GSA 和 BIGD — 填补我国生物信息资源服务的空白

北京大学生命科学学院、北京大学国家蛋白质和植物研究重点实验室、北京大学生物信息中心，北京 100871，luojc@pku.edu.cn

Genomics Proteomics and Bioinformatics 期刊（GPB）2017 年第一期发表了一篇数据库专题论文“基因组序列归档库”（Genome Sequence Archive, 简称 GSA）[1]。作者来自中国科学院北京基因组研究所大数据中心（Big Data Center, Beijing Institute of Genomics, 简称 BIGD），文中对他们开发的 GSA 平台作了简要介绍。该平台旨在收集、整合和发布国内外用户递交的原始序列数据。GSA 项目是基因组所大数据中心正在进行的几个主要研究开发项目之一，该中心由近 50 位年轻的生物信息学研究开发人员组成。除 GSA 项目外，还开展了多项面向生物信息资源服务的课题[2]。应 GPB 编辑部邀请，笔者写了一篇短文，简单回顾国际生物信息数据库创建历史，并向读者推荐 GSA 平台和 BIGD 团队的工作。文章以 Preview 形式发表在同一期的 GPB 上，原文为英文[3]；特撰写此中文稿，以飨国内读者。

最近半个多世纪以来，分子生物学取得了长足的进展。DNA 双螺旋的发现、遗传密码的破解、中心法则的提出，为分子生物学研究奠定了坚实的理论基础。与此同时，**费雷德里克·桑格**（Frederick Sanger）等先后建立了蛋白质、tRNA 和 DNA 序列测定方法，**约翰·肯德鲁**（John Kendrew）和**马克斯·佩鲁茨**（Max Perutz）解决了 X-射线晶体衍射解析蛋白质三维空间结构的难题。这些开拓性的研究，为日后分子生物学数据积累提供了必不可少的技术储备。

蛋白质序列数据库

最早从事蛋白质序列收集的是美国国家生物医学研究基金会（National Biomedical Research Foundation，简称 NBRF）的生物信息学先驱**玛格蕾特·戴霍芙**（Margaret Dayhoff）博士（https://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff）。1965 年，她把当时能收集到的 65 个蛋白质信息编纂成册，并以《**蛋白质序列和结构图册**》（Atlas of protein sequence and structure）为名公开发表，并在以后的几年中不断更新再版。这就是国际上第一个蛋白质序列数据库“蛋白质信息资源”（Protein Information Resource，简称 PIR）的雏形。基于收集到的蛋白质家族序列，**戴霍芙**构建了氨基酸替换计分矩阵 PAM，至今仍广泛用于序列比对和数据库相似性搜索。PIR 于 1984 年正式上线，用户可通过电话网络进行查询。两年后，瑞士日内瓦大学在读研究生**埃姆斯·贝洛克**（Amos Bairoch）开始对蛋白质序列进行人工注释（https://en.wikipedia.org/wiki/Amos_Bairoch），为每个序列条目添加功能和相关文献等信息，并在此基础上创建了著名的“瑞士蛋白质序列数据库”（Swiss-Prot）。

蛋白质结构数据库

第一个蛋白质结构数据库（Protein Data Bank，简称 PDB）创建于 1971 年。与蛋白质序列数据库分别诞生于美国和欧洲不同，PDB 的建立是欧美两国合作者共同努力的结果。1971 年，英国剑桥晶体学数据中心（Crystallographic Data Center）和美国布鲁克海文国家实验室（Brookhaven National Laboratory）在《自然：新生物学》（Nature: New Biology）发布短讯，宣告该数据库系统开始运行[4]。双方各自保存相同的数据文件，并免费向用户发布。1998 年，美国结构生物信息学研究协作组（Research Collaboratory for Structural Bioinformatics，简称 RSCB）成立，负责蛋白质结构数据库运行，称 RSCB PDB。

核酸序列数据库

70 年代末，由桑格等建立的 DNA 测序方法日趋成熟，核酸序列开始积累。欧美各国有识之士敏锐地意识到，大规模测序很快就会到来，建立核酸序列数据库的任务已经提上议事日程。1979 年，美国能源部下属洛斯阿拉莫斯国家实验室（Los Alamos National Laboratory）**沃特·高德**（Walter Goad）领导的计算生物学研究组开始利用计算机收集核酸序列，并开发序列分析计算机软件，著名的序列局部比对 Smith-Waterman 算法也因此应运而生。获美国国立健康研究院（National Institute of Health, 简称 NIH）以及科学基金会（National Science Foundation，简称 NSF）、能源部（Department of Energy，简称 DOE)和国防部（Department of Defense，简称 DOD）等部门资助，核酸序列数据库 GenBank 开始运行。就在同一年，位于德国海德堡的欧洲分子生物学实验室（European Molecular Biology Laboratory，简称 EMBL）发布了欧洲版的核酸序列数据库 EMBL-Bank，有时也简称 EMBL。

美国国家生物技术信息中心 NCBI

八十年代中后期，核酸、蛋白质序列和蛋白质结构数据库已经积累了相当可观的数据，而基于中小型和微型计算机的序列和结构分析软件也不断涌现。与此同时，由美国科学基金会资助的为科研教育服务的计算机网络 NSFNet 也开始投入使用。1988 年 11 月，由已故参议员**克劳德·裴帕尔**（Claude Pepper）提议，位于美国首都华盛顿北郊的美国国家生物技术信息中心（National Center for Biotechnology Information，简称 NCBI）成立。NCBI 隶属美国国家医学图书馆（National Library of Medicine，简称 NLM），而 NLM 则是美国国家健康研究院（National Institutes of Health，简称 NIH）的一个下属机构。NCBI 成立初期，仅 8 名人员，经过近 30 年的建设，NCBI 已发展成国际上最大的生物信息中心，著名的数据库搜索软件 BLAST 主要开发者之一**大卫·李普曼**（David Lipman）担任主任至今。NCBI 拥有上百个数据库和软件工具，包括著名的生物医学文献摘要数据库 PubMed、参考序列数据库 RefSeq、数据库相似性搜索软件 BLAST 等。1989 年，核酸序列数据库 GenBank 也由 NCBI 接管。

欧洲生物信息学研究所 EBI

欧洲生物信息学研究所成立于 1994 年，坐落在英国剑桥南部 12 英里维康基金会（Wellcome Trust）基因组园区内。EBI 是欧洲分子生物学实验室 EMBL 的一个下属单位，主要经费来自欧盟，研究人员主要来自西欧各国。经过 20 多年的建设，EBI 已经成为仅次于 NCBI 的国际生物信息中心，为欧洲各国和世界各地用户提供生物信息资源服务，并从事生物信息研究开发。除核酸序列数据库 EMBL 外，EBI 还有许多特色数据库，如基因组数据库 ENSEMBL、蛋白质家族和结构域数据库 InterPro、基因本体数据库 Gene Ontology 等。

三大国际数据库联盟

由美国政府部门资助的国家级生物信息中心 NCBI 和由欧盟资助的生物信息机构 EBI 的成立，为生物信息资源服务提供了人员和经费保障，促成了国际数据库联盟的建立。2003 年，EBI 的蛋白质结构数据库 PDBe，日本蛋白质结构数据库 PDBj 和美国蛋白质结构数据库 RSCB PDB 共同组成国际蛋白质结构数据库联盟 wwPDB（http://www.wwpdb.org/）。2005 年，NCBI、EBI 和 1987 年成立的日本核酸序列数据库 DDBJ 达成协议，建立国际核酸序列数据库联盟（International Nucleotide Sequence Database Collaboration，简称 INSDC，http://www.insdc.org/）。同年，EBI 的 TrEMBL 与 Swiss-Prot 和 PIR 一起，组成了国际上统一的蛋白质序列数据库 UniProt（http://www.uniprot.org/）。TrEMBL 是核酸序列数据库 EMBL 中的编码区翻译所得的蛋白质序列。

互联网诞生和大数据时代到来

20世纪90年代诞生的国际互连网，标志着信息时代的到来。正如诺贝尔奖获得者**沃特·吉尔伯特**（Walter Gilbert）于1991年1月发表在Nature上的卓有远见的文章中指出的那样，“**我们必须把各自的个人电脑接入全球互联网，以便充分利用日新月异的数据库资源，并通过网络进行直接交流**”。 他明确指出，生命科学研究面临着一个模式的改变[5]。十年后的2001年2月，由政府资助的人类基因组计划协作组和美国Celera公司分别发布了人类基因组草图[6,7]，标志着基因组学研究进入了一个新阶段。得益于高通量、低成本的新一代测序技术的快速发展，数以万计的基因组和宏基因组已经测定。根据基因组在线数据库的统计数据[8]，265,734个不同个体的基因组测序已经完成或正在进行（GOLD, https://gold.jgi.doe.gov/）。毋庸置疑，大数据革命将在未来几年中极大影响分子生物学研究，而数据收集和发布是必不可少的重要步骤[9]。

GSA 项目和基因组所大数据中心 BIGD

近三十年来，尽管我国生物信息学研究开发取得了一定成绩[10]，但在生物信息资源建设方面，却几乎还是空白。历史是最好的镜子，上述历史回顾告诉我们，在提供生物信息资源服务方面，我国已远远落后于欧美各国；三大国际数据库联盟中，根本就没有中国的踪影。

为应对即将到来的大数据浪潮，建立国家级的生物信息资源和服务体系势在必行。遗憾的是，过去十多年来，尽管**郝柏林**院士等国内许多有识之士大声疾呼，我国的国家级生物信息中心依然渺无音讯（http://blog.sciencenet.cn/blog-1248-237322.html）。

值得庆幸的是，由中国科学院北京基因组研究所大数据中心 BIGD 开发的“基因组序列归档系统”GSA 项目已经启动。自 2015 年 12 月上线以来，国内 39 个研究机构近 200 个研究课题已经把他们的数据汇交到 GSA 平台。更加令人欣喜的是，该系统也得到了国际上的认可，美国科学院院报 PNAS 等多个期刊已经发表了汇交到 GSA 的学术论文。GSA 系统只是该大数据中心 BIGD 的主要项目之一，数据库构建、基因组变异图谱等其它多个项目也已经开始，其特色数据库涵盖了基因组、转录组、甲基化组等各个方面，而若干重要动植物的基因组变异数据库也已经上线。此外，国际生物信息数据库目录（Database Common）、水稻信息资源维基（Rice Wiki）等也是该中心开发的特色平台。

在国际合作方面，BIGD 也已经迈出了重要的一步。2016 年年底，BIGD 举办生物信息大数据讨论会，NCBI 和欧洲分子生物学网络组织（European Molecular Biology Network，简称 EMBNet, http://www.embnet.org/）等机构的学者应邀参加，与中心成员交流生物信息研究、开发、服务的经验。此外，中心聘请了 NCBI、EBI、DDBJ 等国际著名生物信息中心的资深人士担任科学顾问，并于 2017 年春节前召开了第一届国际科学顾问委员会会议。

当然，BIGD 还刚刚建立，需要得到政府部门的资助和用户群体的支持，才能不断发展壮大，为建立我国国家级的生物信息中心奠定基础。值得深思的是，BIGD 从事的公益性、服务性的工作，在目前国内“以学术论文论英雄、以影响因子排座次”的评价体系下，很难得到足够重视，希望 BIGD 近 50 位年轻的生物信息研究开发人员要有“板凳坐得十年冷”的思想准备。在此，借用英国学者**安澜·布里斯比**（Alan Bleasby）的话，聊以共勉：“*I don't think we can get a Nobel prize by what we are doing so, but the Nobel prize winners know what we are doing for*”。

参考文献

1. Wang YQ, Song FH, Zhu JW, Zhang SS, Yang YD, Chen TT, et al. GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics.* 2017;**15**:14-18.
2. BIG Data Center Members. The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.* 2017;**45**:D18-D24.
3. Luo JC, GSA and BIGD: Filling the gap of bioinformatics resource and service in China. *Genomics Proteomics Bioinformatics.* 2017; **15**:11-13.
4. Berman HM, Kleywegt GJ, Nakamura H, Markley JL. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure*. 2012;**20**:391–396.
5. Gilbert W. Towards a paradigm shift in biology. *Nature*. 1991;**3496**:99.

6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001; **409**:860-921.
7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science*. 2001;**291**:1304-1351.
8. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res*. 2017;**45**:D446-D456.
9. Toronto International Data Release Workshop Authors. Prepublication data sharing. *Nature*. 2009;**461**:168-170.
10. Wei L, Yu J. Bioinformatics in China: a personal perspective. *PLoS Comp Biol*. 2008;**4**:e1000020.