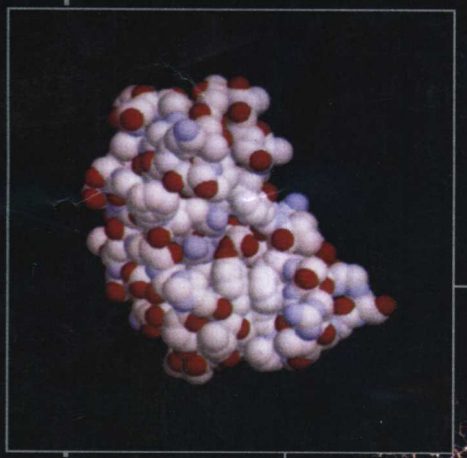
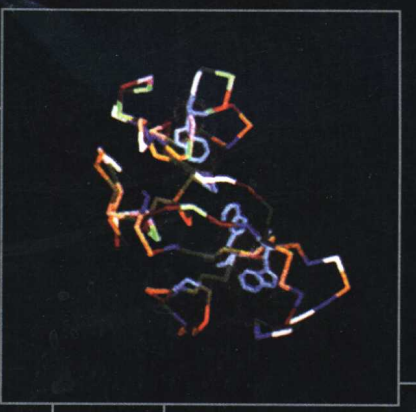
生物信息学

基因和蛋白质分析的实用指南

Andreas D.Baxevanis
B.F.Francis Ouellette 著

李衍达 孙之荣 等 译





清华大学出版社
http://www.tup.tsinghua.edu.cn

生 物 信 息 学

基因和蛋白质分析的实用指南

Andreas D. Baxevanis
B. F. Francis Ouellette
李衍达　孙之荣等　译

清 华 大 学 出 版 社

(京)新登字 158 号

内 容 简 介

人类基因组计划即将完成,这对人类了解生命和人类自身具有非常重要的意义。基因组研究的热点逐渐从基因组测序转向对基因组表达的分析和对蛋白质结构与功能的预测。大量生物分子的数据需要综合运用数学、物理、计算科学和信息科学等进行处理和分析,从而产生了生物信息学这门崭新的交叉学科。

生物信息学集成了分子生物学和计算方法,彻底变革了基因探索及相关的研究,为科学家提供了崭新的工具,可以用来处理由人类基因组计划产生的海量生物数据、原始 DNA 和蛋白质序列信息。

关于生物信息学的研究正深入开展,但是目前国内尚未见到有关生物信息学的图书。1998 年美国国家人类基因组研究所和国家生物技术信息中心的两位教授 Andreas D. Baxevanis 和 B. F. Francis Ouellette 出版了一本生物信息学专著(Bioinformatics—A Practical Guide to the Analysis of Genes and Proteins)。本书是其中译本,是由清华大学生物信息研究所李衍达院士和孙之荣教授组织翻译的。

由前卫计算生物学家撰写的这本书,贯穿了已有的工具和数据库,包括应用软件、因特网资源、向数据库提交 DNA 序列以及进行序列分析和利用核酸序列与蛋白质序列进行预测的方法,使读者从中体验到生物信息学这一崭新的、有待开发的学科的魅力。

读者对象:高等院校的师生和从事生物工程研究的科技工作者。

北京市版权局著作权合同登记号:01-99-2518

书　名:生物信息学基因和蛋白质分析的实用指南
作　者:李衍达　孙之荣 等　译
出版者:清华大学出版社(北京清华大学学研大厦,邮编 100084)
　　http://www.tup.tsinghua.edu.cn
印刷者:北京市清华园胶印厂
发行者:新华书店总店北京发行所
开　本:787×1092 1/16　印张:21.75　彩插:2 页　字数:500 千字
版　次:2000 年 8 月第 1 版　2000 年 8 月第 1 次印刷
书　号:ISBN 7-302-03997-6/Q·10
印　数:0001~5000
定　价:39.00 元

译者序

随着人类基因组计划的实施，通过基因组测序、蛋白质序列测定和结构解析等实验，分子生物学家提供了大量的有关生物分子的原始数据，需要利用现代计算技术对这些原始数据进行收集、整理、管理以便于检索使用，因而出现了生物信息学。为了解释和理解这些数据，还需要对数据进行比对、分析，建立计算模型，进行仿真、预测与验证等，这些也促进了生物信息学的发展。

生物信息学是分子生物学和信息科学与技术、物理、数学等学科交叉、结合的产物，它的出现极大地推动了分子生物学的发展。

现在人类基因组计划接近完成，人们的注意力已从基因组测序转向对基因组表达的分析，对蛋白质组结构与功能的预测。这也是生物信息学面临的主要课题。人们注意到，无论是基因的表达，还是蛋白质的功能，在很多情况下，都是多个基因、多种蛋白质相互作用的结果，要对它进行分析与预测，必然涉及数学、物理、计算科学、系统科学、控制科学、信息科学与生物学的综合应用，因而生物信息学是一门多学科交叉的学科，它需要多个领域的专家通力合作。不仅如此，由于它涉及面是如此广泛，而难度又是如此之大，它更需要全国、全世界的科学家的通力合作。

近年来，由于国际互联网的迅速发展，为这种世界性的合作提供了网络基础，从而大大地促进了世界上生物信息学家之间的交流，他们共享已有的数据、资源，相互交流各自提出的分析方法。相互交流、共同协作，形成了生物信息学界的一股新风气。

人们已经意识到，生物信息学的发展将对我们了解生命，了解人类自身；对于医药、保健、农业等都将起极大的作用，因而生物信息学已引起世界各国的高度重视，纷纷加大投入，发展十分迅速。同样地，生物信息学在我国也正在兴起。生物信息学是一门崭新的学科，目前国内已有大学招收生物信息学的本科生，硕士研究生和博士研究生的培养也在日益加强。但国内尚无一本完好的生物信息学的读本，1998 年美国国家人类基因组研究所和国家生物技术信息中心的两位教授出版的这本生物信息学专著，是一本难得的讲述生物信息学的图书，我们将其介绍给中国读者。我们期望通过本书的翻译出版，对我国生物信息学的发展有所裨益。

本书的翻译工作由几位译者共同完成，具体分工如下：第 1,2 章卢欣；第 3 章蔡军；第 4 章闻芳；第 5 章胡驰峰；第 6 章李萍；第 7,8,9 章季星来；序言，第 10,11 章过涛；第 12 章廖心清；第 13 章罗霄；第 14 章李泽。李衍达和孙之荣负责组织和审校工作。

李衍达　孙之荣

1999 年 11 月

ADB 将本书奉献给他的母亲 Anastasia，并纪念他的父亲 Demetrios，他们的智慧和爱是守护他一生的力量。

BFFO 以本书纪念 Angelos Kalogeropoulos，一位朋友和生物信息学科学家，他的风采令人深深怀念。

序　言

过去 10 年中，全世界的分子生物学家们所收集的原始信息不断激增。在不太久之前，这些信息的分析整理工作只有研究生去做，因为他们对摆弄试管比敲击键盘更有兴趣，而现在有很多人已全身心地投入这个领域。生物信息学正处于萌芽中，它可以不严格地定义为分子生物学和计算生物学的交叉。这个领域中已有了大量重要的发现，并有希望揭示更多大自然的奥秘。对大多数人而言，生物信息学的吸引力在于它是生物学中崭新和有待开垦的领域；而对其他人，其吸引力蕴藏在还原论者对化学层次上的细节的热爱和系统遗传学家对了解各物种体系之间内在关系的兴趣之中。

生物信息学的好处早已作为谈论的话题得以广泛宣扬，它被宣称是能解决一切痼疾的仙丹，或是分解序列数据的强大工具，或简称之为研究科学的一条迷人途径。而实际上，生物信息学是艰难而有意义的工作中的一种新的方法。这一领域中，研究方法大多在不断变化并有待发展和完善，与当年生物化学的黄金年代并无不同，那时人们选择各种能溶解和分析目标分子的手段，而如今生化实验室中所用技术要成熟和精巧得多。然而，在生物信息学被推向前进的竞赛中，一些人曾企图将其从科学分支降级为购买了合适工具包就完成的功能。而维护了生物信息学在科学领域中地位的正是生物信息学研究工作者群体本身，无论他们是在私立大学里还是在政府赞助的研究中心里。生物信息学已取得的卓越进展就蕴含在从收集、整理原始数据，到开发更新、更强的数据处理方法的工作之中，而且一切均处于信息和技术自由共享的环境里。生物信息学群体的独特之处在于，超越商业范畴，其“团体精神”比生物学中许多竞争性领域要开放得多。由此想法，本书试图能让那些想了解更多序列分析方法的科学家跳进书中，来体验令人着迷的科学旅途。通过这本书，我们希望读者能认识到这些方法的严格性，并且明白，与实验并无不同，控制器的运转和弄清哪种方法能解决或不能解决何种问题，是至关重要的。总之，我们鼓励读者不妨一试。

这里，我们要感谢许多同仁，若非他们的帮助本书就难以完成。首先要感谢的是分别完成此书各章节的诸位作者。他们专业的见解与专家的观点，以及他们在十分繁忙中仍友善地配合，使我们感到与这些女士和先生们合作十分愉快。

本书中极大部分内容均得益于在 NIH 的国家生物技术信息中心（NCBI）所开发的工具和数据库，我们要感谢 NCBI 的全体成员，感谢他们辛勤的工作，感谢他们耐心负责地维护着这些公共数据库，特别是他们使这些数据库具有最高的质量和便于科学工作者们访问。那些从事计算生物学的科学家们一贯对他们的成果十分慷慨，制作了自由访问的工具和专门数据库，否则，连最基本的序列分析研究也是无法做到的。我们要感谢参与这些项目的所有成员，包括这里未曾特别提及的，因为是他们造就了这个以序列为基础的新生物学时代的许多辉煌成就。

我们还要感谢本书编辑 Ann Boyle 的耐心帮助、鼓励和支持。这本书也包含了我们的很多第一次，因而在学习出书的里里外外过程中，我们与她建立起了深厚的友谊。我们期

待着未来能与她的再次合作。

以下来自 BFFO:我要向一贯支持、热爱和信任我的妻子 Nancy Ryder 致谢。她对这项工作的尊重,以及她给予了我这项工作所需的空间已超越了任何誓言,我只有以更多的爱作为回报。我还要感谢 Mark Boguski 在 4 年前将我介绍给 NCBI。Mark 还不断引导我留意各种新鲜有趣的项目,这种持续的热情和兴趣将不断赋予我更多灵感。

以下来自 ADB:我诚意地感激 David Landsman 的极大支持和鼓励。我与 David 的交情已有多年,那时他作为一个物理化学家被接受作博士后,而当时他对一门称为生物信息学的领域几乎一无所知。我认为 1992 年的我们都无法猜到我们在教学和研究中的合作最后会导致一本专著的产生。如果不是他在如何认识计算生物学的问题上更加强调生物学对我产生了强烈影响,就不会有这本书。

Andreas D. Baxevanis

B. F. Francis Ouellette

目　录

1

因特网与生物学家

Andreas D. Baxevanis
Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

随着研究者可用的序列与结构信息的爆炸式增长,生物信息学领域,或更确切地说是计算生物学领域,在基础生物医学问题的研究中起着越来越大的作用。计算生物学家面临的挑战,尤其是由人类基因组计划以及其它测序工作生成的大量数据带来的挑战,将对发现基因和设计分子模型、定点突变,以及设计其它有可能发现基因与蛋白质的结构与功能的未知关系的实验有所帮助。

在开始实际讨论解决生物问题的计算方法之前,必须先明确一个共同的背景,从而使用户可以访问和使用本书中讨论的算法和工具。我们首先回顾了因特网及其有关术语,并讨论了4种主要的因特网协议,但不深入涉及协议的技术细节。关于这些协议的内部处理过程的详细描述可以查阅章末的参考文献(Falk,1994;Krol,1994),这些给外行人看的好书。

1.1 因特网基础

尽管"因特网"听起来像一个单独的实体,但它实际上是一个网络的网络,由超过20 000个分布在100多个国家中的相互连接的本地网或地区网构成。虽然有关远程通信的工作在60年代初就已经开始,但因特网的真正起源还是1969年美国国防部远景研究规划局(Advanced Research Projects Agency, ARPA)的网络研究计划ARPANET。ARPANET最初连接了美国西海岸的4个节点,其直接目的是在实验室之间传输有关国防的研究信息。随后又开展了一系列的网络研究项目,并在10年后达到了另一个里程碑式的阶段。1981年引入的BITNET(Because It's Time)在大学之间提供点对点的电子邮件和

文件传输,1982 年 ARPA 引入了传输控制协议(TCP)以及因特网协议(IP),TCP/IP 使得不同的网络可以连接起来并进行通信,从而形成了现在的系统。很多文献详细介绍了因特网的发展历程和通信协议,但大多数用户关心的只是因特网在工作,而非具体的工作原理。

当网络中的计算机连接在一起的时候,需要有一种方法来明确表示每一台计算机,从而使消息和文件真正找到它们的接收者。为此,所有与因特网直接相连的计算机都必须有一个 IP 地址,IP 地址是唯一的,标识一台且只能标识一台计算机。IP 地址由 4 个以点号分隔的数字构成,如美国国立卫生研究院(NIH)生物技术信息中心(NCBI)的主文件服务器的 IP 地址是 130.14.25.1。从左到右这些数字表示的是:主域(130.14 表示 NIH)、子网(.25 表示NIH 的国家药物实验室),以及这台计算机(.1)。虽然采用数字式的 IP 地址可以帮助计算机定位数据,但用户记忆起来却非常困难,所以 IP 地址通常都有相对应的正式域名(FQDN),由域名服务器在后台动态地将其翻译成 IP 地址。回到前面 NCBI 的例子,用户更愿意用 ncbi.nlm.nih.gov 而不是 130.14.25.1 来访问NCBI 的那台计算机。必须注意的是,从左向右 IP 地址的定位范围由大到小,而 FQDN 的定位范围则由小到大。因此,任意指定的计算机的名称都可以看成是具有如下的格式:计算机.域,其中顶级域名(FQDN 中最后一个点号后面的部分)分为 6 个大类(见表 1.1)。在美国之外的国家,顶级域名则是用两个字符表示计算机所在国家(例如,.ca 表示加拿大,.uk 表示联合王国)。

表 1.1 顶级域名表

美国顶级域名	
.com	商业站点
.edu	教育站点
.gov	政府站点
.mil	军方站点
.net	网关或网络主机
.org	私营(通常非赢利)组织
美国以外顶级域名的例子	
.ca	加拿大站点
.ac.uk	联合王国的科学站点
.co.uk	联合王国的商业站点

对因特网规模(即因特网的成功程度)的最具体度量,就是计算物理上接入因特网的计算机的数量。网络 Wizards 通过运行一个探测器去尽可能地寻找主机,并把探测结果返回到运行探测器的计算机上,从而定期地计算这些计算机(或主机)的数量。主机数量的增长速度非常显著,大约每 12 个月增加一倍,目前主机总数已经超过了 1 200万台。这一增长的绝大部分来自商业部门,例如圾万维网等投资于日益大众化的新

多媒体广告与通信平台(图 1.1)。由于可能有许多探测器找不到的主机,这一统计数字的绝对数目不会很精确,它只适用于考察因特网的发展趋势,以及和其它数据进行比较研究。例如:有许多计算机被设置在防火墙后面,出于安全的考虑而阻止了公司内部与外部的通信;其它一些计算机,尤其是家用计算机,只通过调制解调器与因特网短暂相连。所以最好把网络 Wizard 的搜索结果只看成是代表某一时刻因特网的最小规模。

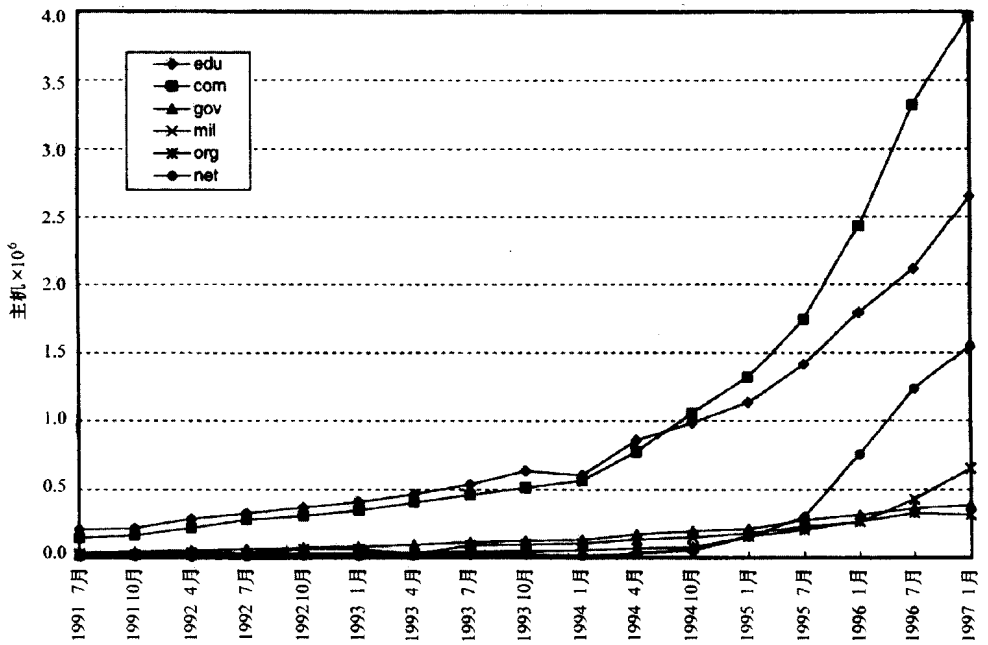


图 1.1 因特网上各域名主机数量的增长。因特网上主机的总数已经超过1 200万台,商业站点(.com)数量在 1994 年首次超过教育站点(.edu)。[Data Network Wizards (*http://www.nw.com*)]

1.2 与因特网连接

那些不能将他们的计算机通过以太网,10BaseT 或类似方式连接到因特网的用户可以有两种主要的方式访问网络:在线服务(online service)或因特网提供商(Internet service provider, ISP)。在线服务,例如美国在线(AOL)、Compuserve 以及 Prodigy,提供了大量的交互式数字服务,包括信息提取、电子邮件(E-mail)、公告牌,以及"聊天室"。在聊天室里同时上线的用户可以就任何话题进行交流。虽然现在在线服务也可以访问万维网,但大多数在这一系统中提供的信息服务依然是通过独占的、封闭的网络进行的。一旦用户计算机和在线服务器之间建立了连接,用户就可以不离开在线系统的主机而访问系统中特定的信息资源。特定的内容可以包括从访问在线旅游预定系统到经常更新的大百科全书,这些项目对于那些没有订购在线服务的用户是无法得到的。服务的价

格随内容而异,大多数这些服务都是按小时记费,即使正常的使用也可能会积累很高的费用。

因特网提供商采用相反的方式。ISP 注重的不是提供内容,而是提供给用户必要的工具以发送和接收邮件,上载和下载文件,以及浏览万维网,发现远程的信息。尽管像 AT&T 和 MCI 这样的大公司占据了 ISP 的主要角色,但不要求提供内容导致了家庭手工业式企业的快速发展,许多小的本地公司也提供与因特网的可靠连接,位于马里兰州巴尔的摩(Baltimore)郊外的 ClarkNet 就是其中之一。从一个 500 英亩农场的谷仓中的一组调制解调器开始,ClarkNet 现在无论从规模,还是服务质量上都已发展成为了这个国家最好的地区服务提供商之一。ISP 的最大优势在于连接速度,通常小的提供商可以提供比在线服务还快的连接速度。一般 ISP 按月收费,可以无限使用。

现在在线服务与 ISP 之间的界限已经逐渐模糊,并向在线服务倾斜。AOL 最近更改了其收费策略,变为按月而不是按小时收费。使得用户花费和大多数 ISP 相等的开销,就可以得到全部 AOL 的专有内容,以及通过 ISP 所能得到的全部因特网工具。在美国的大多数州,AOL 网络的密集程度已经使得访问 AOL 变得像打本机电话一样方便,用户无论在哪里都可以方便地访问电子邮件,这是本地小 ISP 不能比拟的。像这样的发展趋势,加上本地电话和电缆公司也开始通过新的高速光纤网提供访问因特网的服务,使得将来终端用户访问因特网将越来越便宜,而且性能越来越好。

1.3 电子邮件

许多用户是通过使用电子邮件(E-mail)认识因特网的。电子邮件是一个方便快捷的发送、接收及回复消息的媒介,在许多地方实际上已成为不可或缺的工具。它的优势主要有:

- 比邮政服务快得多。
- 传送消息比传统电话或面对面交谈要更为清晰明确。
- 接收者可以有很大的自由度来决定是马上回复、过一会儿回复还是根本不回复,从而更好地控制自己的工作流程。
- 提供了一种方便的整理、保存消息的途径。
- 发送电子邮件的成本很低,或根本不需花费。

虽然这些优势已经使得电子邮件成为了工业界和科学界一种十分重要的个人通信手段,但用户也必须清楚它的两个主要缺点。第一是安全性问题。邮件在传递到接收者的过程中,可能经过一系列远程节点,在其中任何一处邮件都可能被有较高权限的人(例如系统管理员)所截取和阅读。第二是保密问题。在工业界中,电子邮件通常被认为只用于办公室通信的公司财产,因此就必须在管理者的监控之下。而在学院、准学院及研究领域则相反,例如,NIH 鼓励个人在公布的准则范围内使用电子邮件,这里的关键是"公布的准则"。无论在什么情况下,电子邮件系统的用户都应该了解本机构的电子邮件使用规则,从而正确有效地使用这一工具。我们竭诚推荐 Lamb and Peek(1995)所著的一本关于如何有效使用电子邮件的杰出指南。

发送电子邮件

电子邮件地址的格式一般为:用户@计算机.域。其中"用户"是个人用户名,"计算机.域"指向电子邮件账号所在的计算机。和普通邮件一样,电子邮件消息包括"信封"(或称为"信头"(*header*))和"正文"(*body*)。信头的内容包括:收发信人的地址、电子邮件主题行以及邮件如何从发信人到收信人的信息。信头下面是真正的消息(或"正文"),如同普通信件信封里的内容一样。图 1.2 显示了一个电子邮件消息的全部组成部分。

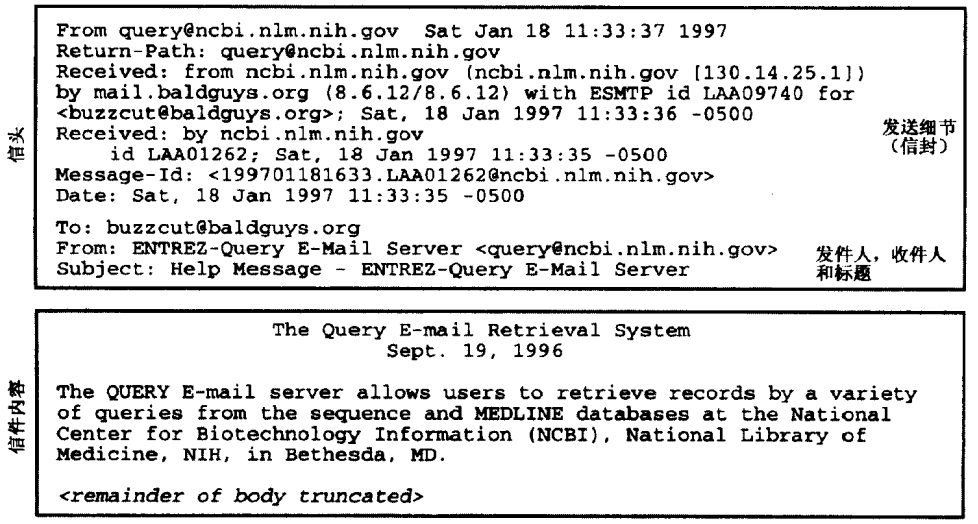


图 1.2 剖析一个电子邮件(E-mail)消息(各部分已经标示出来)。这是一个 NCBI 查询服务器对求助消息的自动答复。

电子邮件程序千变万化,随使用平台以及用户的需求而不同。通常局域网(LAN)的属性决定了可以使用哪些邮件程序,而且这一决定往往是由系统管理员而非个人用户做出的。最广泛使用的带有图形用户界面的电子邮件软件包有:用于 Macintosh 的 Eudora 和用于 Mac,Windows 和 UNIX 平台的 Microsoft Exchange。基于文本的电子邮件程序有 Elm 和 Pine,这通常需要注册到 UNIX 账号来使用。

新闻组

在图 1.2 的例子中,电子邮件被发给一个单独的收信人。电子邮件系统的一大优点就是可以把一封电子邮件发给很多人。第一种方法是通过使用"别名"来实现这一目的。用户可以在邮件程序中定义一组人,并给这个组一个名字("别名")。用户把电子邮件发给这个别名后,邮件程序就会自动地把消息广播给组中的每个人,而不是逐一发送。即使在小规模组中设立别名也可以节省大量时间。这样也可以保证组中的每个人都能真正接收发到组里的每一封信。

"新闻组"则是广播电子邮件消息的另一种方法。和"别名"方法稍为不同的是,订阅

新闻组时电子邮件地址列表由远端计算机编排维护,就像杂志维护一份订阅者名单一样。例如,BIOSCI 新闻组是流通量最大的新闻组之一,提供了一个在相当大的生物学主题范围内讨论和交换思想的论坛。要开始接收发表在 BIOSCI 的自动测序讨论组的邮件,用户需要发送一个消息给 *biosci-server@net.bio.net*,并在正文中写上 subscribe autoseq。新闻组中的全部文章就会发送到新的订户手中,该用户就可以参与讨论了。用户想要退出新闻组时,只需要给相同地址发送消息 unsubscribe autoseq 就可以了。要得到包括讨论组完整列表在内的 BIOSCI 的更多信息,给 *biosci-server@net.bio.net* 发一封邮件,清空邮件主题行,并在消息正文中写上 info faq。BIOSCI 服务器会发送一份常见问题表(FAQ)作为回复,其内容包含了 BIOSCI 管理下的每个新闻组的详细信息。

就像邮政信件一样,电子邮件中最近也有一股"垃圾邮件"的浪潮,这些邮件来自某些公司出于商业宣传的目的而编排的大量邮件地址列表。大多数这样的列表都是从联机注册或相似渠道得来的,所以避开这种邮件列表的最好办法就是有选择地给出你的电子邮件地址。大多数新闻组的电子邮件地址是保密的,如果你有疑问的话,不妨先询问一下。

电子邮件服务器

到此为止,讨论仅限于发送消息,接收者可以是一个或是很多。电子邮件还可以用于从生物数据库中进行预测或读取记录。用户可以用电子邮件给远端的服务器发送消息,以预先定义的格式说明希望进行的操作,服务器就会执行这些操作,并将结果用电子邮件返回给用户。图 1.2 显示的是电子邮件查询结果的示例,其中服务器是 NCBI 的查询电子邮件服务器。虽然这种方式不是交互式的,但它将硬件维护和软件升级的工作交给了维护服务器的管理员,使用户更专心于结果而非程序本身。后面章节中将有一些电子邮件服务器的详细论述。日内瓦大学的 Amos Bairoch 维护着一个优秀的最新电子邮件服务器列表,此列表可以通过匿名文件传输协议(下文介绍)的方式得到。具体做法是:访问 *expasy.hcuge.ch* 站点,进入 */database/info* 目录,下载文件 *serv-ema.txt*。对于大多数这样的服务器,给服务器的电子邮件地址发送一个 help 消息就可以得到服务器的详细指令集,其中包括查询的正确格式。

1.4 文件传输协议

虽然电子邮件传送消息存在很多优点,但是有经验的用户在传输附加文件时都曾遇到麻烦。问题在于仅仅将文件附加在邮件上并发送出去,并不意味着接收者能真正得到、解码并使用这个文件。虽然已经开发了许多跨平台电子邮件软件(例如 Microsoft Exchange),但不同地点的人使用不同的电子邮件软件使得通过邮件发送文件并非有效、安全的方法,至少在近期内如此。对此问题的解决方法之一是使用文件传输协议(FTP)。FTP 的使用十分简单,在用户计算机(客户)和远端服务器之间建立起连接,并在整个 FTP