

1859

12 Ettner N, Hillen W. J Am Chem Soc, 1993; **115**: 254613 Chin J. Acc Chem Res, 1991; **24**: 14514 Sutton P A, Buckingham D A. Acc Chem Res, 1987; **20**: 35715 Burgeson I E, Kostic N M. Inorg Chem, 1991; **30**: 429816 Zhu L G, Kostic N M. Inorg Chem, 1992; **31**: 399417 Zhu L G, Kostic N M. J Am Chem Soc, 1993; **115**: 456618 Zhu L G, Lin Q, Kostic N M. J Am Chem Soc, 1994; **116**: 5218

Research on Protein Specific-Cleavage Reagents. Sun Ziyong, Zhou Song, Hu Jianzhong, Zhu Dexu (*Biochemistry Department, Nanjing University, Nanjing 210093, China*); Chen Xiaohua, Zhu Longgeng (*Cooperation Chemistry Institute, Nanjing University, Nanjing 210093, China*).

Abstract Protein specific-cleavage reagents refer to those chemical tools which can cut off peptide-bonds specifically. These reagents can

be divided into two classes. The first is called oxidative cleavage reagents. Oxidative cleavage has the property of stereo-specificity. Only the peptide bonds which close to cleavage system in tertiary structure are cleaved. The second is called hydrolytic cleavage reagents. These complexes can catalyze the hydrolysis of specific peptide bonds directed by protein sequence specificity. Protein specific-cleavage reagents are useful for protein sequence analysis, for studies on the relationship between structure and function of protein and for synthesis of new chemical therapy drugs. The development of these reagents will improve the studies on protein chemistry and other relative subjects.

Key words protein specific-cleavage, oxidative cleavage, hydrolytic cleavage

电子通讯与生物大分子序列分析

王槐春

(军事医学科学院医学情报研究所, 北京 100850)

摘要 交互网络 (Internet) 的发展为联网的计算机用户之间进行信息交流提供了有效途径. 就分子生物学家而言, 他们不仅可以利用电子邮件系统发送和接收信息, 而且更重要的是能够存取大量的分子生物学数据库和软件. 利用 Internet 可以开展多种序列分析作业, 包括序列数据库的类似性检索、基因编码区鉴定和蛋白质二级结构分析等. 一个数据库, 例如 GenBank, 可以通过多种方式来存取: a. 电子邮件文件服务器, b. 文件传送协议 (FTP), c. Gopher, WAIS 或 WWW 等服务器-客户机 (Server-Client) 系统. 专为分子生物学家设计的 BIOSCI 电子公告牌为研究人员开展学术讨论、寻求别人帮助和与数据库人员交流提供了极大的方便.

关键词 电子通讯, 交互网络, 序列分析, 数据库

随着核酸和蛋白质序列数据的急剧增加和分子生物学家对最新序列数据的需要, 用磁带、磁盘甚至光盘已不能满足大量数据的存贮和数据库迅速更新的要求, 另一方面长期维持订购一套 (或几套) 核酸和蛋白质数据库、购买不

断涌现的新的序列分析软件也是一项巨大的开支^[1]. 近年来, 随着全球性信息高速公路的建设, 越来越多的分子生物学数据库和软件与国际计算机网络系统连结, 任何一台与之连网的

计算机都可以利用这些软件和信息资源. 用户不但能检索到最新的数据库, 而且还能利用网络上的软件开展多种序列分析研究. 因此, 掌握计算机与网络的通讯技术已成为90年代分子生物学家必须具备的基本手段.

1 使计算机连网

进行电子通讯的前提是使你的计算机入网. 几乎任何一家大的科研机构或大学的计算中心都与某个国际网 (Internet, Bitnet 或 Janet) 相连, 因此如果你在或邻近这样的单位, 最简单的办法是询问当地计算中心能否使你的计算机通过他们入网. 例如, 军事医学科学院计算中心计算机已通过中国科学院高能物理研究所的计算机与 Internet 连网, 而该院其它的计算机用户 (包括本文作者的 PC) 可以通过与本院计算中心连网而连结到 Internet. 如果用户所在的区域不具备这样的条件, 他可以通过电话线路和调制解调器直接与高能所的计算机连网后进入 Internet. 一旦你的计算机与 Internet (或其它国际性网络) 连通, 它就与世界各地的1000多万台电脑联系起来, 这样在你的面前就呈现了一个巨大的全球性电子图书馆, 而你则拥有进入图书馆的“阅览证”^[1].

2 通过电子邮件开展序列分析

电子邮件 (e-mail) 是一种最简单的电子通讯方式, 电子邮件具有与传真相同 (甚至更快) 的速度, 而且费用比传真要便宜得多. 此外, 由于 e-mail 是计算机与计算机之间的通讯, 因此它还是更加交互式的信息交流方式, 而且是以电脑可读的文件方式传递信息, 因而它在数据库服务和软件交流方面有着 FAX 或邮政通信所不能取代的功能. 这些特点正是电子通讯用于序列分析的基础.

能接受、处理和发送电子邮件的主计算机称为 e-mail 服务器 (server). 目前, 几乎所有的分子生物学数据库中心都建立了这样的服务器, 向 e-mail 用户免费提供序列检索和序列分析服务, 服务器的唯一要求是用户必须按规定

的格式发送与序列有关的请求. 而这些格式可以通过向服务器发出要求“帮助”的信息来获得. 对于大多数服务器, 该信息只有一个词即“help”. 当服务器接收到“help”信息后, 即返回给用户有关如何编辑一个格式化序列检索或分析请求的文件.

美国国家医学图书馆的全国生物技术信息中心 (NCBI) 提供了2个能检索 GenBank, EMBL 和 PIR 等数据库的 e-mail 服务器: RETRIEVE server (检索数据库记录) 和 BLAST server (序列对数据库的类似性检索). 假设用户 zhousb (e-mail 地址是 zhousb@bepc2.ihep.ac.cn) 要通过 NCBI 检索人 IL-2 的 DNA 序列, 他事先不知道发这样请求的格式, 他就可以发送一封 e-mail 到 RETRIEVE server 请求帮助:

```
From: zhousb@bepc2.ihep.ac.cn
Subject: For help
To: retrieve@ncbi.nlm.nih.gov
help
```

正文部分只有一个词“help”. 第二天他就会收到该服务器发回的邮件, 告诉他应该按如下方式准备检索序列的文件:

```
DATALIB genbank
BEGIN
interleukin-2 AND human
```

其中用大写字母表示的词为保留字.

当向 RETRIEVE server 发送上述文件后, 很快, zhousb 就收到了检索结果, 共292个有关 IL-2 的序列记录.

同样, 为了利用 BLAST server 进行序列对数据库的类似性检索, 也可以先发一个“help”请求到 blast@ncbi.nlm.nih.gov, 它会告诉你按如下方式准备序列:

```
PROGRAM blastp
DATALIB pir
EXPECT 0.40
BEGIN
>matpgsrlmnstwcw.....
```

当把编辑好的文件发送到 BLAST server,

后者在2~3 d 内返回数据库类似性检索的结果。

除掉上述两项序列服务功能外,一些 e-mail 服务器还提供其它的序列分析作业,如基因编码区预测、蛋白质二级结构预测和进化树分析等(表1)^[2]。

从表1还可以看出,用 e-mail 进行序列检索或分析实际上把作业提交给服务器来完成,用户可以无需知道有关程序的具体操作步骤,这显然对分子生物学家提供了极大的方便,然而用户对程序的运行不能直接控制,因而对结果会产生某些影响。

表1 用于序列分析的 e-mail servers

主题	Server *	功能
类似性检索	blast@ncbi. nlm. nih. gov	蛋白质/DNA 序列关于蛋白质/DNA 数据库检索
	blitz@embl-heidelberg. de	蛋白质序列关于 SWISS-PROT 数据库检索
	block@howard. fhcrc. org	蛋白质/DNA 序列关于蛋白质家族数据库检索
	fasta@embl-heidelberg. de	蛋白质/DNA 序列关于蛋白质/DNA 数据库检索
	fileserv@nbf. georgetown. edu	蛋白质/DNA 序列关于 PIR 数据库检索
	pythia@anl. gov	蛋白质序列关于人类重复序列数据库检索
	quick@embl-heidelberg. de	DNA 序列关于 DNA 序列数据库检索
	geneid@bir. cedb. uwf. edu	基因预测
基因预测	grail@ornl. gov	编码区预测
	netgene@virus. fki. dth. dk	内含子剪接位点预测
蛋白质分析	predictprotein@embl-heidelberg. de	二级结构预测
	cbrg@inf. ethz. ch	序列对准比较, 进化树, SWISS-PROT 数据库检索, 等等

*欲了解更详细的信息,可发“help”请求到该服务器。

3 文件传送协议

电子通讯的另一种方式是文件传送协议

(FTP),它允许用户与网上远距离的计算机联结并传送文件。进行 FTP 远程登录时,一般使用“anonymous”(不记名)作为用户名(因此

表2 一些主要分子数据库中心的 FTP 位点

地址	内容	联系人
ncbi. nlm. nih. gov 130.14.25.1*		Scott Federhen federhen@ncbi. nlm. gov
/repository	20多个数据库,包括 SWISS-PROT, PROSITE, CarBank/CCSD, REBASE, EPD, TFD, LiMB, FlyBase, ACeDB (A C. elegans database)	
/toolbox	NCBI 用于开发软件的一套软件和数据交换规定,包括 ASN.1工具和分子序列数据的规定	
/pub	各种公共域软件,如 Authorin, BLAST, FASTA, MACAW 等	
embl-heidelberg. de 192.54.41.20	多个数据库,类似于 NCBI 的/repository	Rainer Fuchs fuchs@embl. bitnet
ddbj. nig. ac. jp 133.39.128.2	GenBank, EMBL, DDBJ, PIR, SWISS-PROT, LiMB	Takashi Gojobori tgojobor@genes. nig. ac. jp
pdb. pdb. bnl. gov 130.199.144.1	蛋白质结构数据库 (PDB)	
mendel. welch. jhu. edu 128.220.59.42	人类基因组数据库 (GDB)	

*用数字表示的 Internet 地址,下同。

FTP 常常被称为 anonymous FTP), 用户的 e-mail 地址为保密字. 当进入远程计算机的 FTP 目录后, 用 get 命令就可输出该目录下的文件, 用 put 命令向该目录输入文件. FTP 是传送软件以及数据库的十分方便的途径. 许多分子生物学数据库中心或软件开发者建立了 FTP server (表2)^[3], 用户可以通过它们免费得到完整的数据库或软件.

4 网络信息工具

为了更有效地利用 Internet 网上各种信息资源, 计算机专家开发了网络信息检索工具, 其中为生物学界所广泛采用的有3个: Gopher, WAIS (wide area information server) 和 WWW (world-wide web), Gopher 最初于1991年由美国明尼苏达大学开发的旨在该校范围内发布信息的软件工具, 很快它发展成为在整个 Internet 网上传播信息的工具. 与 e-mail 或 FTP 服务器相似, 贮存信息并可通过 Inter-

net 提供 Gopher 服务的主计算机称为 Gopher server, 而用户的计算机上必须装有 Gopher client 软件. Gopher 服务器以菜单方式在网上传送信息, 而用户的 gopher client 上呈现的是按层次分布的菜单. 这样, 用户通过逐级选择菜单上的命令而获得所需要的信息, 无需了解如何使用 FTP 或不必记住众多的计算机地址和保密字. 例如, 当用户进入蛋白质数据库 PDB Gopher 后, 用户的计算机即显示一系列的命令, 如果选择 “The PDB Anonymous FTP” 就可以象使用 FTP 一样从 PDB 调出蛋白质原子坐标的记录; 如果选择 “An (almost) full-text search of the PDB Bibliographic Headers”, 便可使用任何关键词检索 PDB, 如用数据库标识符、作者或复合物名. 同样, 你可退出 PDB Gopher 而进入 GDB、EMBL 或 PIR Gopher—许多分子生物学数据库或其它软件资源都可以利用相应的 Gopher server-client 系统存取. 表3列出其中的一些服务器位点^[4].

表3 与序列分析有关的 Gopher 位点 (BioGopher)

Server 地址	主题	评论
gopher. gdb. org	基因组信息 (GDB)	在 Internet 网上 “漫游” BioGopher 的好起点 检索基金项目和资助信息 查找科学家的 e-mail 地址和电话号码
ftp. bio. indiana. edu	Indiana 大学 生物学库	分子生物学信息和软件 网络消息组如 Bionet 了解 Gopher 服务的好地方
gopher. nih. gov	NIH	NIH 信息, 包括卫生和临床信息、 基金和序列数据库
pdb. pdb. bnl. gov	PDB 信息	检索和传输蛋白质结构数据
ftp. embl-heidelberg. de	EMBL 生物信息资源	调出公共域中的科学软件
genome. stanford. edu	酵母数据库信息	
eatworms. swmed. utexas. edu	线虫	种株、等位基因和图象

WAIS、WWW 和 Gopher 相似, 也分别是一对 server-client 网络信息检索系统, 其中 WWW client 不仅能存取 WWW 服务器所贮存的文件和信息, 而且还能检索 WAIS、Go-

pher、X500等系统上的数据, 因而具有更强的检索 Internet 信息的功能. 目前, NCBI 和瑞士日内瓦大学分别建立了 NCBI 数据库和 SWISS-PROT、SWISS-2D PAGE 数据库的

WWW 服务器^[5].

5 电子公告牌

我们都知道“黑板报”或“公告牌”是张贴各种通知和消息的地方,其内容五花八门,从书讯、影讯、会议消息到招聘、求职、寻人和招领启事,无所不包,而且人人都可以张贴. 一则消息可以保留一段时间,然后被刷新,换上新的消息. 同样,电子公告牌(Bulletin Board)是用电子通讯手段“张贴”各种公告或消息,它的优势在于能迅速接近范围更广和远距离的“读者”,使之成为极有力的信息交流工具. 用户在第一次使用电子公告系统时,必须先报名参加某一个他所感兴趣的公告牌,这一过程叫订阅(subscription). 对于分子生物学家,最有用的是生物学家公告牌(Biologists' Bulletin Boards,简称 BIOSCI BB),其 Internet 地址是 biosci@net. bio. net, Janet 地址是 biosci@daresbury. ac. uk.

BIOSCI 公告牌还根据生物学研究人员的兴趣和需要,把它张贴的信息分成一系列的消息组(BIOSCI newsgroups). 表4列出与序列分析有关的消息组^[6]. 事实上,用户正是通过张贴某一论题,吸引众多的读者参加讨论,从中得到帮助或问题的解决. 对分子生物学家有用的消息组还有 Bio-journals、Methods-Reagents 和 Employment 等. Bio-journals 消息组直接从出版者那里向用户发布早于实际发行的32种生物学研究期刊的目录; Employment 消息组提供的是招聘或求职的信息.

当一个实验者急需某一种关键的实验试剂,而他所在的实验室或周围的实验室都没有,而更不幸的是他被告知该试剂的生产厂商已不再制造这种试剂且没有库存,此时,如果他张贴一条请求帮助的信息到 Methods-Reagents 消息组,在一、两天之内,就会有人发来 e-mail 告诉他什么地方有这种试剂,甚至还有人愿意提供给他一些. 同样,当研究人员急需质粒载体或某种免费软件时,也可经张贴请求信息,并能很快地得到答复. 与其它形式的广告所不同

的是,电子公告牌所提供的任何一项服务都是免费的. 此外,电子公告牌上的消息资源也可由 Gopher、WAIS 或 WWW 等系统来检索和存取.

表4 与序列分析有关的 BIOSCI newsgroups

Newsgroup	主 题
Biological information theory	信息理论用于生物学
Bio-journals	生物学期刊的目录列表
Bio-matrix	计算机用于生物学
Bio-software	生物科学软件信息
Chromosome 22	人类第22号染色体制图和测序
Computational biology	数学和计算机用于生物学
EMBL databank	与 EMBL 数据库人员的往来信息
Employment	生物学领域的工作机会
GDB	与 GDB 数据库人员的往来信息
GenBank BB	与 GenBank 数据库人员的往来信息
Human genome program	NIH 资助的关于人类基因组问题的消息
info GCG	关于 GCG 序列分析软件的讨论
Methods and reagents	方法和实验室试剂
Molecular evolution	关于分子进化研究的讨论
Protein analysis	关于蛋白质研究的讨论以及给 PIR 和 SWISS-PROT 人员的信息
Protein crystallography	关于大分子晶体学的讨论以及给 PDB 人员的信息

6 小 结

Internet 为连网的计算机用户之间铺设了进行信息交流的有效通道(现在,还不能说是高速公路). 就分子生物学家而言,它不但具有图文传真那样的快速通信能力,而且更重要的是籍此可以存取最新的分子生物学数据库和软件资源. 通过 Internet 还可开展多种序列分析,例如序列对数据库的类似性检索、基因编码区鉴定和蛋白质二级结构预测等等. 接受用户发来的信息、对信息进行处理并将结果返回给用户的计算机称为服务器. 一个数据库,如 GenBank,可以用多种方式来存取:①e-mail 文件服务器. 如 NCBI 的 RETRIEVE server 根据用

户的检索词检索 GenBank, 并把检出的序列记录通过 e-mail 返回给用户; BLAST server 进行序列对 GenBank 的类似性检索. ②FTP 实现文件在用户和服务器之间的传送. 据此用户可免费获得一个完整的最新版本 GenBank 或最近一周释放的 GenBank 新记录 (GenBank update). ③Gopher、WAIS 和 WWW 服务器. 它们与用户计算机上安装的 client 软件组成相应的 server-client 软件系统, 进行数据库检索或者实现文件传送. Internet 的电子公告牌, 例如生物学家感兴趣的 BIOSCI 公告牌为分子生物学家进行专题学术讨论、寻求别人帮助以及与数据库人员的交流提供了极大的方便.

致谢 本项工作得到本院计算中心周士波高级工程师的帮助. 作者的 e-mail 地址是: zhousb@bepc2. ihp. ac. cn.

参 考 文 献

- 1 王槐春编著. 蛋白质与核酸序列分析基础. 北京: 人民军医出版社, 1994: 1
- 2 Henikoff S. Trends Biochem Sci, 1993; 18: 267
- 3 Gribskov M, Devereux J. Sequence Analysis Primer. New York: Stockton Press, 1991: 1
- 4 Parker M. Trends Biochem Sci, 1993; 18: 485
- 5 Apple R D, Bairoch A, Hochstrasser D F. Trends Biochem Sci, 1994; 19: 258
- 6 Bleasby A, Griffiths P, Hines D *et al.* Trends Biochem Sci, 1993; 18: 310

Electronic Communication and Sequence Analysis. Wang Huaichun (*Institute of Medical Information, Beijing 100850, China*).

Abstract Electronic communication through Internet has provided on-line computer users an effective way for information exchange. For molecular biologists, they can not only use the electronic mail (e-mail) system to send and receive message as fast as fax, but also more importantly, can access large numbers of new molecular biology databases and softwares. They can perform various sorts of sequence analysis tasks, including homology search against databases, gene coding region identification and protein secondary structure analysis, etc. A database, such as GeneBank, can be accessed through the following ways: (1) e-mail file servers, (2) file transfer protocol (FTP), and (3) Gopher, wide area information server (WAIS) or world-wide web (WWW). The BIOSCI bulletin board, specially for the interest of molecular biologists, gives great convenience for molecular biologists to hold discussions of scientific topics, ask for others' help and communicate with database staffs.

Key words electronic communication, Internet, sequence analysis, database

内含子对提高转基因动物基因表达效率的影响

卢一凡 邓继先 肖成祖 马清钧

(军事医学科学院生物工程研究所, 北京 100071)

摘要 综述了内含子在转基因动物基因表达中的作用, 对内源性内含子和异源内含子对转基因表达的影响进行分析, 讨论了内含子提高转基因动物基因表达效率的三种可能机制, 并指出在表达载体构建中应考虑到内含子的重要性.