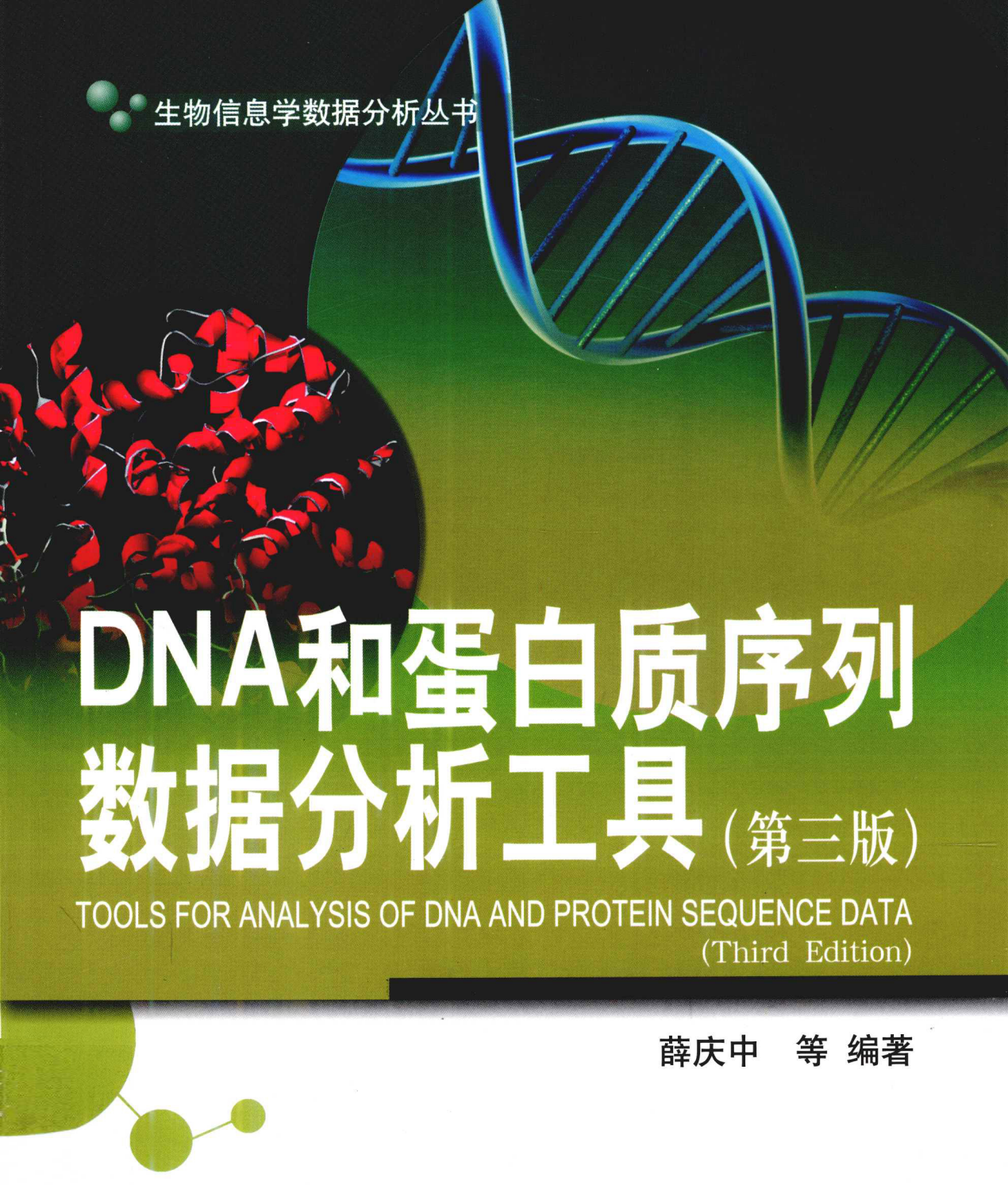

生物信息学数据分析丛书

DNA和蛋白质序列数据分析工具（第三版）

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

薛庆中　等 编著

科学出版社

生物信息学数据分析丛书

Tools for Analysis of DNA and Protein Sequence Data

（Third Edition）

DNA 和蛋白质序列数据分析工具

（第三版）

薛庆中 等 编著

科学出版社

北京

内 容 简 介

近年来新一代测序技术的研发和应用，极大地推动了基因组科学的发展，也给基因组数据分析带来巨大的新挑战。第三版对前两版原有内容做了大量更新和补充，全书17章，分别从基因组学、蛋白质组学、系统生物学三个层次详细介绍了常用的基因数据库和网络工具；为适应Windows7的环境，将BioPerl程序包的数据分析做了重排使其更易操作。尤其是增添了新一代测序数据分析实例，包括SNVs和Indel识别、小RNA-seq分析、枯草杆菌全基因组序列拼接；并对Bowtie等读序列定位工具和UCSC浏览器的使用做介绍。

本书内容深入浅出、图文并茂。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。可作为对基因组学、蛋白质组学、生物信息学感兴趣的本科生、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. —3版. —北京：科学出版社，2012
（生物信息学数据分析丛书）
ISBN 978-7-03-034509-7

Ⅰ. ①D… 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（2012）第114782号

责任编辑：李 悦 刘 晶/责任校对：林青梅
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

北京佳艺恒彩印刷有限公司 印刷
科学出版社编务公司排版制作
科学出版社发行 各地新华书店经销

*

2012年6月第 一 版 开本：B5（720×1000）
2012年6月第一次印刷 印张：23 1/4 插页：6
字数：452 000

定价：75.00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　胡望雄　华大颂　黄鹏宇　蒋　琰

蒋华蔚　刘　杰　骆迎峰　莫　凡　阮　陟

单　杲　王　珺　王瑞娴　王庭璋　薛庆中

叶　琳　张　婧　张维一　周国艳

第三版前言

十几年前，人类、水稻等模式生物全基因组序列图的建立，开启了基因组学的新纪元。近年来新一代测序（next generation sequencing, NGS）技术的研发和应用，其速度之迅猛是始料未及。现已阐明人类遗传性疾病发生与基因组的编码的错误引发的突变紧密关联，科学家预言通过个人基因组测序将有可能发现、捕捉患者体内的这些突变，从而提出个性化的医疗方案，这将为破解威胁人类生命的重大疾病提供可能。同样，一旦数据库中拥有了生物体全基因组参考序列（RefSeq）信息，再对其不同种或品种的重测序就将变得十分便捷和廉价，这将为生物学家们揭示各种生物体遗传、发育、疾病、进化的机制打开科学之门。

浙江大学和中国科学院北京基因组研究所协作举办的“基因组科学研习班”已走过了十年。来自全国30余省市的学员近4500余人，来浙江大学华家池校区参加了学习和上机培训。十年前，基因组学，这门科学对于许多科学工作者而言还是十分陌生的。如今，这个科学盲点已转化为热点，基因组学的知识和工具正在渗透到各个科研领域中。基因组学的发展与新一代测序仪，如454、illumina、SOLiD的研发紧密相关，它们极大地提升了测序速度，每次测序能产生数百万个读序列（read），但是，这些高通量的读序列每条却只有30~400 bp长，它给数据分析带来了巨大的挑战。为此，近年来，一大批适合新一代测序分析的软件和工具应运而生，使得基因组领域的知识更为丰富、更为多元化。为了与时俱进，我们决定编写《DNA 和蛋白质序列数据分析工具》第三版，在编写过程中，我们不仅注重及时更新基因组数据库和工具版本，同时也增添了新一代测序技术数据分析的内容，期望能给学员和读者一些启示和指导。

第三版的内容已从原有的12章扩充到17章。第1~10章除了对有关工具做了更新外，还增添了一些新软件，例如，第6章MEME程序包的软件已从原有的7个增加到12个；第9章推介了DAVID工具，使用户能专注高通量数据功能的挖掘。第10章中补充了对基因表达数据做一致性分析的BioQuali插件。为适应Windows7的环境，第11章将原来BioPerl程序包的安装和数据分析两章内容合为一章，并做了修整使其更易操作。第12章和第13章分别简介了读序列定位工具和UCSC浏览器的使用。第14~17章是我们应用新一代基因组测序数据所做的初步分析，包括SNVs和Indel识别、小RNA-seq分析、枯草杆菌全基因组序列拼接。

新一代基因组测序数据分析内容为本书新版增色不少，我们特别感谢王庭

璋、单杲、莫凡的辛勤劳作，他们刻苦专研、不断修正程序，直至较好地完成数据分析。我们也吸纳了蒋华蔚、张婧、王瑞娴、周国艳、刘杰、胡望雄等研究生参与写作。科学出版社李悦、罗静等编辑对本书样稿做了认真审阅和校对，使稿件在标准化上有了明显的提升。本书的出版得到了浙江大学浙江加州国际纳米技术研究院领导的关心和支持。此外，徐建红、陈爱华、刘秋香、王斌、周忠静、张好富为此书的出版做了很多具体工作，在此一并致谢。如上所述，新一代基因组测序数据分析面临着巨大的挑战，我们的知识水平远远赶不上基因组科学的发展。我们改版的过程也是在不断学习的过程，书中错误、不妥之处，望请读者给予批评指正。

薛庆中

2012 年 3 月于浙江大学

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，这对我们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。

全书共 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 1 章　序列比对工具 BLAST 和 ClustalX

骆迎峰　丁文超　程尹　陈辰　薛庆中

序列比对是基因组学研究的核心手段之一，从测序拼接到基因表达分析都需要将未知序列和数据库中的已知序列进行相似性比较。序列比对工具很多，其中以基本局部比对搜索工具（BLAST，basic local alignment search tool）最为常用。生物不同基因的 DNA 序列或氨基酸序列通过比对，可以在相应数据库中找到相同或相似序列。本章主要介绍美国国家生物技术信息中心（The National Center for Biotechnology Information，NCBI）数据库提供的 BLAST 搜索在线服务及本地运行程序，用户可以通过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，进行序列相似性（Sequence similarity）搜索。本章还将介绍多序列比对工具 ClustalX 的使用方法，以便预测基因的功能，探索物种的亲缘关系及其进化。

1.1　BLAST 搜索程序

NCBI 的 BLAST 搜索程序（http://www.ncbi.nlm.nih.gov/blast/Blast.cgi）下设 3 个部分（图 1.1）：用 BLAST 拼接的参考基因组（BLAST Assembled RefSeq Genome）、基础的 BLAST（Basic BLAST）、特殊的 BLAST（Specialized BLAST）。

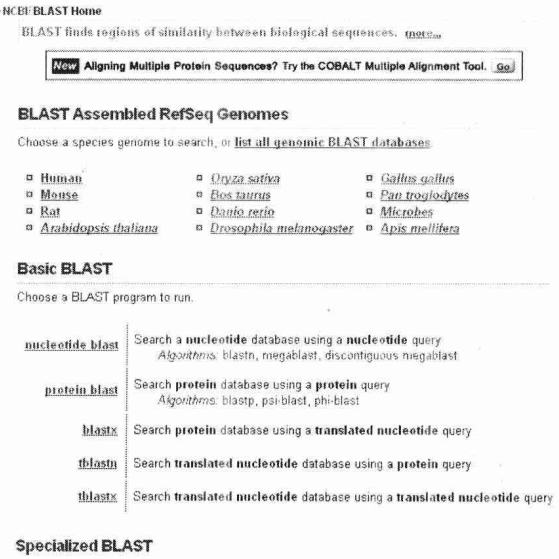

图 1.1　NCBI/BLAST 主界面

1.1.1 用 BLAST 拼接的参考基因组

在做 BLAST 搜索前，用户可根据自己的需求，选择与某个特定物种（special genome）基因组数据库或所有拼接的基因组参考序列数据库 BLAST。如选择后者，点击 list all genomic BLAST databases 后，从图 1.2 可知目前正在测序或已完成测序的物种及其数量，包括：脊椎动物（Vertebrates）26 种、无脊椎动物（Invertebrates）16 种、原生动物（Protozoa）18 种、植物（Plants）47 种、真菌（Fungi）17 种。

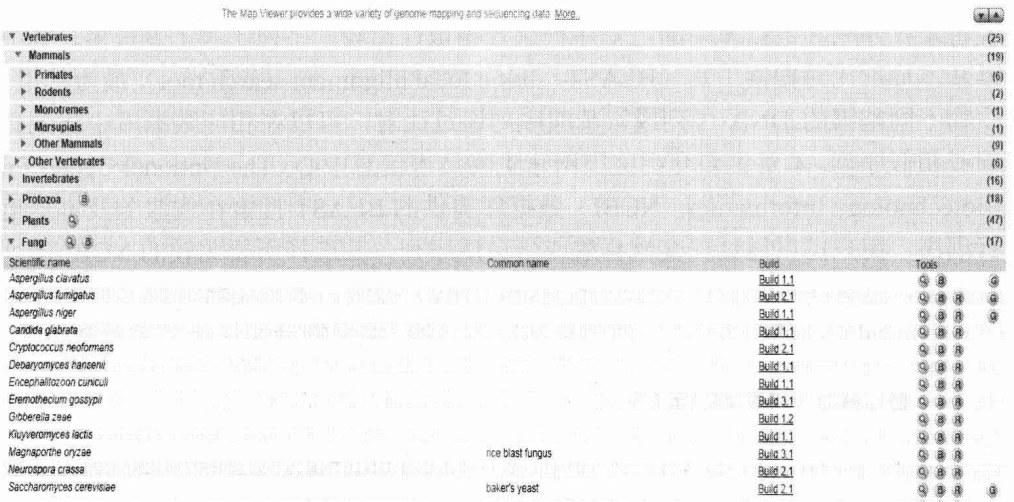

图 1.2 基因组参考序列数据库

1.1.2 基础的 BLAST

确定了相应的数据库，接下来是选择搜索方法。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。子程序 nucleotide blast（blastn）和 protein blast（blastp）最为常用，使用也较简便，可以直接进行比对，搜索时只需将查询序列粘贴到搜索框中，点击 BLAST 即可完成。其中，blastn 用来发现高分值匹配的核酸序列，而 blastp 能发现氨基酸残基的相似性和找到其同源蛋白。

与前两个子程序相比，后三个子程序（blastx、tblastn 和 tblastx）搜索过程较为复杂，在比对前需要先经过“翻译”。例如，运行 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，而 tblastx 需对查询序列和数据库中的核酸序列都进行翻译。现以 blastx 为例（图 1.3），说明核苷酸序列翻译后可能生成 6 种蛋白质序列。

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

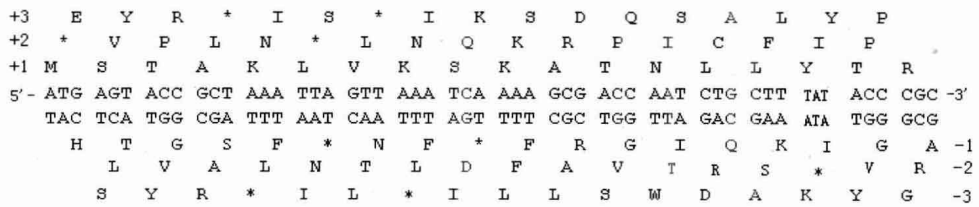

图 1.3　核苷酸序列翻译后可能生成 6 种蛋白质序列

假设目标序列为 ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC，理论上此核苷酸序列翻译时，可以分别从查询序列的正向链或反向互补链的 1、2、3 相位起始。

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