生物信息学

简明教程

第一章 概论

当前人类基因组研究已进入一个重要时期，2000 年将获得人类基因组的全部序列，这是基因组研究的转折点和关键时刻，意味着人类基因组的研究将全面进入信息提取和数据分析阶段，即生物信息学发挥重要作用的阶段。到 1999 年 12 月 15 日发布的第 115 版为止，GenBank 中的 DNA 碱基数目已达 46 亿 5 千万，DNA 序列数目达到 535 万；其中 EST 序列超过 339 万条； UniGene 的数目已达到 7 万个；已有 25 个模式生物的完整基因组被测序完成，另外的 70 个模式生物基因组正在测序当中；到 2000 年 1 月 28 日为止，人类基因组已有 16%的序列完成测定，另外 37.7%的序列已经初步完成；同时功能基因组和蛋白质组的大量数据已开始涌现。如何分析这些数据，从中获得生物结构、功能的相关信息是基因组研究取得成果的决定性步骤。

生物信息学是在此背景下发展起来的综合运用生物学、数学、物理学、信息科学以及计算机科学等诸多学科的理论方法的崭新交叉学科。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重要组成部分，成为生命科学研究的前沿。

近来的研究表明，基因组不仅是基因的简单排列，它有其特有的组织结构和信息结构，这种结构是在长期的演化过程中产生的，也是基因发挥其功能所必须的。弄清楚生物体基因组特有的组织结构和信息结构，解译生命的遗传语言的关键。

目前在数据库中已经有越来越多的模式生物全基因组序列，第一个人类染色体全序列--第 22 号染色体的测序工作已经在 1999 年 12 月完成，整个人类基因组计划工作草图将在最近完成。这无疑给基因组组织结构和信息结构的研究工作提供了大量的第一手材料，同时也为基因组研究取得突破性进展提供了可能。人类对基因的认识，将从以往的对单个基因的了解，上升到在整个基因组水平上考察基因的组织结构和信息结构，考察基因之间在位置、结构和功能上的相互关系。

从目前生物信息学的研究情况来看，国际上公认的生物信息学的研究内容，大致包括以下几个方面：

1. 生物信息的收集、存储、管理与提供。包括建立国际基本生物信息库和生物信息传输的国际联网系统；建立生物信息数据质量的评估与检测系统；生物信息的在线服务；生物信息可视化和专家系统。

2. 基因组序列信息的提取和分析。包括基因的发现与鉴定，如利用国际 EST 数据库 (dbEST) 和各自实验室测定的相应数据，经过大规模 并行计算发现新基因和新 SNPs 以及各种功能位点；基因组中非编码区的信息结构分析，提出理论模型，阐明该区域的重要生物学功能；进行模式生物完整基因组的信息结构分析和比较研究；利用生物信息研究遗传密码起源、基因组结构的演化、基因组空间结构与 DNA 折叠的关系以及基因组信息与生物进化关系等生物学的重大问题。
3. 功能基因组相关信息分析。包括与大规模基因表达谱分析相关的算法、软件研究，基因表达调控网络的研究；与基因组信息相关的核酸、蛋白质空间结构的预测和模拟，以及蛋白质功能预测的研究。
4. 生物大分子结构模拟和药物设计。包括 RNA(核糖核酸)的结构模拟和反义 RNA 的分子设计；蛋白质空间结构模拟和分子设计；具有不同功能域的复合蛋白质以及连接肽的设计；生物活性分子的电子结构计算和设计；纳米生物材料的模拟与设计；基于酶和功能蛋白质结构、细胞表面受体结构的药物设计；基于 DNA 结构的药物设计等。
5. 生物信息分析的技术与方法研究。包括发展有效的能支持大尺度作图与测序需要的软件、数据库以及若干数据库工具，诸如电子网络等远程通讯工具；改进现有的理论分析方法，如统计方法、模式识别方法、隐马尔科夫过程方法、分维方法、神经网络方法、复杂性分析方法、密码学方法、多序列比较方法等；创建一切适用于基因组信息分析的新方法、新技术。包括引入复杂系统分析技术、信息系统分析技术等；建立严格的多序列比较方法；发展与应用密码学方法以及其他算法和分析技术，用于解释基因组的信息，探索 DNA 序列及其空间结构信息的新表征；发展研究基因组完整信息结构和信息网络的研究方法等；发展生物大分子空间结构模拟、电子结构模拟和药物设计的新方法与新技术。
6. 应用与发展研究。汇集与疾病相关的人类基因信息，发展患者样品序列信息检测技术和基于序列信息选择表达载体、引物的技术，建立与动植物良种繁育相关的数据库以及与大分子设计和药物设计相关的数据库。

利用生物信息学方法进行结构功能预测要注意的是同一问题采用不同算法，可能产生相同或不同的结果。因此，必要弄清楚某种方法的基本原理，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。因此，本章采用原理和实用方法并重的原则进行介绍。因生物信息学覆盖面广，限于篇幅，本章并未将生物信息学的全部内容详细加以讲述，仅针对与目前分子生物学实验数据分析密切相关的生物信息学策略及实用工具进行扼要介绍，文中涉及问题的更详细信息可参考相关网站。生物信息学是新兴发展中的学科，该领域的研究日新月异，书中的描述可能滞后于生物信息学的最新发展为在所难免，作者期望本章的介绍对读者的研究工作有所助益。

第二章 生物信息数据库与查询

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，如人类基因组图谱库 GDB、转录因子和结合位点库 TRANSFAC、蛋白质结构家族分类库 SCOP 等等。

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。

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。到 1999 年 8 月，Genbank 中收集的序列数量达到 460 万条，34 亿个碱基，而且数据增长的速度还在不断加快。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 55,000 个物种，其中 56%是人类的基因组序列(所有序列中的 34%是人类的 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 16 类，其中 EST 数据等又被各自分成若干个文件。

(1)Genbank 数据检索

NCBI 的数据库检索查询系统是 Entrez。Entrez 是基于 Web 界面的综合生物信息数据库检索系统。利用 Entrez 系统，用户不仅可以方便地检索 Genbank 的核酸数据，还可以检索来自 Genbank 和其它数据库的蛋白质序列数据、基因组图谱数据、来自分子模型数据库(MMDB)的蛋白质三维结构数据、种群序列数据集、以及由 PubMed 获得 Medline 的文献数据。

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。更详细的 Entrez 使用说明可以在该主页上获得。

(2)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。

NCBI 的网址是：http://www.ncbi.nlm.nih.gov。

Entrez 的网址是：http://www.ncbi.nlm.nih.gov/entrez/。

BankIt 的网址是：http://www.ncbi.nlm.nih.gov/BankIt。

Sequin 的相关网址是：http://www.ncbi.nlm.nih.gov/Sequin/。

2. 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，也可以用 Sequin 软件来完成。

数据库网址是：http://www.ebi.ac.uk/embl/。

SRS 的网址是：http://srs.ebi.ac.uk/。

WEBIN 的网址是：http://www.ebi.ac.uk/embl/Submission/webin.html。

3. 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。

DDBJ 的网址是：http://www.ddbj.nig.ac.jp/。

4. GDB

基因组数据库(GDB)为人类基因组计划(HGP)保存和处理基因组图谱数据。GDB 的目标是构建关于人类基因组的百科全书，除了构建基因组图谱之外，还开发了描述序列水平的基因组内容的方法，包括序列变异和其它对功能和表型的描述。目前 GDB 中有：人类基因组区域(包括基因、克隆、amplimers PCR 标记、断点 breakpoints、细胞遗传标记 cytogenetic markers、易碎位点 fragile sites、EST 序列、综合区域 syndromic regions、contigs 和重复序列)；人类基因组图谱(包括细胞遗传图谱、连接图谱、放射性杂交图谱、content contig 图谱和综合图谱等)；人类基因组内的变异(包括突变和多态性，加上等位基因频率数据)。GDB 数据库以对象模型来保存数据，提供基于 Web 的数据对象检索服务，用户可以搜索各种类型的对象，并以图形方式观看基因组图谱。

GDB 的网址是：http://www.gdb.org。

GDB 的国内镜像是：http://gdb.pku.edu.cn/gdb/。

2.2 蛋白质数据库

1. PIR 和 PSD

PIR 国际蛋白质序列数据库(PSD)是由蛋白质信息资源(PIR)、慕尼黑蛋白质序列信息中心(MIPS)和日本国际蛋白质序列数据库(JIPID)共同维护的国际上最大的公共蛋白质序列数据库。这是一个全面的、经过注释的、非冗余的蛋白质序列数据库，包含超过 142,000 条蛋白质序列(至 99 年 9 月)，其中包括来自几十个完整基因组的蛋白质序列。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。PSD 的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。每季度都发行一次完整的数据库，每周可以得到更新部分。

PSD 数据库有几个辅助数据库，如基于超家族的非冗余库等。PIR 提供三类序列搜索服务：基于文本的交互式检索；标准的序列相似性搜索，包括 BLAST、FASTA 等；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、结构域搜索 GeneFIND 等。

PIR 和 PSD 的网址是：http://pir.georgetown.edu/。

数据库下载地址是：ftp://nbrfa.georgetown.edu/pir/。

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。

SWISS-PROT 的网址是：http://www.ebi.ac.uk/swissprot/。

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。

PROSITE 的网址是：http://www.expasy.ch/prosite/。

4. 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。

RCSB 的 PDB 数据库网址是：http://www.rcsb.org/pdb/。

5. SCOP

蛋白质结构分类(SCOP)数据库详细描述了已知的蛋白质结构之间的关系。分类基于若干层次：家族，描述相近的进化关系；超家族，描述远源的进化关系；折叠子(fold)，描述空间几何结构的关系；折叠类，所有折叠子被归于全 α、全 β、α/β、α＋β 和多结构域等几个大类。SCOP 还提供一个非冗余的 ASTRAIL 序列库，这个库通常被用来评估各种序列比对算法。此外，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，可以找到与未知结构序列远缘的已知结构序列。

SCOP 的网址是：http://scop.mrc-lmb.cam.ac.uk/scop/。

6. COG

蛋白质直系同源簇(COGs)数据库是对细菌、藻类和真核生物的 21 个完整基因组的编码蛋白，根据系统进化关系分类构建而成。COG 库对于预测单个蛋白质的功能

和整个新基因组中蛋白质的功能都很有用。利用 COGNITOR 程序，可以把某个蛋白质与所有 COGs 中的蛋白质进行比对，并把它归入适当的 COG 簇。COG 库提供了对 COG 分类数据的检索和查询，基于 Web 的 COGNITOR 服务，系统进化模式的查询服务等。

COG 库的网址是：http://www.ncbi.nlm.nih.gov/COG。

下载 COG 库和 COGNITOR 程序在：ftp://ncbi.nlm.nih.gov/pub/COG。

2.3 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。

KEGG 的网址是：http://www.genome.ad.jp/kegg/。

2. DIP

相互作用的蛋白质数据库(DIP)收集了由实验验证的蛋白质－蛋白质相互作用。数据库包括蛋白质的信息、相互作用的信息和检测相互作用的实验技术三个部分。用户可以根据蛋白质、生物物种、蛋白质超家族、关键词、实验技术或引用文献来查询 DIP 数据库。

DIP 的网址是：http://dip.doe-mbi.ucla.edu/。

3. ASDB

可变剪接数据库(ASDB)包括蛋白质库和核酸库两部分。ASDB(蛋白质)部分来源于 SWISS-PROT 蛋白质序列库，通过选取有可变剪接注释的序列，搜索相关可变剪接的序列，经过序列比对、筛选和分类构建而成。ASDB(核酸)部分来自 Genbank 中提及和注释的可变剪接的完整基因构成。数据库提供了方便的搜索服务。

ASDB 的网址是：http://cbcg.nersc.gov/asdb。

4. TRRD

转录调控区数据库(TRRD)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、静默子、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(包括所有注释涉及的参考文献)。TRRD 主页提供了对这几个数据表的检索服务。

TRRD 的网址是：http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4/。

5. TRANSFAC

TRANSFAC 数据库是关于转录因子、它们在基因组上的结合位点和与 DNA 结合的 profiles 的数据库。由 SITE、GENE、FACTOR、CLASS、MATRIX、CELLS、METHOD 和 REFERENCE 等数据表构成。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的突变的转录因子和结合位点；S/MART DB 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的网络；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。TRANSFAC 及其相关数据库可以免费下载，也可以通过 Web 进行检索和查询。

TRANSFAC 的网址是：http://transfac.gbf.de/TRANSFAC/。

2.4 其它数据库资源

1. DBCat

DBCat 是生物信息数据库的目录数据库，它收集了 500 多个生物信息学数据库的信息，并根据它们的应用领域进行了分类。包括 DNA、RNA、蛋白质、基因组、图谱、蛋白质结构、文献著作等基本类型。数据库可以免费下载或在网络上检索查询。

DBCat 的网址是：http://www.infobiogen.fr/services/dbcat/。

下载 DBCat 在：ftp://ftp.infobiogen.fr/pub/db/dbcat。

2. PubMed

PubMed 是 NCBI 维护的文献引用数据库，提供对 MEDLINE、Pre-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。

PubMed 的网址是：http://www.ncbi.nlm.nih.gov/。

除了以上提及的数据之外，还有许许多多的专门生物信息数据库，涉及了目前生物学研究的各个层面和领域，由于篇幅所限无法一一详述。国内也有一些大数据库的镜像站点和自己开发的有特色的数据库，如欧洲分子生物学网络组织 EMBNet 中国节点北京大学分子生物信息镜像系统，上海博容基因公司与上海嘉瑞软件公司合作开发的国产汉化基因数据库及分析管理系统，同时国家级的生物信息学中心也在筹建之中。我们期待国内能有更多高质量和使用便利的数据库资源，推动我国生物信息学和整个生命科学的发展。

清华大学生物信息学研究所网址：http://bioinfo.tsinghua.edu.cn

北京大学生物信息镜像系统网址：http://cbi.pku.edu.cn

第三章　序列比对和数据库搜索

比较是科学研究中最常见的方法，通过将研究对象相互比较来寻找对象可能具备的特性。在生物信息学研究中，比对是最常用和最经典的研究手段。

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，通过比较两个序列之间的相似区域和保守性位点，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共同的保守区域、位点和 profile，从而探索导致它们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。

比对还是数据库搜索算法的基础，将查询序列与整个数据库]的所有序列进行比对，从数据库中获得与其最相似序列的已有的数据，能最快速的获得有关查询序列的大量有价值的参考信息，对于进一步分析其结构和功能都会有很大的帮助。近年来随着生物信息学数据大量积累和生物学知识的整理，通过比对方法可以有效地分析和预测一些新发现基因的功能。

3.1 序列两两比对

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，但由于蛋白质具有的模块性质，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。一般的处理方法是用两个罚分值，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 FASTA 工具和 BLAST 工具。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 为了提供序列搜索的速度，会先建立序列片段的"字典"，查询序列先会在字典里搜索可能的匹配序列，字典中的序列长度由 ktup 参数控制，缺省的 ktup=2。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。

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比对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。

表 1. BLAST 程序：

程序	数据库	查 询	简 述
blastp	蛋白质	蛋白质	可能找到具有远源进化关系的匹配序列
blastn	核酸	核苷酸	适合寻找分值较高的匹配，不适合远源关系
blastx	蛋白质	核酸(翻译)	适合新 DNA 序列和 EST 序列的分析
tblastn	核苷酸(翻译)	蛋白质	适合寻找数据库中尚未标注的编码区
tblastx	核酸(翻译)	核酸(翻译)	适合分析 EST 序列

表 2. BLAST 的蛋白质数据库：

数据库	简 述

nr	汇集了 SWISS-PROT,PIR,PRF 以及从 GenBank 序列编码区中得到的
month	蛋白质和 PDB 中拥有原子坐标的蛋白质，并去除了冗余的序列
swissprot	nr 中过去 30 天内的最新序列
pdb	SWISS-PROT 数据库
yeast	PDB 结构数据库中的蛋白质序列
E.coli	酵母基因组中编码的全部蛋白质
Kabat	大肠杆菌基因组中编码的全部蛋白质
alu	Kabat 的免疫学相关蛋白质序列
	由 REPBASE 中的 Alu 重复序列翻译而来，用来遮蔽查询序列中的
	重复片段

表 3. BLAST 的核酸数据库：

数据库	简 述
nr	非冗余的 GenBank＋EMBL＋DDBJ＋PDB 序列，除了 EST、STS、
month	GSS 和 0,1,2 阶段的 HTGS 序列
dbest	nr 中过去 30 天的最新序列
dbsts	非冗余的 Genbank＋EMBL＋DDBJ＋PDB 的 EST 部分
htgs	非冗余的 Genbank＋EMBL＋DDBJ＋PDB 的 STS 部分
yeast	0,1,2 阶段的高产量基因组序列(3 阶段完成的 HTG 序列在 nr 库里)
E.coli	

pdb	酵母的全基因组序列
kabat	大肠杆菌的全基因组序列
vector	由三维结构库来的核酸序列
mito	Kabat 的免疫学相关序列库
alu	Genbank 的载体子集
gss	线粒体核酸序列
	REPBASE 中 Alu 重复序列翻译而来，用来遮蔽查询序列中的重复片段
	基因组勘测序列(Genome Survey Sequence)

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。

BLAST 的当前版本是 2.0，它的新发展是位点特异性反复 BLAST(PSI-BLAST)。PSI-BLAST 的特色是每次用 profile 搜索数据库后再利用搜索的结果重新构建 profile，然后用新的 profile 再次搜索数据库，如此反复直至没有新的结果产生为止。PSI-BLAST 先用带空位的 BLAST 搜索数据库，将获得的序列通过多序列比对来构建第一个 profile。PSI-BLAST 自然地拓展了 BLAST 方法，能寻找蛋白质序列中的隐含模式，有研究表明这种方法可以有效的找到很多序列差异较大而结构功能相似的相关蛋白，甚至可以与一些结构比对方法，如 threading 相媲美。PSI-BLAST 服务可以在 NCBI 的 BLAST 主页上找到，还可以从 NCBI 的 FTP 服务器上下载 PSI-BLAST 的独立程序。

NCBI 的 BLUST 网址是：http://www.ncbi.nlm.nih.gov/BLAST/。

下载 BLUST 的网址是：ftp://ncbi.nlm.nih.gov/blast/。

下载 FASTA 的网址是：ftp://ftp.virginia.edu/pub/fasta/。

3.2 多序列比对

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。

目前使用最广泛的多序列比对程序是 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、PHYLIP 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/。

第四章 核酸与蛋白质结构和功能的预测分析

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。

4.1 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有多个证据的支持。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 TATA Box 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 CENSOR 和 RepeatMasker 等，可以在 Web 界面上使用这些程序，或者用 Email 来进行。如果有大量序列需要处理，可以使用 XBLAST 程序，它可以从 Internet 上下载得到。XBLAST 中以及

包含了由程序作者收集整理的一些重复序列，此外还可以从 Repbase 中找到更多的重复序列。还可以把克隆载体也加入重复序列中，这样就可以在处理重复序列时顺便把克隆载体也一同除去。经处理的序列中重复序列所在位置会一律由“X”代替。

CENSOR 和 Repbase 的网址是：http://www.girinst.org/。

CENSOR 的 Email 服务地址是：censor@sharon.lpi.org。

RepeatMasker 的网址是：http://ftp.genome.washington.edu/cgi-bin/RepeatMasker。

下载 XBLAST 的网址是：ftp://ncbi.nlm.nih.gov/pub/jmc。

下载 Repbase 的网址是：ftp://ncbi/nlm.nih.gov/repository/repbase/REF。

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。

GRAIL 的网址是：http://compbio.ornl.gov/Grail-1.3/。

4. 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。

5. 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。

常见的基因识别工具很多都包含了剪接位点识别功能，独立的剪接位点识别工具有 NetGene 等。

NetGene 服务的 Email 地址是：netgene@cbs.dtu.dk。

6. 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，可以参考下面的综述。

Gelfand, M. S. (1995). Prediction of function in DNA sequence analyis. J. Comput. Biol. 2, 87-115.

7. 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。

8. 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，可以通过 Web 页面或 Email 获得 GENSCAN 服务。

GENSCAN 的网址是：http://ccr-081.mit.edu/GENSCAN.html。

9. tRNA 基因识别

tRNA 基因识别比编码蛋白质的基因识别简单，目前基本已经解决了用理论方法预测 tRNA 基因的问题。tRNAscan-SE 工具中综合了多个识别和分析程序，通过分析启动子元件的保守序列模式、tRNA 二级结构的分析、转录控制元件分析和除去绝大多数假阳性的筛选过程，据称能识别 99%的真 tRNA 基因。可以在 Web 上使用这个工具，也可以下载这个程序。

tRNAscan-SE 的网址是：http://www.genetics.wustl.edu/eddy/tRNAscan-SE/。

4.2 针对蛋白质的预测方法

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。

1. 从氨基酸组成辨识蛋白质

根据组成蛋白质的 20 种氨基酸的物理和化学性质可以分析电泳等实验中的未知蛋白质，也可以分析已知蛋白质的物化性质。ExPASy 工具包中提供了一系列相应程序：

AACompIdent：根据氨基酸组成辨识蛋白质。这个程序需要的信息包括：氨基酸组成、蛋白质的名称(在结果中有用)、pI 和 Mw(如果已知)以及它们的估算误差、所属物种或物种种类或“全部(ALL)”、标准蛋白的氨基酸组成、标准蛋白的 SWISS-PROT 编号、用户的 Email 地址等，其中一些信息可以没有。这个程序在 SWISS-PROT 和(或)TrEMBL 数据库中搜索组成相似蛋白。

AACompSim：与前者类似，但比较在 SWISS-PROT 条目之间进行。这个程序可以用于发现蛋白质之间较弱的相似关系。

除了 ExPASy 中的工具外，PROPSEARCH 也提供基于氨基酸组成的蛋白质辨识功能。程序作者用 144 种不同的物化性质来分析蛋白质，包括分子量、巨大残基的含量、平均疏水性、平均电荷等，把查询序列的这些属性构成的“查询向量”与 SWISS-PROT 和 PIR 中预先计算好的各个已知蛋白质的属性向量进行比较。这个工具能有效的发现同一蛋白质家族的成员。可以通过 Web 使用这个工具，用户只需输入查询序列本身。

ExPASy 的网址是：http://www.expasy.ch/tools/。

PROSEARCH 的网址是：http://www.embl-heidelberg.de/prs.html。

2. 预测蛋白质的物理性质

从蛋白质序列出发，可以预测出蛋白质的许多物理性质，包括等电点、分子量、酶切特性、疏水性、电荷分布等。相关工具有：

Compute pI/MW：是 ExPASy 工具包中的程序，计算蛋白质的等电点和分子量。对于碱性蛋白质，计算出的等电点可能不准确。

PeptideMass：是 ExPASy 工具包中的程序，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。

TGREASE：是 FASTA 工具包中的程序，分析蛋白质序列的疏水性。这个程序延序列计算每个残基位点的移动平均疏水性，并给出疏水性-序列曲线，用这个程序可以发现膜蛋白的跨膜区和高疏水性区的明显相关性。

SAPS：蛋白质序列统计分析，对提交的序列给出大量全面的分析数据，包括氨基酸组成统计、电荷分布分析、电荷聚集区域、高度疏水区域、跨膜区段等等。

ExPASy 的网址是：http://www.expasy.ch/tools/。

下载 FASTA 的网址是：ftp://ftp.virginia.edu/pub/fasta/。

SAPS 的网址是：http://www.isrec.isb-sib.ch/software/SAPS_form.html。

3. 蛋白质二级结构预测

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、α+β 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。

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。

PredictProtein：提供了序列搜索和结构预测服务。它先在 SWISS-PROT 中搜索相似序列，用 MaxHom 算法构建多序列比对的 profile，再在数据库中搜索相似的 profile，然后用一套 PHD 程序来预测相应的结构特征，包括二级结构。返回的结果包含大量预测过程中产生的信息，还包含每个残基位点的预测可信度。这个方法的平均预测准确率达到 72%。

SOPMA：带比对的自优化预测方法，将几种独立二级结构预测方法汇集成"一致预测结果"，采用的二级结构预测方法包括 GOR 方法、Levin 同源预测方法、双重预测方法、PHD 方法和 SOPMA 方法。多种方法的综合应用平均效果比单个方法更好。

nnPredict 的网址是：http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html。

PredictProtein 的网址是：http://cubic.bioc.columbia.edu/predictprotein/。

PredictProtein 的国内镜像在：http://www.cbi.pku.edu.cn/predictprotein/。

SOPMA 的网址是：http://pbil.ibcp.fr/。

4. 其它特殊局部结构

其它特殊局部结构包括膜蛋白的跨膜螺旋、信号肽、卷曲螺旋(Coiled Coils)等，具有明显的序列特征和结构特征，也可以用计算方法加以预测。

COILS：卷曲螺旋预测方法，将序列与已知的平行双链卷曲螺旋数据库进行比较，得到相似性得分，并据此算出序列形成卷曲螺旋的概率。

TMpred：预测蛋白质的跨膜区段和在膜上的取向，它根据来自 SWISS-PROT 的跨膜蛋白数据库 Tmbase，利用跨膜结构区段的数量、位置以及侧翼信息，通过加权打分进行预测。

SignalP：预测蛋白质序列中信号肽的剪切位点。

COILS 的网址是：http://www.ch.embnet.org/software/COILS_form.html。

TMpred 的网址是：http://www.ch.embnet.org/software/TMPRED_form.html。

SignalP 的网址是：http://www.cbs.dtu.dk/services/SignalP/。

5. 蛋白质的三维结构

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。

SWISS-MODEL：自动蛋白质同源模建服务器，有两个工作模式：第一步模式(First Approach mode)和优化模式(Optimise mode)。程序先把提交的序列在 ExPdb 晶体图像数据库中搜索相似性足够高的同源序列，建立最初的原子模型，再对这个模型进行优化产生预测的结构模型。

CPHmodels：也是利用神经网络进行同源模建预测蛋白质结构的方法。

SWISS-MODEL 的网址是：http://www.expasy.ch/swissmod/SWISS-MODEL.html。

CPHmodels 的网址是：http://www.cbs.dtu.dk/services/CPHmodels/。

第五章　分子进化

分子进化钟的发现与中性理论的提出，极大地推动了进化尤其是分子进化研究，填补了人们对分子进化即微观进化认识上的空白，推动进化论的研究进入分子水平，并建立了一套依赖于核酸、蛋白质序列信息的理论方法。分子进化研究有助于进一步阐明物种进化的分子基础，探索基因起源机制，从基因进化的角度研究基因序列与功能的关系。

5.1 分子进化钟与中性理论

60 年代早期“分子进化钟”的发现与 60 年代末期“中性理论”的提出是本世纪进化学的重大事件，是古老的进化学与新生的分子生物学两者“杂交”的产物。它们的相继问世极大地推动了进化尤其是分子进化研究，填补了人们对分子进化即微观进化认识上的空白，并在生物医学等领域产生了广泛影响。

随着不同生物来源的大量蛋白质序列的确定，Zucherkandl 等发现：某一蛋白在不同物种间的取代数与所研究物种间的分歧时间接近正线性关系，进而将分子水平的这种恒速变异称为“分子钟”。

支持进化钟存在的证据来自哺乳动物与其它脊椎动物诸如血清白蛋与转铁蛋白等的免疫学（如微量补体固定）定量比较。人们发现多肽间的免疫距离（如抗原性）与其氨基酸取代百分数成良好的线性相关，如鸟溶菌酶、哺乳动物 RNase、细胞色素 C 与白蛋白、大肠杆菌色氨酸合成酶等。虽然这种相关性的分子基础尚不清楚，但这种客观存在经过反复验证后是不容置疑的。免抗血清由此成为初步估算球形单体蛋白间序列差异的有效工具，但其适用范围 0-30%的氨基酸差异。

自从进化钟假设提出之后，存在许多反驳它的相反事实与异议。这些异议主要针对序列进化的恒速。分子进化钟的最明显的例外之一是分子序列证据与化石证据在人类起源时间上的差异。60 年代中期，许多人类学家认为人类在 3000 年前与我们最近的亲属-- 非洲猿分歧。根据分子钟假设，分歧 3000 万年的物种氨基酸序列差异的应达 4-5%、非重复序列 DNA 差异应约为 8%，但实测值分别为 0.8%与 1.1%。对这种 6 倍左右的差别有两种解释。许多人类学家倾向于怀疑钟的存在，并认为在高等灵长类中分子进化速率下降。总之，虽然大部分分子进化学家同意序列进化与分歧时间密相关，但进化是以年限还是以代限为刻度则仍有分歧与争议；而且因为纵多因素的影响，与进化钟相左的数据，无论是用氨基酸、核苷酸序列差异、免疫学距离，还是用 DNA 杂交复性等参数，均不断有所报道，其论争预计将继续下去。

. 自从 60 年代初发现分子进化钟--“分子进化速率在不同种系中恒定”以来，人们又陆续发现蛋白质中氨基酸的置换是随机而非模式性的；DNA 在哺乳动物种系的总变异速率远远高于形态上的变异速率并远远超出人们的预期的大于 0.5 核苷酸/ 基因组/ 年；蛋白质电泳表明物种内存在大量的变异即广泛的种内多态性，且这些多态性并无可见的表型效应，与环境条件亦无明显相关。以上这些都是新达尔文主义与综合进化理论所难以解释的。

面对上述问题，日本群体遗传学家木村资生（Motoo Kimura）提出：(1) 进化过程中的核苷酸置换其绝大部分是中性或近似中性的突变随机固定的结果而不是正向达尔文选择的结果：(2) 许多蛋白质多态性必须在选择上为中性或近中性，并在群体中由突变引入与随机灭绝间两者的平衡维持。

上述论著问世遭遇到经典进化学家的强烈批判。他们认为新的分子生物学数据完全可以用新达尔文主义的原理来解释。直至现在，选择论者与中性论者的议争仍在继续。这两大学派的本质区别可通过它们各自对突变基因如何在物种内置换老基因这一进化过程的不同解释来洞悉。每一置换刚出现时在群体内均为稀有的突变等位基因，随后扩散至个群体并被固定，即频率达 100%。选择论者认为：一个突变的等位基因在物种内扩散，就必需具有某些选择上的优势，如在选择上为中性，就必需与一选择上具优势的基因紧密连锁，通过“搭车”而达到较高频率。与此相反，中性

论者认为：一些突变在没有任何选择优势的情况下也能自身在群体中扩散。如果一突变体在选择上等同于已存在的等位基因，其命运将取决于机会-随机，其频率存在上下起伏，因为在每代每一雌、雄个体所生的大量配子中只有很少数配子最终被“采用”以形成合子以及相应的个体，并出现在下一代中。在这种随机漂变（random drift）中，绝大部分突变等位基因随机丢失，但有一少部分在群体中被固定下来。如果中性突变在分子水平上普遍存在，且随机漂变在很长时间（如百万年）一直延续，群体的遗传组成将发生显著性改变。群体中出现的任何中性突变其最终固定的概率都等于其原始频率，其固定的平均时间四倍于有效群体的大小（它近似等于每一代参与繁殖的个体数，通常远小于物种的个体总数）。中性理论并非认为中性基因无功能，而仅是认为不同的等位基因在促进个体的生存与生殖方面是有等同的效果。此外，还需强调个体基因突变与群体基因置换的差别，因为只有后者才与分子进化相关。

自 Zuckerkandl 与 Pauling 的早期工作以来，已经知道在蛋白质进化中结构和化学性质上相似的氨基酸间的替换比不相似间的替换更为频繁。他们认为，这种“保守的”的替换看来只造成分子功能的微小改变，因而更容易“被自然选择接受”。同时他们指出，关于氨基酸残基的最重要性质是什么，“化学家和生物学家间显然没有同样的见解”。从中性学说的立场看，保守替换的性质，只需注意到两种氨基酸间的差异越小，它们等于选择等价而不是突变有害的概率就越大，就很容易加以解释。因此，选择上呈中性的替换在得类似的氨基酸间则概率越高，而这类氨基酸的进化替换由于随机遗传漂变则出现得更为频繁。

在阐明分子进化中突变型替换的保守性的同时，有越来越多的证据表明，功能上较不重要的分子或某一分子较不重要的部分，其进化（以突变型替换表示）比那些较重要的要快些。中性论和选择论间的差别，在它们对快速进化的分子（如血纤蛋白肽）或分子的某部分（如胰岛素原的 C 肽）进行解释时，可以最清楚地看出，按中性学说解释，它们在功能上不重要，因而大多数突变是中性的，突变通过随机漂变而迅速积累。另一方，选择论的解释是，快速进化的分子或分子的某部分或许有某些尚不知道的功能，并且通过积累许多由正达尔文选择产生的较微有利的突变，而经历了迅速的适应性方面的改善。这两种解释那一种更为恰当还有待积累更多数据以后才能判定。为了加深我们对分子进化机制的理解，很有必要研究突变型替换的模式与分子的三级结构和功能的相互关系。

综上，中性学说（或者更确切地说是中性突变-随机漂变假说）是分子生物学与群体遗传学交融的产物。它不象传统的综合理论（或新达尔文派的观点），它明确主张：进化中大多数突变型的置换，不是由于正达尔文选择，而是由选择上呈中性或近中性的突变型的随机固定所致。它还断言，分子水平上大多数种内遗传多态性，象以蛋白质多态性形式展现出来的那样，是选择上呈中性或近中性的，并靠着突变输入和等位基因的随机清除或固定这两者之间的平衡而在物种中维持。应该说，这一理论对于人们所认识的分子进化众多现象与规律的阐释比新达尔文更为科学，且提出的多项预测被随后的实验研究所证实。问题是，它作为一种更基本层次－分子

水平的进化理论未能给更高层次的进化提供理性阐释与描写。中性论者过多地注目于与功能无关的分子进化，而忽视了与功能相关的分子进化现象与规律的探索，这恐怕是中性理论之所以能问世，但同时又先天性地带上无视宏观进化，对宏观进化束手无策这一天然缺陷的症结所在。

5.2 进化树

分子钟的发现对于进化研究具有十分重要的意义。它不仅能用于粗略估计不同类群生物间的进化时间，亦可用于构建进化树。实际上，分子钟发现不久，蛋白质序列分析即被广泛用于生物的长时进化研究。

根据蛋白质的序列或结构差异关系可构建分子进化树(evolutionary tree)或种系发生树(phylogenetic tree)。进化树给出分支层次或拓扑图形，它是产生新的基因复制或享有共同祖先的生物体的歧异点的一种反映，树枝的长度反映当这些事件发生时就存在的蛋白质与现在的蛋白质之间的进化距离。根据进化树不仅可以研究从单细胞有机体到多细胞有机体的生物进化过程，而且可以粗略估计现存的各类种属生物的分歧时间。通过蛋白质的分子进化树分析，为从分子水平研究物种进化提供了新的手段，可以比较精确的确定某物种的进化地位。对于物种分类问题，蛋白质的分子进化树亦可作为一个重要的依据。

构建进化树的方法包括两种：一类是序列类似性比较，主要是基于氨基酸相对突变率矩阵（常用 PAM250）计算不同序列差异性积分作为它们的差异性量度（序列进化树）；另一类在难以通过序列比较构建序列进化树的情况下，通过蛋白质结构比较包括刚体结构叠合和多结构特征比较等方法建立结构进化树。

序列进化树

构建序列进化树的主要步骤是比对，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。

距离方阵方法简单的计算两个序列的差异数量。这个数量被看作进化距离，而其准确大小依赖于进化模型的选择。然后运行一个聚类算法，从最相似（也就是说，两者之间的距离最短）的序列开始，通过距离值方阵计算出实际的进化树，或者通过将总的树枝长度最小化而优化出进化树。用最大节约方法搜索进化树的原理是要求用最小的改变来解释所要研究的分类群之间的观察到的差异。最大似然方法评估所选定的进化模型能够产生实际观察到的数据的可能性。进化模型可能只是简单地假定所有核苷酸（或者氨基酸）之间相互转变的概率一样。程序会把所有可能的核苷酸轮流置于进化树的内部节点上，并且计算每一个这样的序列产生实际数据的可能性（如果两个姐妹分类群都有核苷酸“A”，那么，如果假定原先的核苷酸是“C”，得到现在的“A”的可能性比起假定原先就是“A”的可能性要小得多）。所有可能的再现（不仅仅是比较可能的再现）的几率被加总，产生一个特定位点的似然值，然后这个数据集的所有比对位点的似然值的加和就是整个进化树的似然值。

4・进化树搜索

单一的进化树的数量会随着分类群数量的增长而呈指数增长，从而变为一个天文数字。由于计算能力的限制，现在一般只允许对很小一部分的可能的进化树进行搜索。具体的数目主要依赖于分类群的数量、优化标准、参数设定、数据结构、计算机硬件以及计算机软件。

有两种搜索方法保证可以找到最优化的进化树：穷举法和树枝&#0;跳跃法（BB）。对于一个很大的数据集，这两种方法都很不实用。对分类群数量的限制主要取决于数据结构和计算机速度，但是对于超过 20 个分类群的数据集，BB 方法

很少会得到应用。穷举法要根据优化标准，对每一个可能的进化树进行评估。BB 方法提供一个逻辑方法，以确定那些进化树值得评估，而另一些进化树可被简单屏蔽。因此 BB 方法通常要比穷举法快得多。

绝大多数分析方法都使用“启发式”的搜索。启发式现搜索出相近的次优化的进化树家族（“岛屿”），然后从中得到优化解（“山顶”）。不同的算法用不同程度的精确性搜索这些岛屿和山顶。最彻底也是最慢的程序（TBR，tree bisection-reconnection，进化树对分重接）先把进化树在每一个内部树枝处劈开，然后以任意方式将劈开的碎片重新组合起来。最快的算法只是检查一下相邻终端的不太重要的重新组合，因此倾向于找到最近的岛屿的山顶。

降低搜索代价的最好方法是对数据集进行剪除。影响优化搜索策略选择的因素（数据量，数据结构，时间量，硬件，分析目的）太复杂，无法推荐一个简单可行的处方。因此进行搜索的用户必须对数据非常熟悉且有明确的目标，了解各种各样的搜索程序及自己硬件设备和软件的能力。

除上述当前应用最广的方法外，还有大量的建立和搜索进化树的其它方法。这些方法包括 Wagner 距离方法和亲近方法（距离转化方法）；Lake 的不变式方法（一个基于特征符的方法，它选择的拓扑结构包含一个意义重大的正数以支持颠换）；Hadamard 结合方法（一个精细的代数方阵方法，对距离数据或者观察到的特征符进行修正）；裂解方法（这个方法决定在数据中应该支持哪一个基于距离的可选的拓扑结构）；四重奏迷惑（Quartet puzzling）方法可以为 ML 建树方法所应用，这个算法相对而言是个较快的进化树搜索算法。

5・确定树根

上述的建树方法所产生的都是无根树（进化树没有进化的极性）。为了评估进化假说，通常必须要确定进化树的树根。确定系统发育进化树的树根并不简单问题。一种确定树根的好方法就是分析时加入一个复制的基因。如果来自绝大多数物种或者所有物种的所有的平行基因在分析时都被包含进去，那么从逻辑上我们就可以把进化树的树根定位于平行基因进化树的交汇处，当然要假定在所有进化树中都没有长树枝问题。

6・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。

5.3 结构进化树

随着 X－ray、NMR 等实验技术的的进步，蛋白质结构数据的数量日益增多，结构精度也越来越高，使得结构比较更为可行。目前已经发现许多蛋白的一级序列差异很大，难以通过序列比对进行分子进化的研究，但它们的空间拓扑结构仍然很相似，可以进行结构叠合比较、分析它们之间的进化关系，这表明结构比较可以比序列比较获得更多更精确的结构信息。研究发现蛋白质结构比序列的保守性更强，进化过程中蛋白质序列可能发生变化，但它的折叠模式更为保守，即使是 70%的序列发生变化，它的折叠模式也不会有很大的改变[1]。蛋白质分子的结构比较与蛋白质一级序列比较法相比，具有更高的优越性。

目前有关蛋白质结构比较的研究方法很多，主要有刚体结构叠合比较、多特征的结构比较等方法。前者用比较后确定的拓扑等价位点的个数或等价位点 C_α 原子距离的均方根值作为不同结构间差异性的量度（结构进化树）；后者用蛋白质结构的多项特征如残基的物理特性、残基的空间倾向性、主侧链的方向、主链的二面角、二级结构类型和主侧链的可接近性等综合指标作为结构的差异性量度，有时称此类方法构建的结构进化树为“类结构”进化树。

刚体叠合所构建的进化树适用于同源蛋白质结构预测的骨架结构的选择，基于序列的进化树便于描述类似性较大的蛋白质的进化关系，而结构的多特征比较则适用于分析分歧较大的蛋白质结构。

1．刚体结构叠合比较

当已知 2 个以上同源蛋白质的晶体结构时，可将每两套结构的原子坐标进行最佳叠合，确定类似结构中完整的一套残基等价位点，并使等价位点间的距离平方和最小，这样便得到各结构的拓扑等价区。通常将结构简化为一系列 C_α 位置，等价位点被定义为在重叠结构中位于某个特定距离范围（不大于 3 埃）之内的 C_α 原子。通过计算不同结构等价位点的个数或计算多个结构的等价位点 C_α 距离的均方根值作为不同结构间差异性的度量。再根据一般的建树方法，给出几个结构的进化关系。

刚体结构叠合方法需要蛋白质的晶体结构数据的质量要高。事实上，相对于序列而言，已测定的蛋白质晶体结构很少，许多同源蛋白质的结构并不知道。其次，尽管同源蛋白质具有相同的折叠结构，它们的二级结构成分则经历形变、相对平移和旋转使侧链达到优化的包装以适应进化的压力。对于序列相同率为 30%的两个蛋白质，由刚体叠合所确定的拓扑等残基的均方根差大约为 1.5 埃，而且残基数可能只占全部残基数的 1/3。它可能不足以进行结构比较。因此需要发展一种更灵活的确定拓扑等价位点的方法，并且要把二级结构成分的相对运动和变形也考虑进去。

2．多特征结构比较

多特征结构比较以及构建“类结构”进化树的原理与基于残基匹配记分方法（常用PAM250 矩阵）进行多序列比较和构建序列进化树的原理相同。包括以下步骤：（1）动态规划配准和结构匹配；（2）多个结构的多特征比较；（3）多特征结构比较的距离量度；（4）绘制进化树图。

3．相关软件

Phylip

PHYLIP 是一个包含了大约 30 个程序的软件包，这些程序基本上囊括了系统发育的所有方面。PHYLIP 是免费软件，并且可以在很多平台上运行（Mac, DOS, Unix, VAX/VMS, 及其它）。PHYLIP 目前已经是最广泛使用的系统发育程序。

PAUP

开发 PAUP 的目的是为系统发育分析提供一个简单的，带有菜单界面的，与平台无关的，拥有多种功能（包括进化树图）的程序。PAUP 3.0 只建立于 MP 相关的进化树及其分析功能；而 PAUP 4.0 已经可以针对核苷酸数据进行与距离方法和 ML 方法相关的分析功能，以及其它一些特色。

除了 PAUP 和 PHYLIP 以外，还有其它一些系统发育程序，这些程序包括 FastDNAml, MACCLADE, MEGA plus METREE, MOLPHY 和 PAML。

PHYLOGENETIC RESOURCES

http://www.ucmp.berkeley.edu/subway/phylogen.html

PHYLOGENY PROGRAMS

http://evolution.genetics.washington.edu/phylip/software.html

PHYLOGENETIC ANALYSIS COMPUTER PROGRAMS

http://phylogeny.arizona.edu/tree/programs/programs.html

BIOCATALOG MOLECULAR EVOLUTION
http://www.ebi.ac.uk:/biocat/phylogeny.html

PHYLIP http://evolution.genetics.washington.edu/phylip.html

第六章 基因组序列信息分析

DNA 序列自身编码特征的分析是基因组信息学研究的基础，特别是随着大规模测序的日益增加，它的每一个环节都与信息分析紧密相关。从测序仪的光密度采样与分析、碱基读出、载体标识与去除、拼接、填补序列间隙、到重复序列标识、读框预测和基因标注的每一步都是紧密依赖基因组信息学的软件和数据库。特别是拼接和填补序列间隙更需要把实验设计和信息分析时刻联系在一起。

基因组不仅是基因的简单排列，更重要的是它有其特有的组织结构和信息结构，这种结构是在长期的演化过程中产生的，也是基因发挥其功能所必须的。利用国际 EST 数据库 (dbEST) 和各实验室测定的相应数据，经过大规模并行计算识别并预测新基因，新 SNPs 以及各种功能位点，如剪接与可变剪接位点等。

到 1998 年底在人类的约 10 万个基因中有 3 万多个已被发现，尚有约 7 万个未被发现。由于新基因带来的显著经济效益和社会效益，它们成为了各国科学家当前争夺的热点。EST 序列 (Expressed Sequence Tags) 到 1999 年 12 月已搜集了约 200 万条，它大约覆盖了人类基因的 90%，因此如何利用这些信息发现新基因成了近几年的重要研究课题。同时 1998 年国际上又开展了以 EST 为主发现新 SNPs 的研究。因此利用 EST 数据库发现新基因、新 SNPs 以及各种功能位点是近几年的重要研究方向。

虽然对约占人类基因组 95%的非编码区的作用人们还不清楚，但从生物进化的观点看来，这部分序列必定具有重要的生物功能。普遍的认识是，它们与基因在四维时空的表达调控有关。寻找这些区域的编码特征，信息调节与表达规律是未来相当长时间内的热点，是取得重要成果的源泉。

在不同物种、不同进化水平的生物的相关基因之间进行比较分析，是基因研究的重要手段。目前，模式生物全基因组序列数据越来越多，因此，基因的比较研究，也必须从基因的比较，上升到对不同进化水平的生物在全基因组水平上的比较研究。这样的研究将更有效地揭示基因在生命系统中的地位和作用，解释整个生命系统的组成和作用方式。

6.1 基因组序列分析工具

1. Wisconsin 软件包（GCG）

Genetics Computer Group 公司开发的 Wisconsin 软件包，是一组综合性的序列分析程序，使用公用的核酸和蛋白质数据库。SeqLab 是其图形用户界面（GUI），通过它可以使用所有 Wisconsin 软件包中的程序及其支持的数据库。此外，它还提供

了一个环境用于创建、显示、编辑和注释序列。SeqLab 也可以被扩展使其可以包括其它公用或非公用的程序和数据库。

Wisconsin 软件包由 120 多个独立的程序组成，每个程序进行一项单一的分析任务。包括所有程序的完整目录以及详细的描述可以在 Wisconsin 软件包的程序使用文档中找到。GCG 支持两种核酸数据库(GenBank 数据库, 简化版的 EMBL 核酸序列数据库)和三种蛋白质数据库(PIR,SWISS-PROT, SP-TrEMBL)。这些数据库既有 GCG 格式的（供大多数 Wisconsin 软件包程序使用），也有 BLAST 格式的（供 BLAST 数据库搜索程序使用）。同时还提供了用于 LookUp 程序以及数据库参考搜索的索引。

关于 GCG，Wisconsin 软件包，支持的平台以及硬件需求的一般性信息可以在 GCG 的主页以及 Wisconsin 软件包的用户手册中找到。GCG 主页提供了更新信息以及 Wisconsin 软件包程序的完整列表。

SeqLab 中可以使用多个序列分析程序的特性使用户可以应用这些程序顺序地回答相关问题或在对输入序列进行编辑后重复某项分析。而可以同时访问公用数据库和本机序列的优点使用户可以在一个分析中使用其中任意一种而不用先进行转换或格式化的工作。SeqLab 可以解决的序列分析问题：

(1)在两条 mRNA 中寻找开放阅读框架，翻译并对比 RNA 与蛋白质序列

对两条相关的 mRNA 进行测序的用户可能希望寻找开放阅读框架（ORF）、翻译以及进行核酸与氨基酸序列间的两两对比。

把序列加入 SeqLab Editor 中，从 Functions 菜单中选中 Map 选项运行 Map 程序。Map 输出文件包含了限制性酶切图和 6 种可能的翻译框架的 ORF 的显示。这些 ORF 的起始和终止位置可进行标记并选为 SeqLab Editor 中序列显示的范围，然后可用 Edit 菜单的 Translate 操作进行翻译。翻译结果自动出现在 SeqLab Editor 中。

两条相关的核酸或蛋白质序列可用 Gap 程序或 BestFit 程序进行对比。Gap 程序寻找两条序列间的全局最优对比结果。适用于两条待比对的序列是进化相关的情况。BestFit 程序寻找两条序列的局部最优对比结果，它适用于两条序列不是进化相关而是功能相关的情况。

(2)通过参考搜索寻找数据库中的相关条目并进行对比

研究一个特征序列家族成员的用户可能希望寻找这个家族中的其它成员并建立它们的多序列对比。

从 Functions 菜单中选取 LookUp 程序。LookUp 在数据库条目的参考信息部分搜索描述词并建立匹配条目的列表。在参考部分的 Definiton, Author, Keyword 和

Organism 域中搜索描述词并在词之间使用“and”（&）、“or”（|）以及“but not”（！）布尔表达式。例如，在 SWISS-PROT 条目的 Description 域搜索“lactate & dehydrogenase & h & chain”将产生一个输出文件，其中列出了乳酸脱氢酶 H 链（lactate dehydrogenase H chain）条目。这个输出文件可以从 Output Manager 窗口中加以显示，然后与用户的序列一起添加到 SeqLab Editor 中。

要创建所有这些序列的多序列对比，只要根据序列名称选中这些序列并从 Functions 菜单中运行 PileUp 程序。由 PileUp 产生的多序列文件也列在 Output Manager 窗口中并可以直接添加到 SeqLab Editor 中。推荐采用这一步的原因在于数据库条目的特征表格（Features table）信息可与对比结果一起被包括进来。必要时对比结果是可以被编辑的，并且如果数据库条目有相似的特征，这些特征可被附加给用户序列。

(3)用查询序列搜索数据库，将找到的条目与查询序列进行对比并产生进化系统树

克隆并测序一个未知功能基因的用户可能希望在一个数据库中搜索相似的序列。如果搜索到了，用户可能进一步希望创建与查询序列最相似的序列的多序列对比并产生数据的种系图。

往 SeqLab Editor 中添加一个查询序列并从 Functions 菜单中选取 FASTA 程序。FASTA 程序在数据库中搜索与查询序列相似的序列。输出文件可从 Output Manager 窗口中加以显示并直接添加到 SeqLab Editor 中。在这个输出文件中数据库条目与查询序列局部相似性最好的区域被加以标记。如果要显示的话，每个数据库条目只有这种区域可以显示在 SeqLab Editor 中。不要的条目可以从 SeqLab Editor 中一起被删除。

从 Functions 菜单中选中 PileUp 程序创建这些序列的多序列对比。输出可从 Output Manager 窗口中加以显示并添加到 SeqLab Editor 中更新已经存在的未对比序列。必要时可对这一对比结果进行编辑，并且数据库条目的有用的特征表格信息也可以添加给查询序列。

从 Functions 菜单中选取 PaupSearch 程序，程序提供了一个 PAUP（进化系统简约性分析（Phylogenetic Analysis Using Parsimony））中树搜索方式的 GCG 接口。PaupDisplay 程序为 PAUP 中的树操作，鉴定以及显示方式提供了一个 GCG 接口。

(4)拼接交叠序列片段产生一连续序列，寻找并翻译这一序列的编码区域并在数据库中搜索相似序列

克隆了一个基因，把它分解克隆为一组有交叠的序列片段并进行了测序的用户可能希望把这些序列片段重新组装为一条连续的序列。一旦 contig 拼接完成，用户可能希望在序列中寻找阅读框架，翻译并在数据库中搜索相似序列。

Fragment Assmbly System 的程序可用于拼接交叠序列片段。GelStart 程序创建一个项目。GelEnter 程序把序列片段复制到项目中。GelMerge 程序寻找片段之间的交叠并把它们拼接成 contig。GelAssemble 程序是一个编辑器，可用于编辑这些连续的部分并解决片段之间的冲突问题。所有这些程序都可以从 Functions 菜单中选取。一旦拼接完成，最终构成此 contig 的连续序列可以被保存为一个序列文件并添加到 SeqLab Editor 中。

使用 Map、Frames、TestCode 或 Codon Preference 程序可预测序列中的编码区（所有这些程序可以从 Functions 菜单中选中）。使用 Edit 菜单的 Select Range 功能选择这些程序预测的区域并使用 Edit 菜单中的翻译操作把它们翻译为蛋白质。这些提出的翻译区域也可以作为核酸共有序列的特征被加入。

选取蛋白质序列然后选择 Functions 菜单中 BLAST。BLAST 程序在数据库中搜索与查询序列相似的条目，此程序既可以进行远程搜索也可以进行本机搜索。搜索结果可以从 Output Manager 窗口中加以显示。如果被搜索的是一个本机的数据库，结果文件可以加入 SeqLab Editor 或 Main List 窗口中，并允许对找到的序列进行进一步分析。

(5)对比相关的蛋白质序列，计算对比结果的共有序列，辨识序列中新的特征序列模式，在数据库中搜索包含此模式的序列或在对比结果的共有序列中搜索已知的蛋白质模式

辨识了一组相关序列的用户可能希望对其进行对比并计算对比结果的共有序列。如果可以在对比结果中找到保守模式，用户可能希望在数据库中搜索包含这种模式的其它序列。用户可能还希望在计算出的共有序列搜索已知的蛋白质模式。

选取待对比的序列，从 Functions 菜单中选取 PileUp 程序创建多序列对比，PileUp 程序的输出文件可从 Output Manager 窗口中加以显示并添加到 SeqLab Editor 中。用户可以对对比结果的某个区域重新加以对比并以此替换原有的对比结果。只要选取一个区域并重新运行 PileUp 即可。从 PileUp Options 窗口中选取"realign a portion of an existing alignment（重新对比一个已存在的对比结果的一部分）"，这可能有利于选择一个替代评分矩阵或不同的创建和扩展处罚。新的输出文件将包含最初的对比结果以及替换原始对比结果的重新对比的区域。

用 Edit 菜单中 Consensus 操作计算对比结果的共有序列。如果保守模式可被辨识，从 Functions 菜单中选取 FindPatterns 选项。从共有序列中剪切下此特征序列模式并把它粘贴到 FindPatterns 模式选择器中，并在数据库中搜索包含这一模式的序列。

此外，运行 Motif 程序可在共有序列中搜索已知的蛋白质模式。Motif 在蛋白质序列中搜索在 PROSITE，蛋白质位点和模式的 PROSITE 字典中已知的蛋白质模式。如果辨识出一个 Motif，则给所有序列增加一个特征，并标出它的位置。图 4.9 显示了一个蛋白质序列的匹配、一个共有序列以及 Motif 搜索的结果。

(6)使用 Profile 进行相似性搜索并对比相关序列

序列分析的一个新的扩展领域是 Profile 技术。一个 profile 是一个位置特定的评分矩阵，它包含了一个序列对比结果中每个位置的所有残基信息。这一点与共有序列不同，共有序列中只包含每个位置的保守残基的信息。Profile 做好后可用于搜索数据库、数据库划分或在一个集合中搜索与原始对比结果中的序列相似的序列。它也可以用于把一条单独的序列与一个对比结果进行对比。

使用 ProfileMake 程序可创建一个序列对比结果的 profile。使用 ProfileSearch 程序可用 profile 对数据库进行搜索，ProfileSegment 程序可以显示搜索结果。使用 ProfileGap 程序可将一个序列与 profile 进行对比。ProfileMake, ProfileSearch, ProfileSegments 以及 ProfileGap 程序都可以从 Functions 菜单中启动。

GCG 的主页 http://www.gcg.com

2. ACEDB

ACEDB 是一种被广泛应用的管理和提供基因组数据的工具组,适用于许多动物和植物的基因组计划。该软件是免费的，并且可运行在 Unix 和 Macintosh OS 系统下，Windows 版本马上就会推出。数据库以丰富的图形界面提供信息，包括有具体显示的基因图谱，物理图谱，新陈代谢的途径和序列等。数据用流行的对象的形式进行组织，使用大家熟悉的类别如，相关的文献，基因，描述，和克隆的 DNA 等。可用于专用的数据分析以及许多永久性数据的采集,而且使用者不需要经过专门的计算机和数据库的训练就可以使用 ACEDB。对于资源有限的计划，这往往是决定使用 ACEDB 的关键因素。

3・其它工具

不同的基因组测序中心都有其特有的一套序列管理分析方案及工具，并且在不断发展完善之中，具体细节可访问这些测序中心的网站了解。

6.2 人类和鼠类公共物理图谱数据库的使用

1・物理图谱的类型

物理图谱有许多结构和形式。限制性图谱（restriction map），用于对小区域、如 kb 量级做精细结构制图，细胞遗传学图（cytogenetic map），用于对以 10^4 kb 为长度量级的区域制图。最常用的两种类型是 STS 含量图（STS content map）和放射性杂交图（radiation hybrid map），它们的分辨区域都大于 1Mb，并且有能使用简易 PCR 中的定位标记物的优点。

在 STS 含量图中，STS 标记物通过多聚酶链反应所监测，在反应中它与一个大的插入克隆基因库反应，如酵母人工染色体（TACs），细菌人工染色体（BACs）和粘粒等。如果两个或多个 STS 被发现是存在于同一个克隆之中，那么这些标记位点紧密相邻的机会就很高（不是 100%，因为在制图过程中存在一些假象，如出现嵌合克隆体）。一段时期以来，根据 STS 含量图已经建立起一系列重叠群，如含有 STS 的重叠簇克隆。这样一张图的分辨率和覆盖度由一些因子决定，如 STS 的密度、克隆群体的大小、以及克隆文库的深度。通常 STS 含量图以长 1Mb 的插入 YAC 库为基础，分辨率为几百个 bp。如果使用插入部分较小的克隆载体，图谱就会有一个更高的理论分辨率，但是覆盖基因组同样大小面积就需要更多的 STS。虽然一般有可能从 STS 含量图上得到标记物的相对顺序，但是相邻标记物之间的距离还是无法精确测得。尽管如此，STS 含量图还是有与克隆原相关的优点，并且可将其用于更进一步的研究，如次级克隆或 DNA 测序。到目前为止，STS 含量图制图简单而使用最多的来源是巴黎的 CEPH（centre d&#0;Etudes du Polymorphisme Humain）中的 YAC 库。它是一个 10×覆盖率的文库，平均插入长度为~1Mb。

放射性杂交图（对片段 DNA 的断点作图。在此技术中，一个人体细胞系被致死性的 gamma 射线照射，染色体 DNA 分成片段。然后该细胞系与一个仓鼠细胞系融合而被救，并能繁殖几代。在这期间，人类细胞和仓鼠细胞的杂合体随机丢失其人类染色体片段。这样一百个或更多的杂合细胞系克隆体中，每一个都有不同数量的染色体片段，筛选生长后，就可以形成一套杂合组，供接下来的制图实验用了。

如果要在一个放射性杂交组中对一个 STS 作图，那就要将每种杂交组细胞系中的 DNA 进行 STS 的 PCR 操作。细胞系中如果含有该 STS 的染色体片段，那么就能得到一个正的 PCR 信号。在基因组中相邻很近的 STS 有相似的固位模式（retention pattern），因为放射性引起的断点落在它们中间的几率很小。相邻较远的 STS 固位模式相似性降低，相邻很远的 STS 的固位模式将会截然不同。与基因图谱所用方法类似，算法类的软件也能推出 STS 在放射性杂交图上的相对顺序，并通过断点落在其中间的可能性，用某一距离系统计算相邻标记物之间的距离。放射性杂交图还能提供一个标记物位于某一个特殊位点的可能值（优势对数值）。一个放射性杂交图的分辨率依赖于杂交体片断的大小，而这又依赖于人体细胞系所受的辐射量。一般对基因组大小作图的细胞系分辨率为～1M。

除 STS 含量图和放射性杂交图外还有几个方法可用于制作人类物理图谱。克隆图谱使用与 STS 含量图不同的技术来决定克隆体的接近程度。例如，CEPH YAC 图谱法综合利用指纹法（fingerprinting）、间－Alu 产物杂交法（inter-Alu product hybridization）和 STS 含量图法来制作一张重叠的 YAC 克隆体图谱。缺失和体细胞杂交图依赖于大型基因组重组（可以人工引进或由实验本身引起），从而将标记物放在由染色体断点所限定的 bin?中。FISH 图谱使用一个荧光信号来探测克隆体的间期 DNA 扩散时的杂交情况，从而以细胞遗传学图中一条带的位置定出克隆体的位置。

研究者捕捉致病基因时对转录序列图谱有特别的兴趣。这些序列是由已表达序列，和那些从已转化成 STS 并置于传统物理图谱的已知基因衍生而来的。近来一些制作大量 EST 的工程已经使制图实验室能够得到数以万计的单一表达序列。一旦一个致病位点被鉴定出来后，这些转录序列图谱就能明显加快对目标基因的研究速度。

YAC 库可用于 STS 的排序，但其克隆体中的高嵌合率和高删除率使它们不能用于 DNA 测序。去年高分辨率、可用于测序的质粒和 BAC 图谱则发展很快。因为它们所需的克隆工艺水平很低。除了几个特例，如染色体 19 的 Lawrence Livemore 实验室质粒图外，其它图谱都还只处在初级阶段。

2.大型公用数据库中的基因组图谱

人类基因组物理图谱信息的主要来源是由 NCBI 和 GDB 提供的大型公用数据库。这些数据库提供各种图谱的来源，使研究者能够用一个多用户界面交互系统在图谱中进行比较。在一定程度下，这些数据库还能进行图谱的综合及分析。

（1）NCBI Entrez 的染色体图谱

Entrez 的基因组部分是最容易获得物理图谱信息的来源之一。此服务由 NCBI 所提供。Entrez 试图以一种可理解的方式将几种遗传学图谱和物理图谱、DNA 和蛋白序列信息、以及一个目录型引用数据库和三维晶体结构信息融合起来。因为它的内部连接多，而且界面简单，Entrez 可作为搜索图谱的一个起始点。

除人类基因组，Entrez 还提供关于鼠类、果蝇、C.*elegans*、酵母以及一些原生动物的图谱。尽管可比较的（同线性）图仍不可获得，但它代表了现在最大和最完整的一套多生物体的图谱信息。

（2）GDB 的浏览染色体图谱

另一种常见的人类物理图谱数据的来源是 GDB。尽管 GDB 是基于当时基因图谱的重要性才构建起来的，但是最近几年来，GDB 也已经进行了扩建重组，现在同样可以算是物理图谱数据的仓库。不象 NCBI，GDB 只限于人类图谱数据。它不含序列数据，也没有其它种类生物的信息。同 NCBI 一样，GDB 可以由 WWW 上得到。GDB 提供了一种全功能的对其数据库的查询式界面。

（3）来自个体来源的基因组图谱

尽管一级数据库，如 Entrez 和 GDB 是已发表的图谱的重要来源，但是它们还没有能替代原始数据的东西。有能力制作自己的物理图谱的实验室一般都有自己的网址，连向它们的图谱数据库。通过从这一渠道直接获取资料，我们可以看到制图实验室所使用的图的形式、下载原始数据、并且了解实验室制图时的协议。另外，一

些图在出现于 Entrez 和 GDB 前经常被丢掉。Entrez 和 GDB 数据库选择的表达方式，对那些希望将新的标记物定位于已知物理图谱上的研究者来说，只提供了最小的帮助。

基因组的基因图谱

基因图谱是制作许多物理图谱时工作的基本骨架，也是许多制图项目的起点。有两种基因组范围的基因图谱可供选择。Genethon 图含 5264 个多样性微卫星重复片断，间隔 1.6cM。完整的数据库文件，以及图谱的 PostScript 方式图形表示，在 Genethon 的 FTP 站点上均可获得，这些图通过 GDB 也可以获得。

第二大基因图谱由人类连锁合作中心（Cooperative Human Linkage Center）制造，CHLC 图由 10775 个标记物组成，大多数为微卫星重复片断，间隔 3.7cM。

人类基因组的转录物图

在 1996 年 10 月，*Horno sapiens* 的一个全基因组转录物图由一个国际合作的研究实验室发表于 *Science* 上。这个图由～15000 个不同的表达序列组成，由放射性杂交法定位，与 Genethon 基因图谱衍生的框架相近。通过对酵母人工染色体作 STS 含量法又增添了 1000 个表达序列。在这张图中，大约 1/5 的标记物有已知的或是假定的功能，而余下的代表了未知功能的表达序列。制成图的序列一般由 UniGeneset 衍生而来，它是一个由 NCBI 管理的公用重复 ESTs 数据库。

转录物图是通过将八家不同实验室的图谱数据综合而得到的。为协调制图方法的些微不同，表达序列被放在由 Genethon 基因图谱衍生的框架上。结果，该图的最大分辨率为～2cM。很多情况下，可以从各个实验室的数据库里得到针对某一部分数据更好的制图信息，特别是 the Whitehead Institute 和 Stanford University 的。

浏览 NCBI 转录物图

转录物图可在两个网址上得到。数据的"亲本"站点为 NCBI。在那儿可以找到含有全基因组转录物图的 *Science* 文章的全文，以及彩色的图象，但一般都只有装饰性的墙面图案。另外，也有搜索页可以让浏览者对特别感兴趣的基因进行查询，或是通过对功能未知，但其读码框与某已知功能的蛋白质相近的表达序列图谱进行搜索。

NCBI 网址的一个限制就是它不能在低分辨率标记物分布柱形图上提供转录物图的图形。但是通过 Mapview 微程序就可以得到其图形显示。从 GDB 的首页，沿着 What&#0;s New 的链接，可找到全基因组转录物图（到本书出版时链接形式可能已有所不同）。同样，可以认为转录物图也是 Entrez 网将要制作的一部分。

White head Institute 提供的人类物理图谱

The Whitehead Intitute/MIT Center for Genome Research 是两张基因组范围物理图谱的最初来源。其中一张是 STS 含量图，内含指定为 YAC 的 10000 多个标记物，以及一张含 12000 个左右标记物的放射性杂交图。Whitehead 所用的 G4 杂交板（Genebridge 4 radiation hybrid panel）分辨率为～1Mbp，而以 YAC 为基础作的图分辨率大约为 200kbp。这些图已经和 Genethon 基因图相结合，产生了一张合图，在平均 150kb 范围内有 20000 个 STSs。Whitehead 图上大约有一半的标记物是表达序列，它们在人类转录物图上也会出现。

WI（Whitehead Institute）图可通过网络从 Whitehead Center for Genome Research 的主页上得到。沿着“人类物理图项目”（Human Physical Mapping Project）的链接就可以得到感兴趣的图，这些图可通过几种方法浏览。选择一系列 pop-up 菜单可以产生所选染色体的图，选择选项按钮可以综合放射性杂交图、STS 含量图和基因图。与 Entrez 一样，这些图不是固定不变的。点击一个 STS 或是重叠群，会弹出关于该图素详细信息的页面。图形式图谱在网址上可按 GIF 或 Macintosh 最初模式（PICT）下载。Whitehead 网址上还提供了对图谱数据库进行查询的搜索页。这些搜索数据的链接可按名称、GenBank 通道号、STS 型号、染色体分配进行搜索。另外，Whitehead 网页也可根据功能关键字搜索制图转录序列，并提供与 NCBI 中的主转录物图的链接。

Whitehead 也为那些希望建立他们自己的 STS 的研究者提供服务，并将之放在一个或多个图上，这些服务包括：

一个在线的引物选择程序，引物 3

将一个 STS 放在 STS/YAC 含量图上的服务

将一个 STS 放在放射性杂交图上的服务

Whitehead 图远未完善，对合图进行监督性测试就能显示出在基因图、放射性杂交图和 STS/YAC 图上的 STSs 位置间存在矛盾。这些矛盾表现在合图上仍存在交叉线。解释这些图的一个关键点在于理解这些图在可靠性与分辨率水平不一。基因图骨架在数十兆时能可靠地连接标记物，但在低于约 2 兆时就无法准确解决两个 STS 的顺序问题了。放射性杂交图能够测知约 10Mb 的连接，有效分辨率达～1Mb（更小的间隔也能排序，但是不可靠性逐步增加）。STS/YAC 图可以测知两个相互间隔 1Mb 的 STS 的连接，估计分辨力达 100～300kb。理解图谱时头脑中应有这些尺度上的差异。一般在 1Mb 的范围以下，STS/YAC 图是说明顺序的图谱中最可靠的一种。

在 STS 含量图中，由于 STS 和 YAC 的不等分布，可靠性也会有地域差异。在 YAC 密集的区域（每一个 STS 有 5 个或更多的 YAC），在排序信息的重要性上，图谱结果是相对更可靠的。在低密度区，图谱结果中就会有几种同时可能替代的 STS 顺序，并会附上数据。假定的错误的反面情况，如图 12.8 中，表示为图中的

空白框。这一点也会严重降低图谱的准确性。最后，因为在所有 YAC 库中都存在嵌合现象的问题，双键（例如，一对 STS 同时与 2 个或更多 YAC 连接）比单键（STS 只由 1 个 YAC 连接）更能可靠说明相邻关系。尽管只有在基因图或放射性杂交图中存在支持性数据时，图上才能构建单键信息，但单由两个 STS 相连形成的连接仍保留怀疑。这些元素在任何制图区域被详细检查的时候都应考虑在内。

下面的部分介绍如何在 Whitehead 图上，通过 Whitehead 网址安置新的 STS。从 STS 设计和针对 Whitehead 和放射性杂交图进行制图开始。

设计一个 STS，置于 Whitehead 上

设计一个 STS 需要一个高质量的 DNA 序列，至少长达所需的 PCR 产物。为得到最好的结果，这些序列应不含重复元素和载体序列，并且质量相对高些。任何支持一个 WWW 浏览器的计算机系统都可以使用该程序，支持 TCP/IP 的网络连接也是必须的。

首先，将浏览器连到 Whitehead Genome Center 的主页。寻找并点击指向 WWW Primer Picking 的链接。接着出现一页，在其上方有一个很大的输入框。剪切原始序列并粘贴到该处，只用粘贴原始序列，不需用名称或其它标记词。这些碱基可以小写或大写，而白色空格可以忽略。

现在，向下滚动窗口，将 PCR 的条件调至需要值。那些关于盐浓度、温度和产物大小范围等的默认值均是 WI 所设定的。如果有必要的改变需输入时，按标有 Pick Primers 键返回一套引物处进行特定设定。这些引物现在在对感兴趣的序列的审查实验中用得上。通过放大基因组 DNA 中的一条特定带，可以对这些引物的能力进行经验性鉴定。引物的失败主要与引物扫描区域中的重复元素有关。相反，通过进行 BLAST 或 FASTA 搜索，再选择引物对，来对输入序列中的重复序列进行筛选则是比较明智的，如果 STS 成功地放大了一条特定带，它就可以与 Whitehead STS/TAC 含量图或放射性杂交图相联系，被制成图。

与 Whitehead STS/YAC 含量图联系对 STS 制图

一旦被制出后，一个 STS 就可以通过对 CEPT mega-YAC 库的扫描确定在 STS/YAC 含量图上的位置。而对含有超过 30000 个克隆，其中又有 1200 个排列、板块和柱池（row、plate 和 column pool）的 YAC 库进行搜索，实在是一件头疼的任务。可喜的是，几个生物技术公司已经提供了 CEPH YAC 的复本和（或）筛选系统，包括 Research Genetics Corporation。Whitehead 图就是仅从 YAC 库的后一部分构建起来的。这意味着库模块中位于 709－972 的范围仍需筛选。STS 接着就可以用以下步骤放在图上了。

使浏览器连向 Whitehead 的主页，并点击标有 Human Physical Mapping Project 的链接以跳到该组织的物理制图页。从这儿，再找到并选择“Search for a YAC to its

address”，接着出现一页，内有一系列 pop-up 菜单，能用于输入单个 YAC 的地址、或一个输入单个 YAC 名称的主题栏、或一个能粘贴一列 YAC 地址的大型区域。后者适用于将多个 YAC 用于研究的时候。在这个地方输入 YAC 列表，再使用“plate_row_column”形式，这里是用“_”号分离板块、排和列这三维（如 709_A_1），也可输入多个 YAC 地址，用空格或 carriage 回车隔开。搜索过程输入格式并不固定，它也可识别多个 YAC 模式（包括 709_a_1 和 709a1）。

当 YAC 表完成后，按 Search 键，得到一个表，列有各个 YAC，其重叠群位置和染色体分配，以及附近 STS 的位置。这些 STS 位于放射性杂交图和（或）基因图上。

要理解该搜索结果，应该知道 CEPH 库中相当数量（40－50%）的克隆都是嵌合体，这意味着单个 YAC 可能存在于位于基因组不同部分的重叠群中。由于这个原因，需要找到多个 YAC 来证明单个 STS 分配到了某一特定重叠群中，或是从其它方法来证明（比如 FISH，体细胞杂交制图，放射性杂交图制图数据）。

每张图对应输入的一个 YAC 地址，每个表包括已知 YAC 中的 STS 表，以及 STS 制图信息。对于每个 STS，染色体分配、基因图位置和放射性杂交图位置只要已知就会给出。另外，STS 所属的已命名的重叠群也列成表，这些表中大多数元素是超文字链接，选择合适的链接可以获得关于一个 STS 或一个重叠群更多的信息。由于历史原因，许多 STS 有两个重叠群。双链接重叠群（例如由成对 YAC 共有的重叠群）短一些，在构图的起始阶段中是可创造的更可靠的重叠群，它们可以被放心地忽略。单个重叠群长一些，在不同方式下也应承认其合理性。

Whitehead 放射性杂交图

STS 也能被置于 Whitehead 放射性杂交图中，这比 STS/YAC 含量图的问题简单很多，因为在放射性杂交图上搜索一个 STS 只用 93 次 PCR，而不是 1000 次。Whitehead 放射性杂交图使用 Genebridge 4 radiation hybrid panel。与 CEPH YAC 库一样，这些细胞谱系的 DNA 也可以从一些生物技术公司那儿得到。而有些公司还提供搜索服务。为得到最好的结果，PCR 必须在与制作 Whitehead 图的相同条件下进行，并应在复制时进行。复制 PCR 间出现的不同结果说明应继续重复或以未知物对待。

首先，将杂交模板筛选结果重定为“rhv”格式，看上去如下：

sts_name1
0010010110000010000000110100011011100111001010012110011101010101001010000

sts_name2
0000011110000010000000110100000011100111001010012110011101010101001000000

每个数字代表每个放射性杂交细胞系的 PCR 结果：0 说明 PCR 结果为负（无反应产物），1 说明为正，2 说明为“未知”或“未完成”。载体上数字的顺序是很重要的，必须与 G4rhp 中的正式顺序相对应。为找到该顺序，可沿（Whitehead 物理图页上）标有“How the radiation hybrid maps were constructed”（如何构建放射性杂交图）的链接，再按下标有“G40”的链接。该顺序与它们由 Research Genetics 运输时包装的 DNA 顺序相同，所以它一般还不是结果。要增加可读性，可在载体内加入空格，用一个或多个空格、或 Tab 键就可以将 STS 名称与扫描数据分离开了。

从 Whitehead 物理图页上，按下标有“Place your own STSs on the genome framework map”（将你自己的 STS 放入基因组框架图中）的链接，再输入提示的合适的 Email 地址，并将 PCR 值粘贴至位于该页上的大型主题框。输入正确的 Email 地址很重要，否则制图结果将有可能被误解。

默认时，制图数据会以正文形式返回。为产生放在 Whitehead 图上的 STS 的图形，选择一个标有 Mac PICT（针对 Macintosh 系统）或 GIF（针对 Windows 和 Uinx 系统）的选项按钮。

当设置完成时，按下“提交”键。当数据已被转交或正在制图时，你会得到一个证明，在一小时内结果将会通过 Email 回执给你。

对于大量的筛选数据，如果用剪切和粘贴来向服务器提交这些文件就不太方便了。这时可以将数据以纯文本形式存在用户盘上，然后用 RH 制图页中的浏览键来定义并提交此文件给服务器，同样，Email 地址也要手工输入。

对于～98％的提交的标记物，Whitehead 放射性杂交图制图服务器都会找到特定的位置。如果安置成功，软件将会给一回执，包括该标记物的染色体分布和在染色体连接群中的位置、对标记物的表格式说明、和在 Whitehead 放射性杂交图上两侧标记物的存在时其数据情况。按要求将会得到一张 Macintosh 图或 GIF 格式图。这些图由 Whitehead 框架图组成，所提交 STS 的位置以红色标明。

如果发现标记物连接的染色体多于一个或是根本就没有连接，制图过程也可能失败。在前一种情况中，可以重新提交并设置高优势对数值，这样服务器将会认为其连接一个染色体，在后一种情况中，你可以试着利用放射性杂交图页上的一个 pop-up 菜单将限制性降低。如果一个标记物确实连向多个染色体，那么有可能用 STS 探测出重复序列。

Stanford University 放射性杂交图

Stanford Human Genome Center 已经用 G3 制图板发展了一张基因组放射性杂交图。由于比 G4 板所用放射量更高，G3 板的分辨率更高，但是代价是在探测长距离连接时限制很大。Stanford 图一般在平均 375kb 的范围内存在～8000 个 STS，这些标记物中，3700 个左右是表达序列，存在于 NCBI 转录物图中。同以往一样，

在基因组很多部分中，Stanford 图中的表达序列比“全包容”NCBI 图中的准确性更高。

Stanford 提供一个放射性杂交图制图服务器。如同 Whitehead 服务，这个服务器允许对从 Research Genetics 和其它业主处得到的 G3 板进行 STS 扫描。输入数据，服务器将会尝试将 STS 与 Stanford 图相连，并用 Email 返回结果。因为 G3 板不能探测长距离连接，在无其它图谱信息时，Stanford 服务器只能将 75%的 STS 定位在一条染色体上。但是如果要在可选区域内提供标记物的染色体分布。服务器就能够在一个低优势对数连接值时进行分析，并可对 90%的情况作出分布图谱。

当使用 PCR 时，STS 应对 83G3 板 DNA 扫描。为得到最好的结果，可使用 Stanford 的 RH Protocol 主页给出的 PCR 协议，每次分析结果都应该复制，并且复制品间有分析差异就应该重复或标为未知。

Stanford 服务器返回的制图结果由一系列相应的标记物分布组成。对于每一个 STS，服务器都会报告离其最近的基因标记物、染色体、和标记物到 STS 的距离，以 centiray（cR）为单位。尽管对于制图结果并不提供图形显示，图谱信息还是可以用来与以上讨论的浏览图形结合来说明所提交 STS 相对于 Stanford 图上其它 STS 的位置。

要提交这一数据，连接 Stanford 的主页，并按下 RH 服务器的链接，然后是 RH Server Web Submission。输入 Email 地址和提交号的区域已被说明。Email 地址对于保证收到制图结果是很重要的。提交号是一个可选择栏，它会同结果一起回执给用户，并且用于帮助工作人员使结果组织化。如果 STS 的染色体分布已知，那么应输入到标有 Chromosome Number 的区域。这个信息会增加制图软件测出一个正确连接的能力。

现在，将筛选数据粘到大型正文栏中，并按提交键。制图结果一般在几分钟内通过 Email 回执。Stanford 服务器以一系列相对基因标记物的位置返回制图结果。对于每个 STS，服务器会报告离其最近的基因标记物、其所在染色体和 STS 到标记物的距离（以 centirays 为单位）。尽管并不提供制图结果的图形显示，制图信息仍可用于和以上标出了用户的 STS 相对 Stanford 图谱上的其它 STS 的位置的可浏览型图谱相结合。

CEPH YAC 图

1993 年，巴黎的 CEPH（Centre d&#0;Études du Polymorphisme Humain），与 Genethon 合作，发表了人类基因组的第一张物理图谱。这张图由几套重叠 YAC 组成，形成连接邻近基因标记物的途径。YAC 重叠可由几种技术鉴定，包括 YAC 指纹印迹法（YAC fingerprinting）、与 inter-Alu PCR 结果杂交法、荧光原位杂交（FISH）和 STS 含量图。尽管 YAC 克隆图大部分已被更方便的以 STS 为基础的

图谱替代，对于要包括 CEPH YAC 库或以克隆为基础的反应物的制图项目还是有用的。

由于 YAC 库中的高嵌合率，在两个通过指纹法或 inter-Alu PCR 杂交法确定相互重叠的 YAC 之间，每一小步可能都很可能跨过基因组的一个物理距离。基于这一点，短距离比长距离更可靠，这一概念已植入 CEPH 的词条“level”中。一个 1 级（level）途径，由两个锚定 STS 组成，它们应至少有一个 YAC 直接连接。这类途径，与平面 STS 含量图中用于确定相邻关系的键或单键相类同。可以让研究者从一个 STS 跳到另一个，而无需跳过任何 YAC/YAC 连接点。相反，一个 2 级途径，由两个锚定 STS 组成，不直接由单个 YAC 连接，而是由 inter-Alu PCR 或指纹法确定在包含它们的两个或多个 YAC 间有一个重叠，所以 2 级途径需要跳过一个 YAC/YAC 连接点。3 级途径需跳过 2 个。4 级需跳过 3 个，等等。尽管每一种的可靠性尚未经验性证明，通过对一套 CEPH 数据的分析暗示 4 级或更高时可能不精确。而幸好 CEPH 途径中近 90%的基于间距为 3 级的或更低。

从 CEPH 服务器得到 YAC 重叠

CEPH 图可以在其单位的网址上在线获得。这里可找到的链接有 YAC 库信息，也有一系列图谱的后转录文件，用于制图的 QuickMap 软件，以及含原始图谱数据的文件。浏览 CEPH 图最好的作用方法为下载 QuickMap 文件，安装并利用它来观看数据文件。然而，由于 QuickMap 只在 Sun 工作站工作，这种方法已经不可行。CEPH 也提供针对 QuickMap 的一种在线界面，在通过标有 Infoclone 的链接处可以获得。这时会弹出一页，可以提交一个 STS、或一个基因标记物或一个 YAC 的名称。提交名称后会回执所有关于它的原始图谱数据。该文本是超链接，可以从一个 YAC 的单一 inter-Alu PCR 杂交跳至另一个。

要得到数据，将浏览器连到 CEPH 的网址上。这会弹出 ECPH Genethon 网页。现在找到并选择 I 链接，接下来的一页会要你在一个小文本栏中输入一个 YAC 或一个 STS 的名称。YAC 应遵循简便的 plate_row_column（板块_排_列）格式，如 923_f_6。对于 STS，可以用 GDB 分配的 D－片断名（如果可得的话）或是实验室分配的研究名称。该文件只针对特定事例，所以输入 AFM20ZE3 不会得到正确的名为 AFM220ZE3 的 STS。也应注意 YAC 地址中排的名称应小写。

按下 Query（查询）键，如果该名称存在于 CEPH 数据库中，那么含相似信息的页面将会出现。第一部分包括一些关于 STS 的总体信息，如引物序列和基因图谱信息。第二部分给出 STS 的 YAC 搜索数据。该部分列表中的所有 YAC 通过直接 PAC 扫描均发现含有该 STS，注释 Alu-PCR probe（探针）说明这个 YAC 在 inter-Alu PCR 杂交实验中被选用为探针。第三部分包含与 STS 相邻的 YAC 的信息，它们与 STS 相隔一个 inter-Alu PCR 的距离。

为得到一个 YAC 上的制图信息，可在文本栏输入其名称并按下 Query 键，出现的界面将会给出 YAC、FISH 和 STS 含量图数据的尺寸信息，以及 inter-Alu PCR 和指纹印迹实验中衍生出的重叠信息。

每个 YAC 词条有几个编码与之相关。例如，在直接 PCR 扫描表中，c 说明 CEPH 进行实验的无分歧结果，而 E 说明为单个已证明的 YAC，来源于外在（非 CEPH）实验室。在 YAC/YAC 重叠表中，a 说明为一个 A－PCR 关系，而 f 说明为一个指纹印迹关系。完整的编码表从位于该页上的不同帮助链接中而得到。

CEPH YAC 库的一个子集已由脉冲区凝胶电泳法限定了大小。如果可以得到它，就能得到 YAC 的大小。在某些情况下，可以找到多带，这是污染的结果，或是因为在 YAC 插入区和克隆生长时 DNA 的随机删除所造成的。这种情况下，多 YAC 的大小也会演示出来。

特定人类染色体图谱

除基因组图谱外，许多个体染色体物理图谱也由研究实验室和基因组中心构建起来了。在很多情况下，这些图谱能比相应基因组范围图谱提供更详尽的信息。在 GDB 的来源页面上可得到一个最新的表。另一张表由 NHGRI 的网址保存。

3·鼠类图谱来源

现在对鼠类作物理图活动最多的地点是 Whitehead Institute/MIT Center for Genome Research，而且一张 murine STS/YAC 含量图已经被构建起来了。这张图，最终将在 24000 个 YAC 上含有 10000 个 STS。

MIT 的物理图谱可以在 Whitehead 的主页上在线浏览。先按下 Mouse Genetic and Physical Mapping Project（鼠类基因图和物理图制图项目）的链接，然后向下滚动到标有鼠类 STS 物理图谱的部分。这一部分与 Whitehead 人类物理图谱有相同的搜索项和用户界面，但是放射性杂交图数据还不可得。

在 Whitehead 网址上还可以得到基于 6331 个简单相邻长度多态性的鼠类物理图谱，以及这张图与 Copeland/Jenkins 限制性片断长度多态性图的整合。这些 RFLP 图，分辨率为 1.1cM。分辨率更高的鼠类基因图正由 European Collaborative Interspecific Mouse BackCros 项目得到。该图最大的理论分辨率将会达 0.3cM，并且可以在 ECJMBC 的主页上在线得到。到 1997 年 5 月已完成 5 条染色体。

The Mouse Genome Database（MGD）是由 Bar Harbor 的 Jackson Laboratory 维持的一个大型鼠类基因信息的公用数据库。尽管它基本上还是一个基因图库，MGD 还是保留了很多物理图谱信息，包括细胞遗传图谱和 synteny 图，将来一旦得到数据就会加进去。MGD 可在 Jackson Laboratory 的主页上得到。按下标有 Mouse Genome Informatics 的链接，然后是标有 Mouse Genome Database 的链接，可得到

用于不同研究的一个起始网页。在所列选项中包括目录检索、基因和标记物符号检索、以及多态性检索。

CEPH YAC 图	*http://www.cephb.fr/ceph-genethon-map.html*
CHLC 图	*http://www.chlc.org*
ECIMBC 主页	*http://www.hgmp.mrc.ac.uk/MBx/MbxHomepage.html*
Entrez 主页	*http://www.ncbi.nlm.nih.gov/Entrez/*
Entrez 全览页	*http://www.ncbi.nlm.nih.gov/Entrez/nentrez.overview.html*
GDB 主页	*http://gdbwww.gdb.org/*
GDB 来源页	*http://gdbwww.gdb.org/gdb/hgp_resources.html*
Genethon FTP 站点	*ftp://ftp.genethon.fr/pub/Gmap/Nature-1995*
I.M.A.G.E. Consortium	*http://www.bio.llnl.gov/bbrp/image/iresources.html*
Jackson 实验室	*http://www.jax.org/*
NHGRI 来源页	*http://www.nhgri.nih.gov/Data/*
Science 转录物图谱	*http://www.ncbi.nlm.nih.gov/Science96/*
Stanford 主页	*http://shgc.stanford.edu/*
Stanford RH 协议	*http://shgc.stanford.edu/Mapping/rh/procedure/*
Whitehead 主页	*http://www.genome.wi.mit.edu/*
Whitehead FTP 站点	*ftp://www.genome.wi.mit.edu/pub/human_STS_releases*

C.elegans	ACEDB	*http://probe.nalusda.gov:8300/other/*
E.coli	University of Wisonsin	*http://www.genetics.wisc.edu/*
D.melanogaster	FlyBase	*http://flybase.indiana.edu:82/*
S.cerevisiae	SGD,Stanford	*http://genome-www.stanford.edu/Saccharomyces*

6.3 全基因组比较

在不同物种、不同进化水平的生物的相关基因之间进行比较分析，是基因研究的重要手段。目前，我们有了越来越多的模式生物全基因组序列数据，因此，基因的比较研究，也必须从基因的比较，上升到对不同进化水平的生物在全基因组水平上的比较研究。这样的研究将更有效地揭示基因在生命系统中的地位和作用，解释整个生命系统的组成和作用方式。

对伴随人类基因组而完成的大量微生物完整基因组的信息分析，不仅将直接帮助破译人类遗传密码，其本身也可能解决重大的科学问题。因此，由完整基因组研究所导致的比较基因组学必将为后基因组研究开辟新的领域。

6.4 SNP 的发现

人类基因组计划持续产生大量序列数据，清楚表明不同个体在整个基因组有许多点存在 DNA 序列的基本变异。最常见的变异发生在分散的单个核苷酸位置，即单核苷酸多态性（SNPs），估计发生频率大约每 1000 个核苷酸有 1 个。那么，没每 1000 个核苷酸，具有一个群体的基本频率的任何一个双拷贝染色体之间的在任一个位置平均核苷酸的一致性是不同的。SNPs 是双等位基因多态性，即多原则上态性位点的核苷酸一致性通常在人类中倾向于二分之一的机率，而不是四核苷酸机率。

SNPs 在人类遗传学研究中有重要意义。首先，一组 SNPs 发生在蛋白质编码区。特定的 SNPs 等位基因可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。其次，SNPs 可作为遗传作图研究中的遗传标记，帮助定位和鉴定功能基因。推算 3000 个双等位 SNP 标记将足够进行人类全基因组作图；100,000 或更多的 SNPs 能够在更大的群体中进行有效的遗传作图研究。因此，需要发展进行大量 SNP 分析的廉价高效技术，包括 DNA 芯片技术，MALDI-TOF 质谱等。

SNPs 是人类遗传多样性最丰富的形式，可用做复杂遗传性状作图。通过高通量的测序项目的得到的大量数据是丰富的大部分没接上的 SNP 来源。这里介绍一种认一 DNA 来源的遗传序列数据变异发现的整体途径。计划用迅速出现的基因组序列作为模板放置没有作图片段化的序列数据，并用碱基质量数值区别真正的等位基因变异与测序错误。

第七章　功能基因组相关信息分析

功能基因组学是后基因组研究的核心内容，它强调发展和应用整体的（基因组水平或系统水平）实验方法分析基因组序列信息阐明基因功能，特点是采用高通量的实验方法结合的大规模数据统计计算方法进行研究，基本策略是从研究单一基因或蛋

白上升到从系统角度一次研究所有基因或蛋白。随着功能基因组实验研究的深入，大量的数据不断涌现，生物信息学将在功能基因组学研究中的扮演关键角色。

7.1 大规模基因表达谱分析

随着人类基因组测序逐渐接近完成，科学家发现即使获得了完整基因图谱，对了解生命活动还有很大距离。我们从基因图谱不知道基因表达的产物是否出现与何时出现；基因表达产物的浓度是多少；是否存在翻译后的修饰过程，若存在是如何修饰的，等一系列问题。这些问题的实质是不了解按照特定的时间、空间进行的基因表达谱。获得基因表达的信息是比 DNA 序列测定艰巨得多的任务，因为基因表达是依赖于许多因素的动态过程。

国际上在核酸和蛋白质两个层次上发展了分析基因表达谱的新技术，即核酸层次上的 cDNA 芯片（cDNA 微阵列）技术和蛋白质层次上的二维凝胶电泳和测序质谱技术，即蛋白质组(proteome)技术。DNA 芯片技术能够在基因组水平分析基因表达，检测许多基因的转录水平。

对大规模基因表达谱的分析存在新的方法学问题，它们从数学角度看不是简单的 NP 问题、动力系统问题或不确定性问题，而是基因表达网络，因此需要发展新的方法和工具。同时，在芯片等的设计上，也需要从理论到软件的支持

下面主要围绕 cDNA 芯片相关的数据管理和分析问题进行讨论。

1．实验室信息管理系统

cDNA 芯片实验的目的是要在一次实验中同时得到成千上万个基因的表达行为，这样的实验需要有管理实验前后大量数据的能力。设计构建检测基因表达的微阵列需要获得生物体基因的所有序列、注释和克隆。在杂交反应和扫描后，收集到的数据必须以某种方式保存，以便很容易进行图象处理和统计及生物学分析。因此需要建立与大规模高通量实验方法相匹配的实验材料和信息管理系统。

该系统除用来定位和跟踪材料来源（例如，克隆，微阵列，探针）外，还必须管理实验前后大量的数据。此外，还包括实验室设备软件系统，如斯坦福大学 Brown 实验室免费的控制自制机器点样设备软件（http://cmgm.standford.edu/pbrown）

芯片图象处理已有各种软件工具，基本的功能是将不同信号强度点的图像转换为每个点的强度数值。这方面没有一致的方法，许多研究小组仍在开发这类软件。图象分析软件的质量对精确解释玻片和膜上的信号非常关键。NHGRI 的 Yidong Chen 开发了一种复杂的图象分析程序，deArray,可免费获取。

美国国立卫生研究院人类基因组研究所（NHGRI）开发的免费的 cDNA 芯片数据管理分析系统 ArrayDB，涉及微阵列的设计、实验室信息管理、实验结果的处理和解释。下面加以简单介绍。

ArrayDB

ArrayDB 是用来储存、查询和分析 cDNA 芯片实验信息的实验室管理系统。ArrayDB 整合了 cDNA 芯片实验中的多个方面，包括数据管理、用户介面、机器自动点样、扫描和图象处理。ArrayDB 中保存的数据包括实验来源、实验参数和条件以及原始的和经处理的杂交结果。ArrayDB 依托的关系数据库储存了芯片上每个克隆的相关信息，包括基因的简单描述、GenBank 号、IMAGE 克隆识别号、代谢途径号和实验室内部克隆号。ArrayDB 还储存了与 cDNA 芯片制造和实验条件的信息。包括点样相关数据（点样机器的参数）、环境条件（温度、湿度、点样针冲洗条件）等数据。此外，还保存了杂交探针和实验条件，包括研究者的姓名，研究目的和实验条件、组织细胞类型的文本描述。有关杂交的结果的信息包括扫描图象（“原始”结果）、信号强度数据、信号强度比值和本底值。

ArrayDB 的设计允许灵活地提取数据信息。设计策略允许不同来源的数据输入，大多数克隆信息来自 Unigene 数据库(包括序列的命名和获取号)。也允许新分离的还没有获取号及名称的克隆的输入。许多数据输入和处理过程是自动的。软件会自动扫描目录查找新输入数据库中的信号强度数据无须人工辅助，其它自动处理包括很方便地整合信号强度数据和克隆数据。

ArrayDB 的 Web 界面能很方便地进行不同类型信息的查询，从克隆信息到信号强度值到分析结果。ArrayDB 支持各种字段的数据查询，例如克隆 ID、标题、实验编号、序列获取号、微量滴定板编号以及相关克隆的结果。每个克隆的更多信息通过超文本链接至其他数据库如 dbEST、GenBank 或 Unigene，代谢途径信息也可通过链接至 KEGG 得到。

通过序列相似性搜索可以有效地寻找目的基因。ArrayDB 支持对 10K/15K 数据（软件自带数据）进行 BLASTN 搜索以便确定目的基因是否已包含在芯片中。

ArrayDB 能分析单个和多个实验产生的信号强度比值的类型和关系。ArrayViewer 工具支持查询和分析单个实验；MultiExperiment viewer 工具支持多个实验数据。在下述网站可得到更详细信息和相关软件。

DeArray 和 ArrayDB 网址： http://www.nhgri.nih.gov/DIR/LCG/15K/HTML

2・基因表达公共数据库

数据库用途

（1）基础研究 将来自各种生物的表达数据与其它各种分子生物学数据资源，如经注释的基因组序列、启动子、代谢途径数据库等结合，有助于理解基因调控网络、代谢途径、细胞分化和组织发育。例如，比较未知基因与已知基因表达谱的相似性能帮助推测未知基因的功能。

（2）医学及药学研究 例如，如果特定的一些基因的高表达与某种肿瘤密切相关，可以研究这些或其它有相似表达谱的基因的表达的影响条件，或研究能降低表达水平的化合物（潜在药物）。

（3）诊断研究 通过对数据库数据进行基因表达谱的相似性比较对疾病早期诊断具有临床价值。

（4）毒理学研究 例如，了解大鼠某种基因对特定毒剂的反应可帮助预测人的同源性基因的反应情况。

（5）实验质量控制和研究参考 实验室样本与数据库中标准对照样本比较能找出方法和设备问题。此外，还能提供其他研究者的研究现状，避免重复实验，节约经费。

数据库的特点和难点

目前急需建立标准注释的公共数据库，但这是生物信息学迄今面临的最复杂且富有挑战性的工作之一。主要困难来自对实验条件细节的描述，不精确的表达水平相对定量方法以及不断增长的庞大数据量。

目前所有的基因表达水平定量都是相对的：哪些基因差异表达仅仅是与另外一个实验比较而言，或者与相同实验的另一个基因的相比而言。这种方法不能确定 mRNA 的拷贝数，转录水平是总的细胞群的平均水平。结果导致采用不同技术进行基因表达的检测，甚至不同实验室采用相同技术，都有可能不能进行比较。对不同来源数据的进行比较有必要采取两个步骤：首先，原始数据应避免任何改动，比如采取数据标准化（data-normalization）的方法。其次，在实验中设计使用标准化的对照探针和样本以便给出参考点至少使来自同一实验平台的数据标准化。

另一难点是对实验条件的描述，解决方法是对实验方法用采用规范化词汇的文件描述：如基因名称，物种，发育阶段，组织或细胞系。还要考虑偶然的不受控制实验因素也可能影响表达：例如空气湿度，甚至实验室的噪音水平。目前建立一种结构能对将来实验设计的所有细节进行描述显然是不可能的。比较现实的解决办法是大部分采用自由文本描述实验，同时尽可能加上有实用价值的结构。DNA 芯片实验

的标准注释必须采用一致的术语，这有待时间去发展。但目前，就应采用尽可能合理的标准用于 DNA 芯片数据及其注释。

标准化的基因表达公共数据库要有五类必要的信息：

（1）联系信息：提交数据的实验室或研究人员的信息。

（2）杂交靶探针信息：对阵列上的每个"点"，应有相应的 DNA 序列在公共数据库中的编号。对 cDNA 阵列，克隆识别号（如 IMAGE clone_id）应给出。

（3）杂交样本：细胞类型和组织来源用标准语言描述。常规诊断病理中使用的组织和组织病理词汇可被采用，还可采用胚胎发育和器官发生中的标准词汇。样本来源种属的分类学名称（如 *Saccharomyces cerevisiae,Homo sapiens*），应当提供。对有些生物体如啮齿类动物和微生物，品系资料需要提供。关于实验中生物体状况的资料，如用药或未用药非常关键，也需提供。"肿瘤与正常"或不同发育阶段也该注明。细胞或生物体的遗传背景或基因型在特定例子中也应是重要的，如酵母基因缺失和转基因鼠。最后，由于组织处理的会引起差别，故应包括相关的详细处理方法。

（4）mRNA 转录定量：这方面非常关键，很难通过一组"持家基因"做内参照进行标准化，有关的具体定量方法应提供。

（5）统计学意义：理想地，应经济合理地有足够的次数重复一个实验以便给出基因表达测定的变异情况，最好能提供合理的可信度值。

上述表达数据记录的前两个要求是简单的，第三个要求较困难需有标准术语协议，但这并不只是表达数据的要求，类似的要求已在公共序列数据库或专业化的数据库中得到成功解决。目前基因表达数据最富有挑战性的方面是最后两个方面。

现状和计划

几个大的芯片实验室如斯坦福大学和麻省理工学院 Whitehead 研究所等，在发展实验室内部数据库；大的商业化芯片公司如 Affymetrix, Incyte,GeneLogic，正在开发基于 Affymetrix 芯片技术平台的商业化基因表达数据库。哈佛大学已经建立了一个的数据库，数据来自几个公共来源并统一格式。宾夕法尼亚大学计算生物学和信息学实验室正在整合描述样本的术语。

目前至少有 3 个大的公共基因表达数据库项目：美国基因组资源国家中心的 GeneX；美国国家生物技术信息中心（NCBI）的 Gene Expression Omnibus;欧洲生物信息学研究所（EBI）的 ArrayExpress.

欧美专家合作提出有关数据库的初步标准：实验描述和数据表示的标准；芯片数据 XML 交换格式；样本描述的术语；标准化、质量控制和跨平台比较；数据查询语言和数据挖掘途径。（http://www.ebi.ac.uk/microarray/）。EBI 与德国癌症研究中心正在开发 ArrayExpress，一种与目前推荐标准兼容的基因表达数据库。该数据库将利用来自合作方的的数据，可操作的数据库将于近期建立（http://www.ebi.ac.uk/arrayexpress）。

3．大规模基因表达谱数据分析方法

芯片分析能够检测不同条件下的基因转录变化，能够显示反映特征组织类型、发育阶段、环境条件应答、遗传改变的基因谱。当芯片数据大量出现，产生了新的问题：如果将所有获得的数据集中起来，我们能否将未知功能的新基因归类到已知功能分类中？能否将基因表达与基因功能联系起来？能否发现新类型的共调控基因？能否从芯片表达数据中得出完整的基因调控网络？这些唯有通过计算的方法。基因制图及测序所面临的问题与大规模基因表达分析的数学问题相比要小的多。这种新类型的表达数据使我们直接面对生物系统和基因组水平功能的复杂性，从生物系统单个成分的定性发展到完整生物系统行为的描述上来，这方面困难很多，目前只有很少的分析工具。

聚类分析（clustering analysis）是大规模基因表达谱目前最广泛使用的统计技术，最近又发展了一种机器学习方法-支持向量机（support vector machines,SVMs）。这些分析方法均处在研究的初级阶段，随着大量数据及标准化数据库的出现，其它数据挖掘技术包括神经网络和遗传算法将在基因表达数据分析中得到应用。

聚类分析

聚类通过把目标数据放入少数相对同源的组或“类”（cluster）里。分析表达数据，（1）通过一系列的检测将待测的一组基因的变异标准化，然后成对比较线性协方差。（2）通过把用最紧密关联的谱来放基因进行样本聚类，例如用简单的层级聚类（hierarchical clustering）方法。这种聚类亦可扩展到每个实验样本，利用一组基因总的线性相关进行聚类。（3）多维等级分析（multidimensional scaling analysis,MDS）是一种在二维 Euclidean “距离”中显示实验样本相关的大约程度。（4）K-means 方法聚类，通过重复再分配类成员来使“类”内分散度最小化的方法。

聚类方法有两个显著的局限：首先，要聚类结果要明确就需分离度很好（well-separated）的数据。几乎所有现存的算法都是从互相区别的不重叠的类数据中产生同样的聚类。但是，如果类是扩散且互相渗透，那么每种算法的的结果将有点不同。结果，每种算法界定的边界不清，每种聚类算法得到各自的最适结果，每个数据部分将产生单一的信息。为解释因不同算法使同样数据产生不同结果，必须注意判断不同的方式。对遗传学家来说，正确解释来自任一算法的聚类内容的实际结果

是困难的（特别是边界）。最终，将需要经验可信度通过序列比较来指导聚类解释。

第二个局限由线性相关产生。上述的所有聚类方法分析的仅是简单的一对一的关系。因为只是成对的线性比较，大大减少发现表达类型关系的计算量，但忽视了生物系统多因素和非线性的特点。

斯坦福大学的 Michael Eisen 开发的 Windows 平台免费芯片数据分析软件 CLUSTER 和 TREEVIEW，采用配对平均连锁（pairwise average-linkage）聚类分析。这种方法中，每个不同的基因与其它的基因比较，鉴定最相关的基因对。这种基因对的数据用平均数替代，再重新计算关系矩阵，不断重复这个过程。TREEVIEW 对 CLUSTER 计算结果进行图形输出，将芯片中的每个基因的表达比值用彩色方块表示。

尽管 CLUSTER 软件易于使用且直观，但其算法仍有缺陷之处：实际数据由每次重复的平均数据替代；相似性测定的选择（相关性/Eluclidean 距离）；将等级模型用于非等级过程；成对比较矩阵的计算负担。因此，出现了其它方法，包括自组织图（self organizing maps,SOMs），二进制决定-退火算法（binary deterministic-annealing algorithm）,k-means 聚类等。Tamayo 等提供 Windows 平台的 SOMs 软件包。

CLUSTER 和 TREEVIE 下载网址：http://www.genome.standford.edu

基于知识挖掘的机器学习方法

最近发展了一种的有监督的机器学习方法-支持向量机（support vector machines,SVMs）来分析表达数据，它通过训练一种“分类器”来辨识与已知的共调控基因表达类型相似的的新基因。与经典的无监督聚类方法（unsupervised clustering）和自组织图（self-organizing maps）不同，该方法建立在已有的知识上并有改进现有知识的潜力。

无监督的聚类方法，例如层级（hierarchical）和 K-means 聚类，假设每个基因仅属于一“类”（cluster）。这在生物学意义上当然不是真实的。而且，事实上同一类基因不是必然意味着有相似的表达类型。比如，k-means 聚类方法事先指定产生的“类”的数量及并将每个基因放在其最优“类”，并不总是有意义。需要对类（cluster）进行质量评价，“类”的“严谨性”和外围基因的存在（如果存在，它们与下一类的接近度）以及一组核心特征基因应在质量上保证。最重要的是应考虑“类”是否有生物学意义。

与无监督的方法产生基因的“类”相比，有监督的学习方法是向已知的“类”学习。训练者必须提供 SVMs 以每个“类”正反两方面的例子。SVMs 提供一种层级的方法来分析芯片数据。首先，对每个基因，应询问最近的邻居是否它与它们的关系是有生

物学意义的。其次，对已知共调控基因，应该询问它们的表达类型是否相似，如果是这样，还有哪些其它的基因有相同类型。这些在监督阶段可通过 SVMs 或优化的 SOMs 来判断。第三，应该通过无监督的学习方法进行基因分类并询问是否聚类有生物学意义并且包括外围基因。最后，“类”可通过每个无监督的“类”的核心基因训练 SVMs 的方法来检测和优化。

可视化

大规模基因表达数据挖掘另一重要方面是发展有力的数据可视化方法和工具。已经发展了用简单图形显示提供聚类结果的途径，如上述的 TREEVIEW 软件。对大规模基因表达原始数据的进行不失真的可视化并链接的标注过的序列数据库，可为基因表达分析提供非常有价值的工具，有助于从新的视角看待基因组水平的转录调控并建立模型。

7.2 基因组水平蛋白质功能综合预测

蛋白质之间的功能联系

基因组测序计划在产生完全的组成多个亚单位装配和信号通路的蛋白质列表方面取得里程碑式的业绩。这些装配和通路现在必然被制图，Marcotte 等和 Enright 等在此方面走了显著一步。这两个研究小组发展了不是通过氨基酸序列相似性比较的其他特性联系起蛋白质的计算方法。通过比较系统发育（进化）谱和表达类型，以及通过分析结构域融合（domain fusions）新方法识别在代谢通路、信号通路或结构复合体上功能相关的蛋白质。酵母未定性蛋白大约一半&#0;总蛋白数约四分之一可用此方法进行功能注释。因为不依赖于直接的序列相似性，这种方法可预测与已知功能蛋白质缺乏同源性的蛋白质功能。将会发现它们在基因组学中的许多应用，与大规模蛋白质功能实验互为补充。

构建通路和专配有用模型的信息来自实验，最重要的通过蛋白质组学和结构基因组学。蛋白质组学的目标是对所有的蛋白质和蛋白相互作用进行鉴定和定性。它包括采用大规模实验方法如双杂交系统（two-hybrid system）、质谱法(mass spectrometry,MS)、二维凝胶电泳（2D PAGE）和 DNA 芯片杂交（DNA microarray hybridization）。任务大小和复杂性可由下面的假定理解：每个蛋白质有 5-50 个功能连锁，结果在一个酵母细胞中就有 30,000-300,000 个连锁。虽然实验已确定了约 30%的酵母的功能，但是它们有时不是迅速廉价的，且不完全。因此需要用计算的方法来预测功能。

计算方法传统上预测功能是通过与性质明确蛋白质的序列相似性比较。这样标注的可行性是因为进化产生享有共同祖先的的同源性蛋白家族，因此有相似的序列、结构，经常还有功能。蛋白质比较允许对酵母另 30%的蛋白质功能进行研究。但是，通过同源性进行功能预测受两方面的因素制约。首先，它只能用于与已知功能蛋白质有同源性的未知蛋白质的功能预测。其次，不是总清楚匹配的蛋白质何种功能特性为其共享，尤其对那些距离较远的匹配。

Marcotte 等和 Enright 等并未受此限制，因为他们不依赖与未知蛋白质与已知功能蛋白质的序列相似性。而代替的是，将同样通路和装配的蛋白质分组，定义为"功能连锁"（functionally linked）.Marcotte 等针对出芽酵亩基因组蛋白质采用了三种不同的方法：系统发育谱（phylogenetic profiles），结构域融合(domain-fusion analysis)和相关 mRNA 表达类型(correlated messenger RNA expression patterns)。Enright 等独立发展了结构域融合分析，采用新的聚类算法用于三个原核基因组分析。

系统发育谱依赖于蛋白质相关进化。两个蛋白质是进化相关的当它们共有一个系统发育谱，定义为蛋白质在一组基因组中的发生率类型。仅当几个完整的基因组比较时系统发育表达谱才能精确计算。两个蛋白质享有相似的系统发育谱被认为是功能连锁（functionally linked）。因此，根据系统发育谱进行的蛋白质聚类，当未知蛋白质与一个或更多的功能已知的蛋白质归为一组时能够提供未知蛋白质的功能信息。

结构域融合的方法鉴定含有两个分别在其它基因组的非同源性成分蛋白（component proteins）组成的融合蛋白(fusion proteins)。这样的成分蛋白被认为彼此物理上有相互作用。在两个相互作用成分蛋白之间的界面（interface）更有可能进化当两个蛋白融合为一条单一链。著名的例子是，从细菌到真菌的色氨酸合成酶的 α 和 β 亚单位。在一些方面，结构域融合分析与从基因邻近效应（gene proximity）推测功能连锁相似。

Marcotte 等也通过关联它们的 mRNA 表达类型来对酵母蛋白质进行分类。这些类型来自 97 组公共 DNA 芯片数据，显示了大多数酵母蛋白质在正常生长、葡萄糖缺乏孢子形成和突变基因表达的条件下的表达变化。分析建立在认为在一系列相同条件下表达水平相互关联的蛋白质是功能连锁的。

新的功能注释经常是广义的，限制蛋白质的功能为，"代谢"或"转录"。即使随机的一对蛋白质也有 50%的相似机率在这样广义的水平上。但是因为注释一般来自许多连锁，比随机连锁信息量大 3-8 倍，在一些例子中与蛋白-蛋白相互作用的实验决定相比。 例如，Marcotte 等建立了新的 MSH6 的连锁，在某些结肠癌中的 DNA 错配修复蛋白，属于 PMS1 错配修复家族，其中的突变也与人结肠癌、嘌呤生物合成途径、RNA 修饰酶和一个未知的蛋白质家族相关，这样它们可以通过核酸修复或修饰来研究。

这样的注释精确度如何？能覆盖多少比例的蛋白质？这些问题只能部分提出，因为参考的功能连锁蛋白质不是很容易得到。Marcotte 和同事给酵母 2,557 个未知蛋白的一半预测了一般功能。他们估计成对预测来确定功能的近 30%是错误的，虽然两到三种方法联合应用使错误率降到 15%。

Enright 等通过结构域融合在三个原核基因组中仅功能连锁 215 个蛋白，但是非常少的估计假阳性。较少的功能连锁率可能由于没有系统发育谱和 mRNA 表达方法丢失了连锁（作者没有做这两种方法），融合事件更严格的定义以及用较少的蛋白检测融合。尽管假阳性和显得粗糙的功能注释，计算方法使得实验者将注意力集中在有希望的相互作用上。当得到更多的基因组数据，结构域融合和系统发育谱的方法的预测数和精度将增加。

下一步将是提高方法预测蛋白质功能的范围、准确度和精确性。这可能在理论上，通过考虑三维结构来做，因为蛋白质的功能更多直接由它的结构和动力学而不是它的序列来决定。那么为什么在基因组学上结构没有序列用的广泛呢？至少有两个原因。首先，只有一部分蛋白质有三维结构数据。这种限制在几年内随着结构基因组学（structural genomics）的进展而减少。结构基因组学的目标是确定大约 10,000 经仔细挑选的蛋白质结构域的结构，以便所有其它的蛋白质序列能够有很好的精确性建模。其次，能够从结构而不是从序列提取的功能细节依赖于细胞环境下的那种结构的细节，同样也依赖于它的动力学和能量，所有这些在现有的实验和理论技术下难以获得。