



四川轻化工大学课程实施大纲

课程名称：生物信息学

授课班级：2022 级生物制药 1、2、3、4 班

任课教师：陈咏梅、周明、李诗琦

工作部门：化学工程学院

联系方式：电话：13558910780
QQ：676555924

四川轻化工大学制
2025 年 8 月

《生物信息学》课程实施大纲

基本信息

课程代码：16541022

课程名称：生物信息学

学分：2

总学时：32

学期：2025-2026第1学期

上课时间：6-13周

上课地点：N5-1803

答疑时间和方式：课堂、学习通、电子邮件、QQ、电话

答疑地点：课堂、办公室、学习通、班级 QQ 群

授课班级：生物制药2022级1、2、3、4班

任课教师：陈咏梅、周明、李诗琦

学院：化学工程学院

邮箱：676555924@qq.com

联系电话：13558910780

目录

1. 教学理念	1
1.1 关注学生的发展	1
1.2 关注教学的有效性	1
1.3 关注教学的策略	2
1.4 关注教学价值观	2
2. 课程介绍	2
2.1 课程的性质	2
2.2 课程在学科专业结构中的地位、作用	2
2.3 课程的前沿及发展趋势	3
2.4 课程与经济社会发展的关系	4
2.5 学习本课程的必要性	4
3. 教师简介	4
3.1 教师的职称、学历	4
3.2 研究兴趣（方向）	4
4. 先修课程	5
5. 课程目标	5
5.1 知识与技能方面	5
5.2 过程与方法方面	5
5.3 情感、态度与价值观方面	5
6. 课程内容	6
6.1 课程的内容概要	6
6.2 教学重点、难点	6
6.3 学时安排	7
7. 课程实施	7
7.1 第一章 绪论	7
7.2 第二章生物数据库	12
7.3 第三章关键词和词组为基础的数据库检索	18
7.4 第四章 核酸和蛋白质序列为基础的数据库检索	20
7.5 第五章多序列对位排列分析和系谱分析	24
7.6 第六章基因预测和基因结构分析	27
7.7 第七章蛋白质性质和结构分析	29
7.8 第八章农业类数据库的利用	33
7.9 第九章核酸序列的其他分析方法	37
8. 学生课程学习要求	44
8.1 学生自学的要求	44
8.2 课外阅读的要求	44
8.3 课堂讨论的要求	44
8.4 课程实践的要求	44
9. 课程考核方式及评分规程	44
9.1 出勤（迟到、早退等）、作业、报告等的要求	44
9.2 成绩的构成与评分规则说明	44
9.3 考试形式及说明（含补考）	44
10. 学术诚信规定	45

10.1	考试违规与作弊	45
10.2	杜撰数据、信息等	45
10.3	学术剽窃等	45
11.	课堂规范	45
11.1	课堂纪律	45
11.2	课堂礼仪	45
12 .	课程资源	45
12.1	教材与参考书	45
12.2	主要文献资料及相关数据库	45
12.3	课程网站等支持条件	45
13.	教学合约	45
13.1	教师的师德师风承诺	45
13.2	阅读课程实施大纲，理解其内容	46
13.3	同意遵守课程实施大纲中阐述的标准和期望	46
14.	其他说明	46

1. 教学理念

1.1 关注学生的发展

作为一名高校教师，应该明确认识到教学过程中真正的主体是学生。在课堂教学的过程中，始终关注学生的全面发展，帮助学生把握自身专业的特点和发展方向，把学生的个人发展作为教学的重中之重。对于制药专业的本科学生，经过数年的大学教育后，大多数将会进入到制药相关的行业，成为我国医药事业建设中的一员，因此学生的全面发展不仅与其自身的职业发展密切相关，而且会影响整个医药行业的发展。作为大学教师，我们需要给予学生的不仅仅是专业知识和技能，更需为其未来的发展进行筹谋，培养其综合能力，使其在毕业后能够在相关领域有一番建树。

1.2 关注教学的有效性

随着国际竞争的日趋激烈，教育的竞争，尤其是针对肩负重任的青年一代的大学教育职责越来越大，而对于当代大学生而言，要在将来有很好的发展，必须在其学生阶段打下良好的知识基础和能力基础，无论从个人的发展还是民族的振兴而言，实现全民知识化，增强个人创造力都是一种必然的趋势。因此，大学本科教育必须强化教学的有效性，包括增强学生的知识储备和能力的提升。

知识基础：《生物信息学》作为现代生物科学、制药、生物技术、计算机等相关学科的交叉学科，又是处于生命科学前沿的极具潜力的独立学科。因此，在课堂的教学过程中，教师应该仔细研读教材，使学生通过学习生物信息软件预测生物大分子的结构与功能、物质代谢、遗传信息传递等方面的基本理论和基本知识，了解现代生物信息学在医药科学中的新进展。熟悉最基本的生物信息学实验方法和操作技术，并适当地学习最新的生物信息学分析技术；使学生具备运用生物信息学知识理解、分析、解决医学、药物研发等问题，解释相应病理生理现象的能力。同时，增加当今生物信息学研究领域的前沿、热点或工程实践中的难点等相关知识，讲解时做到逐渐展开，层层深入，为学生步入社会后较快、较好的适应工作岗位，真正做到学以致用打下良好基础。

能力基础：在课堂教学过程中教师除了传授专业知识，更重要的是要培养学生的综合能力（学习能力、分析问题能力、解决问题能力、为人处世能力等）。目前，存在大学本科毕业的学生找工作相对较难，且在进入工作岗位后存在与企业需求脱节的现象。反应出高等教育工作者在大学教学过程中往往过多的纠结于纷繁复杂的公式推导或理论知识讲授，而忽视了公式或理论知识的灵活实际应用，导致学生在进入工作岗位后，发现学到的许多知识似乎仅是“课本上的知识”，实际上并没有实际应用价值，而一些需要学习的知识却没有学到。但从长远来看，大学本科毕业的学生在专业道路上明显要比专科或高职院校毕业的学生走得更远，这必然与大学本科教育注重综合能力的培养密切相关。针对这个问题，作为专业课的教师，除了在授课过程中向学生传授理论知识，对学生的综合能力产生潜移默化的影响外；还需有针对性地设计一些相应的教学环节，例如：增加当今研究热点或工程实践中具体案例的分析；引导学生针对具体应用中遇到的问题，查阅资料，分析问题产生的原因并提出解决方案；鼓励学生针对授课过程中感兴趣的点进行资料查阅并制作 PPT 进行汇报，扩充知识面；增加相配套的实验（实践）课程的比

重，尤其是其中的综合设计类型的比重，让学生充分实践分析问题和解决问题的能力，使其在毕业后的生活和工作中有更高层次的发展潜力。

1.3 关注教学的策略

采用合理的教学策略对于教学的有效性起着至关重要的作用。在《生物信息学》这门课程的讲授中，课堂教学仍以教师讲授为主，辅以提问、作业、学生报告、在学习通平台上进行每章节结束后的测验、期中考试，课堂内容整理记录笔记等多种教学方式。在课堂授课中不求贪多，把一些重点知识、难点知识，尤其是与现代研究热点或实际生产相关的知识点讲透。对于具体的应用多列举一些案例加深学生印象；围绕多种疾病的致病机制，以及如何利用生物信息学对“结构、性质和功能”这一主要脉络进行生物药物的设计和开发两个核心问题，采用启发式教学，组织学生查阅相关的资料后进行分组讨论，提高学生的课堂参与度，进一步激发学生的学习热情和兴趣，增长学生的见识。除了课程理论教学外，本门课程配套进行了实验教学环节，使学生通过相关的理论学习和操作，能够对课堂讲授的内容有更直观、深刻的认识，从而提高学生的实践动手能力。

1.4 关注教学价值观

通过大学本科的教育，除了使我们所培养的学生具有较强的生物制药专业知识，能够在将来的研究或工作中学以致用外；更重要的则是培养他们学习知识的能力，分析问题、解决问题的能力 and 为人处世的态度等；这样才能让他们在将来能够较好的适应工作岗位，在日益激烈的竞争中脱颖而出，闯出属于自己的一片天地。

2. 课程介绍

2.1 课程的性质

生物信息学（Bioinformatics）是一门交叉学科。它包含了生物信息的获取、管理、分析、解释和应用在内的所有方面。它综合运用生物学、计算机科学和数学等多方面知识与方法，来阐明和理解大量生物数据所包含的生物学意义，并应用于解决生命科学研究和生物技术相关产业中的各种问题。生物信息学主要有三个组成部分：建立可以存放和管理大量生物信息学数据的数据库；研究开发可用于有效分析与挖掘生物学数据的方法、算法和软件工具；使用这些工具去分析和解释不同类型的生物学数据，包括 DNA、RNA 和蛋白质序列、蛋白质结构、基因表达以及生化途径等。生物信息学这个术语从20世纪90年代开始使用，最初主要指的是 DNA、RNA 及蛋白质序列的数据管理和分析。自从20世纪60年代就有了序列分析的计算机工具，但是那时并未引起人们很大的关注，直到测序技术的发展使 GenBank 之类的数据库中存放的序列数量出现了迅猛的增长。现在该术语已扩展到几乎覆盖各种类型的生物学数据，如蛋白质结构、基因表达和蛋白质互作等。

2.2 课程在学科专业结构中的地位、作用

四川轻化工大学开设的生物制药专业旨在培养具有扎实的生物技术和药学基础理论、基本知识，熟练掌握现代生物制药生产的原理、技术和方法，了解生物制药企业设备、生产、储存、销售和管理等环节的基本知识和技能，具有良好的开拓精神、创新意识和实践能力，能够胜任现代生物制药企业及其相关的科研

院所岗位基本要求的德、智、体、美全面发展的应用型高级专业人才。而《生物信息学》在生物制药专业学生的整个培养发展中起到了重要的作用。一方面,《生物信息学》与人类生活息息相关,随着新冠肺炎疫情的爆发,更需要学生对疾病的致病机理以及药物开发的机制等有更深入的了解,作为相关专业学生从理论课程到专业理论课程的过度课程,能够充分地调动学生的学习积极性,起到较好的承上启下作用;另一方面,《生物信息学》作为《微生物学》、《细胞生物学》、《基因工程》和《药理学》等课程的延续,与其紧密相连,并与其一起构成了生物制药专业完整的专业课程体系。

2.3 课程的前沿及发展趋势

生物信息学是建立在分子生物学的基础上的,因此,要了解生物信息学,就必须先对分子生物学的发展有一个简单的了解。研究生物细胞的生物大分子的结构与功能很早就已经开始,1866年孟德尔从实验上提出了假设:遗传因子是以生物成分存在,1871年 Miescher 从死的白细胞核中分离出脱氧核糖核酸(DNA),在 Avery 和 McCarty 于1944年证明了 DNA 是生命器官的遗传物质以前,人们仍然认为染色体蛋白质携带基因,而 DNA 是一个次要的角色。1944年 Chargaff 发现了著名的 Chargaff 规律,即 DNA 中鸟嘌呤的量与胞嘧啶的量总是相等,腺嘌呤与胸腺嘧啶的量相等。与此同时, Wilkins 与 Franklin 用 X 射线衍射技术测定了 DNA 纤维的结构。1953年 James Watson 和 Francis Crick 在 Nature 杂志上推测出 DNA 的三维结构(双螺旋)。DNA 以磷酸糖链形成发双股螺旋,脱氧核糖上的碱基按 Chargaff 规律构成双股磷酸糖链之间的碱基对。这个模型表明 DNA 具有自身互补的结构,根据碱基对原则, DNA 中贮存的遗传信息可以精确地进行复制。他们的理论奠定了分子生物学的基础。DNA 双螺旋模型已经预示出了 DNA 复制的规则, Kornberg 于1956年从大肠杆菌(E.coli)中分离出 DNA 聚合酶 I (DNA polymerase I), 能使4种 dNTP 连接成 DNA。DNA 的复制需要一个 DNA 作为模板。Meselson 与 Stahl (1958) 用实验方法证明了 DNA 复制是一种半保留复制。Crick 于1954年提出了遗传信息传递的规律, DNA 是合成 RNA 的模板, RNA 又是合成蛋白质的模板,称之为中心法则(Central dogma), 这一中心法则对以后分子生物学和生物信息学的发展都起到了极其重要的指导作用。经过 Nirenberg 和 Matthai (1963) 的努力研究, 编码20氨基酸的遗传密码得到了破译。限制性内切酶的发现和重组 DNA 的克隆(clone) 奠定了基因工程的技术基础。正是由于分子生物学的研究对生命科学的发展有巨大的推动作用, 生物信息学的出现也就成了一种必然。2001年2月, 人类基因组工程测序的完成, 使生物信息学走向了一个高潮。由于 DNA 自动测序技术的快速发展, DNA 数据库中的核酸序列公共数据量以每天106bp 速度增长, 生物信息迅速地膨胀成数据的海洋。毫无疑问, 我们正从一个积累数据向解释数据的时代转变, 数据量的巨大积累往往蕴含着潜在突破性发现的可能, “生物信息学”正是从这一前提产生的交叉学科。粗略地说, 该领域的核心内容是研究如何通过对 DNA 序列的统计计算分析, 更加深入地理解 DNA 序列, 结构, 演化及其与生物功能之间的关系, 其研究课题涉及到分子生物学, 分子演化及结构生物学, 统计学及计算机科学等许多领域。生物信息学是内涵非常丰富的学科, 其核心是基因组信息学, 包括基因组信息的获取, 处理, 存储, 分配和解释。基因组信息学的关键是“读懂”基因组的核苷酸顺序, 即全部基因在染色体上的确切位置以及各 DNA 片段的功能; 同时在发现了新基因信息之后进行蛋白质空间结构模拟和预测, 然后依据特定蛋白质的功能进行药物设计。了解基因表达的调控机理也是生物信息学的重要内容, 根据生物

分子在基因调控中的作用，描述人类疾病的诊断，治疗内在规律。它的研究目标是揭示“基因组信息结构的复杂性及遗传语言的根本规律”，解释生命的遗传语言。生物信息学已成为整个生命科学发展的的重要组成部分，成为生命科学研究的前沿。

2.4 课程与经济社会发展的关系

众所周知，当今经济社会面临着多种危机，如粮食危机、能源匮乏、资源紧缺、生态恶化、人口爆炸、老龄化等。而生物信息学的基础知识和研究技术因为其独特的性质，在解决当今经济社会中人类面临的的各种危机中发挥了其不可替代的作用。具体而言，利用生物信息学的方法和技术进行微生物改造的筛选，可提高土壤的肥力、改进作物的特性、促进粮食增产、防治病虫害，甚至在各种食品的生产过程中都可发挥重要的作用（粮食）；生物信息学方法可用于再生资源产乙醇、甲烷、氢气，提高石油采收率等（能源）以及将地球上的可再生资源转化为各种化工及制药等所需的原料（能源）、菌群的筛选；用于生产肥料、杀虫剂，可用于净化污水，可用于制造可降解材料等环境保护方面（生态恶化）；生物信息学与人类健康密切相关，尤其是抗生素、抗癌药物的开发和发展取得了瞩目的成绩，从而推动了生物信息学的快速发展，进一步为经济社会的发展贡献力量。同时，此次新冠肺炎大流行也亟待科学家们在对病毒机制的深入研究基础之上尽快开发出特效药物以及预防和治疗疫苗。

2.5 学习本课程的必要性

生物信息学是在生命科学的研究中，以计算机为工具对生物信息进行储存、检索和分析的科学。它是当今生命科学和自然科学的重大前沿领域之一，同时也将是21世纪自然科学的核心领域之一。其研究重点主要体现在基因组学

（Genomics）和蛋白质组学（Proteomics）两方面，具体说就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。在生命科学大发展的今天，生物信息学作为生物制药专业多门主干课程的基础支撑课程，起着承上启下、多学科贯穿的重要作用。通过本课程学习，要求学生掌握生物信息学的基本理论知识，使学生掌握预测生物大分子的结构与功能、物质代谢、遗传信息传递等方面的基本理论和基本知识，了解现代生物信息学在医药科学中的新进展。熟悉最基本的生物信息学操作技术，并适当地学习最新的生物信息学分析技术；使学生具备运用生物信息学知识理解、分析、解决医学问题，解释相应病理生理现象的能力，并进一步运用这些知识和技能去进行药物研发和生产，从而造福人类和推动社会的进步。

3. 教师简介

3.1 教师的职称、学历

陈咏梅，博士，讲师；周明，博士，讲师；李诗琦，博士，讲师；

3.2 研究兴趣（方向）

陈咏梅：群体遗传学；进化基因组学；生物制药；

周明：主要基于多组学（基因组、转录组、代谢组）联合分析对动物特异表型（肥胖、抗癌、长寿、极端环境适应等）适应性进化分子机制中关键基因的挖掘及功能验证；天然药物关键有效成分基因的挖掘及新药开发；动/植物适应性进化的分子机制；

李诗琦：天然药物关键有效成分基因的挖掘及新药开发；动/植物适应性进化的分子机制，群体遗传学。

4. 先修课程

《细胞生物学》、《基因工程》、《生物信息学》等基础学科。学生如若要学习并掌握好本门课程，需要提前复习这些相关课程，这样才能够在进行本课程的学习时进行较好的运用。

5. 课程目标

5.1 知识与技能方面

(1) 掌握核酸和蛋白质序列分析的基本原理和常用方法，了解分子系统进化的理论，熟悉常用的生物信息学数据库，了解基因组信息分析、基因表达数据分析和大分子结构预测等方面研究热点。

(2) 掌握常用生物信息学序列分析软件的使用和数据分析方法，掌握常用生物信息学数据库的检索方法；熟悉一些生物信息学在线分析工具的功能。

(3) 加强家国情怀和科技强国的意识，能认识不断探索和学习的必要性，具有自主学习和终身学习的意识。

5.2 过程与方法方面

从与学生生活息息相关的方面入手，调动学生学习生物信息学课程的兴趣，使其了解学习生物信息学的重要意义。让学生知道生物信息学很有用，从而喜欢上生物信息学。从生物信息学课中，得到很多与健康、疾病、医药、营养、保健、防病和治病等有关的知识。这些知识可以受用一辈子，而且也可以将这些生物信息学知识传播给家人和朋友。电视上每天都充斥着各种骗人的医药、保健品的广告，这些广告利用的就是大众缺乏生化知识这一点。爱斯基摩人为什么少得心血管疾病？正常的人需要补脑黄金（DHA）吗？为什么过夜的韭菜不能吃？骆驼为什么几个月可以不喝水？为什么狗急会跳墙、人急会生智？蜘蛛丝和钢筋相比，哪一个强度更强？生化中的“近朱者赤近墨者黑”是指什么？在学完这门课程以后，你自然就知道这些问题的答案了。

其次，学好生物信息学还是学好生命科学其他课程的基础，比如细胞生物学、遗传学和分子生物学、植物生理学。而在教学过程中，除了采用传统的讲授法向学生传授相关知识外，本课程还将采用启发性的教学方式，具体来讲即对学生进行分组，各组针对某一重点问题，查阅最新的文献资料后进行总结归纳和讲解，全班同学进行提问或补充。这样不仅能充分调动学生的思考能力和获取新知识的能力，有效启发其学习兴趣，而且能够使学生了解生物信息学领域相关的最新发展动态，明白生物信息学理论和技术在制药领域的重要意义，从而为其将来进入工作岗位或从事科学研究打下基础。

生物信息学作为生物制药专业多门主干课程的基础支撑课程，起着承上启下、多学科贯穿的重要作用。课程从分子水平揭示生命本质，为药物研发、工艺优化提供理论基础。通过本课程学习，通过本课程教学，要求学生掌握预测生物大分子结构功能、代谢调控及基因表达机制，培养学生理解、解释相应病理生理现象，并能分析和解决生物制药复杂工程问题的能力。具备从事生物药物研究和分析的基本能力，进而提升创新、创业意识和能力。

5.3 情感、态度与价值观方面

通过本课程的学习，不仅使学生掌握生物信息学的基础理论、知识和技术相关，而且培养学生获取新知识的能力和实事求是的科学作风。与课程相关的实验环节，不仅很好地加深了学生对理论知识的理解程度，培养了学生的动手能力；而且培养了学生科学严谨、实事求是、尊重实验结果的科研道德；不弄虚作假、

尊重他人成果的科学态度。而这种能力、道德及态度将继续深深地影响着学生的发展，使其在将来的学习、生活和工作中都受益匪浅。

6. 课程内容

6.1 课程的内容概要

生物信息学不仅仅是生物学知识的简单整理和数学、物理学、信息科学等学科知识的简单应用。海量数据和复杂的背景导致机器学习、统计数据分析和系统描述等方法需要在生物信息学所面临的背景之中迅速发展。巨大的计算量、复杂的噪声模式、海量的时变数据给传统的统计分析带来了巨大的困难，需要像非参数统计（BMC Bioinformatics, 2007, 339）、聚类分析（Qual Life Res, 2007, 1655-63）等更加灵活的数据分析技术。高维数据的分析需要偏最小二乘（partial least squares, PLS）等特征空间的压缩技术。在计算机算法的开发中，需要充分考虑算法的时间和空间复杂度，使用并行计算、网格计算等技术来拓展算法的可实现性。如基因表达谱分析，代谢网络分析；基因芯片设计和蛋白质组学数据分析等，逐渐成为生物信息学中新兴的重要研究领域；在学科方面，由生物信息学衍生的学科包括结构基因组学，功能基因组学，比较基因组学，蛋白质组学，药物基因组学，中药基因组学，肿瘤基因组学，分子流行病学和环境基因组学，成为系统生物学的重要研究方法。从发展不难看出，基因工程已经进入了后基因组时代。

6.2 教学重点、难点

6.2.1. 生物信息学数据库概述及检索

教学重点：（1）生物信息学数据库的分类；（2）基于关键词的检索方法；（3）基于序列相似性的检索方法。

教学难点：（1）基于序列相似性的检索方法。

6.2.2. 多重比对和系统发育分析

教学重点：（1）多重比对的方法；（2）系统分析的方法；（3）进化树的构建。

教学难点：（1）多重比对的算法原理；（2）系统分析的原理。

6.2.3. 蛋白质结构预测与分析

教学重点：（1）蛋白质三级结构预测的方法；（2）蛋白质跨膜结构域、信号肽、亚细

胞定位的预测工具使用方法。

教学难点：（1）PDB 数据库的使用；（2）蛋白质跨膜结构域、信号肽、亚细胞定位的预测工具使用方法。

6.2.4. 生物信息学常用计算机技术

教学重点：（1）MEGA 软件；（2）Python 语言和 R 语言基础；（3）引物设计等在线分析工具；（4）Origin 作图。

教学难点：（1）Python 语言和 R 语言基础。

6.3 学时安排

章（或节）	主要内容	课程思政映射与融入点	学时安排
第一章	(1)数据库目录及分类;(2)基于关键词的检索方法;(3)基于序列相似性的检索方法。(4)数据两两比对。	家国情怀; 科技强国	4
第二章	(1)多重比对的方法;(2)系统分析的方法。	学无止境; 自主学习; 终身学习	6
第三章	(1)蛋白质一、二、三级结构预测;(2)蛋白质跨膜结构域、信号肽、亚细胞定位的预测。	科技创新; 勇攀高峰	6
第四章	(1)MEGA 软件;(2)Python 语言和 R 语言基础;(3)引物设计等在线分析工具;(4)Origin 作图。	自主学习, 质疑权威, 实践是检验真理的唯一标准	16
合计		32	

7. 课程实施

7.1 第一章 绪论

7.1.1 什么是生物信息学?

生物信息学是一门交叉学科。它包含了生物信息的获取、管理、分析、解释和应用在内的所有方面。它综合运用生物学、计算机科学和数学等多方面知识与方法,来阐明和理解大量生物数据所包含的生物学意义,并应用于解决生命科学研究和生物技术相关产业中的各种问题。生物信息学主要有三个组成部分:建立可以存放和管理大量生物信息学数据的数据库;研究开发可用于有效分析与挖掘生物学数据的方法、算法和软件工具;使用这些工具去分析和解释不同类型的生物学数据,包括 DNA、RNA 和蛋白质序列、蛋白质结构、基因表达以及生化途径等。生物信息学这个术语从20世纪90年代开始使用,最初主要指的是 DNA、RNA 及蛋白质序列的数据管理和分析。自从20世纪60年代就有了序列分析的计算机工具,但是那时并未引起人们很大的关注,直到测序技术的发展使 GenBank 之类的数据库中存放的序列数量出现了迅猛的增长。现在该术语已扩展到几乎覆盖各种类型的生物学数据,如蛋白质结构、基因表达和蛋白质互作等。

7.1.2 生物信息学的发展历史

生物信息学早期的研究对象主要限于 DNA 序列的存储和分析，而其最近的迅速发展主要缘于基因组计划及相关转录组、蛋白质组、代谢组、相互作用组等计划的实施和高通量生物实验技术的发展，使生物学实验数据出现了爆炸性增长。生物信息学作为一门独立的学科只有近20年的历史，但事实上，与生物信息学相关的研究可以追溯到远至上世纪中期对蛋白质和 DNA 结构预测的模型研究。

7.1.3 生物信息学的主要研究领域、基本问题和方法

目前的生物信息学研究，已从早期以数据库的建立和 DNA 序列分析为主的阶段，转移到后基因组学时代以比较基因组学（comparativegenomics）、功能基因组学（functionalgenomics）和整合基因组学（integrativegenomics）为中心的新阶段。生物信息学的研究领域也迅速扩大。生物信息学涉及生物学、计算机学、数学、统计学等多门学科，从事生物信息学研究的工作者或生物信息学家可以来自以上任何一个领域而侧重于生物信息学的不同方面。事实上，我们今天正需要具备各种背景知识、才能和研究思路的研究人员，集思广益来共同面对生物信息学给我们的这史无前例的挑战。以下简要归纳当前生物信息学研究中的基本问题。

(1) 生物学数据库的建立和搜寻

生物学数据库贮存生物信息学研究的原始数据，是生物信息学存在和发展的基础。从 Dayhoff 及其同事20世纪60年代建立第一个已知蛋白质序列的分子生物学数据库到今天经历了突飞猛进的发展。80年代 GenBank、EMBL（EuropeanMolecularBiologyLaboratory）和 DDBJ（DNADatabankofJapan）以 DNA 序列为主的世界三大标准数据库的建立为分子生物学数据库的发展奠定了基础，并发挥了核心作用。计算机网络的发展与迅速普及和使用极大地促进了数据库的发展，并保证其数据为广大的用户方便地获取，而计算机储存技术的发展和储存量的快速增长满足了生物数据指数增长的需求。其中同样关键的是关系数据库技术（relationaldatabase）的发展促进了对数据库的使用。多年前在所有的分子生物学学术相关杂志中确立统一标准，要求所有新发表的分子序列在正式发表之前必须储存到 GenBank、EMBL 或 DDBJ 中的任何一个数据库并获得一个统一的序列登记号码（accessionnumber）。这对分子生物学序列数据的标准化和保证数据库所含数据的公开起到关键的作用。目前这三大数据库实行每天进行数据互相交换，使得3个数据库所包含的核心数据相同，极大地方便了用户对数据库的使用。初期的数据库以单纯 DNA 和蛋白质序列为主，每一个数据条目仅包含文件名和序列。但这些数据库大多都已扩展到包含与序列相关的多种信息，包括功能、突变、编码产物、调节因子和参考文献等。除经典的 DNA 和蛋白质序列数据库外，还有生物大分子三维结构数据库（如 PDB）、文献数据库（如 PubMed）、与生物学相关的知识数据库（如 KEGG 和 GeneOntology）及基因组数据库等多种类型。其中以包含多种数据类型的综合型数据库为今后的发展重点。像 UCSC 的基因组浏览器就是这一类型的很好的例子。它集序列、多种基因注释、比较基因组、功能基因组和许多其他数据类型于一体。这类数据库通常具备方便的图形界面，便于不具备生物信息学技能的一般用户使用。但建立这类数据库要求对多种数据类型进行有效的整合，其中不仅需要考虑如何建立数据之间的联系，也对相关的软件技术开发提出新的挑战。数据格式的建立、数据的准确性和质量控制、方便的数据搜寻方式以及数据的及时更新是数据库建立和维持中的重要问题。目

前最为成功和使用最广的序列数据库提取系统当首推 NCBI 的 ENTREZ 系统 (<http://www.ncbi.nlm.nih.gov/Entrez/>)。另外,为达到数据库搜寻的最高效率,数据库中数据的重复必须达到最低水平。但要真正做到没有数据重复有较大的难度。NCBI 中的蛋白质序列数据库 nr (non-redundant) 就是这样的数据库。但事实上这个数据库已经包含相当多的重复序列。另外,还有为数不少的非完整序列的继续存在。而与其相应的 DNA 序列数据库 nt (在 NCBI 的 BLAST 网页上为 nr) 已早就公布不再是冗余的了。用户在使用这类数据库时有必要了解这些情况,才能对搜寻结果做出正确判断。尽管数据库的建立不再是当前生物信息学的焦点,但对现有数据库的扩展、维护仍然非常必要。与此同时,开发新型数据库、研究有效数据格式和类型、促进数据交换和提取也非常关键。

(2) 序列比较与相似序列搜索

在现有的序列数据库找出与用户序列相同或相近序列,可以提供与此序列特征和功能相关的重要信息,涉及的问题是 DNA 和蛋白质序列相似性的分析。其中蛋白质的相似性分析因涉及不同的氨基酸的结构和功能的影响不同而远比 DNA 的序列分析来的复杂。序列相似性分析是生物信息学最早所涉及的问题,也是现今生物信息学研究中的日常工作之一。由于数据库的数据量日益增长,其中主要的计算问题是找到一种快速而灵敏的计算机运算法则。可以说以 Needleman-Wunsch 和 Smith-Waterman 为代表的运算法则较为满意地解决了这一问题,但对于新的运算法则的发展仍然十分迫切,以满足数据库快速增长和大规模序列分析所提出的新的要求。NCBI 的 BLAST 是现今序列相似性分析中使用最广的常用工具。除此以外,还有最近有 David Haussler 和 Jim Kent 领导的 UCSC 研究组所创立的 BLAT 方法,以极其快速的优点而在其基因组浏览器中得到成功的应用。

(3) 基因组结构注释

预测一段 DNA 序列或一个物种的基因组序列中具体哪些区域代表用于编码蛋白质的功能基因是生物信息学研究中的另一个经典问题。相对而言,原核生物基因的预测较为简单,因为原核基因没有内含子,因而只需寻找达到一定长度而具有起始密码子的开放阅读框 (openreadingframe, ORF)。其中较为复杂的情形包括:同一转录子编码多个蛋白;不同的基因间相同方向或相反方向互相重叠等。尽管如此,可以说,对于原核基因组中的基因预测,已经获得较为满意的成就。然而,真核生物基因因为有内含子和外显子之分,加之选择性转录本的存在,其基因结构的预测成为生物信息学研究中的一大挑战。其中的难点之一为外显子和内含子交界位点的确定,而第一个包含起始密码子的外显子的预测难度为最大。现有的预测方法通常借助于已知蛋白质序列的比对 (DNA-蛋白质序列),与已知 cDNA 及表达序列标签 (EST) 的比对 (DNA-DNA),及相近物种基因组序列间的比对 (在翻译上的 DNA-DNA 比对)。除此以外,预测方法利用已知基因结构序列进行训练,采用包括神经网络 (neuralnetwork) 和 HMM (HiddenMarkovModel) 在内的机器学习方法来识别外显子,尤其是外显子与内含子交界区域序列模式特征,已经取得了长足的进展,产生了包括 FGENES、GeneFinder、GeneMark、GeneParse、GenScan 及 FirstEF 等在内的一系列预测真核生物基因的软件工具。然而,所有这些成为 *abinitio* (意为从头开始,预测不直接依赖实验和其他数据) 的预测方法,其准确率还远未达到令人满意的程度。真核生物基因预测的另一难点是有许多基因利用多种可能的外显子进行不同的组合获得不同的基因表达产物。这是真核生物利用有限的基因组产生的复杂的基

因功能，达到适应不同发育、生理及环境条件的一种有效途径。目前对于什么因素控制或操纵不同外显子的选择仍然知之甚少。很显然，这一研究领域可提高和完善的空间还很大。这不仅有待于发展更智能化的运算法则，也依赖对于包括RNA剪接在内的许多真核生物基本现象的进一步认识。除蛋白编码基因的预测外，高等生物基因组中占绝大多数的非编码序列的分析所带来的挑战更大，近年来已逐渐成为生物信息学的研究热点之一。

(4) 蛋白质结构和功能的预测

到目前为止，尽管已知的蛋白质序列已过百万计，但结构已知的仍为少数。目前储存在蛋白质结构数据库 PDB 中的条目还只有2万多个。因此结构已知的蛋白质数目仍只占已知序列蛋白质总数的2%不到。其原因在于测定蛋白质的结构需要使用非常费时和昂贵的实验方法，如 X-晶体衍射或核磁共振（nuclear magnetic resonance, NMR）。很显然，仅仅依靠实验的方法很难测定所有已知蛋白质的结构。因而，根据蛋白质的一级结构来预测其高级结构，包括预测蛋白质间的相互作用，蛋白质与受体和药物的作用等，具有很大的应用价值。蛋白质在被合成时为线状氨基酸肽链，然后链的不同区域形成包括螺旋（ α -helix）、片层（ β -sheet）和转折（turn/loop）在内的二级结构。二级结构再互相重叠形成高级结构。蛋白质折叠（folding）的结果一般是把疏水的氨基酸放置在折叠后的内部区域，而把与水和其他分子相作用的极性氨基酸留在表面。一级结构引导折叠的过程，其中有时需要伴护蛋白（chaperone）的帮助。我们目前已经积累了相当数量的已知蛋白质结构，因而可以采用计算的方法来找到或预测哪些序列与已知结构具有相似的结构团（structural fold，指的是通过相似的折叠连接的相同二级结构方式）。对已知蛋白结构的统计分析表明，蛋白质结构团只有有限的约500种不同类型，而现有的已知结构已经包括自然界所存在的蛋白质结构的90%以上。其中的原因可能是蛋白质折叠过程中化学的限制或是蛋白质结构存在单一的进化途径。由于序列不同的蛋白质可以具有相同的结构，这使得蛋白质结构的预测难度增加。新型的蛋白结构预测方法采用复杂的统计和机器学习方法，其准确率在逐步地提高。现有的预测方法对于较小的分子（<300氨基酸）的结构预测已经有相当高的成功率，但对于包含多个结构域的大分子的预测则仍然相当困难。这一领域新近的进展包括根据一级结构把蛋白质分成不同的家族，然后用统计的方法来找到每一个家族特定的共同模式（consensus pattern）。

(5) 基因组数据的分析

由于过去10年内大规模工业化自动基因预测技术的发展和不断成熟，加之全基因组鸟枪法（whole genome shotgun）在基因测序上的成功应用，使得基因组的测序速度得到迅速提高。以目前最先进的设备和技术，一个细菌基因组测序的全部过程可以在一周内完成，而一个平均大小的高等真核生物基因组也可在1~2年内完成。相信在不远的将来，随着新技术的诞生，其中包括纳米技术的应用，基因组测序的速度将进一步加快，成本将成倍下降。很有可能会实现最近有人提出的千元人类基因组测序的设想。大量基因组序列的测定，对生物信息学提出新的挑战。其中包括如何提高基因组组装的效率和准确性，如何有效地储存、显示基因组的数据和相关信息，以及如何发展新的软件工具、新的运算法则来比较大量的基因组数据。

(6) 比较基因组和系统发生遗传学分析

大量基因组序列的完成，给我们提供了空前多的 DNA 数据。对于不同物种基因组的比较分析成为比较基因组学（comparative genomics）。它是基因组学领

域里最强有力的、也最具挑战性的研究方法。其研究的着眼点包括：序列的保守与差异、基因组的结构、基因与基因间物理位置的保守、非编码 DNA 的数量和种类差异等。通过将人类基因组与小鼠及其他模式动物基因组的比较，极大地增进了我们对人类基因组中功能基因的了解。通过比较致病菌和同一物种非致病菌的基因组，可以了解其致病的机制，便于发展控制致病细菌传染的新方法、新疫苗和新药物。对于同一物种中不同个体基因组的比较，可以全面地观测群体中的个体差异（polymorphism）或多样性。利用基因组数据来研究物种间的进化关系和系统发生，可以克服使用单个基因所存在的片面性，能够从整个基因组的水平上来更为全面地理解物种的进化关系及基因组演变的规律。因而，比较基因组学将是生物信息学中今后的发展重点。与单个或多个基因的序列比较不同，基因组间因涉及的信息量常常是单基因信息量的上千倍、甚至上万倍，而对所用工具和计算机的内存和速度都提出更高的要求。

(7) 功能基因组和蛋白组学数据的分析

基因组的完成给基因表达的研究带来了最为深远的革命。后基因组时代的到来使得生物学的研究重点从以前的以 DNA 测序为主转移到以系统了解基因组内所有基因的生物学功能即功能基因组（functional genomics）为中心。以 DNA 芯片技术为代表的新技术能把一个基因组中所有的基因安放在一张小小的玻璃片上，因而，使得我们可以同时研究所有基因的表达，从而从整个基因组的水平上来研究基因的表达。从 DNA 芯片的设计，信号的定量分析，到根据基因的表达谱进行分类和组合，找到与表型相关的基因，都包含统计生物信息学方法的应用。更为重要的是根据基因芯片的分析结果，将其已知的基因和所有相关及相似的多种数据进行整合，对涉及的基因、基因调节、信号和代谢途径进行预测和模型研究。蛋白质组学（proteomics）以一个细胞或一个物种所具有的全部蛋白质为研究对象，尽管目前由于技术的限制，无法进行蛋白质的大规模测序分析，其研究的广度和深度仍然有限。但蛋白组代表基因组表达的最后结果和效果，与基因和细胞功能更加直接相关。因此，将会是今后发展的重点之一。其中主要的生物信息学问题是如何有效地根据蛋白质的分子质量和等电点等物理化学特征，以及蛋白质质谱分析、多肽指纹分析（peptide mass fingerprinting）等技术获得的结果，结合已知的基因组数据来预测每一蛋白分离样品的实际身份。

(8) 信号传导、代谢和基因调节途径的构建与描述

细胞内的基因和蛋白质都不是独立行使功能。蛋白间相互作用，基因间互相调节，形成一个由信号传导、蛋白网络、代谢途径及细胞间的互相作用等组成的极其复杂和微妙精巧的网络来完成一个细胞的生物功能。过去，我们对生物现象的认识和了解大多仍局限于单个基因（蛋白）或单个信号传导或代谢途径范畴上一种静态的理解。基因组时代和后基因组时代的到来，不仅让我们清楚地认识到从系统和整体水平上来理解细胞功能的重要性，同时也第一次使得在这一水平上的研究成为可能。生物信息学今后的一大任务就是根据比较基因组学、功能基因组学、蛋白质组学研究的结果，结合我们从实验生物学积累的所有生物学数据来构建完整的代谢途径及与基因调节和各类信号传导相关的网络系统，包括了解它们之间的相互作用。更具挑战性的是如何有效地表示和利用这类研究的结果和对生物现象在整体和系统水平上进行模拟研究，提出新的理论和学说。

7.1.4 生物信息学今后的发展方向和趋势

人类基因组计划的完成标志着基因组时代进入高潮和后基因组(post-genome)时代的到来。到目前为止,上千种病毒基因组和近百种细菌基因组及数十种真核生物基因组的测序已完成,更多的物种被列入基因组测序的计划之中。与基因组生物信息学(genomicsinformatics)不同,后基因组生物信息学以从基因组信息获取的生物学知识来了解生命的基本原则,同时也有其在生物医学应用中的实际目的。后基因组生物信息学与功能基因组中以基因芯片和其他高通量(high-throughput)技术为基础的系统实验学相连,但生物信息学无疑将在实验设计和预计中起着更为主导的作用。大量生物物种基因组序列的完成和分析及生物信息学研究所带来的新型研究手段和成果正在迅速地改变生物医学的研究方法。其中最显著的一个革命性的改变就是反向遗传学(reversegenetics)研究策略的诞生和大量使用。许多情形下基因的功能可以通过生物信息学的研究方法预测,实验研究的目的是在许多情形下只是证实预测的功能。而在其他的情形,生物信息学为实验生物学的设计提供信息和思路,缩小研究的对象。因而,我们对基因功能的研究进程将大大加快。与此同时,生物信息学研究将根据比较基因组学、功能基因组学等分支学科的研究成果,运用大规模高度复杂和智能的数学统计模型,从而对生物学研究产生深远的革命性影响。

7.2 第二章生物数据库

近年来大量生物学实验的数据积累,形成了当前数以百计的生物信息数据库。它们各自按一定的目标收集和整理生物学实验数据,并提供相关的数据查询、数据处理的服务。随着因特网的普及,这些数据库大多可以通过网络来访问,或者通过网络下载。

一般而言,这些生物信息数据库可以分为一级数据库和二级数据库。一级数据库的数据都直接来源于实验获得的原始数据,只经过简单的归类整理和注释;二级数据库是在一级数据库、实验数据和理论分析的基础上针对特定目标衍生而来,是对生物学知识和信息的进一步整理。国际上著名的一级核酸数据库有Genbank数据库、EMBL核酸库和DDBJ库等;蛋白质序列数据库有SWISS-PROT、PIR等;蛋白质结构库有PDB等。国际上二级生物学数据库非常多,它们因针对不同的研究内容和需要而各具特色。下面将顺序简要介绍一些著名和有特色的生物信息数据库。

7.2.1 基因和基因组数据库

(1) Genbank

Genbank库包含了所有已知的核酸序列和蛋白质序列,以及与它们相关的文献著作和生物学注释。它是由美国国立生物技术信息中心(NCBI)建立和维护的。它的数据直接来源于测序工作者提交的序列;由测序中心提交的大量EST序列和其它测序数据;以及与其它数据机构协作交换数据而来。Genbank每天都会与欧洲分子生物学实验室(EMBL)的数据库,和日本的DNA数据库(DDBJ)交换数据,使这三个数据库的数据同步。Genbank的数据可以从NCBI的FTP服务器上免费下载完整的库,或下载积累的新数据。NCBI还提供广泛的数据查询、序列相似性搜索以及其它分析服务,用户可以从NCBI的主页上找到这些服务。

Genbank库里的数据按来源于约160,000个物种,其中约17%是人类的基因组

序列(所有序列中的64%是 EST 序列)。每条 Genbank 数据记录包含了对序列的简要描述, 它的科学命名, 物种分类名称, 参考文献, 序列特征表, 以及序列本身。序列特征表里包含对序列生物学特征注释如: 编码区、转录单元、重复区域、突变位点或修饰位点等。所有数据记录被划分在若干个文件里, 如细菌类、病毒类、灵长类、啮齿类, 以及 EST 数据、基因组测序数据、大规模基因组序列数据等 18 类, 其中 EST 数据等又被各自分成若干个文件。

GenBankflatfile (GBFF) 是 GenBank 数据库的基本信息单位, 也是最广泛地用以表示生物序列的格式之一。DDBJflatfile 格式与 GBFF 格式是相同的, EMBL 格式则与之有所差异。所有这些格式实际上都是由更结构化的 ASN.1 生成的。但是主要由于历史的原因, 许多用户在工作中使用 GBFF。

GBFF 可以分成三个部分, 头部包含关于整个记录的信息(描述符)。第二部分包含了注释这一记录的特性, 第三部分是核苷酸序列自身。所有的核苷酸数据库记录(DDBJ/EMBL/GenBank)都在最后一行以//结尾。

头部是记录中与数据库关联最大的部分。各个的数据库并不一定在这一部分包含相同的信息, 而可能存在着微小的差别。但各数据库已作出努力以在彼此之间保证信息兼容。所有的 GenBankflatfile 开始于 LOCUS 行。这一行中的第一项是 LOCUS 名称。历史上曾用这个名称来表示本记录描述的基因座, 提交者和数据库工作人员花费了无数的时间来设计这一名称。这一成分开始于一个英文字母, 总长度不能超过 10 个字符。第二个字符以后可以是数字或字母, 所有字符均要大写。LOCUS 名称在以前是最为有用的, 那时大多数 DNA 序列记录只表示一个基因座, 这样在 GenBank 中寻找一个可以用少数几个字母和数字来代表生物体的独特的名字是很容易的事。为了可用起见, LOCUS 名称在数据库中必须是独一的。因为几乎所有有意义的命名符都被使用过了, 所以今天 LOCUS 名称已不再是一个有用的成分。LOCUS 行中的下一项表明生物分子的类型。“分子类型”通常是 DNA 或 RNA, 但也有少量其他类型出现, 以表明生物分子的最初来源。LOCUS 行中的日期是数据最后被公开的日期。在许多情况下, 也是第一次被公开的日期。记录中包含的另一个日期是序列提交给数据库的日期。

DEFINITION 行(也称为“DEF”行)在 GenBank 记录中用以总结记录的生物意义。这一行将出现在 NCBI 的 FASTA 文件中, 这样任何人进行 BLAST 相似性搜索时都会看到这些信息。但是, 用一行文字来说明生物背景并不总是可行的, 对此不同的数据库采用了各自的解决方法。其中有一些共识, 并且每个数据库也都了解其他数据库的解决方法, 并尽力与之一致。

检索号在记录的第三行, 是从数据库中检索一个记录的主要关键词。这个号码将在参考文献中被引用, 并始终和序列在一起。就是说, 当序列被更新(例如更正一个核苷酸)时, 这个号码不会改变。检索号码采取下列两种方式之一: 1+5 或 2+6 格式。1+5 格式是指 1 个大写字母后跟 5 位数字; 2+6 格式是指 2 个大写字母后跟 6 位数字。绝大多数新近加入数据库的记录采取后一种方式。所有的 GenBank 记录都只有一个单独的 ACCESSION 行, 行中可能有多个检索号码, 但绝大多数情况只有一个检索号。这通常称为主检索号码, 其余的是二级检索号码。例如: AF010325.1, 这表明序列第 1 版, 检索号为 AF010325, gi 号为 2245686。

KEYWORDS 是另一个有趣的历史遗留物, 并且不幸地在很多情况下被误用了。给一个记录加上关键词通常并不十分有效, 因为在过去的年月中有许多作者选用了不在受控词表中的词, 并且在整个数据库中用法也不一致。因此, NCBI 不鼓励使用关键词, 但在查询时加入关键词是可以的, 特别是那些没有在其它记

录中出现的过词,或以一种受控的方式来使用的词(例如:对于 EST, STS, GSS, HTG 记录)。

SOURCE 行中有生物的通用名或科学名称。有些情况下也有其它来源的信息。现在正在一致努力以保证来源特性中包含所有必须的信息(不同于现在的 SOURCE 行),并且所有关于分类的信息(SOURCE 行和 ORGANISMS 行)可以从来源特性以及 NCBI 分类服务器中获得。对于系统族或关于分类的其它方面感兴趣的读者可以访问 NCBI 的分类主页。这一分类被所有核苷酸序列数据库以及蛋白质数据库 Swiss-Prot 所采用。

每个 GenBank 记录至少要有一篇参考文献。许多情况下有多篇。未发表的论文标记为“未发表”或是“已投”),如果将来文章发表的话则将代替于此。参考文献提供了科学证据以及一个背景来解释这个特定的序列为何会这样确定。当参考文献发表时,通常会有一个 MEDLINE 标识符,正如下面例子中一样,提供了指向 MEDLINE/PubMed 数据库的链接。在1998年末,又加入了一个新的行,以及其标识符 PUBMED,允许指向 PubMed 数据库以及发表者在线全文电子版的链接。

GBFF 记录的中间部分,也是最重要的一部分,就是注释,它直接表达了记录的生物背景知识。也许有人争辩说生物背景在记录所引用的参考文献中有最好的表现,但不论怎样,记录中的一整套注释有助于快速地抽取相关生物信息,并允许提交者指出这一记录当时为什么会被提交到这个数据库中。这里对于注释的选择就十分关键了。特性表文档详细描述了合法的特性(允许使用的注释),以及这些特性的允许限制词。不幸的是,这里经常有一些非法的,推测性的或由计算得出的注释。如果一个注释是仅由计算得到的,它作为记录说明的可用性就大打折扣了。

来源(source)是唯一一个必须在所有 GenBank 记录中出现的特性。所有的特性都有一系列合法的限定词,有些是强制性的(例如来源中的/organism(生物体))。所有的 DNA 序列记录都有出处,即使是合成序列这样极端的特例也一样。大多数情况下一个记录只能有一个来源特性,并带有/organism 限定词。限定词 organism 包含属和种的科学名称,有些情况下还可以在亚种水平描述。

CDS 指示读者如何将两个序列连接在一起,或如何根据核苷酸序列以及基因编码得到氨基酸序列。GBFF 以 DNA 为核心,通过 DNA 序列坐标系统映射所有特性,而不是从氨基酸的角度。在分析这些数据时,我们必须从 DNA 坐标推导出氨基酸位置,并且我们对于所编码蛋白质的了解也将仅限于从对 DNA 特性的描述中获得。这一限制可被 Sequin 克服。这一例子也显示了数据库交叉索引(db_xref)的使用。这一受控限制词允许数据库将另一个外部数据库的序列(第一个标识符)与一个在本数据库中使用的标识符交叉索引。允许 db_xref 的数据库都是合作数据库所维护的。正如上面提到的,NCBI 给每个记录赋予一个 gi(geninfo)标识符。这意味着翻译产物蛋白质序列(不是简单附属于 DNA 记录,如同在 GenBank 记录中显示的),也有自己的 gi 号码。一个特定的标识符当且仅当序列更改时才更改。蛋白质 gi 号码现在作为 PIDdb_xref 或蛋白质标识符出现。

7.2.2 Genbank 二级数据库

(1) 表达序列标记数据库 dbEST

EST (ExpressedSequenceTags) 方法已被证明是识别转录序列的最有效方法。

在1990以前,关于人类基因序列的数据主要来自于对单个基因的研究,EST数据的出现是生物信息学发展历史上的一块里程碑。EST序列大约覆盖了人类基因的90%。EST序列中含有大量的基因信息,利用这些信息可以发现新的基因,阐明基因的功能。dbEST是GenBank的一个部分,该数据库包括不同生物的EST序列数据及其它相关信息,主要是从大量不同组织和器官得到的短mRNA片段。通过WEB页面可以查询有关EST的数据和相关报道,也可以通过FTP下载dbEST数据库。EST数据库的主要作用是通过搜索比较,给实验新得到的一条cDNA序列或基因组序列赋予公认的功能。通过对EST数据库的逆向分析,能识别与疾病相联系的基因。

(2) 基因聚类数据库 UniGene

UniGene数据库将GenBank中的序列进行自动分类,形成面向基因群的非冗余集合。每个UniGene群包含代表一个唯一基因的多个序列,附有该基因相关的信息,如基因表达的组织类型、定位图谱。除了基因的序列之外,还包括大量的EST序列。UniGene既可以作为发现新基因的数据源,也可以作为生物学家进行大规模表达分析的辅助工具。需要指出的是,自动分类的过程还有待于进一步发展和完善。目前,UniGene中包括人类、小鼠、水稻、小麦等生物的相关数据,因为这些生物有大量的EST数据。

(3) 序列标记位点数据库 dbSTS, UniSTS

STS(SequenceTaggedSite)是序列标记位点。dbSTS是NCBI的一个数据源,也是GenBank的一个部分,包含已知的序列标记位点组成和定位信息。可以通过BLAST搜索STS序列,或者直接通过FTP下载序列。

(4) 基因组数据库

随着核酸测序技术的迅速发展,人类已经得到一部分生物的全基因组数据,如人、小鼠、大鼠等。这些数据对于我们认识基因组信息组织的奥秘、了解生物体的生长发育的规律是非常重要的。国际上有专门的组织收集和管理这些数据。NCBI基因组数据库EntrezGenomes所收集的基因组数据量非常大。该数据库还提供了—个基因组数据浏览工具MapView,利用这个工具,用户可以很方便地得到所需要的数据。

(5) 单碱基多态性数据库 dbSNP

遗传学研究的一个重要方面是建立生物分子序列变化与可遗传表型之间的联系,其中最常见的序列变化就是单核苷酸多态性SNPs(Singlenucleotidepolymorphisms)。在人类基因组中,大约在500到1000碱基长度范围内,就会出现一次单碱基的变化。SNPs对人类遗传学研究和医学应用具有重要的意义,无论对于人类种群遗传学的研究,还是疾病易感性分析、药物基因组研究或个体化医疗,都需要深入地研究SNPs。找出人类基因组中所有的SNPs是基因组研究的一个组成部分。某些特定的SNPs等位基因被认为是人类遗传疾病的致病因子,在个体中筛选这类等位基因可以检查其对疾病的遗传易感性。SNPs也可以作为遗传作图的遗传标记,帮助定位和鉴定功能基因。目前,科学家在SNPs筛选和发现方面正在做大量的工作,由于大规模基因组序列分析及其相关技术(特别是基因芯片技术)的不断提高,同时,也由于生物信息学及

计算机技术的发展,使得检测和分析 SNPs 成为可能。
单核苷酸多态性数据库 dbSNP 是由 NCBI 与人类基因组研究所合作建立的,它是关于单碱基替换以及短插入、删除多态性的资源库。

7.2.3 EMBL 核酸序列数据库

EMBL 核酸序列数据库由欧洲生物信息学研究所(EBI)维护的核酸序列数据库构成,由于与 Genbank 和 DDBJ 的数据合作交换,它也是一个全面的核酸序列数据库。该数据库由 Oracal 数据库系统管理维护,查询检索可以通过因特网上的序列提取系统(SRS)服务完成。向 EMBL 核酸序列数据库提交序列可以通过基于 Web 的 WEBIN 工具。

7.2.4. DDBJ 数据库

日本 DNA 数据仓库(DDBJ)也是一个全面的核酸序列数据库,与 Genbank 和 EMBL 核酸库合作交换数据。可以使用其主页上提供的 SRS 工具进行数据检索和序列分析。可以用 Sequin 软件向该数据库提交序列。

7.2.5 蛋白质数据库

(1) PIR

这是一个全面的、经过注释的、非冗余的蛋白质序列数据库。所有序列数据都经过整理,超过99%的序列已按蛋白质家族分类,一半以上还按蛋白质超家族进行了分类。注释中还包括对许多序列、结构、基因组和文献数据库的交叉索引,以及数据库内部条目之间的索引,这些内部索引帮助用户在包括复合物、酶-底物相互作用、活化和调控级联和具有共同特征的条目之间方便的检索。

(2) SWISS-PROT

SWISS-PROT 是经过注释的蛋白质序列数据库,由欧洲生物信息学研究所(EBI)维护。数据库由蛋白质序列条目构成,每个条目包含蛋白质序列、引用文献信息、分类学信息、注释等,注释中包括蛋白质的功能、转录后修饰、特殊位点和区域、二级结构、四级结构、与其它序列的相似性、序列残缺与疾病的关系、序列变异体和冲突等信息。SWISS-PROT 中尽可能减少了冗余序列,并与其它30多个数据建立了交叉引用,其中包括核酸序列库、蛋白质序列库和蛋白质结构库等。利用序列提取系统(SRS)可以方便地检索 SWISS-PROT 和其它 EBI 的数据库。SWISS-PROT 只接受直接测序获得的蛋白质序列,序列提交可以在其 Web 页面上完成。

(3) PROSITE

PROSITE 数据库收集了生物学有显著意义的蛋白质位点和序列模式,并能根据这些位点和模式快速和可靠地鉴别一个未知功能的蛋白质序列应该属于哪一个蛋白质家族。有的情况下,某个蛋白质与已知功能蛋白质的整体序列相似性很低,但由于功能的需要保留了与功能密切相关的序列模式,这样就可能通过 PROSITE 的搜索找到隐含的功能 motif,因此是序列分析的有效工具。PROSITE 中涉及的序列模式包括酶的催化位点、配体结合位点、与金属离子结合的残基、二硫键的半胱氨酸、与小分子或其它蛋白质结合的区域等;除了序列模式之外,PROSITE 还包括由多序列比对构建的 profile,能更敏感地发现序列与 profile 的

相似性。PROSITE 的主页上提供各种相关检索服务。

7.2.6 结构数据库

(1) PDB

蛋白质数据仓库(PDB)是国际上唯一的生物大分子结构数据档案库, 由美国 Brookhaven 国家实验室建立。PDB 收集的数据来源于 X 光晶体衍射和核磁共振(NMR)的数据, 经过整理和确认后存档而成。目前 PDB 数据库的维护由结构生物信息学研究合作组织(RCSB)负责。RCSB 的主服务器和世界各地的镜像服务器提供数据库的检索和下载服务, 以及关于 PDB 数据文件格式和其它文档的说明, PDB 数据还可以从发行的光盘获得。使用 Rasmol 等软件可以在计算机上按 PDB 文件显示生物大分子的三维结构。

7.2.7 功能数据库

(1) KEGG

京都基因和基因组百科全书(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。基因组信息存储在 GENES 数据库里, 包括完整和部分测序的基因组序列; 更高级的功能信息存储在 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, 还包括同系保守的子通路等信息; KEGG 的另一个数据库是 LIGAND, 包含关于化学物质、酶分子、酶反应等信息。KEGG 提供了 Java 的图形工具来访问基因组图谱, 比较基因组图谱和操作表达图谱, 以及其它序列比较、图形比较和通路计算的工具, 可以免费获取。

7.2.8 文献数据库

(1) PubMed

PubMed 是 NCBI 维护的文献引用数据库, 提供对 MEDLINE 等文献数据库的引用查询和对大量网络科学类电子期刊的链接。利用 Entrez 系统可以对 PubMed 进行方便的查询检索。

(2) 人类遗传数据库 OMIM

OMIM(OnlineMendelianInheritanceinMan)是关于人类基因和遗传疾病的分类数据库, 由约翰霍普金斯大学开发。该数据库收集了已知的人类基因及由于这些基因突变或者缺失而导致的遗传疾病。OMIM 主要的服务对象是医师、遗传疾病研究人员、生物医学专业高年级学生。在 OMIM 中, 可以按照基因搜索数据库, 也可以按照遗传疾病搜索数据库。OMIM 的网络服务器位于 NCBI, 每条记录引用的参考资料都有到 Entrez 系统的链接。OMIM 的使用非常方便。查询程序根据输入到检索窗口的一个或几个词执行简单的查询, 返回含有该词的文档的列表, 用户可以在列表选择一个或更多的记录查看其 OMIM 数据的全文。记录含有各种信息, 如基因符号、病变的名称、对病变的描述(包括临床的, 生物化学的, 细胞遗传学的特征)、遗传模式上的细节(包括图谱信息)、临床的说明等, 还有参考文献。用户也可以选择特定的染色体, 浏览染色体上相关的基因及病变信息。

7.2.9 向 Genbank 提交序列数据

测序工作者可以把自己工作中获得的新序列提交给 NCBI, 添加到 Genbank

数据库。这个任务可以由基于 Web 界面的 BankIt 或独立程序 Sequin 来完成。BankIt 是一系列表单,包括联络信息、发布要求、引用参考信息、序列来源信息、以及序列本身的信息等。用户提交序列后,会从电子邮件收到自动生成的数据条目, Genbank 的新序列编号,以及完成注释后的完整的数据记录。用户还可以在 BankIt 页面下修改已经发布序列的信息。BankIt 适合于独立测序工作者提交少量序列,而不适合大量序列的提交,也不适合提交很长的序列, EST 序列和 GSS 序列也不应用 BankIt 提交。BankIt 使用说明和对序列的要求可详见其主页面。大量的序列提交可以由 Sequin 程序完成。Sequin 程序能方便的编辑和处理复杂注释,并包含一系列内建的检查函数来提高序列的质量保证。它还被设计用于提交来自系统进化、种群和突变研究的序列,可以加入比对的数据。Sequin 除了用于编辑和修改序列数据记录,还可以用于序列的分析,任何以 FASTA 或 ASN.1 格式序列为输入数据的序列分析程序都可以整合到 Sequin 程序下。在不同操作系统下运行的 Sequin 程序都可以在 <ftp://ncbi.nlm.nih.gov/sequin/> 下找到, Sequin 的使用说明可详见其网页。

7.3 第三章关键词和词组为基础的数据库检索

随着大量生物学实验数据的积累,众多的生物学数据库也相继出现,它们各自按照一定的标准收集和处理生物学实验数据,并提供相关的数据查询、处理等服务。如何从浩瀚的数据库中获取有用信息,怎样处理提取的数据,进而从中获得与生物结构、功能相关的信息成为科学工作者面临的一个急待解决的问题。用户想要有效、迅速的获取生物信息,首先必须对因特网上的生物信息资源相当了解。我们在上一章中已经详细讲述了重要的生物学一级和二级数据库。在正确选择了可能包含要查询信息的数据库后,我们就要选用合适的检索工具对其进行检索。在生物信息数据库发展的同时,各数据库开发和维护单位也在同时进行高效率的数据库检索系统的研发。检索体系可分为两大类:以关键词或词组为基础进行检索和以核苷酸或蛋白质序列为基础进行检索。前者的代表是 NCBI 开发的 Entrez 系统和 EBI 开发的 SRS 系统,而后的代表则是 NCBI 的 BLAST 和 EBI 的 FASTA。

7.3.1 Entrez 检索系统

Entrez 是 NCBI 提供的以关键词和词组为基础的数据库检索系统。与 Entrez 体系相连的数据库有8大类29个,其包括文献数据库如 Pubmed、OMIM、Books 等;核苷酸序列数据库如 Genbank、Gene、SNP、UniSTS 等;蛋白质序列数据库如 Proteins 等;结构数据库如 Struture、3DDomains 等;生物分类数据库如 Taxonomy 等;基因组数据库如 Genome、GenomeProject 等;表达数据库如 UniGene、GEOprofiles 等;以及其他数据库如 PubChemSubstance、CancerChromosomes 等。所有的数据库既可独立检索,也可同时检索,数据库之间建有超级链接,可直接进行交互访问使用。

Entrez 提供了方便实用的检索服务,所有操作都可以在网络浏览器上完成。用户可以利用 Entrez 界面上提供的限制条件(Limits)、索引(Index)、检索历史(History)和剪贴板(Clipboard)等功能来实现复杂的检索查询工作。对于检索获得的记录,用户可以选择需要显示的数据,保存查询结果,甚至以图形方式观看检索获得的序列。详细的 Entrez 使用说明可以在该主页上获得。

(1) Entrez 检索方法

用户登录 NCBI 网站后可在检索栏的下拉列表中选择相应的数据库, 在检索提问栏内输入检索词开始检索; 也可通过“Limits”设定限定项后再进行检索。

可在下列检索领域(SearchFields)中选择关键词或词组进行检索:

登陆号(Accession): 也可为 GI 号

物种(Organism): 包含与该蛋白或核酸序列相关物种的学名和俗名。

基因名(GeneName): 基因的标准名称。

特性(Genedescription): 一个或几个关键词, 用来描述该序列的类型。

片段长度 (Sequencelength)

特色 (Featurekey): 基因特性。

关键词(Keywords): 可以使用较特定的索引条目来检索以上数据库。

作者姓名(AuthorName): 文章作者名单, 通常名字为首个字母的缩写。

附属机构(Affiliation): 包括该检索领域建立时的相关信息, 原作者地址, 有时亦有其他作者地址

杂志名(JournalTitle): 为检索条目第一次发表时的杂志名, 该杂志名是以缩写形式储存于数据库中, 如果不清楚杂志是如何缩写的可采用 ListTerms 来查看。

杂志期号 (issue)

出版日期 (PublicationDate)

关键词或词组可为

主题词: 如专业术语

短语: 要加双引号, 如“16SRNA”, 表示单词的特定排列顺序

对关键词可使用通配符*, 如 wan*表示所有以 wan 开头的单词, 扩大了检索面但专一性降低。也可对多个关键词联合检索, 使用 AND, NOT 或 OR 进行连接。

(2) 特征栏介绍

在检索提问栏下方的特征栏上设置了条件限定(Limits)、预览/索引(Preview/Index)、检索史(History)、粘贴板(Clipboard)和细节(Datails)5种功能按钮。

(3) Limits

点击 Limits, 系统显示多种限定条件供用户选择。1、将检索词限定在某一特定字段(Limitedto); 2、将检索限定在某一特定年龄组(Ages)、性别(Gender)、人或动物(HumanorAnimal); 3、将检出的文章限定在某一指定的语言(Languages), 以及某一指定的出版物类型(Publicationtypes), 如综述; 4、可以用输入到数据库的日期(EntrezDate)或期刊出版年代(PublicationDate)限定; 5、可以将检索限定在 PubMed 中的某一子数据库(Subsets)。Entrez 的常用限定检索字段有 Affiliation、AllField(默认)、Author、EC/RNNumber(酶学编码字段)、EntrezDate(录入 Entrez 系统的日期)、Issue(期刊期号)、Journal(期刊名)、Language(语种)、PublicationTypes(出版物类型字段)、GeneName、ProteinName 等。限制功能(Limits)可缩小检索面, 缩短检索时间。

通过使用逻辑算符(AND、OR、NOT)和使用各种限制功能, 可大大提供检索的成功率, 缩短检索时间。

(4) Preview/Index

其作用是在显示条目之前预览检索策略和检索结果的记录数、修改检索策略以及从索引表中选词检索。

(5) History

点击特征栏上的 History 可显示检索式的检索序号、检索词、检索时间以及

检索结果数量, 以便于利用检索序号进行检索式之间的组配检索。

(6) Clipboard

该功能允许用户将检索结果进行临时保存,并对添加到 Clipboard 中的文献进行进一步的筛选、保存、打印以及二次检索等处理。要把检索结果全部存入粘贴板中,只需直接点击“AddtoClipboard”键。如果要把部分检索结果存入粘贴板,只要点击该记录左边的选择框后,再点击“AddtoClipboard”键。该功能使用户不必对每次检索结果进行筛选保存,而是可以将几次检索的结果同时进行批处理,简化操作过程。

7.3.2 检索结果展示方式

Entrez 系统自动执行检索,并将结果显示出来。显示屏上有检索提问框中的当前检索式、检出的记录总数、每页显示条数,共显示页数,以及命中记录初始简要格式。检索结果可采用多用方式进行展示。

对于 PubMed 文章,检索结果中通常包含题目、期刊年份期号和页码、作者姓名、文摘, MeSH 主题词,指向全文的链接等。

对于蛋白和核酸序列文件,检索结果可按照标准的 GenBank 或 GenPept 格式, ASN.1格式、FASTA 格式、图形格式(GraphicView)等方式进行展示。

对于结构文件可以查看其三维结构。对于基因组文件可按照图形格式或序列格式进行展示。

7.3.3 SRS 检索系统

是由欧洲生物信息研究所(EBI)开发的以 WWW 界面运行的数据库检索及导航系统,检索面宽但操作较为复杂。共有17大类194个数据库与 SRS 体系相连,可选择快速检索和高级检索两种模式。

7.3.4 DBGET 检索系统

与 KyotoEncyclopediaofGenesandGenomes(KEGG)database 相连,操作较 SRS 简单,但检索面较 SRS 和 Entrez 窄。DBGET 与36个数据库相连,可选择单库检索和多库检索两种模式。

7.4 第四章 核酸和蛋白质序列为基础的数据库检索

序列比对的理论基础是进化学说,如果两个序列之间具有足够的相似性,就推测二者可能有共同的进化祖先,经过序列内残基的替换、残基或序列片段的缺失、以及序列重组等遗传变异过程分别演化而来。序列相似和序列同源是不同的概念,序列之间的相似程度是可以量化的参数,而序列是否同源需要有进化事实的验证。在残基-残基比对中,可以明显看到序列中某些氨基酸残基比其它位置上的残基更保守,这些信息揭示了这些保守位点上的残基对蛋白质的结构和功能是至关重要的,例如它们可能是酶的活性位点残基,形成二硫键的半胱氨酸残基,与配体结合部位的残基,与金属离子结合的残基,形成特定结构 motif 的残基等等。但并不是所有保守的残基都一定是结构功能重要的,可能它们只是由于历史的原因被保留下来,而不是由于进化压力而保留下来。因此,如果两个序列有显著的保守性,要确定二者具有共同的进化历史,进而认为二者有近似的结构和功能还需要更多实验和信息的支持。通过大量实验和序列比对的分析,一般认为蛋

白质的结构和功能比序列具有更大的保守性，因此粗略的说，如果序列之间的相似性超过30%，它们就很可能是同源的。

早期的序列比对是全局的序列比较，但由于蛋白质具有模块性质，可能由于外显子的交换而产生新蛋白质，因此局部比对会更加合理。通常用打分矩阵描述序列两两比对，两条序列分别作为矩阵的两维，矩阵点是两维上对应两个残基的相似性分数，分数越高则说明两个残基越相似。因此，序列比对问题变成在矩阵里寻找最佳比对路径，目前最有效的方法是 Needleman-Wunsch 动态规划算法，在此基础上又改良产生了 Smith-Waterman 算法和 SIM 算法。在 FASTA 程序包中可以找到用动态规划算法进行序列比对的工具 LALIGN，它能给出多个不相互交叉的最佳比对结果。

在进行序列两两比对时，有两方面问题直接影响相似性分值：取代矩阵和空位罚分。粗糙的比对方法仅仅用相同/不同来描述两个残基的关系，显然这种方法无法描述残基取代对结构和功能的不同影响效果，缬氨酸对异亮氨酸的取代与谷氨酸对异亮氨酸的取代应该给予不同的打分。因此如果用一个取代矩阵来描述氨基酸残基两两取代的分值会大大提高比对的敏感性和生物学意义。虽然针对不同的研究目标和对象应该构建适宜的取代矩阵，但国际上常用的取代矩阵有 PAM 和 BLOSUM 等，它们来源于不同的构建方法和不同的参数选择，包括 PAM250、BLOSUM62、BLOSUM90、BLOSUM30等。对于不同的对象可以采用不同的取代矩阵以获得更多信息，例如对同源性较高的序列可以采用 BLOSUM90矩阵，而对同源性较低的序列可采用 BLOSUM30矩阵。

空位罚分是为了补偿插入和缺失对序列相似性的影响，由于没有什么合适的理论模型能很好地描述空位问题，因此空位罚分缺乏理论依据而更多的带有主观特色。一般的处理方法是给两个罚分值，一个对插入的第一个空位罚分，如10—15；另一个对空位的延伸罚分，如1—2。对于具体的比对问题，采用不同的罚分方法会取得不同的效果。

对于比对计算产生的分值，到底多大才能说明两个序列是同源的，对此有统计学方法加以说明，主要的思想是把具有相同长度的随机序列进行比对，把分值与最初的比对分值相比，看看比对结果是否具有显著性。相关的参数 E 代表随机比对分值不低于实际比对分值的概率。对于严格的比对，必须 E 值低于一定阈值才能说明比对的结果具有足够的统计学显著性，这样就排除了由于偶然的因素产生高比对得分的可能。

Genbank、SWISS-PROT 等序列数据库提供的序列搜索服务都是以序列两两比对为基础的。不同之处在于为了提高搜索的速度和效率，通常的序列搜索算法都进行了一定程度的优化，如最常见的 BLAST 工具和 FASTA 工具。

7.4.1 BLAST

大多数研究目前都通过国际互联网 Internet 应用 NCBI 研制的 BLAST 程序 (BasicLocalAlignmentSearchTool)来进行 DNA 和蛋白质序列相似性搜索。用一组 BLAST 程序联配可以快速进行核酸和蛋白质序列库的相似性检索。采用 BLAST 的基本算法编成了若干各不同的程序，分别使用特定的序列库和用于特定类型的输入序列。BLASTN 是在核苷酸序列库搜索核苷酸序列。BLASTP 是在蛋白质序列库中搜索氨基酸序列。TBLASTN 则可以在核酸序列库中搜索氨基酸序列，此时序列库在搜索之前要按所有6种读框即时翻译。与此相反的一项分析则由 BLASTX 来完成，它要将所输入的核酸序列按所有6种读框翻译，然后再以之搜

索蛋白质序列库。近期 Altschul S.F. 等人 (1997) 提出了一个通过寻找蛋白质家族保守序列来提高算法敏感性的 PSI-BLAST (Position-Specific Iterated BLAST) 算法, 并开发了相应的软件。PSI-BLAST 可以对数据库进行多轮循环检索, 每一轮的检索速度都大约是 BLAST 的两倍, 但每一轮都能提高检索的敏感性。它是目前 BLAST 程序家族中敏感性最高的成员。

如果目的序列中有蛋白质编码区, 则用翻译的蛋白质序列来搜索蛋白质序列库要比用 DNA 序列搜索核酸序列库更有价值。由于蛋白质序列的进化要比 DNA 序列慢一些, 在蛋白质序列水平上的远缘关系在 DNA 水平上可能被错过。如果无法确定编码区, 则可利用 BLASTX 按所有 6 种读框来翻译 DNA 序列, 然后用它搜索蛋白质序列库。由于蛋白质序列库仅包含已鉴定的蛋白质, 所以必须采用 TBLASTN 程序在现有的 GenBank、EMBL 或 DDBJ DNA 序列库中检索新确定的氨基酸或翻译过的 DNA 序列。这种检索有时可以找到一些显著相似的 DNA 序列, 而原本并不知道这些序列可编码蛋白质。

BLAST 的一项重要特性就是所报告的匹配序列的统计学显著性评分。这一统计学显著性评分是用 Karlin-Altschul 算法决定的, 所算出的 Poisson 概率表明所得到的序列相似性随机出现的可能性。

BLAST 是现在应用最广泛的序列相似性搜索工具, 相比 FASTA 有更多改进, 速度更快, 并建立在严格的统计学基础之上。NCBI 提供了基于 Web 的 BLAST 服务, 用户可以把序列填入网页上的表单里, 选择相应的参数后提交到数据服务器上进行搜索, 从电子邮件中获得序列搜索的结果。BLAST 包含五个程序和若干个相应的数据库, 分别针对不同的查询序列和要搜索的数据库类型。其中翻译的核酸库指搜索时会把核酸数据按密码子按所有可能的阅读框架转换成蛋白质序列。

BLAST 对序列格式的要求是常见的 FASTA 格式。FASTA 格式第一行是描述行, 第一个字符必须是“>”字符; 随后的行是序列本身, 一般每行序列不要超过 80 个字符, 回车符不会影响程序对序列连续性的看法。序列由标准的 IUB/IUPAC 氨基酸和核酸代码代表; 小写字符会全部转换成大写; 单个“-”号代表不明长度的空位; 在氨基酸序列里允许出现“U”和“*”号; 任何数字都应该被去掉或换成字母(如, 不明核酸用“N”, 不明氨基酸用“X”)。此外, 对于核酸序列, 除了 A、C、G、T、U 分别代表各种核酸之外, R 代表 G 或 A(嘌呤); Y 代表 T 或 C(嘧啶); K 代表 G 或 T(带酮基); M 代表 A 或 C(带氨基); S 代表 G 或 C(强); W 代表 A 或 T(弱); B 代表 G、T 或 C; D 代表 G、A 或 T; H 代表 A、C 或 T; V 代表 G、C 或 A; N 代表 A、G、C、T 中任意一种。对于氨基酸序列, 除了 20 种常见氨基酸的标准单字符标识之外, B 代表 Asp 或 Asn; U 代表硒代半胱氨酸; Z 代表 Glu 或 Gln; X 代表任意氨基酸; “*”代表翻译结束标志。

数据库相似性搜索程序 BLAST 和 FASTA 程序清单

程序	待检序列类型	数据库类型	说明
BLASTP	p	p	在蛋白质序列库中比对待检蛋白质序列

BLASTN	n	n	在核酸序列库中比对待核酸序列
BLASTX	n	p	在蛋白质序列库中比对待检核酸序列(用所有6种读框翻译)
TBLASTN	p	n	在核酸序列库(用6种读框即时翻译)中比对待检蛋白质序列
TBLASTX	n	n	在核酸序列库(用6种读框即时翻译)中比对待检核酸序列(同样用所有6种读框翻译)
FASTA3	p	p	在某一蛋白质序列库中搜索蛋白质相似序列
	n	n	在某一核酸序列库中搜索核酸相似序列
TFASTA3	p	n	在核酸序列库(已被即时翻译)中比对待检蛋白质序列
FASTX3	n	p	在蛋白质序列库中比对待检核酸序列(用6种读框翻译)
TFASTX3	p	n	在核酸序列库中比对待检蛋白质序列

注：n：核酸序列或核酸序列库；p：蛋白质序列或蛋白质序列库

(1) BLAST 选项

“WORDLENGTH”(字长)选项：BLAST 程序是通过比对未知序列与数据库序列中的短序列来发现最佳匹配序列的。最初进行“扫描”(scanning)就是确定匹配片段。序列的匹配程序由短序列(定义为“word”,即字)的联配得分总和来决定。联配时,“字”的每个碱基均被计分：如果碱基对完全相同(如 A 与 A),得某一正值；如果碱基对不很匹配(W 与 A 或 T),则得某一略小的正值；如果两个碱基不匹配,则得一负值。总的合计得分便决定了序列间的相似程度。得分高的匹配序列被称为高比值片段对(high-scoring segment pairs, HSP)。BLAST 程序在两个方向扩展 HSP,直至序列结束或联配已变为不显著。替换矩阵在扫描(scanning)和扩展过程被应用。最后在 BLAST 报告中被列出的序列都是所有得分最高的序列。以上述及的初始字长便是由 W(WORDLENGTH)值设定。BLAST 只对字长为 W 的“字”进行扩展联配。BLAST 的字长缺省值为11,即 BLASTN 将扫描数据库,直到发现那些与未知序列的11个连续碱基完全匹配的11个连续碱基长度片段为止。然后这些片段(即字)被扩展。11个碱基的字长已能有效地排除中等分叉的同源性和几乎所有随机产生的显著联配。

(2) “Filter”(过滤器)选项

BLAST2.0版本已有序列过滤器功能。过滤器将锁定诸如组成低复杂(low compositional complexity)序列区(如 Alu 序列),用一系列 N(NNNNNN)替代这些程序。N 代表任意碱基(IUB-code)。只有未知待检序列被过滤替代,而数据库的序列将不被过滤。过滤对绝大多数序列都是有益的,“Filter”项的缺省选项为 ON。例如,多 A 碱基的尾部和脯氨酸富积的序列,会得到人为的高联配得分而误导分析。这是因为这类序列数量极大,遍布整个基因组,直至整个数据库。

(3) “EXPECT”选项：

你可能会想为搜索设定一个期望值阈值(EXPECT),例如缺省值设为10。这

一设置则表示联配结果中将有10个匹配序列是由随机产生,如果联配的统计显著性值(E 值)小于该值(10),则该联配将被检出,换句话说,比较低的阈值将使搜索的匹配要求更严格,结果报告中随机产生的匹配序列减少。

(4) 输入框选项:

在序列的输入框内可以键入 EMBL 的身份号(ID)或 GenBank 的登录号(accessionnumber)。这样的输入选择将仅返回数据库中的某一序列资料(最新版本),该序列与键入的记录号相对应。在不少情况下需要类似检索,例如核对 PCR 产物。其它一些选项情况可参阅 BLAST 的在线使用手册。

7.4.2 FASTA

FASTA 是第一个被广泛应用的序列比对和搜索工具包,包含若干个独立的程序。FASTA 首先在序列库中进行快速的初检,找出与待检序列高度相似的序列。这一快速检索局限于待检序列和序列库序列之间较短的完全相同序列区段上。FASTA 的结果报告中会给出每个搜索到的序列与查询序列的最佳比对结果,以及这个比对的统计学显著性评估 E 值。FASTA 工具包可以在大多提供下载服务的生物信息学站点上找到。

7.4.3 BLITZ

BLAST 和 FASTA 检索体系有时不能检测出某些远缘序列的相关性,基于 Smith-Waterman 算法的 Blitz 检索体系在发现家族成员方面比上述两种检索体系更灵敏和更可靠。BLITZ 被设计在大型机上运行,速度慢,最好使用 email 服务。EBI 提供了两种分析方法: MPsrch4和 ScanPS。

7.5 第五章多序列对位排列分析和系谱分析

双序列比对是序列分析的基础。与序列两两比对不一样,序列多重比对(MultipleAlignment)的目标是发现多条序列的共性。如果说序列两两比对主要用于建立两条序列的同源关系和推测它们的结构、功能,那么,同时比对一组序列对于研究分子结构、功能及进化关系更为有用。例如,某些在生物学上有重要意义的相似性只能通过将多个序列对比排列起来才能识别。同样,只有在多序列比对之后,才能发现与结构域或功能相关的保守序列片段。对于一系列同源蛋白质,人们希望研究隐含在蛋白质序列中的系统发育的关系,以便更好地理解这些蛋白质的进化。在实际研究中,生物学家并不是仅仅分析单个蛋白质,而是更着重于研究蛋白质之间的关系,研究一个家族中的相关蛋白质,研究相关蛋白质序列中的保守区域,进而分析蛋白质的结构和功能。序列两两比对往往不能满足这样的需要,难以发现多个序列的共性,必须同时比对多条同源序列。目前对多序列比对的研究还在不断前进中,现有的大多数算法都基于渐进的比对的思想,在序列两两比对的基础上逐步优化多序列比对的结果。通过序列的多重比对,可以得到一个序列家族的序列特征。当给定一个新序列时,根据序列特征,可以判断这个序列是否属于该家族。对于多序列比对,现有的大多数算法都基于渐进比对的思想,在序列两两比对的基础上逐步优化多序列比对的结果。进行多序列比对后,可以对比对结果进行进一步处理,例如构建序列的特征模式,将序列聚类,构建分子进化树等。

7.5.1 多序列比对的意义

多序列比对有时用来区分一组序列之间的差异,但其主要用于描述一组序列

之间的相似性关系，以便对一个基因家族的特征有一个简明扼要的了解。与双序列比对一样，多序列比对的方法建立在某个数学或生物学模型之上。因此，正如我们不能对双序列比对的结果得出“正确或错误”的简单结论一样，多序列比对的结果也没有绝对正确和绝对错误之分，而只能认为所使用的模型在多大程度上反映了序列之间的相似性关系以及它们的生物学特征。显然，多序列比对需要使用许多专门的分析工具。除了一些已经广泛使用并仍在不断改进的多序列计算机程序外，还需要有一个开发方便实用的多序列比对手工编辑工具。可以从多个不同角度出发构建多序列比对模型。这里，主要指建立比对模型的生物学基础，而不仅是具体的比对方法，如自动比对或手动比对等。目前，构建多序列比对模型的方法大体可以分为两大类。第一类是基于氨基酸残基的相似性，如物化性质、残基之间的可突变性等。另一类方法则主要利用蛋白质分子的二级结构和三级结构信息，也就是说根据序列的高级结构特征确定比对结果。显然，这两种方法所得结果可能有很大差别。一般说来，很难断定哪种方法所得结果一定正确，应该说，它们从不同角度反映蛋白质序列中所包含的生物学信息。基于序列信息和基于结构信息的比对都是非常重要的比对模型，但它们都有不可避免的局限性，因为这两种方法都不能完全反映蛋白质分子所携带的全部信息。我们知道，蛋白质序列是经过 DNA 序列转录翻译得到的。从信息论的角度看，它应该与 DNA 分子所携带的信息更为“接近”。而蛋白质结构除了序列本身带来的信息外，还包括经过翻译后加工修饰所增加的结构信息，包括残基的修饰，分子间的相互作用等，最终形成稳定的天然蛋白质结构。因此，这也是对完全基于序列数据比对方法批评的主要原因。显然，如果能够利用结构数据，对于序列比对无疑有很大帮助。不幸的是，与大量的序列数据相比，实验测得的蛋白质三维结构数据实在少得可怜。在大多数情况下，并没有结构数据可以利用，我们只能依靠序列的相似性和一些生物化学特性建立一个比较满意的多序列比对模型。

7.5.2 多序列比对的定义

顾名思义，多序列比对就是把两条以上可能有系统进化关系的序列进行比对的方法。目前对多序列比对的研究还在不断前进中，现有的大多数算法都基于渐进的比对的思想，在序列两两比对的基础上逐步优化多序列比对的结果。进行多序列比对后可以对比对结果进行进一步处理，例如构建序列模式的 profile，将序列聚类构建分子进化树等等。

7.5.3 多序列比对的方法

目前使用最广泛的多序列比对程序是 Clustal，它是由 Feng 和 Doolittle 于 1987 年提出的。Clustal 的基本思想是基于相似序列通常具有进化相关性这一假设。作为程序的一部分，Clustal 可以输出用于构建进化树的数据。Clustal 程序有许多版本，ClustalW(它的 PC 版本是 CLUSTALX)。CLUSTALW 是一种渐进的比对方法，先将多个序列两两比对构建距离矩阵，反应序列之间两两关系；然后根据距离矩阵计算产生系统进化指导树，对关系密切的序列进行加权；然后从最紧密的两条序列开始，逐步引入临近的序列并不断重新构建比对，直到所有序列都被加入为止。CLUSTALW 的程序可以自由使用，在 NCBI 的 FTP 服务器上可以找到下载的软件包。CLUSTALW 程序用选项单逐步指导用户进行操作，用户可根据需要选择打分矩阵、设置空位罚分等。EBI 的主页还提供了基于 Web 的 CLUSTALW 服务，用户可以把序列和各种要求通过表单提交到服务器上，服务

器把计算的结果用 Email 返回用户。CLUSTALW 对输入序列的格式比较灵活，可以是前面介绍过的 FASTA 格式，还可以是 PIR、SWISS-PROT、GDE、Clustal、GCG/MSF、RSF 等格式。输出格式也可以选择，有 ALN、GCG、PHYLP 和 GDE 等，用户可以根据自己的需要选择合适的输出格式。用 CLUSTALW 得到的多序列比对结果中，所有序列排列在一起，并以特定的符号代表各个位点上残基的保守性，“*”号表示保守性极高的残基位点；“.”号代表保守性略低的残基位点。

EBI 的 CLUSTALW 网址是：<http://www.ebi.ac.uk/clustalw/>。

下载 CLUSTALW 的网址是：<ftp://ftp.ebi.ac.uk/pub/software/>。

7.5.4 系统进化树

系统发育学研究的是进化关系，系统发育分析就是要推断或者评估这些进化关系。通过系统发育分析所推断出来的进化关系一般用分枝图表（进化树）来描述，这个进化树就描述了同一谱系的进化关系，包括了分子进化（基因树）、物种进化以及分子进化和物种进化的综合。因为“clade”这个词（拥有共同祖先的同一谱系）在希腊文中的本意是分支，所以系统发育学有时被称为遗传分类学（cladistics）。在现代系统发育学研究中，研究的重点已经不再是生物的形态学特征或者其他特性，而是生物大分子尤其是序列。构建系统进化树的主要步骤是比对序列，建立取代模型，建立进化树以及进化树评估。

(1) 建立数据模型（比对）

建立一个比对模型的基本步骤包括：选择合适的比对程序；然后从比对结果中提取系统发育的数据集，至于如何提取有效数据，取决于所选择的建树程序如何处理容易引起歧义的比对区域和插入/删除序列（即所谓的 indel 状态或者空位状态）。

一个典型的比对过程包括：首先应用 ClustalW 程序，然后进行手工比对，最后提交给一个建树程序。这个过程有如下特征选项：（1）部分依赖于计算机（也就是说，需要手工调整）；（2）需要一个先验的系统发育标准（即需要一个前导树）；（3）使用先验评估方法和动态评估方法（推荐）对比对参数进行评估；（4）对基本结构（序列）进行比对（对于亲水氨基酸，推荐引入部分二级结构特征）；（5）应用非统计数学优化。这些特征选项的取舍依赖于系统发育分析方法。

(2) 决定取代模型

取代模型既影响比对，也影响建树；因此需要采用递归方法。对于核酸数据而言，可以通过取代模型中的两个要素进行计算机评估，但是对于氨基酸和密码子数据而言，没有什么评估方案。其中一个要素是碱基之间相互取代的模型；另外一个要素是序列中不同位点的所有取代的相对速率。还没有一种简单的计算机程序可以对较复杂的变量（比如，位点特异性或者系统特异性取代模型）进行评估，同样，现有的建树软件也不可能理解这些复杂变量。

(3) 建树方法

三种主要的建树方法分别是距离、最大节约（maximum parsimony, MP）和最大似然（maximum likelihood, ML）。最大似然方法考察数据组中序列的多重比对结果，优化出拥有一定拓扑结构和树枝长度的进化树，这个进化树能够以最大

的概率导致考察的多重比对结果。距离树考察数据组中所有序列的两两比对结果，通过序列两两之间的差异决定进化树的拓扑结构和树枝长度。最大节约方法考察数据组中序列的多重比对结果，优化出的进化树能够利用最少的离散步骤去解释多重比对中的碱基差异。

(4) 评估进化树和数据

现在已经有一些程序可以用来评估数据中的系统发育信号和进化树的健壮性。对于前者，最流行的方法是用数据信号和随机数据作对比实验（偏斜和排列实验）；对于后者，可以对观察到的数据重新取样，进行进化树的支持实验（非参数自引导和对折方法）。似然比例实验可以对取代模型和进化树都进行评估。

7.6 第六章基因预测和基因结构分析

人们获得各种核酸和蛋白质序列的目的是了解这个序列在生物体中充当了怎样的角色。例如，DNA 序列中重复片段、编码区、启动子、内含子/外显子、转录调控因子结合位点等信息；蛋白质的分子量、等电点、二级结构、三级结构、四级结构、膜蛋白的跨膜区段、酶的活性位点、以及蛋白质之间相互作用等结构和功能信息。虽然用实验的方法是多年以来解决这类问题的主要途径，但新的思路是利用已有的对生物大分子结构和功能特性的认识，用生物信息学的方法通过计算机模拟和计算来“预测”出这些信息或提供与之相关的辅助信息。由于生物信息学的特点，可以用较低的成本和较快的时间就能获得可靠的结果。近10年来生物学序列信息的爆炸性增长大大促进了各种序列分析和预测技术的发展，目前已经可以用理论预测的方法获得大量的结构和功能信息。要注意的是，尽管各种预测方法都基于现有的生物学数据和已有的生物学知识，但在不同模型或算法基础上建立的不同分析程序有其一定的适用范围和相应的限制条件，因此最好对同一个生物学问题尽量多用几种分析程序，综合分析各种方法得到的结果和结果的可靠性。此外，生物信息学的分析只是为生物学研究提供参考，这些信息能提高研究的效率或提供研究的思路，但很多问题还需要通过实验的方法得到验证。在构建一个基因结构预测模型时，一些主要问题是值得注意的：（1）对真核生物序列，遮蔽重复序列应先于其它分析过程；（2）大多程序都有特定生物物种适用性；（3）许多程序只能特定适用于基因组 DNA 数据或者只适用于 cDNA 的数据；（4）序列的长度也是一个重要因素。

7.6.1 针对核酸序列的预测方法

针对核酸序列的预测就是在核酸序列中寻找基因，找出基因的位置和功能位点的位置，以及标记已知的序列模式等过程。在此过程中，确认一段 DNA 序列是一个基因需要多个证据的支持。一般而言，在重复片段频繁出现的区域里，基因编码区和调控区不太可能出现；如果某段 DNA 片段的假想产物与某个已知的蛋白质或其它基因的产物具有较高序列相似性的话，那么这个 DNA 片段就非常可能属于外显子片段；在一段 DNA 序列上出现统计上的规律性，即所谓的“密码子偏好性”，也是说明这段 DNA 是蛋白质编码区的有力证据；其它的证据包括与“模板”序列的模式相匹配、简单序列模式如 TATABox 等相匹配等。一般而言，确定基因的位置和结构需要多个方法综合运用，而且需要遵循一定的规则：对于真核生物序列，在进行预测之前先要进行重复序列分析，把重复序列标记出来并

除去；选用预测程序时要注意程序的物种特异性；要弄清程序适用的是基因组序列还是 cDNA 序列；很多程序对序列长度也有要求，有的程序只适用于长序列，而对 EST 这类残缺的序列则不适用。

(1) 重复序列分析

对于真核生物的核酸序列而言，在进行基因辨识之前都应该把简单的大量的重复序列标记出来并除去，因为很多情况下重复序列会对预测程序产生很大的扰乱，尤其是涉及数据库搜索的程序。常见的重复序列分析程序有 GrailEXP 等，可以在 Web 界面上使用这些程序，或者用 Email 来进行。

(2) 数据库搜索

把未知核酸序列作为查询序列，在数据库里搜索与之相似的已有序列是序列分析预测的有效手段，在上一节中已经专门介绍了序列比对和搜索的原理和技术。但值得注意的是，由相似性分析作出的结论可能导致错误的流传；有一定比例的序列很难在数据库里找到合适的同源伙伴。对于 EST 序列而言，序列搜索将是非常有效的预测手段。

(3) 编码区统计特性分析

统计获得的经验说明，DNA 中密码子的使用频率不是平均分布的，某些密码子会以较高的频率使用而另一些则较少出现。这样就使得编码区的序列呈现出可察觉的统计特异性，即所谓的“密码子偏好性”。利用这一特性对未知序列进行统计学分析可以发现编码区的粗略位置。这一类技术包括：双密码子计数(统计连续两个密码子的出现频率)；核苷酸周期性分析(分析同一个核苷酸在3,6,9,...位置上周期性出现的规律)；均一/复杂性分析(长同聚物的统计计数)；开放可读框架分析等。常见的编码区统计特性分析工具将多种统计分析技术组合起来，给出对编码区的综合判别。著名的程序有 GRAIL 和 GenMark 等，GRAIL 提供了基于 Web 的服务。

7.6.2 启动子分析

启动子是基因表达所必需的重要序列信号，识别出启动子对于基因辨识十分重要。有一些程序根据实验获得的转录因子结合特性来描述启动子的序列特征，并依次作为启动子预测的依据，但实际的效果并不十分理想，遗漏和假阳性都比较严重。总的来说，启动子仍是值得继续研究探索的难题。

7.6.3 内含子/外显子剪接位点

剪接位点一般具有较明显的序列特征，但是要注意可变剪接的问题。由于可变剪接在数据库里的注释非常不完整，因此很难评估剪接位点识别程序预测剪接位点的敏感性和精度。如果把剪接位点和两侧的编码特性结合起来分析则有助于提供剪接位点的识别效果。

7.6.4 翻译起始位点

对于真核生物，如果已知转录起始点，并且没有内含子打断5'非翻译区的话，“Kozak 规则”可以在大多数情况下定位起始密码子。原核生物一般没有剪接过程，但在开放阅读框中找正确的起始密码子仍很困难。这时由于多顺反操纵子的

存在，启动子定位不象在真核生物中起关键作用。对于原核生物，关键是核糖体结合点的定位，可以由多个程序提供解决方案。

7.6.5 翻译终止信号

PolyA 和翻译终止信号不象起始信号那么重要，但也可以辅助划分基因的范围。

7.6.7 其它综合基因预测工具

除了上面提到的程序之外，还有许多用于基因预测的工具，它们大多把各个方面的分析综合起来，对基因进行整体的分析和预测。多种信息的综合分析有助于提高预测的可靠性，但也有一些局限：物种适用范围的局限；对多基因或部分基因，有的预测出的基因结构不可靠；预测的精度对许多新发现基因比较低；对序列中的错误很敏感；对可变剪接、重叠基因和启动子等复杂基因语法效果不佳。相对不错的工具有 GENSCAN 和 GeneFinding，可以通过 Web 页面或 Email 获得服务。

这些程序的主要局限性在于：（1）复合的算法目前只适用少数物种；（2）所有的程序（除了 GENSCAN）在输入序列中包含多基因或者部分基因时，所预测的外显子仍可靠，但所预测的基因结构就不一定了；（3）由于尚不完全清楚的原因，预测精度可能比原先想象的低得多，尤其是对新发现的基因。（Burset 和 Guigó, 1996，用百来个简单实例来标定了能得到的程序，结果无一能正确预测出多于一半的外显子）；（4）大多复合算法都明显对测序错误十分敏感（Burset 和 Guigó, 1996）；以及（5）交替剪接、重叠基因和启动子结构这样的基因语法结构仍超出当前程序的处理能力。

既然这些程序中没有一个十全十美，它们都覆盖了一些不同算法，都在迅速进步，因此强烈建议分析每个序列时采用3到4个不同程序，并仔细对比其结果。如果某个工具会经常用到，就值得用大量已知结果的序列对其进行测试，以便对算法适用性有所了解。

7.7 第七章蛋白质性质和结构分析

传统的生物学认为，蛋白质的序列决定了它的三维结构，也就决定了它的功能。由于用 X 光晶体衍射和 NMR 核磁共振技术测定蛋白质的三维结构，以及用生化方法研究蛋白质的功能效率不高，无法适应蛋白质序列数量飞速增长的需要，因此近几十年来许多科学家致力于研究用理论计算的方法预测蛋白质的三维结构和功能，经过多年努力取得了一定的成果。

7.7.1 分析蛋白质的一级结构

蛋白质序列一级结构分析包括了对理化性质和序列模式的分析。蛋白质理化性质的分析通常包括：蛋白质的分子量、等电点(pI)、氨基酸组成、疏水性和亲水性分析等。目前已经开发了众多的蛋白质序列理化性质计算工具，且大多提供网络服务或允许自由下载安装。除了下表中列出的由 ExPASy 整理的蛋白质序列理化性质计算的工具体外，ANTHEPROT、DNAMAN、BioEdit 等也是较好的理化性质计算工具。理化性质对于进一步确定蛋白质的亚细胞定位、功能等非常有用，比如利用疏水残基与跨膜螺旋间的关系可以预测蛋白质序列是否跨膜，利用氨基

酸组成成份可以预测蛋白质序列的亚细胞定位等。

(1) ProtParam

ProtParam 是计算输入序列的等电点(pI)、分子量(Mw)、氨基酸组成、消光系数、半衰期等理化性质的工具。对 pI 的确定基于早期研究中将蛋白质从由中性到酸性变性条件下迁移过程中所获得的 pK 值,对于碱性蛋白质所得到的 pI 值可能不准确。分子量的计算是把序列中每个氨基酸的同位素平均分子量加在一起,再加上一个水分子的分子量。用户可以把序列整理为 FASTA 格式,或提供 SwissProt 标识。若用户提供了序列,该工具会自动计算全序列的 pI 和分子量;若用户提供的是 SwissProt 标识,程序会显示该条目的描述和物种记录,此时计算将在片段上进行,而不是针对整个序列。

(2) ProtScale

可以计算给出蛋白质序列的氨基酸疏水区和亲水区,用户可以把序列整理为 FASTA 格式,或提供 SwissProt 标识,需选择合适的 aminoacidscale。

(3) REP

可以分析指定蛋白质序列中的重复序列模块,用户可以把序列整理为 FASTA 格式,或提供 SwissProt 标识。已有的可供查询的模块包括: Ankyrin, Armadillo, HAT, HEAT, HEAT_AAA, HEAT_ADB, HEAT_IMB, Kelch, Leucine Rich Repeats, PFTA, PFTB, RCC1, TPR, WD40 等。

7.7.2 分析蛋白质的二级结构

二级结构是指 α 螺旋和 β 折叠等规则的蛋白质局部结构元件。不同的氨基酸残基对于形成不同的二级结构元件具有不同的倾向性。按蛋白质中二级结构的成分可以把球形蛋白分为全 α 蛋白、全 β 蛋白、 $\alpha+\beta$ 蛋白和 α/β 蛋白等四个折叠类型。预测蛋白质二级结构的算法大多以已知三维结构和二级结构的蛋白质为依据,用过人工神经网络、遗传算法等技术构建预测方法。还有将多种预测方法结合起来,获得“一致序列”。总的来说,二级结构预测仍是未能完全解决的问题,一般对于 α 螺旋预测精度较好,对 β 折叠差些,而对除 α 螺旋和 β 折叠等之外的无规则二级结构则效果很差。

(1) nnPredict

用神经网络方法预测二级结构,蛋白质结构类型分为全 α 蛋白、全 β 蛋白和 α/β 蛋白,输出结果包括“H”(螺旋)、“E”(折叠)和“-”(转角)。这个方法对全 α 蛋白能达到79%的准确率。

(2) SOPMA

位于法国里昂的 CNRS (Centre National de la Recherche Scientifique) 使用独特的方法进行蛋白质二级结构预测。它不是用一种,而是5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 (GOR) 方法 (Garnier 等, 1996)、Levin 同源预测方法 (Levin 等, 1986)、双重预测方法 (Deléage 和 Roux, 1987)、作为前面 PredictProtein 一部分的 PHD 方法和 CNRS 自己的 SOPMA 方法 (Geourjon 和 Deléage, 1995)。简单的说, SOPMA 这种自优化的预测方法建立了已知二级结构序列的次级数据

库，库中的每个蛋白质都经过基于相似性的二级结构预测。然后用次级库中得到的信息去对查询序列进行二级结构预测。

(3) Paircoil

Paircoil 将输入序列与数据库中 coiled-coils 比较产生相似性分值。通过比较这个分值与球蛋白，卷曲螺旋蛋白的分值的分布，可以计算提交序列将会采取的卷曲螺旋构象的概率分值。

7.7.3 分析蛋白质的三级结构

蛋白质三维结构预测时最复杂和最困难的预测技术。研究发现，序列差异较大的蛋白质序列也可能折叠成类似的三维构象，自然界里的蛋白质结构骨架的多样性远少于蛋白质序列的多样性。由于蛋白质的折叠过程仍然不十分明了，从理论上解决蛋白质折叠的问题还有待进一步的科学发展，但也有了一些有一定作用的三维结构预测方法。最常见的是“同源模建”和“Threading”方法。前者先在蛋白质结构数据库中寻找未知结构蛋白的同源伙伴，再利用一定计算方法把同源蛋白的结构优化构建出预测的结果。后者将序列“穿”入已知的各种蛋白质的折叠子骨架内，计算出未知结构序列折叠成各种已知折叠子的可能性，由此为预测序列分配最合适的折叠子结构。除了“Threading”方法之外，用 PSI-BLAST 方法也可以把查询序列分配到合适的蛋白质折叠家族，实际应用中发现这个方法的效果也不错。

在谈及这种或那种预测技术之前要预先说明的是，无论用哪种方法，这些结果都是预测。不同的方法，采用了不同的算法，可能产生相同或不同的结果。但有一点很重要：弄清楚某种方法的原理，而不是仅把算法当作一个“黑箱”。因为一种方法可能对特定实例很合适，而对另一个则完全不对。虽然如此，存在一种强大合作的潜力：正确应用这些预测技术，参照以主要的生化数据，就能提供有关蛋白质结构与功能的有价值信息。

7.7.4 分析膜蛋白质

膜蛋白可分为膜附着蛋白和膜镶嵌蛋白两大类，膜蛋白的跨膜区一般形成 α -螺旋，膜附着蛋白通常由膜镶嵌蛋白剪切形成。

(1) SOSUI

东京农业科技大学(Tokyo University of Agriculture and Technology)提供的膜蛋白分类和二级结构预测在线工具，可分析蛋白质跨膜区。预测结果分为两类：membrane protein（显示 transmembrane helix）和 soluble protein（非膜蛋白）。

(2) DGPI

Prediction of GPI-anchor and cleavage sites，预测 GPI 的切点与接点位置，判断是否是膜附着蛋白。

7.7.5 分析蛋白质的翻译后修饰

蛋白质序列的翻译后修饰分析包括信号肽分析、亚细胞定位、糖基化修饰等。这些性质多半可直接由分析其序列而获得，并且这些性质大多与蛋白质的结构和功能相关，比如羧基端含有 KDEL 序列特征的蛋白质将被引向内质网，从而通

过确定其亚细胞预测其功能。又如，亚细胞定位与功能相关，因为蛋白质必须在一定的细胞位置与一定的蛋白质相互作用才能完成特定的功能。蛋白质如果没有定位到正确的细胞位置还可能引起比如 cancer 和 Alzheimer 等疾病。另外，细菌的 extracellular 蛋白可以沿着类似的 default pathway 进入到真核生物细胞中。更进一步，通过与某些真核生物蛋白形成复合体进入特定的亚细胞位置，从而激发或抑制某些特定的细胞功能，甚至直接导致细胞死亡。这对于疾病治疗和预防，如治疗癌症，以及寻找疾病疫苗或生物制剂药物，如构造细菌武器及其治疗疫苗等具有重要意义。再如，蛋白质的磷酸化和去磷酸化几乎调节着生命活动的所有过程，包括细胞的增殖、发育和分化，神经活动，肌肉收缩，新陈代谢，肿瘤发生等。尤其在细胞应答外界刺激时，蛋白质磷酸化是目前所知道的最主要的信号传递方式。如在中枢神经系统中，几乎所有的兴奋性神经递质信号受体调控的方式都是磷酸化与去磷酸化。

(1) SignalP

SignalP 可以对革兰氏阳性菌，革兰氏阴性菌和真核生物的蛋白质序列进行信号肽分析。预测采用了隐马尔可夫和人工神经网络技术的融合。其训练数据是从 SwisProt 第29版中挑选来的，分为真核生物，革兰氏阳性菌，革兰氏阴性菌三组。

(2) NetOGlyc 和 NetNGlyc

分别分析 O-连接糖链和 N-连接糖链的连接位点，判断待分析蛋白是否可能发生糖基化。

7.7.6 分析蛋白质的亚细胞定位

蛋白质必须在一定的亚细胞器上才能正确行使其功能。同时也只有在相同或相近的亚细胞位置上蛋白质间才会有相互作用。亚细胞位置异常的蛋白质通常还会引起如癌症、老年痴呆症等疾病。亚细胞定位的预测目前有同源传递、基于序列 motifs 识别的方法、abinitio 方法、蛋白质-蛋白质相互作用方法等。这里对 PSORTB 的预测原理和方法作简单介绍。PSORTB1.0forbacteria 预测5类亚细胞位置：

cytoplasm, the inner membrane, the periplasm, the outer membrane, the extracellular space

（针对革兰氏阴性菌）。采用的训练数据集是从 SWISS-PROT40.29中收集的注释了亚细胞位置的革兰氏阴性菌序列。工具的使用方法见网页 <http://www.psort.org/>。

7.7.7 分析化学因子作用蛋白质的位点

PeptideCutter: 分析蛋白质在各种蛋白酶和化学试剂处理后的内切产物。蛋白酶和化学试剂包括胰蛋白酶、糜蛋白酶、LysC、溴化氰、ArgC、AspN 和 GluC 等。

7.8 第八章农业类数据库的利用

生物信息学在农作物基因组分析中的深入应用无疑会加速农业生产的发展。全世界的农学家、生物信息学家已充分认识到这一点。近年来,以“水稻基因组计划”为代表的农作物基因图谱研究为生物信息学的农业应用打下了良好基础。一方面,通过比较基因组学、表达分析和功能基因组分析,识别重要基因、发现新基因、加快基因克隆的速度,为培育转基因作物、改良作物的质量和数量性状奠定基础。农业生物信息学与常规育种技术相结合,提高育种效率、创新遗传资源、加快育种进程,已成为农作物育种的发展趋势。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。以信号受体和转录途径组分分析为基础,进行农业化合物设计,结合化学信息学方法,鉴定可用于杀虫剂和除草剂的潜在化学成分,将成为生物信息学在农业上的另一推动力,这将保证农作物高产优质和绿色环保的市场要求。

另一方面,生物信息学应用于农业可以充分利用植物遗传资源,保护农作物遗传多样性。对于我国这样的农业大国,运用先进有效的生物信息学研究手段,结合我国丰富的特有的遗传资源,开展中国优良农作物资源的单核苷酸多态性和插入缺失多态性的研究,分离、克隆有自主知识产权的有重要经济价值的新基因及重要的基因表达调控元件,发现控制优良性状基因的分子标记,将极大的加快我国的农作物应用研究工作,实现我国农业持续发展。

通过生物信息学推动农业基础研究及应用研究的关键在于获取主要农作物和家畜家禽的完整基因组建立基因组数据库。利用生物信息数据库对基因、基因的结构、基因产物的功能分析将成为农业基础与应用研究工作中必不可少的技术手段。本章对国际上已经建立了一些重要的农作物基因组数据库进行介绍。

7.8.1 Gramene

网址: <http://www.gramene.org/>

谷类比较图谱资源的网站。是一个协助性的、以网络为基础的公开性数据资源,致力于稻科植物类的比较基因组分析。他们的目标是使用公用工程信息促进交叉物种的同源关系研究,这些公用工程包括基因组、EST 序列、蛋白质结构和功能分析、遗传学和物理图谱、生物化学通路的阐述、表型特征和突变的 QTL 定位及描述。作为一个信息源, Gramene 的目的是在公共资源中为资料提供更多的价值,便于研究者用水稻基因组序列来鉴定和阐述稻科作物的相应基因、通路和表型。

7.8.2 Soybase

网址: <http://soybase.agron.iastate.edu>

美国农业部大豆基因数据库,包含了大豆的遗产、表型及其他信息。可以查到大豆的各种遗传图谱和物理图谱,以及某些功能基因的信息。

7.8.3 GrainGenes

网址: <http://www.graingenes.org>

GrainGenes 是美国农业部和国家农业图书馆的植物基因组计划支持的麦、燕

麦和甘蔗遗传数据库。数据组成和检索方法与 Soybase 相似。

7.8.4 ArkDB

网址: <http://www.thearkdb.org>

农业相关和其他动物的基因组数据库。其中包含有常见家畜, 如: 猪、牛、羊等物种的遗传图谱和物理图谱。

其他网络公共农业数据库:

1. [美国农业部\(USDA\)研究数据库 USDA Research Database](#) 这是由美国农业部 (USDA) 创建的可检索数据库网页, 该数据库系统提供了美国农业部正在进行的研究项目和已完成项目的资助人。由该网页可超链进入美国农业部的各个部门进行查询, 如 USDA 的营养研究, 生物化学与生物物理研究, 微生物学研究, 遗传学, 病毒学, 分子生物学, 药理学, 工程学, 化学等领域。
2. [美国农业研究局\(ARS\)数据库 ARS Database](#) 这个网页给出了24个与农业有关的数据库的超链接, 可分别查询: ARS 水数据库, 牲畜[基因组](#)图, 国家农业图书馆数据库, 食品成分数据, 真菌保藏与信息, 玉米基因数据库, 小谷物库, 国家植物种质系统, 油料作物种子库, 大豆库, 猪[基因组](#)图等。
3. [美国食品与农业局统计数据库 FAO STAT Statistics Database](#)
4. [美国农业部作物数据库 USDA National PLANTS Database and Projects](#)
5. [加利福尼亚植物数据库 Cal Flora Database](#)
6. [宾夕法尼亚植物数据库 Pennsylvania Flora Project](#)
7. [食肉植物数据库 Carnivorous Plant Database](#)
8. [食肉植物知识数据库 Carnivorous Plant Knowledge Database](#)
9. [食用园艺数据库 Edible Landscaping Database](#)
10. [基因组数据库 Bovine Genome Database](#)
11. [植物病毒数据库 Plant Viruses Online-VIDE Database](#) 爱达荷大学。
12. [裸子植物数据库 Gymnosperm Database](#) 所有裸子植物: 针叶树、凤尾松等。
13. [BGRG 降水模拟数据库 BGRG Rainfall Simulation Database](#)
14. [植物化学和人文植物数据库 Phytochemical and Ethnobotanical Databases](#)
15. [食物辐射 \[国家食物安全数据库\] Facts About Food Irradiation \[National Food Safety Database\]](#) 检查辐射食物和消费者。
16. [CAB 国际 CAB International](#)
17. [水生植物数据库 Center for Aquatic and Invasive Plants](#)
18. [植物病害常用名词 Common Names of Plant Diseases](#) 美国植物病理学学会的数据库
19. [密苏里大学, 哥伦比亚 - 农学院, 食物与自然资源 University of Missouri, Columbia-College of Agriculture, Food and Natural Resources](#)
20. [谷物数据库 Grain Genes](#) 美国农业部资助, 有小麦、大麦、黑麦、燕麦、甘蔗的分子和表形信息。
21. [花粉和孢粉学数据库 PalDat](#)
22. [花卉数据库 Flowerweb](#) 荷兰花卉商会
23. [林业世界 Forestworld](#)
24. [菲律宾国家草药数码库 Philippine National Herbarium Digital Library](#)
25. [工业化农业数据库 Industrial Agriculture-USAClearinghouse](#)

26. [花卉数据库 Flowerbase](#)
27. [非洲-荷兰农业合作数据库 SPAAR&CGIAR](#)
28. [猪基因数据库 PigGenomeMapping](#)
29. [园艺与作物科学数据库 HorticultureandCropScienceinVirtualPerspective](#)
30. [豆类基因数据库 BeanGenes](#)
31. [康涅狄格草药学院 ConnecticutCollegeHerbarium\(CCNL\)](#)
32. [有毒植物数据库 GuidetoPoisonousPlants](#) 科罗拉多州立大学兽医与生物医学学院
33. [世界濒危植物数据库 ThreatenedPlantsoftheWorld](#)
34. [东北食品合作系统 NortheastFoodSystemPartnership](#)
35. [渔业数据库 FishIndustryNet](#)
36. [食品数据库 OurFood](#)
37. [BSEatCABI](#)
38. [俄亥俄州立大学土壤品质鉴定实验室 OhioStateUniversitySoilCharacterizationLab](#) 有俄亥俄土壤资源数据库
39. [加拿大猪改良中心 CanadianCenterforSwineImprovement](#)
40. [植物信息国际组织 InternationalOrganizationforPlantInformation](#) 国际合作建立植物分类信息数据库
41. [兰花数据库 Orchidlink](#) 20多个国家的兰花种植商及兰花学会。
42. [植物数据库 PlantMaster](#) 电子园艺杂志，供园林设计专家选用植物。
43. [植物数据库 Floraguide](#) 园林和园艺信息、公司、讲座数据库和软件。
44. [土豆数据库 PotatoResearchandExtension](#) 土豆品种、土豆蚜虫、土豆交易信息、最新虫害信息。
45. [澳大利亚羊驼协会 AustralianAlpacaAssociation](#)
46. [国际花商词典 InternationalFloristDirectory](#)
47. [树林数据库 SmartForest](#) 不到6000棵树的交互式可视化树木数据库。成为树木生长、死亡、虫害简化的生态模型。
48. [威斯康辛大学麦迪逊合作公园研究项目 UniversityofWisconsin,MadisonCooperativeParkStudiesUnit\(WICPSU\)](#) 美国地质调查署资助，有国家公园地衣和菌类数据库。
49. [Birkner 欧洲和国际纸业数据库 BirknerEuropeanandInternationalPaperWorld](#)
50. [蕨类植物数据库 FernsALaAkrón,OH](#)
51. [美洲驼数据库 LlamaSearchandResourceCenter](#)
52. [美国农业部森林全球定位系统 USDAForestServiceGlobalPositioningSystem\(GPS\)Page](#) 全球定位系统（GPS）和地理信息系统（GIS）用于森林覆盖地区和偏远/荒野地区。
53. [医用和毒性植物数据库 MedicinalandPoisonousPlantDatabases](#)
54. [综合灾害管理资源数据库 DatabaseofIPMResources\(DIR\)](#)
55. [国家食品安全数据库 NationalFoodSafetyDatabase](#) 为消费者、制造商提供食品安全资料。包括家用罐头指南。
56. [日本园林数据库 JGarden:TheJapaneseGardenDatabase](#) 有关日本园林的历史、构造、人物、语言、图案等。
57. [INVADERS（侵入者）数据库 INVADERSDatabaseSystem](#) 美国西北部外来植物名称、分类登记、管理信息。

58. [美国印第安人文植物学数据库 AmericanIndianEthnobotanyDatabase](#)
59. [世界牛市 CattleOfferingsWorldwide\(C.O.W\)](#) 互联网上进行牛及其胚胎的买卖交易，数据库适应买卖双方。
60. [大豆数据库 SoyBase](#) 美国农业部大豆基因数据库
61. [园艺数据库 GardenGuides](#)
62. [园艺大事记 CalendarofGardenEvents](#) 可按地点、时间、关键词检索的园艺大事数据库，用户可加入自己的数据。
63. [詹森斯华伦草药库 JasonSwallenHerbarium-OhioWesleyanUniversity](#) 俄亥俄卫斯理大学
64. [雪茄数据库 CigarAficionado](#)
65. [麻种数据库 Hempseed.com](#)
66. [农作物数据管理系统 CropDataManagementSystems,Inc.](#) 农业化学数据库，信息来自71家化学厂商。
67. [棕榈树数据库 PalmsOnline](#)
68. [种植数据库 PlantMaster](#)
69. [谷物仓库管理数据库 GanaraskaSystems](#)
70. [农场设备数据库 Iron-Web](#)
71. [蝼蛄知识库和教程 MoleCricketKnowledgebaseandTutorials](#) 佛罗里达大学
72. [兰花数据库 OrchidWeblopedia,The](#)
73. [农业信息数据库 AgDB](#)
74. [牧牛场、牧马场数据库 AgDirect](#)

7.9 第九章核酸序列的其他分析方法

7.9.1 序列综合分析软件

(1) Bioedit

BioEdit 软件是一个性能优良的免费分子生物学应用软件，可在 Windows95/98/NT/2000/XP 中运行，它的基本功能是提供蛋白质核酸序列的编辑排列处理和分析。该软件有一个简单亲切的界面，集成很多其他的序列比对分析软件，功能强大，另外该软件还有很多有用的相关站点连接。虽然该软件看起来结构简单，但却又很强的扩充性，可以自由整合许多软件。

(2) EBItools

欧洲生物信息学研究所 (EBI) 提供了许多生物信息学在线分析工具，包括如下几大类：

相似性和同源性分析：如 BLAST 和 FASTA 程序

蛋白质功能分析：如 InterProScan 可查询蛋白质序列中的特殊模块

序列分析：如 ClustalW 程序

结构分析：如 MSDfold 和 DALI 可分析蛋白质的结构并与 PDB 数据库的信息进行比较

(3) 其他商业软件：

DNAStar 即著名的 LasergeneSuite，由 EditSeqMegAlign、GeneQuestMapDrawPrimerSelectProteanSeqManII 七个模块组成，该软件的 MegAlign 模块，可以对多达64000的片段进行拼装。整个拼装过程即时显示，并提示可能的完成时间。拼装结果采用序列、策略等方式显示。DNAStar 是哈佛大学医学院是使用的序列分析软件，可见其功能强大。

(4) Omiga

大部分对核酸蛋白的序列分析功能，在 Omiga2.0中都能找到；而且界面非常友好。Omiga 作为强大的蛋白质、核酸分析软件，它还兼有引物设计的功能。主要功能：编辑、浏览、蛋白质或核酸序列，分析序列组成。用 Clustal.W 进行同源序列比较，发现同源区。实现了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。查找核酸限制性酶切位点、基元 (Motif) 及开放阅读框 (ORF)，设计并评估 PCR、测序引物。查找蛋白质解蛋白位点 (ProteolyticSites)、基元、二级结构等。查寻结果可以以图谱及表格的显示，表格设有多种分类显示形式。利用 Mange 快捷键，用户可以向限制性内切酶、蛋白质或核酸基元、开放阅读框及蛋白位点等数据库中添加或移去某些信息。每一数据库中都设有多种查寻参数，可供选择使用。用户也可以添加、编辑或自定义某些查寻参数。可从 MacVectorTM、WisconsinPackageTM 等数据库中输入或输出序列。另外，该软件还提供了一个很有特色的类似于核酸限制酶分析的蛋白分析，对蛋白进行有关的多肽酶处理后产生多肽片段。

VectorNTISuite 不喜欢装备各种专业性强的软件，而希望用一个综合性的软件代替的同志可以选择本软件。本阶段的大部分功能它都有。该软件具体特有良

好的数据库管理（增加、修改、查找），对要操作的数据放在一个界面相同的数据库中统一管理。软件中的大部分分析可以通过在数据库中进行选定（数据）->分析->结果（显示、保存和入库）三步完成。在分析主界面，软件可以对核酸蛋白分子进行限制酶分析、结构域查找等多种分析和操作，生成重组分子策略和实验方法，进行限制酶片段的虚拟电泳，新建输入各种格式的分子数据、加以注释，输出高质量的图像。VectorNTISuite 还有以下独立的分析程序，完成相关分析。这些独立的程序，可以通过选定->分析->结果三步调用。

（4）DNATools

与 Omiga, DNAsis, PCgene 等软件属于同一类的综合性软件,操作简单功能多。DNATools 设计的用户友好、强壮，以便快速、方便地获取、贮藏和分析序列及数据库查询获得的序列相关信息。DNATools 包容性很好，能把几乎所有文本文件打开作为序列。当程序不能辨别序列的格式时（通过寻找常用序列格式的特征），会显示这个文件的文本形式，以便你编辑生成正确的蛋白质或 DNA 序列，编辑后可以再被载入程序。若你的序列是 DNATools 格式时（DNA 或寡核苷酸序列），程序不加注解的载入序列，程序模式调整成可以接受载入的数据类型（蛋白质、DNA 和寡核苷酸引物序列）。在一个项目中可以加入几千个序列或引物，并在整个项目中分析这些序列及标题。这个程序的一个特点是给每个序列或引物添加文本标题。这样就可以用自定义的标题识别序列，而不必通过它们的文件名。

7.9.2 引物分析软件

（1）**PrimerPremier** 顾名思义，该软件就是用来进行引物设计的。可以简单地通过手动拖动鼠标以扩增出相应片段所需的引物，而在手动的任何时候，下面显示各种参数的改变和可能的二聚体、异二聚体、发夹结构等。也可以给定条件，让软件自动搜索引物，并将引物分析结果显示出来。而且进行这些操作非常简单

（2）**Oligo6.57** 引物分析著名软件，主要应用于核酸序列引物分析设计软件，同时计算核酸序列的杂交温度（ T_m ）和理论预测序列二级结构。

7.9.3 基因定位软件

Forward electronic PCR & Reverse electronic PCR 查找 DNA 序列的 STSs (sequence tagged sites, 序列标签位点) 的在线工具软件。

（1）在线信息数据库部分

SDBS 光谱数据库: http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi
简介: 很好的有机化合物光谱数据库, 包含六类光谱: EI-MS、FT-IR、H-NMR、C13-NMR、ESR、Raman。含 3 万余个化合物, 其中以商业化学试剂为主, 约 2/3 是 6 碳至 16 碳的化合物。数据大部分是其自行测定的, 并不断添加。可以通过化合物、分子式、分子量、CAS/SDBS 注册号、元素组成、光谱峰值位置/强度方式搜索。

生物核磁共振数据库: <http://bmrb.protein.osaka-u.ac.jp/deposit>

CRYSTAL 程序基因组数据库: <http://www.tcm.phy.cam.ac.uk/~mdt26/crystal.html>

计算化学比较和基准数据库(CCCBDB): <http://cccbdb.nist.gov>

简介: 此数据库包括各种量子化学方法、各种基组下对不同分子的各种属性的计算结果, 也包含实验数据。可用来对比不同方法计算结果优劣, 此数据库内容在

不断增加。

量化频率计算校正因子: <http://cccbdb.nist.gov/vibscale.asp>

简介: 实际上就是 CCCBDB 的一个子页面, 比较重要故单独列出。

IUPAC 金属络合物稳定常数数据库: <http://www.acadsoft.co.uk>

注: 需要付费, 可免费下载试用版。

NIST 化学数据库: <http://webbook.nist.gov/chemistry>

简介: 是美国国家标准与技术研究院 NIST 的基于 Web 的物性数据库。输入分子查找条件, 可获得分子量、CAS 登记号、各种热力学数据、谱图等信息, 部分分子包含 3D 结构。

RESPESPchargeDDatabase(REDDB)

<http://q4md-forcefieldtools.org/REDDB/index.php>

简介: 分子的 RESP 电荷的数据库

UppsalaElectronDensityServer: <http://eds.bmc.uu.se/eds>

简介: 用于评价蛋白质数据库中晶体结构电子密度。输入 pdbID (比如 1cbs) 进入后可以对各种内容做图。点击 EDSSummary 下面的 Go 按钮可以自动启动基于 java 的电子密度图可视化程序观看电子密度图, 注意不要开启浏览器的弹出窗口过滤。

上海有机所化学专业数据库: <http://202.127.145.134/scdb/default.htm>

简介: 十分有用的数据库, 免费注册。可获得分子的红外、质谱谱图、结构、物化性质、毒性、生物活性以及相关反应等。还包括中英互译、药品名称检索等功能。

EMSL 基因组数据库: <https://bse.pnl.gov/bse/portal>

Clarkson 大学相对论有效势数据库: <http://people.clarkson.edu/~pac/reps.html>

含重原子全电子 STO 基因组数据库: <http://www.scm.com/Downloads/zorabasis/Welcome.html>

原子间势参数数据库: <http://www.dfri.ucl.ac.uk/Potentials>

Stuttgart 赝势参数数据库: <http://www.theochem.uni-stuttgart.de/clickpse.en.html>

ChemBioFinder: <http://chembiofinder.cambridgesoft.com/SimpleSearch.aspx>

简介: 根据分子质量、名称或者自行绘制结构, 从几十万分子中搜索, 得到二维结构、Smiles、InCHI 字符串、分子量等简单信息。

Sigma-Aldrich 公司产品数据库: http://www.sigmaaldrich.com/Area/United_States.html

简介: 主要用来获得化合物 IR、NMR 谱图。右上角输入化合物的名字, 搜索到后进入相应条目, 如果在左侧有 FT-IR/Raman、FT-NMR 字样, 就可以进入察看, 没有则说明此化合物无光谱数据。也可以获得化合物的一些物性数据, 但不全面。

基本物理常数数据库: <http://physics.nist.gov/cuu/Constants/index.html>

简介: 可以查到精确的计算化学中涉及的物理常数及换算关系, 如 hartree→eV

百奥知识数据库: <http://tong.bioknow.cn/html/sites/database/index.htm>

简介: 一个生物信息数据库的比较全的列表, 每个数据库有简单官方介绍。

(2) 结构数据库部分

GLYCAM 寡糖数据库: <http://glycam.ccruc.uga.edu/CCRC/Library/index.jsp>

ICSD 无机晶体数据库: <http://icsd.ill.fr/icsd/index.php>

简介: 免费在线提供部分晶体结构信息及 cif 文件。

NDB 核酸数据库: <http://ndbserver.rutgers.edu>

简介：根据 NDBID，获得核酸坐标文件、出处等信息，类似 RCSB 蛋白质数据库。

PDBbind-CN: <http://www.pdbbind.org.cn/index.asp>

简介：收集了 pdb 数据库中的生物分子复合物。给出受体、配体名称、亲和性、序列等信息，可在线观看或下载结构，可根据配体名称、结构搜索含有此配体的复合物。

PDBWiki: http://pdbwiki.org/index.php/Main_Page

简介：基于 PDB 数据库，对蛋白质进行了简单分类，访问者可以给每个蛋白添加注释。

sc-PDB: <http://bioinfo-pharma.u-strasbg.fr/scPDB>

简介：收集了 PDB 数据库中含有可以为药物结合的位点的蛋白。可根据配体、蛋白、结合方式为特征进行搜索。

RCSBPDB 数据库: <http://www.rcsb.org/pdb/home/home.do>

HPDB 蛋白质数据库: <http://hpdb.hbu.edu.cn>

简介：河北大学的蛋白质数据库，可通过此库间接下载到 RCSB 蛋白质数据库的文件。有中文 PDB 文件格式的介绍，比较有用，还有另外一介绍蛋白质的些文章。

ZINC 化合物虚拟筛选数据库: <http://zinc.docking.org/index.shtml>

简介：根据自定义的化合物的性质，在 800 万种以上可买到的产品中进行筛选。可以下载到结构文件。

PubChem: <http://pubchem.ncbi.nlm.nih.gov>

简介：NCBI 下属的小分子数据库，包括化合物、物质、生物活性三大数据库，含上千万条目并不断增加。可通过分子结构、名称、分子式、分子量、XLogP、氢键信息方式查询。可以得到分子的简介、化学结构、XLogP（自动计算）、同义词、生物活性、毒性、药理学信息及分类、SMILE 和 InChI/key 字符串、相似化合物、2D 的 SDF 文件。一些结构还有 3DSDF 结构文件，进入条目后可在页面最下面点 SDF 按钮保存，可被一些软件直接读取，如 ChemBio3D。是一个很有用的获得小分子三维结构的方法。

SuperNature 天然产物数据库: <http://bioinformatics.charite.de/supernatural>

简介：几万种天然产物数据库，可通过名称、结构、相似度、LogP、分子量、分子式等信息搜索，也可以绘制结构或根据结构模版搜索，可在线观看结构，并获得净电荷、偶极矩、手形中心数目、可旋转键数目、氢键受/配体等信息。

蛋白质 pKa 数据库(PPD): <http://www.jenner.ac.uk/PPD>

蛋白质分类数据库(CATH): <http://www.cathdb.info>

简介：其中结构来自 PDB 数据库，半自动地对每个结构根据二级结构、形状、拓扑、同源性进行了分类，可以根据这些特点进行分类查询。

蛋白质结构分类数据库(SCOP): <http://scop.mrc-lmb.cam.ac.uk/scop>

简介：和 CATH 的功能类似，包含几万个蛋白质结构。但是是人工对每个蛋白结构进行分类，比 CATH 的分类更为合理。结构分类基于四个层次：class、fold、superfamily、family。

结构相似蛋白质家族数据库 (FSSP) :
<http://srs.ebi.ac.uk/srsbin/cgi-...i2u1RffMj+-lib+FSSP>

简介：此数据库对 pdb 数据库中的结构使用 Dali 算法进行了相似度计算，用以找到相似蛋白质。

(3) 在线工具部分

Dalserver: http://ekhidna.biocenter.helsinki.fi/dali_server

简介: 输入 PDBID 或者上传 PDB, 服务器会对此结构与 PDB 数据库中的结构用 dali 算法计算, 将其中有一定结构相似度的 PDB 列出。通过复选框选择几个结构, 可以对比序列, 以及在线观看它们重叠后的 3D 结构。

DaliDatabase(<http://ekhidna.biocenter.helsinki.fi/dali/start>)每年更新两次, 如果是在更新日期以前发布的 pdb, 可以直接在 DaliDatabase 里面查询之前已算好的结果, 而不必在 Daliserver 里面重新计算。

在线生成金刚石结构 (包括同晶体结构的硅、锗):

<http://turin.nss.udel.edu/research/diamondonline.html>

在线生成石墨结构: <http://turin.nss.udel.edu/research/graphiteonline.html>

在线生成碳纳米管结构: <http://turin.nss.udel.edu/research/tubegonline.html>

ADIT 蛋白质检查工具: <http://deposit.rcsb.org/adit>

简介: 可以自行上传 pdb 文件, 通过 Validate 操作, 自动通过 PROCHECK 程序绘制出各种图表用以检查蛋白质。包括 ramachandran 图, chi1-chi2 图, 主链/侧链信息图 (解析度标准偏差、不合理接触等)、二级结构图、扭转角分布、键长距离分布、侧链上平面结构偏差分布、主链键长键角扭曲情况。

webPIPSA(ProteinInteractionPropertySimilarityAnalysis): <http://pipsa.embl.org/pipsa>

简介: 上传 pdb/pqr 或输入 pdbID, 自动调用 APBS/UHBD 计算它们的静电势, 根据一定规则绘制成距离矩阵, 并做簇分析。可以分析不同蛋白质在相互作用上的相似性。可以下载到运行期间的中间文件, 包括转化后的 pqr 和计算得到的 grid 格点文件, 可被 vmd 等软件读取。

CASTp: <http://sts-fw.bioengr.uic.edu/castp/calculation.php>

简介: 找出某蛋白所有口袋或孔洞, 并得到它们的容积和表面积, 有助于研究潜在的配体结合位点。

SFoldRate 预 测 蛋 白 质 折 叠 速 率 :

<http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html>

ALOGPS: <http://www.vcclab.org/lab/alogs>

简介: 在线上传分子结构, 计算 LogP、水溶性、PKa、SMILE 字符串等信息。支持分子结构格式十分多。

CORINA: http://www.molecular-networks.com/online_demos/corina_demo.html

简介: 通过 SMILE 字符串得到分子的结构文件

一些有用的晶体学工具: <http://www.cryst.ehu.es>

GETAREA: <http://curie.utmb.edu/getarea.html>

简介: 计算分子 SASA 和溶解能, 服务器不太稳健

DockingServer: <http://www.dockingserver.com/web/>

简介: 在线分子对接, 须注册, 对免费用户功能有限制

GLYCAM 在线构建糖、糖蛋白结构:

http://glycam.ccrcc.uga.edu/ccrc/biombuilder/biomb_index.jsp

MolEdit: <http://159.149.163.21/moedit.htm>

简介: 在线绘制 2D 结构, 自动转化为三维坐标文件, 支持格式很多。

E-Babel: <http://www.vcclab.org/lab/babel>

简介: 相当于在线版的 Babel, 可以支持几十种结构文件格式的转换。注意不要打开浏览器弹出窗口过滤功能。

OpalDashboard: <http://ws.nbcr.net/opal2/GetServicesList.do>

简介: 一大批软件的在线计算工具, 包括 MEME (搜索一组 [dna](#)/蛋白序列中的基序), APBS (解 PB 方程得到静电势分布、溶解自由能), PDB2PQR (往 pdb 格式中添加原子半径信息, 转为 apbs 等软件所需的 pqr 文件), Preparereceptor/GPF (创建 pdbqt、GPF 文件), Autogrid (计算 autodock 所需的格点文件), Autodock (分子对接), FIMO, GLAM2, GLAM, GOMO。

在线版 PDB2PQR: <http://nbcr.sdsc.edu/pdb2pqr>

Prodrgr2.5: http://davapc1.bioch.dundee.ac.uk/cgi-bin/prodrgr_beta

简介: 输入结构或者在线绘制结构, 生成 gromacs 等软件的拓扑文件, 以及加过氢的 pdb、gro、mol 结构文件。支持 gromos87/96 力场, 支持结构优化。

Karlsberg+: <http://agknapp.chemie.fu-berlin.de/karlsberg/index.php>

简介: 基于线性 PB 方程在线计算蛋白质 Pka

PROPKA: <http://propka.ki.ku.dk>

简介: 输入 PDBID 或者上传 pdb 文件, 计算 PKa。输出结果包括每个残基的 Pka, 不同 PH 下的蛋白去折叠化能、最稳定时的 PH 值, 折叠与去折叠时在不同 PH 下所带电荷、等电点 PH、缓冲能力。虽然独立状态的氨基酸的 PKa 是已知的, 但在蛋白中由于受到周围其它氨基酸的影响 PKa 会发生改变, 故此程序有助于正确判断在不同 PH 环境下模拟蛋白质时氨基酸所应处的质子化态。

H++: <http://biophysics.cs.vt.edu/H++>

简介: 通过隐式溶剂(GB/PB)和分子力学模型计算 Pk, 并根据指定 PH 自动将结构质子化

ProBuilder: <http://159.149.163.21/probuilder.htm>

简介: 输入蛋白质序列和预期的二级结构生成蛋白结构文件

PDBsum: <http://www.ebi.ac.uk/pdbsum>

简介: 输入 PDBID 或者序列, 显示蛋白质信息、基本结构特征、结构出处的文献摘要, 可调用 PROCHECK 分析结构, 可在线观看 3D 结构。

PLATINUM: <http://model.nmr.ru/platinum>

简介: 此程序基于分子疏水势的概念。上传受体/配体结构文件后计算, 会显示分子总面积、极性面积、非极性面积的大小。得到的疏水势格点文件可保存也可以在线自动调用 Jmol 观看, 以不同颜色描述分子的疏水/亲水性质, 可以直观了解受、配体不同方式的结合能力。对于脂体系可以显示 2D 疏水图。

MarvinSketch: <http://www.chemaxon.com/marvin/sketch/index.jsp>

简介: 一款功能强大的绘制分子结构、反应式并计算相关性质的软件的在线版, 需要 java 运行环境, 载入较慢。可自行绘制也可以直接通过名字生成结构(edit-importname), 可以直接从模版库/基团库中插入结构(insert-templatelibrary/group), 可以保存结构文件到本地。可以显示周期表(view-periodictable), 获得 smile 字符串(选中分子, edit-saveassmile, 然后随便找个文本框 paste), 显示 3D 结构(view-OpenMarvinView3D/Space), 给出结构的名称(tools-Naming), 得到分子式、分子量、元素组成(tools-Elementalanalysis), 绘制滴定曲线、计算 PKa、等电点、获得指定 PH 环境下的被质子化/去质子化后的结构(tools-Protonation), 计算 LogP、LogD(tools-partitioning), 计算原子电荷、极化率、轨道电负性(tools-charge), 获得互变异构体、立体异构体(tools-isomers), 计算各个异构体的能量、做简单分子动力学(tools-conformation), 结构拓扑分析、优化并计算能量、计算 SASA(tools-geometry), 显示氢键供体/受体原子数目、

Huckel 分析、计算折射率、显示共振结构、获得结构框架(tools-other)。在程序下方文本框中可以使用 Chemicalterm 语言编写表达式来通过性质筛选分子、计算属性。亦可免费下载此软件的单机版。

MarvinSpace: <http://www.chemaxon.com/marvinspace/applet.html>

简介: 在线的基于 java 的分子可视化程序, 使用 Opengl 库, 支持 pdb、mol 和 cub 格点文件。可用 NewCartoon 等方式显示蛋白质骨架结构, 可以用几种方式显示分子表面, 并在上面用颜色显示包括静电势在内的几种信息。缺点是程序载入很慢。

REDS(RESPEPchargeDeriveServer): <http://q4md-forcefieldtools.org/REDS>

简介: 在线计算 RESP 电荷, 注册十分麻烦。

计算 RRKM 反应速率: <http://phd.marginean.net/rrkm.html>

DynDom: <http://fizz.cmp.uea.ac.uk/dyndom/runDescription.jsp>

简介: 输入蛋白质分子的两个构象, 可分析出构象变化所绕着的旋转轴, 以及导致结构变化的关键残基。结果可以用 rasmol 程序显示。

StrucTools: <http://helixweb.nih.gov/structbio/basic.html>

简介: 可以绘制蛋白质二级序列图, 计算主链氢键, 绘制 B 因子-残基图, 计算残基所占体积并绘图, 计算残基 SASA, 做 Ramachandran 图, 做蛋白质旋转的 gif/mpeg 动画。

TarFisDock(TargetFishingDock): <http://www.dddc.ac.cn/tarfisdock>

简介: 反向对接程序, 提供小分子结构, 寻找受体蛋白。国内可能需要国外代理才能访问, 速度比较慢。

VRMLFileCreator: <http://cactus.nci.nih.gov/vrmlcreator>

简介: 通过 smile 字符串或者结构文件, 创建 VRML(.wrl)文件。比如可以导入 Acrobat3D, 在 pdf 文档中演示分子的立体结构。

w3DNA: <http://w3dna.rutgers.edu>

简介: 输入核酸 PDBID 或上传结构文件, 分析其碱基结构参数。

WebMO: <http://www.webmo.net/demo/index.html>

简介: 提供了友好的 GUI 界面, 可以在线绘制或导入结构并调用服务器上的 Gaussian、Gamess、Mopac、Molpro、NWChem、QChem、Tinker 程序进行量化/分子力学计算。对于免费用户 WebMO 提供了 guest 帐户登入, 但运算时间限制在 60 秒以内, 在缺乏计算条件下可以应急使用。

估算 REMD 模拟适宜的温度设定: <http://folding.bmc.uu.se/remd>

在线蛋白质分析工具列表: <http://www.bioinf.org.uk/servers>

PBTProfiler: <http://www.pbtprofiler.net>

简介: 输入化合物代码或在线绘制结构, 快速预测此化合物对环境污染的情况, 包括在各种环境下的半衰期和分布状况、生物累积性、毒性等。

WHATIF: <http://swift.cmbi.ru.nl/servers/html/index.html>

简介: 提供了十分丰富的蛋白质模拟相关工具。包括同源建模、检查和修复蛋白结构、残基突变、蛋白结构分析、可及表面计算、分析氢键、补全质子、原子间不正当接触检测、计算盐桥、生成 FlexDock 输入文件等等。

8. 学生课程学习要求

8.1 学生自学的要求

学生上课前，需对课本进行预习。预习时可参考本大纲的内容进行快速阅读以及中国大学 MOOC 网中《生物信息学》相关内容进行预习。课后，学生需对课堂上重点强调的内容进行复习，以及老师给的练习题进行练习，以达到熟练掌握理论知识的目的。

8.2 课外阅读的要求

课外，对于参考教材中的内容，特别是课堂上进行重点强调、补充的内容可通过查阅相关的书籍，或者通过网络（如中国大学 MOOC、中国知网、万方等）以及学习运用 AI 工具（DeepSeek、Kimi、豆包等）进行相关知识的延伸阅读和了解，以达到扩充知识面的目的。

8.3 课堂讨论的要求

对老师提出的讨论题目结合所学知识、自身经验等进行认真思考，积极参与，踊跃发言。在整个讨论过程中，教师不得限制学生的发言，可适当地进行点拨，从而达到最大限度地调动学生学习本门课程积极性，启发学生的思考能力的目的。

8.4 课程实践的要求

按照课程的安排要求，学生需准时参加，不得无故迟到、早退甚至旷课，认真完成课程相关的专题汇报和实验工作。对于专题汇报，需先进行大组讨论，确定总的中心思想和具体实施途径后，再查阅文献具体实施。对于实践课程，在实践操作前需对项目进行认真预习，了解其原理和基本的分析操作过程；在操作过程中需积极思考，认真动手，对于操作过程中遇到的不确定因素应先查阅相关资料或向老师提问，不能肆意揣测。

9. 课程考核方式及评分规程

9.1 出勤（迟到、早退等）、作业、报告等的要求

对于教师：不得无故调课、停课、迟到和早退，且至少需在每堂课开始前 15-20 分钟到达上课地点，检查多媒体教学设备（腾讯会议）及课件播放情况是否正常，若有问题需及时调整。

对于学生：要求提前至少 5 分钟到达教室，每堂课严格考勤。若无故缺课达到本门课程 1/3 学时的，取消其考试资格，该门课成绩为不合格。

课堂专题操作实践和讨论以班级大作业的形式布置，鼓励学生积极认真地准备，教师需鼓励大家积极发言、点评，并对学生发言过程中错误的知识点和认知进行纠正和解释。

9.2 成绩的构成与评分规则说明

课程考核采用平时考核和期末考核。平时考核成绩由课堂、课后作业和期中作业成绩组成。期中作业成绩可按 1 次平时作业成绩计算或者授课老师根据实际情况自行确定其在平时成绩中的比例。平时考核成绩占课程成绩的 40%，期末作业成绩占课程成绩的 60%。

9.3 考试形式及说明（含补考）

考试形式为考查形式，相关要求按照四川轻化工大学考试相关要求执行。

10. 学术诚信规定

10.1 考试违规与作弊

考试违规和作弊者，按照四川轻化工大学有关规定进行处理。

10.2 杜撰数据、信息等

杜撰数据和信息者，按照四川轻化工大学有关规定，交学校学术委员会讨论处理。

10.3 学术剽窃等

学术剽窃者，按照四川轻化工大学有关规定，交学校学术委员会讨论处理。

11. 课堂规范

11.1 课堂纪律

按照四川轻化工大学关于课堂纪律的要求执行。

教师认真授课，上课时不得接听或拨打电话，不得讲授与课程无关的内容，在整个教学过程中需维持课堂良好的纪律，以保证教学质量。

学生认真听讲，积极踊跃发言，在教师授课时，对于不懂的或有争议的问题，可以随时举手打断老师，进行讨论式的学习和讲解。不得在上课时打闹，吃零食，玩手机，做任何与课程无关的事。

11.2 课堂礼仪

教师和学生的课堂礼仪按照四川轻化工大学关于课堂礼仪的规定执行。总的要求是教师应衣着规范，干净整洁，普通话标准，给人为人师表的形象，如无特殊情况，不得坐着授课；学生同样应衣着整齐，不奇装异服，应具备大学生应有的青春风貌。

12 . 课程资源

12.1 教材与参考书

生物信息学（第四版），陈铭主编，科学出版社

12.2 主要文献资料及相关数据库

《基因组学》、《生物信息学》、维普、万方和中国知网；ScienceDirect、Wiley、Springer、NCBI 等数据库中有关生物信息学的期刊和论文。

12.3 课程网站等支持条件

学习通；中国大学 MOOC：生物信息学；主要文献资料或相关数据库主要包括 NCBI (<https://www.ncbi.nlm.nih.gov>)、EBI (<https://www.ebi.ac.uk>)、EMBLnet (<http://www.emblnet.org>)、国家基因库生命大数据平台 (China National GeneBank DataBase, CNGBdb, <https://db.cngb.org/>) 等。

13. 教学合约

13.1 教师的师德师风承诺

为了更好地贯彻国家的相关规定，履行教师的职业道德，塑造良好的教师形象，我承诺在整个教学过程中将始终遵守《教师职业道德规范》，教书育人，爱

岗敬业；认真执行《中国教育改革和发展纲要》及《教师法》等有关法律法规；积极参加教改实验和科研，探索更好的教育教学方法；关爱学生，尊重学生，理解和亲近学生，不对学生进行体罚，杜绝任何有损学生身心健康的行为；自觉遵守学校各项规章制度和工作纪律，以德立身。

13.2 阅读课程实施大纲，理解其内容

学生应认真阅读课程实施大纲，如有异议或建议，可以向授课教师提出，教师根据实际情况做修改和调整；如无异议，则视为同意遵守课程实施大纲当中所确定的责任与义务。

13.3 同意遵守课程实施大纲中阐述的标准和期望

课程实施大纲编写完成后旨在提高教学规范和效率，学生需按照达到本课程实施大纲所要求的标准进行学习。

14. 其他说明

无