Bioinformatics Methods and Protocols

生命科学实验指南系列

生物信息学方法指南

〔加〕S. 米塞诺　〔美〕S.A. 克拉维茨　著
欧阳红生　阮承迈　李慎涛　等译

科学出版社
www.sciencep.com

生命科学实验指南系列

生物信息学方法指南

〔加〕S. 米塞诺　〔美〕S. A. 克拉维茨　著

欧阳红生　阮承迈　李慎涛等　译

科学出版社

北京

图字：01-2003-0476

内 容 简 介

计算机在分析生物学日益增长的海量数据方面起到了不可估量的作用，并推进了现代生物学的快速发展。本书详细介绍了一些重要生物学软件和数据库的使用，同时提供了一些实用的技巧和最新研究进展。全书分为五部分，包括序列分析软件包、分子生物学软件、网络信息资源、计算机和分子生物学的关系、生物信息学教学与最新文献跟踪。内容全面，实用性较强，可帮助生物信息学人员对该学科有更深入地了解。

本书可作为大专院校、科研机构的分子生物学、生物信息学等相关专业的研究生、科研和教学人员的参考书。

图书在版编目(CIP)数据

生物信息学方法指南/欧阳红生，阮承迈，李慎涛等译. —北京：科学出版社，2005
(生命科学实验指南系列)
ISBN 7-03-014465-1

I. 生… II. ①欧… ②阮… ③李… III. 生物信息论 IV. Q811.4

中国版本图书馆 CIP 数据核字(2004)第 114450 号

责任编辑：马学海 庞在堂 彭克里 席 慧/责任校对：朱光光
责任印制：钱玉芬/封面设计：王 浩

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
源海印刷有限责任公司印刷
排版制作：科学出版社编务公司
科学出版社发行 各地新华书店经销
*
2005 年 2 月第 一 版 开本：B5 (720×1000)
2005 年 2 月第一次印刷 印张：26
印数：1—3 000 字数：517 000
定价：58.00 元
(如有印装质量问题，我社负责调换〈环伟〉)

前　言

计算机已成为现代生物学的一个基本组成部分，它帮助管理大量的并日益增长的生物学资料，并在发现新的生物学关系方面继续起着综合的作用，这种基于计算机的生物学方法帮助重塑了现代生物学。我们正置身于生物学革命中，每个科学家都必须提高和训练当今的生物信息学技能，但达到初级水平即可。《生物信息学方法指南》就是要满足这种挑战，并向富有经验的用户提供实用的技巧和目前最新进展的概观。本书以 1994 年出版的两卷套的《序列数据的计算机分析》为基础编著。我们把《生物信息学方法指南》分成五部分，包括在大多数机构都可得到的基本序列分析软件包的全面综述，也有基础入门级的生物信息学课程的设计和内容。此外，本书还介绍了非商业化的专业软件、数据库及因特网上的其他资源，以及针对生物学家现在面临的计算挑战及将来可能的解决方案的讨论。

第一部分，序列分析软件包，介绍目前可得到的一些分析软件包的资源指南，包括在大多数机构可见到的客户服务器 GCG 软件包，几个基于 PC 机和 Mac 机(Macintosh)的适于独立计算的软件包。Staden 软件包也很有特色，因为它是最广泛使用的整套序列分析和装配的软件工具，且针对学术研究可以免费得到。这里也介绍了免费软件的使用，这些软件用于建立一些解决特定需求的分析方案。

第二部分，分子生物学软件，收集了用于完成一些基本生物信息学任务的软件资源。本部分从目前各种计算机平台可得到的免费软件的概观开始，接着是一些特定的例子，其中有用 FASTA、CLUSTAL 多重序列比对进行的序列相似性搜索和种系发生分析。之后，讨论了 Genotator，一个功能非常强大的序列注释和展示套件，它整合了多种不同分析输出适于出版的格式。本部分最后讨论了常见图像分析技术。

第三部分，网络信息资源，主要介绍了基本的一级(primary)序列数据库和各种可得到的分析工具。本部分也有对临床资源进行的独特描述，临床资源正快速成为分子医学新兴领域的整体部分。一级序列分析方法有用 MatInspector 鉴定转录控制区的手段，也有对目前基因鉴定方法的评述。最后讨论了寡聚核苷酸和 PCR 引物的设计及通过万维网分配实验方法和试剂的非常实用的模式。

第四部分，计算机和分子生物学，作者直接阐述了基于计算机分析的局限性和可能存在的解决办法。本部分最后提出了仍然不能回答的问题，即怎样从 A、C、G 和 T 序列串中检测有生物学意义的模式。

生物信息学教学很快成为大多数大学核心课程的组成部分。本书第五部分，作者推荐了生物信息学入门课程的设计。本部分深入分析了如何跟踪日益增加的

文献。简而言之，为在现代生物学成功生涯中日常遇到的问题提供了基本的、实用的答案。

生物信息学为目前的现代生物学革命成为可能提供了帮助。只有了解并明智地使用这些资源，我们才能向前推进。在本书中，对每个学科领域广泛的了解主要是为了帮助那些刚刚开始用计算工具诠释生物学问题的人把握方向。我们相信，在这个独特的软件集和解释例子的指导下，即使初学者也能很快应付每个计算问题，并获得满意的结果。

Stephen A. Krawetz
Stephen Misener

(欧阳红生　译)

编 写 成 员

Ashok Aiyar • *University of Wisconsin-Madison, Madison, WI*
Roger Anderson • *Anderson Unicom Group, Inc., Yorba Linda, CA*
Kathryn F. Beal • *MRC Laboratory of Molecular Biology,Cambridge, UK*
James K. Bonfield • *MRC Laboratory of Molecular Biology, Cambridge, UK*
Timothy G. Burland • *DNASTAR, Madison, WI*
Brian Fristensky • *University of Manitoba, Winnipeg, Manitoba, Canada*
Don Gilbert • *Indiana University, Bloomington, IN*
Nomi L. Harris • *Lawrence Berkeley National Laboratory, Berkeley, CA*
Jack P. Jenuth • *Base4 Bioinformatics, Mississauga, Ontario, Canada*
Lila Kari • *University of Western Ontario, London, Ontario, Canada*
Jeffrey A. Kramer • *Monsanto Life Science Company, St. Louis, MO*
Stephen A. Krawetz • *Wayne State University School of Medicine, Detroit, MI*
Maryann Labant • *Anderson Unicom Group, Inc., Yorba Linda, CA*
Laura F. Landweber • *Princeton University, Princeton, NJ*
Avi Orr-Urtreger • *Genetic Institute, Tel Aviv, Israel*
William R. Pearson • *University of Virginia, Charlottesville, VA*
Promila A. Rastogi • *Oxford Molecular Group, Campbell, CA*
Keir Reavie • *Wayne State University, Detroit, MI*
Jeffry A. Reidler • *Scion Corporation, Frederick MD*
Jacques D. Retief • *University of Virginia, Charlottesville, VA*
Patricia Rodriguez-Tomé • *EMBL European Bioinformatics Institute, Hinxton, Cambridge, UK*
Steve Rozen • *Whitehead Institute for Biomedical Research, Cambridge, MA*
Helen Skaletsky • *Whitehead Institute for Biomedical Research, Cambridge, MA*
Gautam B. Singh • *Oakland University, Rochester, MI*
Rodger Staden • *MRC Laboratory of Molecular Biology, Cambridge, UK*
Paul Stothard • *University of Alberta, Edmonton, Alberta, Canada*
Gary H. Van Domselaar • *University of Alberta, Edmonton, Alberta, Canada*
Thomas Werner • *Institute of Mammalian Genetics, Neuherberg, Germany*
David S. Wishart • *University of Alberta, Edmonton, Alberta, Canada*
David D. Womble • *Wayne State University School of Medicine, Detroit, MI*
Yuval Yaron • *Genetic Institute, Tel Aviv, Israel*

目　录

前言
编写成员
第一部分　序列分析软件包 …… 1
1　GCG：序列分析程序威斯康星软件包 …… 3
2　GCG 序列分析程序基于网页的界面 …… 19
3　Omiga：一种基于 PC 机的序列分析工具 …… 26
4　MacVector：Macintosh 计算机集成序列分析软件 …… 38
5　DNASTAR 的 Lasergene 序列分析软件 …… 56
6　PepTool™ 和 GeneTool™：非平台依赖性的生物序列分析工具 …… 74
7　Staden 软件包，1998 …… 91
8　利用免费软件建立多用户序列分析系统 …… 104
第二部分　分子生物学软件 …… 117
9　Macintosh 和 MS Windows 计算机分子生物学方面的免费软件 …… 119
10　用 FASTA3 程序软件包进行灵活的序列相似性搜索 …… 158
11　采用 CLUSTAL W 和 CLUSTAL X 进行多序列比对 …… 185
12　用 PHYLIP 进行系统发生学分析 …… 204
13　使用 Genotator 注释序列数据 …… 218
14　低价位的凝胶分析系统 …… 233
第三部分　网络信息资源 …… 243
15　供临床遗传学者和分子遗传学者使用的计算机资源 …… 245
16　NCBI 网页上的公用工具和资源 …… 253
17　EBI 上的资源 …… 264
18　计算机辅助分析转录调控区域：MatInspector 和其他程序 …… 284
19　计算机辅助的基因鉴定 …… 294
20　万维网上适用于一般用户和生物学工作者的 Primer3 程序 …… 306
21　利用万维网装备分子生物学实验室 …… 327
第四部分　计算机和分子生物学：信息发布与限制 …… 337
22　网络计算 …… 339
23　利用 DNA 进行计算 …… 349
24　检测生物模式：整合数据库、模型和算法 …… 363

第五部分　生物信息学教学与最新文献跟踪 …… 375
25　分子生物学和遗传学的计算机应用入门课程的设计与实施 …… 377
26　虚拟图书馆 I：MEDLINE 搜索 …… 387
27　虚拟图书馆 II：科学引文索引和更新通告服务 …… 395
28　虚拟图书馆 III：电子期刊、赠款、基金资助信息 …… 402

第一部分　序列分析软件包

1 GCG：序列分析程序威斯康星软件包

David D. Womble

1.1 引言

GCG 程序，又称为“威斯康星软件包”，是具有强大功能的操作、分析和比较核苷酸和蛋白质序列的整套软件工具[1]。GCG 是遗传学计算机小组(Genetics Computer Group)的缩写，该小组隶属于牛津分子小组(加州坎贝尔)。威斯康星软件包含有 130 多个程序，每个程序都可以作为完成特定任务的工具，例如，翻译核苷酸的编码序列、分析限制酶切位点。大多数 GCG 程序用文件的方式输入数据，并将分析结果写到另一个文件中。很多 GCG 程序的输出文件可作为其他 GCG(或另一些软件包)程序的数据输入文件。很多情况下，复杂的问题需通过连续使用几个 GCG 程序得到解决。

威斯康星软件包通常安装在网络上的共享计算机上，如安装在含 UNIX 操作系统的服务器上，这样，用户可在远程终端上通过自己的个人计算机或其他终端访问 GCG 程序。有几种不同的方法运行 GCG 程序，软件包中包括了两种方法：一种是命令行界面，它是一种传统方法，用户键入一个 GCG 程序名开始交互式程序应用；另一种是图形用户界面(Graphical User Interface，GUI)，称为 SeqLab。此界面中，用户打开一套 GCG 程序的窗口，采用图形交互式方式选择序列和程序功能。SeqLab 也包括一个功能强大的用不同色彩作标记的图形界面用户序列编辑器。但在每一种界面中，所有程序操作方式都相似。用户一旦熟悉怎样运行软件包中的一个程序，所有的其他程序都能用同样的模式运行。根据作者的经验，刚开始接触 GCG 程序的学生常用易于使用的 SeqLab 图形界面，而有经验的 GCG 用户常用命令行，因为采用命令行运行更快捷，特别是在远程终端上通过网络运行更是如此。这两种界面都能很好运行。最近引入的基于网页的界面，称为 SeqWeb，也可以从 GCG 得到。这种界面允许用户通过 Netscape Communicator 和 Internet Explore 等网页浏览器运行 GCG 程序和操作序列文件。GCG 软件包基于互联网网页的界面见第 2 章。

1.2 材料

本章描述的方法是基于安装在与 TCP/IP 网络相连的 UNIX 操作系统共享计算机上的第 9.1 版 GCG 程序软件包[2]。该软件包能安装在几种不同的计算机系统上，如运行数字 UNIX4.0 的 Digital Alpha 机、运行 6.2、6.3 或 6.4 版 IRIX 的基于 RISC 的 Silicon Graphics 机和运行 2.51 或 2.6 版 Solaris 的基于 SPARC 的 Sun 机。该软件包也可运行在以 6.2 版 OpenVMS 为运行环境的 Digital Alpha 机上。安装维护含全套数据库的威斯康星软件包至少需要 15G 的硬盘空间。随着数据库的扩展，所需硬盘空间需要快速增加。个人用户文件也需要额外的硬盘空间。建议至少应有 128M 核心内存和 200M 的虚拟内存。程序通常运行于 UNIX 环境中的 C 环境中。软件包可以从遗传学计算机小组得到，它们的地址是：3575 Science Drive, Madison, WI 53711, 电话：(608)231-5200, 传真：(608)231-5202, 电子邮件地址：`info@gcg.com`，网站地址：`http://www.gcg.com`。

GCG 程序能在 UNIX 计算机控制台(console)或远程工作站(即运行 Windows 或 MacOS 个人计算机)上直接操作。如果使用命令行操作程序，应该使用含 VT100 终端仿真器的运行远程登录软件的终端或 PC 机。如果使用 SeqLab，应该使用 X-Windows 终端或运行 X-Windows 服务器软件的个人计算机。

大多数 GCG 程序的结果保存为常见的文本(ASCII)文件。文本文件可以使用任何文本和文字编辑器进行进一步的操作。此外，很多 GCG 程序的结果以图形方式输出，如限制性内切核酸酶图谱、RNA 二级结构预测。图形输出的方式有：在终端屏幕上显示、在与终端相连的打印机或绘图仪上打印或保存成文件用于以后显示或打印。为了在屏幕上显示图形，需要图形终端或仿真器。X-Windows 仿真器可在屏幕上显示图形，也可用 SeqLab 图形界面。GCG 程序包所带的说明指出各种终端和图形软件都适用于本软件包。作者的建议见 1.4 节。

要打印 GCG 图形，PostScript 或 HPGL 图形语言的机器都可使用。为了在用户终端的打印机上直接从远程服务器上打印图形，需要一个使打印处于透明方式(transparent mode)的终端程序，以便文件直接从打印机输出，而无需个人计算机处理。

1.3 方法

1.3.1 程序描述

威斯康星软件包中共有 130 多个程序。尽管各个程序都能作为独立的工具使

用，但为了便于描述，这些程序可根据功能进行分组。本节介绍软件包中一些通用程序的功能，同时对某些程序进行简要描述，并含有一些例子。尽管对 GCG 程序的完整介绍不是本章的范围，但这些例子可提供足够的信息，使读者了解本软件包中的工具箱的总体内容。

1.3.1.1 比较

1) 配对比较

这些程序可以将一个序列与第二个序列进行比较。可选项有生成两个序列最优的全局(global)比对，找到两个序列的最相似的片段(bestfit)，或者形成序列相似性的 X/Y 图(compare/dotplot)。

2) 多重比较

PileUp 程序采用渐进和配对比对(pairwise alignment)的方法对多组相关序列进行多重序列比对分析。本组中的另一些程序(SeqLab)可手动编辑比对的序列，显示比对序列的各种属性或从比对序列生成用于数据库检索的流程。

1.3.1.2 数据库检索

1) 文献检索

LookUp、StringSearch 程序可通过名称、登录号、作者以及其他关键词查找序列。

2) 序列分析

在这些程序组(BLAST、NetBLAST、FAST 等)中的程序可以在数据库中检索与待查序列相似的序列。NetBLAST 可直接检索美国国家生物技术信息中心(NCBI)的数据库，其他程序可检索本地安装的数据库。

1.3.1.3 编辑和发表

该组程序中有编辑单个序列文件的程序(SeqEd)，也有编辑多个序列文件的程序(LineUp、SeqLab)，同时也可对将要发表的序列数据或质粒图谱作准备。

1.3.1.4 进化关系分析

程序 PAUPSearch、PAUPDisplay、Distances、GrowTree 及 Diverge 可以进行多重比对比较，分析序列的相似性和进化关系。

1.3.1.5 片段组装

GCG 片段组装系统是一套将测序项目得到的序列数据组装成连续序列的程序。

1.3.1.6 基因查找和模式识别

此组程序超过 12 个，有 TestCode、Frames、Motifs 等，这些程序可以帮助识别蛋白质的编码区、蛋白质的结合基序、直接的重复、其他模式以及其他类似的任务。

1.3.1.7 输入和导出

该组有 15 个这种程序，可辅助输入序列数据和对各种格式的序列文件进行格式转换，可转换的格式有 GCG、Staden、EMBL、GenBank、IntelliGenetics、PIR 和 FASTA。

1.3.1.8 绘图

绘图程序(Map、MapPlot、MapSort 等)生成和显示限制酶切图、可读框图、肽消化图、T1 核酸酶消化图、质粒图等。

1.3.1.9 引物挑选

Prime 程序挑选用于聚合酶链反应(PCR)实验和 DNA 测序所用的寡聚核苷酸引物。

1.3.1.10 蛋白质分析

蛋白质分析程序(PeptideMap、PepPlot、PeptideStructure 等)能辅助确定蛋白质氨基酸序列有关的信息，如确定等电点、定位功能基序、预测蛋白质二级结构、分析抗原性质和分泌信号。

1.3.1.11 RNA 二级结构

该组程序(Mfold、StemLoop 等)能按 Zuker 法[3]及确定反向重复序列位置的方法预测 RNA 二级结构并以多种格式显示 RNA 二级结构。

1.3.1.12 翻译

翻译程序(Translate、BackTranslate、PepData 等)将核苷酸序列翻译成肽序列或进行相反的工作。

1.3.1.13 工具程序

1) 序列工具

该部分有几个实用程序(Reverse、Shuffle、Simlify 等)，功能有生成反向的核苷酸序列、随机化序列或用 X 字母取代低复杂度区序列。

2) 数据库工具

用这些程序，可以从任何 GCG 格式的序列中生成 GCG 个人数据库，将任何 GCG 序列连接到一个数据库中，所形成的数据库可以被 Blast 检索，也可从序列中随机提取序列片段。

3) 打印和绘图工具

这些程序(Lprint、ListFile 等)用于显示、打印和绘制 GCG 结果文件，将文本文件或图形文件连接到各种显示、打印或绘图装置。更多的显示或打印 GCG 结果文件的信息见 1.3.5 节。

4) 文件和其他小工具

许多其他小工具程序(ChopUp、Replace、Reformat 等)辅助操作文本文件，打印 GCG 文件及其他任务。

1.3.2 数据库

威斯康星软件包中含有一套综合序列数据库。其中有 GenBank 和 EMBL 核酸序列数据库(EMBL 数据库有删节，以避免与 GenBank 重复)、PIR 和 SwissProt 蛋白质序列数据库。数据库中的序列是 GCG 文件格式的，因此它们能直接作为 GCG 程序的输入文件。因为大多数序列存在于数据库中，因此每个用户无须自行搜集这些序列的拷贝，只要查阅数据库中的拷贝就可以搜集到序列数据。还包括了 GCG 程序所用的各种数据库，如限制酶、分值矩阵(scoring matrix)、水解酶及试剂、蛋白质分析数据文件、转录因子数据库(TFD)、密码子使用频度表、翻译表和蛋白质位点和模式字典 PROSITE。这些数据以文本文件形式保存，个人可以按照自己的需要或特定目的进行检索和编辑。

1.3.3 界面

威斯康星软件包有两种界面：命令行界面和名为 SeqLab 的图形用户界面。用命令行界面，用户键入一个 GCG 程序名称就开始一个程序的交互式对话。然后，提示用户运行程序所需的信息，如输入序列文件的名称、让用户从各种备选项的菜单中选择程序怎样操作。在最后一次按回车键后，程序运行。运行后通常将结果存放到一个文件中。所有的 GCG 程序从命令行的操作相似。因此，一旦熟悉一个程序的操作过程，可以用所熟悉的方式操作其他程序。从命令行操作程序也

可脚本化运行，脚本化运行时含命令行开关，这是用多个输入文件多次运行 GCG 程序的高效运行方法。为了从远程终端中使用命令行界面，需要将终端模拟程序经远程登录，与 GCG 服务器相连。终端模拟程序使用 VT100 终端功能。图 1.1 举例示意了命令行界面。

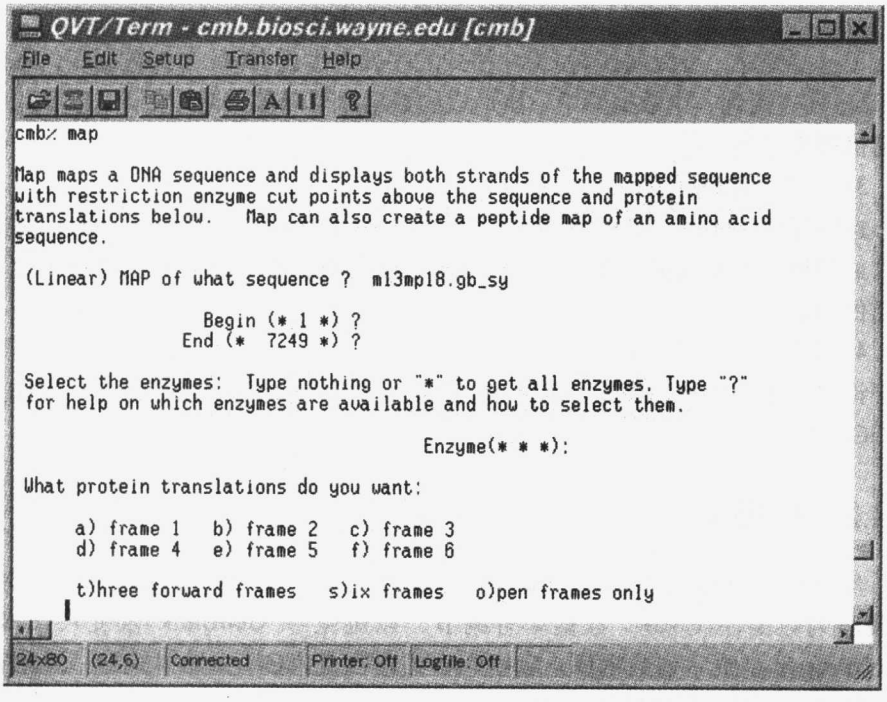

图 1.1 GCG 命令行界面

SeqLab 是 GCG 图形用户界面，它提供了操作威斯康星软件包更方便的使用方法。使用 SeqLab 的下拉菜单可选择程序对序列进行操作。当从下拉菜单中选择一个 GCG 程序时，出现程序专用的一个独立的窗口。然后用鼠标点击选项，可选项有分析哪个序列，接着按 Run 按钮。所选的 GCG 程序的结果列在另一个称为 Output Manager Window(输出管理窗口)中。然而 SeqLab 程序功能超过命令行界面程序，对各碱基或残基或已知的序列性质有更丰富的视觉显示。这种视觉显示使得手动编辑序列或生成和处理多重序列比对更加容易。SeqLab 所用的图形用户界面称为 X-Windows，它是运行 UNIX 操作系统计算机的一种窗口系统。使用 SeqLab，需要一种 X-Windows 显示，如运行在 Windows PC 机或 Mac 机上的 X 服务器程序，或运行 X-Windows 的工作站。图 1.2 显示的是 SeqLab 界面的一个例子。

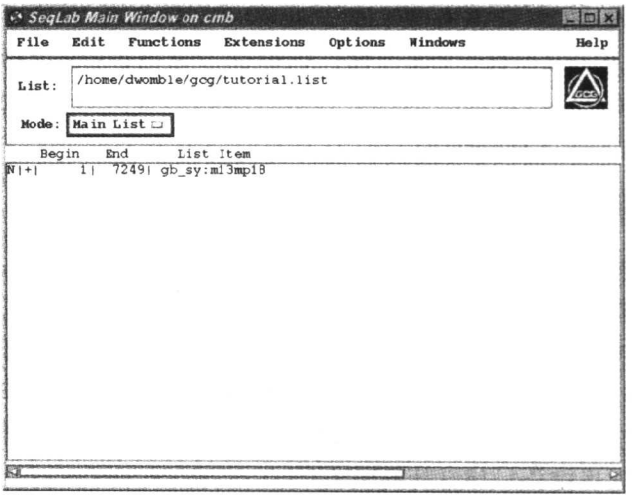

图 1.2　GCG SeqLab X-Windows 界面

一个名为HYGCGmenu 的程序(GCG 超文本菜单)能用于增强命令行界面的功能。HYGCGmenu 生成一套 GCG 程序屏幕菜单，这样可以用“点击”的方法操作 GCG 程序。箭头指针键被用于选择一个序列并启动交互式 GCG 程序对话。在 HYGCGmenu 中，GCG 程序按功能组织成菜单，这样不用记住各个 GCG 程序的名称也能容易选定程序。HYGCGmenu 也有一套目录浏览和文件管理工具，如拷贝、改名、编辑等，以增强其效能。用 HYGCGmenu 的一个好处是在用户端不需要额外的软件，通过远程登录终端的 VT100 模拟器进行操作。HYGCGmenu 不由 GCG 生产，也不由威斯康星软件包提供，但能被各教育用户或系统管理员免费下载(见 1.4 节)。图 1.3 显示的是 HYGCGmenu 的一个例子。

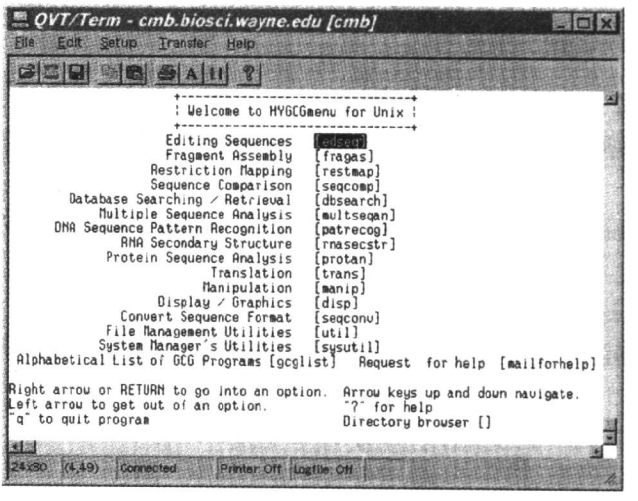

图 1.3　GCG 的超文本菜单 HYGCGmenu