doi:10.3969/j.issn.1672-5565.2015.01.06

生物数据标准化研究进展

操利超,陈凤珍,严志祥*
(深圳华大基因研究院,深圳 518083)

摘　要:随着生物测序技术的快速发展,积累了海量的生物数据。生物数据资源作为生物分析研究及应用的核心和源头,为保证数据的正确性、可用性和安全性,对生物数据资源进行标准化的管理非常重要和迫切。本文综述了目前国内外生物数据标准化研制进展,目前国内外对生物数据缺少一个总体的规划,生物数据语义存在大量的不兼容性,数据格式多种多样,在生物数据收集、处理、存储和共享等方面缺乏统一的标准。国内外生物数据标准化处于起步阶段,但各国生物专家都在努力进行标准研制工作。文章最后从生物数据术语、生物数据资源收集、处理和交换、存储、生物数据库建设和生物数据伦理规范等方面出发,对标准研制工作进行一一探讨,期望能为生物数据标准制定提供一定的参考和依据。

关键词:生物数据 ; 标准化 ; 标准研制
中图分类号:Q-1　　**文献标志码**:A　　**文章编号**:1672-5565(2015)-01-031-04

Research development of biological data Standardization

CAO Lichao, CHEN Fengzhen, YAN Zhixiang*
(*BGI-Shenzhen*, *Shenzhen* 518083, *China*)

Abstract: Vast amounts of biological data have been accumulated with the rapid development of bio-sequencing technology. Meanwhile, biological data resources are essential for biological research and application, the standardization of biological is very important and urgent in terms of ensuring data accuracy, availability and security. This paper reviews the research progress of biological data standards. At present, there are still many unsolved problems , such as incompatibility of relevant biological data semantic, varied data formats and uniform standards in biological data collection, processing, storage and sharing and so on. Although the standardization of biological data is at the beginning stage, relevant experts are trying to draft the standardization sheme. Finally, this paper discusses some topics in the future such as the terms of biological data, the collection, processing, exchange and storage of biological data resources, the construction of biological databases, the ethics of biological data. We hope it will provide a guide for the research of biological data standardization.

Keywords: Biological data; Standardization; Standardization research

随着测序技术的快速发展,特别是 HiSeq X 10、Complete Genomics(CG)等高通量测序仪的应用,基因组测序的费用越来越低[1]。据 GOLD(Genomes Online Database)不完全统计,截至到 2014 年 5 月全球正在进行的基因组测序项目有 24 189 个,已完成的基因组测序项目有 19 093 个,这些项目都会产生海量的基因组学数据。截止到 2014 年 3 月,GenBank/EMBL/DDBJ 核苷酸数据库核苷数量达 202 392 167 431,核苷条数达 171 164 046,SRA(Sequence Read Archive)数据库碱基数量超过 2.5 PB。目前,基因测序技术已从科研服务走向健康医疗、农业和环境能源等产业,从实验室走向个人,影响力将越来越大。因此,生物数据资源已成为 21 世纪重要的战略资源。

然而,目前生物数据来源广泛,测序仪器种类众多,数据类型和格式各异;测序数据量大,大型存储

收稿日期:2014-10-28;**修回日期**:2014-11-17.
作者简介:操利超,男,硕士研究生,研究方向:生物数据分析;E-mail:caolichao@genomics.cn.
*通信作者:严志祥,男,高级工程师,博士,研究方向:生物数据分析;E-mail:yanzhixiang@genomics.cn.

设备和存储结构不完善,很难保证数据的延续性、可用性、完整性和安全性;在数据共享和管理方面,每个科研机构甚至同一机构内部各成体系,并涉及伦理和知识产权等问题,导致数据共享困难。因此,建立共同的生物数据标准十分重要和迫切[2]。

1 生物数据标准现状

1.1 国内数据标准现状

中国具有大量的人口和丰富的物种资源,随着测序技术的发展,生物数据在爆炸性增长,政府也已制定了很多相关政策来支持生物数据平台构建和标准制定工作。例如,在《国家中长期科学和技术发展规划纲要》中指出"充分利用现代信息技术手段,建设基于科技条件资源信息化的数字科技平台,促进科学数据与文献资源的共享";在《标准化事业十二五发展规划》里指出"研制人口管理、人类遗传资源、计划生育、生殖健康等领域的标准"和"整体规划和整合标准化信息资源,统一管理标准化资源数据"等。

在政府的支持和科学家们的共同努力下,国内一些科研单位在生物数据的采集、存储、使用和共享等方面做了很多工作,生物数据中心已经初具规模,建立了大规模测序、生物信息和医学健康等技术平台,并已经开始摸索着从事构建生物数据平台和标准方面的工作,取得了一些进展。例如,在华大基因,至少已经对5.7万个人类基因组进行了测序。专攻生命科学的纽约投资研究公司的 Ross J. Muken 认为,华大基因在全球基因测序服务市场上的份额至少为25%,数据规模达到 PB 级别。2013 年,华大基因与中国标准化研究院等共同制定了《生物信息学术语》国家标准。在生物信息数据库建设方面也取得了重要进展,如深圳国家基因库构建和完善了覆盖人类资源、动物资源、植物资源、微生物资源和海洋资源等各方面资源的数据库。这些生物数据库的建立,积累了大量的相关标准规范化工作的实践经验,2013 年,深圳华大基因研究院制定并通过了《生物基因信息数据库建设与管理规范》地方标准。

另外,国内科学家和相关机构积极参加到国际标准组织,参与数据标准工作。例如,国内蛋白质组研究工作组与欧盟研究机构合作共同制定蛋白质组学数据标准;2014 年,深圳华大基因研究院参与生物技术(ISO/TC 276 Biotechnology)标准工作委员会,并向该技术委员会提交了"The collection, processing, storage and usage specification for the biological information data"、"methods to evaluate the quality of the massive sequencing data"等国际标准草案,积极参与国际标准的制定工作,该单位还参与 FDA 牵头的高通量基因测序 SEQC 国际标准的编制等。

然而,目前中国在生物数据标准化工作中存在诸多问题,突出表现在以下两个方面:

(1)缺乏生物数据标准化的总体规划和一个可以覆盖生物数据采集到共享使用各环节的标准体系;

(2)生物数据语义、数据格式等存在大量的不兼容性,在数据采集、存储和共享使用也缺乏统一协调的标准。

因此,要解决这些问题就必须研究和制定生物数据标准。这需要中国生物科学工作者和各科研单位相互合作,加快标准工作进程。

1.2 国外数据标准现状

近年来,随着生物信息科学领域的快速发展,国际上也成立了很多生物相关标准组织。例如,2013 年,在德国成立了 ISO/TC276 生物技术标准工作委员会;2013 年 7 月,在加拿大多伦多成立了全球基因和健康联盟,该联盟旨在建立统一的管理和操作方法,以促进基因研究和人类健康,加速信息广泛传播,该联盟成立了四个工作组:临床工作组(Clinical Working Group)、数据工作组(Data Working Group)、管理和伦理工作组(Regulatory and Ethics Working Group)和安全工作组(Security Working Group),旨在建立相关标准规范。

为促进生物数据资源标准化,加强生物数据的交流与共享,来自全球 30 多家科研机构的 50 多名研究人员共同建立了 ISA Commons (www.isacommons.org/)标准联盟,该项目发表的评论文章中提到,目前世界上拥有一些较成熟的数据库,但是没有对入库数据进行统一标准化,导致数据交流共享困难,因而该联盟制定了 ISA-Tab file format 数据格式标准[3]。同时,为规范化国际上基因组数据的描述、交换和整合,成立了基因标准联盟(The Genomic Standards Consortium, GSC),该联盟制定了一系列基因序列格式标准,如 minimum information about a genome sequence (MIGS)[4]、Minimum information about a marker gene sequence (MIMARKS)[5];在转录组方面,国际上成立了 RNAi Global,并制定了 Minimum Information About an RNAi Experiment (MIARE)等标准,方便 RNA 数据的共享;在蛋白质组学方面,成立了蛋白质组学标准组织(PSI, Proteomics Standards Initiative),并成立相关工作组,包括分子相互作用(Molecular

Interactions,MI)、质谱(Mass Spectrometry, MS)、蛋白质组学信息学(Proteomics Informatics, PI)、蛋白质的修改(Protein Modifications, MOD)和蛋白质分离(Protein Separation, PS)工作组,制定了一系列蛋白质数据格式标准,如 The Minimum Information About a Proteomics Experiment (MIAPE)[6]、The minimum information required for reporting a molecular interaction experiment (MIMIx)[7]、质谱鉴定的肽段或蛋白质数据交换格式 mzIdentML、mzQuantML 等和质谱数据格式 mzML、mzData 等标准;在代谢组学方面,成立了代谢组学标准组织(Metabolomics Standards Initiative, MSI);在数据质量控制方面,国际上成立了 MAQC(MicroArray Quality Control)项目组,旨在建立相关数据质量标准,提高微芯片和二代测序技术的数据质量。另外,Biosharing(http://www.biosharing.org/)汇总和发现已有的标准信息,广泛涵盖生物、自然科学和生物医学方面的标准,识别重复的标准,促进协调标准的制定,并协调停止重复的标准制定工作。据统计,截止至 2014 年 10 月,biosharing 收集的标准类型里术语文件标准(Terminology artifact)336 篇,交换格式(Exchange format)157 篇,报告指南(Reporting guideline)72 篇。

然而,目前这些标准组织成立时间都较短,标准的内容主要涉及生物数据术语、生物数据交换格式等,数据分析、存储、使用和共享等方面的标准较为欠缺。总体而言,国际上正努力进行生物数据标准工作的研制,但仍然还有很长的一段路要走。

2　生物数据标准前景分析

2.1　生物数据术语规范

生物学由遗传学、数学和信息学等各个学科相互交叉融合,因而导致大量新概念,新术语出现,并存在同义词以及一词多义等模糊性现象。生物数据方面急需要制定相关的术语标准,便于生物数据相关概念的统一、协调和学术交流,有利于生物数据的共享、使用。生物术语的定义应具备准确性、适度性和简明性,避免循环定义。生物学起源于欧美等发达国家,在翻译成汉语的时候,需遵从汉语的造词习惯,表达简单清晰,减少多义和同义现象。

2.2　生物数据资源的收集规范

当前,各个科研单位、企业、学校等产生的生物数据以各种不同的数据格式和存储方式进行收集和管理,为规范收集和管理不同单位产生的数据资源,需确立科研机构数据资源收集规范,如规定所有申请科研资助项目必须提交一个数据管理计划,数据管理计划作为基金或课题申请书的一部分等。

在生物数据收集的过程中,数据质量直接影响到数据本身的价值,因此,生物数据质量控制标准的制定十分重要。例如,2014 年,深圳华大基因研究院参与制定的《高通量测序质量评估方法》国际标准,包括碱基的质量控制、物种间的交叉污染评估和 index-adaptor 污染评估等。然而,在制定数据质量规范过程中,需考虑更多的因素,比如不同的测序平台,质量评估的方法可能各异,各种不同的分析工具,如 FastQC、NGSQC[8]、QC-Chain[9] 等可能造成测序质量评估有所差异。这些因素需要由更多的标准来规范,因此,还需要更多的该领域内的工作者共同努力去完善。

2.3　数据处理和交换规范

要实现生物数据在信息系统之间进行快速便捷的处理和交换,需统一生物数据的信息分类与编码。如通过规范的元数据、生物分类表和主题词表,运用合适的生物信息数据表示方法,如使用 XML 标识语言将生物数据进行合理的组织,便于公众快速检索,交换和使用数据。然而,目前该方面的标准规范主要来源于计算机科学行业,而生物数据具有其自身的特性,我们需要结合这些特性来制定相关标准,提高生物数据处理和交换的效率。

2.4　生物数据存储规范

随着测序行业的发展,全球每天都会产生巨量的生物数据,这将会产生庞大的数据存储需求。当前,很多公司正在打造生物数据存储平台,如 DNAnexus、Flatiron Health、BaseSpace、EasyGenomics 等,但是生物存储仍然存在很多问题,如语义异构、模式异构和生物数据安全问题等,急需制定相关的标准。针对生物数据模式异构问题,可考虑在存储之前,对数据中包含的信息进行抽象,如对生物数据基本信息、物种、类别、功能和测序进行抽象[10],然后采用适当统一的存储模式进行存储。XML 结构化的特点使其成为最佳的生物信息描述语言,而且数据各部分具有独立性,当前很多大型的生物数据库都已经使用了 XML 数据格式[11-12]。为确保数据安全性,需从存储系统、数据管理、存储网络和人员等各方面考虑,制定生物数据安全标准,保证生物数据的安全性。

2.5　生物信息数据库建设规范

随着生物数据增加,各种生物数据库也随之增多,但各生物数据库之间数据格式不统一,数据类型各异,缺乏相关的数据库建设标准,导致在搜集和整理生物数据资源时缺乏依据,在建设生物数据库时无法保证数据的准确性、完整性和安全性。最终导

致数据共享和使用困难。为规范生物数据库建设，需要从生物数据采集、处理、存储、管理、使用和共享等各方面进行全面的调研和研究，制定细致全面的数据建设标准。

2.6 生物数据伦理规范

生物大数据很大一部分是关于个人的或者私有的，很多人都会担心隐私问题。较典型的例子是Lars Steinmetz与他的研究小组公开发表世界上最著名的人体细胞系海拉细胞基因组的研究成果时，引来很多伦理方面的争论，最后不得不将基因组数据从公共数据库中移除。因此，生物数据标准的制定需要充分考虑患者的隐私、知情同意、数据的发布和使用等问题。

3 展 望

基于快速增长的生物数据，依照标准和规范先行原则，对生物数据标准进行分析和研究，是生物科学发展所必需。目前，生物数据标准化研制还处于初步阶段，未来需从总体规划构建生物数据标准体系框架；从生物数据定义、采集、分析、存储、共享和利用等各个环节出发，针对生物数据多种多样、格式不一和存储不规范等方面进行研究，以实用性、共享性和方便性等为原则，研制更详细的数据格式、存储和利用等标准；更重要的是，生物数据标准建设与实施需要生物领域研究工作者的共同参与；同时，国内外专家及组织需共同合作，加强沟通交流，早日形成共识，加速生物数据标准化工作的研制，指导生物数据资源合理使用和共享，助推生物产业发展。

参考文献(References)

[1] DRMANAC R, SPARKS A B, CALLOW M J, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays[J]. Science, 2010, 327(5961): 78-81.

[2] SANSONE S A, ROCCA-SERRA P, FIELD D, et al. Toward interoperable bioscience data [J]. Nature Genetics, 2012, 44:121-126.

[3] NATHAN A, JULI D, STACEY L, et al. Standardizing data[J]. Nature Nanotechnology, 2013, 8:73-74.

[4] FIELD D, GARRITY G, GRAY T, et al. The minimum information about a genome sequence (MIGS) specification[J]. Nature Biotechnology, 2008, 26:541-547.

[5] YILMAZ P, KOTTMANN R, FIELD D, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications[J]. Nature Biotechnology, 2011, 29:415-420.

[6] TAYLOR C F, PATON N W, LILLEY K S, et al. The minimum information about a proteomics experiment (MIAPE)[J]. Nature Biotechnology, 2007, 25:887-893.

[7] ORCHARD S, SALWINSKI L, KERRIEN S, et al. The minimum information required for reporting a molecular interaction experiment (MIMIx) [J]. Nature Biotechnology, 2007, 25:894-898.

[8] PATEL R K, JAIN M. NGS Q C Toolkit: A toolkit for quality control of next generation sequencing data[J]. PLoS One, 2012, 7(2): e30619.

[9] ZHOU Q, SU X, WANG A, et al. QC-Chain: fast and holistic quality control method for next-generation sequencing data[J]. PLoS One, 2013, 8(4):e60234.

[10] 杨进才,赵森,刘小姣, 等.一个基于软件设计模式的生物信息存储模式[J].计算机应用研究, 2010, 27(7):2598-2601.
YANG Jincai, ZHAO Sen, LIU Xiaojiao, et al. Storage pattern of bio-information based on software design patterns[J]. Application Research of Computers, 2010, 27(7):2598-2601.

[11] WANG L, RIETHOVEN J J, ROBINSON A. XEMBL: distributing EMBL data in XML format [J]. Bioinformatics, 2002, 18(8):1147-1148.

[12] MIYAZAKI S, SUGAWARA H, GOJOBORI T, et al. DNA data bank of Japan (DDBJ) in XML[J]. Nucleic Acids Res, 2003, 31(1):13-16.