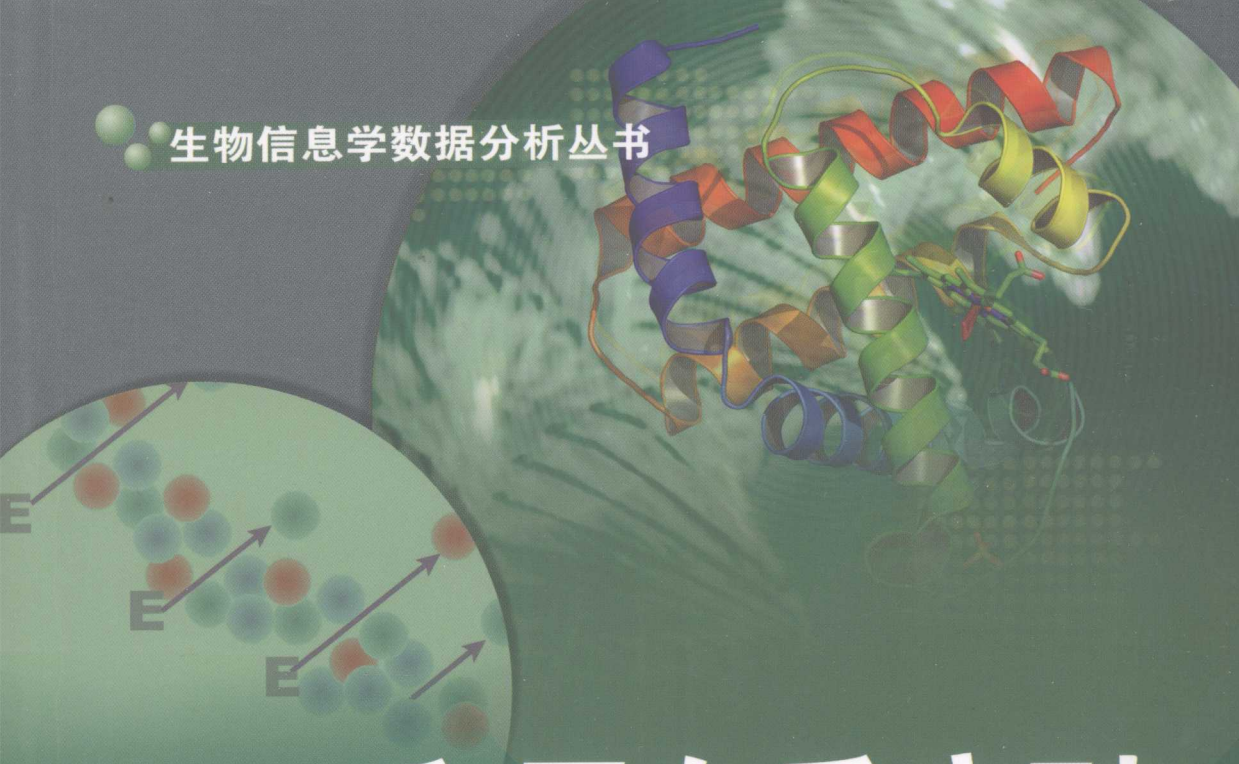
生物信息学数据分析丛书



DNA和蛋白质序列数据分析工具

TOOLS FOR ANALYSIS OF DNA AND PROTEIN SEQUENCE DATA

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

DNA 和蛋白质序列数据分析工具

薛庆中 等 编著

科学出版社

北京

内 容 简 介

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。书后附有专业词中英文对照。

本书是为从事生物学、医学、农学及计算机科学等领域研究的研究生、大学生、教师、医生、研究人员、计算机工作人员提供的通俗易学的手册和工具。

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. —北京：科学出版社，2009

ISBN 978-7-03-022633-4

（生物信息学数据分析丛书）

I.DNA… II.薛… III.①脱氧核糖核酸–研究②蛋白质–研究 IV.Q5

中国版本图书馆 CIP 数据核字（2008）第 123483 号

责任编辑：李 悦 席 慧/责任校对：陈玉凤

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

铭浩彩色印装有限公司印刷

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)

2009 年 1 月第一次印刷 印张：13 1/2 插页：4

印数：1—3 000 字数：258 000

定价：48.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

编委会名单

（按汉语拼音排序）

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

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（她）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

此为试读，需要完整PDF请访问：www.ertongbook.com

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做出贡献。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。对陈爱华在对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 的使用

骆迎峰　程尹　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。通过输入序列从美国国家生物技术信息中心（NCBI）数据库中发现其相似的序列；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。并从相应数据库中找到相同或相似序列。

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，其中包括匹配的序列、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　网上运行 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 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；表 1.1 列出了 BLAST 家族的 5 个子程序的查询序列、数据库及搜索方法。

其中使用 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 将核酸数据库中的序列翻译成蛋白质序列，tblastx 则对查询序列和数据库中的核酸序列都进行翻译。

此为试读，需要完整PDF请访问：www.ertongbook.com

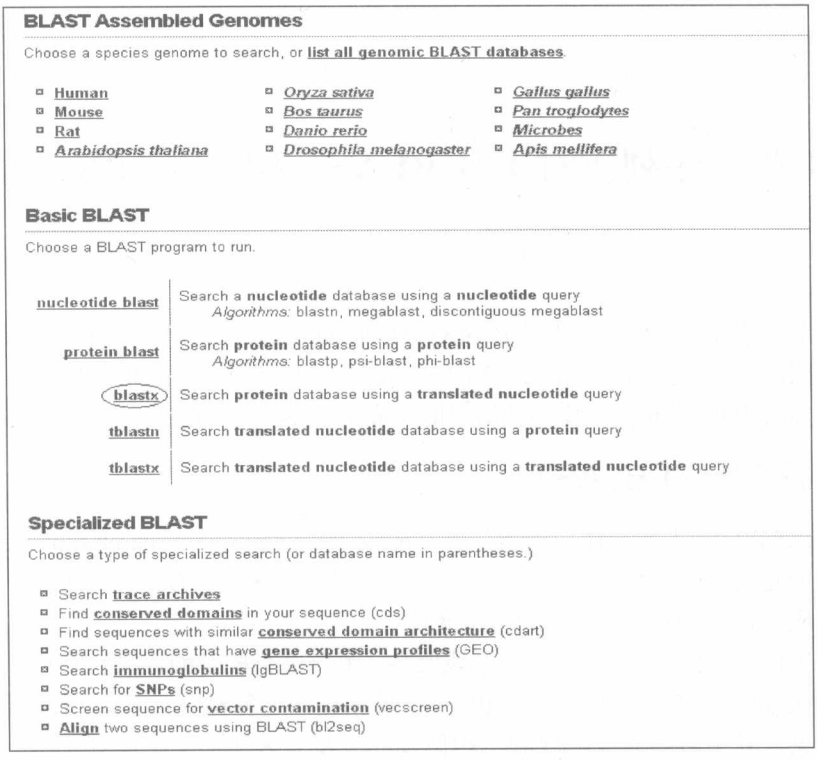


图 1.1　网上 BLAST 运行主界面

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 框翻译成的蛋白质序列比对

此外，还有特殊比对（Specialized BLAST）程序，如 cds，发现序列的保守域；GEO，搜索基因表达序列谱序列；lg BLAST 搜索免疫球蛋白（immunoglobulin）；bl2seq，做两条序列比对；vecscreen，筛选载体污染序列等。

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）。

```
+3   E   Y   R   *   I   S   *   I   K   S   D   Q   S   A   L   Y   P
+2   *   V   P   L   N   *   L   N   Q   K   R   P   I   C   F   I   P
+1 M   S   T   A   K   L   V   K   S   K   A   T   N   L   L   Y   T   R
5'- ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC -3'
    TAC TCA TGG CGA TTT AAT CAA TTT AGT TTT CGC TGG TTA GAC GAA ATA TGG GCG
      H   T   G   S   F   *   N   F   *   F   R   G   I   Q   K   I   G   A -1
        L   V   A   L   N   T   L   D   F   A   V   T   R   S   *   V   R  -2
          S   Y   R   *   I   L   *   I   L   L   S   W   D   A   K   Y   G  -3
```

图 1.2　核苷酸序列双链上的 6 个理论可读框

如图 1.2，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 个可读框

（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

（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

本例中，目标序列翻译所得 6 条蛋白序列分别为（注：“—”为终止子。）：

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。

因此，blastx 程序需要将翻译后的 6 条氨基酸序列在蛋白质数据库中进行逐一比对，得到最佳比对的翻译序列，将此序列视为生物体中最有可能表达的目标

图 1.3　目标序列 6 个可读框及翻译结果

核苷酸序列。本例 blastx 结果显示，最佳比对结果为正向第一位起始的翻译，MSTAKLVKSKATNLLYTR（图 1.4）。

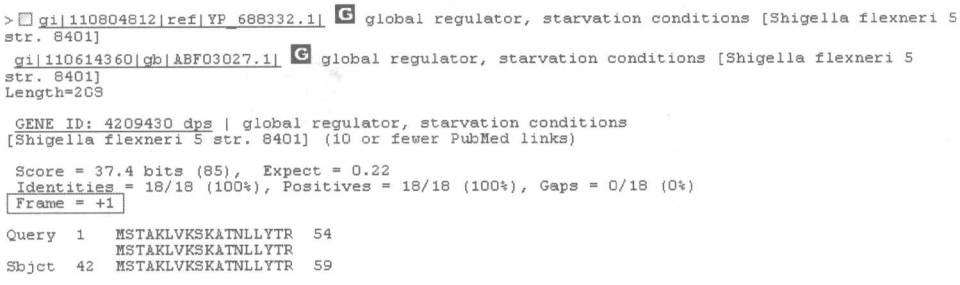


图 1.4　blastx 比对最优结果

1.1.1.1　网上运行 blastx

在 BLAST 主界面点击“blastx”（图 1.1），进入序列提交界面。它由查询序列输入框、基本搜索设置选项（图 1.5）和比对参数设置选项（图 1.6）三部分组成。

网上运行 BLAST 时有两种序列提交方式：用户可以直接把 NCBI 数据库 GI 号、纯序列或 fasta 格式序列粘贴到序列提交框中；也可以将多条序列以 fasta 格

式保存到本地文件中，然后点击“浏览”按钮上传该文件。最新的网上运行 BLAST 服务可同时比对多条序列，若以序列形式提交数据，序列必须是 fasta 格式，而以 GI 号形式提交数据时，则必须采用一行一个 GI 号一行的格式。

“基本搜索设置选项”为用户设置搜索序列及待搜索的数据库（图 1.5）。用户先为 BLAST 搜索任务命名（Job Title），在本例中提交的是单条 fasta 格式序列，默认搜索任务就是该序列名称“lesson.seq.screen.Contig17”，相应地搜索任务名称自动变为“lesson.seq. screen.Contig17”。数据库（Database）默认为非冗余蛋白库（nr），用户在“Organism”选项需选择物种来源，如填入“human”，则搜索过程中就只比对人类数据库的序列。也可以选择密码子表（Genetic code），默认为标准密码子。还可以在“Entrez Query”选项中使用布尔表达式（Boolean expression）。基本搜索选项设置完成后，点击“BLAST”即可运行搜索。

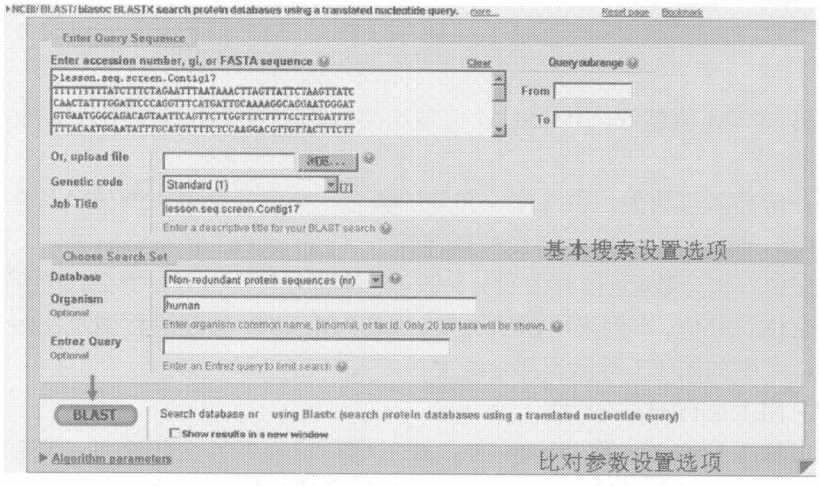


图 1.5　blastx 序列提交界面中的基本搜索设置选项

比对参数设置包括最多目标序列条数（Max target）和期望阈值（Expect threshold，E 值），打分矩阵（Matrix），空位罚分（Gap costs），过滤低复杂性区域（Filter），屏蔽（Mask）等选项。E 值表示在数据库搜索时与期望值随机匹配的可能性，E=1 表示匹配是随机产生的；反之，E=0 表示匹配不是随机产生的，因此 E 值越小置信度越高（图 1.6）。

对蛋白质序列相似性计分通常采用突变数据（Mutation data，MD）和 BLOSUM 两种矩阵。突变数据基于可接受点突变（PAM，Point Accepted Mutation）值。PAM1 表示一个进化变异单位，即有 1%的氨基酸变异。常用的矩阵 PAM250 相似性计分值相当于两个序列间保留 20%匹配。在测定远距离序列相关性时可采用 BLOSUM 矩阵，BLOSUM 值表示相同序列的百分比（如常用的 BLOSUM62 表示比对结果中至少有 62%的氨基酸相同）。

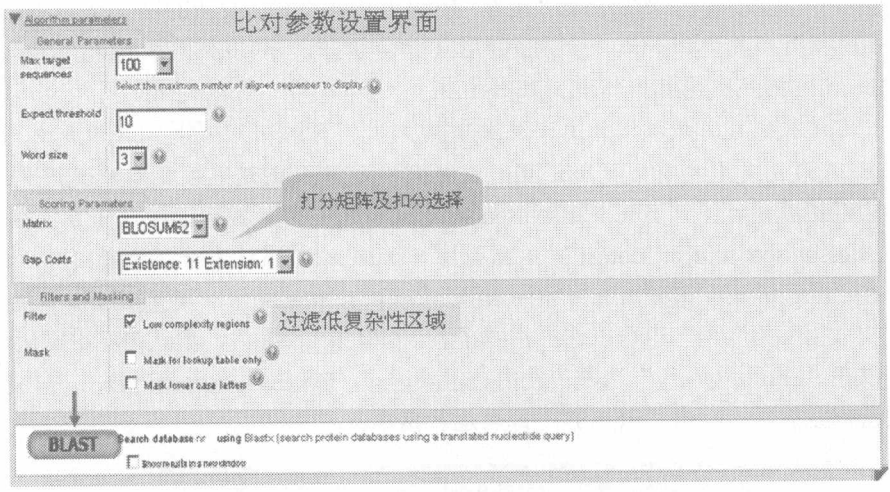


图 1.6　blastx 序列提交界面中的比对参数设置选项

一般而言，PAM 值越低，BLOSUM 值越高的矩阵适用于相似性较高序列的比对；反之亦然。

为补偿插入与缺失，引入空位罚分的概念。空位罚分包含两部分：空位开放罚分（gap opening penalty）和空位延伸罚分（gap extension penalty）。一个长度为 n 的空位，扣分数=空位开放罚分+空位延伸罚分*（n–1）。每个打分矩阵都有默认的空位罚分值。本例采用默认选项，点击“BLAST”按钮进行网上比对。blastx 运行后，产生如图 1.7 所示 4 部分结果。

1.1.1.2　网上运行 PSI-Blast

蛋白质序列比对，除 blastp 外，还可选择位点特异迭代比对（PSI-Blast，position-specific iterated blast），该程序是在蛋白质数据库中搜索查询蛋白质，所有前一次被 PSI-Blast 发现的统计显著蛋白质信息将整合成新打分矩阵，进行二次比对；该步骤可以根据需要循环操作，直到不再发现统计显著的新蛋白质。这种相似蛋白质搜索程序敏感性很高。此外，还可选择模式发现迭代比对（PHI-Blast，pattern-hit initiated blast），用户输入感兴趣的蛋白质序列模式，程序就在相应数据库中查找含有该模式的序列。

如果想在不同物种中寻找同源蛋白，就可以选择位点特异迭代比对。现以果蝇线粒体 RNA 聚合酶为搜索序列，说明 PSI-Blast 的使用方法。

首先在输入框内输入果蝇线粒体 RNA 聚合酶蛋白序列号（AAF51421），选择子程序“PSI-Blast”，采用默认的非冗余蛋白数据库（nr），并在参数设置选项限定结果输出 E 值（1e-10），点击“BLAST”按钮进行数据搜索（图 1.8）。

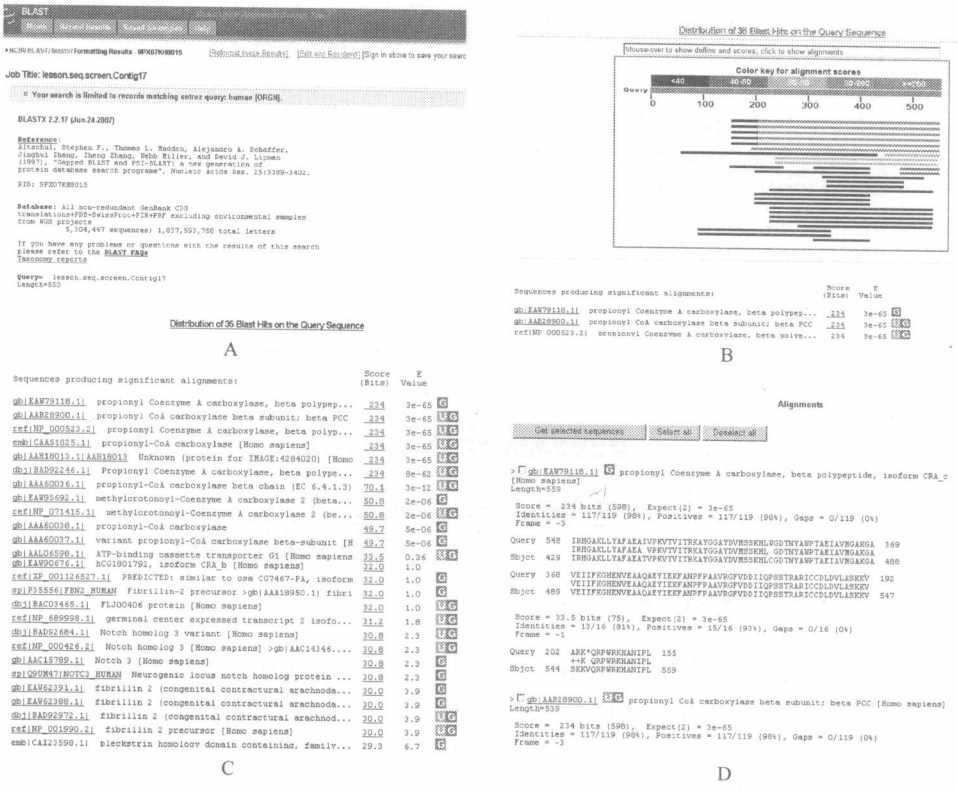
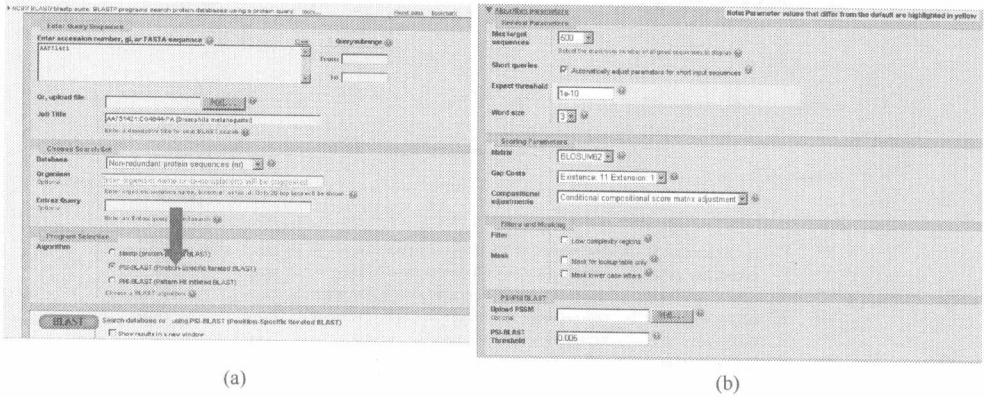


图 1.7　blastx 的输出结果界面

A. BLAST 程序的参考文献，搜索任务名称，提交序列名称和长度，搜索数据库名称及序列条目信息，并对数据库中来自“human”的序列进行搜索；B. 展示比对图形，包括匹配的序列数量，相对位置及其得分高低，深色分值最低，浅色分值最高；C. 列表显示匹配序列的数据库号，序列名称，得分（Score）情况及期望值（E Value）；D. 提交序列及其匹配序列两两比对的详细信息，包括得分（Score）、期望值（Expect）、一致性（Identities）、读框（Frame）等



(a)　(b)

图 1.8　PSI-Blast 检索输入及参数设置

此为试读，需要完整PDF请访问：www.ertongbook.com