文章编号: 1007- 2985(2008)02- 0042- 03

数据库技术在生物信息学中的应用*

谢民主[1], 刘新求[2]
(1. 湖南师范大学物理信息学院, 湖南 长沙 410081; 2. 湖南工程职业技术学院, 湖南 长沙 410151)

摘 要: 阐述了生物信息学研究中重要的数据库及其应用, 并对其发展进行了展望.

关键词: 生物信息学; 数据库技术; 生物数据
中图分类号: TP311 文献标识码: A

"生物信息学"(Bioinformatics)最早是出现在 1956 年美国田纳西州 Gatlinburg 召开的首次"生物学中的信息理论讨论会"上. 随着基因测序技术的不断发展, 通过实验获得的生物分子数据呈指数增长, 急需利用计算科学和信息科学对其进行分析和处理, 生物信息学这门崭新的学科应运而生. 生物信息学是建立在数学、计算机科学和生命科学基础之上的一门交叉学科, 它主要研究如何获取、分析、处理、存储和利用生物信息. 为了高效处理日益增长的海量的生物数据, 使全世界的研究人员能共享已有的研究成果, 数据库技术在生物数据的处理和储存上有越来越重要的应用.

1 核酸序列数据库

用于核酸分析的数据库主要有 GenBank、EMBL、DDBJ 等 3 大数据库. GenBank 的数据来源于约 140 000 个物种, 包含了所有已知的核酸序列和蛋白质序列, 以及与它们相关的文献著作和生物学注释. GenBank 数据主要来源于数据提交, 可通过 BankIt 或 Sequin 向其提交数据. GenBank 数据记录被分为 17 类, 例如灵长类、细菌、病毒等; 最近几年又加入了 EST (Expressed Sequence Tag)、GSS (Genome Survey Sequence)、HTG (High Throughput Genomic)、HTC(High Throughput cDNA) 序列[1].

EMBL 现由 EBI (European Bioinformatics Institute) 管理, 它的数据可通过 WEBIN 和 Sequin 等软件来提交. 新近有全基因组鸟枪 (Whole Genome Shotgun, WGS) 序列和序列版本文档(Sequence Version Archive, SVA)[2] 加入.

DDBJ 近来开发了 SQmatch 工具, 用来搜索基因或蛋白质中短的碱基或氨基酸序列区域, 并建立了简便且易操作的 SOAP (Simple Object Access Protocol) 服务器. 它的数据主要通过 Sakura 和 MST 工具来完成[3]. 这 3 大数据中心各自收集序列数据, 并通过网络每天进行数据交换. 近来 3 大数据库合作的项目主要包括 TPA(Third Party Annotation)、CON(struct) 或 CON(tig) 和 XML 数据交换格式的建立.

* 收稿日期: 2007- 10- 22
基金项目: 湖南省教育厅科学研究项目(06C526)
作者简介: 谢民主(1969-), 男, 湖南涟源人, 湖南师范大学物理信息学院讲师, 博士研究生, 主要从事生物信息学、数据仓库、数据挖掘等研究.

2　蛋白质序列数据库

蛋白质序列数据库主要有 SWISS-PROT, 它经过人工校验, 只收录已知蛋白质序列, 并且每一条数据均有详细注释, 包括功能、结构域、翻译后修饰以及一些相关的综述. TrEMBL 是从 EMBL 库中的核酸序列翻译出来的氨基酸序列, 并可以作为 SWISS-PROT 的一种补充[4]. PIR (Protein Information Resource) 通过合作建立了 PIR-PSD(Protein Sequence Database) 蛋白质序列数据库, 该库收录了大量非冗余的分类与注释详尽的序列信息, 可按注释种类、地域、交互式文本进行搜索[5]. UniProt (Universal Protein Resource) 是个综合数据库, 它集成了 SWISS-PROT、TrEMBL 和 PIR 数据库. 在蛋白质结构域方面最为详尽的是 InterPro, 它集成了 Pfam、PROSITE、PRINTS、SMART、ProDom、TIGRFAMs 等数据库, 收录了大量关于蛋白质家族、结构域和功能位点方面的数据[6].

3　结构数据库

结构数据库主要包括蛋白质结构、核酸结构、小分子数据库等, 这里就重要的蛋白质结构数据库加以介绍. PDB(Protein Data Bank) 是最为详尽的蛋白质结构数据库, 它收录由 X 射线晶体衍射和核磁共振得到的三维结构数据. 可以从 PDB 检索得到原子坐标数据, 然后通过 RasMol、Chime 等浏览器插件进行三维图像显示[7]. 结构分类数据库主要是 SCOP (Structural Classification of Proteins), 它将已知蛋白质的结构按照进化与结构关系进行了全面的分类. SCOP 将蛋白质结构域按家族 (Families)、超家族 (Superfamilies)、折叠家族(Fold Families)、折叠类(Fold Classes) 进行了分类, 并且 SCOP 的分类还在不断完善中[7].

4　基因组数据库

基因组数据库根据物种类别可以分为很多种, 而文中提及的主要是人类基因组. TIGR 数据库包含大量正在测定中的基因组数据, 特别是 EST 序列; TIGR 还拥有大量的 cDNA 数据库. 人类基因组数据库主要是 GDB(Genome Database), 它保存了大量人类基因图谱和疾病数据, 用户可通过基因符号、GenBank 注册号或关键词进行搜索. Ensembl 数据库拥有大量基因序列注释信息, 目前包括脊椎动物、蠕虫和昆虫共 9 个基因组数据. 与疾病相关的数据库有 OMIM (Online Mendelian Inheritance in Man)[8], 它收录了大量人类正常基因与致病相关基因数据.

5　数据库检索与分析系统

为了更好地利用数据库资源, 数据库一般均有自己的检索系统, 其中最为综合的是 Entrez 和 SRS 系统[9]. Entrez 是 NCBI 所提供的集成检索工具, 可通过网络或下到本地使用. Entrez 整合了核酸序列数据库、蛋白质序列数据库、蛋白质结构与结构域数据库、基因组图谱数据库以及通过 PubMed 检索的文献库 MEDLINE. 用户可以通过这些数据库用作者名字、序列索取号、基因或蛋白质名称等多个关键词进行检索. SRS (Sequence Retrieval System) 主要是 EMBL 和 DDBJ 的检索系统, 检索的内容包括 EMBL 最新官方核酸数据库 EMBLRELEASE、更新数据库 EMBLNEW、TPA 入口 EMBLTPA、WGS 入口 EMBLWGS、CON 入口 EMBLCON, 并链接其他重要的分子生物学数据库; 与 Entrez 最大的不同是, SRS 允许用户将自己的数据库整合到该系统中.

6　展望

生物信息学作为一门新兴学科已经成为生命科学研究中必不可少的研究手段. 生物信息学的发展在国内外基本上都处在起步阶段, 也远非成熟. 未来生物学领域高效的研究发现将有赖于生物信息学的发展, 而目前生物信息学存在不少的难题有待解决. 首先, 生物信息学理论研究明显薄弱. 生物信息学对许多学科都提出了巨大的挑战, 包括分子进化遗传学、群体遗传学、统计生物学、基因组学以及计算机科学和应用数学等相关学科. 如果基础理论研究得不到应有的发展, 那么生物信息学的发展将受到严重制约. 其次, 生物学领域中各种不同来源数据的有效整合处理将面临 3 个方面的挑战: 计算基础设施、数据模式和预测

分析模式. 计算基础设施包含了数据存储和数据处理能力2个方面. 数据建模的挑战是如何建立一个可用的、可发展的生物学数据模式. 而预测分析模式的挑战则是如何高效、自动化地获取有用的科学假设. 另外, 如何监控生物数据的质量是摆在生物信息学家面前的另一大难题. 监控已有生物数据的可信度对于生物遗传、物理图谱的构建具有十分重要的意义. 人类科学研究史表明, 科学数据的大量积累将导致重大的科学规律的发现. 例如, 对数百颗天体运行数据的分析导致了开普勒二大定律和万有引力定律的发现; 数十种元素和上万种化合物数据的积累导致了元素周期表的发现; 氢原子光谱学数据的积累促成了量子理论的提出, 为量子力学的建立奠定了基础. 历史的经验值得注意, 有理由认为, 今日生物学数据的巨大积累也将导致重大生物学规律的发现.

参考文献:

[1] BENSON D A, KARSCH-MIZRACHI I, LIPMAN D J, et al. GenBank: Update [J]. Nucl. Acids Res., 2004, 32: 23- 26.

[2] KULIKOVA T, ALDEBERT P, ALTHORPE N, et al. The EMBL Nucleotide Sequence Database [J]. Nucl. Acids Res., 2004, 32: 27 - 30.

[3] MIYAZAKI S, SUGAWARA H, IKEO K, et al. DDBJ in the Stream of Various Biological Data [J]. Nucl. Acids Res., 2004, 32: 31 - 34.

[4] BOECKMANN B, BAIROCH A, APWEILER R, et al. The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003 [J]. Nucl. Acids Res., 2003, 31(1): 365- 370.

[5] WU C H, HUANG H Z, ARMINSKI L, et al. The Protein Information Resource: An Integrated Public Resource of Functional Annotation of Proteins [J]. Nucl. Acids Res., 2002, 30(1): 35- 37.

[6] MULDER N J, APWEILER R, ATTWOOD T K, et al. The InterPro Database, Brings Increased Coverage and New Features [J]. Nucl. Acids Res., 2003, 31(1): 315- 318.

[7] WESTBROOK J, FENG Z K, CHEN L, et al. The Protein Data Bank and Structural Genomics [J]. Nucl. Acids Res., 2003, 31(1): 489- 491.

[8] HAMOSH A, SCOTT A F, AMBERGER J, et al. Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders [J]. Nucl. Acids Res., 2003, 30(1): 52- 55.

[9] 郝柏林, 张淑誉. 生物信息学手册 [M]. 第2版. 上海: 上海科学技术出版社, 2002.

Application of Database Technology in Bioinformatics

XIE Min-zhu[1], LIU Xin-qiu[2]

(1. College of Physics and Information Science, Hunan Normal University, Changsha 410081, China;
2. Hunan Vocational College of Engineering Technology, Changsha 410151, China)

Abstract: This paper addresses some important databases and their applications in the research of Bioinformatics, and discusses the research direction.

Key words: Bioinformatics; database technology; biologic data

(责任编辑 向阳洁)