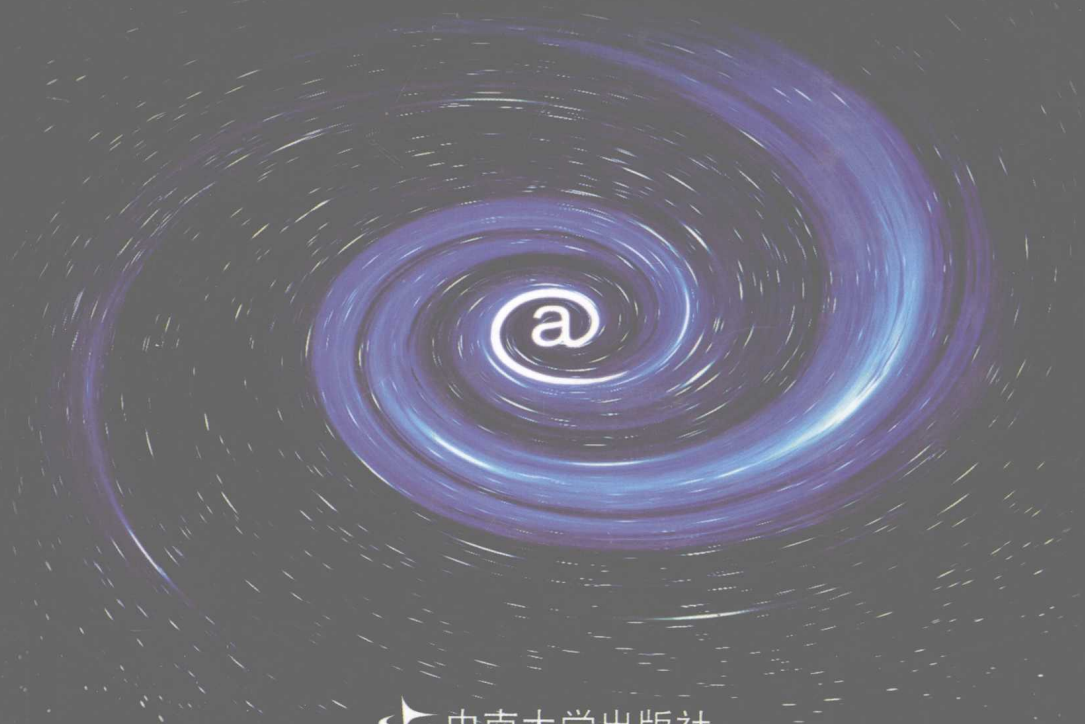
BIOINFORMATICS MANUAL
Based on World-Wide-Web

基于WWW的
生物信息学应用指南

主　编　李桂源　钱　骏



中南大学出版社

图书在版编目(CIP)数据

基于 WWW 的生物信息学应用指南/李桂源,钱骏主编.
长沙:中南大学出版社,2004.4
ISBN 7-81061-843-1

Ⅰ.基... Ⅱ.①李...②钱... Ⅲ.因特网-生物信息论-情报检索-指南 Ⅳ.G252.7-62

中国版本图书馆 CIP 数据核字(2004)第 034792 号

基于 WWW 的生物信息学应用指南

主编 李桂源 钱 骏

□**责任编辑** 谢新元 李 娴
□**出版发行** 中南大学出版社
社址:长沙市麓山南路 邮编:410083
发行科电话:0731-8876770 传真:0731-8710482
电子邮件:csucbs @ public. cs. hn. cn
□**经 销** 湖南省新华书店
□**印 装** 中南大学湘雅印刷厂

□**开 本** 787×1092 1/16 □**印张** 21 □**字数** 514 千字
□**版 次** 2004 年 4 月第 1 版 □2004 年 4 月第 1 次印刷
□**书 号** ISBN 7-81061-843-1/TP·031
□**定 价** 35.00 元

内 容 提 要

在当今的信息时代，失去信息的利用，等于失去许多重要的一切。本书则是教会从事生物、医学等生命科学的人们如何利用国际互联网掌握更多、更前卫的生物信息。环球网（WWW）是当前国际互联网中重要的信息网，它涵盖核苷酸和蛋白质序列查询、递交、基因的计算机克隆、序列相似性搜索，蛋白质基序和结构域识别、进化树构建、蛋白三维同源模型构建以及基因数字化差异表达分析等主流生物信息学的应用，以上内容在本书中均作了详细阐述。同时，本书最大的特色是结合了中南大学肿瘤研究所以及肿瘤基因组研究中心精心设计了多达43个具体生物信息学应用练习题，并按实例进行深入浅出的表达，讲解清晰，使读者能够直接参考这些练习来解决科研中的实际问题，避免不必要的重复研究，少走弯路，有助于提高我国生物科学研究水平。本书很适合从事生物学、医学、分子生物学、肿瘤学、病理生理学等专业的科研人员及工作者阅读参考，也可作为本科、研究生教育教材使用。

前　言

21 世纪是生命科学的世纪，是一个全球化的信息时代，而最能够体现这一时代特点的则莫过于著名的人类基因组计划（Human Genome Project，HGP）的实施和即将完成。除了人类基因组外，其他模式生物，诸如啤酒酵母、线虫、果蝇、小鼠、拟南芥、水稻等的基因组测序计划在最近几年里也均相继完成或全面实施。正是这些测序计划导致了当前有关核酸、蛋白质的序列和结构数据呈爆炸性指数增长。全世界的科学家们也正在积极运用数学、计算机科学和生物学的各种工具，来阐明和理解基因组学获得的大量数据中所包含的生物学意义。从 20 世纪 80 年代末开始，一门新兴学科——生物信息学应运而生。近年来，计算机和因特网的发展更是为生物信息学的迅速发展提供了硬件基础和便利。无庸讳言的是，当前我们也面临着生物信息数据增长过快，各种各样数据库种类愈来愈多的困境，但这也同时意味着成长迅速、充满活力的生物信息学给我们带来的无限施展创造力的机会。

本书重点介绍了如何充分利用当前基于国际互联网的各类免费生物信息学资源来进行核苷酸序列和蛋白质序列的生物信息学分析，内容涵盖了序列查询和递交、基因的计算机克隆、序列相似性搜索、蛋白质基序和结构域识别、进化树构建、蛋白三维同源模型构建以及基因数字化差异表达分析等主流的生物信息学应用。书中最大的特色在于精心设计了多达 43 个具体的生物信息学应用练习，这些练习中的绝大多数内容均结合了中南大学肿瘤研究所及中南大学肿瘤基因组研究中心在近年来实践过程中的具体实例进行表述，讲解极为清晰，并配有大量图表，内容深入浅出，避免了使用大量复杂的生物信息学专业理论词汇，使读者能够直接参考这些练习来解决实际问题。实际上，我们认为不论是对于实验室还是分子生物学家个人，学会充分利用互联网上的各种生物信息资源进行分析、归类与重组，发现新线索、新现象和新规律，用以指导实验工作的设计，这应该是一条切实可行的、既快又省的科研路线，可避免不必要的重复，少走弯路，有助于提高我国生物科学的研究水平。

最后需要特别指出的是，本书中所介绍的某些内容均来自于基于互联网的免费生物信息学数据库及相应的分析工具，而这些均得益于全世界各个生物信息学研究机构的全体成员在崇尚信息和技术自由共享的精神指导下所做出的默默无闻的辛勤工作和无私奉献。事实上，没有他们耐心负责地维护着这些生物数据库，没有他们慷慨地制作出许多能够被自由访问的生物信息分析工具，我们将连最基本的序列分析研究都无法做到。尤为重要的是，这些数据库和分析工具在提供给我们免费使用的同时，也都具备了他们所能做到的最高质量。我们要感谢所有参与这些项目的人们，没有他们也就不可能有本书的诞生。

我们还要感谢国家自然科学基金委和国家 863、973 项目的支持，没有他们多年来对我们巨大的基金资助，我们也不可能得以在充满荆棘的科学道路上进行各种有益的探索。

编　者

2004 年 4 月 8 日于中南大学

目 录

第1章 引 论

人类基因组计划（Human Genome Project，HGP）及其他模式生物的测序计划带来了无法预料的庞大的序列数据信息。面对这些测序计划所产生的海量信息，作为生命科学领域中的新兴学科——生物信息学（Bioinformatics）的重要性越来越突出。它无疑将为21世纪生命科学的研究带来革命性的变革。尽管早在上个世纪的60年代就已经开始了生物信息学的某些应用例如关于DNA和蛋白序列数据库的建立以及序列的基本分析等，但生物信息学这个重组词汇却直到20世纪90年代初才正式出现。实际上，正是人类基因组计划的启动和实施促使了生物信息学从一个简单的重组词汇向一门全新的充满悬念的前沿学科跨越。也正是随着人类基因组计划过程中出现的爆炸性增长的各类序列信息加速了生物信息学的迅猛发展。作为一门前沿性的交叉学科，生物信息学涉及到计算机科学、数学、物理学、生物学、医学等多个学科领域（图1-1）。与传统的基于实验室的生物学相比，在被称为信息时代的21世纪，我们渴望利用计算机从海量的数据中挖掘出更多的科学知识。因此，建立在计算科学基础上的生物信息学又被称为是21世纪的“新生物学”。这种基于序列及其相关信息的新生物学的终极目的，就是破译在生命起源及演进过程中的信息密码。

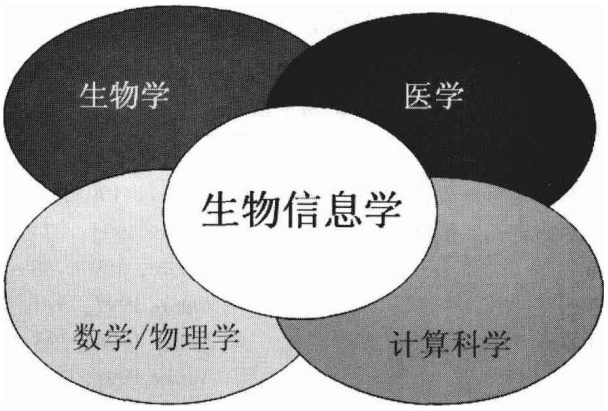


图1-1 生物信息学与其他多个学科的交叉

在实际工作中，生物信息学不仅仅是一门前沿理论学科，更重要的是它已经成为一个融入到实验室的具体实践方法。生物信息学就如同一个向导，实验室中的生物学家可以通过这个向导对将要开展的实验研究作出更周密的实验设计、预测研究的结果和对实验结果做规律分析，以减少实验失败率、缩短实验周期、获得实验成功。正如1991年诺贝尔奖获得者W. Gilbert所指出的：“传统生物学解决问题的方式是实验的。现在，基于全部基因都将知晓，并以电子可操作的方式储存在数据库中，新的生物学研究模式的出发点应是理论的。一个科

学家将从理论推测出发，然后再回到实验中去，追踪或验证这些理论假设。”如果我们曾经不得不花上几个月的时间在实验室运用从电泳到杂交等传统实验手段来发现一个基因或蛋白的可能功能，那么在基于序列信息的“新生物学”的今天，相信我们将更会愿意在计算机面前多呆上几个小时，以期望获得更多从传统实验中无法获得的信息——生物信息学的诱惑力和潜力也正是在此。

1.1 基因组/蛋白质组信息学

目前生物信息学并没有一个严格的学科分类，随着人类基因组计划测序目标的即将完成和一个更为长期、任务更为艰巨的后基因组时代（post - genome era）的到来，大致可以将当前的生物信息学分为基因组信息学（genome informatics）和后基因组信息学（post - genome informatics）两大类，后者目前主要的内容即蛋白质组信息学（proteomicinformatics）。

1.1.1 基因组信息学

人类基因组草图在 2001 年的发表成为生物信息学发展历史上最近的里程碑之一。自从人类基因组计划实施以来，除了对人类基因组进行测序外，事实上，在上个世纪的最后 10 年还完成了许多重要模式生物的基因组全序列测定或基因组草图，包括果蝇、酵母、小鼠及多种与人类疾病有关的细菌等（表 1 - 1）。这些测序计划的实施和快速发展产生了基因组信息学，这是目前生物信息学的主流，对这些生物基因组序列的分析也已成为生物信息学近年来的主要任务。

表 1 - 1 已经完成的若干重要生物基因组序列

时间	物种	参考文献
1995	第一个细菌基因组——流感嗜血杆菌	*Science*, 1995, 269: 496 - 512
1996	第一个真核生物——芽殖酵母	*Science* 1996, 274: 546 - 567
1997	幽门螺杆菌	*Nature*, 1997, 388: 539 - 547
1997	枯草芽胞杆菌	*Nature* 1997, 390: 249 - 256
1998	结核分枝杆菌	*Nature* 1998, 393: 537 - 544
1998	立克次体	*Nature* 1998, 396: 133 - 140
1998	第一个多细胞生物——美丽线虫	*Science*, 1998, 282: 2011 - 2046
2000	果蝇	*Science*, 2000, 287: 2181 - 2215
2000	第一个植物基因组——拟南芥	*Nature* 2000, 408: 791
2000	霍乱弧菌	*Nature*, 2000, 406: 477 - 483
2001	大肠埃希菌	*Nature* 2001, 409: 529 - 533
2001	鼠疫耶氏菌	*Nature* 2001, 413: 523 - 27
		Science, 2001, 291: 1304 - 1351
2001	人类基因组草图	*Nature*, 2001, 409: 6822
2002	水稻基因组草图	*Science*, 2002, 296: 79 - 92
2002	小鼠基因组草图	*Nature*, 2002, 417: 141 - 148
2003	人类基因组测序计划完成	

基因组信息学的研究内容涵盖了基因组信息的获取、处理、储存、分发、分析和注释等所有方面，主要包括有：① 发展有效的信息分析工具，构建适合于基因组研究的数据库，用于搜集、管理和使用人类基因组和模式生物基因组的大量信息；② 配合实验研究，确定约30亿个碱基对的人类基因组完整的核苷酸顺序，找出全部3～4万个人类基因在染色体上的位置以及包括基因在内的各种DNA片段的功能，也就是“读懂”人类基因组。这些工作包括DNA序列相似性搜索和比较、序列模式识别（开放阅读框、外显子、内含子等的识别）。其中主要任务之一是发现新基因及其相应的新功能；③ 研究基因产物即蛋白质和多肽的结构信息，如基序（motif）和结构域（domain）的识别、二级结构预测、亲疏水性分析、三维结构研究等；④ 构建基因或蛋白质家族的系统发生进化研究，即进化树的构建。

尽管建立更合理的算法和开发相应的计算程序本身就是基因组信息学的一个重要组成部分，但对于一个比较专注于实验室数据的分子生物学家而言，在他眼里的基因组信息学往往转化为非常具有实践性的实验室问题。例如，如何利用计算机来克隆新基因，或者如何了解一个实验序列是否已经被他人所克隆过等等。因此从实际应用出发，在很大程度上，目前所谓的基因组信息学的中心内容就是对各类基因或蛋白序列进行计算机分析，从而帮助生物学家寻找上述问题的答案。一个较为完整的序列分析过程大致可以归纳如图1－2所示。

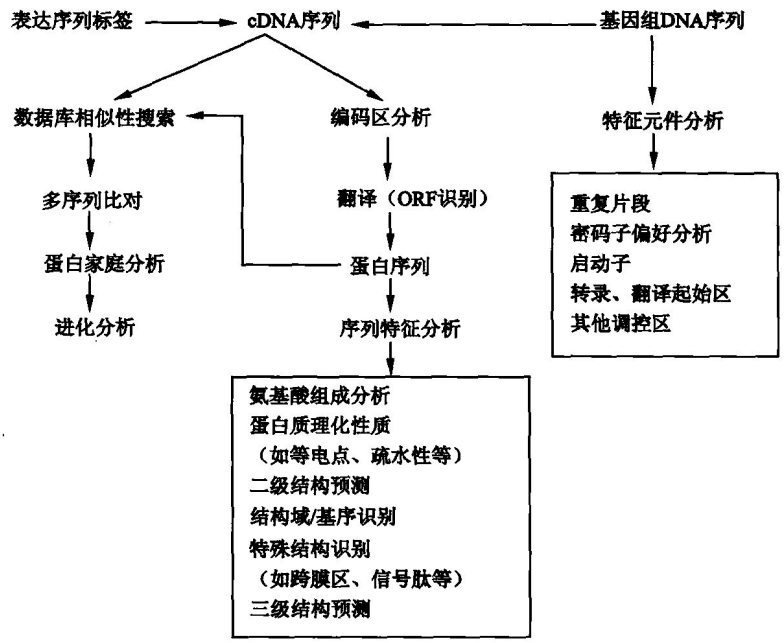


图1－2 基因组信息学序列分析流程

图1－2包括了从基因的电子克隆到核苷酸和氨基酸序列的生物信息学分析全过程。在这个流程图中可以看出最重要的一个方法或工作就是序列的数据库相似性搜索。通过对不同类型的数据库搜索，我们可以确定一个序列是否是新基因、该序列是否可能属于某个已知的基因家族、该基因的编码产物的重要保守结构域/基序的分析，以及在此基础上的基因系统进化分析和在不同组织的基因表达谱分析等等。这些内容是当前生物信息学应用的主流，也

正是本书所要介绍的重点所在。

1.1.2 蛋白质组信息学

随着大量基因序列的产生，生物信息学的发展已经不再局限于序列分析。在后基因组计划中，分析人类全部 3 ~ 4 万个基因的功能，阐明基因组所表达的真正执行生命活动的全部蛋白质的表达规律和生物功能是主要任务，其中最直接的就是蛋白质组研究。然而，蛋白质组比基因组具有更大的复杂性，蛋白质组信息学更有挑战性。初级阶段主要涉及蛋白质数据库，特别是人或其他哺乳动物蛋白质组数据库的建立、相关软件的开发和应用，进而研究蛋白质组成员的序列、结构、功能及定位分类和基于生化途径、遗传网络等构建蛋白质组功能系统即蛋白质连锁图，高等生物基因组中蛋白质编码基因的识别及算法研究，基于蛋白质数据库与知识库的知识与规律发现，新型蛋白质结构、功能预测方法及程序等（详见第 13 章）。

1.2 互联网与生物信息学

如果说大量生物序列数据是生物信息学诞生的基础，那么计算机和互联网（Internet）就是生物信息学腾飞的翅膀。事实上，离开了计算机的发展，人类基因组及其他模式生物基因组便不可能在过去短短的 10 年内获得完成；而没有互联网，这些序列信息便无法有效地被组织、存储、传递给全世界的科学家共享。非常有趣的一个事实是，人类基因组计划从一开始就申明了所有人类基因的序列信息应该属于全人类，反对将这些信息用于商业性质的任何企图。信息的共享成为人类基因组计划实施过程中一个非常重要的理念。而互联网的快速发展不仅很好地体现了这种自由共享的精神，更重要的是它提供了实现这种自由共享的舞台。生物信息学群体的独特之处在于，在商业部门之外，它比生物学中许多竞争性领域更提倡“团队精神”。生物信息学中已取得的卓越进展就蕴藏在从收集、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，而且这一切均处于信息和技术自由共享的环境内。生物信息学正是通过互联网来增强和拓展其应用，从而使得传统的生物学从实验室演变为一门信息科学。在这个意义上，生物信息学正成为一门全球化的学科，借助计算机网络，我们可以非常方便地找到各种生物数据以及发展各种软件。目前，绝大多数的国家和研究机构均致力于将生物信息向全人类公开，尤其是通过互联网让全世界的科学家共享。在这个意义上，本书所涉及到的数据库和相应的分析软件均是来自于所谓的公共数据库以及一些生物信息学团队所开发的免费工具，尤其是以基于 WWW 操作界面的软件为主。

尽管本书的大部分读者被假定为已经掌握了基于互联网的计算机有关基本操作，例如采用 WEB 浏览器浏览互联网信息和收发电子邮件等日常工作，但在我们正式开始利用互联网进行第一个生物信息学应用教程之前，仍然有必要了解一些关于互联网和 WWW 的基本概念，特别是对于那些曾经只专注于实验室，对于计算机和互联网的了解也仅仅止于用浏览器浏览信息和收发电子邮件等操作的分子生物学家，我们确信你知道得不够。例如，实际上，有些时候你不能确定是否确切知道 WWW 和 Internet 的区别，特别在如果你是个生物信息学的初学者的情况下。本节将帮助你更多地了解一些 Internet，尤其是目前互联网最大的应用

——环球网（World Wide Web，WWW）。

1.2.1 互联网基础

1.2.1.1 INTERNET 和 WWW

1946 年，第一代计算机的诞生是人类科学发展史上一个重要的里程碑。计算机的出现实现了人类自文艺复兴以来最渴望的一个梦想——制造一台能帮助人进行计算的机器。遗憾的是，当时的计算机体积之大、能耗之高、故障之多、价格之贵让今天的人们难以想象，因而也大大地制约了它的普及应用。“未来的计算机不会超过 1.5 吨。”当年的科学杂志对于计算机的将来作了非常乐观的预测。幸运的是，没过多久，这样的预测便宣告失败。随着晶体管和集成电路的出现，计算机终于找到了腾飞的起点，一发而不可收拾。尤其是集成电路的发明使得人们能够开始制造革命性的微型/个人计算机。20 世纪 80 年代微型计算机的出现，改变了主机模式的集中管理和运行方式，把强大的计算和处理能力交到了个人手里，这为计算机进入各行各业乃至普通老百姓的家庭奠定了基础。几乎是在个人计算机蓬勃发展的同时，一项被称为 20 世纪人类最伟大发明之一的互联网（Internet）已经初现端倪。后来的事实证明，与其说是计算机，不如说是由于互联网的出现，计算机才真正开始改变人们的工作和生活。在互联网时代，整个互联网实际上就是一台巨大的超级计算机，实现了计算资源、存储资源、数据资源、信息资源、知识资源、专家资源的全面共享。人们常说网络就是计算机，深刻地反映了网络在计算机发展史中极为重要的作用和影响。

虽然在世界范围内，互联网早在 20 世纪 80 年代就存在了。然而在中国，当 1994 年之前如果你跟实验室的同事提起 Internet，我相信他肯定还不知道 Internet 是什么。因为中国直到 1993 年才由中国科学院高能物理研究所建立了第一条国际互联网专线。时至今日，短短不到 10 年的时间，在中国兴起了一种研究、学习和使用 Internet 的浪潮，Internet 已经越来越成为中国人科研工作甚至日常生活的一个重要组成部分。互联网正在悄悄改变人们的生活及行为方式，越来越多的人每天都体验着互联网的魅力，这中间也包括那些每天在实验室里呆上 8 小时以上的分子生物学家们——他们早就习惯了在等待实验结果出来之前的一小会时间里，上网打开自己的免费邮箱，看看有没有最新的电子邮件。毫无疑问，作为一种全新的通信方式，电子邮件业已成为在网络上最常用的人与人交流的方式之一。在这个基础上，加上受到某些媒体的错误引导，大多数人们便会轻易地相信电子邮件就是当今互联网上最多的应用。的确，从互联网的历史发展来看，正是由于电子邮件服务的蓬勃兴起，互联网才得以成为一种现实。然而促使互联网真正成为每个普通老百姓都能接触的却是环球网（World Wide Web，WWW）的出现。

环球网（WWW）有时也简称为 Web。它是一种基于互联网的采用交互式图形界面的网络应用。从包含电子邮件在内的各种互联网信息应用所占的信息流量比例来看，WWW 才是目前全世界应用最广的一项服务。要知道，即使在电子邮件应用中，有超过 90% 的用户是通过访问基于 Web 的电子信箱来发送和接收电子邮件的。WWW 的出现使得全世界的用户均可通过简单的图形界面就可以访问各个大学、组织、公司等最新信息和各种服务。事实上，目前 WWW 已经成为很多人在网上查找、浏览信息的主要手段。然而，由于互联网的

最初宗旨是支持教育和研究活动，因此，早期的互联网充满了学术研究的庄严与神秘，枯燥单调的字符为互联网的普及设置了高高的门槛。1992 年，一个在欧洲核物理研究所（CERN）工作的物理学家惕莫伯纳斯利（Tim Berners - Lee）发明了 WWW。他在 CERN 工作的时候写了第一个 WWW 客户程序和第一个服务器程序，并且制定了一些至今仍在广泛使用的标准，如 URL、HTML 和 HTTP 等。1993 年，美国伊利诺宜斯州立大学的国家超级计算中心（NCSA）发布了第一个图形界面浏览器 Mosaic。Mosaic 浏览器的诞生使包含图形的网页设计成为可能。正是 WWW 以及随后的浏览 WWW 的工具——Web 图形浏览器的发明带来了互联网的春天。从此，互联网的表现力越来越真实，越来越强大，带来了人们对 Web 浏览的无限向往。

WWW 对于互联网的革命，就好比微软的视窗（windows）操作系统的出现对个人计算机普及的贡献。WWW 这种基于超文本链接的界面简化了互联网中资源获取的方法，用户无需去记忆繁琐的操作命令，用户界面更为友好。就拿收发电子邮件这一最简单的网络实践来说吧，现在我们已经习惯了直接在提供电子邮件服务的 Web 页面上输入我们的用户名和密码进入我们的信箱。而如果没有 WWW，那么我们要想成功地收发一封电子邮件，首先我们必须安装一个类似于 Outlook Express 的电子邮件终端软件，然后设置好发送邮件的地址和接受邮件的地址，如果发送邮件的服务器还需要身份验证，那么我们还要进一步设置好我们的邮箱账户和密码，否则，你将无法把邮件发送出去。毫无疑问，在上述两种方式之间，通过 WWW 直接收发邮件更加简单、方便和傻瓜化——即使是你的一条狗也能成功地做到，虽然它从来就未曾比第二种更快或更安全。

1.2.1.2 HTML 和超文本链接

简单地讲，所谓的环球网主要是由两部分组成：一是 Web 服务器（Web Server），用来存储和发布各种信息；一个是客户端的 Web 浏览器（Web Browser），用来获得由各种服务器所提供的信息。用户可以方便地通过浏览器来浏览位于 Web 服务器上的文件，而不管文件是在哪一台电脑上，这就是所谓的超文本链接技术（Hypertext）。所有的 WWW 文档都是用超文本标记语言（Hypertext Markup Language，HTML）来写的。这些文档称之为 HTML 文档，通常以 HTML 或 HTM 为文件扩展名。尽管其中某些文件有不同的扩展名（如 .cfm 或 .asp），但它们的核心仍旧是 HTML。HTML 不是像 C + + 或 Pascal 那样真正的计算机编程语言，它是一种描述文档的系统，用来描述文件的结构和超文本链接信息。因此 HTML 文档可以用任何文本编辑工具来书写。这是一个很好的优点，即意味着 HTML 文档内容能被任何系统读入（包括 Macintosh 和 UNIX 系统等），而不用担心所谓的兼容问题。这样只要有浏览器在的时候，WWW 就能被任何平台使用。

WWW 浏览器通过解释 HTML 并显示它，这就成为平常我们所看到的一个 Web 页面。图 1 - 3 是用微软的 Web 浏览器 Internet Explorer 5.0 浏览位于北京大学生物信息学中心的首页。在这一页中包含了以下划线、文字或图形标识的超文本链接。浏览主页时，用户只需用鼠标点击感兴趣的超文本链接（通常这时鼠标指针会发生变化，即变成一个手指的模样，表示这是一个可以点击的超链接），就可以获得该链接所指向的进一步信息。通过这样的超链接，用户可以随意地从一个页面跳转到另一个页面，从中国北京访问到美国纽约，这种基于超文本链接的浏览过程被形象地称为“网上冲浪”——那就意味着你可以从任何地方开

始，而且什么使你产生了兴趣你就可以跳到什么地方去，就象一个在狗身上的多动的跳蚤一样在 Internet 里跳来跳去^_ ^。这正是 WWW 迷人的强大之处，其超文本链接不仅能够指向统一目录下的文件、同一计算机不同目录和文件，还可以指向世界上任何地方、任何计算机及其任何目录文件。而这些对于我们用户是透明的，即我们无须知道这些超链接背后的信息是如何被组织及管理的，我们所要作的就是滑动我们手中的鼠标，尽情地在网上冲浪。

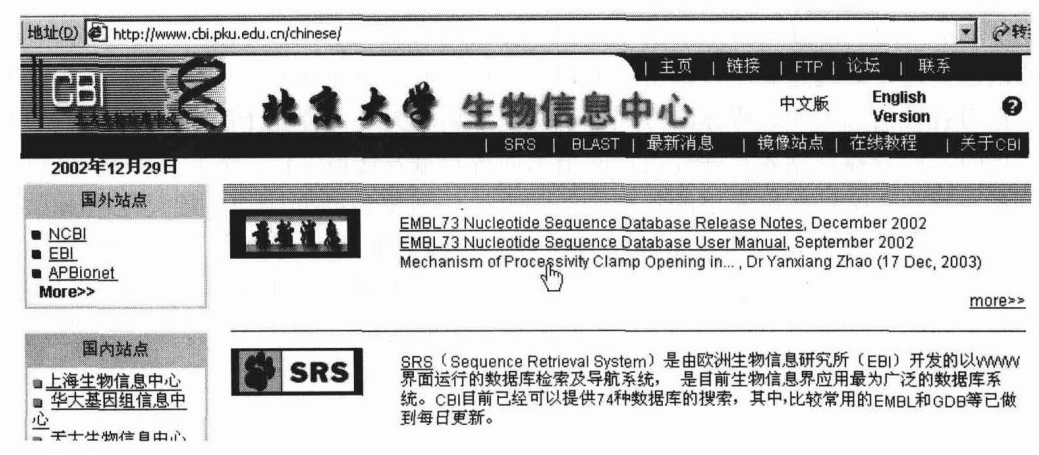


图 1-3 基于超文本链接 Web 浏览界面

目前 HTML 版本有 2.0、3.0、4.0、Netscape 扩展版本及 Microsoft 扩展版本等。HTML 3.0 是所有浏览器都普遍支持的标准。而对于用 HTML 4.0 所编写的 Web 页面，不同的浏览器有可能产生不同的效果。而在未来，一个决定 Web 命运的新的标准将可能是可扩展标记语言（EXtensible Markup Language，XML）。如果说，HTML 提供了显示全球数据的通用方法，那么 XML 进一步提供了处理全球数据的通用方法。XML 有着 HTML 语言所欠缺的巨大伸缩性与灵活性。XML 不再像 HTML 一样有着一成不变的格式。XML 实际上是一种定义语言，即使用者可以定义无穷无尽的标记来描述文件中的任何数据元素，从而突破了 HTML 固定标记集合的约束，使文件的内容更丰富更复杂并组成一个完整的信息体系。良好的数据存储格式、可扩展性、高度结构化、便于网络传输是 XML 四大主要的特点，决定了其卓越的性能表现。它代表着人们编写标准 HTML 习惯的最大转变。XML 定义文档的结构而不是定义浏览器应该如何显示文档，这将给 Web 开发商提供许多灵活性。它改变了浏览器的显示、组织和搜索信息的方法。

XML 自 1996 年开发以来已取得了巨大进展。微软的 Web 浏览器 Internet Explorer 4.0 以上版本已经开始支持 XML，Netscape 计划中的新版本也肯定会支持 XML。其他公司，包括 Adobe 和 Sun 公司也宣布支持 XML。XML 无疑将成为在 Web 上发布基于 Standard General Markup Language（SGML）规范的信息的工具。特别是对于生物信息学，XML 可能更加重要。例如，目前生物信息学数据库多如牛毛，他们都有各自不同的复杂格式。在不久的将来，使用者与这些数据库间将有可望只通过一种标准语言进行交互，那就是 XML。由于 XML 的自定义性及可扩展性，它足以表达各种类型的数据。客户收到数据后可以进行处理，也可以在不同数据库间进行传递。总之，在这类应用中，XML 解决了数据的统一接口问题。

目前，IBM、Sun 和 Oracl 等公司宣布将利用 XML 与 JAVA 语言联合开发出大型生物信息平台，旨在帮助生物技术研究人员处理基因、蛋白质信息，以推动人类基因组学研究成果的商业化进程。世界上最大的两个生物信息中心美国生物信息学中心（NCBI）和欧洲生物信息研究所（EBI）也即将在不久的将来开始向 XML 迁移。

1.2.1.3 HTTP 协议、URL 和 WEB 浏览器

WWW 系统的工作方式主要是基于 HTTP 协议的客户/服务器模式，即客户端和服务器之间的信息传递是通过超文本传输协议 HTTP（HyperText Tromsmission Protocol）实现的。服务器负责对各种信息按照超媒体的方式进行组织、形成文件并存储于服务器上，当用户进行信息查询时，服务器根据客户端提出访问请求向客户端发送该文件，客户端收到文件后，通过浏览器对超文本文件作出完整的解释。实际上，任何单独的程序均可通过这种分散方式，通过协议某个恰当的网关 CGI 来被客户端执行。正因为如此，这也成为大多数基于 WWW 的生物信息学服务所采用最多的一个模型（图 1－4）。CGI 代表 Common Gateway Interface，即公共网关界面，它是 HTTP 协议与其他程序和系统的接口，和 HTTP 的客户与服务器采用同样的数据协议。

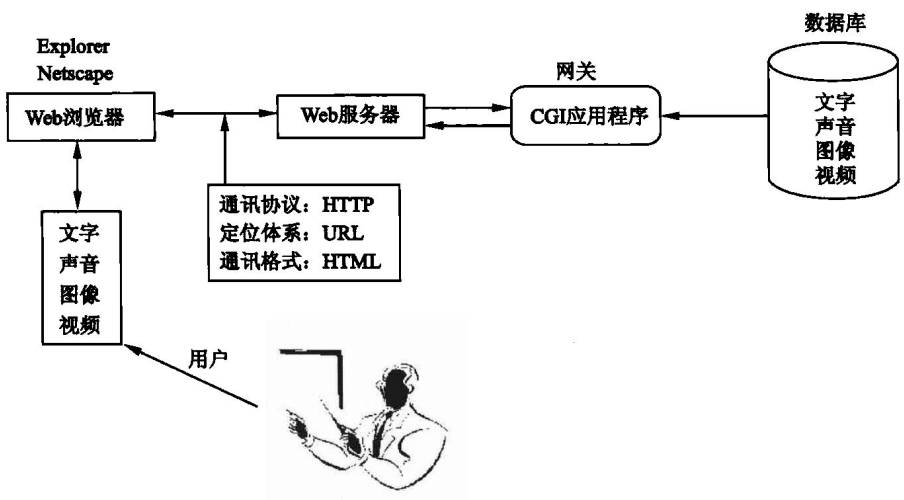


图 1－4 基于 HTTP 协议的 Web 客户/服务器模式

用户在利用 Web 浏览器漫游 WWW 时，Web 浏览器的主要任务就是使用 HTML 中的一个起始统一资源定位符（Uniform Resource Locator，URL）来获取某个 Web 服务器上的 Web 文档，解释这个 HTML，并将文档内容以用户所许可的效果最大限度地显示出来。所谓的 URL 是 Internet 上用来描述信息资源的字符串，主要用在各种 WWW 客户程序和服务器程序上。URL 完整地描述了互联网上文件的地址，俗称网址。用户在进行信息查询时，首先应正确地输入 URL 的地址，其地址书写格式为：HTTP：//服务器名/文件的路径及名字。可以看出，URL 由三部分组成，其中第一部分表示访问信息的方式或使用的协议，如 HTTP 是指超文本传输协议；第二部分表示提供服务的主机名；第三部分是所访问主机的端口号、文件、目录或检索数据库的关键词等。例如我们在浏览器的地址栏输入以下 URL：http：//

www. cbi. pku. edu. cn/chinese/documents/index. html，所表示的含义就是在北京大学生物信息中心中文版 WWW 服务器上文档文献子目录下名为 index. html 的文件（图 1 - 5）。Internet 上每个资源都可以用一个 URL 来指定。这种资源的通用命名规则使得 WWW 成为一个很好的信息集成环境，所以许多用户以 Web 浏览器作为访问 Internet 的通用工具。

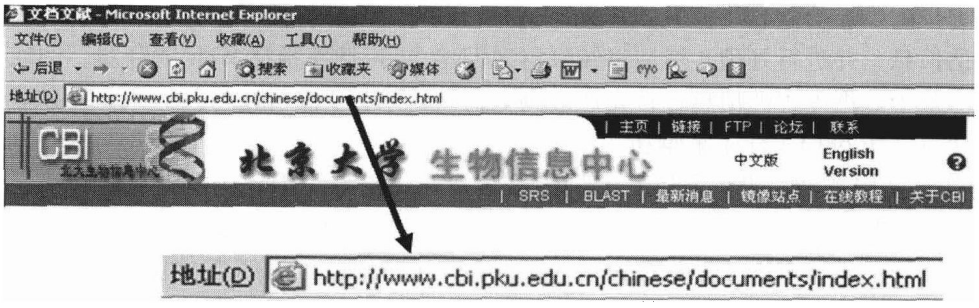


图 1 - 5 北京生物信息中心的在线教程的 Web 地址

目前最流行的浏览器莫过于微软的 Internet Explorer 和网景公司的 Netscap Navigator，尽管后者的市场占有率正在下降。随着技术的发展，这些浏览器的最新版本除了支持浏览 WWW、电子邮件和其他网络应用外，浏览器已经进化到可以处理流式音频、视频和各种其他令人爽快的特征。新版本的浏览器都支持最常用的 HTML 元素。当然，两种浏览器都支持 HTML4 标准。本书所介绍的绝大多数生物信息学应用均只需采用浏览器即可完成。

1.2.1.4 重要的 Web 技术

前面我们已经了解了关于 Web 的一些核心组成部分，例如 HTTP、HTML、URL 和浏览器等。此外，Web 的发展离不开一些重要的技术应用。正是这些不断更新的技术推动了 Web 向我们无法预料的方向发展。下面我们就简要介绍一下这些技术，其中绝大多数也在当前的生物信息学服务中提供了主流的支持和应用。

▶CGI、Perl 和 Bioperl

在图 1 - 4 中，我们概括了基于客户机/服务器计算模型的 Web 工作框架。用户通过客户端的计算机采用浏览器对 Web 服务器提出请求，服务器则根据客户端要求抓取数据以作回应，浏览器再根据服务器所返回的各类数据包括文本、图像、声音和视频等属性以超文本的方式显示出来。在这个过程中，我们已经了解到服务器和客户端之间采用 HTTP（Hyper-Text Transmission Protocal）为通讯协议，以统一资源定位器（Uniform Resource Locator，URL）为地址体系（Addressing Scheme），以便在互联网中要取得数据时能快速确定数据的位置与取得方式。WWW 中的各类资源和服务则通过 HTML 文件内的超链接来彼此联系。问题在于，有时候我们不仅仅只需要查看某些文本或图片，更多的时候我们会希望 Web 服务器能够对我们输入的一些数据作出响应。例如我们通过搜索引擎输入一些关键词查询信息，搜索引擎服务器则返回符合这些关键词的查询结果，这个过程便是一个最简单的客户端和服

务器彼此互动的例子。在这种情况下，光靠 HTML 已经不能解决问题。因为 HTML 超文本是一种不可编程的静态文本，它的设计原则是显示数据和文件，而不是处理数据，因而无法用来产生动态的信息，缺乏交互性。为了解决通过 Web 产生的动态服务，产生了 CGI 技术。CGI（Common Gateway Interface），即所谓公共网关界面，其实是一种编程标准，它规定了 Web 服务器调用其他可执行程序的接口协议标准，是 HTML 的功能扩展。灵活易用的 CGI 程序与 HTML 的结合实现了交互式的动态通信。简单地说，它接受 Web 浏览器发送给 Web 服务器的信息，并进行处理，然后将结果再送回给 Web 服务器及 Web 浏览器。例如，用户为了存取某个数据库中的信息，首先通过浏览器将请求以 html 格式传给 Web 服务器，而服务器则通过专用的 CGI 程序来承担了所有的查询、计算工作，将结果构建成一个 HTML 文档反馈给 Web 服务器，再将 HTML 文档传给客户端浏览器。因此，CGI 实际上就成为连接用户和数据库服务器的一座桥梁。通过这座桥梁，我们通过 Web 已经不再局限于观察一些静态的文字信息，还可以一种简洁、可控制的方式来进行一些诸如查询数据库、生成定制的图像以及更复杂的生物信息学计算等工作。

今天，大多数 Web 站点都采用 CGI 技术来生成和传递动态内容。作为 Web 服务器端的一个通用接口程序，基本上所有的计算机语言都可以用来扩展 CGI 程序，最常用的几种包括 C/C + +、Perl 和 Visual Basic。但是到目前为止，Perl（Practical Extraction and Report Language 实用摘录和报告语言）可以说是最受欢迎的 CGI 编程语言。在最初的十年间，Perl 主要是面向文本处理，现在它已经成为一种强大的面向对象语言，为 Web 开发者所青睐。CGI 程序员们喜爱 Perl 的文本处理能力和 CGI. pm 模块，它们为几乎所有 CGI 相关的任务提供了良好集成的、面向对象的接口。很多人认为 CGI 和 Perl 是推动 Web 发展的工具，这种语言已得到全世界开发人员和 Web 编码者狂热的支持。而对于生物信息学而言，Perl 不仅是生物信息学界中很热门的一种编程语言，甚至已经被夸大为人类基因组计划的“救星”。著名的生物信息学家 Lincoln Stein 发表了一篇题为《Perl 是如何拯救人类基因组计划的?》的文章，在该文中，作者认为 Perl 语言是编写 CGI 引擎的完美语言，并竭力描绘了 Perl 在基因组序列分析、数据整合以及不同测序小组之间的数据交换之间的地位和作用。他说当基因组计划奠基在一片不兼容的资料格式的海洋，Perl 拯救了这一切。虽然它并不完美。Perl 似乎明显地满足了许多基因组测序中心的需要，而且经常是当我们遇到问题时第一个想到的工具。

当然，对于大多数分子生物学家而言，或许并不需要对 CGI 了解得比浏览器更多，但对于那些打算致力于开发一些生物信息学软件并通过 Web 提供给其他科学家分享的人而言，应该花点时间对 CGI 有更多的了解和学习。感兴趣的读者可以通过互联网找到更多的关于 CGI 和 Perl 的学习资料及应用范例。关于 Perl 在生物信息学的应用，还可参考一本在生物信息学爱好者中广为流传的 Perl 经典教程《Beginning Perl for Bioinformatics》（James Tisdall 著）以及一个叫做 Bioperl 的计划。Bioperl 是一个非盈利的学术组织，正式成立于 1995 年，旨在开发用于生物信息学、基因组学和生命科学研究的开放源码的 Perl 工具，促进 Perl 在生物信息学中应用。从 1995 年至今，经过 8 年的发展，Bioperl 现在已成为一个令人瞩目的国际性自由软件开发计划，并且获得了国际开放生物信息学基金（Open Bioinformatics Foundation）全力支持。通过访问 Bioperl 的 Web 服务器 www. bioperl. org，我们可以获得许多免费的可用于生物信息学开发的 Perl 模块和脚本（图 1 - 6）。最近 Bioperl 更是推出了包含大量生物信息学工具的 Bioperl 1.2 版本软件包。Bioperl 1.2 包括 1091 个文件，116 脚本