生物信息学:生物实验数据和计算技术结合的新领域

欧阳曙光 贺福初*
(军事医学科学院放射医学研究所,北京 100850. * 联系人)

摘要 大量的蛋白质和核酸数据的积累与理性地分析这些数据中所蕴涵的生物学意义的双重需要,产生了综合生物学研究与计算技术研究等领域最新成果的交叉性学科"生物信息学".分别从基因序列或蛋白质结构等生物信息数据库、基因组分析或蛋白质结构分析等常规生物学计算软件、基因组数据库检索或蛋白质空间结构识别与预测等在线生物学计算服务器、人工生命等几个方面,概述了发展中的生物信息学的最近动态和有关信息,并同时提供了相关的热门生物信息学站点和资源在互联网上的超文本或文件传输协议地址. 此外,还介绍了讨论组、新闻组等其他形式的医学、生物学、信息学资源. 最后,就生物信息学存在的问题与前景进行了讨论,指出生物信息学将是一次国际性的科学大协作,也是我国生命科学振兴的一个新契机.

关键词 **生物信息学 生物信息 分子生物学 计算机科学 互联网**

1 概述

蛋白质与核酸测序技术应用以来,已积累了极大量数据. 同时,基于典型西方哲学演绎与解析的分析思路而建立的组合化学数据库已经成为合理分子设计(rational molecular design)的重要支柱,为创造全新的非自然产物提供了可能. 所以,将新颖的计算技术与方法应用于经验和理论生物学研究的时代已经到来,生物信息学由此诞生. 但生物数据的海量性和复杂性又都是组合化学等其他数据密集型科学所不及的,这也是生物信息学所面临的更大挑战.

一般意义上,生物信息学研究生物信息的采集、处理、存储、传布、分析和解释等各个方面,它通过综合数学、计算机科学与工程和生物学的工具与技术而揭示大量而复杂的生物数据所赋有的生物学奥秘. 它作为一个交叉学科领域而荟萃了数学、统计学、计算机科学和分子生物学的科学家,目标就是要发展和利用先进的计算技术解决生物学难题. 这里所说的计算技术至少包括机器学习(machine learning)、模式识别(pattern recognition)、知识重现(knowledge representation)、数据库、组合学(combinatorics)、随机模型(stochastic modeling)、字符串和图形算法、语言学方法、机器人学(robotics)、局限条件下的最适推演(constraint satisfaction)和并行计算等. 而生物学方面的研究对象覆盖了分子结构、基因组学、分子序列分析、进化和种系发生、代谢途径、调节网络等诸多方面.

许多研究与发展组织都预测:基因组学研究将会彻底革新未来鉴定生物学产物和选择更佳目标用于小分子生物功能筛选的过程. 随着基因组研究规模扩大,生物信息学将原始序列数据转换为有意义的生物学信息之重要性也随之增长.

严峻的挑战和巨大的机会往往出现在相同的时间和地点. 生物信息学各个分支都亟待改进和提高的3个方面是:更加有效地处理大规模的数据、建立通用的智能型工具、使所有的操作程序自动化.

2 生物信息数据库

目前,国际性合作的几个基因组计划已经积累了超大量的生物信息并以不同组织形式构成许多数据库. 其中一些属于商业数据库需要预先注册和付费才能检索,而更多数据库是公开和免费的并可通过互联网络(Internet)访问. 随着研究深入,公共数据库越来越成为世界各地生物学家的重要给养.

美国国家实验室(Brookhaven National Laboratory, BNL)的蛋白质数据库(Protein data bank, PDB)可同时提供蛋白质序列及其三维空间晶体学原子坐标. 其中受体-配体、抗原-抗体、底物-酶复合物等相互作用分子的共结晶图谱是基于同源比较的分子设计所需的最佳模型,因此PDB为初步的蛋白质合理设计提供了无价的知识来源. 其超文本传输(hyper text transfer protocol)地址为http://www.pdb.bnl.gov/,文件传输(file transfer protocol)地址为ftp://ftp.pdb.bnl.gov/pub/databases/pdb/all_entries/compressed_files/. PDB在几个世界著名科研机构所在地设有镜象站点(mirror site),如欧洲生物信息学研究所(European Bioinformatics Institute, EBI)的http://www.ebi.ac.uk/pdb/和ftp://ftp.ebi.ac.uk/pub/databases/pdb/,北京大学物理化学研究所的http://162.105.177.12/npdb/和ftp://162.105.177.12/fullrelease/compressed_files/等.

超文本版本的细胞系数据库(Hypertext version of the cell line data base, HyperCLDB)专门提供欧洲各家实验室和捐献站的人和动物细胞系的信息超过. 目前已有3100种以上的品系,在其说明中能查到可以从哪些实验室获得,并显示每个术语或数值在总词汇表和索引表中的出现频率. 还有指向在线人类孟德尔遗传(Online mendelian inheritance in Man, OMIM)记录的链接,提供较为深入的病理学知识,从病理学家名录到与某个特定病理过程相关的细胞系资料. 直接指向URL提醒系统(reminder system)的链接可在所注册的网页更新时就用电子邮件提醒用户. HyperCLDB的搜索引擎在http://www.biotech.ist.unige.it/tab/HyperSearch.html.

OWL混合蛋白质序列数据库(Composite protein sequences databases)是一非重复蛋白质序列数据库,其数据来源包括(截止到1998年6月以前的统计):(1) 含有69 110个分子25 083 142个残基的第35版Swiss-Prot; (2) NBRF的含有393个分子235 554个残基的第55版PIR1, 45 067个分子12 796 251个残基的第55版PIR2, 357个分子69 696个残基的第55版PIR3, 164个分子27 699个残基的第55版PIR4; (3) 含有134 190个分子41 324 437个残基的第105.0版GenBank; (4) 含有1 233个分子236 843个残基的第23.0版NRL_3D,每项条目都可以在BNL的X线晶体结构数据库中查到,其代码为NRL_开头再加上4个字符的PDB代码. 全部入库序列数已达到25 0514个分子79 773 622个残基. 它的WWW地址为http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/owlcontents.html.

欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL)的TREMBL是对Swiss-Prot蛋白质序列数据库的增补,含有EMBL核酸序列数据库中尚未出现于Swiss-Prot的所有编码区(CDS)的翻译序列,可以看作是Swiss-Prot的前言部分,今后都可能升级到标准Swiss-Prot中,故而全分配有Swiss-Prot访问代码. 目前的第3版TREMBL源于第50版EMBL核酸序列数据库,有126 995条序列34 178 645个氨基酸残基. 它分成两个部分: SP-TREMBL(104 865)是肯定要转入Swiss-Prot的,包含fun.dat(真菌)、hum.dat(人)、inv.dat(无脊椎动物)、mam.dat(其他哺乳动物)、mhc.dat(MHC蛋白)、org.dat(细胞器)、phg.dat(噬菌体)、pln.dat(植物)、pro.dat(原核生物)、rod.dat(啮齿动物)、vrl.dat(病毒)、vrt.dat(其他脊椎动物)等

文件, 已经可以在 EBI 的 FASTA 服务器上搜索, 不久也将能在 BLITZ 服务器上搜索; REMTREMBL 则是不准备收入 Swiss-Prot 的其他数据. TREMBL 站点位于 http://www.ebi.ac.uk/srs/srsc/ 和 ftp://ftp.ebi.ac.uk/pub/databases/trembl/.

与生物催化和生物降解相关的数据库站点有: UM-BBD, 即 Minnesota 大学生物催化和生物降解数据库(University of minnesota biocatalysis/biodegradation database), 提供关于微生物酶与代谢通路的信息, 位于 http://dragon.labmed.umn.edu/~lynda/index.html; EcoCyc, 大肠杆菌基因和代谢百科全书(Encyclopedia of *Escherichia coli* Genes and Metabolism), 是一个汇集了所有已知的关于大肠杆菌基因和中间代谢的数据的大型知识库, 它位于 http://www.ai.sri.com/ecocyc/ecocyc.html; GenoBase Selkov EMP, 是 GenoBase 数据库通道(GenoBase Database Gateway) 中一个经过索引的、关于酶与代谢通路(Enzymes and Metabolic Pathways) 的数据库, 处于 http://specter.dcrt.nih.gov:8004/Pathway/pathway_toc_by_name.html; KEGG, 日本的基因和基因组京都百科全书(Kyoto Encyclopedia of Genes and Genomes), 内容包括代谢通路图谱、分子编目表、基因编目表、基因组图谱等数据, 它被放置于 http://www.genome.ad.jp/kegg/kegg1.html; SoyBase, 是植物基因组计划(Plant genome program) 中的一部分 ——花生计划(Soybean roject) 研究数据的集合, 可以在 http://probe.nal.usda.gov: 8000/plant/aboutsoybase.html 看到详细内容; Swiss-Prot, 是带有注释的、具有最小冗余的、与其他数据库的整合度很高的蛋白质序列数据库, 在 http://www.expasy.ch/sprot/sprot-top.html; 以及 WIT (What is there), 是一个基于最近的关于细菌全基因组序列的足够了解、在 WWW 上设计实现的交互式代谢重构模型, 它位于 http://www.cme.msu.edu/WIT/.

最新的整合型鼠基因组的遗传图谱和物理图谱数据库(Genetic and physical maps of mouse genome data) 第 14 版已经被放在了 http://www.genome.wi.mit.edu/cgi-bin/mouse/ index. 位于右侧的鼠遗传图谱包括了定位于 Ob x Cast F2 杂交系的 6331 种简单序列长度多态性(Simple sequence length polymorphism, SSLP), 平均分辨率 1.1 cM. 位于左侧的 Copeland/Jenkins 图谱包括了定位于 Spretus 回交系的 2 342 个分子标记, 将近多一半的是 SSLP, 另一半的是 RFLP, 既可以分子标记的名称, 也可以分子标记的位置/多态性进行检索. 鼠 STS 物理图谱包含了来源于平均插入片段长度约为 820 kb 的酵母人工染色体(YAC) 克隆文库的超过 6 000 种的 STS, 可分别以分子标记的名称、YAC 的位置或 YAC 的名称进行检索.

位于 http://www.mpimg-berlin-dahlem.mpg.de/~andy/GN/ 的基因组导航者(Genome navigator) 是提供到达含有关于人类基因组、鼠基因组和酵母基因组等的物理图谱和遗传图谱信息的主要数据库的视化的交互式通道. 它使用基于 Java 小控件(applet) 的通用性程序 DerBrowser 来显示和导引这些生物的多种不同类型的基因组图谱. 除了常规功能以外, 它的一个特别之处就是还能让用户查询外部的相关数据库中存在的任一图谱, 目前的数据来源已经包括: 麻省理工学院(Massachusetts Institute of Technology, MIT) 基因组研究中心的 Whitehead 生物医学研究所(Whitehead Institute for Biomedical Research, Whitehead/MIT), 约翰·霍普金斯大学医学院(Johns Hopkins University School of Medicine) 的基因组数据库(Genome database, GDB), Jean Dausset 基金会(Fondation Jean Dausset) 的人类基因组多态性研究中心(Centre d'Etudes du Polymorphisme Humain, CEPH) 和 Genethon 研究所的 infoclone, 人类基因连锁研究合作中心(Cooperative Human Linkage Center, CHLC), 美国国家生物技术信息中心(National Center for Biotechnology In-

formation, NCBI)的人类转录本图谱(Human transcript map, HTM),以及其他一些专门收录人类染色体信息的数据库;欧洲合作种间鼠回交(European collaborative interspecific mouse Backcross, EUCIB)计划的鼠回交数据库(Mouse Backcross Database, MBx),斯坦福大学医学院(Stanford University School of Medicine)的酵母基因组数据库(*Saccharomyces* Genome Database, SGD),Proteome公司的酵母蛋白质数据库(Yeast Protein Database, YPD),Max-Planck研究所的慕尼黑蛋白质序列信息中心(Munich Information Centre for Protein Sequences, MIPS),全自动分析生物序列的GeneQuiz服务器等.

能提供啤酒酵母(*Saccharomyces cerevisiae*)蛋白质三维结构信息的酵母基因组数据库SGD已经可以在http://genome-www.stanford.edu/Sacch3D/找到. 其特性包括:⑴以基因名称、开放读码框架(ORF)名称、染色体编号、文字等形式检索酵母基因组中任一蛋白质的潜在的结构信息; 使用RasMol或基于Java的显示程序交互地观察结构信息;⑵浏览全部已经收录于PDB结构数据库中的啤酒酵母蛋白质;⑶可以到达NCBI的MMDB, SCOP, Swiss-Prot等面向结构的其他数据库的链接. 总体上看,已知的结构信息还是相当稀少的,目前在酵母基因组中只有12%的蛋白质与已知结构的蛋白质之间呈现显著的序列相似性. 但随着更多的新的蛋白质结构被测定和检测结构相似性技术的改进,这个数字必将增长.

IUBio档案是一个生物数据和软件的档案库,囊括了各种各样的大众化的浏览、检索和传输软件、分子数据、生物学新闻和文件,其互联网地址是iubio.bio.indiana.edu (magpie 129.79.225.200). 分子生物学是这里的焦点,它也同时是果蝇研究数据的一个大本营. 这里维护着可在所有计算机上运行的一些对于生物学挺重要的软件:公共软件使用的分类包括了生物学、化学、科学、应用程序等;而分子生物学部分使用的分类包括了对齐、密码子、自动测序、浏览、一致序列、进化、模式、引物、限制酶、RNA折叠、检索、IBM-PC, Mac, M$win, Unix, Vax等;检索服务包括了GenBank核酸数据库、Swiss-Prot和PIR蛋白质数据库、Bionet新闻组、序列检索系统SRS和SRS-FASTA. 它的专门的果蝇基因组数据库FlyBase位于http://flybase.bio.indiana.edu/ (firefly 129.79.225.202).

3 生物计算

就目前的数学和计算机科学的能力而言,对数据容量达到上十亿字节的数据库进行生物计算仍然是一项很艰巨的任务. 虽然最简单的序列比较可以被简化成字符串匹配的算法,以及将模式识别和神经网络等先进算法也运用其中,但是扩展的和多重的序列比较还是处于试验摸索中. 理论上有希望的、通过量子化学算法预测蛋白质的空间折叠的方法靠现有的计算能力尚无法成为现实, 因为这些都需要数学与纯计算机效能上的新突破.

大分子设计和模建算法让曾经致力于分子力学和分子模型构建的应用数学家、物理学家、化学家和生物学家走到了一起. 现在的重点和挑战在于如何获得高增益、高效率、高可信度的蛋白质、核酸和多聚体的模拟算法. 分子力学的高级时间步长法(advanced time-stepping)、静电学、经典量子力学、结构确定(structure determination)、自由能和整体集群计算(ensemble calculations)等,都是可能的突破点.

分子图形和模型学是生物信息学和药物设计的重要部分. 当基因组学的成就被应用于合理目标鉴别时,蛋白质结构相似性和结构预测、确定蛋白质-蛋白质相互作用、识别类似的和同源的蛋白质折叠等方法都会显著地影响最后的结果. 自动同源模建和结构-功能预测也需要

更多的努力,并利用趋于成熟的神经网络方法来实现.

先是可在本地的个人机或工作站上运行的生物计算软件和程序,下一部分侧重于通过互联网络的在线计算.

(1) 日常数据维护. 为生命科学研究人员实现全方位计算能力的软件工具 Prophet 5.0,提供适合于数据管理和视化、包括从简单描述性的统计处理到多元方差分析(Multi-factor ANOVA),logistic 回归和非线性模型分析等多种统计分析. 它配备了多序列对齐、翻译、限制酶和蛋白水解酶酶切分析、PCR 引物设计、BLAST 检索、远程数据库检索等生物序列分析工具. 全功能的 Prophet 5.0 程序可以从 http://www-prophet.bbn.com/ 下载,可免费使用 60 d. 其支持和即将支持的 Unix 平台包括 SUN/Solaris 2.4, DEC Alpha/Digital Unix 和 Silicon Graphics/Irix 6.2 等. 这个综合性的数据分析软件包以快速易用为特点: 会用鼠标就会用 Prophet.

由 http://www.unizh.ch/vetvir/plugin.html 可下载一些能加快实验室日常工作、用于苹果 Mac 机或 Windows 系统的浏览器(Netscape 2.x 和 Internet Explorer 2.x)的免费插件(plugin)程序,安装后再连接到 http://www.unizh.ch/vetvir/programs.html. 其功能包括: 酶切预览(根据用户使用的限制性酶和 DNA 的核苷酸序列在虚拟的琼脂糖凝胶上电泳,用户可以在到紫外灯箱拍照前知道凝胶上出现条带的理论位置);稀释计算(任何浓度的溶液稀释配比计算);接头设计(得到一个用于插入序列连接的、无自连末端的接头序列)等.

质粒处理器(Plasmid processor)是专门绘制科研与教育用质粒图谱的简单程序. 可以输入线形或圆形质粒,任意定义限制位点、基因位点和多克隆位点,任意插入或缺失部分片段. 输出的质粒图谱可复制到剪贴板上,也可以存盘以便后用,或通过程序内置的打印模块打印. 压缩的程序包 plasp102.zip (约 239 kB)可以从 http://www.uku.fi/~kiviraum/plasmid/plasmid.html 下载.

(2) 序列对齐. 基于"近似字符串匹配(Approximate string matching)"算法的 Cleanup 1.8 能够确定从核苷酸序列数据库中指定的任何一对序列间的整体同源性,并自动从冗余数据库中生成一组纯化的无冗余的核苷酸序列集萃. 冗余问题一向是序列组间比较的关键概念,无冗余序列无疑对进行统计学分析和加快广泛性检索核苷酸序列数据库的速度非常有益. 所有公开的数据库都会存有同一序列或近似于相同序列的多个不同条目,基于这种偏倚数据的统计学分析往往会有很高的将不显著视为显著的危险性. 为了实现无偏倚的统计学分析和进行更有效的数据库检索,必须使用经过纯化的无冗余序列数据. 然而实际操作中对生物序列数据冗余性的定义难免含混、不易确切,Cleanup 就使用了一个基于序列相似性程度的定量指标来描述冗余性: 一旦用户给出一个阈值,那么显示出一定的相似性而且与数据库中的另一较长序列间存在重叠的序列就被认为是冗余序列. 从互联网上下载此程序的地址是 ftp://area.ba.cnr.it/pub/embnet/software/Cleanup/.

大规模序列比较软件包(Large Scale Sequence Comparison Package) LASSAP (位于 http://www-rocq.inria.fr/genome/)是一个跨越多种 Unix 平台(SGI/Irix, SUN/Solaris, IBM/AIX, DEC/Digital Unix 等)的新颖而全面的序列比较软件包. 它使用了目前所有主要的序列比较算法: BLAST, FASTA, Smith-Waterman 动态变程、Needleman/Wunsch 法、K-best 对齐法、字符串匹配(主要针对冗余问题)、模式匹配算法(譬如搜索 ProSite 特征模式)等. LASSAP 中的所有算法都是基于成对比较、且不同算法间的优势能共享以外,还具备: 1) 数据库内或库间比较(数据库既可以是来源于一个大数据库的一套序列,也可以是单独的一条序列);2) 直接计算(选择和计

算部分还有待完善); 3) 序列翻译(可使用不同遗传密码); 4) 结构化的计算结果和强大的再分析能力(支持 3 种输出格式: 含有对齐序列的全文本; 每行一对结果的压缩文本, 便于使用 grep, awk 或 perl 等过滤程序; 结构化的标准格式, 便于继续进行 cluster 等复杂而深入的分析); 5) 并行计算和利用特殊硬件设备而使性能加强(基础版本的 LASSAP 适合一般应用, 优化算法的并行版本则适合处理复杂的大规模问题, 特别是专门用于 Smith-Waterman 算法的优化还利用了 SUN 的视频指令集). 同时, 它提供的应用编程接口(API) 允许用户植入任何其他基于成对比较的算法(公用 API 不久就会发行). 因此, LASSAP 是为满足大规模序列数据分析、克服目前序列比较程序所受限制而设计的可编程的高效应用系统. 使用 LASSAP 的成功范例已有: 蛋白质结构域分析 ProDom 的建立(http://protein.toulouse.inra.fr/), 微生物基因组的穷举比较(Protein Science, Vol 6, Suppl 1, April 1997), TREMBL 中的亚片段匹配问题(Proceedings of ISMB 97 Conference, June, Greece), 等.

蛋白质多序列编辑器(Protein multiple sequence editor) ProMSED2 是运行于 Windows 3.11/95 平台的能自动或手动完成 DNA 和蛋白质序列对齐、编辑、比较和分析的应用程序. 它能读入几种常见格式(NBRF/PIR, FASTA, MSF, EMBL/Swiss-Prot, Intelligenetics 和 Clustal 等) 的序列数据, 自动进行对齐、对齐结果的视化和编辑, 还可以在保持原来对齐区域不变的同时交互地对齐其他部分. 其用户界面友好, 手动对齐和序列分析时用不同的颜色组表示氨基酸序列在突变、理化等性质上相似的位点, 是一套能方便地完成序列的对齐、分析、视化、编辑和制图的小巧而灵活的工具程序. 它的下载地址是 ftp://ftp.ebi.ac.uk/pub/software/dos/promsed/prsed2_.exe.

(3) 分子结构视化. LoopDloop 是一个描绘分子生物学中 RNA 二级结构的程序, 它读入含有碱基配对信息的生物序列数据, 显示出 RNA 分子的二级结构, 并允许对结构进行修饰、美化等编辑. 但是这个程序自己没有预测二级结构中碱基配对的功能, 因为通过 RNAFold, MulFold 和多序列对齐编辑器等其他软件可以完成这种工作. 该软件的下载地址在 http://iubio.bio.indiana.edu/IUBio-Software+Data/molbio/loopdloop/java/, 或 ftp://iubio.bio.indiana.edu/molbio/loopdloop/java/.

仅仅从平面图形提供的信息是很难清楚蛋白质、DNA、RNA 的三维立体结构以及它们之间相互作用的, 而深刻掌握结构又是对理解功能相当重要的. 现在许多免费软件已经有了显示生物分子的醒目而具备深度感与动感的三维立体的空间填充(space filling) 彩色视图的能力. RasMol (http://www.umass.edu/microbio/rasmol/) 就是其中之一, 它能在多种 Windows 平台和 Mac 机上运行, 同时免费提供全部的源程序代码以鼓励改进和自行开发. Chime (http://www.umass.edu/microbio/chime/) 则可以将预先定制的带有注解的分子图谱转换成为新颖的网上教程. 其他各种网上应用的图形工具可以在 http://www.umass.edu/microbio/rasmol/emr-web.htm 得到.

(4) 基因组分析. 蛋白质展开、描述和分析工具(Protein extraction, description and analysis tool) PEDANT 是专门为了实现对全基因组的序列进行计算分析而设计的, 它位于 http://pedant.mips.biochem.mpg.de/frishman/pedant.html. 目前它已经分析了 9 套全部的、1 套质粒的和 2 套部分的基因组: 啤酒酵母(*S. cerevisiae*)、生殖道支原体(*Mycoplasma genitalium*)、肺炎支原体(*Mycoplasma pneumoniae*)、甲烷球菌(*Methanococcus jannaschii*)、粘囊菌(*Synechocystis sp.*)、流感杆菌(*Haemophilus influenzae*)、大肠杆菌(*E. coli*)、幽门螺旋杆菌(*Helicobacter pylori*)、甲烷

细菌(*Methanobacterium thermoautotrophicum*)、根瘤菌(*Rhizobium sp.*)的质粒、部分枯草杆菌(*Bacillus subtilis*)、部分硫叶菌(*Sulfolobus solfataricus*)等. 它以序列比较和序列预测结合起来的组合判别法为工具, 对已经全部测序的基因组上所预测的 ORF 进行穷举形式的功能性和结构性分类; 其 ORF 的功能性预测主要依据于 FASTA2 相似性搜索, 并辅以 ProSite 模式和 motif 检测、与保守序列块的比较等; 最后将序列与最显著相关的 PIR 条目相偶联, 从而并入某一个 PIR 超家族之中. 它还能功能性地依据对几个经过手工归入功能性类群的定性的细菌和酵母的主基因集合(curated master gene set)的相似性检索将基因产物分类, 并通过对赋有二级结构的 STRIDE 数据库的每一个序列运用 Smith-Waterman 相似性比较算法, 预测二级结构、跨膜区域、低复杂性区域和无规卷区区域, 以及抽取出可知的三维结构信息.

(5) 基因模式识别. Procrustes 4.01 是为支持实验性基因判定和提示性的定性基因预测的计算工作而设计的, 它在 http://www-hto.usc.edu/software/procrustes/. 其主要性能有: 运用 Las Vegas 基因预测法的准确无误的基因和 exon 判定; 容错性的基因识别; 基于 GenePrimer 软件的能满足大规模测序工程中利用 PCR 技术进行基因判定的引物构建; 基于 Cassandra 软件的能指导探针和 PCR 引物选择的高特异性 exon 识别; 通过局部切割后对齐从未完成的 cosmid 大小的基因组序列中识别出不完整基因; 新颖的图形输出显示多基因预测和实验性基因判定的结果; 给基因预测打分以体现其可信程度; 利用部分优化的切割后对齐进行多基因预测; 基于相似功能区域而不是整个蛋白质的基因识别; 不同种系的基因识别.

(6) 蛋白质分析. Windows 版的蛋白质分析专家(Protein analyst for windows) ProAnWin 是用于多个蛋白质序列对齐、比较性序列分析、研究蛋白质结构-功能(基因型-属性)关系和设计点突变的一个新程序. 它试图找出蛋白质或多肽的活性(或属性或相关表现型)与分子的一级结构或三级结构中某些特征的关系, 其依据包括: 从序列上看所归属的蛋白质家族, 与蛋白质活性相关的一些参数(pK 值、ED_{50}、K_m 值等), 和尽可能的、至少其中之一的三维结构数据(假设全部同源蛋白质都以共同的方式形成空间折叠). 主要目的就是要找出与蛋白质活性变化相关联的影响因子: 活性调节位点的位置和该位点在结构上的重要特性. ProAnalyst 是为 ProAnWin 提供多功能的蛋白质序列和结构分析的扩展模块, 它可以搜索 motif、绘制理化关系图、对蛋白质的序列变异进行语义分析和理化分析、绘出结构-活性关系的剖析图等. 这一套功能相关的软件的下载地点有: [ProAnWin] ftp://ftp.ebi.ac.uk/pub/software/dos/proanwin 或 ftp://ftp.bionet.nsc.ru/pub/biology/vector/proanwin.dem/paw$.exe, [ProAnalyst] ftp://ftp.ebi.ac.uk/pub/software/dos/proanalyst, ftp://iubio.bio.indiana.edu/molbio/ibmpc/panalys1 或 ftp://ftp.bionet.nsc.ru/pub/biology/vector/proanaly.dem/panalys$, [ProMSED] ftp://ftp.ebi.ac.uk/pub/software/dos/ promsed, ftp://iubio.bio.indiana.edu/molbio/ibmpc/promsed1 或 ftp://ftp.bionet.nsc.ru/pub/biology/ vector/promsed.dem/promsed$.

(7) 蛋白质结构模建. 可以从 http://www.nimr.mrc.ac.uk/~mathbio/a-aszodi/dragon.html 下载的 SGI 版 Dragon 4.17.7 是一个基于"距离几何学(Distance Geometry)"的蛋白质模建程序. 它可以根据所给定的蛋白质序列、二级结构和一套残基间距离的限定矩阵(如果有的话), 预测小分子量可溶蛋白质的三级结构. 如果序列中的一部分结构在多序列对齐中能够找到同源, 就可以试着对比模建(comparative modeling). 它以一个简单的命令行作为人机交互界面, 接受参数和输入文件名等.

(8) 神经网络. 神经网络通过编程模拟神经元的行为, 是生物计算中较新的技术之一. 开始的工作往往是先利用 Genesis, Neuron, XPP 或其他可以在 Unix 工作站上运行的类似的软件包, 建立许多单一的神经元模型, 联接为网络, 并组成神经系统. 接下来是建立亚细胞处理模型, 从模拟简单回路直到大型神经元网络, 甚至构建系统水平的整个大脑的模型. 然后这个神经网络就可以根据需要通过不断的训练和学习来加以完善, 直至可以预测出满意的结果.

4 在线生物计算

(1) 蛋白质家族鉴定. 基因家族鉴定程序网络版(Gene family identification network design) GeneFIND (http://diana.uthct.edu/genefind.html) 是一个综合了几种检索/对齐程序、基于 ProClass 数据库(http://diana.uthct.edu/proclass.html)、提供快速而有意义的、带有充足的家族分类信息的检索结果的数据库检索系统. 它应用了多层次的过滤程序: 先从最快速的 MotiFind 神经网络开始, 接着是 BLAST 搜索、Smith-Waterman 序列对齐(SSearch) 和 motif 模式搜索. 该服务器目前提供了多达 942 种不同蛋白质家族的大规模在线序列鉴定. HTML 形式的检索结果包括: 全局和 motif 得分、针对所有 ProSite 蛋白质种属的所有最为匹配的成员清单、所属 PIR 超家族、motif 模式匹配情况和指向对应 ProClass 家族数据记录的链接.

(2) 蛋白质空间折叠识别. FEBS 蛋白质结构预测 1997 (http://predict.sanger.ac.uk/irbm-course97/) 的先驱者们希望能尽可能多地利用最新的折叠识别和从头预测(*ab initio* prediction) 等方法学上的进展, 对一些具有生物学价值的蛋白质结构进行预测. 如果有个蛋白质还没有任何实验性的结构信息, 也没有与已知结构的任何蛋白质表现出同源性, 不如将其序列呈送到 http://predict.sanger.ac.uk/irbm-course97/ 看看是否会有所帮助. 想看看对目前已收到的 113 个目标样本的自动分析和对其中 17 个已经作出的预测结果, 可以浏览 http://predict.sanger.ac.uk/irbm-course95/.

(3) 快速数据库检索. 最新发行的 FASTA (ver 3.0) 及其以前材料现在都可以在 http://www.techfak.uni-bielefeld.de/bcd/Lectures/pearson3.html, http://www.biotech.ist.unige.it/bcd/Lectures/pearson3.html 或者 http://merlin.mbcr.bcm.tmc.edu:8001/bcd/Lectures/pearson3.html 找到. 与流行的 NCBI 的 BLAST 算法相比, FASTA 3.0 已经修正了序列长度对数据库相似性得分的影响; 下一步的发展将是在快速数据库检索中加入对长程蛋白质间同源性识别的策略, 以及对各种方法输出结果的解释所需的一些参考提示信息.

(4) 基因组数据库检索. 大肠杆菌(*E. coli*) 全基因组测序于一月份的最后一个星期完成了, 这对于分子生物学家有着特殊的意义, 因为大肠杆菌的绝大部分基因的功能已经被实验研究所确定了, 而其他基因组则还主要依赖于同源性来确定功能. 已经和接近完成的基因组包括: 啤酒酵母(*S. cerevisiae*)、甲烷球菌(*M. jannaschii*)、大肠杆菌(*E. coli*) 和枯草杆菌(*B. subtilis*). 在 http://bmerc-www.bu.edu/genome/genomeblastp.html, http://bmerc-www.bu.edu/genome/ecoli-keyword.html 或 http://www.tiac.net/users/mammon/index.html, 使用 BLAST 接口程序, 可以将您的序列提交 BLASTP 并针对这些基因组的两套公认的 ORF 进行搜索: 针对注释的 ORF (Against Annotated ORF) 或针对未注释的 ORF (Against Unannotated ORF). 输出结果包括原始的 BLAST 输出和对呈现显著 BLAST 匹配的详细参考信息(一般含有一个参考号码, 如果有注释的话还带有蛋白质序列和 ORF 处的 DNA 序列等).

(5) 蛋白质结构预测. 位于 http://www.biokemi.su.se/~server/DAS/ 的服务器使用基于

“密度对齐的表面(Dense alignment surface, DAS)”算法的预测方法定位蛋白质中的跨膜区域. 其特点是无需多序列对齐或是正电荷内置法则(positive inside rule)的任何信息,就可以达到其他最有效的预测方法的效能.

5 人工生命

因为构造一个活细胞的知识目前尚未具备,这里所谓“人工生命”指的是机器人学的机械模型及其所配备的人工智能的计算机大脑. 今日的人工智能机器人学受到了生物学和心理学的许多概念的启发,故而将构造或者行为设计上受生物学启发的机器人称为“动物机器人(Animats)”. “计算神经生态学(Computational neuroethology)”和“合成心理学(Synthetic psychology)”等术语在机器人学中越来越频繁出现的今天,探讨一下生物学和心理学的知识如何应用于机器人学和机器人学实验将会反过来带给生物学和心理学什么样的顿悟,也许会很有意思. 有兴趣就请到 http://www.cogs.susx.ac.uk/ecal97/.

6 生物信息

除了关于DNA和蛋白质的分子生物学数据库和有关生物计算的软件与在线服务以外,还有其他形式的生物信息可供利用,从一般的医疗话题到简单易用的讨论组、为每一位关心生物学进展的人提供论坛的新闻组等.

离子通道毒剂(ion channel toxin)、生物物理软件、在线的离子通道文章等内容已经都加入到了位于 http://qlink.queensu.ca/~4jch3/ 的“离子通道网页(Ion channel webpage)”. 其上安装的分子显示程序提供离子通道毒剂的三维投射图象,且能按指令旋转. 更有特色的是它的离子通道研究者之页、离子通道论坛、离子通道文献列表和序列分析等,是离子通道研究的信息之窗.

位于 http://www.graylab.ac.uk/cancerweb.html 的癌症研究网页 CancerWEB 为患者、临床医师和科学研究人员提供了很多有用的信息和资源. 它组织有序并自带一个快速搜索引擎用以检索和定位目标文档. 它也是 NCI PDQ 数据库在英国的 redistributor,即 CancerNET UK,因此拥有 CancerNET 和 CancerLIT 文件. 全部章节都可以比较容易地从主页、给临床医师的信息(http://www.graylab.ac.uk/cancerweb/clinical.html)、给患者的信息(http://www.graylab.ac.uk/cancerweb/patients.html)等部分找到. CancerWEB 的 SiteNET (http://www.graylab.ac.uk/cancerweb/sitenet.html)是一个综合性的、按照地理位置排布的与癌症有关的学术研究所、医院的列表和匿名 FTP 站点. CancerWEB 图书馆(http://www.graylab.ac.uk/cancerweb/library.html)则指向 CancerLIT 文件和其他可供研究人员参考的信息资源. CancerWEB 教育资源(http://www.graylab.ac.uk/cancerweb/educate.html)链接到其他含有高质量的、与一些如肿瘤学等医学专业教育有关的信息的站点. 全球癌症研究(http://www.graylab.ac.uk/cancerweb/further.html)按照肿瘤类型将与癌症研究相关的 WWW 链接作成了一张大表.

CCP11 计划是英国协作计算计划(Collaborative Computational Project, CCP)中的生物序列和结构分析部分,它的新主页位于 http://www.dl.ac.uk/CCP/CCP11/,对于从事计算分子生物学的科学工作者很有益处.

分子科学虚拟学校(Virtual School of Molecular Sciences, VSMS)已经开始提供 Java 和 XML 这两个在今天的 WWW 上最具动感的新技术的虚拟课程(http://www.vsms.nottingham.ac.uk/

vsms/ java/), 目的是让科学界各学科的信息发布者和编程人员充分发挥 Java 和 XML 的威力, 改进科技信息发表、传播、转化、应用和存贮的方式方法. 所以这个课程不仅是为了学习这两种新技术, 更是要告诉人们不远的将来信息革命的前景. VSMS 是基于 Nottingham 大学、拥有一大批咨询专家、合作者、教师和赞助者的虚拟社群, 不断地将最新出现的技术成果以虚拟教育这样一种优于常规手段的形式推广到受过大学教育的人群中.

欧洲结构生物学(Structural biology in europe, STRUBE) 讨论组和新药发现与蛋白质科学(Drug discovery and protein science, DDPS) 会议年表已经有了在线服务, 在 http: / / www. biodigm. com/ strube. htm.

最近一次的关于开放式外壳计算的量子化学大会(Quantum chemistry symposium on open-shell calculations) 上, 量子化学界领头的专家们共同讨论了在这个正高速发展的领域内的一些最新进展. 包括所有张贴报告、会谈摘要和电子张贴在内的全部活动都已由分子设计电子会议(The molecular modeling e-conference, TMMeC, ISSN 0797-9274) 记录在案. 若欲了解详细情况或访问此次大会的 WWW 网页, 您可以到 http: / / 129. 43. 50. 12/ tmmec/ 或 http: / / 129. 43. 50. 11/ tmmec/ mirrors. html (美国) 、http: / / 164. 73. 160. 8/ tmmec/ mirrors. html (乌拉圭) 、http: / / 130. 206. 125. 40/ tmmec/ mirrors. html (西班牙) 、http: / / 192. 54. 49. 75/ tmmec/ mirrors. html (德国) .

著名的 GenStructure 新闻组的宗旨是为围绕和涉及基因组与染色质结构和功能的话题提供一个讨论的论坛, 让从事于基因组-染色质结构或相关领域的研究者们交流信息和思想, 并拓展国内与国际研究组织进行合作的机会. 当前的讨论话题包括: 1) 基因组-染色质的可操作性和重组; 2) 细胞核的空间立体组织结构; 3) DNA 超螺旋和拓扑结构(三链、Z-DNA、十字、弯折等) 对生物学过程的作用; 4) 组蛋白、核小体和染色质的结构与功能; 5) 环区结构域模型(Loop Domain Model) 、隧道模型、百万碱基巨型环区结构域模型(Megabase giant loop model) 等基因组结构模型; 6) 经典的染色体部件及其与基因功能的关系; 7) 基因组进化; 8) 影响基因组-染色质结构的生物学意义重大的基因突变和基因敲除; 9) 基因组-染色质分析技术; 10) 染色质-DNA 结合蛋白及其对染色质结构和基因表达的作用; 11) 核质(NM) 与核内膜(NL) ; 12) 基质附着区(matrix attachment region, MAR) 、结构域边界和基因座位; 13) 位置效应和拟等位反式(transvection) 等现象; 14) 后成(epigenetic) 效应对基因功能的作用; 15) 剂量补偿机制和 X 染色体失活; 16) 染色质结构与 DNA 复制; 17) 核包装的特别技术; 18) DNA 修复与染色质结构; 19) 基因组不稳定性的机制等. 此外, 该新闻组还提供关于本专业的会议消息、教材、网络资源、可视资料、计算机程序、疑难解答、实践指南等的分论坛.

美洲药物治疗和生产组织(Pharmecutical Research and Manufacturers of America, PhRMA http: / / www. phrma. org/) 与美洲生物科学研究所(American Institute of Biological Sciences, AIBS http: / / www. aibs. org/) 最近一起共同建立了一个提供最新基因组研究信息、名为“基因组学 —— 全球的资源(Genomics —— A Global Resource) ”的互联网站点(http: / / www. phrma. org/ genomics/) . 它随时将新得到的、有用的、关于基因组学研究的信息发布出来, 并维护一些经过选择的、指向世界各地的信息源的链接, 为决策者和普通大众提供一个动态的、易于访问的信息资源: 基因治疗和遗传学取样、检测和筛选, 以及关于生物多样性遗传学和保护濒危物种的数据.

7 生物信息学存在的问题与前景

获得完全的序列和基因组成为可能以后, 如何分析、解释和可视化基因组序列的数据又提出

了新的挑战. 非常必要的一件事是将各自的、独立的、分散的基因组信息整合到一起来. 这些信息可以是计算性的或抽象性的, 譬如关于生物学功能的解释, 像蛋白质的功能, 既不能计算出来也不能被验证. 这使得对这些复杂数据的整合和全面分析变得既耗时又依赖于技巧和知识. 按照交叉学科建设的要求, 计算机科学的技术和概念是解决这些难题所必需的. 分子生物学非常需要并行算法和并行数据库系统的辅助, 以及其他数学的、计算和实验方法方面的新工具.

对基因组部分或全面的序列测定结果迫切需要解释和破译它们的技术. 通用和专用数据库在过去的十年里扩增很快, 要用日益高深的计算机技术来解释这些数据就要求分子生物学、化学、计算机科学、数学和统计学的各个不同学科的密切合作. 这些卓有成效的合作已经取得重大进展的方面包括: 序列搜索和比较、基因组图谱的构建、进化和系统发生; 有望不久取得进展的还有: 序列的统计学分析、多序列比较、遗传图谱、DNA 和蛋白质分析、新的计算和数学工具等.

当前的一些研究新热点包括: (1) 基因表达和遗传网络: 监控、分析和模建 RNA 与蛋白质表达的计算方法; 遗传调控网络模型和采集与分析大规模基因表达数据的新方法. 要了解有关这方面的信息可以到 http://www.cgl.ucsf.edu/psb/sessions/expression.html. (2) 从分子到图象的视化工具和交互工具: 通过视化和用户交互行为帮助科学家权衡、吸收、导引和关联序列、结构和功能数据的新工具和新技术. 这个主要是软件的问题可以到 http://www.cgl.ucsf.edu/psb/sessions/visualization.html 看一看. (3) 大规模基因组序列中的基因结构鉴定: 计算方法寻找新基因的任何一个方面, 其重点是如何全效地发挥出目前已经可以得到的 EST/蛋白质序列等生物信息, 能够在大规模的基因组序列中自动完成基因鉴定和注释的统计学和数学工具. 这方面的详细信息还可以到 http://www.cgl.ucsf.edu/psb/sessions/gene.html 获取. (4) 药物设计和生物技术中的分子设计: 最崭新而时髦的分子设计方法已经在小分子和基于结构的药物设计中崭露头角, 人们正期待着它在蛋白质工程中更伟大的辉煌. 详情可参阅 http://www.cgl.ucsf.edu/psb/sessions/modeling.html. (5) 蛋白质结构预测: 蛋白质结构预测的任何方面, 但着重于可验证的蛋白质结构预测方法, 以及能将实验结果泛化到一个较大的蛋白质类群的方法. 有关问题位于 http://www.cgl.ucsf.edu/psb/sessions/psp.html. (6) 蛋白质结构和功能关系、蛋白质是如何形成功能分化的: 解决“结构-功能”问题的计算策略, 但着重于自动结构分析、进化改变和生物学内涵等这些问题的焦点上. 详情请看 http://www.cgl.ucsf.edu/psb/sessions/function.html. (7) 基于生物分子的计算: 无论是人工还是自然界发生的计算过程, 其中生物大分子都是作为计算部件的. 这方面的研究将导致生物计算机(biocomputer)最终成为现实. 请到 http://www.cgl.ucsf.edu/psb/sessions/compute.html 了解详情. (8) 混沌学(Complexity)和信息论方法应用于生物学: 利用信息论和混沌学的概念与方法来解决生物学上的问题, 包括算法概率(Algorithmic probability)、最小信息长度(Minimum message length)和最小描述长度(Minimum description length)等. 两个容易想象得到的可能应用就是用信息学的方式和手段探讨蛋白质折叠和生物信息处理的问题. 不妨看看 http://www.cgl.ucsf.edu/psb/sessions/info.html. (9) 分布式智能型数据库: 可以形成更加智能的、互相联系的、容易访问的分子生物学数据库的新型计算机和新算法. 这必将有助于对生物语言学的深刻理解. 详细的资料正在 http://www.cgl.ucsf.edu/psb/sessions/database.html 等着您. (10) 在太平洋地区建成一个生物信息大构架(Bioinformation infrastructure): 以太平洋地区为中心的世界各国共同协力合作, 创作出一个共享的生物信息大构架, 这就保证了能为生物计算和生物信息学资源的用户提供高质

量的服务. 此中的重点自然是如何向发展中国家的那些难以利用生物计算和生物信息学服务的研究组织提供周到的技术支持. 该方面的资料请参阅 http://www.cgl.ucsf.edu/psb/sessions/pacific.html.

我国学者也看到了生物信息学所带来的契机,专门为生物信息学在我国的发展而组织了香山会议,为已经、正在和即将在世界生物信息学的前沿阵地冲刺的年轻学者们发展抱负指明了方向,提供了策略. 可以相信,在生物信息学为揭示生命的本质而逐步成熟和完善的过程中,我国生物信息学者一定会取得可喜的成就.

致谢 本工作为国家杰出青年科学基金(编号: 39625014)与国家自然科学基金(批准号: 39730310)重点资助项目.

参 考 文 献

1 Rodriguesz-Tome P. The BioCatalog. Bioinformatics, 1998, 14(5): 469~470
2 Cooper D N, Ball E V, Krawczak M. The human gene mutation database. Nucleic Acids Res, 1998, 26(1): 285~287
3 Sugawara H, Miyazaki S. Towards the Asia-Pacific Bioinformatics Network. Pac Symp Biocomput, 1998(4): 759~764
4 Attwood T K, Beck M E, Flower D R, et al. The PRINTS protein fingerprint database in its fifth year. Nucleic Acids Res, 1998, 26(1): 304~308
5 Bono H, Ogata H, Goto S, et al. Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. Genome Res, 1998, 8(3): 203~210
6 Glemet E, Codani J J. LASSAP. A large scale sequence comparison package. CABIOS, 1997, 13(2): 137~143
7 Wozniak A. Using video oriented instructions to speed-up sequence comparison. CABIOS, 1997, 13(2): 145~150
8 Proutski V, Holmes E. SWAN: sliding windows analysis of nucleotide sequence variability. Bioinformatics, 1998, 14(5): 467~468
9 Trelles O, Andrade M A, Valencia A, et al. Computational space reduction and parallelization of a new clustering approach for large groups of sequences. Bioinformatics, 1998, 14(5): 439~451
10 Krause A, Vingron M. A set-theoretic approach to database searching and clustering. Bioinformatics, 1998, 14(5): 430~438
11 Lijnzaad P, Helgesen C, Rodriguez-Tome P. The radiation hybrid database. Nucleic Acids Res, 1998, 26(1): 102~105
12 Stoesser G, Moseley M A, Sleep J, et al. The EMBL nucleotide sequence database. Nucleic Acids Res, 1998, 26(1): 8~15
13 Van Ommen G J. The Human Genome Project and the role of genetics in health care. Clin Chem Lab Med, 1998, 36(8): 515~517
14 Rebhan M, Chalifa-Caspi V, Prilusky J, et al. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics, 1998, 14(8): 656~664
15 Sorace J M, Canfield K. Collaborative bioinformatics: data warehouses for targeted experimental results. J Interferon Cytokine Res, 1998, 18(9): 799~802
16 Borsani G, Ballabio A, Banfi S. A practical guide to orient yourself in the labyrinth of genome databases. Hum Mol Genet, 1998, 7(10): 1641~1648
17 Goodman N, Rozen S, Stein L D, et al. The LabBase system for data management in large scale biology research laboratories. Bioinformatics, 1998, 14(7): 562~574
18 Baxevanis A D. The Internet and the biologist. Methods Biochem Anal, 1998, 39: 1~15
19 Gaasterland T. Structural genomics: bioinformatics in the driver's seat. Nat Biotechnol, 1998, 16(7): 625~627
20 Wu C H, Shivakumar S, Shivakumar C V, et al. GeneFIND web server for protein family identification and information retrieval. Bioinformatics, 1998, 14(2): 223~224
21 Kastin S, Wexler J. Bioinformatics: searching the Net. Semin Nucl Med, 1998, 28(2): 177~187
22 Stein L D, Cartinhour S, Thierry-Mieg D, et al. JADE: An approach for interconnecting bioinformatics databases. Gene, 1998, 209(1-2): 39~43
23 Baker P G, Brass A. Recent developments in biological sequence databases. Curr Opin Biotechnol, 1998, 9(1): 54~58

(1998-11-12 收稿, 1999-02-29 收修改稿)