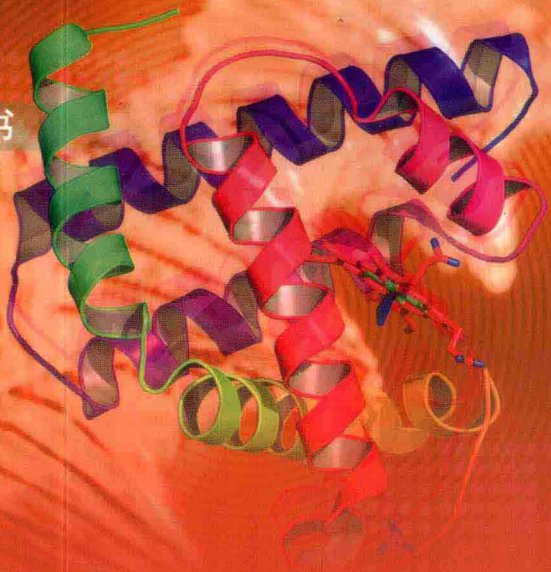
生物信息学数据分析丛书



DNA和蛋白质序列数据分析工具（第二版）

TOOLS FOR ANALYSIS OF DNA AND PROTEIN SEQUENCE DATA
(Second Edition)

薛庆中　等 编著

科学出版社
www.sciencep.com

生物信息学数据分析丛书

Tools for Analysis of DNA and Protein Sequence Data

（Second Edition）

DNA 和蛋白质序列数据分析工具

（第二版）

薛庆中 等 编著

科学出版社

北京

内 容 简 介

在众多生物基因组测序项目完成之际，我们面临的最大挑战是如何对 DNA 和蛋白质数据进行科学的分析和注释。本书分三个层次解读基因数据库和网络工具：基因组学层面重点介绍序列比对工具 BLAST 和 ClustalX 的使用、真核生物基因结构的预测、电子克隆及分子进化遗传分析工具（MEGA 4）的使用；蛋白质组学层面介绍了蛋白质结构与功能预测、序列模体的识别和解析、蛋白质谱数据分析、基因芯片数据处理和分析，以及应用 GO 注释基因功能和通过 KEGG 分析代谢途径；系统生物学层面从网络结构分析阐述了蛋白质与蛋白质的相互作用；此外，还增添了使用 Bioperl 模块进行数据分析和 Windows 操作系统下 Bioperl 程序包的安装等内容。书中提及的各种方法均有充实的例证并附上相关数据和图表，供读者理解和参考；书后还附有中英文的专业术语和词汇。

本书可作为对生物信息学专业感兴趣的本科生、研究生和研究人员学习、研究的重要工具手册。

图书在版编目（CIP）数据

DNA 和蛋白质序列数据分析工具/薛庆中等编著. —2 版. —北京：科学出版社，2010

（生物信息学数据分析丛书）

ISBN 978-7-03-027052-8

Ⅰ. ①DNA… Ⅱ. ①薛… Ⅲ.①脱氧核糖核酸–数据–分析②蛋白质–数据–分析 Ⅳ. ①Q523②Q51

中国版本图书馆 CIP 数据核字（2010）第 047908 号

责任编辑：李 悦 刘 晶/责任校对：桂伟利

责任印制：钱玉芬/封面设计：耕者设计工作室

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

骏走印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2009 年 1 月第 一 版 开本：B5 (720×1000)

2010 年 4 月第 二 版 印张：18 1/4 插页：4

2010 年 4 月第一次印刷 字数：347 000

印数：1—3 500

定价：48.00 元

（如有印装质量问题，我社负责调换）

编委会名单

（按姓氏汉语拼音排序）

陈　辰　陈晓龙　程　尹　丁文超

冯　晔　韩　序　华大颂　黄鹏宇

蒋　琰　骆迎峰　莫　凡　阮　陟

王　珺　王庭璋　薛庆中　叶　琳

张维一

第二版前言

第一版《DNA 和蛋白质序列数据分析工具》一书发行后三个月，出版社就来信告诉我“该书销售实在是太火爆了，现在库房就剩 24 本，准备重印”。与此同时，参加我们的“基因组科学研习班”的人员也在激增。这在一定程度上反映了国内读者对这一科学“盲点”开始重视并产生了强烈兴趣，对我们这些作者来说，这无疑是莫大的鼓励和鞭策。随着新一代测序仪的问世，生物数据库中的 DNA 和蛋白质序列数据量直线飙升，计算机专家们开发的新工具不断涌现，更新速度之快令人瞩目，我们很快就发现完成不久的工具书已有落伍的迹象。为此，我们打算借重印的机会，对其中部分内容加以修改，并将其中一章的修改稿送交出版社审阅。责任编辑阅后，建议补充些内容重新出版一本新书，作为该书的再版。于是我们就开始着手组织人员写作新书。

第二版中我们坚持“兼顾学术思想的前沿性和写作的通俗性”的原则，主要面向初学的读者群，尽量使用 Windows 操作系统下的在线工具，配以详细的图文注释，同时对英文的专业术语和词汇进行了翻译，便于读者自学和操作。即使是缺乏基因组知识的读者，通过参加“基因组科学研习班”学习也能较快入门。

第二版中我们对书稿进行了较大修整，不仅丰富了新书的科学内容，同时加深了对数据的诠释，并适当增加了相关生物学背景知识的介绍。与第一版相比，我们主要做了以下几个方面的修改和补充。

其一，软件普遍升级，网页版面全面更新。例如，用 GPM Tornago XE（X! 系列图形界面）替代了原来的 X!Tandem 软件后，使文件输入、参数设置、结果输出等信息都实现了可视化。在系统进化树构建时，使用 MEGA 升级版后就可以免除对多序列比对结果文件格式转换的过程，应用更为便捷。原来介绍使用的 ClustalX1.83 也已升级为 ClustalX2。

其二，补充介绍了新内容。例如，在基因芯片数据处理和分析时，增加了 MeV 的聚类、差异表达基因筛选内容；在 KEGG 数据库中添加软件 KegArray 实现了 KEGG 数据库和基因芯片数据整合分析。系统生物学网络结构分析中，增补了应用插件 Cytoprophet 预测潜在蛋白质和结构域的相互作用。

其三，对原有内容重新进行梳理。例如，在蛋白质结构与功能预测中，以 ExPASy 数据库提供的蛋白质分析系列工具为引线，整合了 InterPro 各相关数据库介绍。

其四，增加了三章内容，分别是：序列模体的识别和解析（第 6 章）、使用 Bioperl 模块作数据分析（第 11 章）、Windows 操作系统下 Bioperl 程序包的安

装（第 12 章）。其中利用 Bioperl 网站的模块可以免去自编程序的困难，可能会使读者感兴趣并带来方便。

这次再版我们吸纳了几位博士生参与，使编委人数增加到 17 位，他们是：陈辰、陈晓龙、程尹、丁文超、冯晔、韩序、华大颂、黄鹏宇、蒋琰、骆迎峰、莫凡、阮陟、王珺、王庭璋、薛庆中、叶琳、张维一。其中部分编者虽已离开浙江大学去国外深造或转到其他岗位工作，但仍然积极参与本书的工作或予以关心。在美国学习的博士生程尹在结束期末考试后就急忙着手文稿修改；在法国的博士生王庭璋为实现 VISTA 操作系统下的 Bioperl 程序包的安装，煞费苦心对每个操作细节反复调试和摸索，直到模块顺畅运行。在稿件修订过程中，我与各章编写人员多次讨论、切磋，尽量使差错减少。有些内容还提前在近期的培训班上试讲和使用，征求学员们的意见直至取得良好效果。经过编委们三个多月的共同努力，使再版工作顺利完成。

本书的编写工作依旧得到了浙江大学浙江加州国际纳米技术研究院领导的大力支持，特此表示衷心感谢。还要感谢科学出版社李悦女士的热心和认真。第一版中发现的一些错误虽已进行了纠正，然而，不免又会增生新的差错和缺点，还望读者雅正。我们期待第二版的发行更加顺畅，使更多读者了解基因组，走近基因组科学。

薛庆中

2009 年 9 月于浙江大学

第一版前言

当今生物基因组 DNA 测序数据总量正在以指数倍的速度增长。如何对数据库的海量数据进行科学的搜集、管理、挖掘、注释已成为基因组学和蛋白组学研究的热点。为普及和提高我国科学工作者基因组科学知识，学习并掌握序列数据分析的实用操作技能，及时了解该领域的最新进展，自 2003 年以来，浙江大学和中国科学院基因组研究所紧密协作已举办了 24 期基因组科学培训班。培训学员来自全国各地 29 个省市，人次多达 1800 余人，每次培训班中都不仅常见到较多教授和副教授们的身影，年轻的研究生更是踊跃参加。他们的专业背景虽然各自不同，涵盖理学、工学、医学、农学等不同门类和学科，但渴求知识、不断进取的态度却是一样的。

基因组科学培训班由杨焕明、于军、郑树、林标扬、胡松年、薛庆中、徐宇虹等教授担当主讲教师。他们不仅对基因组科学的基本概念加以正确诠释，对当今的最新进展进行全面介绍，并能结合自己的科研工作，分别讲解他们在医学、农学、微生物等领域具体的应用实例。学员们反映，通过这些生动趣味的讲座加深了他们对 DNA 数据挖掘的理解，有助于开阔研究视野和工作思路，同时激发了学习这门前沿科学的兴趣和热情。

培训班主要学习数据库搜索和实用工具的操作，采用“跟我学”的教学方式，指导教师边讲边示范，学员们每人备有电脑，跟着大屏幕一步步操作；辅导教员随时在旁帮助解难，使学员们在较短时间内尽快初步掌握基本操作程序。为满足培训的需要，我们先后编写了《基因组数据分析手册》（胡松年和薛庆中，2003）和《EST 数据分析手册》（胡松年，2005），得到良好的反映和发行量。教学内容的不断更新是培训班久盛不衰的保证。近期培训内容中我们又新增芯片数据、蛋白质数据和系统生物学网络结构显示与分析等内容。为此，在前两本书的基础上，我们新编写了《DNA 和蛋白质序列数据分析工具》一书。

全书分 9 章。第 1 章，阐述序列比较的核心方法，即运用 BLAST 和 ClustalX 等工具做序列比对。第 2 章，重点介绍核苷酸序列分析工具，主要包括：基因可读框的识别，CpG 岛、转录终止信号和启动子区域的预测分析，用 mRNA 序列预测基因等。第 3 章，介绍电子克隆的概念和具体操作方法。第 4 章，用 MEGA4 做分子进化遗传分析，绘制系统进化树，为研究基因进化打好基础。第 5 章，对蛋白质基本理化性质、二级结构、结构域和三维空间结构、预测目标蛋白的生物学功能等工具做逐一介绍。第 6 章，通过 Gene Ontology 和 KEGG 两个数据库，挖掘基因和蛋白质的功能并做代谢途径分析。第 7 章，利用 X!Tandem 软件鉴定蛋白质的串联质谱数据，进而预测蛋白质；同时借助 TPP 软件包进行蛋白质组学数据统计学分析，

优化检索结果。第 8 章，使用 TM4 软件实现芯片的数据采集和标准化处理，并借助 GenMAPP 软件挖掘芯片数据的生物学意义。第 9 章，通过 Cytoscape 软件演示，介绍系统生物学分析概况，展示蛋白质-蛋白质相互作用，应用插件做网络结构分析。

本书的特点是较好地兼顾了学术思想前沿性和写作的通俗性。其前沿性体现在汇集了现代 DNA 和蛋白质序列分析内容精髓，对包括芯片数据、质谱数据处理和分析、系统生物学分析等各类数据分析工具进行扫描和重点介绍，而通俗性则通过较多使用网上在线工具配以详细的图文注释实现，同时写作上力求通俗渐进，有助于科研及教学人员，通过培训结合网络自学，掌握数据库搜索及其常用工具的操作方法，从中感悟 DNA 和蛋白质数据分析方法的要领。

本书内容已在浙江大学基因组科学培训班中试用四期，学员们反映，经过培训可以初步掌握上述方法，并结合阅读教材复习巩固并能用于自己的课题中。“快节奏、高效率”的会务组织、安排，也给学员们在浙江大学培训期间营造了良好的学习氛围，深受学员赞誉。培训期间还组织学员参观了浙江大学纳米研究院的科研设施，如质谱仪、芯片点样、分子影像等国际一流的仪器设备使他们大开眼界，增长见识，产生协作研究和学习的愿望。

值得庆幸的是，2006 年我们的培训班迎来了沃森博士，他因与克里克博士共同发现 DNA 双螺旋结构而荣获诺贝尔奖。他到培训课堂与学员们亲切交流，极大鼓舞了大家的学习信心，也为我们力争将培训班办成“东方冷泉港”模式增添了动力。我们期待通过培训班和本书的发行，对宣传基因组科学在国民经济建设中的作用，普及生物学知识，培养年轻科学人才等方面作出贡献。例如，浙江大学博士生苏志熙曾在培训班当过辅导教师，他撰写的基因组论文发表在 *Genome Research* 上，2007 年被评为国内 100 篇最具影响力的论文。还有不少科研单位通过培训班和浙江大学建立科研协作，共同撰写发表了 SCI 论文。

本书由陈辰、陈晓龙、程尹、冯晔、洪旭、黄鹏宇、蒋琰、骆迎峰、莫凡、王珺、薛庆中、叶琳 12 位编者共同完成，为使全书文笔流畅连贯，我们在汇总时，对全部文字和图表进行了认真修正和统一处理。对网上工具的术语和名称尽量备注了英文，并在书后附有中英文专业词汇对照供读者查阅。为指导计算机操作，我们还在相应图表中做了文字注释和符号标记。但是，书中的错误和缺点仍在所难免，恳请读者指正。

本书的编写得到了浙江大学浙江加州国际纳米技术研究院领导的大力支持，胡松年研究员和徐宇虹教授对本书给予了热忱的推荐和鼓励，同时得到浙江省政府项目和国家自然科学基金（30571146）资助。对陈爱华在培训班和书稿的组织贡献和鲁平在培训班的努力工作，在此一并表示衷心感谢。

薛庆中

2008 年 8 月于浙江大学

目　录

第二版前言

第一版前言

第 1 章　序列比对工具 BLAST 和 ClustalX 的使用 ……1

1.1　序列比对 BLAST ……1

1.1.1　Basic BLAST ……1

1.1.2　网上 blastx 比对 ……4

1.1.3　网上 PSI-Blast 比对 ……7

1.1.4　Specialized BLAST ……9

1.1.5　网上 Blast2 比对 ……10

1.2　本地运行 BLAST（Windows 系统） ……11

1.2.1　BLAST 程序下载 ……11

1.2.2　BLAST 程序安装 ……11

1.2.3　进入 DOS 命令行提示符状态 ……13

1.2.4　搜索数据库的格式化 ……14

1.2.5　BLAST 搜索程序运行 ……14

1.2.6　本地化 BLAST 搜索结果查看 ……15

1.3　多序列比对（ClustalX） ……15

1.3.1　ClustalX 的使用 ……16

1.3.2　数据的输入 ……17

1.3.3　数据的输出 ……20

主要参考文献 ……22

第 2 章　真核生物基因结构的预测 ……23

2.1　基因可读框的识别 ……23

2.2　CpG 岛、转录终止信号和启动子区域的预测 ……24

2.2.1　CpG 岛的预测 ……24

2.2.2　转录终止信号的预测 ……26

2.2.3　启动子区域的预测 ……27

2.3　基因密码子偏好性计算：CodonW 的使用 ……29

2.4　采用 mRNA 序列预测基因：Spidey 的使用 ……31

2.5　ASTD 数据库简介 ……33

主要参考文献 ……37

第 3 章　电子克隆 ……38
3.1　利用 UniGene 数据库进行电子延伸 ……38
3.1.1　目标序列的检索 ……39
3.1.2　UniGene 数据库检索 ……40
3.2　从数据库中获取 cDNA 全长序列 ……43
3.3　本地序列拼接 ……44
3.3.1　CAP3 序列拼接程序 ……45
3.3.2　Velvet 序列拼接程序 ……47
3.4　基因的电子表达谱分析 ……52
3.5　核酸序列的电子基因定位分析 ……54
主要参考文献 ……57
第 4 章　分子进化遗传分析工具（MEGA 4）的使用 ……58
4.1　序列数据的获取和比对 ……58
4.1.1　数据库直接检索 ……59
4.1.2　多序列比对 ……60
4.2　进化距离的估计 ……62
4.3　分子钟假说的检验 ……64
4.4　系统进化树构建 ……66
4.4.1　系统进化树构建方法选择 ……66
4.4.2　进化树的树形选择 ……68
4.4.3　进化树的拓扑结构调整 ……70
4.4.4　进化树树枝形态的优化 ……71
4.4.5　进化树的保存 ……73
主要参考文献 ……74
第 5 章　蛋白质结构与功能预测 ……75
5.1　蛋白质一级结构分析 ……76
5.1.1　ProtParam：蛋白质序列理化参数检索 ……76
5.1.2　ProtScale：蛋白质亲疏水性分析 ……78
5.1.3　COILS：卷曲螺旋预测 ……81
5.2　蛋白质二级结构预测 ……83
5.2.1　PredictProtein：蛋白质结构预测 ……83
5.2.2　PSIPRED：不同蛋白质结构预测方法 ……87
5.3　InterProScan：模式和序列谱研究 ……88
5.3.1　InterProScan 简介 ……88
5.3.2　PROSITE：蛋白质结构域、家族和功能位点数据库 ……91

5.3.3 Pfam：蛋白质家族比对和 HMM 数据库……95
5.3.4 BLOCKS：模块搜索数据库……97
5.3.5 SMART：简单模块构架搜索工具……98
5.3.6 TMHMM：跨膜区结构预测服务器……100
5.4 蛋白质三级结构预测……101
5.4.1 Swiss-Model Workspace：同源建模的网络综合服务器……102
5.4.2 Phyre（successor of 3D-PSSM）：线串法预测蛋白质折叠……104
5.4.3 HMMSTR/Rosetta：从头预测蛋白质结构……106
5.4.4 Swiss-PdbViewer：分子建模和可视化工具……106
主要参考文献……109
第 6 章 序列模体的识别和解析……110
6.1 MEME 程序包……110
6.2 通过 MEME 识别 DNA 或蛋白质序列组中模体……111
6.3 通过 MAST 搜索序列中的已知模体……114
6.4 通过 GLAM2 识别有空位的模体……117
6.5 通过 GLAM2SCAN 搜索序列中的已知模体……120
6.6 应用 TOMTOM 与数据库中的已知模体进行搜索比对……122
6.7 应用 GOMO 鉴定模体的功能……123
主要参考文献……124
第 7 章 蛋白质谱数据分析……125
7.1 生物质谱技术介绍……125
7.1.1 质谱技术的基本原理……125
7.1.2 X!Tandem 软件……129
7.1.3 Mascot 软件……135
7.1.4 Sequest 软件……139
7.2 蛋白质组学数据统计分析软件……142
7.2.1 TPP 简介……142
7.2.2 TPP 的安装与配置……143
7.2.3 样本数据准备……144
7.2.4 将 RAW 文件转换成 mzXML 文件……145
7.2.5 由 out 数据文件夹生成 pepXML 文件……147
7.2.6 运行 PeptideProphet……151
7.2.7 PeptideProphet 处理后的结果分析……154
7.2.8 运行 ProteinProphet……154
7.2.9 数据的过滤筛选和将结果保存成 Excel 文件……158

主要参考文献……160
第 8 章 基因芯片数据处理和分析……161
8.1 芯片数据的获取和处理……161
8.1.1 Express Converter……161
8.1.2 MIDAS……163
8.2 芯片数据聚类分析和差异表达基因筛选……168
8.2.1 MeV……168
8.2.2 Cluster……174
8.2.3 TreeView……175
8.3 芯片数据的可视化……176
8.3.1 GenMAPP 的概念……176
8.3.2 GenMAPP 的安装……177
8.3.3 GenMAPP 的使用……177
8.4 芯片数据的检索和提交……182
8.4.1 GEO 检索……182
8.4.2 Platform 信息……183
8.4.3 Series 信息……184
8.4.4 Samples 信息……184
8.4.5 芯片数据的提交……185
主要参考文献……186
第 9 章 应用 GO 注释基因功能和通过 KEGG 分析代谢途径……187
9.1 Gene Ontology 数据库……187
9.1.1 简介……187
9.1.2 用关键词检索 GO 数据库……188
9.1.3 用序列检索 GO 数据库……193
9.2 KEGG 数据库……194
9.2.1 简介……194
9.2.2 根据代谢途径名称检索……196
9.2.3 根据基因名称检索……198
9.2.4 根据序列检索……199
9.2.5 利用 KAAS 工具作批量注释……201
9.2.6 基因芯片数据的代谢途径分析……204
主要参考文献……207
第 10 章 系统生物学网络结构分析……208
10.1 Cytoscape 软件简介……208

10.1.1 概况 ……208
10.1.2 主要功能 ……208
10.2 软件安装 ……209
10.3 Cytoscape 基本操作 ……209
10.3.1 信息输入 ……209
10.3.2 插件安装 ……215
10.4 应用 Cytoscape 进行基因注释 ……215
10.4.1 BiNGO 插件的安装 ……215
10.4.2 使用实例 ……215
10.5 应用 Cytoscape 进行亚细胞定位 ……218
10.5.1 Cerebral 插件的安装 ……218
10.5.2 使用实例 ……218
10.6 应用 Cytoscape 搜索基因相互作用文献 ……222
10.6.1 Agilent Literature Search 插件安装 ……222
10.6.2 使用实例 ……222
10.7 应用 Cytoscape 做网络分析 ……225
10.7.1 将 BOND 网络数据库的信息输入 Cytoscape ……225
10.7.2 网络分析 ……227
10.8 应用插件 Cytoprophet 预测潜在蛋白和结构域的相互作用 ……230
10.8.1 Cytoprophet 插件的安装 ……230
10.8.2 使用实例 ……230
主要参考文献 ……234
第 11 章 使用 Bioperl 模块作数据分析 ……235
11.1 概述 ……235
11.2 Bioperl 安装 ……236
11.2.1 Bioperl 的组成 ……236
11.2.2 Unix/Linux 系统下 Bioperl 的安装步骤 ……236
11.2.3 Windows 系统下 Bioperl 的安装 ……236
11.3 Bioperl 重要模块简介和脚本实例 ……236
11.3.1 实现文件格式转换脚本（实例 1） ……236
11.3.2 实现 DNA 序列的翻译脚本（实例 2） ……237
11.3.3 计算序列长度脚本（实例 3） ……238
11.3.4 GenBank 文件解析脚本（实例 4） ……239
11.3.5 图形化显示序列特征脚本（实例 5） ……240
11.3.6 从公共数据库获取序列脚本（实例 6） ……242

11.3.7 应用 AlignIO 模块实现文件格式转换脚本（实例 7）……242
11.3.8 计算比对序列 JukesCantor 距离脚本（实例 8）……243
11.3.9 计算同义替换率（D_s）和非同义替换率（D_n）脚本（实例 9）……244
11.3.10 Bioperl 调用 Clustalw 脚本实例（实例 10）……244
主要参考文献……245
第 12 章 Windows 环境下 Bioperl 程序包的安装……246
12.1 Vista 下 Perl 语言环境的安装……246
12.1.1 下载 Perl 文件……246
12.1.2 Perl 安装文件……246
12.2 Bioperl 的安装……247
12.2.1 启动 Perl Package Manager……247
12.2.2 丰富 Bioperl 资源库……247
12.2.3 安装 Bioperl 核心包和工具盒……249
12.3 局域网通过代理服务器用户的安装……250
12.4 在 Windows XP 系统下安装……252
12.4.1 下载 Perl 文件……252
12.4.2 安装 Bioperl……252
12.4.3 局域网通过代理服务器用户的安装……252
12.5 从 CPAN 安装 Perl 文件……252
12.5.1 调试 Bioperl 程序……252
12.5.2 个别下载 Bioperl 模块文件……254
12.5.3 个别安装 Perl 模块 Bio::Graphics……254
12.5.4 调试 Bio::Graphics 模块……254
12.6 安装多序列比对程序 Clustalw 模块……256
12.6.1 下载并安装 Clustalw 程序……256
12.6.2 设置系统环境变量……258
12.6.3 测试 Bioperl 与 Clustalw 之间的接口……259
主要参考文献……261
英汉对照词汇……262
彩图

第 1 章　序列比对工具 BLAST 和 ClustalX 的使用

骆迎峰　程尹　陈辰　薛庆中

本章主要介绍序列比对工具 BLAST 和 ClustalX 的概念和使用方法。通过美国国家生物技术信息中心（The National Center for Biotechnology Information，NCBI）数据库搜索与输入序列相似的序列，进而通过不同物种多条相似序列的比较，预测基因的功能，探索物种的亲缘关系及其进化。

1.1　序列比对 BLAST

BLAST 是基本局部比对搜索工具（basic local alignment search tool）的缩写。它的功能是对生物不同蛋白质的氨基酸序列或不同基因的 DNA 序列进行比对，在相应数据库中进行序列相似性搜索，寻找相同或相似序列。

NCBI 提供了 BLAST 搜索的在线服务，用户提交核苷酸或蛋白质序列，并选择所要比较的 NCBI 序列数据库，BLAST 搜索程序运行结束后会自动以网页形式返回结果，其中包括比对上的序列、相似性程度及显著性水平等信息。

NCBI 还提供 BLAST 搜索程序和所有 BLAST 序列数据库的下载接口，用户下载相应软件后就可以在本地系统（Windows 或者 Unix）运行 BLAST 搜索。这样更有利于用户根据研究需要，整理出小型化且带有注释的自定义数据库，进行快速搜索，获得相似序列等信息。

1.1.1　Basic BLAST

访问网址 http://www.ncbi.nlm.nih.gov/blast/Blast.cgi，进入 NCBI 网页 BLAST 主界面，它下设 3 个部分（图 1.1）：BLAST Assembled Genome、Basic BLAST、Specialized BLAST。

BLAST 汇集的基因组（BLAST Assembled Genome）。用户可点击特定物种或所有基因组 BLAST 数据库，并选择一个 BLAST 子程序，对该物种所有形式的数据（基因组、EST、峰图等）进行比对。

Basic BLAST。表 1.1 列出了 BLAST 家族的 5 个子程序及其查询序列、数据库、搜索方法。使用 nucleotide blast 和 protein blast 时，由于查询序列与数据库中序列类型相同，可以直接进行比对；而其余 3 个程序——blastx、tblastn 和 tblastx，在比对前要先经过“翻译”。例如，blastx 需先将查询序列翻译成蛋白质序列，

tblastn 需将核酸数据库中的序列翻译成蛋白质序列，tblastx 则需对查询序列和数据库中的核酸序列都进行翻译。

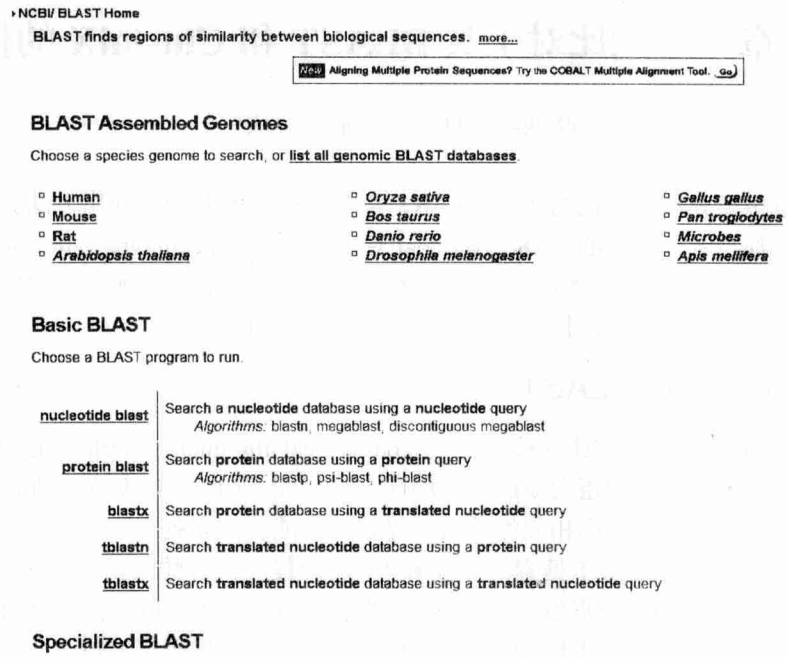


图 1.1　NCBI/BLAST 主界面

表 1.1　BLAST 的 5 个子程序及其搜索方法

程序名称	查询序列	数据库	搜索方法
nucleotide blast	核酸	核酸	用查询核酸序列搜索核酸数据库中的序列。算法：blastn，megablast，discontiguous megablast
protein blast	蛋白质	蛋白质	用查询蛋白质序列搜索蛋白质数据库中的序列。算法：blastp，psi-blast，phi-blast
blastx	核酸（翻译）	蛋白质	用查询核酸序列翻译成蛋白质序列后再对蛋白质数据库中的序列搜索
tblastn	蛋白质	核酸（翻译）	用查询蛋白质序列和核酸数据库中的核酸序列翻译后的蛋白质序列比对
tblastx	核酸（翻译）	核酸（翻译）	用查询核酸序列翻译成蛋白质序列，再和核酸数据库中的核酸序列 6 个框翻译成的蛋白质序列比对

blastn 和 blastp 是最常用的 BLAST 的子程序。前者用于发现高分值匹配的核酸序列；后者能发现氨基酸残基的相似性和寻找同源蛋白。这两种方法搜索较为简便，只需将查询序列粘贴到搜索框中，点击 BLAST 即可完成。其余三个子程

序操作较为复杂。

现以 blastx 为例说明（图 1.2）核苷酸序列翻译后可能形成的 6 种蛋白质序列（分别从查询序列的正向链或反向互补链的 1、2、3 位起始的可读框）。

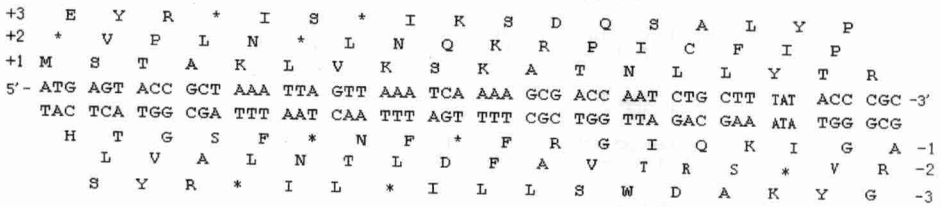


图 1.2　核苷酸序列双链上的 6 个理论可读框

如图 1.2 所示，目标序列为 ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC，可能产生以下 6 个不同的可读框：

正向链（5′→3′端）

（1）第一位起始：ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC

（2）第二位起始：TG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC

（3）第三位起始：G AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC

反向链（3′→5′端）

（4）第一位起始：GCG GGT ATA AAG CAG ATT GGT CGC TTT TGA TTT AAC TAA TTT AGC GGT ACT CAT

（5）第二位起始：CG GGT ATA AAG CAG ATT GGT CGC TTT TGA TTT AAC TAA TTT AGC GGT ACT CAT

（6）第三位起始：G GGT ATA AAG CAG ATT GGT CGC TTT TGA TTT AAC TAA TTT AGC GGT ACT CAT

上述目标序列翻译后会产生相应的 6 个不同相位（phase）蛋白序列：

（1）M S T A K L V K S K A T N L L Y T R

（2）— V P L N — L N Q K R P I C F I P

（3）E Y R — I S — I K S D Q S A L Y P

（4）A G I K Q I G R F — F N — F S G T H

（5）R V — S R L V A F D L T N L A V L

（6）G Y K A D W S L L I — L I — R Y S

结果如图 1.3 所示（注：“—”为终止子）。