 数据库相比有哪些异同点？
6. CGAP 数据库包括哪些主要内容？请以白血病为例在 CGAP 数据库中获取白血病相关的信息。这些信息对在 OMIM 数据库中获得的信息有哪些补充？
7. 简述 ICD 体系的主要内容和应用领域。
8. 研究 Disease Ontology(DO)的体系结构，探讨和 ICD 的联系。
9. 讨论二部网络重构方法在复杂疾病研究中的应用。
10. 浅谈你对“生物体是一个复杂的网络”这句话的认识。

主要参考文献

1. Becker K. G., Barnes K. C., Bright T. J., et al. *The genetic association database.* Nat Genet, 2004. 36(5): p. 431-432.
2. Cooper D. N., Krawczak M. *Human Gene Mutation Database.* Hum Genet, 1996. 98(5): p. 629.
3. Ding L., Getz G., Wheeler D. A., et al. *Somatic mutations affect key pathways in lung adenocarcinoma.* Nature, 2008. 455(7216): p. 1069-1075.
4. Goh K. I., Cusick M. E., Valle D., et al. *The human disease network.* Proc Natl Acad Sci U S A, 2007. 104(21): p. 8685-8690.
5. Hamosh A., Scott A. F., Amberger J., et al. *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2002. 30(1): p. 52-55.
6. Hamosh A., Scott A. F., Amberger J., et al. *Online Mendelian Inheritance in Man (OMIM).* Hum Mutat, 2000. 15(1): p. 57-61.
7. Lee D. S., Park J., Kay K. A., et al. *The implications of human metabolic network topology for disease comorbidity.* Proc Natl Acad Sci U S A, 2008. 105(29): p. 9880-9885.
8. McKusick V. A. *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. 80(4): p. 588-604.
9. Osborne J. D., Flatow J., Holko M., et al. *Annotating the human genome with Disease Ontology.* BMC Genomics, 2009.10 Suppl 1: p. S6.
10. Rebhan M., Chalifa-Caspi V., Prilusky J., et al. *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet, 1997. 13(4): p. 163.
11. Safran M., Chalifa-Caspi V., Shmueli O., et al. *Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE.* Nucleic Acids Res, 2003. 31(1): p. 142-146.
12. Stenson P. D., Mort M., Ball E. V., et al. *The Human Gene Mutation Database: 2008 update.* Genome Med, 2009. 1(1): p. 13.
13. Wang J., Zhang S., Wang Y., et al. *Disease-aging network reveals significant roles of aging genes in connecting genetic diseases.* PLoS Comput Biol, 2009. 5(9): p. e1000521.
14. Wu X., Jiang R., Zhang M. Q., et al. *Network-based global inference of human disease genes.* Mol Syst Biol, 2008. 4: p. 189.

第十五章 单核苷酸多态与人类疾病
CHAPTER 15 SNP IN HUMAN DISEASES

第一节 引 言
Section 1 Introduction

人类疾病的发生是多种因素共同作用的结果。绝大多数常见疾病，如糖尿病、癌症、心脏病、精神性疾病等具有非常强的家族聚集特征，表明遗传因素在疾病形成中有重要作用；而同一家族某些成员发病，另一些成员不发病，以及同一种疾病在不同个体中具有不同的严重程度和表现症状，这些又体现了常见疾病的多因素特征。事实上，现有研究提供了大量的证据显示常见疾病遗传上的复杂性，认为常见疾病是众多基因共同作用的结果，人与人之间在疾病发生中的差异很大程度上可以通过遗传变异来解释，并在此基础上提出了著名的“常见疾病，常见变异”假说。

任意两个不相关个体的 DNA 序列有 99.8% 是一致的，而剩下的 0.2% 由于包含了遗传上的差异因素，造成人们不同的生理表型、罹患疾病的风险及不同的药物反应，这些差异在人类多样性形成中具有同等重要意义。这 0.2% 的差异在基因组序列中具有不同类型和作用形式。其中，不同个体 DNA 序列上的单个碱基的差异，称作单核苷酸多态性(single nucleotide polymorphism，SNP)，如图 15-1(A)。例如，某些人的染色体上某个位置的碱基是 A，而另一些人的染色体的相同位置上的碱基则是 G，而同一位置上的每个碱基类型叫做一个等位(allele)。除性染色体外，每个人体内的染色体都有两份，即同源染色体，一对同源染色体上的两个等位的组合叫做基因型(genotype)，如图 15-1(B)。对上述 SNP 位点而言，一个人的基因型有三种可能性，分别是 AA、AG 和 GG。而检定基因型的过程，称作基因分型(genotyping)。由于 SNP 在人群中具有最大的数量和最广泛的分布，且易于分型，它已经成为现代遗传变异与复杂性状研究中最重要的研究对象，也是生物医学、农业、畜牧业研究中非常重要的研究工具。

如果将世界上所有人看作一个群体，那么全人类中大约存在一千万个 SNP 位点，这些 SNP 绝大多数呈现二态性，并且具有不同的等位频率，人们将在某个研究群体中出现较少的等位频率称作最小等位频率(minor allele frequency，MAF)，并以此将 SNP 划分为常见和罕见两类，一般说来，常见的 SNP 最小等位频率应当大于 5%(也有文献定为 1%)，具有比较广泛的群体分布，与个体表型差异和疾病易感有关；而罕见的 SNP 往往是某些单基因病或偶发疾病的承载者。由于减数分裂过程中，染色体发生重组的位置具有选择性，染色体上距离越近的 SNP 越倾向于以一个整体遗传给后代，这样，把位于染色体上某一区域的一组相互关联的 SNP 称作一个连锁块(linkage block)，这是将 SNP 作为一种重要的遗传标记进行复杂性状和复杂疾病定位的分子基础。

除了从频率的角度对 SNP 进行划分，并在此基础上进行基于统计思想的遗传定位分析外，由于 SNP 本身数量众多、分布广泛等特点，它还具有非常重要的功能特性。人们习惯于将分布在基因(编码或非编码)区域，并且能够直接影响基因表达数量或基因产物(蛋白质或 RNA)结构的 SNP 称为非同义 SNP(non-synonymous SNP)。在实际研究中，还发现不同 SNP 之间具有潜在的相互联系，同一个基因或同一个生物学过程中多个 SNP 的互相作用能够起到从量变到质变的效果，直接影响生理指

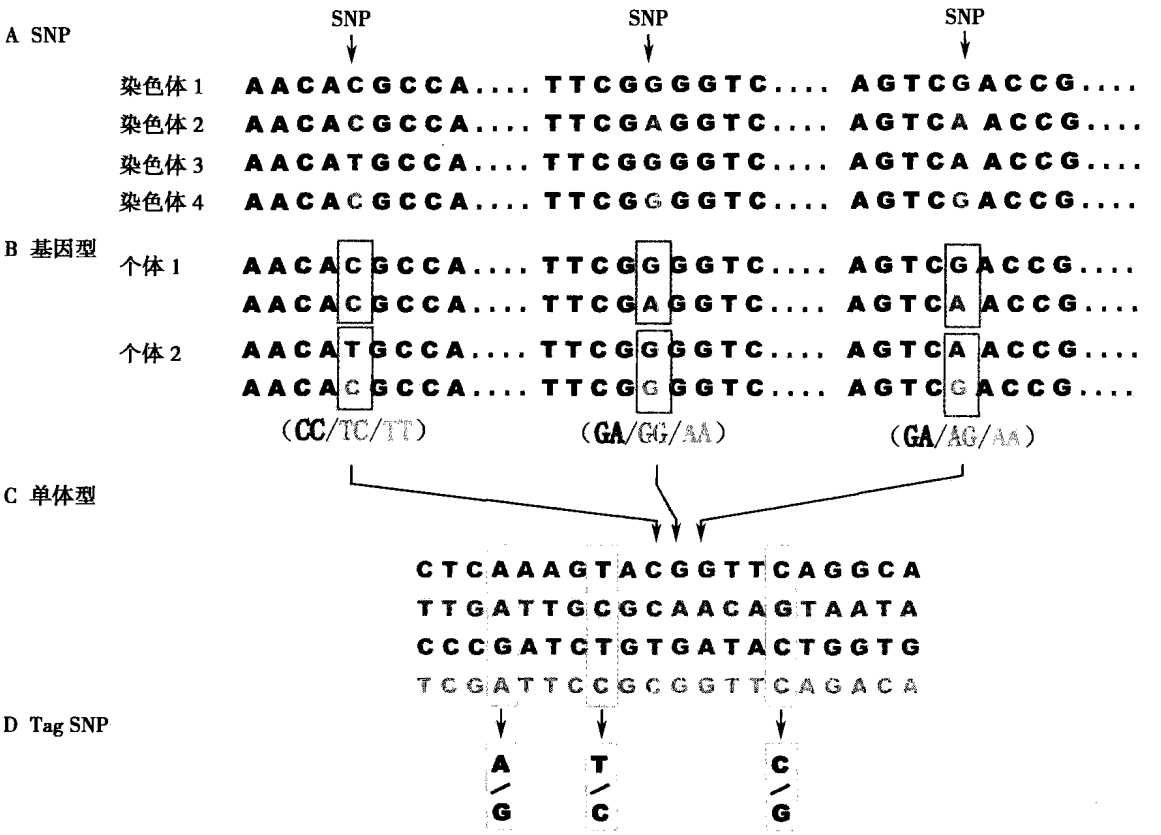


图 15-1 SNP、基因型、单体型与 Tag SNP

A 图中彩色标记出不同的 SNP 位点，及其在不同个体中的等位情况；B 图显示同一个体某个基因座上两个等位位点组合，即基因型；C 图中将某个个体的同一条染色体上的 SNP 放在一起，将其定义为单体型，这里的单体型是一个狭义的概念，也是本章研究的单体型含义；D 图是在单体型基础上提出的基于群体分布的单体型标签，即 Tag SNP

标、病理发生和药物反应的差异性。因此从功能和生物学系统的角度研究 SNP 在复杂性状和复杂疾病中的作用非常重要。

本章将从 SNP 用于复杂疾病研究的基本概念出发，介绍传统的以 SNP 作为分子标记的实验设计，以及采用统计学、机器学习方法进行风险 SNP 遗传定位(或称遗传做图)的基本技术；同时，从 SNP 与复杂性状形成的角度，共同探讨 SNP 用于复杂性状和分子定位的研究方法；在此基础上，融合功能信息学和系统生物学知识，介绍面向 SNP 功能和生物学过程研究复杂疾病的基本思想；最后，将 SNP 与复杂疾病的研究向现在有一定认识的多种人类遗传变异进行扩展，展开一幅比较完整的人类变异组与复杂性状形成和复杂疾病发生的相互关系画面。

第二节 SNP 分型技术与数据资源

Section 2 SNP Genotyping Technologies and Resources

一、SNP 检测和分型技术

围绕 SNP 展开的研究工作，首要任务是实现目标 SNP 的分型。与为数有限的蛋白质测序和 DNA 序列分析方法相比，SNP 测定的基本方法在数量上已达几十种，按其研究对象主要分为两大类，第一类是对未知 SNP 进行分析，即找寻未知的 SNP 或确定某一未知 SNP 与某遗传病的关系；第二类是对已知 SNP 进行分析，即对不同生物群 SNP 遗传多样性检测或在临床上对已知致病基因的遗

传病进行基因诊断。在实际应用中许多检测未知 SNP 的方法也可用来对已知 SNP 进行检测，而对已知 SNP 检测的方法也可用于对未知 SNP 的粗筛，筛选后再用测序方法确定 SNP 突变类型及其位置。具体而言，有以下一些主要的 SNP 检测方法。

（一）基于分子杂交的方法

1. 等位基因特异寡核苷酸片段分析(allele-specific oligonucleotide, ASO) 运用 PCR 和 ASO 方法相结合，设计一段 20bp 左右的寡核苷酸片段，其中包含了发生突变的部位，以此为探针，与固定在膜上经 PCR 扩增的样品 DNA 杂交。可以用各种突变类型的寡核苷酸探针，同时以野生型探针为对照，如出现阳性杂交带，则表示样品中存在与该 ASO 探针相应的点突变，ASO 须严格控制杂交条件和设置标准对照，避免假阳性和假阴性。目前，已有商品化的检测盒检测部分癌基因 ASO 突变。

2. 基因芯片方法 基因芯片集成了大量的密集排列的已知的序列探针，通过与被标记的若干靶核酸序列互补配对，与芯片特定位点上的探针杂交，利用基因芯片杂交图像，确定杂交探针的位置，便可根据碱基互补配对的原理确定靶基因的序列。对多态性和突变检测型基因芯片采用多色荧光探针杂交技术可以大大提高芯片的准确性、定量及检测范围，应用高密度基因芯片检测单碱基多态性，为分析 SNP 提供了便捷的方法(更详尽的原理参见第七章)。目前 SNP 分型芯片对 SNP 的检测可以自动化、批量化，是基因组范围关联研究的最主要的技术支持。SNP 分型芯片种类很多，目前常用的 SNP 芯片单次测量数量已达到 50 万~100 万个 SNP，数据缺失率一般不超过 5%，单张芯片的价格也已降至 5000 元人民币以下，是一种高效的基因组范围 SNP 分型技术手段。基因芯片用于分型 SNP 的主要缺点是不能区分两个等位的来源，两等位之间不区分先后顺序(图 15-2)，实际研究中如要用到多个 SNP 的组合信息，需要进行算法推断。

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	…	SNP_n
样本$_1$	AG	TT	AG	AC	GT	CC	AT		CC
样本$_2$	AA	TT	AG	CC	GG	CG	TT		CG
样本$_3$	GG	CC	AG	CC	GT	GG	TT		GG
…									…
样本m	AG	TC	GG	AA	GT	GG	AA	…	GG

图 15-2 SNP 分型芯片获得的数据信息

（二）以 SNP 影响核酸构象为基础的方法

1. 变性梯度凝胶电泳(denaturing gradient gel electrophoresis, DGGE)和温度梯度凝胶电泳(temperature gradient gel electrophoresis, TGGE)法 DGGE 法分析 PCR 产物，如果突变发生在最先解链的 DNA 区域，检出率可达 100%，检测片段可达 1kb，最适范围为 100~500bp。基本原理基于当双链 DNA 在变性梯度凝胶中进行到与 DNA 变性湿度一致的凝胶位置时，DNA 发生部分解链，电泳适移率下降，当解链的 DNA 链中有一个碱基改变时，会在不同时间发生解链，因影响电泳速度变化程度而被分离。由于本法是利用温度和梯度凝胶迁移率来检测，需要一套专用的电泳装置，合成的 PCR 引物最好在 5′末端加一段 40~50bpGC 夹，以利于检测发生于高熔点区的突变。在 DGGE 的基础上，又发展了用湿度梯度代替化学变性剂的 TGGE 法。DGGE 和 TGGE 均有商品化的电泳装置，操作简便，适合于大样本的检测筛选。

2. 单链构象多态性(single strand conformation polymorphism, SSCP) SSCP 是一种基于单链 DNA 构象差别的点突变检测方法。相同长度的单链 DNA 如果碱基顺序不同，甚至单个碱基不同，就会形成不同的构象，在电泳时泳动的速度将产生差异。将 PCR 产物经变性后，进行单链 DNA 凝胶电泳时，靶 DNA 中若发生单个碱基替换等改变时，就会出现泳动变位。这种方法多用于鉴定是否存在突变及检测未知突变。

（三）基于酶切的方法

限制性片段长度多态性(restriction fragment length polymorphism, RFLP)是实验室中最常用的低

通量 SNP 分型方法之一。由于 DNA 上的多态性致使 DNA 分子的限制酶切位点及数目发生改变，用限制酶切割基因组时，所产生的片段数目和每个片段的长度不同，即所谓的限制性片段长度多态性。导致限制片段长度发生改变的酶切位点，又称为多态性位点，最早是用 Southern Blot/RFLP 方法检测，后来采用 PCR 与限制酶酶切相结合的方法，现在多采用 PCR-RFLP 法进行研究基因的限制性片段长度多态性。

（四）测序方法

1. 直接测序方法　是检测 DNA 序列中 SNP 的重要方法，与基因芯片进行 SNP 分型的方法相比，测序结果能够完全呈现 SNP 等位在人类基因组精确的排列顺序（图 15-3），提供具有更高信息含量的数据，对于研究遗传互作有重要的意义。随着第二代测序技术的兴起，对基因组直接测序获得高通量、全局性的 SNP 图谱将很快成为最经济、最准确，应用最广泛的高通量分型技术。

	相型	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7	…	SNP_n
样本$_1$	相型$_1$	G	T	A	A	T	C	A		C
	相型$_2$	A	T	G	C	G	C	T		C
样本$_2$	相型$_1$	A	T	G	C	G	C	T		C
	相型$_2$	A	T	A	C	G	G	T		G
样本$_3$	相型$_1$	G	C	G	C	G	G	T		G
	相型$_2$	G	C	G	C	T	G	T		G
…										…
样本m	相型$_1$	A	T	G	A	T	G	A		G
	相型$_2$	G	C	G	A	G	G	A	…	G

图 15-3　直接测序方法获得的 SNP 基因型数据

2. SNP-shot-Gene-Scan 技术　是利用 GeneScan 检测 SNP 的一项技术，也称为小测序技术，可以在检测已知 SNP 时替代测序，但比测序方便、省时、经济。

除以上列举的各种常用方法外，质谱技术、序列标签、分子信标、原子探针显微镜、DNA 焦磷酸序列分析和 EST 分析等技术也可用于 SNP 检测或分型。

二、连锁不平衡、单体型与 Tag SNP

（一）连锁不平衡

SNP 与复杂疾病相关性研究中最重要的概念是连锁不平衡。连锁不平衡（linkage disequilibrium，LD）是指相邻基因座上等位基因的非随机相关，当位于某一基因座上的特定等位与同一条染色体另一基因座位上的某等位同时出现的概率高于或低于人群中的随机分布，就称这两个位点处于连锁不平衡状态。假定两个 SNP 1 和 2 各有两个等位型（A，a；B，b，SNP 等位应为 A、C、G、T 四种，这里用 A、B 表示便于描述），那么同一条染色体上将有四种可能的组合方式：A-B，A-b，a-B，a-b。假定等位 A 的频率为 P_A，B 的频率为 P_B，那么在连锁不平衡条件下，等位组合 A-B 的频率 $P_{AB} \neq P_A \times P_B$，而是 $P_A \times P_B + D$（D 表示两位点间的连锁不平衡程度）。正是由于连锁不平衡的存在，才可能将 SNP 原有的单个位点的差异拓展到某个区域或某个基因和生物学过程的研究层面。

导致连锁不平衡的主要因素有遗传漂变、人口增长与群体结构改变、重组率变化、突变率变化和基因转换。人群迁移、隔离、再分能够增加 LD 程度，而人口增长、世代增加、CpG 源新 SNP 发生、基因转换能够减小 LD 程度。相对于短暂的人类史，人口历史因素对 LD 的影响很大。世界上不同地域的群体经历了不同程度的迁移、混合或遗传，造成了不同区域间的 LD 程度差别很大，如欧洲人群的 LD 程度远高于非洲人群。另外，LD 程度在基因组不同区域也有很大差别，某些区域两个相距很近的位点具有很弱的 LD，而另一区域的两个相隔 100kb 的位点却可能具有较强的 LD。

（二）连锁不平衡的量度

连锁不平衡有很多度量方法，它们均可以用来度量 SNP 这样的二态遗传多态位点间的连锁关

系，也有部分度量能够扩展到多个位点及多种状态的情况。目前常用的度量方法主要是 D'、r^2 和 LOD 值，这里主要对前两个度量进行介绍，LOD 值方法将在第五节中介绍。D'、r^2 的取值范围均在 0(连锁平衡)和 1(连锁不平衡)之间，但具有不同的意义。

1. r^2 值量度 LD r^2 代表两位点在统计学上的关系，其表达式为：

$$r^2 = (P_{AB} - P_A P_B)^2 / P_A P_a P_B P_b \qquad \text{式 15-1}$$

式 15-1 中，P_A、P_a、P_B、P_b 分别为 A、a、B、b 等位频率，P_{AB}、P_{Ab}、P_{aB}、P_{ab} 分别是 AB、Ab、aB、ab 四种单体型的频率。r^2 等于 1 说明两位点没有被重组分开，且等位基因频率相同。r^2 的数值表示一个位点可反映另一位点信息量的程度，$r^2=1$ 称为完全连锁不平衡，这时两位点等位基因频率相同，只观察一个标记即可提供另一个标记的全部信息。另外，需要指出的是，r^2 在小样本中不会显著增加。

2. D' 值量度 LD D' 值又称为连锁不平衡系数，其表达式为：

$$D' = D / D_{\max}, D = P_{AB} - P_A P_B \qquad \text{式 15-2}$$

式 15-2 中，P_{AB} 为 A、B 两个等位连锁出现的频率，A、B、a、b 的频率分别为 P_A、P_B、P_a、P_b，$D_{\max} = \max(P_A P_b, P_a P_B)$，即 $D_{\max}$ 取 $P_A P_b$、$P_a P_B$ 当中的最大值。当 $D'=1$ 时，说明两个位点间没有发生重组，与 r^2 相比较，D' 等于 1 时两位点等位基因频率并不需要相同，它只是反映最近一次突变发生后突变位点与邻近多态性位点的关系。如果 $D'<1$，则说明这两个位点间发生过重组或新发生了突变，如果 D' 值接近于 1，则两位点 LD 历史上发生重组的可能性很小，但如果 D' 处于中间值，则不能用它来比较两位点 LD 程度的差别。

仅考虑两个位点的 LD 度量方式在使用和计算上具有优势，但是当有多个位点需要综合考虑的时候，两点的度量方式将损失信息，因此，多点度量具有极大的应用研发价值。D' 度量能够比较方便地扩展到多个位点 LD 情况，有较为广泛地应用。

(三) 单体型、单体型块和 TagSNP

1. 单体型 单体型(haplotype)是指一条染色体上紧密连锁的多个基因的线性排列，图 15-1(C)。SNP 单体型就是不同 SNP 位点上核苷酸碱基的线性排列，每一种线性排列称为一种 SNP 单体型。如果在某一段 DNA 片段上发现 10 个 SNP，理论上可能存在 1024(2^{10})种单体型，但由于 LD 的存在，实际上真实出现的单体型数目远少于理论上的数目。

2. 单体型块与 TagSNP 基因组中单体型呈“块状”分布。这样的块状结构称为单体型块(haplotype block)。同一单体型块中的 SNP 间处于高度的 LD 状态，有共同遗传的趋势。不同单体型块中的 SNP 个数、单体型类型、块跨度是不同的。另外，由于基因组不同区域的重组率大不相同，从而形成间隔不同的单体型块，而单体型块之间的区域重组概率大，这些位置被称为重组热点(recombination hot spot)。

如果单体型块已经确认，就可以精确的查找到其中某些特异的 SNP，利用这些 SNP 与周围 SNP 之间的紧密连锁，可以从中选择一定的组合来代表整个单体型块中的绝大多数单体型类型，这些 SNP 称为单体型标签 SNP(TagSNP)，图 15-1(D)。通过对 TagSNP 的识别和分型，来研究疾病，能够有效的识别疾病相关的染色体位点，并且能够用最少的 SNP 数量实现全基因组范围扫描，极大地节省了花费，因此是目前商业 SNP 分型芯片的主要组成部分。着眼于单体型块可以更好的阐明 LD 结构。如果可以确认单体型，可以从中发现 TagSNP，从而捕获单体型或研究 LD 区域，还可以把每个单体型块看作一个等位基因，来进行 LD 分析，而单体型块方法比单个 SNP 更能精确反映生物基因组的多样性，并能够提供更清晰的 LD 图。

由于现有的高通量 SNP 分型技术主要是基于基因芯片的方法，测定结果不能直接反映 SNP 的相型(phase)特征，所以基于单体型块和 TagSNP 的研究方法需要首先对单体型块和 TagSNP 进行推断，这样的方法已经有很多，但有各自的侧重(参见第五节 Haploview 部分)，开发更为完善的推断方法也是 SNP 研究的一个重要领域。

三、国际人类基因组单体型图计划及其应用

（一）国际人类基因组单体型图计划概况

国际人类基因组单体型图计划（The International HapMap Project，http://hapmap.ncbi.nlm.nih.gov）是继国际人类基因组计划之后，人类基因组研究领域的又一个重大国际合作项目。HapMap 计划起始于2002年，由美、加、中、日、英、尼日利亚等国研究机构发起、参与及完成。中国科学家承担3号、21号和8号染色体短臂单体型图的构建，工作量约占总计划的10%。项目共取样270个正常个体，其中有欧裔美国人和尼日利亚雅鲁巴人（非洲）各30个核心家系（90个个体），还包括中国北京汉族人及日本东京人各45个个体。一期已于2005年完成，成功分型100多万个常见SNP位点，达到平均每3kb一个SNP的测定。由于染色体连锁不平衡的存在，一期数据可以捕获基因组上80%的遗传差异信息。二期计划在一期基础上完成300多万个SNP位点的分型，构建起一张精度更高、信息更完整的多人种遗传多态图谱。三期计划已经开展，在进一步测定原有群体基因型基础上，加入另外7个不同历史遗传背景的人群，部分分型数据已经发布。HapMap 计划期望在全部完成时能够提供一个包括全部人类遗传差异的多态组图谱，同时带动其他人类遗传变异的发现和研究。

（二）HapMap 数据特点与扩展应用

HapMap 计划建立了人类全基因组遗传多态图谱，依据这张图谱可以进一步研究基因组的结构特点以及SNP位点在人群间的分布情况，为群体遗传学、进化遗传学分析提供数据，也为复杂疾病的遗传定位提供高密度的SNP数据参考。HapMap 的构建分为三个步骤：①在多个个体的DNA样品中鉴定单核苷酸多态（SNP）；②将群体中频率大于1%的那些共同遗传的相邻SNP组合成单体型；③在单体型中找出用于识别这些单体型的标签SNP。这样，HapMap 提供的每个研究个体的数据包括SNP等位、基因型、基因型频率、200kb范围内SNP之间的LD量度（r^2、D'）。

伴随 HapMap 计划的进一步拓展，结合群体遗传学的研究手段，人们可以更加深入地去观察和研究基因组。基于大群体、多种群的人类单核苷酸多态数据的重组率推算提供了一张基因组进化痕迹图；连锁不平衡的计算提供了一张基因组块状连锁结构图；种群差异研究让人们看到一张种群间基因组结构差异图；SNP的杂合情况揭示人类基因组上受到选择的区域或区域内的基因；利用SNP位点向两边延伸的长度差异情况，可以观察到一些基因组上近期正在进行的选择事件，甚至是当前正在悄悄进行中的进化，因为新产生的突变位点传代较少，它和周围位点的连锁情况受重组事件的影响较小，另一方面优势突变也会因选择压力的存在使周围的重组受到影响……当然这些不同的指标中也隐藏了人类成长过程中的一些信息，如迁徙、战争、灾难、繁盛等对基因组遗传多态性产生影响的历史事件。

此外，高密度的SNP位点，为进一步加强和完善基因组范围的表型和遗传相关性分析（关联研究或数量性状定位）提供了可能，以往遗传学上定位基因使用较多的工具是微卫星，这些新产生的SNP位点弥补了微卫星在基因组上分布不够均匀、密度不够高的缺点，是一种更为有效的分子标记。目前，已经有很多致病基因借助SNP数据得到定位。另外，根据SNP在基因的不同功能元件中的分布情况和基因在细胞中的表达情况，人们可以研究基因上的不同元件序列是如何控制蛋白质表达进而影响个体表型的。伴随着 HapMap 三期数据的产出、各种实验技术的进一步发展，以及更加大量的基因组序列数据加入到人类的知识库中，与此相关的研究方法和研究手段会不断出现，人类将能够更加完整、更加深入、更加正确地认识自己，揭示生老病死的奥秘，并为人类生存质量的提高提供有益的参考信息。

（三）利用 HapMart 进行科学研究

为了便于科研工作者快速提取感兴趣的SNP数据，在 HapMap 数据基础上，BioMart（一个重要的生物信息学数据分析平台）开发了方便、友好的SNP获取网络平台 HapMart。这个平台支持研究者输入SNP、基因、染色体区段等信息进行限定条件下的SNP查询及相关信息的输出。由于

HapMap 数据本身跨群体的特性，用户可以通过这个平台进行不同群体间的数据提取，如果是候选基因或多 SNP 实验设计，还可以联系其他的连锁不平衡分析工具（如下文将提及的 Haploview）及感兴趣的基因型频率信息进行深层次的 SNP 选择。利用 HapMart 进行 SNP 数据的提取主要分为三个步骤：输入设置、输出设置和结果导出。

1. HapMart 的输入设置　图 15-4 显示了 HapMart 查询过程中的输入和查询限定界面。在这里，可以进行研究群体的选择、SNP 质量限定，以及查询设置。目前，HapMart 主要支持四个群体的查询，后续的群体正在添加中。对于目标 SNP 可以进行最小等位频率、分型机构、分型平台、SNP 类型的限定。可以根据 SNP 的标识符、定位区域（功能区域或染色体位置），及其与基因的位置关系进行单个或高通量的 SNP 查询。

图 15-4　HapMart 的输入设置界面

2. HapMart 的输出设置　图 15-5 显示了 HapMart 的 SNP 输出属性设置界面。可以根据研究者的研究兴趣进行设定，并输出相应的结果。SNP 相关属性主要有标识、遗传定位、等位和基因型状态和频率特征。

3. 查询结果的导出　根据研究者的研究兴趣和输入输出设置，可以以特定的格式显示和导出查询结果。图 15-6 显示的是限定最小等位频率 0.01 时，定位在基因 *IL10* 上的 SNP 位置、等位和基因型频率信息。

HapMart 查询结果以 HapMap 数据为基础，提供的是不同种群特定群体的 SNP 信息，主要用于实验设计者针对特定人群的实验参考。由于计划测定规模的限制，数据本身存在一定的偏差，因此查询结果应当进行一定的预实验和初步分析，才能用于大规模实验。

四、重要的 SNP 数据库

（一）SNP 存储与维护数据库 dbSNP

SNP 作为新一代遗传标记具有数量多、分布广、密度大等特点，已广泛应用于遗传学研究中。重要的 SNP 数据库有 dbSNP 和 dbGap。为了满足对基因组范围总体变异的需求，解决在关联研究、基

图 15-5 HapMart 的输出设置界面

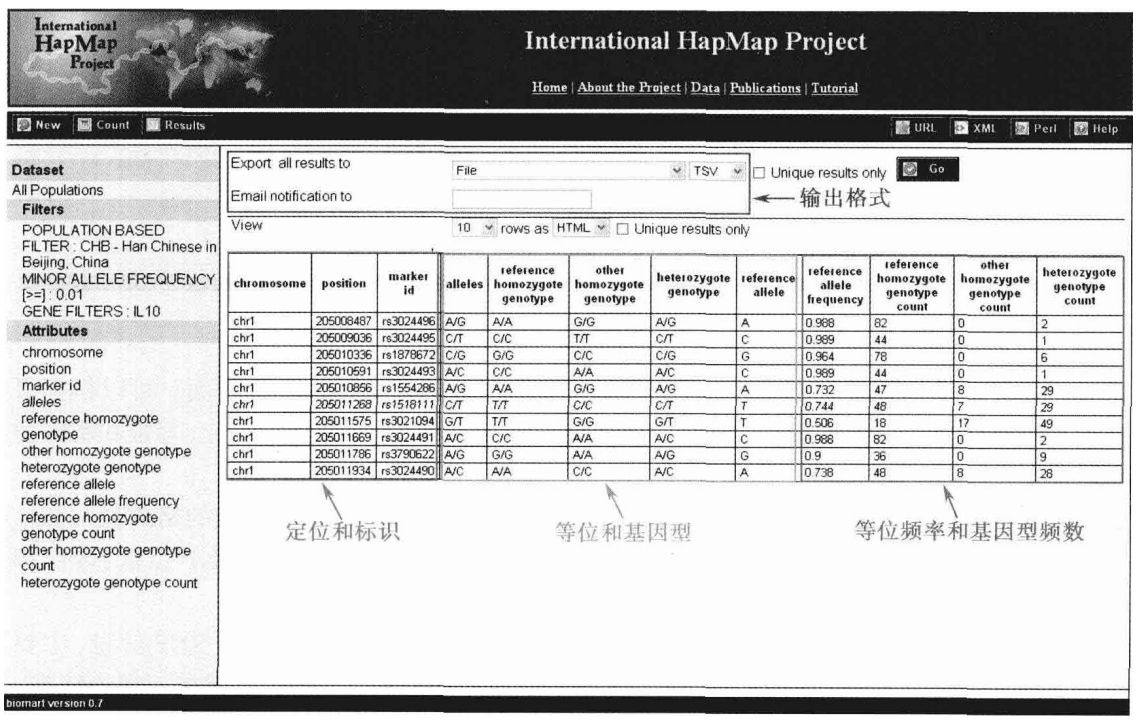


图 15-6 HapMart 的结果显示与导出界面

因定位、功能和药理遗传学、群体遗传学、进化生物学以及定位克隆、物理作图等领域中大规模抽样设计的需求，NCBI 与 NHGRI 协作创建了 dbSNP。通过 dbSNP，由公共和私人组织提交的遗传变异数据与其他信息来源，如 GeneBank、PubMed、LocusLink 及人类基因组数据实现交叉引用，为广大研究者提供了丰富的遗传变异，特别是 SNP 信息，呈现了一幅全面的人类 SNP 的基因组分布图。充分

利用数据库中资源将大幅度降低研究成本、提高研究效率。此处，就 dbSNP 数据库的功能、范围、数据提交、检索进行简要的介绍。

1. dbSNP 的主要功能

（1）遗传变异序列环境分析：dbSNP 通过 BLAST 和 E-PCR 对变异周围序列进行分析，将其链接到其他 NCBI 序列资源，对变异进行交叉注释。用户可直接在 dbSNP 中检索，或在 NCBI 查询空间的任何部分开始，构建一个满足要求的 dbSNP 记录集，该记录集可通过超文本或 URL 与外部信息资源整合。

（2）基于 NCBI 的遗传变异交叉注释：在后基因组时代，对特征序列的注释（如新基因或调控区域）为当前在随机序列中发现的变异提供一个功能背景。随着这些新基因条目的出现，dbSNP 通过链接能够将变异自动注释到恰当的参考序列集或 UniGene 集中。

（3）外部资源整合：dbSNP 具有“LinkOut URL”功能，将变异信息链接到 NCBI 之外的信息资源。这种整合非常重要，尤其是当考虑将变异注释到整个基因组上或考虑其对生物体的意义时。

（4）遗传变异的功能分析：NCBI 没有直接地在序列上注释变异的详细生物化学或者表型信息，而在 dbSNP 中保留了与外部数据库的链接。因此，dbSNP 记录能够链接到那些对个别变异描述更加完整的位点特异突变数据库。

2. dbSNP 数据特征　dbSNP 数据库中不仅收录了人类 SNP 数据，还收录了所有已知的跨物种的 SNP、插入 / 缺失、拷贝数和微卫星多态，且包含种族特异频率和基因型数据、实验条件、分子背景，以及功能特性和临床变异的定位信息。截止到 2009 年 10 月 7 日，dbSNP 已经更新至 130 版本，涉及 55 个物种的 1.5 亿个 SNP，编码区 SNP 已超过 2000 万个，具有频率信息的 SNP 超过 300 万个。

3. 向 dbSNP 提交数据　目前，科研领域出版物中涉及遗传变异信息一般要求提交到 dbSNP 数据库中。所需数据提交信息包括特定位点观察到的等位基因、突变周围的侧翼序列、使用的实验方法，伴有 STS 或 GeneBank 记录的指针。每个实验室具有唯一标识，允许提交的数据与实验室相关联。NCBI 给每个提交的 SNP 分配一个编号 ss#，一个物种基因组 SNP 也将分配一个标识符（人类的 SNP 标识符为 rs#）。所有这些编号或标识符被用于将 SNP 映射到外部资源或数据库中，包括 NCBI 中其他数据库。

4. 利用 dbSNP 进行信息检索　在 dbSNP 中可直接查询，也可通过其他 NCBI 查询框来检索。直接查询可以通过提交实验室、新的批量提交、鉴定方法、群体类型研究、期刊题目、群体变异水平或 STS 映射信息实现。作为 NCBI 中一个整合部分，dbSNP 中的内容与其他信息资源记录是横向链接的。从其中任何来源中查询的结果集合会给用户提供一个返回 dbSNP 相关记录的指针。图 15-7 显示的是以人类 *IL10* 基因相关 SNP 为例的 dbSNP 查询过程，及其显示结果，进一步点击蓝色链接将显示每个 SNP 的详细信息。

5. 提供 dbSNP 交叉引用的模块　BLAST：dbSNP 查询，可通过标准的 BLAST 算法来实现，即将用户提交的序列与 dbSNP 中所有侧翼序列记录进行匹配。除了在 NCBI 首页中提供了一般的 BLAST 功能，dbSNP 中也提供了此功能。LocusLink：dbSNP 也可通过将其与其他 NCBI 资源整合来检索。通过 LocusLink，由基因名字或系统命名来进行检索。从 LocusLink 数据库中检索的结果将呈现为一个紫色的“V”形按钮，该按钮可以指向一个 LocusLink 数据库中任何一个基因上的参考 SNP 记录列表。Entrez：“图形可视化”旁边的工具条有一个链接将 dbSNP 中的 SNP 记录链接到 Entrez Gene 数据库，这样的链接可以直接看到 Entrez Gene 中基因上的 SNP 分布情况，并能够根据需求，如是否具有频率、是否编码等信息进行可视化的 SNP 查询。Genome sequence：重叠视图除了可以设置为显示 STS“marker”和序列组成，还可以显示“变异体”。

图 15-8 显示了 dbSNP 与 Entrez Gene 之间的交叉引用结果，通过 Entrez Gene 向 dbSNP 的超链接，查询到 *IL10* 基因上的 SNP 分布情况。dbSNP 用不同的颜色和柱体长度表示基因上的 SNP 类型及其频率状况，对于深入的选取对研究有影响的 SNP 提供了直观的借鉴信息。从查询结果上看，

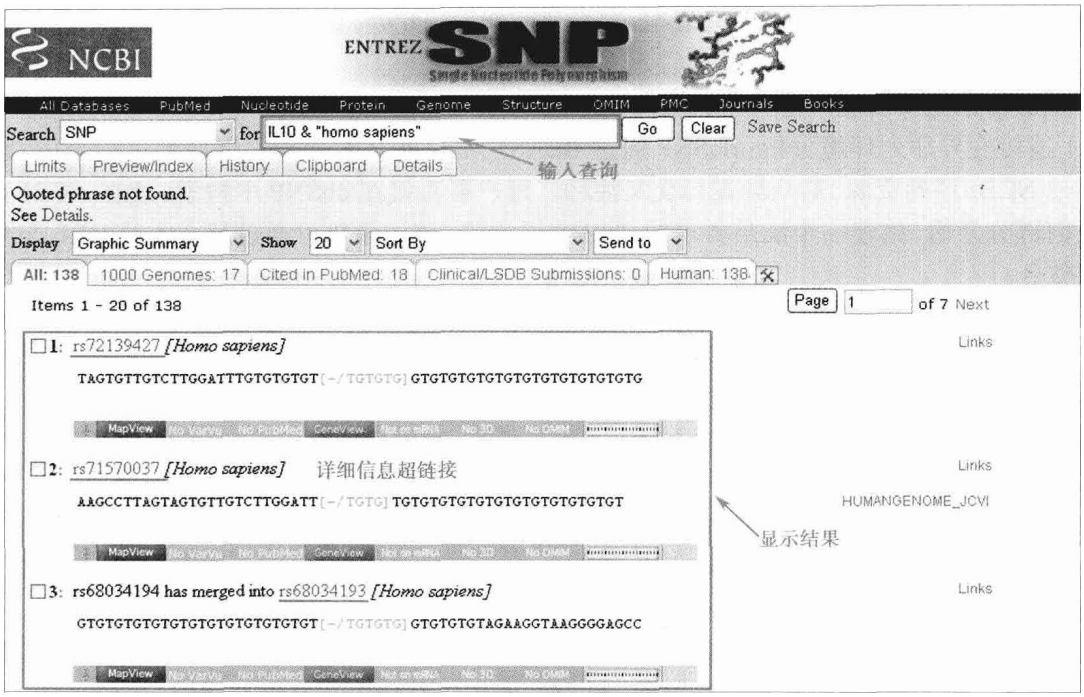


图 15-7 dbSNP 的查询界面

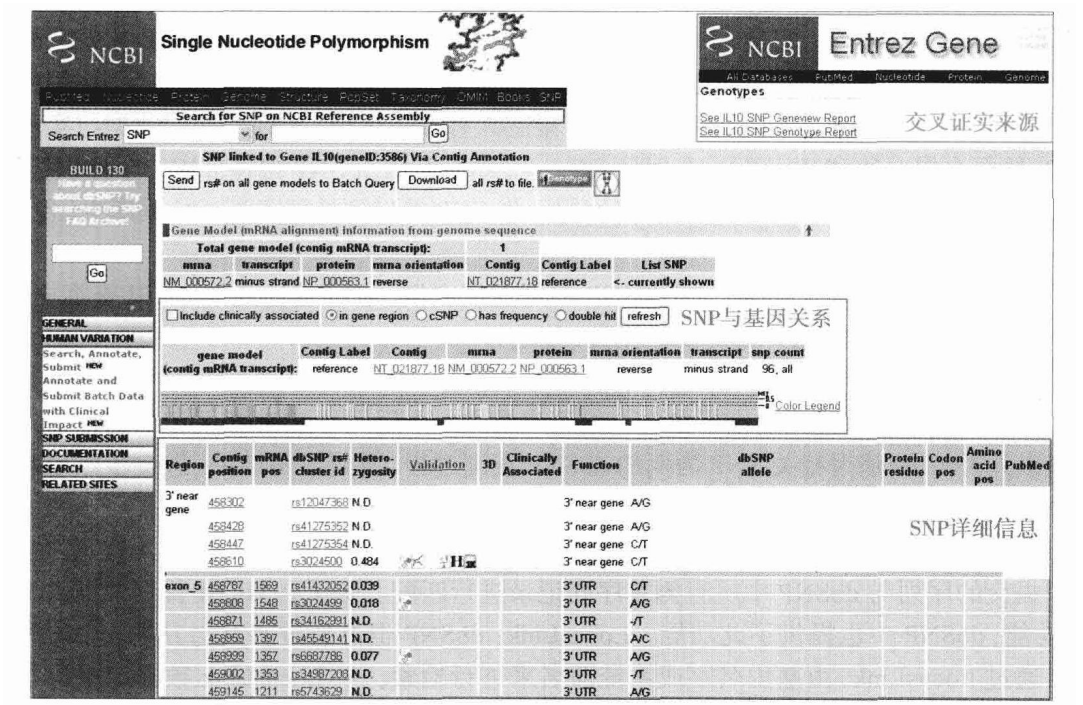


图 15-8 dbSNP 与 Entrez Gene 交叉引用

dbSNP 展开的查询获得的 SNP 数量要比 HapMart 的多，这主要是 dbSNP 本身不限制收录的 SNP 最小等位频率造成的，所以研究过程中还应当进一步考虑频率和相应的群体信息。

（二）关联研究基因型数据的存储与整理 dbGap

1. dbGap 的主要功能 NCBI 建立了基因型和表型数据库 dbGap。dbGap 的开发是为了存储和发布基因型和表型相关的研究数据及研究结果。这些研究包括全基因组关联研究、医疗测序、分子诊断化验，以及基因型与非临床性状之间的关联性。用于基因分型的高通量、低成本、高效率的分析

方法研究，发现海量基因型和表型数据相关性的未知信息提供了强有力的工具。

dbGap 是一个存储了个体水平表型、基因型、序列数据，以及它们之间的关联性的公共知识库。dbGap 收录的数据绝大部分是大规模的基因组范围关联研究数据，对研究过程中得到的信息子集，包括文件、个体表型变量、特征数据表、基因型数据，计算表型与基因型之间的相关性，设定唯一的标识符。其中的部分数据采用直接开放的管理办法，非注册用户即可直接下载相关的数据进行非商业化的科学研究。为了确保被研究者的个人权益及数据检测部门的优先使用权利，dbGap 中的大部分数据的访问和使用需要进行人工申请。

dbGap 包含了多样化的实验设计研究。它包括四个基本类型的数据：①研究文件，包括研究说明，协议文件和数据收集文书，如问卷调查表；②每个被评估变量的表型数据，包括在个体水平上的和以摘要形式进行评估的；③遗传数据，包括研究对象的个体基因型、谱系信息、精细定位结果和重新测序的描述；④统计结果，包括关联和连锁分析结果。

2. dbGap 中保存的数据访问　为了保护研究对象的权益，dbGap 只接受已被查证的数据申请，并要求使用者通过一个授权程序才可以获取个体水平的表型和基因型数据集。总结性的表型和基因型数据，以及研究文件，可以无限制的获取。

dbGap 提供两个访问级别：开放和受控，这么做的目的是为了让非敏感数据广泛开放，同时提供对涉及个人健康信息的敏感数据集进行负责任地监督和调查。研究的总结和测量变量的内容，以及原始研究文件的文本，一般会提供给公众，而要获得个体水平的数据，包括表型数据表和基因型数据就需要不同的授权级别。

（1）开放数据：开放式访问数据可以在线浏览或未经批准或授权就可以从 dbGap 中下载。这些数据包括但并不仅限于表 15-1 所列的内容。

表 15-1　dbGap 中的数据类型

dbGap 数据类型	信息所在位置
研究	当浏览研究时在名为“Study”的列中出现 在标签“Studies”下的一个搜索结果 通往一个变量或一个文件的路径的一部分
研究文件	从“Browse Studies”链接 与“Associated Documents”下的研究报告链接 标签“Study Documents”下的一个搜索结果
表型变量	与“Browse Studies”链接 与“Associated Variables”下的研究报告链接 标签“Variables”下的一个搜索结果
基因型 - 表型分析	与“Associated Analyses”下的变量报告链接 与“Associated Analyses”下的研究报告链接

这是一个可用于开放式进入用户的一般性描述。提供给开放式进入用户的数据可能根据数据维护的需要而有所变化，但一般不会提供书面说明。

（2）受限数据：受控访问数据只能在用户已通过适当的数据访问委员会（DAC）的授权后才能获得。申请受控数据访问的信息或提供给授权的研究人员的数据包括以下内容：①用于个人研究课题的确定的表型和基因型；②家族谱系；③基因型与表型前期处理过程单变量相关性（如果没有在公开网站上提供）。

由于数据访问策略是基于每个研究的基础上确定的，提供给用户的带有受控访问授权的数据在不同的研究之间可能会发生变化，也有可能在没有通知的情况下就与这里所描述的有所不同。关于用于一个特定的研究的数据的访问策略，可以在研究报告页连同适当的授权机构的链接上找到更多的细节。

第三节 基于SNP的复杂疾病遗传定位方法

Section 3 SNP-based Complex Disease Mapping Methods

复杂疾病机制研究是生物医学研究中的重中之重。致病基因的发现是研究复杂疾病机制的重要环节，也是长期困扰科学研究者的一个难题。从20世纪初，人们就在探索基于分子标记的统计分析方法用于致病基因的识别，到20世纪80年代，伴随分子生物学技术的革新，这一研究方案得到了长足的发展。这种方法通过进行标记测定，采用统计学方法研究分子标记的遗传特性与疾病发生之间的相关性，来实现疾病基因的染色体定位，而几乎不需要任何先验的生物学知识，是一种强大的疾病基因识别手段。随着SNP分型技术的发展，SNP作为一种最重要的分子标记，不仅能够成功地应用于孟德尔遗传病的研究，同时被广泛的用来进行复杂疾病的染色体定位。分子标记遗传定位过程包括样本选取、标记分型、数据分析和结果注释几个部分，标记分型方法在上节中已经介绍，本节将简要的介绍基于SNP的复杂疾病遗传定位实验样本选取准则、连锁分析、关联分析、统计结果的取舍等内容。

复杂疾病的遗传定位策略已经形成一个非常系统的研究领域，但由于疾病本身的异质性(genetic heterogeneity)及相关基因或位点的多效性(pleiotropy)和遗传互作(epistasis，也称上位效应)的存在，分子定位研究复杂疾病依然是一个充满挑战的领域，本节中会介绍部分机器学习和系统生物学思想进行风险SNP定位并研究疾病机制的必要性和研究进展。近几年，伴随测序和分型技术的提高，越来越多的研究倾向于将人类全基因组的SNP作为研究对象全面的进行关联分析，即基因组范围关联研究(genome-wide association study，GWA)，以此发现真实、完整、可靠的发现复杂疾病相关位点，实现基因定位，并已经取得了很大的成就。本节最后一部分将重点介绍基因组范围关联研究的发展前景和研究思路，以及由GWA研究衍生而来的基于生物学通路和网络的分析思想，既所谓第二代GWA研究。

一、疾病定义与样本选取偏好

遗传定位以疾病为研究对象，准确的疾病定义，特别是细化疾病的分类层次对于获得有针对性的致病因子有重要的意义，同时也是指导实验样本选取的首要条件。但正如前文所言，复杂疾病本身的特性，限制了疾病定义的准确性，在遗传定位研究中选择疾病和疾病样本一般遵循约定的五个原则，同时也是影响遗传定位结果的五个重要因素。

1. 临床表型　在临床中，同一疾病的不同亚型往往具有不同的临床症状，而特定的症状可能隐含着特定的遗传特征。以结肠癌为例，如果结肠癌患者有严重的结肠息肉，这一型的结肠癌实际上是一种与*APC*基因相关的显性遗传病，其他类型的结肠癌也可以根据临床表型进行区分。而在高血压研究中，由于原发性高血压往往伴发高血脂，可以就此进行原发性高血压患者的选取。由以上的两个例子可以看出，对患者临床表型的深入分析对于疾病分型，从而正确的选择遗传定位样本很重要。

2. 发病年龄　亲属风险(relative risk)能够表述疾病在亲属中发生的相关性，是流行病学中衡量遗传效力的重要参数。根据长期的调查，乳腺癌、阿尔兹海默病等大部分复杂疾病的早发个体具有较高的亲属风险，这表明选取疾病早发个体进行研究有利于进行遗传定位研究。

3. 家族史　某个个体家族中如果有成员罹患某种疾病有助于对其本身的疾病进行诊断，同时，具有家族史也是很多复杂疾病亚型(如息肉型结肠癌)的重要特征，有利于辅助疾病分类。

4. 严重程度　遗传定位实验设计中，偏好选择疾病发生严重程度比较高的个体，一方面患病严重的个体易于正确诊断，另一方面这些个体可能会具有更为典型的遗传特征。

5. 群体分层　由于相同疾病在不同的群体中往往有不同的遗传特性，样本选择过程应该尽量选

择同质性的群体。同时，由于连锁不平衡在群体中的分布特性，选择同质性的群体也有利于进一步获得候选基因。

因此，复杂疾病的遗传定位疾病组样本应该尽可能选择具有明确的临床症状、偏向于早发、具有家族史、病情较严重、同质性的个体；相对的，对照组中就应当避免出现与疾病组个体特征接近的个体。遵循这样的原则，才能最大限度上获得真实可靠的分析结果，打下良好的数据基础。

二、连锁分析进行风险 SNP 定位原理

连锁分析(linkage analysis)是根据家系中遗传标记重组率来计算两等位之间距离的方法，主要是通过分析已知的性状或疾病表型与基因型在家系中遗传模式，来定位新的易感位点和易感区域。正如定义中所述，连锁分析是用于研究家系中标记传递的一种分析策略。根据连锁分析过程中是否依赖于假设模型，将连锁分析方法分为两类：参数连锁分析和非参数连锁分析。

(一) 参数连锁分析方法

对于孟德尔遗传病(单基因病)，人们易于清楚地知道该疾病的遗传方式、外显率、基因频率等指标，从而确定一个准确的遗传模型进行连锁分析(图 15-9)。随着统计方法的不断发展，某些遗传模型并不清楚的疾病也通过改变策略而适用于连锁分析，但无论如何，相对准确的模型建立是参数连锁分析成功的基本条件。直接计分法和 LOD 值法是最常用的参数连锁定位方法，这里以 LOD 值法为例进行简要的介绍。

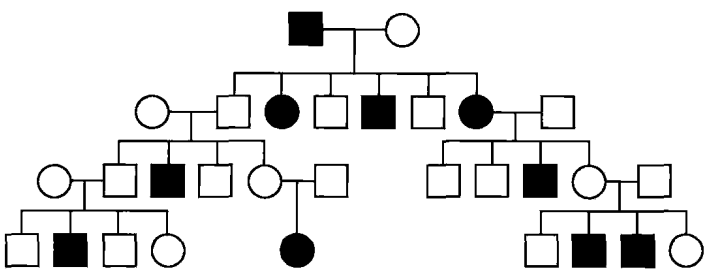


图 15-9 参数连锁分析所依据的家系遗传模型：典型常染色体隐性模型示意

LOD 值法进行连锁分析首先针对某一疾病收集一定数量的家系资料并进行分离分析，确定遗传模型；然后通过文献检索了解其可能的决定性状的染色体区域，并对该区域的 SNP 进行查询和筛选，基于选定的 SNP，对该家系成员进行基因分型；最后通过连锁分析估计疾病与 SNP 在子代中重组的发生率，计算 LOD 值，确定重组分数及相应的遗传距离，并进行假设检验，判断易感基因是否与遗传标记连锁。

LOD 值是指在一定重组率 θ 条件下，两个位点相连锁的似然性和不连锁的似然性比值的对数值。即

$$\text{LOD} = \log_{10} \frac{\text{两位点连锁的似然性}}{\text{两位点不连锁的似然性}} \qquad \text{式 15-3}$$

在进行连锁分析时，要计算 $\theta=0.0$(不重组)到 $\theta=0.5$(随机分配)的一系列 LOD 得分。当 LOD 得分为 +3 或更大时，肯定连锁；当 LOD 得分小于或等于 −2 时，排除连锁。LOD 得分最大时的 θ 值被接受为最大似然估计值。由于现有的 LIPED(http://linkage.rockefeller.edu/ott/liped.html)、LINKAGE(http://linkage.rockefeller.edu/soft/linkage/)、S.A.G.E.(http://darwin.cwru.edu/sage/)等自由软件包提供了包括 LOD 值法在内的多种参数连锁分析工具，这里对具体的算法不再展开。由于早期的连锁分析方法对模型的依赖性较强，主要适用于单基因病，计算速度慢等原因，新的方法也在不断的开发，如“混合模型”方法、多位点连锁分析方法、基于仿真的吉布斯取样及蒙特卡罗方法等。

参数连锁分析方法已经被应用于几百种孟德尔遗传病的遗传定位研究中，同时也在某些复杂疾

病研究，特别是大家系研究中获得成功。当然，实际的疾病家系非常复杂，所以在研究中还应该注意一些特殊的情况：①如果在特定的家系中难以获得明确的连锁关系，还可以收集大量的家系资料进行分析，但并不是说连锁分析结果在某些家系中出现阳性结果就可以忽略阴性结果的家系，背后可能还存在更复杂的遗传机制。同样，在实验样本获取部分曾经提出五个基本的原则，参数连锁分析家系选择过程中也可以考虑以上的因素，做出合理的家系筛选。②对于某些外显率并不明确的疾病，还需要对外显率进行估计，而采用疾病个体特异的分析策略，将无病个体设置为表型未知个体也是一种有效的分析方法。③家系中某些个体的疾病表型并不典型，难以确定是否受累，如某些精神疾病。这时就需要进一步严格疾病定义，将出现某一特定的表型作为诊断的标准，或放宽标准，只要出现疾病某一典型表现即定义为受累。

（二）非参数连锁分析方法

非参数连锁分析是一种在分析前不需要确定疾病遗传模式（如基因型频率、外显率等）或半依赖模型的分析方法。最常用的非参数连锁分析方法是等位共享方法。等位共享方法不依赖于遗传模型的构建，而是一个排除模型的过程。通过显示受累亲属间高于随机情况的共享遗传相同的染色体区域（或位点）概率来证实染色体区域的遗传模式与孟德尔遗传之间的差别。由于等位共享的方法是一种非参数方法，比参数连锁分析方法有更宽泛的应用范围，即使在受累亲属中不完全显性、表型复制、遗传异质性和高频等位等影响因素存在时，也有较好的表现。唯一的缺陷是等位共享方法提供的结果，一般说来没有参数连锁分析方法显著。

等位共享方法研究家系中亲属在共享来源于同一祖先的特定染色体区域或位点的频率，把这种区域或位点也叫做血源一致性（identical-by-descent，IBD），然后将某个位点共享 IBD 的情况与随机进行比较。通常，可以构建一个血缘一致性受累家系成员（identity-By-Descent Affected-Pedigree-Member，IBD-APM）统计量

$$t(s)=\sum_{i,j}x_{ij}(s) \qquad \text{式 15-4}$$

式 15-4 中，$x_{ij}(s)$是指家系中第 i 个和第 j 个亲属在染色体位点 s 处共享 IBD 的个数，加和指的是这个家系中所有亲属对在 s 处共享 IBD 的个数。如果是多个家系的组合研究，那么可以加和成 $T(s)$。在随机分离状态下，$T(s)$趋于均值为 μ，标准差为 σ 的正态分布，μ 和 σ 可以通过计算血缘系数（kinship coefficient）获得。当统计量 $(T-\mu)/\sigma$ 超出了设定的阈值，就可以判定此时的状态与随机分离相偏离，从而得到阳性的结果。

在等位共享分析中，最简单的一种形式是同胞对（sib pairs）分析，同胞对共享 IBD 数为 0，1 或 2（随机情况下，共享频率分别为 25%、50%、25%），（图 15-10），可以采用简单的 χ^2 检验分析疾病状态下的等位共享情况。这样的方法同样可用于受累叔侄对、表兄弟对的研究。

IBD 之外，还有一个与之相似的概念状态一致性（Identical-By-State，IBS）。IBS 用来描述亲属对之间共享同一等位（不区分是否同一祖先来源）的频率。两者的基本分析方法是相通的，但采用 IBS 方法可以避免 IBD-PAM 分析过程中对 IBD 的估计过程，因此应用也非常广泛。随着遗传标记分型技术，特别是 SNP 分型技术的进步，IBD 和 IBS 方法也逐渐应用于基因组范围关联研究中。

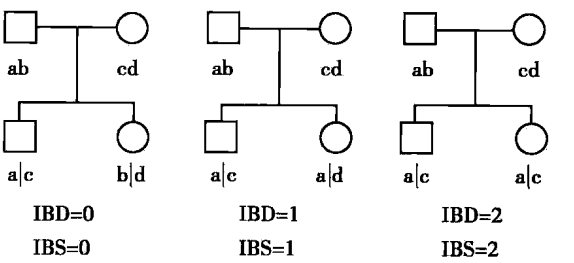


图 15-10 同胞对血缘一致性和状态一致性示意

三、关联研究发现疾病风险 SNP

关联研究（association study）是不依赖于家系信息的一种遗传定位策略，由于资源丰富，分析方法简便，是目前遗传定位研究中最常用的分析方法。关联研究通过检验某个特定的等位在疾病组和对

照组中出现的频率差异来判断此等位是否是疾病易感等位。就 SNP 而言，发现风险 SNP 的过程可以采用四格表 χ^2 检验进行等位频率分析，也可以采用 $2\times3\chi^2$ 检验进行基因型分析。采用简单的统计方法对 SNP 与疾病关联性进行分析，简捷性显而易见。但关联研究也有明显缺点，即对对照组样本选取具有严格的限制，此外，由于关联研究可能针对任何一个分子标记进行，而不存在先验的假设，对关联研究发现的风险 SNP 尚需要进行可靠的功能验证。由此可见，关联研究中对标记信息的分析比研究方法本身更重要，下面将从关联研究机制上来探讨风险 SNP 发现应注意的问题。

关联研究中发现 SNP 与疾病发生之间的显著相关性可能存在三个原因：①SNP 本身就是一个致病的 SNP，图 15-11(A)；②SNP 本身不能导致疾病，但与导致疾病的基因处于连锁不平衡状态，图 15-11(B)；③研究群体选择失误造成的统计显著性。第三种情况是关联研究过程中需要避免的，所以关联研究过程中还应注意三点：①关联分析的样本选取要严格限制在同质性群体中；②关联研究对照组选取应当谨慎，必要时选择未受累亲属作为内对照；③如条件允许，对获得的阳性位点可进行传递不平衡检验(Transmission Disequilibrium Test，TDT)来确认发现的致病等位在家庭遗传中倾向于向患病子代遗传。

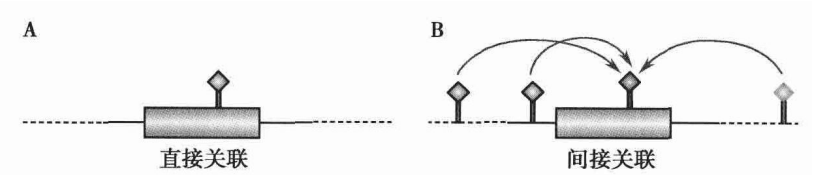


图 15-11 显著关联潜在的生物学机制

由于复杂疾病发生过程中，存在遗传位点间的相互作用，单个位点的关联分析方法有时不能获得足够的信息来发现某些区域与疾病之间的关联性。基于单体型、罗杰斯特回归、主成分分析、随机森林等统计学和机器学习方法的遗传定位方法成为有用的研究手段，得到了比较广泛的应用。

总起来看，关联研究和连锁分析有很多重要的区别。关联研究检验疾病与等位频率在群体中是否存在相关性，连锁分析检验疾病与位点是否在家系中共同传递。当群体中致病因素是多样的，而且致病位点相互独立，散在存在的时候，每个位点与疾病关联都将很弱，遗传定位中往往只能检测到连锁而难以发现关联；相反，当致病位点等位效应较弱，对疾病贡献较小时，但在疾病个体中有较高的等位频率时，基于家系的连锁分析难以发现潜在的传递模式，而关联研究却能识别出这种致病位点。因此，关联研究和连锁研究本身并不存在孰强孰弱，而需要考虑实际解决的问题进行选择。

四、遗传分析中的统计显著性

遗传分析方法虽然笼统的分为两类，但相应的研究方法众多，既有传统的统计分析方法，也有衍生而来的机器学习方法，无论采用何种方法进行复杂疾病的遗传分析，最终都将面对统计结果的取舍问题，即如何进行统计显著性的阈值设定。而且，这个问题还将因为遗传分析中分子标记的增多或检验模型的增加而变得更为严峻。

在进行 SNP 与疾病之间的连锁或关联分析时，要设置一个可以接受的假设检验显著性水平 α(一般为 5%)。这样，每一次检验，都有 5% 的可能引入一个假阳性的结果(Ⅰ类错误)。当进行 n 次独立的连锁或关联检验时，引入的Ⅰ类错误水平将满足 $\alpha=1-(1-\alpha')^n$，当 n 变大时，引入的假阳性结果也将增多，从而使得在进行数以千计的 SNP 关联或连锁分析时，需要对 α 进行 Bonferroni 校正 $\alpha'=\alpha/n$。在这种情况下，如果对 1000 个 SNP 进行检验，且要达到显著性水平 $\alpha=0.05$，需要达到真实的显著性水平为 $\alpha=5\times10^{-5}$，而 100 万个 SNP 进行检验时，所需要达到的真实显著性水平为 $\alpha'=5\times10^{-8}$，这对于高维度 SNP 遗传定位是个灾难性的结果，直接导致单次关联或连锁分析所能获得的显著性结果极少，一方面许多真正相关的 SNP 没有被发现，造成了很大的假阴性，另一方面在发现的极少的显著性结果中依然存在着较大地假阳性。

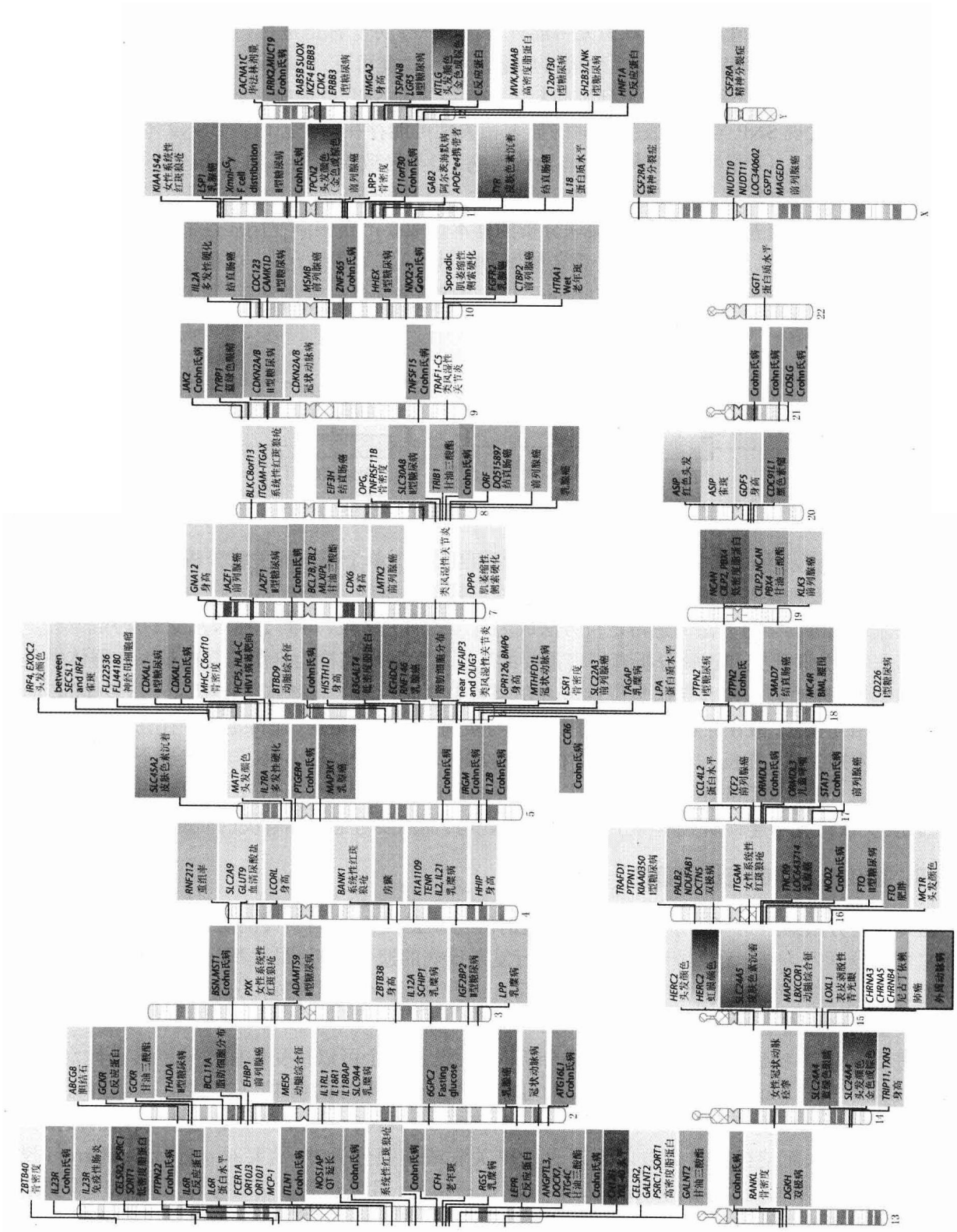


图 15-12 基因组范围关联研究取得的成效

因此，对于遗传定位的结果取舍，特别是多重检验问题一向都是人们关注的重点，采用多次随机 SNP 与疾病相关性检验进行显著性水平选取是目前为回避多重检验校正而广泛采用的一种方法。另外，考虑到基因组中广泛存在的连锁不平衡问题，对待检的 SNP 进行 LD 修正是降低多重检验校正影响的一种有效方法。此外，在芯片分析中采用的 FDR 方法也经常用于遗传定位结果的修正。

五、基因组范围关联研究与系统生物学方法在医学中的应用

随着 HGP 和 HapMap 计划的开展和完成，已识别的人类 SNP 已达到千万，常见 SNP 数量也已经达到 300 万以上，同时 HapMap 计划推动的商业分型芯片发展，已经促使遗传定位研究由最初的几个至数千个分子标记的研究发展到当前 50 万~100 万 SNP 的研究维度，极大地推动了复杂疾病风险定位的研究，遗传分析已经进行了基因组范围关联研究(genome-wide association study，GWA)阶段。目前，基因组范围关联研究已经应用于 40 多种复杂疾病的研究，绝大多数的研究涉及的 SNP 数目已经超过 50 万，并通过 GWA，成功获得了 150 多个致病基因(图 15-12)。这些疾病基因的获得对于复杂疾病，特别是癌症、糖尿病、心脏病等常见病的研究提供了大量的有用信息，也为进一步揭示这些疾病的发生机制做出了贡献。真正意义上的 GWA 开始于 2005 年前后，应该说，现在还只是起步阶段，大规模的 GWA 研究还在酝酿，相应的研究策略也在不断地开发。

但高维度的 SNP 数据也给统计学方法带来了很大的压力，多重检验问题困扰着大规模的遗传定位研究。目前，基因组范围关联研究主要通过两个策略来实现风险 SNP 和风险基因的发现。一方面，采用合并不同实验室样本数据的方法，通过提高研究某个疾病的样本量来加大风险 SNP 的显著性水平，即人们常说的 meta 分析方法，并且成功的应用于乳腺癌、结肠癌和 2 型糖尿病研究中，与之相伴的是基因型推断技术的发展。另一方面，采用候选区域精细定位的方法，在较低样本量情况下采用基因组范围关联分析获得候选风险区域，缩小范围后对候选区域加大样本量，进行精细的 SNP 分型，采用多轮重复策略，最终获得高显著、高精确度的风险位点(图 15-13)。这些策略的实施为发现真实的风险 SNP 提供了可靠的保障，但依然存在花费大、效率低的缺点。

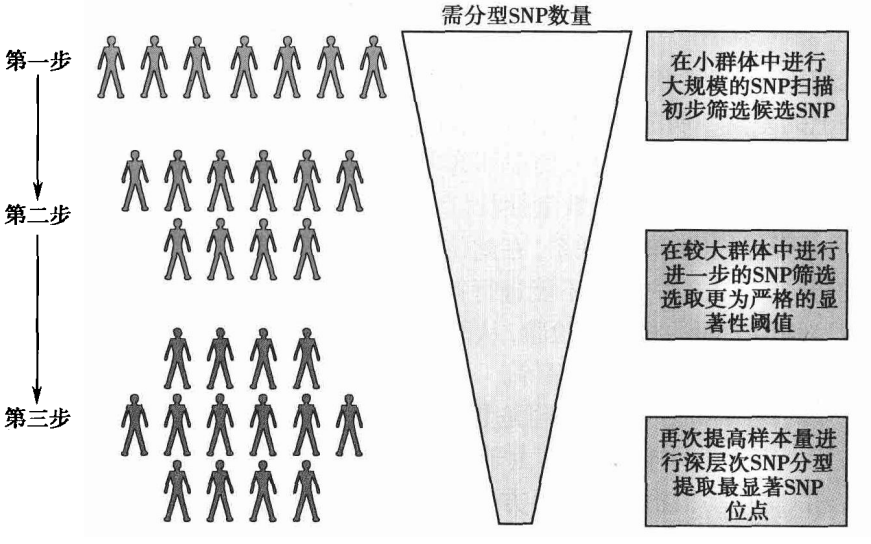


图 15-13 精确定位策略提高关联分析可靠性

在这样的情况下，人们逐渐将目光从统计方法研究和提高统计显著性角度转移到关联分析结果的信息挖掘上，称之为第二代关联分析策略。第二代关联分析策略将关联分析作为疾病风险权重，期望借助于已知的通路、网络、互作、功能等知识进行位点和基因层面之外的更高层次的信息发现。这样的策略不仅坚持了疾病基因层面的发现，同时获得的结果还能够从细胞过程和机制的角度来解

释疾病的发生，相比原有的方法，有着不言而喻的优势。由于作为研究基础的高通量先验知识本身还存在不完整和假阳性，因此第二代关联分析策略还处于起步和摸索阶段，真正意义上的高层面信息发现还需要对现有知识进行深入、系统的梳理和总结。

第四节　数量性状研究与 SNP 的系统遗传学分析

Section 4　Quantitative Traits Analysis and System Genetics Methods in SNP

人与人之间形态、生理指标、行为及疾病易感等表型差异共同构成人类本身的多样性。这些复杂表型的变化往往是由潜在的多位点遗传复杂性及其遗传等位与个体所处环境之间的不同反应造成的。从 DNA 变异与表型差异之间的相关性研究角度，讨论数量或复杂性状产生的原因对于预测疾病发病风险和个性化治疗有重要意义。与某些数量或复杂性状形成相关的 DNA 区域被称为决定这个性状的数量性状位点(quantitative trait loci，QTL)。早在 20 世纪早期，人们就开始了对数量性状的研究，并采用遗传多态标记与 QTL 连锁分析的思想对数量性状进行遗传定位。到 20 世纪 80 年代，数量性状研究得到了空前的发展，但是遗传多态标记的缺乏大大限制了它的进一步发展。直到最近几年，随着测序技术的发明和人类单体型图计划的实施和完成，大量遗传标记被发现，而且分型成本不断降低，基因组范围数量性状的 QTL 定位研究迅速发展，并广泛地应用于人类性状和疾病研究领域。

经过二十多年的努力，人们已经能够从候选基因、不同遗传背景下的等位分离、生态与环境对表型影响、功能等位效应的分子基础、群体致病等位频率等方面对遗传变异与数量性状形成之进行解释。某些研究通过 QTL 定位发现了新的疾病或复杂性状位点，并为揭示疾病生物学机制提供新的视野，但明确指出的导致表型和疾病形成的变异，只占全部表型决定子的一小部分，通过 QTL 定位直接发现表型相关的基因更是少之又少。不过，这一情况并不取决于目前对 QTL 定位的研究方法，而是与现在的 DNA 和 RNA 的测序水平相关的，将会伴随新的高通量、快速、低廉的测序技术的产生而取得新的突破。

一、数量性状定位的基本思想

与质量性状相比，数量性状的遗传研究要困难得多，主要是由于质量性状可以通过表型来辨别，而数量性状表型上的差异不明显，基因型与表型间难以找到准确的对应关系。由于人类群体不可能像在动植物中进行杂交实验，所以对人类群体的数量性状定位更加困难。无论是在人类还是在动植物中，数量性状定位的基本原理都是数量性状位点与可见的分型分子标记之间存在遗传连锁。如果某个 QTL 与某个分子标记(SNP)相联系，在此位点上具有不同等位的个体具有不同的数量性状平均值。基于这样的思考，在人类中虽然不能进行特定的位点杂交实验，但可以通过遗体遗传学方法进行位点与数量表型均值之间的相关性检验，从而完成数量性状定位。常用的数量性状定位分子标记除 SNP 外，还有插入 / 删除多态、微卫星等。

显著相关位点的检测和原因基因克隆是数量性状定位的两个要点。图 15-14 显示了基于 SNP 的 QTL 分析基本过程。在人类样本分析中，由于家系信息难以获得，主要通过关联分析的方法进行检验(图 15-14 右侧)，相对的，动植物研究中可方便地进行杂交实验，一般采用连锁分析的方法(图 15-14 左侧)。

人类遗传学中进行数量性状定位最常用的方法是线性回归和方差分析。方差分析进行数量性状定位类似于自由度为 2 的皮尔森检验，这里，将 0 假设定义为数量性状与 SNP 基因型没有相关性，备选假设为有相关性。而线性回归方法用于数量性状研究主要考虑 SNP 基因型与数量性状平均值之间的关系，自由度为 1(图 15-15)。两种情况下均要求数量性状呈近似正态分布，如果分布有偏差，可以考虑进行对数转换。

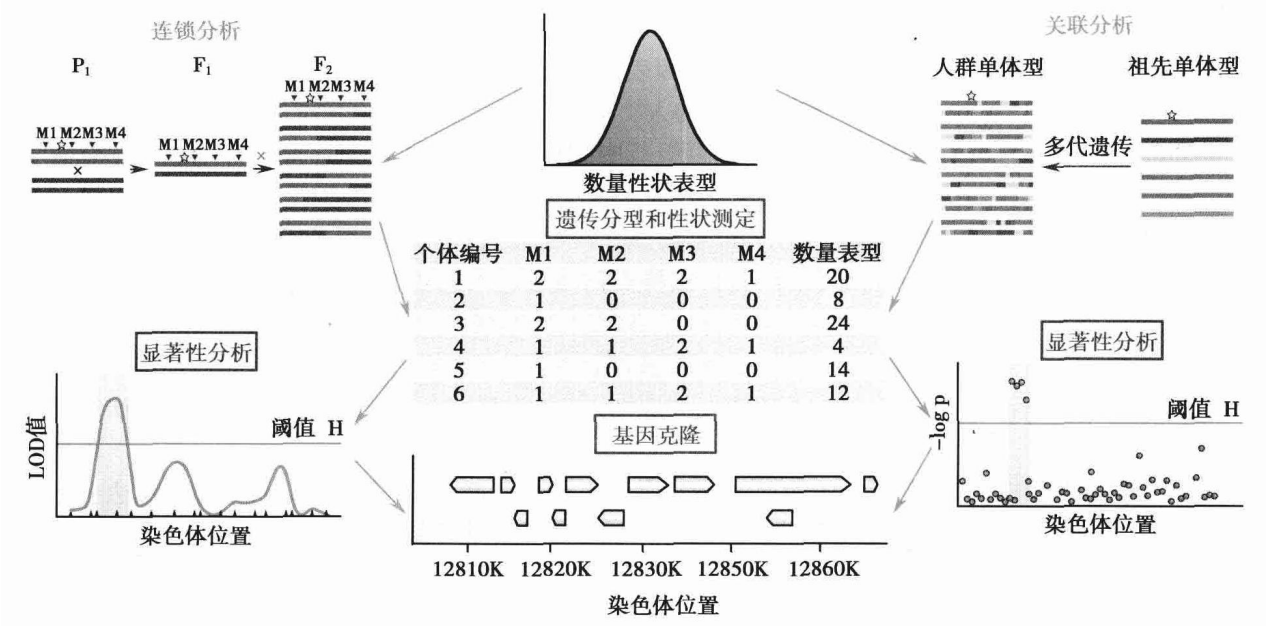


图 15-14 数量性状分析流程图

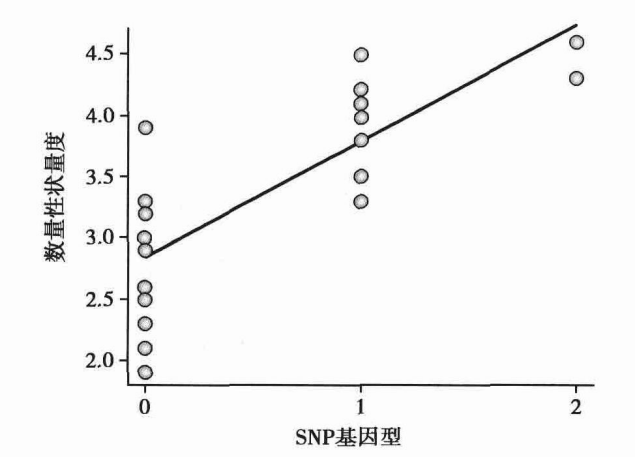


图 15-15 线性回归方法分析数量性状与 SNP 相关性

二、系统遗传学思想研究复杂性状

对于常见的数量性状，人们可以很自然的联系到身高、体重等看得见的研究对象，除此之外，还能够想到血压水平、血糖水平等与人体健康检查有关的症状反应或生化指标。随着 SNP 分型技术和基因表达作谱技术的发展，越来越多的研究把目标锁定在人类基因表达调控子的发现上，通过对同质性样本的 SNP 分型及基因表达作图谱，期望建立分子标记与基因表达之间的联系，这一过程被称作表达数量性状位点(expression quantitative trait loci，eQTL)定位。2002 年，Rachel 等利用芯片表达技术与关联分析相结合的方法研究杂交酵母的遗传变异与基因表达之间的相关性。发现众多的基因表达是受遗传变异影响的，不同的遗传等位有可能导致不同的表达效果。这为从遗传角度揭示表型形成提供了一个有力证据。2004 年，Michael 等将 eQTL 研究引入到人类基因组研究领域，通过对 14 个家系 196 个个体的 2980 个 SNP 与 3553 个转录子表达的测定，发现 170 多个 SNP 与之邻近的基因表达之间存在相关性，并把这种关系称之为 Cis 关联，此外，还存在众多远距离 SNP 与基因表达之间的相关性(称之为 Trans 关联)，这一研究证实人类基因表达受到了广泛的遗传调控，而且可以通过数量性状定位的方法将遗传变异和表达进行关联。

目前，与基因组范围关联研究发展相适应，eQTL 研究已经从最初的数以千计的 SNP 与基因表达规模发展到数以十万计的 SNP 和 2 万多基因表达之间的关系，而且从基于家系和模式生物的研究逐渐过渡到基于不相关个体的研究，发现的人类遗传与表达之间的关系也越来越多。2007 年 10 月，Nature Genetics 连续发表了 3 篇人类基因组范围的 eQTL 研究文章：Barbara 等基于 HapMap 样本，进一步测定 14 000 个基因的表达情况，进行了四个群体的 eQTL 研究，从群体比较方面揭示遗传变异调控基因表达的群体差异性；Harald 等将基于淋巴细胞的研究样本量提高到 1240 个个体，研究的基因数高达 1.9 万；而 Anna 等的研究首次将疾病因素引入到基因组范围关联研究中，通过研究哮喘家系中的遗传变异与基因表达之间的关系，提出可能实现联合 eQTL 与疾病的研究，易化关联研究中的功能元件提取。2008 年 3 月，Valur 等联合基因表达、遗传变异及临床肥胖指标进行合并的 eQTL 研究，进行疾病相关的遗传子及功能元件识别，并在此基础上提出从分子网络的角度研究复杂疾病。

表达数量性状定位的提出为生物医学研究展开了更为广阔的视野，也为从 DNA→表达→分子表型→性状的研究提供了可能。在这样的背景下，科学家们提出系统遗传学(systems genetics)概念，即希望从全面的生物学资源出发，研究遗传因素对人体生理病理的影响。Trudy 等在此基础上提出未来的遗传定位研究的着眼点(图 15-16)，期望借助系统遗传学工具实现从分子到整体的全面了解。而表达与标记、表达与表达之间的相关又能向网络的层面进行转化，对于获取生理学或病理学功能信息将产生直接而有效的影响。

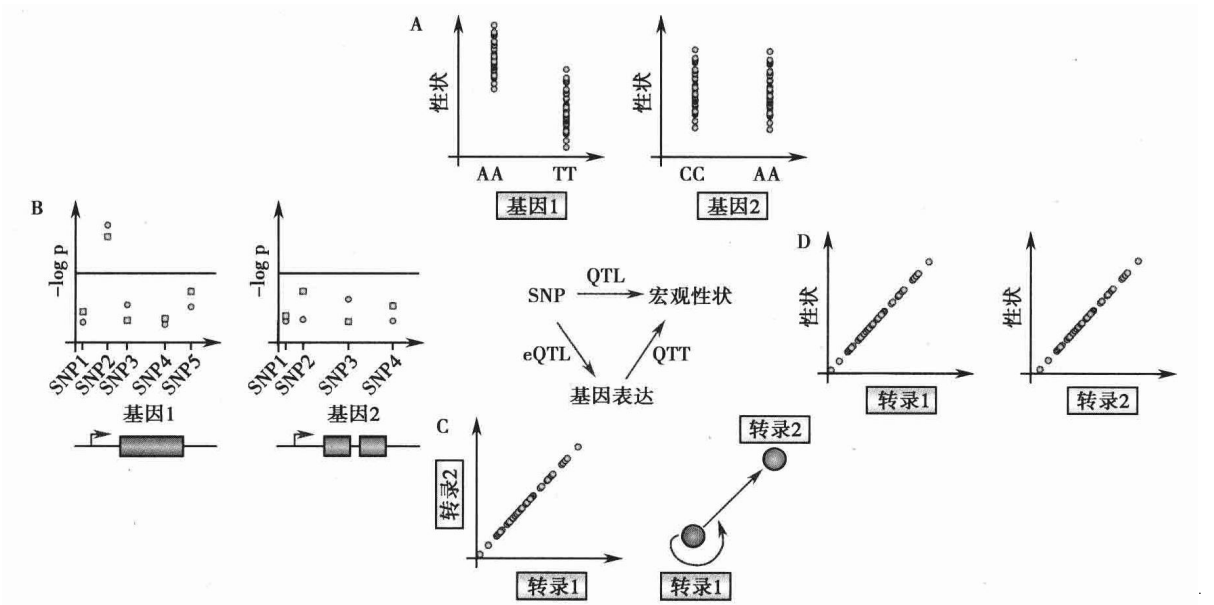


图 15-16　系统遗传学思想构建遗传调控网络揭示性状形成机制

图 15-16(A)中显示了广泛应用的纯合基因型与数量性状平均值线性回归方法获得与性状相关的遗传标记(SNP)的过程，即人们常说的 QTL 定位，将一个宏观的指标与分子层面的标记相互联系。B 图展示的是基因表达与 SNP 之间的关联分析过程，即上文讲到的 eQTL 定位，从而捕获影响基因表达的 SNP，这些 SNP 称为调控 SNP，这个过程将基因表达和 SNP 进行关联。C 图进行的是基因共表达分析(详见第十二章)。通过 B、C 两图可以构建出一个基于遗传分析的调控网络，而由于已经在 SNP 与表型之间建立起联系，借助分子网络的分析手段，能够发现影响性状形成的基因集，甚至发现与性状发生密切相关的网络模块或通路。目前，由于同时进行基因型、基因表达、人类表型的测定和收集过程耗时、耗力，而且花费巨大，在一定程度上限制了系统遗传学的开展，但随着技术的革新，这样一种研究思想将逐步成为人类生理、病理研究的必由之路。

三、变异组学的研究现状

SNP作为最简单的变异形式得到广泛的关注，事实上染色体上的多态性也是组学层面的问题，这里简要地介绍一下人类染色体中其他的遗传变异，涉及最简单的变异形式插入/删除多态(In/Del)、关系碱基数量最大的多态拷贝数变异(copy number variants，CNV)、早期应用的遗传标记微卫星(microsatellite，MS)等。

如图15-17所描述的人类染色体上的各种遗传变异，以1kb长度为界，将遗传变异分为两类，一类自身影响的范围比较小，是包括SNP在内的序列变异，另一类是从微卫星和插入删除多态起到长重复片段的结构变异，更大的染色体变化将之称为染色体畸变，也是遗传学研究中的重要范畴，这里不展开介绍。

微卫星多态目前已发现5000余个，是早期遗传定位研究中非常重要的分子标记，也与癌症等多种疾病的稳定性有关。已经发现的人类插入删除多态已达到586个，这些多态最长能达到70kb，在多种疾病，特别是精神病发生过程中有重要的作用。CNV目前已经识别了1447个，涉及360Mb的染色体范围，占人类染色体总量的12%，是影响核苷酸数最多的变异形式。由于CNV本身的长度超过100kb，能够直接引起基因拷贝数、调控区段的变化，因此对于生理病理有着重要的影响。变异组学的研究证据不断地告诉人们，人类染色体中还有着巨大的未知的秘密，既决定了人类种族的一致性，又决定着人类多样性的产生，由于它们的存在，这个世界变得绚烂多彩，同样由于它们存在，人们对人生的感悟又有所不同。真正全面了解这些遗传变异在人类生理病理中发挥的重要作用，才能够实现从系统的角度揭示人类生命的本质。

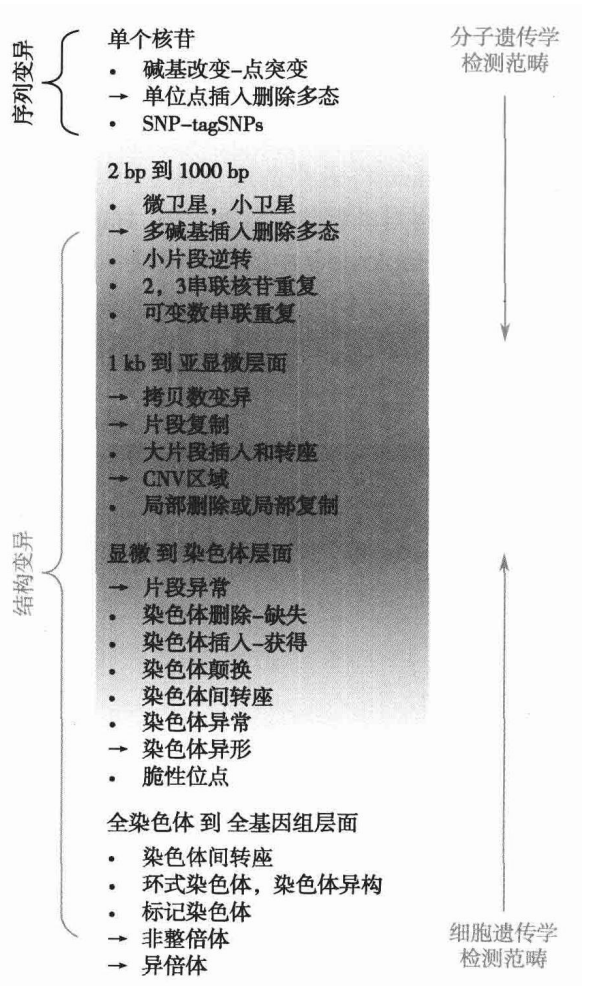


图15-17 人类染色体上的序列和结构变异

第五节 SNP相关的集成软件工具

Section 5 Important Tools in SNP Studies

一、Haploview识别TagSNP及推断单体型

(一)软件介绍

Haploview软件由剑桥大学编写，http://www.broad.mit.edu/mpg/Haploview/index.php，并免费提供给用户使用。它具有以下几种分析功能：①连锁分析机单体型block分析；②单体型群体频率的估计；③单位点SNP或单体型的关联分析；④随机扰动检验结果可靠性；⑤TagSNP推断。可视化的Haploview可以分析数千个样本中的上万个SNP数据，DOS版本可以不受数量的限制。Haploview通过三种不同的方法分析基因型，得到不同的单体型及单体型块，并且分别对不同方法所得到的不同的单体型块进行连锁分析或关联分析，得出群体中传递频率高的单个等位基因及单体型在每代之

间的遗传频率，从而找到与疾病相关性最大的等位基因，单体型及高度连锁分析的等位基因。

（二）数据的统计描述

当载入与 Haploview 对应格式的数据文件时，Haploview 对数据信息进行筛选。对所载入的数据，Haploview 设置了一些指标的阈值来选定具有特征性的 Marker。当数据的 SNP 满足以下默认条件时：

哈代 - 温伯格平衡检验：$HWE \geq 0.001$

最小等位基因频率：$MAF \geq 0.001$

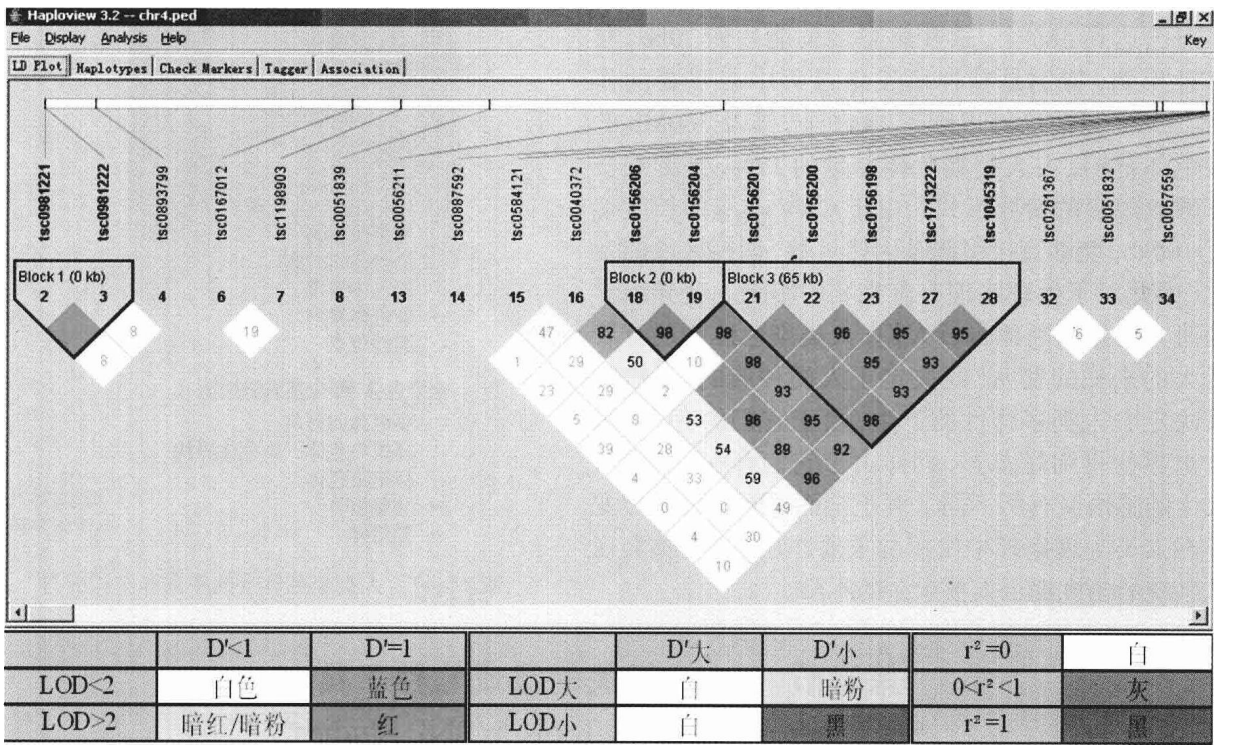
未缺失的基因型频率：$\%Geno \geq 75\%$

孟德尔遗传规律错误的个数：$MendErr \leq 1$

Haploview 将最终选取这些 SNP 进行有效的分析。而对于那些未满足这些衡量标准的 SNP，Haploview 将自动删除。除此之外，Haploview 不会对未被完全检验的 Marker 进行后续的分析及检验。

（三）Haploview 中的分析模块

1. 连锁不平衡分析　在连锁不平衡分析中，Haploview 通过对输入数据的处理，计算 SNP 之间的 LD 量度（D'、r^2、LOD 值），并用不同的颜色表示不同标记之间的连锁不平衡强度，实现可视化（图 15-18）。



	D'<1	D'=1		D'大	D'小	$r^2=0$	白
LOD<2	白色	蓝色	LOD大	白	暗粉	$0<r^2<1$	灰
LOD>2	暗红/暗粉	红	LOD小	白	黑	$r^2=1$	黑

图 15-18　连锁不平衡分析界面及其对应的颜色说明

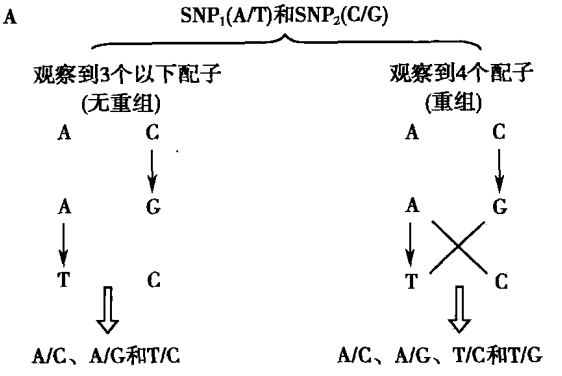
2. 单体型块分析　Haploview 提供了三种不同的方法进行单体型块估计，可以自动产生单体型块，也可以由用户手动自定义形成。

（1）致信区间法（confidence interval）：根据 Gabriel 等的研究，利用 D' 值来判定 SNP 的连锁不平衡程度即高度连锁不平衡，不确定性或高度重组。高度连锁不平衡及高度重组的 SNP 被认为是富含信息量的，即非不确定性的。当 95% 的富含信息量的 SNP 是高度连锁时，则认为这些 SNP 形成一个单体型块。但是首先必须保证这些 SNP 的最小等位基因频率 $MAF \geq 5\%$。

（2）四配子检验（Four-Gamete Test，FGT）：FGT 是通过突变体的重组关系来确定单体型的连锁块。每对 $MAF \geq 1\%$ 的 SNP，当发生突变时，由于重组关系可能形成 4 种由两个 SNP 形成的单体型。

当每个单体型出现的频率 $f \geqslant 0.01$ 时，认为发生了重组。这种情况下，单体型的连锁块由连续的不发生重组的连锁块组成，在这个块中，每个单体型都只可能存在三种组合。

在 m 个 SNP 中，人们对每对 SNP 进行 FGT 来识别过去的重组事件。如图 15-19 所示，图 15-19(A)考虑两个 SNP(SNP_1 和 SNP_2)，每个 SNP 有两个等位 A/T 和 C/G。在群体事件中，A 转变为 T，C 转变为 G，从左边可以看到两个 SNP 形成的三种单体型(A/C，A/G 和 T/C)；右边则因为发生重组，使两个 SNP 之间形成了四种单体型(A/C，A/G，T/C 及 T/G)。图 15-19(B)是单体型块的识别(此时 $m=8$)。用 0 和 1 表示是否在某两个 SNP 间形成四种单体型，确定每个块中包含所有不发生重组的单体型。当出现重组即有标志 1 出现时，就把它认为是另一个块的开始，直到下一个标志 1 出现为止。



B

	1	2	3	4	5	6	7	8
1	0	0	0	1	1	0	1	1
2		0	0	1	0	0	0	0
3			0	1	0	0	0	0
4				0	0	0	1	1
5					0	0	1	1
6						0	1	0
7							0	0
8								0

图 15-19 四配子检验方法识别单体型块的基本原理

(3) 连锁不平衡的稳定连接限制：在连锁不平衡计算过程中，Haploview 寻找两个 SNP 之间高度连锁不平衡的证据，将这两个 SNP 作为连锁块的头尾两个 SNP。即在确定的块中，头尾两个 SNP 必须与中间所有的 SNP 高度连锁不平衡，而中间的 SNP 间则不必要连锁不平衡。

3. 单体型分析 Haploview 运用最大期望算法(expectation-maximization algorithm，EM)对未知位置的单体型进行位置预测。EM 运算法则是在哈代 - 温伯格平衡(Hardy-Weinberg equilibrium，HWE)条件下计算连续的单体型频率的方法。

如图 15-20 所示，Haploview 提供的单体型窗口显示了块中每个单体型的群体频率及单体型块之间的关系。在交叉区域中，计算出了多等位基因的 D' 值，反映了块之间的重组水平；同时显示单体型中的 SNP 标识及 Tag SNP。

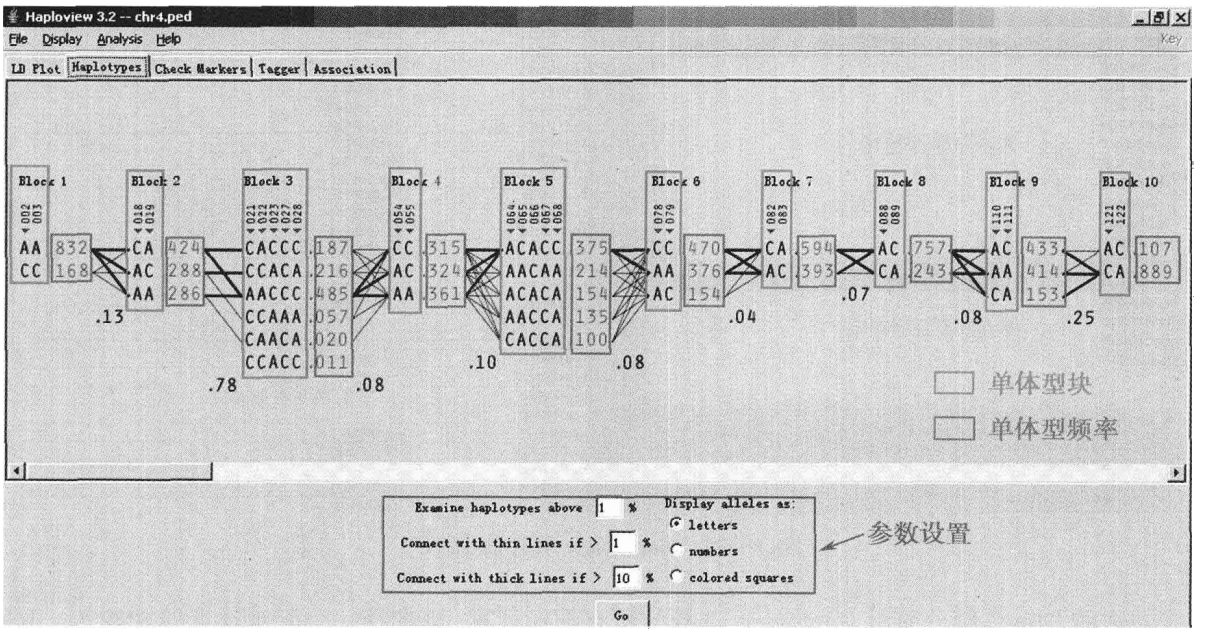


图 15-20 Haploview 估计单体型类型及频率

单体型也可以通过调整以不同的形式来表现。在窗口的底部，有几种选择：等位基因，数字，及红色和蓝色框架。其中，等位基因用 A、C、T、G、X 来表示，X 代表不确定的数据；数字 1、2、3、4 分别表示碱基 A、C、T、G，8 则表示 X；蓝色框架代表最多的等位基因，红色则代表最少的等位基因。

TagSNP 能够代表整个单体型，也可以通过 TagSNP 在新的样本中发现新的变异。在一个连锁块中将 SNP 排序并选择群体频率大于 1% 的单体型，从而保证了四配子方法获得的连锁块的有效集合，但是在发生重组的块中偶尔会产生冗余。

4. TagSNP 分析　Haploviewk 中的 Tagger 模块能够选择等位基因高度连锁不平衡的多等位基因组合，并准确记录可以用于关联性检验的等位基因。Tagger 给定了一些阈值，选择连锁不平衡系数 r^2 大于阈值的 SNP 集合。Tagger 提供了两种分析方法：配对法和捕获法，这里主要介绍一下捕获法。捕获法提取 Tag SNP 包括两个步骤：①获得在配对法中没有被捕获的 SNP，因为这些 SNP 除了与自己的关联性比较大之外，不能与其他的 SNP 配对；②通过取代有多个待检 SNP 的标签来制定标签目录。Tagger 用 SNP 建立多标记检测集，这些 SNP 彼此之间高度连锁不平衡，其配对 LOD 值须大于 3.0。但是，这里的 LOD 值也可以根据需要进行调整。这里的 LOD 值与连锁分析中的 LOD 略有不同，是指在一定重组率(θ)条件下，两个位点相连锁的似然性和不连锁的似然性比值的对数值。

由于涉及标记名称，Tagger 分析要求有信息文件与其相对应。Tagger 的设置平台有三个选项：

Force Include：选择作为 TagSNP 的 SNP

Force Exclude：选择不作为 TagSNP 的 SNP

Capture this Allele：选择想要被 test 捕获的 SNP，则可以得到捕获这些 SNP 的 test

如图 15-21 所示，Tagger结果平台包含所有SNP的目录，捕获这些SNP的test及它们之间的r^2值。其中，未被标记的 SNP 用红色表示，未选择被捕获的 SNP 则显示灰色。

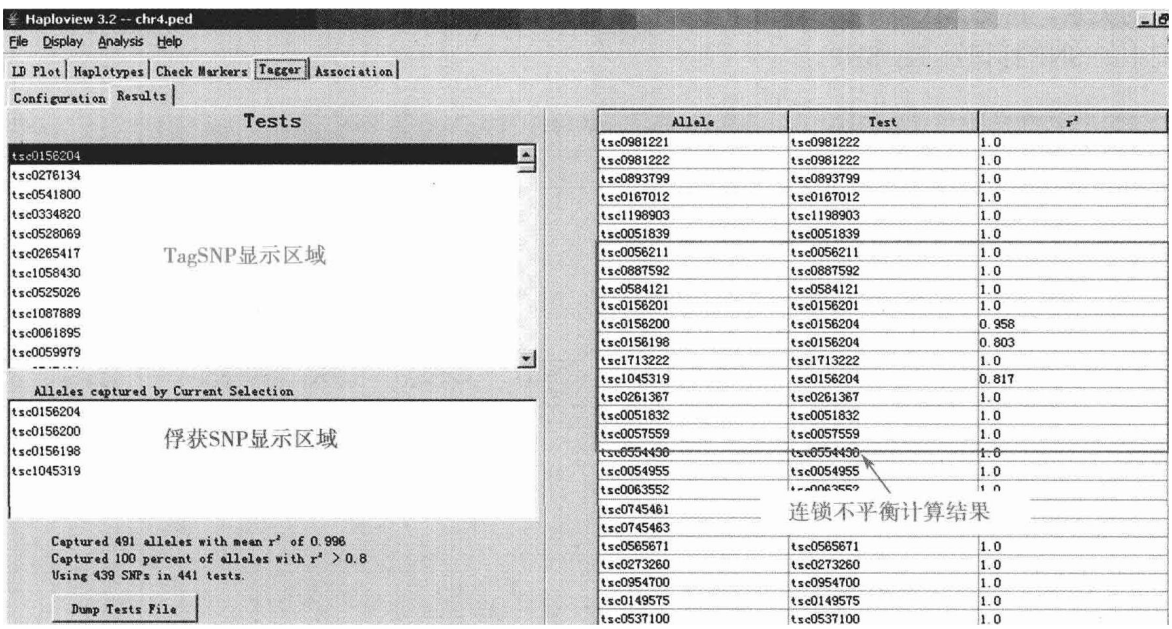


图 15-21　Haploview 识别 Tag SNP

左边的“Test”部分显示 Haploview 选择的所有配对计算，当选择一个配对计算的 SNP 时，下边将给出这个 test 所捕获的等位基因。左下角列出了标签计算的概括：

Captured N alleles with mean r^2 of X.：数据中 N 个 SNP 被所选的 test 标签，这些 SNP 的 r^2 平均值为 X。

Captured N percent of alleles with $r^2>0.8$：被标签的等位基因中 $r^2 \geqslant 0.8$ 的百分数。

Using N SNPs in M tests.：选定 N 个 SNP 建立 M 个 test，这些 test 可以为单个 SNP，也可以是几个 SNP 的组合。

5. 关联研究 Haploview 还可以对数据进行单个基因座及多个基因单体型关联性分析。根据选择，对核心家系及普遍群体进行分析。对于核心家系，计算所有先验者的传递不平衡检验(TDT)值。对群体而言，Haploview 分别计算每个等位基因频率在疾病 - 对照的卡方统计量和 p 值。在计算中，Haploview 将每个单体型的可能分数相加，就可以得到 TDT 值及受累 / 健康的关联性检验的数值。也就是说，如果单个个体含有单体型 A、B 的可能性分别为 60% 和 40%，0.6 和 0.4 将分别加到单体型 A、B 的计数中。所以 Haploview 提供了关联性检验的直接方法，另外，还可以通过载入由标签分析和单体型分析的结果，自定义需要进行关联性检验的等位基因及单体型。

6. Permutation 检验结果稳定性 Haploview 可以对关联性结果进行 permutation 随机扰动试验，衡量关联性检验结果中比较有意义的分析。在进行大量的 permutation 后，Haploview 显示 permution 的分布图，所有有意义的配对计算目录及它们的卡方检验及随机扰动的统计意义。

二、基因组范围关联研究软件包 SNPtest

SNPtest(http://www.stats.ox.ac.uk/%7Emarchini/software/gwas/snptest.html)是一个强大的基因组范围关联研究软件包，它可以对单个 SNP 关联进行频率检验或贝叶斯检验，值得注意的是，它的实施只适合于二进制(病例对照)性状，但该软件可以根据任意的协变量集进行设置，并且能够考虑基因型的不确定情况。目前，被广泛应用的 WTCCC 中，7 套复杂疾病的基因组范围关联研究，就是采用该软件进行的数据分析。SNPtest 同时提供了 2000 个个体中 100 个 SNP 的疾病 - 对照示例文件，与 SNPtest 具有同等价值的关联软件包还有 Plink 等(这里不进行介绍)。

(一) 软件的输入文件

SNPtest 允许分析多组个体。每组数据存为两个文件：第一个文件为样本文件，存储的是 ID 号、关联协变量和每组个体的表型信息；第二个文件为基因型文件，存储的是每组基因型数据。软件当中包括的例子数据集中每组的样本和基因型文件分别有符合要求的 _sample 和 _gen 样文件。

基因型文件格式(_gen)：该文件每行表示一个 SNP 信息，前 5 列分别为：SNP ID、RS ID、SNP 碱基对位置、两个等位基因(M、N)；接下来的 3 个数字表示三种基因型 MM、MN、NN 在第一个个体中出现的概率值，再接下来的 3 个数字表示三种基因型在第二个个体中出现的概率，以此类推。并且个体的顺序应该与 _sample 文件中个体的顺序相同。同时，考虑到缺失基因型情况，因此基因型概率之和不必均为 1。

例如：已知 5 个 SNP 在 2 个个体中的基因型，转化为正确的 _gen 文件，如图 15-22 所示：

```
SNP 1 : AA CC          SNP1 rs1 1000 A C 1 0 0 0 0 1
SNP 2 : GG GT          SNP2 rs2 2000 G T 1 0 0 0 1 0
SNP 3 : CC CT          SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP 4 : CT CT          SNP4 rs4 4000 C T 0 1 0 0 1 0
SNP 5 : AG GG          SNP5 rs5 5000 A G 0 1 0 0 0 1
```

图 15-22 SNPtest 数据格式

红色线内的为个体 1，蓝色线内的为个体 2

当对多组执行 SNPtest 时，假设每组数据的 SNP 集大小相同并且这些 SNP 在每组的基因型文件中的存储顺序相同。

样本文件格式(_sample)：该文件包括三个部分，第一行，表示每一列的名字；第二行，表示每一列所存储变量的类型；接下来的每行表示一个个体的详细相关信息。例如：

ID_1	ID_2	missing	cov_1	cov_2	cov_3	cov_4	phenotype_1
0	0	0	1	2	3	3	p
1	1	0.007	1	2	0.0019	−0.008	1.233
2	2	0.009	1	2	0.0022	−0.001	6.234
3	3	0.005	1	2	0.0025	0.0028	6.121
4	4	0.007	2	1	0.0017	−0.011	3.234
5	5	0.004	3	2	−0.012	0.0236	2.786

第一行分别表示：个体的第一个 ID 号、第二个 ID 号、个体中缺失值的比例，这三个是必须要有的，接下来的分别表示变量的名字。上面的例子中，有 4 个协变量 cov_1、cov_2、cov_3、cov_4 和 1 个表型名字 phenotype_1。

第二行表示每列中变量的类型，前 3 个设置为 0，接下来的位置应遵循下面的规则：

1	离散的协变量(用正整数表示)，对关联进行 Mantel-Haentzel 检验
2	离散的协变量(用正整数表示)，对关联进行跨群体的整合检验
3	连续协变量
p	表型

(二) 软件包中的分析模块

1. 数据的统计描述　SNPTEST 最基本的用途是对 SNP 数据基本信息进行描述，生成包括基因型数目、等位基因频率、SNP 缺失数据比例和优势比等的描述信息，这个功能用以下命令行可以实现：

```
./snptest –cases ./example/cases.gen ./example/cases.sample –controls ./example/controls.gen ./example/controls.sample
–o ./example/ex.out
```

2. 哈代温伯格平衡检验　命令“-hwe”表示在输出结果中显示出每个 SNP 的 HWE 检验结果。例如：

```
./snptest -cases ./example/cases.gen ./example/cases.sample -controls ./example/controls.gen ./example/controls.sample
-o ./example/ex.out -hwe
```

将产生一个输出文件 ./example/ex.out，该文件的列包含的是对每个对照组的精确 HWE 检验的 p 值、对照组的整合集、每个病例组 HWE 检验 p 值、病例组整合集。

3. 基本的关联检验　病例对照检验：对加性、显性、隐性、常规及杂合子 5 个模型的关联进行标准频率病例对照检验，可由命令“-frequentist”来执行。例如，下面的命令行被用来对这四种模型进行检验：

```
./snptest -cases ./example/cases.gen ./example/cases.sample -controls ./example/controls.gen ./example/controls.sample
-o ./example/ex.out -frequentist 1 2 3 4 5
```

五种不同的模型编号为：1- 加性模型、2- 显性模型、3- 隐性模型、4- 常规模型和 5- 杂合子模型。加性模型是对加性遗传效应进行 Cochran-Armitage 检测，显性模型和隐性模型是将 AA 基因型当作起点基因型，常规模型则是对关联进行自由度为 2 的标准检验。

输出文件为 ./example/ex.out，包含了每个 SNP 所有如前面描述的梗概信息。四个检验的 p 值分别在 frequentist_add，frequentist_dom，frequentist_rec，frequentist_gen and frequentist_het 列中给出。

数量性状检验：对 SNP 与一个数量性状关联的检验可以用 -qt 命令来执行。对每个 SNP 的关联该命令是通过 F 检验来执行的。命令“-frequentist”被用来指定每个 SNP 的基因型编码。每个个体的基因型必须出现在样本文件当中。在默认情况下，检验将使用样本文件当中的第一个基因型。用户应当用“-pheno”这一命令来指定你所要检测的表型。例如下面的命令行，是对例子数据集中的第二个表型在五个不同模型中进行检验：

```
./snptest -cases ./example/cases.gen ./example/cases.sample -controls ./example/controls.gen ./example/controls.sample
-o ./example/ex.out -qt -pheno 2 -frequentist 1 2 3 4 5
```

4. 贝叶斯检验　用命令“-bayesian”可对五个标准遗传模型进行贝叶斯检验。例如，下面的命令行：

```
./snptest -cases ./example/cases.gen ./example/cases.sample -controls ./example/controls.gen ./example/controls.sample
-o ./example/ex.out -bayesian 1 2 3 4 5
```

产生一个输出文件，包含以下几列信息：bayesian_add，bayesian_dom，bayesian_rec，bayesian_gen and bayesian_het。

三、连锁分析和数量性状分析工具 Merlin

Merlin（http://www.sph.umich.edu/csg/abecasis/Merlin/index.html）是一个利用稀疏遗传树进行系谱分析的软件包。Merlin 利用稀疏树来代表系谱中的基因，它是最快的谱系分析软件包之一。Merlin 能够被用于参数或非参数的连锁分析，以回归为基础的连锁分析或对数量性状的关联分析，IBD 和亲属关系的估计，单体型分析，错误检测和模拟分析。在大部分分析中标记之间可以存在连锁不平衡状态，并且能够比其他的系谱分析软件包处理更多的标记。

Merlin 进行普遍的家系分析。输入文件描述数据集中个体之间的关系，储存了标记基因型，疾病的状况和数量性状标记信息，并提供了位点定位及等位基因频率信息。Merlin 支持 QTDT 或 LINKAGE 格式的输入文件。这两种格式非常相似，下面将主要关注 QTDT 格式。

（一）群体分层分析

虽然家系会变得非常复杂，在一个家系文件中所有用于重建个体间关系的信息可以概括为 5 个项目：家庭标识符、个体标识符、父亲标识符、母亲标识符及个体性别。

以下是一个虚拟的家系文件：

FAMILY	PERSON	FATHER	MOTHER	SEX
example	granpa	unknown	unknown	m
example	granny	unknown	unknown	f
example	father	unknown	unknown	m
example	mother	granpa	granny	f
example	sister	father	mother	f
example	brother	father	mother	m

这些关键值构成了任何一个家系文件的前五列。由于在早期的遗传程序中存在的限制，文本标识符通常被唯一的数值所取代。每个标识符被唯一的整数所替代且将性别编码为女性为 2，男性为 1 之后，一个基本的以空格分隔的家系文件会是以下这种形式：

<contents of basic.ped>

1	1	0	0	1
1	2	0	0	2
1	3	0	0	1
1	4	1	2	2
1	5	3	4	2
1	6	3	4	1

<end of basic.ped>

一个家系文件可以包括多个家庭。每个家庭都有唯一的结构，在数据集中与其他家庭之间存在独立性。

（二）表型与基因型

通常标准的 5 列之后的各种类型的基因数据，包括离散的表型数据，数量性状数据和标记基因型数据。

疾病状况通常在单独的一列进行编码：

U or 1 for unaffecteds, A or 2 for affecteds, and X or 0 for missing phenotypes.

编码数量性状时用 X 表示缺失值（也可以使用一种特殊的数值表示缺失的表型值，但该程序容易出错，不推荐）。

标记基因型被编码成用两个连续的整数，对于每一个等位基因用一个“/”进行分隔，或自 1.1 版本后使用字母“A”、“C”、“T”和“G”来编码。为了表示缺失的基因，可以用 0、X 或 N。以下是所有有效的基因型项 1/1（等位基因为 1 的纯合子）、0/0（缺失的基因型）及 3/4（等位基因为 3 和 4 的杂合子）。在 Merlin 的较新版本 A/A、A/C 和 C/C 也是有效的基因型，对于 X 染色体，男性假定他们有两个相同的等位基因。

以下为前面的家系文件添加了疾病状况，对数量性状的测量值和两个标记的基因型后所呈现的形式：

<contents of basic2.ped>

1	1	0	0	1	1	x	3	3	x	x
1	2	0	0	2	1	x	4	4	x	x
1	3	0	0	1	1	x	1	2	x	x
1	4	1	2	2	1	x	4	3	x	x
1	5	3	4	2	2	1.234	1	3	2	2
1	6	3	4	1	2	4.321	2	4	2	2

<end of basic2.ped>

注意第 5 个和第 6 个个体，他们都被标记成易感（即第 6 列的值为 2），其他的每个个体都被标记成非易感的（即第 6 列的值为 1），对应的数量性状（第 7 列）值为 1.234 和 4.321。尽管每个个体在第一个标记上都进行了基因分型，但对于第二个标记，只有个体 5 和个体 6 进行了基因分型。

（三）家系数据分析

家系文件所包含的标记基因型，疾病的状况和数量性状变量的个数只受可用内存的限制。由于每个家系文件具有唯一的结构（除了第一个 5 列），其内容必须在与其配对的数据文件中被描述。

数据文件包括家系文件中的每行数据项，显示出了数据类型（将标记编码为 M，将易感状况编码为 A，将数量性状编码为 T，并将相关变量编码为 C）并为每一个数据项提供了一个用一个单词表示

的标签。对应于上述家系的包含有一个易感状况，接下来是一个数量性状和两种标记基因型的数据文件的具体形式，如下所见：

<contents of basic2.dat>

A	some_disease
T	some_trait
M	some_marker
M	another_marker

<end of basic2.dat>

可以利用 pedstats(包含在 Merlin 中)得到任何一组家系文件和数据文件的概括性描述。要运行 pedstats 必须提供数据文件的名称(-d 命令行选项)和家系文件的名称(-p 命令行选项)。在 Merlin 的例子的目录中，尝试下面的命令：

```
prompt> pedstats -d basic2.dat -p basic2.ped
```

小提示：在 Merlin 和 Pedstats 的新版本中，就可以组合多个家系和数据文件。这种方法在分析多个不同的子集或通过染色体或区域划分基因型时非常方便。例如，如果表型数据存储在 pheno.dat 和 pheno.ped 文件中，且基因型数据存储在 geno.dat 和 geno.ped 文件中，可以利用以下命令将它们进行组合：

```
prompt> pedstats -d pheno.dat,geno.dat -p pheno.ped,geno.ped
```

(四) 遗传定位

为了分析遗传标记，Merlin 需要它们在染色体上的定位信息。这通常要提供一个定位文件。如果正在使用的是性别平均定位，此文件中的每个标记占一行三列，显示出染色体，标记名称和位置(以厘摩为单位)。如果正在使用的是性别特异性定位，需要另外两列分别来指定沿女性遗传方向定位的标记位置和沿男性遗传方向定位的标记位置。

数据文件和定位文件可以包含不同的标记集合，但那些在定位文件中缺少标记就会被 Merlin 忽略。下面是一个典型的定位文件，如下所示：

<contents of basic2.map>

CHROMOSOME	MARKER	POSITION
24	some_marker	123.4
24	another_marker	136.2

<end of basic2.map>

这里是一个精密版本的定位文件，包括每个标记的性别特异性定位位置：

<contents of file with sex-specific map>

CHROMOSOME	MARKER	POSITION	FEMALE_POSITION	MALE_POSITION
24	some_marker	123.4	146.8	100.0
24	another_marker	136.2	166.4	103.0

<end of sex-specific map>

使用划分后的数据和定位文件作出了一个非常简单的文件结构，并允许 Merlin 在一个单一的运行中分析多个染色体。

（五）关联分析模块

Merlin 也可以检测一个 SNP 与一个或多个数量性状之间的关联性。在 Merlin 中进行的关联性检测包括一个集成的基因型推理功能，它可以在一些基因型缺失的情况下提高工作效能。在这个例子中，会看到如何利用 Merlin 进行关联分析，以及如何利用集成的基因型推理功能估计缺失的基因型。

Merlin 进行的关联检测可以用于全基因组关联性扫描，或用于候选区域研究。不过，重要的是要注意与标准的以家庭为基础的关联测试的相比，在 Merlin 中进行的检测并不控制群体分层。如果群体分层是一个要关注的方面，那么群体的成员应该作为相关变量被包括在其中或用基因控制的方法来矫正结果。

要运行 Merlin 中的关联分析，需要指定数据集合(-d 参数)，一个家系(-p 参数)和定位文件(-m 参数)。此外需要下列关联性检测之一：打分检测(-fastAssoc)或似然比检验(-assoc)。打分检测(-fastAssoc)能够快速、理想的筛选大量的标记(例如，在一个全基因组范围关联扫描的第一阶段中)，而更精确的似然比检验(-assoc)可以用来评估数量较少的标记(例如，可用于在候选区域进行挑选的后续分析中)。在只包含较小家系的数据集或当被评估的影响较小时，这两项检测会给出类似的结果。

```
prompt> merlin -d assoc.dat -p assoc.ped -m assoc.map –fastAssoc
prompt> merlin -d assoc.dat -p assoc.ped -m assoc.map –assoc
```

“-assoc”和“-fastAssoc”，是两个最常用于检测关联性的命令，上面的命令行是采用这两个命令的输入格式。这些命令在 Merlin 中用于常染色体分析，且在 Minx 中用于 X- 连锁标记分析。命令运行中，还可以采用“-PDF”选项和“-inverseNormal”选项对结果进行了图形化的概括或自动变换性状使它们遵循平稳的正态分布。

小　结

人类复杂疾病机制研究是生物医学研究的重要组成部分，单核苷酸多态(SNP)很早就用于复杂疾病的研究。随着人类单体型计划提出，人类 SNP 分型数量得到飞跃式发展，使得从 SNP 的遗传定位、变异功能分析等层面进行复杂疾病研究得到长足的发展，并有力地促进了临床发现和应用，为个性化医疗提供了可能。本章中：①介绍了 SNP 的分型技术和数据资源；②复杂疾病或复杂性状遗传定位研究中的实验设计和主要分析方法；③系统生物学、系统遗传学思想在复杂疾病(性状)研究中的现状和展望；④多种软件工具在相关问题处理中的应用。对于学生而言，应当对方法学方面的知识有所了解，对于不同的问题能够知道采用何种方法，选取适当的软件进行分析，在理论基础上进行适当的拓展，并能够对结果进行科学的解释。同时在学习过程中也应当领悟生物信息学方法辅助实验设计的技能，结合本章，就是合理的利用遗传变异数据资源，运用统计遗传学、系统遗传学思想进行疾病机制研究中的候选致病因子或靶标筛选，相信在实验和临床工作中会有所收益。

Summary

Etiology research of human complex (or common) disease is one of the most important items in biomedical field. Single nucleotide polymorphisms (SNPs) as molecular marker used in this field are thought as an effective study tool and are proven have a comprehensive association with complex disease and phenotypes. In this chapter, we mainly introduce ①SNP genotyping technologies and resources. ②SNP related experimental design and analysis methods. ③System biology and system genetics strategies in disease research. ④Useful software and tools in SNP and disease relationship studies. Eight-year program students major in clinic should understand

statistical genetics and system genetics theory in human disease research, grasp the basic experimental design and analyzing methods, and have the ability to explain the obtained results.

（李亦学 徐良德 李 霞）

习 题

1. 人类的 DNA 组成碱基有 A、C、G、T 四种，为什么绝大多数 SNP 却是二态的分子标记？

2. 利用 HapMart 网络平台提取中国汉族人群基因 *IL8* 上的 SNP，并列出这些 SNP 各自所处的功能区域。

3. 试结合第一章中对 NCBI 等重要的序列数据资源知识介绍，选取适当的限制性片段长度多态性方法用于基因 *IL8* 上 SNP 分型的内切酶。

4. 结合网络信息，研究不同类型的 SNP 分型芯片特点及其适用的研究范围。

5. 从本章给出的 Haploview 网址下载并安装最新版的 Haploview 软件版本，研究一下可以输入的数据格式。

6. 在 Haploview 中导入 HapMap 格式 *IL8* 基因上的 SNP 基因型，从中提取能够覆盖整个基因的最少数量的 TagSNP。

7. 通过用户手册，了解 Plink 软件包各工作模块及不同的关联分析方法的软件实现（Plink 网址 http://pngu.mgh.harvard.edu/~purcell/plink）。

8. SNPtest 软件包还提供了基因型推断功能模块，通过网站，对其进行了解，探讨基因型推断在实际工作中的意义。

9. 如果你已经进入临床实习阶段，结合了解的临床知识，对本科室收治较多的病例进行记录和统计，预演基因组范围关联研究的取样过程。

10. 通过本章的介绍和你自己的理解，设计一个基于 SNP 的通路与疾病相关性实验设计，并分析其可行性。

主要参考文献

1. The International HapMap Consortium. A haplotype map of the human genome. Nature, 2005, 437(7063): 1299-1320.
2. Abecasis G. R., Cherny S. S., Cookson W. O., et al. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet., 2002, 30(1): 97-101.
3. Balding D. J., A tutorial on statistical methods for population association studies. Nat. Rev. Genet., 2006, 7(10): 781-791.
4. Barrett J.C., Fry B., Maller J., Daly M. J. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 2005, 21(2): 263-265.
5. Carlson C. S., Eberle M. A., Kruglyak L., et al. Mapping complex disease loci in whole-genome association studies. Nature, 2004. 429(6990): 446-452.
6. Conrad D. F., Jakobsson M., Coop G.,et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet., 2006, 38(11): 1251-1260.
7. Cookson W., Liang L., Abecasis G., et al. Mapping complex disease traits with global gene expression. Nat. Rev .Genet., 2009, 10(3): 184-194.
8. Frazer K. A., Ballinger D. G., Cox D. R. et al. A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007, 449(7164): 851-861.

9. Hirschhorn J. N., Daly M. J. Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet., 2005, 6(2): 95-108.
10. Jakobsson M., Scholz S. W., Scheet P. et al. Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 2008. 451(7181): 998-1003.
11. Lander E. S., Schork N. J. Genetic dissection of complex traits. Science, 1994, 265(5181): 2037-2048.
12. Mackay T. F., Stone E. A., Ayroles J. F. The genetics of quantitative traits: challenges and prospects. Nat. Rev. Genet., 2009, 10(8): 565-577.
13. McCarthy M. I., Abecasis G. R., Cardon L. R., et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet., 2008, 9(5): 356-369.
14. Rockman M. V., Kruglyak L. Genetics of global gene expression. Nat. Rev. Genet., 2006, 7(11): 862-872.
15. Slatkin M. Linkage disequilibrium: understanding the evolutionary past and mapping the medical future. Nat. Rev. Genet., 2008, 9(6): 477-485.

第十六章 miRNA与复杂疾病
CHAPTER 16 MIRNA AND COMPLEX DISEASE

第一节 引 言
Section 1 Introduction

微小RNA(microRNA, miRNA)是一类非编码的小RNA分子,其长度约22个核苷酸(nucleotide, nt),通过RNA诱导的沉默复合物(RNA-induced silencing complex, *RISC*)和其靶基因3′非翻译区(3′untranslated region, 3′UTR)结合,导致其靶mRNA的降解或阻碍其靶mRNA的翻译。从1993年在线虫中首次发现miRNA到现在,大量的研究表明miRNA可以通过精细地调节基因的表达进而参与细胞的发育、分化、增殖、凋亡以及应激反应等生物学过程,miRNA受到越来越多的研究人员关注。随着miRNA在复杂疾病中的深入研究,研究者发现其在疾病的发生发展过程中起着巨大的作用,其功能异常能够导致各种人类复杂疾病(例如癌症、心血管疾病等)的发生。这将使miRNA成为疾病诊断、预后的新的biomarker,并为进一步揭示复杂疾病的发病机制提供新的方向。

第二节 miRNA与其靶基因
Section 2 miRNA and Their Targets

一、miRNA概述

miRNA是一类广泛存在于真核生物中的内源性非编码RNA,长度在19~24nt间,在转录后水平上通过调节其靶基因行使功能。研究表明miRNA在诸如生长发育等多种生物学过程中发挥重要调节作用。其表达具有组织特异性和时空特异性,并能够精细地调控基因的表达。据推测,人类有超过三分之一的基因受miRNA调控。

(一) miRNA的发现

第一个miRNA是1993年在对秀丽新小杆线虫发育过程的研究中首次被发现的,当时被命名为*Lin-4*。它通过与*Lin-14*的3′UTR相互作用来调节线虫的发育。随后,在人类、果蝇、斑马鱼、拟南芥和水稻等多种真核生物中找到了上百个类似的小分子RNA,并将其称为miRNA。

(二) miRNA的生物起源

编码miRNA的基因首先在细胞核内产生长度在几百至几万nt的初始miRNA(pri-miRNA)。pri-miRNA被一种称为微处理器的多蛋白质复合物剪切为长度在60~70nt间并具有发夹结构的单链前体miRNA(pre-miRNA)。pre-miRNA通过转运蛋白质*Exportin-5*被转运至细胞质中,经过*Dicer*酶及其辅因子*TRBP*共同加工形成长度在19~24nt的miRNA以及miRNA*。随后,细胞中的*TRBP*募集*Argonaute*蛋白与*Dicer*酶形成三聚体复合物进而启动*RISC*的装配。miRNA链通过5′端互补被整合进*RISC*的同时,miRNA*链被特异的降解掉。miRNA一旦被整合进*RISC*,就会通过碱基配对引导*RISC*到达其靶mRNA从而行使功能(图16-1)。

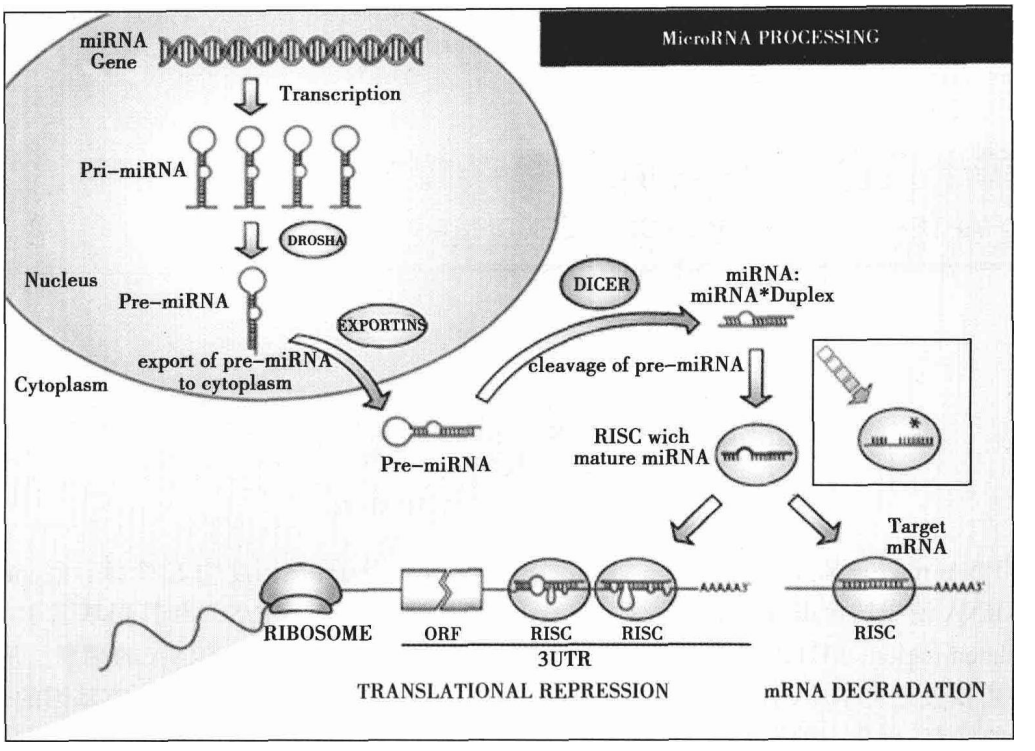


图 16-1 miRNA 的生物起源

（三）miRNA 的特点、作用机制及分类

研究表明 miRNA 在序列、表达、调控、物理位置等方面主要有如下特征：①在序列特征上主要有两方面特点，即 miRNA 本身不具有开放阅读框，不编码蛋白质；成熟的 miRNA 5′端为单一磷酸基团，3′端为羟基；②miRNA 的表达具有时序性以及组织特异性；③miRNA 与其靶基因间呈多对多的调控关系，即一个 miRNA 可能调控多个靶基因，而一个基因也可能受多个 miRNA 调控；④ miRNA 的物理位置倾向于成簇地出现在染色体上；⑤miRNA 还具有在物种间高度保守的特点。

成熟 miRNA 主要通过抑制和降解两种方式调节其靶基因的表达，具体采用哪种机制取决于 miRNA 与靶 mRNA 间的互补程度，即“种子区域”（通常指 miRNA 5′端第二位到第八位的核苷酸序列）与靶 mRNA 3′端的互补性。如果两者完全互补，则 miRNA 直接使 mRNA 降解；若两者不完全互补，则抑制 mRNA 的翻译。根据与靶基因结合方式的不同，miRNA 可大致分为三类：①第一类以线虫中的 *Lin-4* 为代表，该类 miRNA 与其靶基因以不完全互补配对的方式结合，抑制 mRNA 的翻译但不影响其稳定性（目前发现的大部分 miRNA 属于这一类）；②第二类以拟南芥中的 *miR-171* 为代表，该类 miRNA 与其靶基因以完全互补的方式结合，其作用方式和功能与小干扰 RNA（small interfering RNA，siRNA）非常类似，即直接靶向降解 mRNA；③第三类以 *Let-7* 为代表，该类 miRNA 可以通过上述两种方式作用于靶基因。例如，在果蝇和 Hela 细胞中的 *Let-7* 直接介导 *RISC* 降解其靶 mRNA；而线虫中的 *Let-7* 则与其靶 mRNA 3′UTR 以不完全互补配对的方式结合进而抑制其靶基因的翻译。

研究 miRNA 生物学功能和作用机制的关键是准确识别 miRNA 的靶基因。miRNA 通过结合 *RISC* 并作用于 mRNA 的 3′UTR 上，降解其靶 mRNA 或抑制其靶 mRNA 的翻译，广泛参与细胞的增殖、分化、发育、凋亡等多种生物学过程，并对多种疾病的产生有重要影响。但到目前为止，仅有少量的 miRNA 靶基因得到了实验证实，仍旧有很多 miRNA 的靶基因不能确定，导致这些 miRNA 的功能不能得到充分的研究。因此，如何快速准确地鉴定 miRNA 的靶基因是当前研究的一项重要挑

战。靶基因的识别对认识miRNA的功能、参与的生物学过程和疾病的发生，以及最终将miRNA用于临床实践具有十分重要的意义。近年来科研人员开始利用生物信息学的方法对miRNA靶基因进行预测。从2003年开发的第一个基于序列的miRNA靶基因预测算法miRanda开始，已涌现出多种miRNA靶基因预测算法。

二、基于序列的miRNA靶基因预测方法

miRNA的靶基因通常分为两类：5′端主导型和3′端补充型。其中，5′端主导型又分为5′端主导的“标准型”和“种子型”：5′端主导的“标准型”是指miRNA的5′端和3′端都具有较好的碱基互补配对；5′端主导的“种子型”是指miRNA的3′端没有发生较好的碱基互补配对，但miRNA的5′端至少有连续的7个碱基与mRNA的3′UTR完全互补。3′端补充型指miRNA序列3′端有多个碱基和mRNA的3′UTR发生互补配对，允许种子区第4~6位碱基或第7~8位碱基不互补（图16-2）。

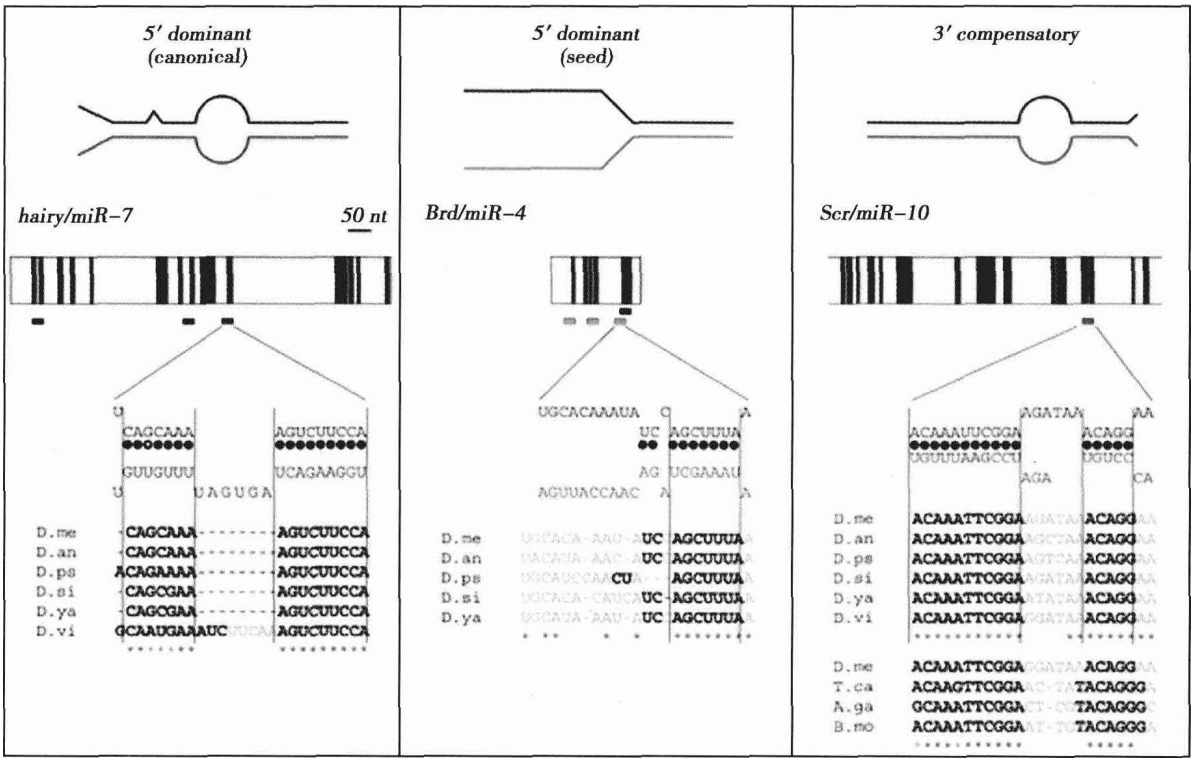


图16-2 三类miRNA靶点

当前基于序列的miRNA靶基因预测算法虽然各不相同，但通常遵循以下几个原则：①miRNA的“种子区”与mRNA的3′UTR序列碱基互补；②靶点在多物种间的序列保守性；③miRNA与mRNA形成双链结构的热力学稳定性；④靶基因二级结构和靶点外的序列对靶基因预测的影响。

基于以上的原则，目前常见的miRNA靶基因预测算法一般都包括以下三个步骤：①在mRNA的3′UTR上探寻和miRNA“种子区”完全互补的序列；②计算miRNA和这些序列结合产生的自由能下降值，对靶点进行筛选；③对靶点进行物种间的序列比对，利用靶点的物种保守性进行进一步筛选。

尽管步骤基本相同，但不同的miRNA靶基因预测算法侧重点有所不同。第一个利用生物信息学方法开发的基于序列的miRNA靶基因预测算法是miRanda，其选取了黑腹果蝇的所有miRNA序列。首先对miRNA和mRNA的3′UTR序列进行碱基互补分析，碱基互补遵循四个规则：①miRNA

的第 2~4 位碱基必须与 mRNA 的 3′UTR 精确匹配；②miRNA 的第 3~12 位碱基错配小于或等于五个；③第 9~25 位碱基至少有一个错配；④miRNA 的最后五个碱基错配不能多于两个。miRanda 采用一种类似于 Smith-Waterman 的算法来构建打分矩阵，允许 G-U 错配。互补的打分规则为：A-U 和 G-C 为 +5，G-U 为 +2，其他错配方式为 −3，空位罚分为 −8，空位延伸罚分为 −2。为了体现出 miRNA 的 5′ 端和 3′ 端在与靶基因结合过程中作用的不均一性，miRanda 软件设定了 scale 参数，即 miRNA 5′ 端前 11 个碱基的互补得分乘以 scale 参数，再和 3′ 端 11 个碱基互补得分相加作为序列的最终碱基互补得分。其次，在 miRNA 与靶基因形成二聚体的热力学稳定性方面，miRanda 利用 Vienna 软件包中的 RNAlib 计算 miRNA 与 mRNA 3′UTR 结合的自由能。最后，miRanda 要求靶点在多物种间保守，即靶点在多物种 3′UTR 序列比对中相同位置具有相同的碱基。

TargetScan 主要考虑物种间保守的 miRNA 靶基因，并且在 TargetScan 中首次提出了“种子匹配”的概念。在 TargetScan 算法中，“种子匹配”被定义为 miRNA 5′ 端的第 2~8 位碱基与 mRNA 3′UTR 上的一段 7nt 种子序列完全互补。从种子区开始向 miRNA 两侧寻找互补碱基，允许 G-U 配对，直到出现碱基错配为止。在物种保守方面，TargetScan 算法发现随着物种数目的增多，预测的靶基因数目逐渐减少，但预测结果的准确率得到提高。2005 年，同一组研究人员在 TargetScan 中添加了更多的物种，改进的算法称为 TargetScanS。与 TargetScan 相比，TargetScanS 在人、小鼠、大鼠三个物种的基础上增加了狗和鸡的数据，并将 miRNA“种子区”由之前定义的 miRNA 5′ 端第 2~8 位 7 个碱基调整为第 2~7 位 6 个碱基，在“种子区”完全互补的前提下，同时要求 miRNA 5′ 端第 8 位碱基与靶基因互补或第 1 位碱基是腺嘌呤(Adenine；A)。研究人员同时检测了一组已知的秀丽新小杆线虫 miRNA 靶点，识别出一种连续的 GC 富集(GC-rich)碱基对模式，并命名为“结合核”(binding nucleus)，这些“结合核”的长度通常为 6~8 个碱基并分布在接近 miRNA 的 5′ 端。针对“结合核”设计的打分机制充分考虑了连续碱基 GC、AU 以及 G-U 对的权重。2007 年，Andrew 等研究了 miRNA“种子区”外的序列特征对 miRNA 靶基因预测的影响，并对 TargetScanS 的算法进行了改进。新的算法加入了 miRNA“种子区”外第 12~17 个碱基通常与 mRNA 的 3′UTR 互补、miRNA 靶点多位于 mRNA 3′UTR 的 AU 富集区、功能相似的 miRNA 靶点距离较近、miRNA 靶点多位于 mRNA 3′UTR 的两端等特征。

在预测 miRNA 靶基因的算法中，RNAhybrid 考虑了靶基因结合自由能对预测结果的影响。该算法利用动态规划算法寻找一条短链 RNA(miRNA)和一条长链 RNA(mRNA 3′UTR)杂交时的最优自由能鉴别 miRNA 的靶点。与其他的 RNA 二级结构预测软件 mfold、RNAfold 等相比，RNAhybrid 除了具有明显的速度优势外，RNAhybrid 算法还禁止 miRNA 分子间和靶基因间杂交产生二聚体。RNAhybrid 没有考虑靶基因的物种间保守性，允许用户自己定义自由能的阈值、P 值，也允许用户自己设置 miRNA“种子区”的位置和长度以及是否允许出现 G-U 错配等。

此外，基于序列的 miRNA 靶基因预测研究中还使用了机器学习的方法，通过在少量实验证实的 miRNA 靶基因集合内提取 miRNA 与靶基因的结合特征，并利用这些特征训练分类器来预测 miRNA 的靶基因。如 TargetBoost 和 miTarget 等算法都是从实验证实的 miRNA 靶基因集出发，评估 miRNA 与靶基因结合的序列特征、二聚体结构特征和热力学特征等参数，最后对预测的靶基因进行打分。

在 miRNA 与靶基因结合的过程中，mRNA 的 3′UTR 二级结构起着重要作用。研究发现，miRNA 靶点几乎都在 3′UTR 的二级结构不稳定区域内，通过计算 mRNA 的 3′UTR 二级结构被破坏、形成或破坏碱基互补配对、形成 miRNA-mRNA 二聚体时获得或损失的自由能，可以鉴别 miRNA 靶基因；同时，通过实验发现，提高靶点附近序列二级结构的稳定性大大降低了 miRNA 对靶基因的作用。

靶点外的序列也对 miRNA 调节靶基因起到重要作用。已有实验表明，靶点后的一段序列对 miRNA 与靶基因的识别起着重要的作用，该段序列的突变将使 miRNA 对靶基因的调控作用明显减

弱，而将该段序列完全删除则miRNA对靶基因的调控作用完全消失。最近的研究表明，在miRNA调控靶基因的过程中，靶点外的其他序列甚至整个3′UTR序列都起到了关键作用，这些序列可能是RNA结合蛋白的作用位点。

三、基于表达信息或实验结果预测miRNA靶基因

在最初的研究中，研究人员认为miRNA结合在mRNA的3′UTR上抑制mRNA翻译成蛋白质，降低蛋白质丰度，并不会影响到相应mRNA的表达水平。但现在已经明确认为：在许多情况下，miRNA还能直接对mRNA的表达产生影响。科研人员已经开发了整合表达信息的miRNA靶基因预测算法，并证明了表达信息在miRNA靶基因预测上的重要价值。

Huang等利用在88个组织中同时检测的miRNA和mRNA表达数据，并结合贝叶斯方法开发了靶基因预测算法GenMiR++，在基于序列算法预测结果的基础上对靶基因进行进一步筛选，提高预测精度。结果共得到了104个人类miRNA的高精度靶基因，并通过实验证实了预测的*let-7b*靶基因。结果表明，与基于序列的方法相比，利用相同样本中同时检测的miRNA和mRNA表达谱可以更准确地预测miRNA靶基因。

Gennarino等通过研究miRNA宿主基因的表达情况，开发了miRNA靶基因预测算法HOCTAR。HOCTAR是第一个利用miRNA宿主基因表达与mRNA表达信息对miRNA靶基因预测的算法，它基于两者表达的逆相关特征对预测的miRNA靶基因进行筛选。通过对178个人类miRNA的宿主基因分析，发现预测准确性优于现存的基于序列的预测方法，HOCTAR减少了基于序列算法预测的靶基因数量。

Bandyopadhyay等利用miRNA的表达谱和mRNA表达谱构建了一组阴性样本集，并利用机器学习方法开发了miRNA靶基因预测算法TargetMiner。由于当前实验证实的miRNA靶基因阴性数据较少，用机器学习方法预测miRNA靶基因常具有较高的假阳性率，作者从miRNA和mRNA的表达谱中得到了300多个组织特异的阴性样本，并结合实验证实的miRNA靶基因数据，利用支持向量机方法开发了新的miRNA靶基因算法。

四、整合已有知识预测miRNA靶基因

在当前的miRNA靶基因预测研究中，研究人员逐渐意识到单一依靠序列信息或表达信息已不足以提高miRNA靶基因的预测效能。因此，整合功能信息、蛋白质互作信息、表达信息、序列信息以及当前实验证实的miRNA靶基因等已有资源预测miRNA靶基因是十分必要的。在过去的研究中，研究人员利用生物信息学的方法整合多种数据资源，已成功对疾病候选基因、药物靶基因等进行了筛选和优化，这些方法将为新的miRNA靶基因预测算法的开发提供重要参考；同时，利用高通量的实验方法对miRNA靶基因的预测也在不断的发展中，这些研究将对最终揭示miRNA功能和参与的生物学过程、找出miRNA诱导的疾病发生机制以及将miRNA用于治疗癌症等疾病具有重要意义。

五、miRNA数据资源

（一）TarBase数据库

TarBase是一个目前使用最广泛的存储真实miRNA与靶基因间关系的数据库，其网址为：http://diana.cslab.ece.ntua.gr/tarbase/，存储来自大约200篇文献，涵盖多种实验方法的1333个miRNA与靶基因关系对。用户可通过选择物种、miRNA名称以及基因名称对miRNA与靶基因的对应关系进行查询，结果将按自动编号顺序列出概要信息，点击结果条目编号旁的加号图标即可展开，得到详细信息（图16-3）。其主要由三部分组成：第一部分为miRNA信息，提供来自miRBase的miRNA序列、mRNA序列等基本信息；第二部分为基因信息，提供靶基因的染色体定位、表达信息以及编码的蛋

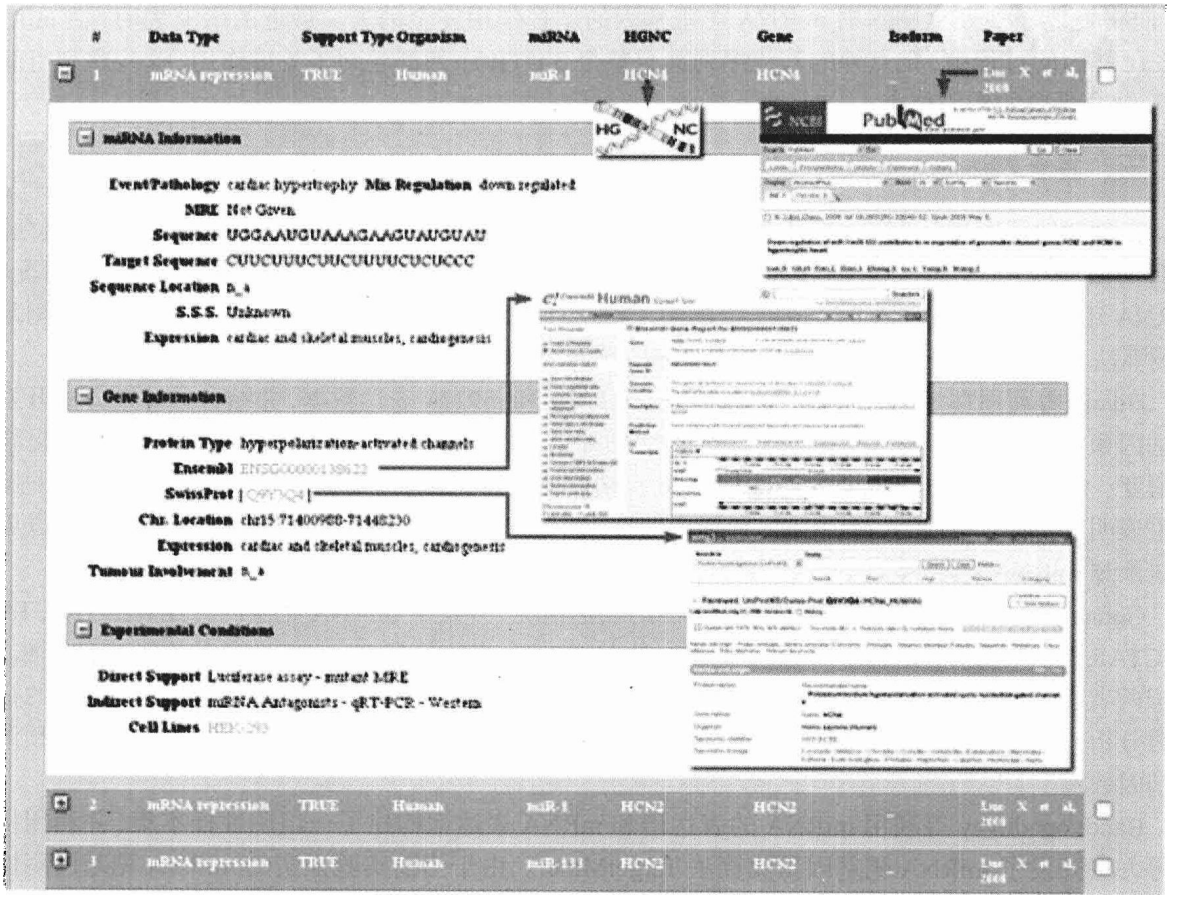


图 16-3 TarBase 数据库查询结果解析图

白质在 SWISS-PROT、Ensembl 数据库的链接；第三部分为实验条件，提供直接或间接的实验技术支持。数据库以 Excel 文件形式存储，可供用户下载使用。

（二）miRBase 数据库

miRBase 是一个集 miRNA 序列、注释信息以及预测的靶基因数据为一体的数据库，是目前存储 miRNA 信息最主要的公共数据库之一（网址：http://www.mirbase.org/）。miRBase 提供便捷的网上查询服务，允许用户使用关键词或序列在线搜索已知的 miRNA 和靶基因信息。该数据库主要包括三部分内容，即 miRBase Registry、miRBase Sequence 以及 miRBase Targets。miRBase Registry 主要是为新发现的 miRNA 命名服务；miRBase Sequence 包含所有已发布的成熟 miRNA 序列，同时提供对应的预测的发卡结构、注释信息以及与其他数据库的链接；miRBase Targets（已更名为 microCosm，网址：http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/）主要采用 miRanda 算法预测 miRNA 靶基因。本节将重点介绍 Targets 部分，首先用户可以点击“Enter”进入数据库浏览数据，通过“Search”按钮进入搜索界面，可通过选择物种、输入 miRNA ID、基因名称、Ensembl 标识符以及关键词进行 miRNA 与靶基因关系的查询，并且可以通过下拉列表选择 Gene Ontology（GO）的不同部分或直接输入 GO 节点名称查询，结果有 GFF 和 TXT 两个格式供用户下载保存。点击“Download”则可以下载相关的数据（各部分功能见图 16-4 所示）。

此外，许多 miRNA 相关数据库被开发，例如 miRGen、MiRNAmap 以及 microRNA.org。这些数据库包括了大量信息（靶基因数据、基因组注释数据以及 miRNA 表达相关数据），为 miRNA 的研究提供了便捷。

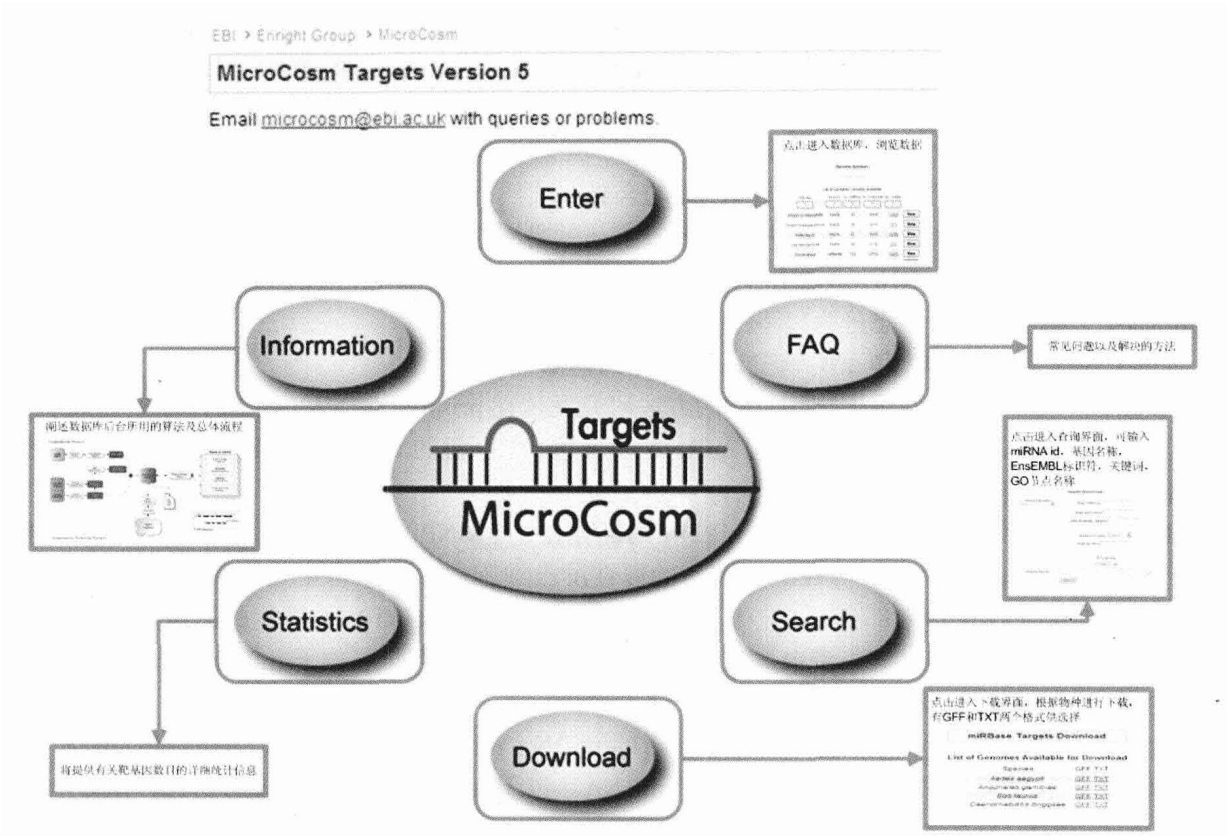


图 16-4 miRBase 靶基因数据各部分功能

第三节 miRNA 多态和复杂疾病

Section 3 miRNA Polymorphism and Complex Disease

miRNA 多态能够从不同的层面影响 miRNA 功能。miRNA 多态导致的异常可能发生在 miRNA 的形成至行使功能过程的任意一个阶段，如图 16-1 所示。在人类基因组中这些多态可以以插入、删除、扩增或者染色体易位的形式出现，最终导致 miRNA 位点或功能的缺失或者获得，这是人类基因组一类新的功能多态。miRNA 多态不仅会影响 miRNA 的产生和表达，而且会影响 miRNA 与靶基因的结合从而影响靶基因的表达。例如，如果 miRNA 多态影响的肿瘤抑制基因表达下调，那么就会导致肿瘤的发生。目前的一些研究表明，miRNA 多态与疾病的发生发展和药物反应存在重要的关系。根据当前已知的研究，将 miRNA 多态主要分为三类：①位于 miRNA 基因内影响 miRNA 形成和功能的多态；②位于 miRNA 靶基因内影响 miRNA 与其靶基因调控关系的多态；③改变药物反应和 miRNA 基因表观遗传调控的多态。

一、miRNA 基因内部的多态

miRNA 通过与 mRNA 的 3′UTR 区域结合对成百上千的 mRNA 进行转录后调控。影响 miRNA 生物学形成的多态不仅会影响 miRNA 自身的表达，也会对 miRNA 调控的基因产生影响。在染色体基因组水平上单个核苷酸变异引起的 miRNA 序列多态性对 miRNA 的转录、形成、输出和调控具有重要影响。Saunders 等对人类 474 个 miRNA 的 SNP 进行系统的分析发现 SNP 在 miRNA 序列内的密度低于其周围的侧翼序列。49 个 pre-miRNA 内存在 65 个 SNP 位点，其中 3 个 miRNA 的种子序列存在 SNP。Duan 等对人类 227 个 miRNA 基因的 SNP 分布研究也发现了相似的结果，这意味着 miRNA 基因内部存在多态。

在miRNA形成的不同阶段有不同的蛋白质和蛋白质复合物参与其中，这些蛋白质包括RNA聚合酶Ⅱ复合物、*Drosha/Pasha*、*Exportin-5\Ran-GTP*、核孔复合体、*Dicer*和*RISC*。如果一些多态影响上述这些蛋白质参与miRNA的形成，那么这些多态也会对miRNA产生效应。这些影响蛋白质表达的多态可能会导致miRNA表达的下调。位于miRNA基因或者靶基因序列的多态可以通过miRNA或者参与miRNA形成的蛋白质影响miRNA的合成和成熟，导致新的miRNA和其相应的靶基因的产生，使得miRNA的功能缺失或者进一步影响疾病的易感性和药物的敏感性。本节主要讲述位于miRNA基因内部影响miRNA形成和生物学功能的多态。根据目前miRNA多态的研究，可以将影响miRNA生物学形成的多态分为以下三类：

（一）位于pri-miRNA和pre-miRNA基因序列内部

Saunders和Duan等利用生物信息学的方法研究发现在pri/pre-miRNA内部存在单核苷酸多态。位于pri/pre-miRNA内部的多态会影响miRNA的表达，产生新的miRNA以及影响miRNA与靶基因的结合，甚至与疾病的风险相关。

Duan等研究发现*miR-146a*的pre-miRNA内部存在一个SNP(rs2910164)，此位点的C等位可以增加miR-146a的表达。Calin等在一些家族的慢性淋巴细胞白血病患者中发现miR-15a/miR-16的初级转录本(pri-miRNA)内部存在C→T改变，此多态与miR-15a/miR-16表达的减少相关，这两个miRNA在几乎70%的白血病患者中具有比较低的表达水平。这意味着位于miR-15a/miR-16初级转录本的多态与白血病的发生有关。

位于pri/pre-miRNA内部的多态也导致新的miRNA的产生。研究发现，miR-146a前体会产生miR-146a(正链)和miR-146a*(负链)两种miRNA。位于miR-146a前体的rs2910164不仅会影响miR-146a的表达，而且会导致miR-146a*C和miR-146a*G两种miRNA的产生。

位于pri/pre-miRNA的多态会影响miRNA的表达或者产生新的miRNA，从而影响与靶基因的结合或表达。这些多态通过影响miRNA与靶基因的结合，从而可能与多种疾病的发病风险相关。之前研究的位于miR-146a前体的多态(rs2910164)会产生两种miRNA，其中miR-146a*C调控*PTC1*基因，miR-146a*G调控*IRAK1*基因。位于该位点的C等位影响miR-146a调控的乳腺癌基因*BRCA1*和*BRCA2*的表达。这就可以说明此多态位点CG杂合子基因型患乳腺癌的风险降低，而CC和GG两种纯合的基因型患病的风险增加。

总之，位于pri/pre-miRNA内部的多态可能会产生新的miRNA，影响miRNA的表达进而影响靶基因的表达，甚至与复杂疾病的发病风险相关。

（二）位于成熟的miRNA序列内部

成熟的miRNA通过与mRNA的3′UTR区域结合对mRNA进行转录后调控。miRNA与mRNA结合的区域包括两部分：一部分是miRNA的5′端第2~7个碱基，称为种子区域，这一部分区域要求与mRNA完全匹配；另一部分是种子区域附近的3′端方向，允许一定程度的错配称为3′容错区域(3′MTR)。位于成熟miRNA序列的多态会影响其对靶基因的调控，消除、弱化、增强或者产生新的结合靶点。Warthmann等在植物中发现位于miR-319a内部的突变会导致该miRNA功能的丢失。

根据miRNA与靶基因结合的两部分区域，可以将位于成熟miRNA上的多态分为以下两类：

1. 位于miRNA的5′种子区域　Saunders等研究结果发现，位于miRNA种子区域的SNP不足1%。而目前的研究发现，位于miRNA种子区域的多态会影响miRNA的表达以及与靶基因的结合。位于*miR-125a*种子区域的多态显著地抑制了pri/pre-miRNA过程，导致miRNA表达的减少。*miR-206*调控*ERα*的表达，*miR-206*存在两个与*ERα*结合的靶点。位于*miR-206*种子区域的多态导致两个靶点都失活，消除了与原来靶的结合。

由于miRNA调控成百上千的mRNA，那么理论上认为位于miRNA种子区域的多态会影响成百上千的基因的表达，但这需进一步的实验证实。

2. 位于miRNA的3′容错区域(3′MTR)　3′MTR区域不同于种子区域，允许一定碱基的错配。

然而，在这一区域存在的多个SNP、插入、删除或者异位可能会对miRNA调控靶基因产生影响。但是，目前没有发现确切的效应，还需要将来进一步的研究。

3. 影响蛋白质参与miRNA的形成 miRNA的生物学形成过程中，存在一些蛋白质参与其中。影响这些蛋白质参与miRNA形成的多态可能会影响miRNA的表达和调控，很可能会导致miRNA的下调。Gottwein等研究疱疹病毒相关的miRNA时发现，位于pri-miR-K5的SNP显著地抑制了Drosha对pri-miR-K5的剪切，使得pre-miR-K5和成熟miR-K5的表达显著降低。由于miRNA表达的降低或者升高在细胞内有严重的后果，因此影响蛋白质参与miRNA形成的多态会广泛地影响miRNA的转录、加工、输出和靶向，在细胞内具有不利的影响。

二、miRNA靶点的多态

miRNA靶点的多态性指的是靶基因上影响miRNA与靶基因结合的序列多态性。相对于生成miRNA的染色体区域，基因的3′UTR区域具有较弱的序列保守性，因此在3′UTR中出现序列变异的频率更高，这表明在人类基因组中，与miRNA自身的多态性相比miRNA靶点的多态性具有更高的分布密度。由于靶点的多态性会影响miRNA对靶基因的调控强度，导致基因调控的失常，所以它们与多种遗传疾病的发病风险有关，成为遗传药理学研究的重要内容之一。根据多态性位点与miRNA靶位点的位置关系，可以进一步把miRNA靶点的多态性分为miRNA结合位点上的多态性和miRNA结合位点上下游的多态性。

（一）miRNA结合位点上的多态性

对miRNA与靶基因的结合机制的研究表明，一个19~22碱基的靶结合位点可以分为种子区域和非种子区域。种子区域一般为结合位点的前2~7个碱基，miRNA结合靶位点时对此区域的序列匹配程度要求很高，SNP的出现会严重影响miRNA与靶基因的结合。非种子区域则允许有一定程度的碱基错配，因此也将其称为错配容忍区，如图16-5所示。

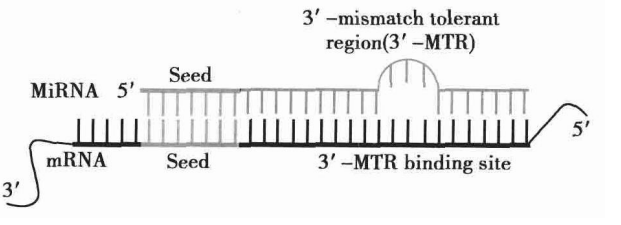


图16-5 miRNA与靶基因结合

（二）miRNA结合位点上下游的多态性

miRNA与靶mRNA的结合事件的发生除了需要miRNA与靶位点的序列匹配外，还需要一些辅助蛋白，如*RISC*蛋白等的参与。然而，转录出来的靶mRNA的3′UTR区域在细胞中不是以单链的形式存在的，它会自我折叠形成由各种茎环和柄组成的二级结构。因此它的某些部分可能是不能或不容易与蛋白质或miRNA结合的，这样mRNA和蛋白质或miRNA是否能相互作用还要依赖于靶mRNA上是否具有某些特定二级结构的元件。例如：当miRNA结合位点位于柄区域时，miRNA与靶mRNA的结合需要破坏柄结构将结合位点暴露出来，这一过程需要辅助蛋白提供较多的能量，因此miRNA较难与靶mRNA在此区域结合。当miRNA结合位点位于茎环区域时，由于结合序列直接以单链的形式存在，所以miRNA与靶mRNA的结合比较容易。Zhao等研究发现大多数的靶mRNA 3′UTR上的miRNA结合位点都具有较简单的二级结构，因此更容易使miRNA-RISC复合物进入与靶点结合，位于靶点附近的多态位点可能会改变mRNA的二级结构从而影响miRNA与靶基因的结合。

Mishra等发现，在针对日本人群的SNP分析中，*DHFR*基因3′UTR上的*miR-24*靶点附近的*SNP 829*(C>T)能够影响*miR-24*与靶位点的结合。当*SNP 829*等位基因型为T时，*miR-24*不能与*DHFR*上的靶点结合。

此外，位于结合靶点上下游的SNP会影响miRNA与3′UTR上其他调控元件的协同作用。3′UTR是mRNA分子3′端的非编码片段，从编码区末端的终止密码子延伸至多聚A尾巴(Poly-A)的末端，除miRNA结合位点外，还包含很多顺式作用元件和功能结构元件，例如CPEARE序列(3′UTR中

的 AUUUA 重复序列)等。位于 miRNA 结合位点附近的多态性会影响 miRNA 与其他位于 3′UTR 的顺式调控元件的作用，从而影响 miRNA 对靶基因的调控，如图 16-6 所示。

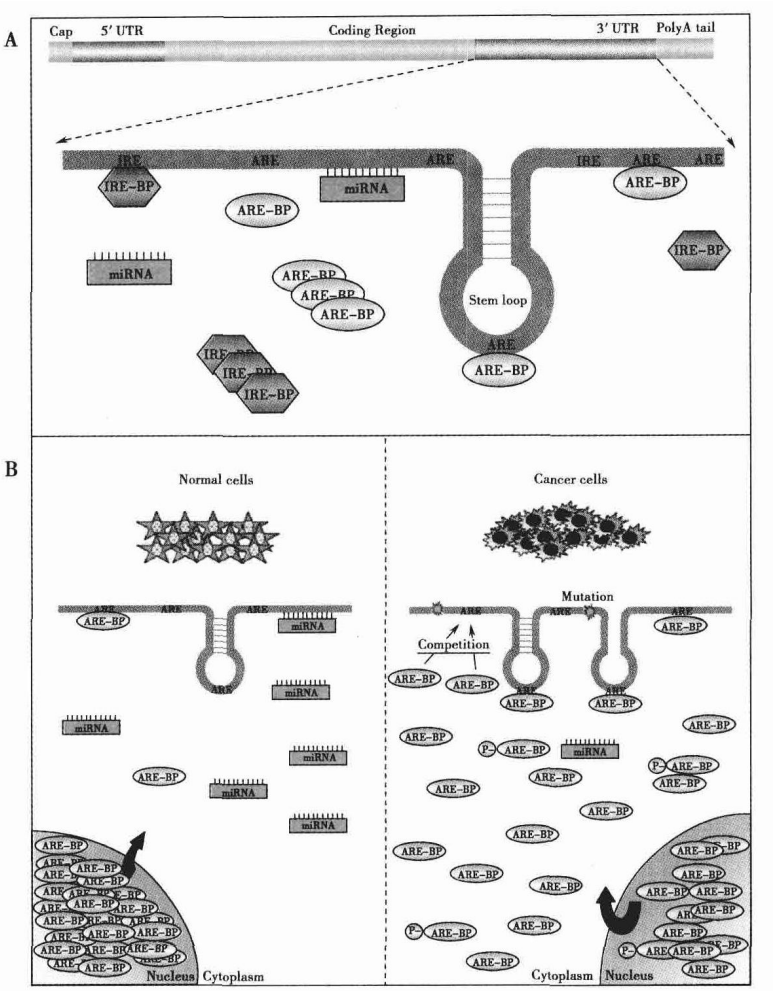


图 16-6 多态性位点影响 miRNA 与 3′UTR 区域其他顺式元件的作用

mRNA 中的特殊序列元件影响 mRNA 的稳定性，不稳定元件往往出现在 3′UTR 区域，它可以缩短 mRNA 的寿命。不稳定 mRNA 的一个普遍特征是其 3′ 尾部存在约长 50 个碱基的富含 AU 的序列(称为 *ARE*)。*ARE* 中的共有序列是重复几次的五聚核苷酸 AUUUA，它控制着 mRNA 的降解。在含有 *ARE* 的 mRNA 的降解过程中，还需要一些蛋白(例如，*Dicer1* 和 *Ago1* 等)的参与，这些蛋白组分也是在 miRNA 合成和行使功能时所必需的。进一步研究发现 *miR16* 含有与 *ARE* 互补的序列，这说明 *miR16* 可能会和 *ARE* 顺式调控元件相互作用，进而控制 mRNA 的降解。因此多态位点会影响 miRNA 与其他位于 3′UTR 的顺式调控元件的相互作用。

三、miRNA 多态影响药物反应

miRNA 多态性位点是一类新的功能多态位点，它们可以通过影响 miRNA 与药物作用蛋白的结合，从而增强药物敏感性或者导致药物抗药性，如图 16-7 所示。

有研究表明，人类基因 3′UTR 区域的 SNP 与 α-地中海贫血、人类乳头瘤病毒感染、胰岛素敏感、尿石症和 5-氟尿嘧啶化疗治疗的敏感性相关。作为人类基因 3′UTR 区域上的调控因子，miRNA 靶点上或者邻近区域的 SNP 能够影响 miRNA 与靶基因的结合，产生新的 miRNA 调控关系或者失活已有的 miRNA 调控关系，从而导致疾病或者抗药性的产生。

Mishra 等研究发现，miRNA 多态性位点的出现与甲氨蝶呤(MTX：一种抗肿瘤药物)的抗药性

有关。当 *DHFR* 基因 3′UTR 上 *miR-24* 靶点附近的 *SNP 829* 等位基因型为 T 时，*miR-24* 将不能与 *DHFR* 上的靶点结合，*miR-24* 的功能失活导致 *DHER* 基因 mRNA 和蛋白的堆积，堆积的 *DHFR* 直接导致了甲氨蝶呤抗药性的产生。在此工作基础上，Mishra 等提出 miRNA 多态性对药物作用产生影响的两种可能途径。首先，miRNA 多态性能够产生新的 miRNA 靶向关系。当药物作用蛋白是药物靶点蛋白时，新产生的 miRNA-药物靶点蛋白调控对能够下调药物靶点蛋白的表达，从而增强药物敏感性；当药物作用蛋白是药物激活蛋白时，新产生的 miRNA-药物激活蛋白调控对会下调药物激活蛋白的表达，从而导致药物抗药性。与之相对应，miRNA 多态性也能够失活已有的 miRNA-药物作用蛋白调控对的靶向关系，当药物作用蛋白是药物靶点蛋白时，失活的调控关系会导致靶点蛋白的过量表达，从而导致药物抗药性；当药物作用蛋白是药物激活蛋白时，失活的调控关系增加激活蛋白的表达，从而增强药物敏感性。同时，人们进一步提出了 miRNA 药物基因组学的概念。miRNA 药物基因组学可以被定义为通过分析 miRNA 和影响 miRNA 功能的多态性，从而预测药物作用效果和提高药物作用有效性的学科。miRNA 药物基因组学有着很强的临床应用前景。miRNA 是很值得研究的药物靶点，因为它们能够调控细胞中一些关键蛋白的表达，同时其自身在癌症组和正常组中也存在差异表达。miRNA 多态性能够导致 miRNA 对药物靶点蛋白调控的失活从而产生抗药性。因此，随着利用病人的基因型值来确定病人对药物的耐受程度的研究方法的不断发展，miRNA 多态性位点能够作为临床药物作用效果的预测因子，使得药物使用过量发生的可能性逐渐减小。

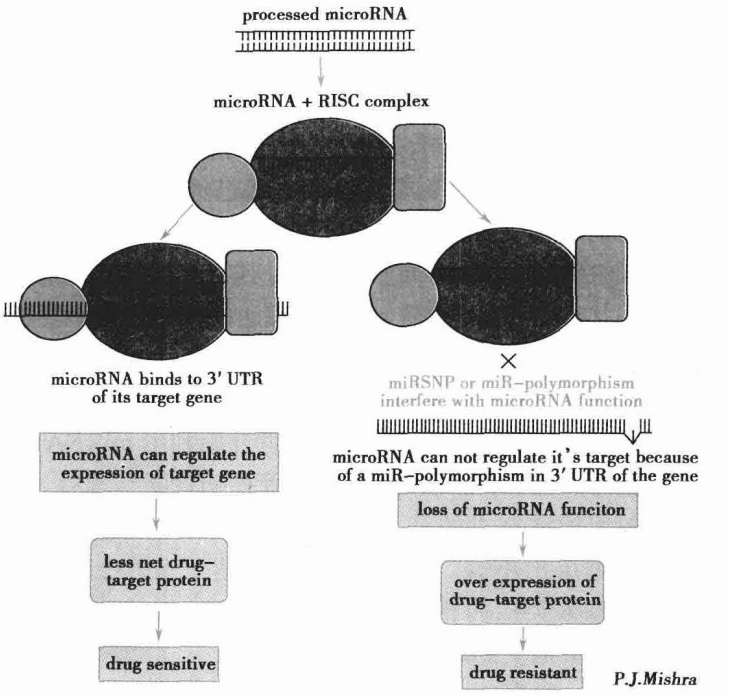


图 16-7 miRNA 多态对药物反应的影响

四、miRNA 多态改变表观遗传调控

很多 miRNA 由于异常的高甲基化受表观遗传修饰沉默的影响。miRNA 的表观沉默最初在乳腺癌的发病过程中被发现。Lehmann 等针对 71 例乳腺癌患者研究发现基因中 *miR-9-1*、*miR-124a3*、*miR-148*、*miR-152* 和 *miR-663* 在 34%~86% 的患者中具有异常的高甲基化。因此，miRNA 的异常高甲基化会导致癌症的发生。引起 miRNA 表观遗传调控改变的多态研究是一个新的还没有探索的领域，由于 miRNA 多态导致的原癌基因或抑癌基因表观遗传调控的缺失或者获得在细胞中可能具有决定性的影响。因此，可以利用改变表观遗传调控的 miRNA 多态研究疾病发生的机制。

第四节 miRNA 表达谱与复杂疾病

Section 4 miRNA Expression Profile and Complex Disease

一、miRNA 表达谱识别癌症相关 miRNA

随着对癌症发病机制的逐渐了解，科学家将癌症的本质归结为各种原因引起的基因结构和功能的异常。这些异常通常表现为致癌基因的高表达以及抑癌基因的低表达。作为一类重要的基因负

调控子，miRNA 通过调节大量靶基因广泛参与各种生物学过程。癌症过程中 miRNA 表达水平的变化会对其靶基因的活动产生深远的影响。正是由于 miRNA 具有使其靶基因高效沉默的调控作用，miRNA 便很自然地被认为参与癌症的发生，并因此被引入到癌症的研究及治疗中。

基因组数据的蓬勃发展大大加快了基因结构和功能的研究，同时也改变了人们对基因调控的认识和理解。毫无疑问的是微阵列技术已经在基础和应用性研究中做出了巨大贡献，为将来实现临床研究中的个性化给药提供了可能。微阵列技术能够同时检测成千上万个基因的 mRNA 表达水平。对于 miRNA 这个相对崭新的领域，其特有的性质（如片段小、低丰度、组织特异性、发育阶段特异性和疾病状态特异性）使得 miRNA 表达的检测受限。尽管如此，目前有许多新的检测方法，包括基于放射性标记探针的 Northern 印迹、克隆、定量 PCR 扩增、SAGE 技术、磁珠技术和寡核苷酸芯片技术，已经被用于 miRNA 表达水平的检测。研究表明，miRNA 芯片能够成功地用于评估大量 miRNA 的表达水平。如今，越来越多的研究者开始利用 miRNA 芯片来揭示 miRNA 的生物合成、组织特异性、干细胞的发育、肿瘤的发病机制以及疾病诊断与预后。

随着检测技术的突飞猛进，miRNA 表达数据层出不穷。基于 miRNA 表达谱来挖掘人类疾病相关的 miRNA 进而解释发病机制受到越来越多研究者的关注。迄今，许多癌症相关的 miRNA 表达谱数据已经被提交到 Gene Expression Omnibus（GEO）和 ArrayExpress 数据库。图 16-8 显示了利用 miRNA 表达谱研究复杂疾病的流程：miRNA 表达谱的产生及获取、数据预处理、差异表达 miRNA 筛选、后期生物学实验的证实以及基于靶基因功能富集的异常生物学过程的识别。

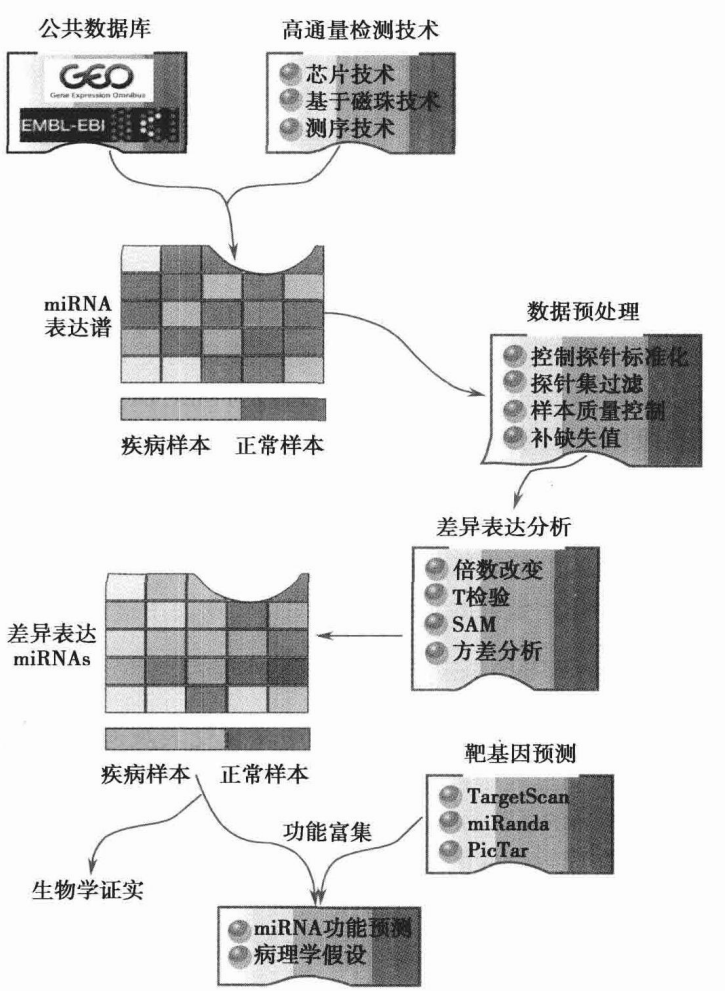


图 16-8 基于 miRNA 挖掘疾病相关基因的流程

首先，与 mRNA 芯片数据的分析过程类似，需要对所获取的 miRNA 表达谱数据进行标准化。然而，许多用于 mRNA 表达谱数据标准化的方法并不能简单地移植到 miRNA 表达谱数据。由于目前并没有统一有效的方法用于 miRNA 表达谱数据的标准化，大多数研究都只是将 miRNA 基因表达数据进行中值化处理。随后，通过比较癌症样本和正常样本中 miRNA 的表达数据，利用常见的差异表达检测算法（如 t 检验等），寻找出显著差异表达的 miRNA。对于这些表达异常的 miRNA，需要后期生物学实验进一步证实。这些异常 miRNA 通过调节大量靶基因的转录变化，导致某些生物学功能失调从而诱导癌症的发生。虽然目前还没有针对 miRNA 功能注释的数据库资源，但是编码蛋白质的基因的注释信息却非常丰富，如 GO、KEGG 等。因此，通过靶 mRNA 的功能来推测 miRNA 的功能是可行并且合理的。利用本章第二节中介绍的 miRNA 靶基因预测方法：Targetscan、miRanda、PicTar 等，获取癌组织中异常表达 miRNA 的靶基因集合，对靶基因进行 GO 功能注释以及 KEGG 通路分析，得到靶基因集合显著富集的生物学过程。由此可推测

异常miRNA正是通过调节这些生物学过程而参与癌症发生。

目前，许多研究利用miRNA表达谱通过比较正常和癌组织中miRNA的表达水平来识别出特定癌症类型中异常的miRNA。Yanaihara等对104个肺癌组织和非肺癌组织的miRNA表达谱进行分析找到了肺癌中43个异常表达的miRNA，其中28个显著下调，15个上调。Murakami等通过表达谱分析发现在肝癌细胞系中，8个差异表达的miRNA中5个显著低表达。Climmino等最近报道，*miR-15a*和*miR-16-1*负调控*BCL2*。*BCL2*是一个已知的抗凋亡基因，它在多种肿瘤(如白血病、淋巴瘤)中过度表达。可见，*miR-15a*和*miR-16-1*的缺失或下调会导致*BCL2*表达的升高，从而促进白血病和淋巴瘤的发生。第一个被发现的由*let-7*家族编码的miRNA负调控癌基因*Ras*。*Ras*蛋白是一个膜相关的*GTPase*信号蛋白，能够调节细胞生长、分化。大约有15%~30%的人类肿瘤都含有*Ras*突变，而该突变导致蛋白表达上升进而引起细胞转化。因此，一个能够调节这些潜在的癌基因表达的miRNA，可以控制细胞的增殖速度。在人类基因组中存在12个由*let-7*同源基因家族编码的miRNA，可能起到抑癌基因的作用。研究者还发现这些miRNA与肺癌、乳腺癌、子宫癌等有关的脆性位点有联系。更直接的证据是Takamizawa发现人的肺癌中*let-7*同源物的表达显著减少，并且导致了更差的预后。临床研究发现非小细胞肺癌患者的*let-7*表达水平越低，其预后越差、术后生存期越短，这说明*let-7*可以用于疾病的诊断。体外组织培养实验发现，在人的肺癌细胞中瞬时的增加*let-7*可以抑制细胞的增殖，这也说明*let-7*在肺组织中可能是一个抑癌基因。因此，可以考虑使用*let-7*来治疗肺癌。

二、miRNA表达谱分类人类癌症

迄今，癌症的分子分型已经取得了巨大的进步。许多研究已经表明编码蛋白的转录本可以有效地区分各种癌症。这些与癌症相关的转录本作为一种可靠的生物学标记已被广泛应用于各种癌症的分型研究。近十几年，随着人们对miRNA的了解以及实验技术的进步，越来越多miRNA被发现。同时，这些miRNA的功能也得到了更深入的研究和证实。重要的是，许多研究表明miRNA的表达异常通常与癌症的发生发展有密切关系。因此，目前一些研究已经开始探索利用miRNA表达谱数据对癌症进行分类的可行性，并且将miRNA作为一种新的生物学分子标记用来判断癌症发生、发展或者预后。

2005年，Lu等成功地利用磁珠流式细胞术检测技术检测334个样本中的217个miRNA的表达水平，并使用该表达谱首次全面的证实了miRNA在癌症分类中的有效性。

首先从http://www.broadinstitute.org/cancer/pub/migcm/获取所有相关数据，其中包括334个样本的原始miRNA表达数据、预处理后的miRNA表达数据、探针信息、样本信息等。该数据集也被提交到GEO数据库，其访问号为GSE2564。该表达谱数据包含用两个芯片平台(两套miRNA探针集)检测的334个样本的miRNA表达谱。所采用的预处理过程包括基于控制探针的标准化(两套miRNA探针集中包含了一些控制探针)、修正表达强度偏低的探针、删除所有控制探针、以及对表达值进行以2为底的对数转换。检测的334个样本中包括多种人类组织，如胃、结肠、肺等，其中某些组织取自癌症患者，例如肺癌、白血病患者等。由于某些miRNA的表达并没有被检测到或其表达值很低，简单地把这些miRNA表达值包含在表达谱数据集中只会增加数据的噪音，会影响后续的分析结果。因此，首先对miRNA表达谱进行过滤，删除那些在所有样本中不表达或表达值很低的miRNA。如果某一miRNA在某一样本中的表达值低于7.25(此阈值是基于一个先行实验所确定的)，则认为该miRNA在该样本中不表达或者其表达值过低。随后，对每个miRNA的表达数据进行均值为0、标准差为1的标准化。

采用层次聚类方法(平均链路算法、皮尔森相关系数)分别对样本和miRNA进行聚类分析，发现几乎所有的miRNA表达值在不同的癌症中都不相同(图16-9)。而且，从聚类图中可以明显看出具有相同组织发育起源的样本被聚到一类。例如，起源于上皮组织的样本几乎都被聚到一起，而造血相关的恶性肿瘤样本明显的分布在另一主要分支上。该结果表明miRNA表达谱能够很好地区分不

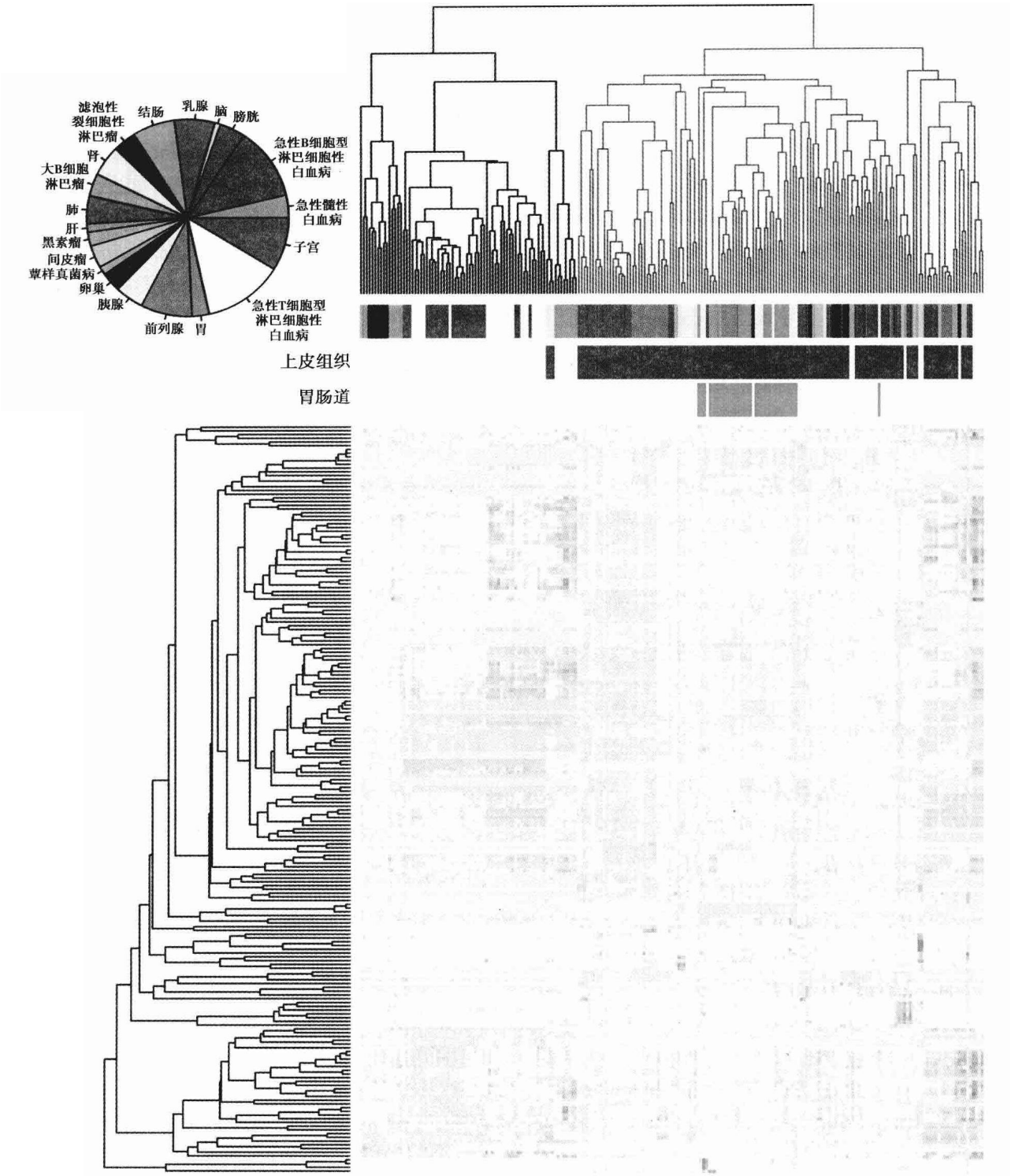


图 16-9　218 个样本的 miRNA 表达谱聚类图

同组织起源的样本。

除了上述所用到 218 个样本，该数据还包括了检测自 73 个急性淋巴细胞白血病患者骨髓样本的 miRNA 表达水平。如图 16-10 所示，经过聚类，这些样本被划分进入三个主要的分支：其中一个分支包含所有 5 个 BCR/ABL 阳性样本以及来自 11 个 TEL/AML1 样本中的 10 个样本；第二个分支包含了 19 个急性 T 细胞淋巴细胞白血病样本中的 13 个。该结果说明即使对于同一组织起源的样本，仍旧能观测到不同的 miRNA 表达模式。

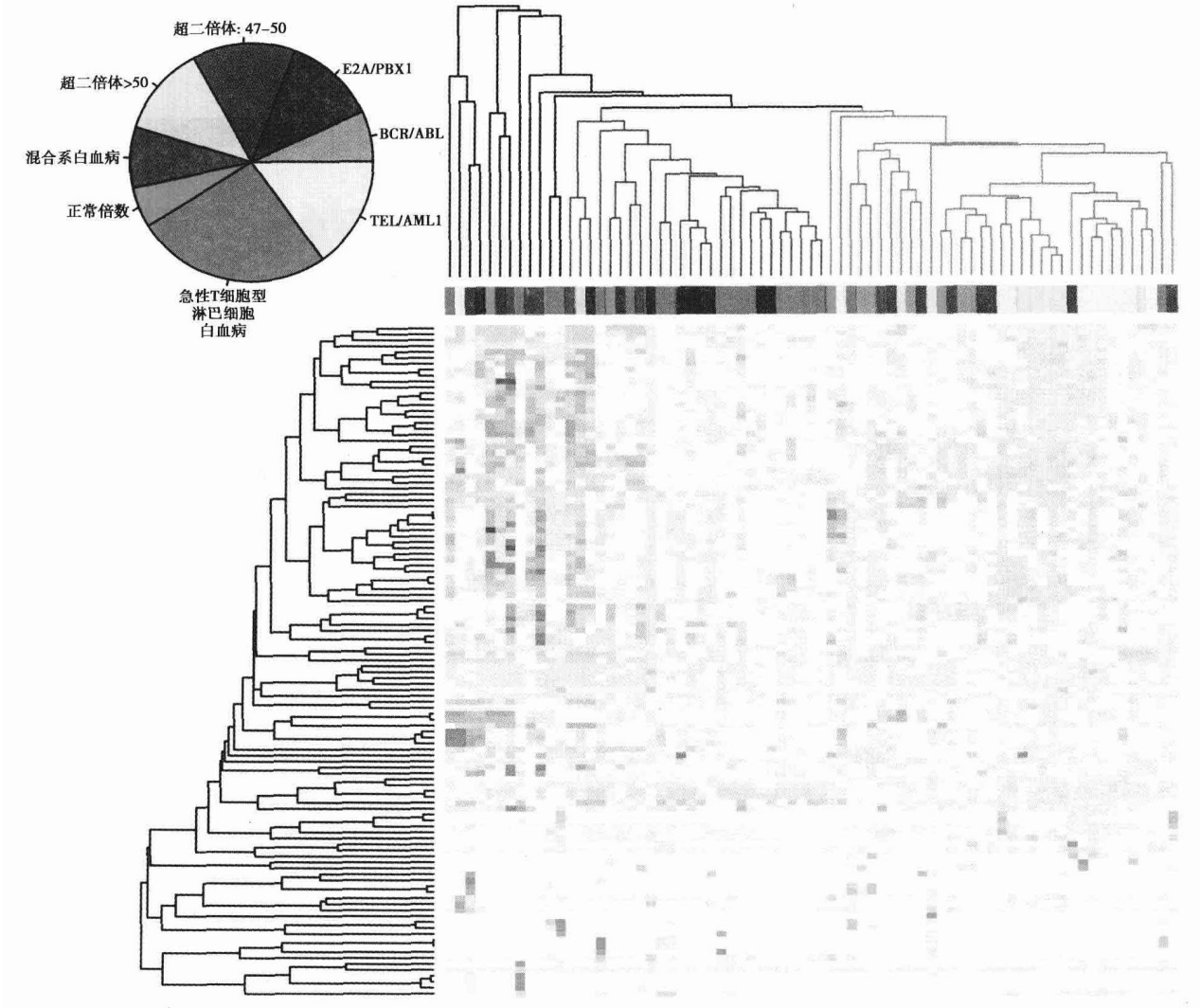


图 16-10 73 个急性淋巴细胞白血病患者的骨髓样本的 miRNA 表达聚类图

从图 16-9 可以发现，来自结肠、肝、胰腺以及胃部的样本被很好的聚在一类。这正好反映出它们共同起源于胚胎的内胚层，进一步表明对样本的 miRNA 表达谱进行聚类分析能够揭示出样本的组织起源。数据集中还包括了来自 218 个样本中的 89 个组织的 mRNA 表达水平。当利用大约 16 000 个 mRNA 表达谱数据对同样的样本聚类时，起源相同的组织并未被聚到一起(图 16-11)。这种现象的发生，很有可能是由于在高维度的 mRNA 表达谱数据中存在大量的噪音或者是不相关的信号。

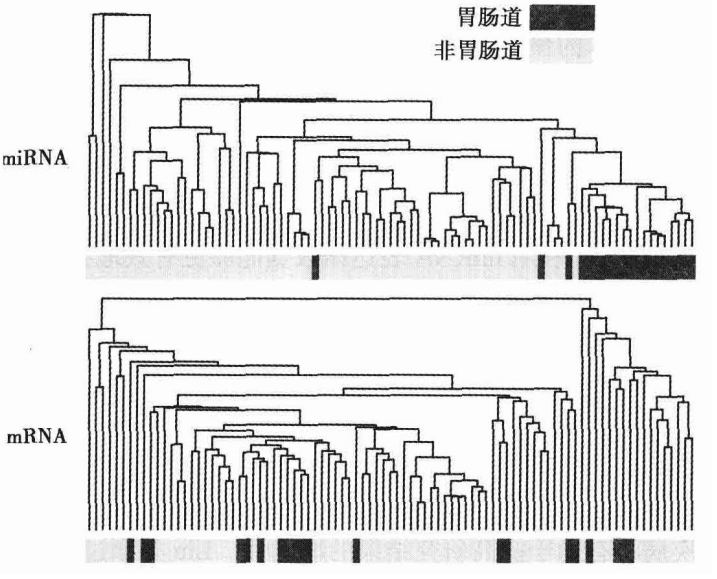


图 16-11 miRNA 与 mRNA 表达谱的比较

利用 t 检验方法对正常组织与肿瘤组织的 miRNA 表达水平进行比较，并使用随机扰动的方法为每个 miRNA 产生一个 p 值，最后对 p 值进行 bonferroni

多重检验校正。可以发现大多数的 miRNA 在肿瘤组织中呈现显著的低表达(217 个 miRNA 中有 129 个 p 值小于 0.05),并且该现象并不依赖于某一种特定的肿瘤类型(图 16-12)。

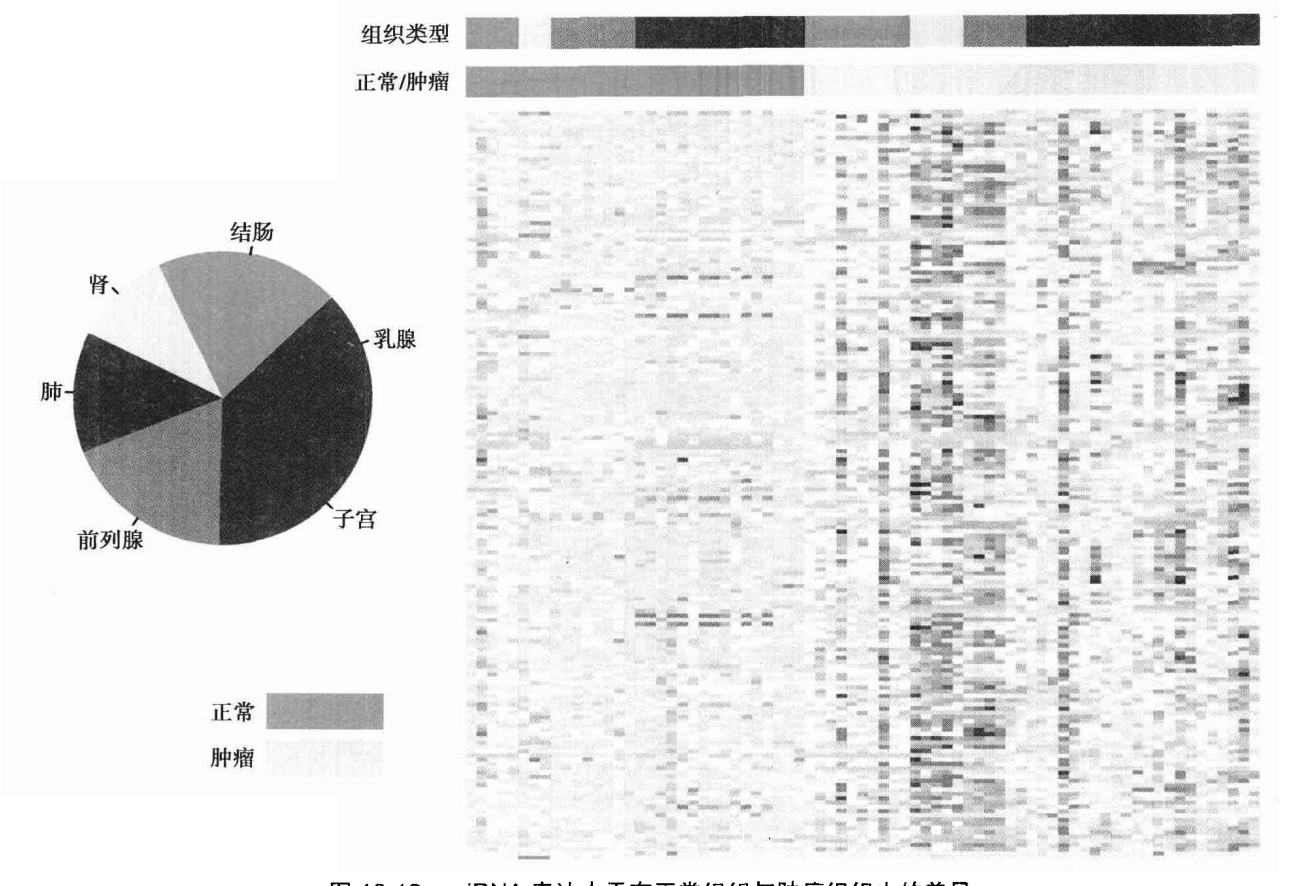


图 16-12 miRNA 表达水平在正常组织与肿瘤组织中的差异

为了进一步的证实 miRNA 表达谱能否用于肿瘤的诊断,选取了 68 个高分化的肿瘤样本(代表 11 种肿瘤类型),并利用概率神经网络算法对这些样本的 miRNA 与 mRNA 表达谱数据分别训练产生相应的多类别分类器。随后,利用所产生的分类器来预测 17 个低分化的肿瘤样本的组织类型。通过上述过程,每个测试样本都能得到 11 个组织类型的预测概率。选取具有最高概率的组织类型作为样本的预测组织类型。尽管 miRNA 表达水平在肿瘤样本中整体偏低,但是基于 miRNA 表达的分类器正确分类了 17 个低分化肿瘤样本中的 12 个样本,而利用 mRNA 构建的分类器只正确地分类了其中的一个样本。

上述实验结果暗示着 miRNA 表达水平蕴含着惊人的信息量,能够有效地反映出组织起源和肿瘤分化状态。同正常组织相比较,大多数 miRNA 在肿瘤样本中呈低表达状态。而且,同 mRNA 数据相比较,利用 miRNA 表达谱数据能够更有效地预测出低分化肿瘤样本的组织类型。总之,miRNA 表达谱数据为癌症的诊断提供了潜在的可能性。

三、miRNA 表达谱与 mRNA 表达谱的整合分析

随着实验检测技术的不断发展,大量的生物学资源涌现。生物信息学研究者已经不仅仅局限于使用一种数据资源,而是通过整合多种类型的生物数据(如各种表达谱数据、互作网络、调控网络、SNP 数据、表型数据等)来解决复杂的生物医学问题。整合 miRNA 和 mRNA 的表达数据研究复杂疾病将有助于提高研究结果的准确性。Liu 等通过整合 mRNA 与 miRNA 表达谱,来预测 miRNA 的功能,从而解释 miRNA 参与疾病发生和发展的过程。

为了阐明利用miRNA和mRNA表达谱整合分析方法来研究复杂疾病的过程，本节将使用2005年Lu等采用磁珠流式细胞式检测技术得到的89个人类多种组织样本中的miRNA和mRNA表达谱数据作为案例进行研究。

首先，获得miRNA和mRNA表达谱数据，并对其进行标准化处理成为可供分析的数据格式。为了降低实验误差，计算表达谱中所有miRNA和mRNA表达水平的方差，删除30%表达变化最小的miRNA和mRNA。数据筛选之后，得到了包含151个miRNA和11 114个mRNA的表达谱。同单独的mRNA表达谱数据分析类似，对于miRNA表达谱，通过计算151个miRNA之间的皮尔森相关系数来识别功能上相关的miRNA。结果中可以看出某些miRNA之间存在着高度的表达一致性(图16-13显示在皮尔森相关系数接近1的位置存在一个较小的波峰)。这些高度表达一致性的miRNA可能属于同一个miRNA簇(许多研究将位于染色体上距离50kb内的miRNA定义为一个miRNA簇)，并且共同发生转录。此外，通过整合miRNA和mRNA表达谱，计算miRNA-mRNA对的相关性，得到了miRNA-mRNA皮尔森相关系数密度曲线(图16-13)。从密度曲线图中发现，虽然miRNA-mRNA的表达相关性没有miRNA-miRNA的强烈，但仍然存在一些miRNA-mRNA对，它们的皮尔森相关系数较大。

为了利用成熟的mRNA的功能来推断miRNA的功能，设定相关显著性阈值为0.4，筛选出了149个表达上显著相关的miRNA-mRNA关系对。图16-14显示了利用149个共表达miRNA-mRNA对构建的网络。图中，*miR-99a*同时与7个基因(*ACTG2*、*MYH11*、*PRUNE2*、*GCSH*、*EDNRA*、*ACSL1*、*CCDN2*)具有较高的表达一致性。通过对这一组基因进行GO功能注释，发现其中4个基因共同参与了催化活性生物学过程，由此可以推测*miR-99a*可能参与催化活性作用。

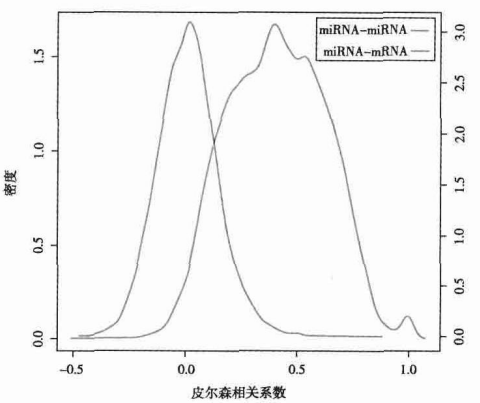


图16-13 miRNA-miRNA与miRNA-mRNA皮尔森相关系数密度曲线

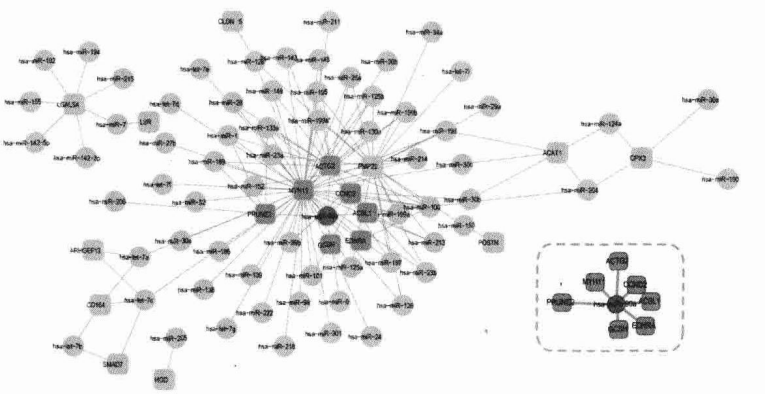


图16-14 miRNA与mRNA共表达网络

表达谱中的89个样本中有12个来自前列腺组织，其中包含6个正常样本和6个前列腺癌样本。对于miRNA表达数据，利用SAM算法寻找前列腺癌和正常样本中差异表达的miRNA(FDR<0.05)，筛选得到60个显著下调的miRNA。对于mRNA表达数据，通过t检验分析($p<0.05$)，识别出485个显著差异表达的mRNA。基于miRNA与靶mRNA的负调控关系，利用本章第二节中的miRNA靶预测算法，来确定60个差异表达的miRNA和485个mRNA的靶向对，这些miRNA-mRNA靶向对是在前列腺癌中共同出现异常的因子，所以，前列腺癌症的发生与这些miRNA-mRNA对密切相关。对异常miRNA靶向的mRNA进行GO功能注释，从而得到异常miRNA在前列腺癌症过程中所参与的生物学过程。

四、miRNA——新的生物标记

特异的临床生物标记能够潜在地应用于各种临床诊断，例如从妊娠异常到心肌梗死和癌症等各

类疾病，并在临床诊断和治疗上具有显著作用和效果。值得的注意的是，现代癌症研究的主要目的是寻找能够检测出早期瘤形成或变化的高敏感生物标记。理论上这些生物标记容易获取，例如可以从非侵入的组织中取样，甚至可以从血液或尿液之类的体液中就能直接获得样本。迄今为止，最显著的生物标记研究是通过各类疾病的基因表达谱数据识别了大量的可以作为生物标记的 mRNA，例如众所周知的肿瘤抑制因子 *P53*。

近年来，研究者发现 miRNA 与 mRNA 相似，也潜在地具有作为生物标记的特征。在人类中已检测到大约 1000 个 miRNA 中很多 miRNA 与癌症的发生、衍化、侵袭和转移有关，而且这些 miRNA 基因大部分位于染色体脆性位点区域，在癌症发生或发展过程中趋向于导致染色体扩增、缺失或异位等异常。同时，许多研究也进一步发现癌症相关的 miRNA 在癌症衍化过程中具有致癌或抑癌的作用，表明这些 miRNA 可以作为一种新的“原癌基因”或“抑癌基因”用于癌症诊断和研究，为研究者更加深入地探索癌症的发病机制提供了新的思路。

随着 miRNA 检测技术的不断发展，miRNA 在癌症中的表达模式逐渐成为许多癌症研究的焦点。特别是 Lu 等的研究结果表明了 miRNA 比 mRNA 能更好地分类癌症样本之后，大量的 miRNA 表达谱研究相继完成并得到了相似的结论。例如，在丙型肝炎患者肝组中，*miR-122* 的表达水平能够预测干扰素治疗的反应；在甲醛固定石蜡包埋的肿瘤样本中，*miR-205* 的表达可以区分鳞状和非鳞状小细胞肺癌；基于 48 个 miRNA 的分类器可以开发用于鉴别胃癌、胰腺癌等不同癌症的组织起源；乳腺癌研究中，通过分析乳腺癌的正常组织样本和恶性肿瘤组织样本的表达谱发现小部分 miRNA(*mir-125b*、*mir-145*、*mir-21* 和 *mir-155*)能够有效地区分恶性乳腺癌组织与正常组织，这些 miRNA 与许多乳腺癌组织病理学特征，例如乳腺肿瘤大小、淋巴结转移、增殖能力和血管侵蚀等密切相关。

同其他已有的生物标记例如 mRNA 相比，miRNA 片段更小并且表达量更大。miRNA 的这些特征将使得研究者在组织样本或者临床上获取的体液样本中利用定量的 PCR 技术检测 miRNA 的表达水平更加容易。已有一些研究者开始通过检测 miRNA 在血液中的表达水平来分析 miRNA 对癌症诊断的可能性。例如，在大 B 细胞淋巴瘤患者的血清中，*miR-155*，*miR-210* 和 *miR-21* 的表达水平被发现显著地上调；在肺癌患者的血清中，利用测序和定量 PCR 技术一致发现 *miR-25* 与 *miR-233* 的表达量增加；*miR-141* 在血清中的表达水平被用来检测早期的前列腺癌患者。最近的一个小鼠研究表明对乙酰氨基酚过量导致的肝损伤能够引起血液中 *miR-122* 和 *miR-192* 的急剧增加。与往常所采用的标准的肝毒性标记物丙氨酸氨基转移酶相比，miRNA 的表达变化反应更加迅速。miRNA 的大小以及表达量可能意味着 miRNA 能够有效的从特定的细胞中释放到体液中。但是，一旦 miRNA 从特定的细胞中释放出来，就很难知道释放出的 miRNA 在大量核酸酶的环境中能存活多久。最近的一些研究表明血浆中 miRNA 能够作为一种稳定的生物分子用于癌症的检测。这有可能是由于 miRNA 能够通过化学修饰或包装进入蛋白质复合物或者整合进入膜颗粒中被保护起来。随着对 miRNA 在血液中稳定性的进一步了解，将来也许能够发展出更加有效的递药系统应用于 RNA 的治疗。

由于 miRNA 是一类特异的并具有高信息量的生物分子，加之各种检测技术的逐渐成熟，miRNA 在肝癌、肺癌、肠癌、卵巢癌和白血病等各类癌症中的特征会得到更深入地研究和证实。这些特征说明 miRNA 可以作为癌症诊断的新的生物标记和治疗药物作用的靶点。

第五节 miRNA 调控分子网络

Section 5 miRNA Regulation of Molecular Networks

在活体细胞中，基因之间不是孤立的，也不是彼此独立地行使某种功能，而是聚集成团共同完成一些特定的生物功能。换句话说，在细胞中这些成团的基因共同起作用，因此可以用这些基因间存在的各种类型互作形成分子网络对成团基因的功能进行概念性的描述。在过去几十年里，基因调控网络和细胞信号网络被认为是细胞中主要的调控系统，而 RNA 仅被认为是传递遗传信息到蛋白质

产物的分子。但是，目前发现的多种非编码RNA小片段，特别是miRNA，挑战了一些传统概念。

miRNA具有许多生物学功能。通过预测算法得到的miRNA靶基因功能类型也很广泛，包括转录因子、信号蛋白、骨架蛋白、代谢酶等。miRNA靶点的多样性和丰富性暗示着miRNA可以通过其靶基因和其他细胞网络，如转录调控网络等，相互交织组成更复杂的调控网络。因此，miRNA通过调控细胞网络行使功能的观点是合理的，而且在系统水平上了解miRNA如何参与细胞过程是具有显著意义的。

蛋白互作网络、代谢网络、基因调控网络和信号网络是目前研究最广泛的四种细胞分子网络。其中，蛋白互作网络包含两种信息，即蛋白信息及其物理互作信息。蛋白互作囊括了从基本的细胞机制，如DNA合成的蛋白质复合物、转录因子复合物，到细胞信号涉及的蛋白质复合物等信息。简单地讲，基因组范围的蛋白互作网络包含了细胞中所有生物过程所涉及的全部蛋白互作关系。基因调控网络描述转录因子和编码蛋白的基因间的调控关系与蛋白互作网络相似，基因调控网络包含细胞中参与所有生物学过程的全部基因调控信息。因此，可以把蛋白互作网络和基因调控网络划分到同一类，称为“一般网络”。第二类网络由代谢网络和信号网络组成，称为“特异细胞网络”，这类网络描述某些特定的细胞活动。细胞代谢网络包含所有代谢反应和代谢流，同时信号网络包含信号流和信号传导中的生化反应。通常这两种信息用线性通路表示，例如代谢通路和信号通路。代谢网络中，代谢通路是相互交织的，因此代谢流可以通过许多不同的通路传递，并且有些代谢产物是被多条代谢通路共享的，因而通过一条或者多条通路可以得到某些最终代谢产物。信号网络涉及细胞内和细胞间的交流及信号蛋白对信息的处理方式。

一、miRNA调控细胞信号网络

信号网络是处理早期细胞内和细胞外信号的最重要的复杂系统，它作为高级交流系统会完成一系列的任务，例如细胞生长、存活和发育等。目前，许多研究者手工注释信号信息，并以信号通路或者信号网络的形式进行组织。还有一些研究者通过高通量实验技术发现新的信号蛋白及其涉及的互作关系。所有这些努力产生了大量信号通路信息，相互交织组成一个大的复杂的信号传递网络。

生化信号，例如磷酸化、乙酰化和泛应素等可以激活或者抑制信号蛋白。如果信号传导中出现错误，将会改变发育或者导致错误的决策行为，这些都能产生发育的异常。因此，信号蛋白间的关系对决定细胞行为和维持细胞稳态起到关键的作用。另外，信号蛋白对应的基因表达及调控子失调都会在信号网络中表现出来，同样引起发育终点的异常，例如导致癌症或其他疾病的发生。miRNA是作为一类重要的转录后调控子被认为参与信号网络的调控。

目前还有很多疑问尚待解决，例如信号蛋白作为一类特殊蛋白，其mRNA是否倾向被miRNA调控；miRNA倾向调控哪种类型的信号蛋白，是细胞表面的信号蛋白还是细胞核内的信号蛋白。另外，信号网络可以用图的方式表示，其中点代表信号蛋白、有向边代表蛋白间的激活或者抑制关系、无向边则代表简单的物理互作。那么，研究者就可以找到信号网络各种类型的局部模块形式，比如网络模体，而miRNA倾向调控哪种类型的模体，或逃避调控哪种类型的模体还需要进一步研究和确认。2006年，Cui等把miRNA靶基因数据映射到信号网络上，揭示了人类中miRNA调控信号网络的一些策略。下面以该研究作为探索miRNA调控信号网络的例子讲解。

首先从TargetscanS和PitTar中获得miRNA的靶基因数据，从中提取出人类靶基因数据，然后利用两种算法的交集作为比较可靠的miRNA调控靶基因的数据。另一方面，人类信号网络的数据主要是通过文本挖掘，经过人工检查校正，最后得到一个包含540个蛋白，1258条边的信号网络，数据下载地址为http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1681519/bin/msb4100089-s2.zip，其中的Text file S1和Text file S2分别储存信号蛋白和信号蛋白间的关联。

依据信号蛋白是否是miRNA的靶基因对其进行标记，结果发现29.4%的信号蛋白是miRNA的靶基因，而人类基因组中只有17%被miRNA调控，表明miRNA的调控在信号网络中有相对更重要

的作用。

下面进一步探讨哪类信号蛋白更倾向被 miRNA 调控。第一种是按照细胞定位信息将信号蛋白分为：配体、细胞表面受体、细胞内信号蛋白和核内的信号蛋白，miRNA 对四类信号蛋白的调控倾向性可以通过计算四类蛋白中 miRNA 靶基因的比例进行衡量。其中，只有 9.1% 的配体被 miRNA 调控，但是一半的核蛋白(大部分是转录因子)被 miRNA 调控，也就是说信号流下游成分中 miRNA 靶基因数比上游成分多五倍。第二种是依据信号蛋白的功能，以连接蛋白为主，许多细胞内的信号活动(例如，添补下游信号成分到邻近的受体)是通过连接蛋白完成的。连接蛋白没有酶的活性，但是和上游及下游的信号蛋白存在物理互作，从而实现信号的传递。依据连接蛋白下游成分的多少可以将连接蛋白分成两组：高连接组(下游有多于 4 个的信号蛋白)和低连接组(下游有不足 4 个的信号蛋白)，分别计算两组中下游信号蛋白中被 miRNA 靶向的比例，结果发现 36.1% 的高连接组下游的信号蛋白被 miRNA 调控，低连接组只有 24.2%，所以 miRNA 倾向调控前者的下游信号成分。例如，连接蛋白 *GRB2* 有 14 个下游信号蛋白，其中一半是 miRNA 的靶点。这些下游成分和不同的信号通路有关，可以导致不同的细胞输出。为了精确响应胞外刺激，连接蛋白需要选择性的添补下游成分。拥有的下游成分越多，其对应的基因的表达动态性就越高。这和 miRNA 具有高时空表达性是一致的，表明 miRNA 通过控制连接蛋白下游成分的浓度，从而精确响应刺激信号。

信号网络存在多种类型的网络模体，网络模体是网络中具有简单结构的单元，它表示信号网络中信号传递的特定小规模模式，不同的网络模体代表不同的信号传递模式。利用 Mfinder 程序可以提取由三到四个蛋白组成的网络模体。依据模体中蛋白被 miRNA 靶向的比例，可以将模体进一步分类。例如，三个点组成的模体中，所有蛋白都不被 miRNA 调控(类别 0)，或者只有一个蛋白是 miRNA 的靶点(类别 1)，或者两个(类别 2)、三个都是(类别 3)。其次，对每种子类模体，计算正向调控边(激活)占所有有向边(激活和抑制)的比例(称为 Ra)。最后把每种子类模体得到的 Ra 平均值和所有该类模体中得到的 Ra 平均值进行比较(图 16-15)。在大部分模体中，子类别 0 对应的 Ra 要显著小于该类模体的 Ra，而子类别 3 对应的 Ra 显著高于该类模体的 Ra，且该子类模体中的有向边都是正向调控的。这个结果表明 miRNA 不倾向靶向负向调控模体中的蛋白，而是倾向调控正向调控模体中的蛋白。

正向反馈环(positive feedback loop)是一类重要的网络模体，通常用来把短暂的信号转变成持续性的细胞响应(图 16-15)。在正向反馈环中，任何成分的噪声或者波动都容易被放大，从而使生物系统任意转换状态。这种情况下，miRNA 的负向控制能够增强对这些噪声或者波动放大的过滤或者缓冲作用。同转录抑制比较，miRNA 能更迅速地在转录后水平上调节靶蛋白的表达水平。因此，通过调控正向调控模体，miRNA 可以提供快速的反馈响应和更有效的噪声过滤。

另外，网络主题是由连接在一起的网络模体组成，即至少共享一个信号蛋白的网络模体，它是比模体更大的子图，代表信号蛋白间更高水平的调控关系，通常和特定的生物功能有关。通过寻找某个 miRNA 调控的网络主题，可以推测该 miRNA 调控的生物功能。为了达到这个目的，首先找到所有的至少有一个蛋白被特定一个 miRNA 调控的模体，分析这些模体中的某些部分是否能聚集成团。结果发现大部分 miRNA 可以找到一个或者两个包含多于 20 个信号蛋白的网络主题。另外，大部分网络主题和五个细胞机器中的一个或者更多是相关的。这五个细胞机器分别是转录机器、翻译机器、分泌机器、活力机制和电荷机器。

下面介绍细胞机器的构建及 miRNA 对不同细胞机器共享蛋白的调控倾向性。因为信号网络最下游的蛋白将会把信号传递给信号网络外的五类基本细胞机器中的蛋白，即信号终结蛋白。信号从细胞外传递到终结蛋白，在这里细胞机器是指信号传递到各类细胞机器的终结蛋白过程中所经过的信号蛋白。具体做法是找到受体蛋白到终结蛋白的所有最短路径及其经过的蛋白，然后按照终结蛋白类别，将对应的最短路径经过的蛋白也分为五类，即实现不同细胞机器功能所需要的信号蛋白，其中最短路径使用的是 Dijkstra 算法。结果发现 miRNA 不倾向调控这五类基本细胞机器的共享蛋白，

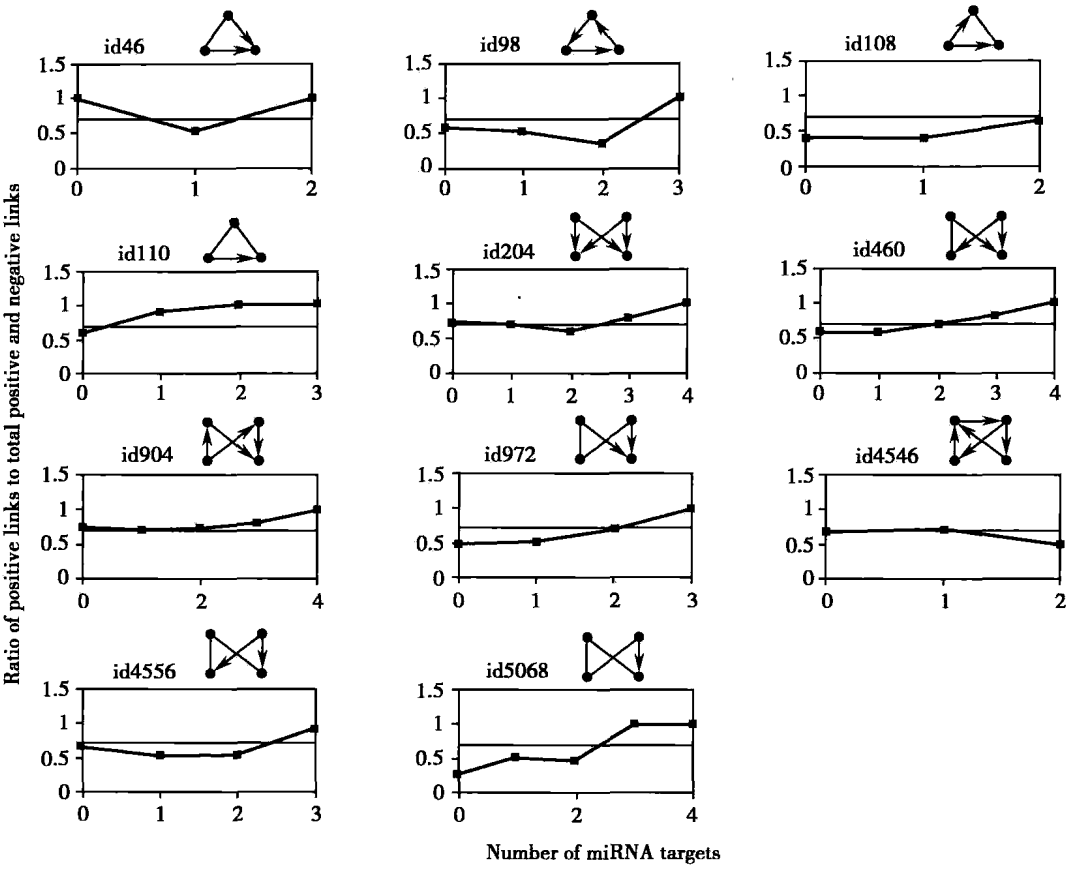


图 16-15 各种类型模体的所有子类中正向边的相对丰度

每类模体依据其中蛋白被 miRNA 调控的数目进一步分为一些子类，比如，三个点的模体基于 miRNA 靶点数目的多少可以被分为四类：0，1，2，3。计算每个子类中正向边占所有正向边和负向边的比值，对每类模体，可以把比值作为 miRNA 靶点数的函数，然后作图。该图来自 Principles of microRNA regulation of a human cellular signaling network

只有 14.3% 的共享蛋白被 miRNA 靶向，显著低于网络蛋白被靶向的比例。由此可以看出，这些机器的公用蛋白发挥着基础作用，它们的表达水平相对比较稳定，不需要过多的调控，所以 miRNA 倾向避免调控这些共用蛋白。

以上这些规律表明在人类中 miRNA 利用多种方式调控信号网络。通过选择性调控网络的大部分下游蛋白、下游成分的连接蛋白、正向调控模体，miRNA 能够终止以前存在的信息并快速促进和稳定对新信号的响应。另一方面，miRNA 不倾向调控网络的上游成分，例如配体、基本细胞机器共享的蛋白和负向调控模体中的信号蛋白。

二、miRNA 调控代谢网络

在生物系统中，代谢产物是很重要的。有些代谢产物例如氨基酸和脂肪酸参与发育、生长和许多细胞过程，同时其他代谢产物可以抵抗寄生作用。许多代谢产物可以被不同的代谢通路共享，且互相交织形成复杂的代谢网络。因为许多细胞活动都伴随着代谢的发生，所以控制代谢过程的速度是必要的，从而机体能对活体细胞内外信号的变化作出响应。代谢网络的控制机制是复杂的，涉及转录水平、转录后水平和翻译水平上的调控。通常认为代谢网络中的酶被转录因子紧紧控制。事实上，实验也验证了 miRNA 确实调控氨基酸代谢、胆固醇的生物合成等代谢过程。

miRNA 对信号网络有一些调控原则，那么 miRNA 对代谢网络的调控是否也存在一些类似或者代谢网络特有的原则呢？ Tibiche 等系统分析了人类代谢网络的 miRNA 调控模式。下面将以该研究的数据为基础进行详细探究。

首先，从KEGG数据库中获取人类代谢通路数据，然后用以反应为中心的模式来描述代谢网络，即有向图的方式，其中一个点代表一个反应或者该反应涉及的所有酶，对两个点（即反应）A和B，当A的代谢产物是B的一种反应底物时，就把A和B连接起来，方向指向B，即A→B。因为在代谢网络中，超过70%的点都被连接到最大网络组分（网络的一个最大子网，该子网中任意两个节点都可以通过其他节点相互连接）中，包含了大部分反应间的关联关系，所以后续分析只在该最大组分中进行。另外，把TargetScan预测得到的人类整个基因组的miRNA靶基因数据映射到该代谢网络上并分两种情况：第一种是反应只有一个酶催化，那么如果酶被miRNA调控，就认为该反应被miRNA调控；如果一个反应包含多个酶（不少于两个酶），统计这些酶中是miRNA靶基因的数目，通过与随机比较计算统计学显著性。随机方法是扰动miRNA与所有酶的调控关系，即真实情况下，有些代谢酶可能不受miRNA调控，但是在随机情况下可能会受到miRNA调控；反过来有些酶可能受到miRNA调控，但是随机扰动后却不受miRNA调控。在扰动后的网络中重新统计每个反应涉及的酶中是miRNA靶基因的数目，重复5000次，然后利用经验概率P估计显著性，即统计5000次随机中，比真实情况高的比例，因为多酶的反应都进行显著性检验，所以需要进行多重检验验证，可以通过Qvalue进行多重检验校正，校正后的显著性水平设为0.25，低于该值的都被认为是被miRNA显著调控的反应。

同信号网络一样，代谢网络中22%的节点都是miRNA的靶点，即和全基因组比较，miRNA的调控在代谢网络中也有相对更重要的作用。由于代谢网络中的节点按照信号网络的方式进行分类是没有意义的，所以采用基于代谢网络的结构特性进行划分。这里分五类，其中把没有入度的点称为上游节点（upstream node，UPN），即该反应的底物不是其他任何代谢反应的产物；同样把没有出度的点称为下游节点（downstream node，DSN），即该反应的产物不是其他任何反应的底物；把那些删除后能够增加网络组分数目的节点称为切割点（cut point，CP），即那些删除后破坏网络连通性的节点；另外一类是有较高出入度加和的节点，即高度连接的节点，称为hub（这里是网络中前5%出入度和大的节点），即这些反应的产物是多个反应的底物，同时这些反应的底物又是很多反应的产物；网络中剩余的所有节点称为中间节点（intermediate node，ITN）。这五类节点可以如图16-16形象地表示。为了计算每类节点被miRNA调控的显著性，通过从整个代谢网络中随机抽取与每类节点数目相同的节点，计算被miRNA调控的比例，重复这种操作5000次，计算两种类型的经验概率，即随机情况中高于真实情况和低于真实情况的比例，取0.05作为显著性水平。结果发现，ITN避免被miRNA调控且和随机比较是显著的；而miRNA靶点显著富集在hub和CP两个集合上。CP实际上是一种瓶颈蛋白，在网络中处于关键位置，控制着代谢流从一部分流动到另一部分。hub是那些被多个反应共享的反应，miRNA调控hub就能对代谢网络有全局的影响。例如miRNA调控柠檬酸合成酶，该酶能够影响78个代谢通路。也就是说miRNA能通过调控hub对代谢网络进行全局调控，另外通过控制CP蛋白对网络进行局部调控。

另外，代谢网络也有局部模块化结构，研究比较多的是代谢流，其对应某种特定的物质代谢过程，下面对这种结构进行分析。代谢流有两种组成方式：线性排列方式，即一个反应的产物是且只是另一个反应的底物，这里只考虑包含两个反应或者三个反应的线性模式；另外一种是支化模式，它可以进一步分为两种子模式：①一个反应的产物是另外两个反应的底物，这种子模式称分叉型；②两个反应的产物是另外一个反应的底物，该模式称为汇集型。这三种模式（图16-17）可以通过枚举网络中连接的两个反应或者三个反应的组合得到。然后依据每种类型中节点被miRNA调控的数目可以进一步进行划分，最后检验这些更细模式在网络中的富集和缺失情况。通过统计真实网络中这些模式的数目，然后和随机情况比较计算显著性。随机方法和上面多酶反应的随机方法类似，只不过是随机的miRNA对反应的调控关系，然后计算随机网络中每种更细模式的数目，共随机10 000次。计算统计两种经验概率，一种是随机情况中高于真实情况的比率，第二种是随机情况中低于真实情况的比率。结果表明所有的由两个或者三个反应组成的线性代谢流都被miRNA富集靶向，且与随

机比较，这种现象也是显著的。这表示代谢网络中的某些局部反应区域被miRNA显著调控。另外发现不论是汇集型还是分叉型，其中不包含任何miRNA靶点的模式在网络中是显著出现的，而至少包含两个miRNA靶点的模式在网络中是显著缺失的。

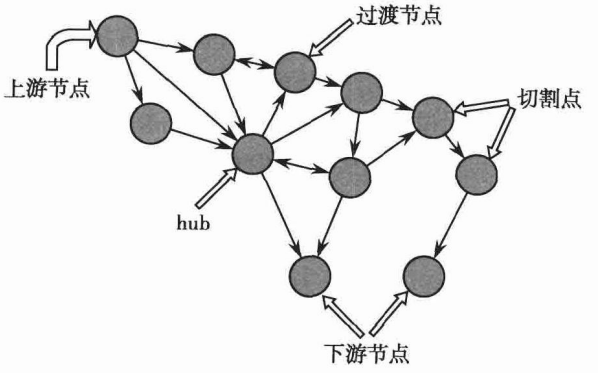


图16-16 代谢网络中五类节点类型

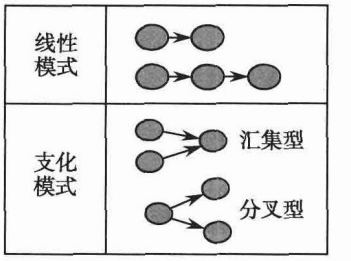


图16-17 代谢网络中出现的三种代谢流模式

在信号网络中，可以找到调控某种细胞机器的miRNA，那么代谢网络中miRNA倾向调控哪些代谢通路呢？下面分两种情况探讨miRNA调控和KEGG代谢通路的功能关联：①在已知的代谢通路上；②在网络背景下，如果一个通路包含miRNA靶向的下游节点或者下游节点的附近节点被miRNA调控，那么该通路就认为被miRNA调控。为了判断miRNA调控的富集性，进行和判断多酶的反应是否显著被miRNA调控类似的随机并进行多重检验校正，并以同样的阈值作为显著性水平。结果，在第一种情况下，六个通路被miRNA显著调控；在第二种情况下找到30个通路，二者总共包含32个通路。结果发现许多基本的代谢通路被miRNA广泛调控，例如氨基酸合成或降解和某些脂类代谢，暗示着miRNA在细胞中心代谢活动中起作用。

上面这些规律表明在人类中miRNA也存在多种方式调控代谢网络。通过选择性调控代谢网络的hub、切割点及线性代谢流，而避免调控代谢网络的过渡节点及支化代谢流，并且miRNA也特异性调控基本的代谢通路。

三、miRNA调控基因转录调控网络

基因转录调控网络描述转录因子及其调控的基因之间的关系。理论上，基因调控网络包含所有可能发生的基因调控关系和实现某种生物学功能的不同调控关系的组合机制。通过手工注释和高通量实验获得的基因调控关系使大范围地分析基因调控网络成为可能。但是，哺乳动物的基因调控信息还是远远不够的。基因调控网络也可以用有向图表示，其中点表示转录因子或者被调控的基因，边表示转录因子对基因的调控关系，箭头指向被调控的基因。

因为人类转录调控数据的特点，不能像前面两种网络那样分析，现在提出一种新的研究模式：即先在有限的试验获得的数据中寻找到miRNA对其的调控规律，然后再在预测的转录调控数据中进行验证和拓展分析。

首先是获得实验和预测的转录调控数据，实验数据是2005年Boyer等通过染色体免疫共沉淀-芯片法(Chromatin Immunoprecipitation-chip，ChIP-chip)检测了三个转录因子在人类胚胎干细胞中的调控靶点，这三个转录因子是OCT4、NANOG和SOX2；另外预测的转录结合位点数据来自Krek等的研究。miRNA靶基因数据来自PicTar的预测结果。

因为转录调控和前面两种网络不同，是一般网络，所以分析转录调控的靶基因是否倾向被miRNA靶向是没有意义的。进一步地，可以将三个转录因子的靶基因按照受多少个转录因子调控分成三组，即被一个转录因子调控、被两个转录因子和被三个转录因子调控。然后统计每组中被miRNA靶向的数目，以及不被miRNA靶向的数目。三组中miRNA比例差异的显著性通过Fisher

精确检验得到。结果发现 miRNA 的靶点富集在被多个(大于或等于 2)转录因子调控的基因集合上，也就是说被越多转录因子调控的基因，就越可能被 miRNA 调控。

上面实验获得的转录调控数据比较小，为了证实结果的鲁棒性(robustness)，需要进一步地在基因组范围内的调控信息中进行同样的分析。由于目前还没有实验获取的全基因组的基因调控数据，所以只能分析通过计算预测的转录因子在启动子上的结合位点信息。这种全基因组的分析也显示基因拥有的转录因子结合位点越多，即这个基因被越多的转录因子调控，该基因的转录后调控机制就越复杂。这个结论和前面实验数据获得的结论是一致的。反过来，也可以先按照 miRNA 的调控数目将基因分组，然后统计每组基因包含的转录因子结合位点的平均数，结果发现 miRNA 的调控数目和转录因子结合位点的平均数是显著正相关的，也就是说如果基因被越多的 miRNA 调控，那么这些基因就有越多的 TFBS 结合位点。上面的结果显示，在人类基因组中，转录后水平上 miRNA 的调控复杂度和转录水平上转录因子对基因的调控复杂度是正相关的，即在转录水平基因受到的调控越复杂，这些基因就越需要频繁开启，而且越有可能在不同时空下表达，因此这些基因也需要更频繁的关闭。miRNA 作为负向调控因子能够在转录后水平上通过抑制 mRNA 翻译或者降解 mRNA 实现基因功能的关闭。这也是基因表达的协同调控机制中一个新的发现。另外，通过对具有较多转录因子结合位点和被较多 miRNA 靶向的基因进行功能富集分析(功能数据来自 GO 数据库，用累计超几何方法计算富集的显著性)，结果发现这些基因富集在某些特定生物过程中，特别是那些和发育相关的生物过程。

在信号网络中，miRNA 倾向于调控信号网络中的正向调控环。但是，目前还没有人类或者啮齿类生物的关于基因的正向或负向转录调控关系的数据，所以现在还不能进行类似信号网络局部模块化的分析。

四、miRNA 调控蛋白互作网络

蛋白互作网络是机体中的许多生物学过程的基础，为更好理解蛋白质组的功能提供了有价值的框架。迄今为止，已经在酵母、大肠杆菌以及其他细菌、线虫、果蝇和人中进行了大规模的蛋白质相互作用的测定。因为蛋白互作网络的数据相对比较容易获得，所以对其进行的分析也比较广泛，从网络全局拓扑属性到网络模体、网络模块和网络进化。通常蛋白互作图用静态网络表示，点表示蛋白，边表示蛋白质之间的互作关系。事实上，蛋白互作网络是动态的：网络的功能状态依赖蛋白的表达，而蛋白质的表达受到时间和空间上不同调控机制的影响。

由于蛋白互作数据和上面三种网络的数据相比是比较全面的，因此蛋白互作网络中 miRNA 的转录后调控机制也是探讨最细致的。在此，将目前的研究方式分为三种：①与上面三种网络的分析方式一样，不区分对待 miRNA，统一平等对待所有的靶基因；②区分对待 miRNA，以单个 miRNA 的靶基因作为一个单元；③以 miRNA 簇(miRNA cluster)为单位，不区分对待簇内的 miRNA，把它们调控的靶基因作为一个单元。尽管下面的结果不是在同一蛋白互作数据和同一的靶基因数据中得到的，但是这些结果是可信的。

(一) 不区分对待 miRNA

在蛋白互作网络中，一个节点的直接互作邻居数目定义为该节点的度，是节点在网络中一个重要的属性，反应蛋白在网络中的局部中心性。结果发现度和 miRNA 靶点类型数是显著正相关的，且这种趋势不是因为具有不同度的蛋白的 3′UTR 长度和保守性差异所导致。通常一个蛋白的互作邻居越多，这个蛋白就会参与越多的生物过程，它的表达越具有动态性，也就越容易被转录因子和 miRNA 调控。蛋白互作网络中另一个重要的拓扑测度是聚类系数。聚类系数可以衡量蛋白互作邻居之间的连接紧密度。在蛋白互作网络中，hub 蛋白是一类重要蛋白质。研究表明在 hub 蛋白中 miRNA 靶点类型数和聚类系数是显著负相关的。推测有较高聚类系数的 hub 蛋白很可能是模块内的 hub 节点，它和它的互作邻居一起通过形成蛋白质复合物才能行使某种功能；而有较低聚类系数

的hub蛋白倾向是模块间的hub，会在不同的时空下选择不同的邻居。因此模块内的hub有相对较小的调控压力。

为了衡量miRNA调控复杂性和基因功能复杂性的关系，一种方法是用基因在组织中的表达广度反映基因功能的复杂性。研究表明，miRNA靶点类型复杂性和基因的功能复杂性显著正相关，即靶mRNA的组织表达广度越大，就有越多的miRNA靶点类型。

另外，因为编码互作蛋白的基因对应的表达是相似的，所以推断互作蛋白的miRNA调控机制也是相似的。研究表明，互作蛋白共享靶点类型数以及共享靶点类型的互作对数都比随机情况显著高，即在蛋白互作网络中，miRNA对互作蛋白进行协同调控。随机方法和前面代谢网络中多酶反应的随机方法是类似的。

（二）区分对待miRNA

这种分析本质是探索至少共享一个miRNA的靶基因，它们在蛋白互作网络中的特性。这是另外一种研究方法，可以丰富第一种研究方法的结果。以单个miRNA靶基因的集合为单位进行分析，发现miRNA的靶基因也倾向于网络中的hub和瓶颈蛋白。这里瓶颈蛋白是指在网络中介数大的蛋白集合。下面将探讨蛋白互作网络中两种相关子网络的模块化性质，第一种子网是单个miRNA靶点集合形成的子网（下面称为子网L0），另外一种是每个miRNA靶点集合与其直接互作邻居形成的子网（下面称为子网L1），这里模块化特征用子网的入度比来衡量。子网的入度比是子网中节点的入度比的平均值，节点的入度比是该节点在子网中边的数目占其在整个网络中边总数的比例，入度比越大，子网内的蛋白之间的连接要比它们与其他蛋白的连接频繁。通过计算，子网L0对应的平均入度比才是0.031，而子网L1的平均入度比是0.448，且二者有显著的区别。这些结果表明，被单个miRNA调控的基因不能形成模块，但是这些基因和它们的直接邻居联合起来就具有显著高的模块化性质。这里的显著性和前面代谢网络中五种蛋白的随机方法类似。

（三）以miRNA簇为单位

这种miRNA分类方式是介于上面第一种和第二种之间的一种过渡方式。考虑是具有共表达特性的miRNA靶基因的特性，这里同样发现互作的蛋白也倾向被miRNA簇中的miRNA共调控。一般地，两个蛋白共享miRNA簇的数目和它们在蛋白互作网络中的最短路径长度成反比，其中可以利用Dijkstra算法计算蛋白间的最短路径，它是典型的最短路径算法，用于计算一个节点到其他所有节点的最短路径。

这里分析了另外一种模块化结构社区（community），社区内的节点彼此紧密连接，但是社区间的连接是比较疏松的。利用软件Cfinder可以找到蛋白互作网络中的社区。当社区中至少50%的蛋白被某个miRNA簇调控，且该簇中的任何一个miRNA最多调控该社区中40%的蛋白，满足上面两种条件，就认为该miRNA簇调控这个社区。最后，可以找到一些miRNA簇显著共调控的社区。

总结三种研究方式，会发现转录后调控复杂度和互作复杂度、功能复杂度都是正相关的。不管是否区分对待miRNA，蛋白互作网络中hub蛋白都是倾向被miRNA调控。互作的蛋白具有一致的转录后调控模式。单个miRNA的靶点不具有模块化性质，加入靶点的蛋白互作邻居后模块化显著增强。同样可以找到miRNA簇显著调控的功能模块。

五、miRNA调控的网络模体

网络中在统计上显著出现的结构模式或者小的子网称为网络模体，在网络水平上它们是趋同进化的结果，并且在细胞内执行重要的功能。目前，实验数据证实了一些miRNA调控的网络模体，例如在线虫外阴细胞中，Notch信号蛋白可以激活*miR-61*的转录，而*miR-61*反过来在转录后水平抑制Notch信号蛋白的一个转录因子，从而稳定了外阴细胞的命运。在果蝇眼睛细胞以及分化的人类粒细胞中也发现了类似的环路。最近，一个转录因子-miRNA调控数据库TransmiR是通过人工阅读近5000篇文献，搜集到243条转录因子调控miRNA的信息，同时也给出调控关系参与疾病的注释信息。

在未来的工作中，转录因子和 miRNA 构成的反馈环或者前馈环信息也会被整合到 TransmiR 中。

目前，人和鼠中超过 80% 的 miRNA 位于编码区或者非编码基因的内含子区域，而且在组织和单个细胞水平中利用表达谱数据也证实了许多 miRNA 和其宿主基因的表达是高度相关的。这表明它们倾向以相同的速度进行共转录。这些事实可以得到一种假设，即在各种条件下，宿主基因的相对转录水平能够精确地表示内嵌的 miRNA 的表达水平。Tsang 等利用这一特性在人和鼠的基因组中系统地研究了 miRNA 的调控环路。

前馈环(Feedforward Loop)是模体的一种类型，它的模式是上游转录因子同时激活 miRNA 及其靶基因的表达，另外转录的 miRNA 也调控该靶基因。Tsang 等将前馈环分为两种，分别命名为Ⅰ型 miRNA 前馈环和Ⅱ型 miRNA 前馈环，如图 16-18。在Ⅰ型 miRNA 前馈环中，上游转录因子同时激活(或抑制)靶基因和 miRNA 的转录，且 miRNA 和靶基因间存在抑制关系。在Ⅱ型 miRNA 前馈环中，上游转录因子能抑制(或激活)靶基因的转录，且同时激活(或抑制)miRNA 的转录，该 miRNA 又抑制靶基因的翻译。下面是实验证实的Ⅰ型 miRNA 前馈环的一个例子：miR-17-5p 抑制 E2F1，而两者在人类细胞中都被 c-Myc 转录激活。这种类型环路可能具有调控和信号处理的功能。

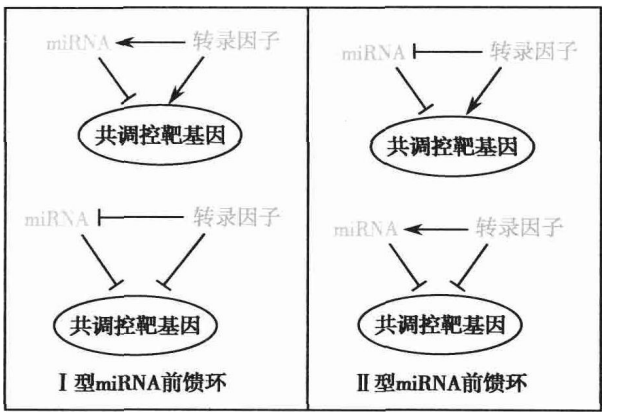


图 16-18 Ⅰ型和Ⅱ型 miRNA 前馈环路

Tsang 等分析了由人和鼠组成的 61 个组织 / 细胞类型的 mRNA 表达谱，发现具有Ⅱ型前馈环的 miRNA 占了位于内含子 miRNA 的很大一部分。这一结果和以前的结果是一致的，即在 miRNA 表达的组织中，一些组织特异表达的 miRNA 靶基因的表达量是比较低的。另外，也发现了相当一部分的Ⅰ型 miRNA 前馈环，特别是在成熟的神经细胞中。

在Ⅱ型 miRNA 前馈环中，miRNA 对其靶基因的调控与转录调控方向是一致的，因此在转录后水平上增强了对靶基因的调控。这种环路可以使细胞有效抑制那些在转录水平被遗漏的靶基因。另一方面，Ⅰ型 miRNA 前馈环可以阻止噪声带来的转变，如 c-Myc/E2F1/miR-17-92 网络。

识别 miRNA 网络模体的另一个方向是识别转录因子和 miRNA 间的协调调控，及它们所调控的共同的靶基因。可以通过统计学检验或者随机方法找到显著共享靶基因的转录因子 -miRNA 对，构建转录因子 -miRNA 的共调控网络。结果发现转录因子 - 转录因子和 miRNA-miRNA 的共调控是常见的，miRNA- 转录因子的协同调控是较少见的，且显著关联的 miRNA- 转录因子对中，有一些是通用转录因子，它们有很多靶基因，这些转录因子几乎和所有的 miRNA 都显著协同调控。最后在该共调控网络中寻找网络模体，与 Tsang 等的结果类似，这种方法也发现了Ⅰ和Ⅱ型前馈回路。同时，还发现了另外两类 miRNA 调控的网络模体：混合调控环和间接前馈环(图 16-19)。在前者中，转录因子在转录水平上同时激活 miRNA 和共调控靶基因，且 miRNA 在转录后水平上同时抑制转录因子和共调控靶基因。在后者中，转录因子 1 同时激活另外一个转录因子 2 和共调控靶基因，另外，转录因子 2 激活一个 miRNA，而这个 miRNA 又抑制共调控靶基因。但是这两种类型模体的功能至今还不清楚。

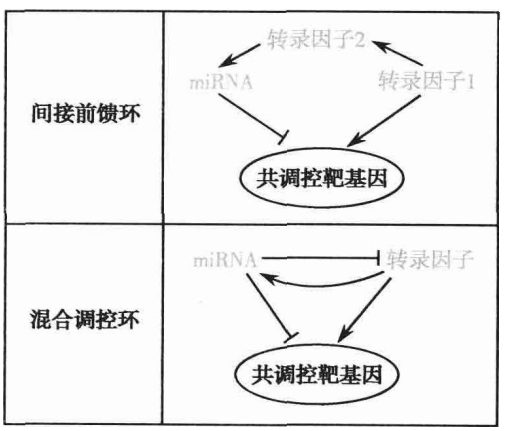


图 16-19 两类 miRNA 调控的网络模体：混合调控环和间接前馈环

总之，在基因组中miRNA作为一类重要的转录后调控因子广泛参与转录后基因调控。通过分析miRNA和“一般网络”间的调控关系，可以得到一个共同的规律：miRNA倾向靶向有高度调控复杂性的基因。这个事实暗示了基因调控中存在转录调控水平和转录后调控水平上的协同调控。另外，对“特异细胞网络”，miRNA有一些不同于“一般网络”的调控策略。例如，miRNA倾向调控信号网络的正向调控模体、高连接的支架蛋白和网络的下游成分，这种方式能够终止预先存在的信号，快速响应新信号。另一方面，miRNA不倾向调控信号网络中的负向调控模体、基本生物机制的共享蛋白以及网络的上游成分。在代谢网络中，miRNA选择性调控有特定网络结构特征的基因，从而有效地调控细胞代谢。

小　结

在这一章里阐述了miRNA和复杂疾病尤其是与癌症的关系，以及miRNA可以作为一种新的生物学标记应用于各类癌症研究。由于每个miRNA可以调控几十个甚至上百个靶基因，所以如何准确地鉴别出miRNA在癌症中的靶点仍然是一项艰巨的任务。在本章的第一节与第二节中总结了一些已有的miRNA靶基因预测算法和相关的数据库资源，例如TarBase、miRBase和miRGen数据库等，期望这些miRNA靶基因预测算法和数据库资源能够被基本了解并学会使用。为了进一步说明miRNA在癌症的发生和发展过程中具有重要作用，第三节从miRNA的序列角度阐述miRNA多态与癌症的关系，例如miRNA靶点的多态、结合位点上下游的多态等。近年来，随着miRNA检测技术的不断发展，大量的miRNA表达谱数据已经广泛应用于癌症诊断或预后，第四节里介绍了miRNA表达谱在人类癌症中的应用。通过讲解miRNA表达谱研究中寻找癌症相关miRNA和对癌症进行分类的基本流程和结果，说明miRNA的表达异常与癌症的发生密切相关及利用miRNA表达谱能够准确对癌症进行分类，并且整合分析miRNA表达谱与mRNA表达谱能够提高癌症分类的效能，之后还介绍了在miRNA表达谱研究中一些miRNA可以作为生物学标记。最后在第五节详细阐述了miRNA如何调控细胞信号网络，代谢网络，基因转录调控网络，蛋白互作网络和miRNA调控的网络模体。从分子功能上说明miRNA作为一种重要的转录后调控因子广泛参与转录后基因调控。

Summary

In this chapter, we are focus on delineating the relations between miRNAs and complicated diseases, especially in cancers. MiRNAs have been served as novel biomarkers that applied to various cancers. As each miRNA can regulate hundreds of genes, how to exactly find targets corresponding to miRNAs is still a challenge. So we firstly summarized several known miRNA target gene prediction algorithms and database resources in the first and the second sections, such as, TarBase, miRBase and miRGen. We hope that these tools and databases could be understanded and used by students. To further explain the fact that miRNAs play the important roles in the process of the onset and development of cancers, we illustrated the relationship of cancer with miRNAs' polymorphisms, for instance, the polymorphisms of miRNA targets, upstream or downstream of binding sites, and so on. Moreover, with the development of detecting technologies of miRNAs, a large number of miRNAs have been used in the diagnosis and prognosis of cancers, and hence in the fourth section we illustrated the utilization of miRNA expression profiles in human cancer researches. We described the basal workflows of identification of cancer-related miRNAs and miRNA-based classification of cancers and applied

them to known miRNA data. Resutls suggested that abnormal expressions of miRNAs occurred in some cancers, utilization of miRNA expression profiles are able to precisely classify the cancers, and the integration of miRNA and mRNA expression profiles can increase the efficacy of classifying cancers. We then introduced that some miRNAs could be biomarkers in the diagnosis of various cancers. Lastly, in the fifth section we detailedly introduced how miRNAs regulate cellular signaling networks, metabolism networks, transcriptional regulatory networks and protein interaction networks, and miRNA regulatory network motifs, suggesting that miRNAs serve as one of the important post-transcriptional regulatory factors widely participate in regulation of genes.

（李 霞 肖 云 徐 娟 王莹莹 宁尚伟）

习 题

1. 简述 miRNA 的起源过程。
2. 简述任意三个本章介绍的基于序列的 miRNA 靶预测方法。
3. miRNA 靶数据库有哪些？并简单介绍。
4. 位于 miRNA 基因内影响 miRNA 形成和功能的多态可以分为几类？并简单介绍。
5. 简述基于 miRNA 表达谱对复杂疾病分析流程。
6. 在 miRNA 调控蛋白互作网络中，如何计算单个 miRNA 靶点集合形成的子网的统计学显著性？
7. 从 KEGG 下载人类通路数据后，如何将网络转化为以反应为中心的网络模式，然后将网络模式包含的节点分成正文提出的五种类型。

主要参考文献

1. Bartel D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 2004, 116(2): 281-297.
2. Cimmino A., Calin G. A., Fabbri M., et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc. Natl. Acad. Sci. U S A., 2005, 102(39): 13944-13949.
3. Cui Q., Yu Z., Purisima E. O., et al. Principles of microRNA regulation of a human cellular signaling network. Mol. Syst. Biol., 2006, 2: 46.
4. John B., Enright A. J., Aravin A., et al. Human MicroRNA targets. PLoS. Biol., 2004, 2(11): e363.
5. Lewis B. P., Shih I. H., Jones-Rhoades M. W., et al. Prediction of mammalian microRNA targets. Cell, 2003, 115(7): 787-798.
6. Liang H., Li W. H. MicroRNA regulation of human protein protein interaction network. RNA, 2007, 13(9): 1402-1408.
7. Liu B., Li J., Tsykin A. Discovery of functional miRNA-mRNA regulatory modules with computational methods. J. Biomed. Inform., 2009, 42(4): 685-691.
8. Lu J., Getz G., Miska E. A., et al. MicroRNA expression profiles classify human cancers. Nature, 2005, 435(7043): 834-838.
9. Mattie M. D., Benz C. C., Bowers J., et al. Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies. Mol. Cancer, 2006, 5: 24.
10. Mendes N. D., Freitas A. T., Sagot M. F. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res., 2009, 37(8): 2419-2433.
11. Tibiche C., Wang E. MicroRNA Regulatory Patterns on the Human Metabolic Network. Open Syst. Biol. J., 2008, 1: 1-8.
12. Tsang J., Zhu J., van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol. Cell, 2007, 26(5): 753-767.

中英文对照索引

A

B

C

D

E

F

G

H

J

K

L

X

Y

Z

总平均疏水性	grand average of hydropathicity	95
组蛋白乙酰转移酶	histone acetyltransferase, HAT	362
组件	module	262
组学	Omics	3
祖先重建	ancestral reconstruction	116
最大距离	complete linkage	192
最大期望算法	expectation-maximization algorithm, EM	421
最简约重建	most parsimonious reconstruction	116
最小等位频率	minor allele frequency, MAF	399
最小二乘法	least-squares, LS	115
最小进化树	minimum evolution tree	116
最小距离	single linkage	192

英中文对照索引

C

D

E

F

G

H

I

N

O

P

S

T

U

[General Information]
书名=生物信息学
作者=李霞主编
页数=478
SS号=12622019
出版日期=2010.08