

国家标准

《生物技术 生命科学数据格式和描述要求》

(征求意见稿)

编制说明

《生物技术 生命科学数据格式和描述要求》

标准起草组

2024 年 11 月

《生物技术 生命科学数据格式和描述要求》

国家标准编制说明

(征求意见稿)

一、工作简况

(一) 任务来源

本项目根据国家标准化管理委员会关于下达 2024 年第一批推荐性国家标准计划及相关标准外文版计划的通知（国标委发【2024】16 号），国家标准《生物技术 生命科学数据格式和描述要求》由全国生化检测标准化技术委员会（SAC/TC 387）提出并归口，本标准等同采用 ISO 20691:2022《Biotechnology — Requirements for data formatting and description in the life sciences》，计划编号为 20240065-T-469。

(二) 起草单位和分工

计划下达后，由中国测试技术研究院组织成立了标准编制工作组。标准编制工作组由中国测试技术研究院、深圳华大生命科学研究院、深圳华大基因科技有限公司组成。

中国测试技术研究院为本文件主办单位，全面负责本文件的管理、评审组织、过程指导等工作；深圳华大生命科学研究院、深圳华大基因科技有限公司负责本文件编制的核心起草单位主导部分内容的编制起草、确认、标准验证工作。

(三) 标准编制过程和主要工作过程

1)、起草阶段

计划正式下达后，起草单位组成了标准起草工作组，根据下达计划明确了本文件主要采用国际 ISO 20691:2022《Biotechnology — Requirements for data formatting and description in the life sciences》的标准内容。标准启动会上，明确了标准牵头单位对照国际标准翻译为国家标准《生物技术 生命科学数据格式和描述要求》的要求。

会后，标准牵头单位、起草单位对照国际标准翻译初步形成了一版较为成熟的翻译文档，并发标准编制组内成员讨论修改。同时，为了对本文件内容进行确认，编制组内按照本文件规定的生命科学数据格式及相应的描述要求对生命科学业务中涉及到的相关格式、描述等开始了验证工作，并积极向编制组反馈了验证结果。2024 年 4 月，标准编制组将标准文本在生命科学数据格式工作组内部会议对标准草案的翻译进行修改完善，并根据国家标准编写规则，对标准整体内容格式进行逐步完善，将其核心术语的翻译进行调整。

2024年6月，标准草案通过了编制工作组标准内审会，内审专家主要对本文件与采标标准的一致性程度提出意见，编制组讨论后经与专家确认，以本文件及其修订的主要内容依旧围绕采用国际标准为主体，同时专家对本文件的部分翻译提出建议，经过与会专家及编制组的反复讨论确认，需将国际标准中晦涩原语进行意译，以便于该标准在国内的理解与实施。标准编制组根据专家意见对草案进行了修改完善，同时对部分翻译进行了修改，最终形成征求意见稿。

二、国家标准编制原则、主要内容、解决的主要问题

（一）编制原则

本文件按照 GB/T 1.1—2020 及 GB/T 1.2—2020 给出的规则起草。本文件编制遵循“科学性、实用性、统一性、规范性”的原则。按照全国生化检测标准化技术委员会（SAC/TC 387）相关章程中标准制修订工作程序的要求开展工作。

本标准是在 ISO 20691:2022《Biotechnology — Requirements for data formatting and description in the life sciences》的基础上，结合我国基因测序行业、生命科学领域生物信息学应用方向数据使用场景，对 ISO 20691:2022 现有的主要内容全部采纳制定而成，制定本标准时遵循以下原则：

- 1) 根据市场应用和行业技术的实际情况出发，最大限度的促进我国生命科学领域、基因测序领域、生物信息领域的测序生产、质量控制、分析方法等环节的研发及应用等方面的提高与发展。本标准对生命科学领域过程中的数据产出、数据存储、数据访问、数据共享提供技术依据，并为行业未来的技术发展留有一定空间，使得本项目标准具有一定前瞻性和开拓性。
- 2) 本标准与现行相关法律法规、标准等协调一致。
- 3) 在确定本标准的数据兼容性、可扩展性、一致性和兼容性、格式验证、数据类型要求、生物数据存储库的要求等内容时，综合考虑行业的需求、以科学研究需要、便利性等方面寻求最大促进行业发展和社会效益角度出发，充分体现了标准在技术上的先进性、科学上的合理性，使标准内容更加全面、完善和易于实施及应用。
- 4) 根据国情，标准制定坚持面向行业、面向市场的原则。结合我国生命科学领域的实际现状，并以引领生命科学研究过程中的数据格式和描述要求的水平提升为目标而制定，提高我国在这一领域标准的综合水平，使标准适应市场需求，满足领域发展，为科学研究、科研生产领域提供标准指导，引导生命科学相关行业领域采用本标准

进行规范化生产、交流，具有一定程度的指导性。

- 5) 对标准的结构编排、编写格式和内容表达方法等按 GB/T 1.1-2020 等系列标准的规定进行编写，使标准规范化。

(二) 国家标准主要内容

本标准全文分为 9 章节和 2 个附录。标准的主要内容如表 1 所示：

表 1 《生物技术 生命科学数据格式和描述要求》主要内容

章节	名称	内容简要
1	范围	明确规定了生命科学领域在数据和相应元数据格式和文档的要求。涵盖了生物技术和生命科学中的基因组学、转录组学、蛋白质组学、代谢组学、合成生物学、系统生物学等相关领域。
2	规范性引用文件	ISO 8601-1, Date and time — Representations for information interchange — Part 1: Basic rules ISO 8601-2, Date and time — Representations for information interchange — Part 2: Extensions
3	术语和定义	对本标准过程中的术语及定义的描述
4	生命科学数据中对实体和概念描述的建议和要求	生命科学数据、数据类型、数据集及相应元数据及中对于实体概念描述的建议和要求。
5	数据格式的技术和组织建议与要求	结构化数据格式。
6	语义建议和数据格式要求	在不同生物数据术语场景下的数据格式要求
7	适用于生物数据注释的术语和本体的要求	生物数据本体是领域内共同认可的概念、实体及其关系的系统，本章节是对上述生物本体的要求。
8	领域特定数据标准的要求	本章节对于一些其他指定技术领域的数据标准进行相关要求。
9	生物数据存储库的要求	本章节对于生物数据存储过程中的存档、索引、搜索、共享等环节进行相关要求。
10	附录 A	本附录对于生命科学领域常见的数据格式进行列举及建议。
11	附录 B	本附录描述了数据、模型、元数据的最低报告标准。

(三) 主要技术差异

等同采用了 ISO 20691:2022 《Biotechnology — Requirements for data formatting and description in the life sciences》，本标准与 ISO 20691:2022 相比，没有技术差异。

(四) 标准解决的主要问题

在生命科学研究及其成果在生物技术中的应用中,诊断学和制药行业比较依赖于从广泛的化验、生物学和功能研究中广泛从复杂数据以及过程描述、实验室中和现场测量等方式获得数据。这之中包括用一些衍生的生物数据进行生物、生物技术和生理过程的计算重建、建模和模拟,以及他们在生物技术工作流程中的应用。数据支持的生命科学和生物技术研究跨越了广泛的生物和生物技术领域和应用(例如人类健康、基因工程生物、环境科学、农业、生物修复、DNA 测序、色谱、显微镜)。生命科学中的数据驱动、数据密集和大数据分析方法只有使用计算方法并通过数据的一致描述、结构化和集成才可能实现。数据的存储、表示、意义、解释、交换和再利用都受到格式设计的影响。通过为生命科学中的数据记录、处理、重用和交换设定基本要求,满足建立可互操作和明确的数据记录、描述和传输框架的关键需求,从而实现最大的数据价值和利用是本标准解决的主要问题。通过 ISO 20691 国际标准《Biotechnology — Requirements for data formatting and description in the life sciences》的采标工作,提供了标准化可互操作生命科学数据格式的要求和建议。它为生物技术和生物领域社区定义的许多不同的子领域特定数据格式和描述标准提供了概念框架和参考。为了通俗易懂地利用所引用的特定领域格式化和描述标准及其协同互动而描述了一个最低要求和规则的独立技术框架。因此,本文件提供了相关子领域总体数据格式和通俗的描述规则与指南,作为跨域数据集成的基础。还提供了创建特殊(子)领域的标准、互操作性及其实现的规则和指南,对于生命科学数据格式必要性做出要求,解决了发挥数据价值方面的基本问题。

三、主要验证情况

本标准在编制过程中,牵头单位同时开展了标准验证工作。验证工作按照本文件规定的数据格式的技术、涉及的相关领域、数据格式的要求、生物数据存储库的要求、生命科学数据常见格式等方面对业务流程中所涉及的相关内容验证。结果显示本文件中对数据格式等方面内容符合国内生命科学、基因组学等产业发展情况。

四、标准中涉及专利的情况

本标准不涉及专利问题。

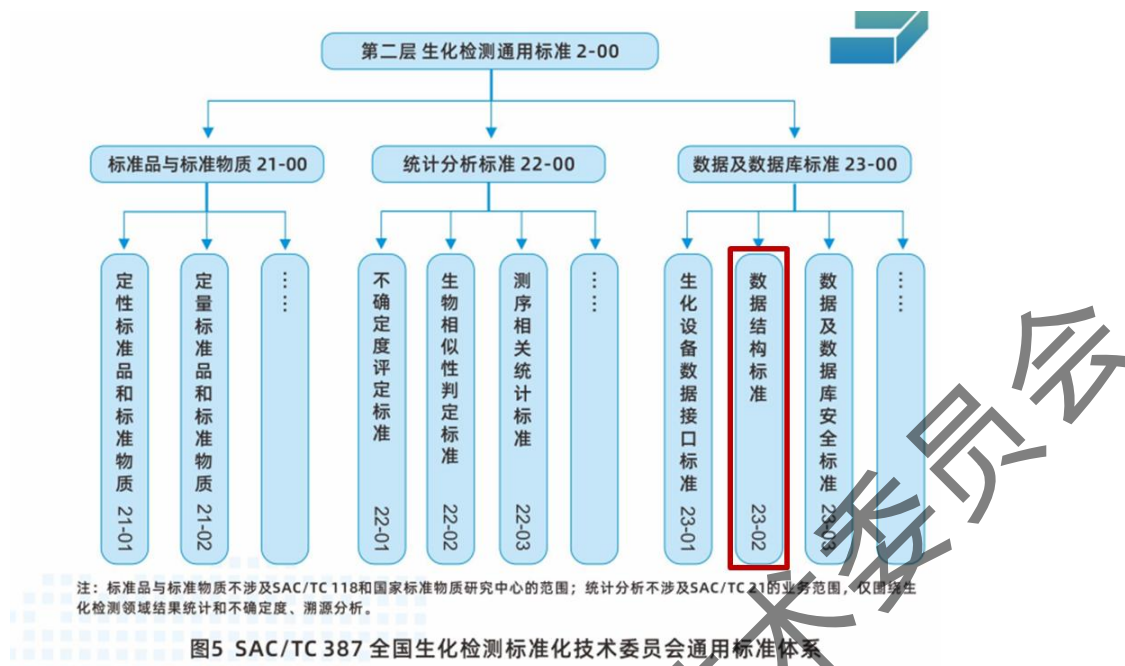
五、预期达到的社会效益、对产业发展的作用等情况

建立生命科学数据格式和描述要求的国家标准，对基因组学、生命科学整个产业的发展和社会效益将产生深远的影响。首先，从产业发展的角度来看，规范数据格式能够促进生命科学数据的整合与共享，这对于推动生命科学领域的研究至关重要。随着高通量测序技术的发展，基因组学数据呈爆炸性增长，海量的数据需要有效的管理和利用。国家标准的制定能够确保数据的一致性和可比性，使得不同来源、不同平台产生的数据能够被有效地整合和分析，这对于跨学科、跨领域的合作研究尤为重要。其次，国家标准的建立能够提高数据的质量和可靠性。通过制定严格的数据采集、处理和存储规范，可以确保数据的准确性和可用性，这对于科学研究和临床应用都是至关重要的。例如，在药物研发和精准医疗中，高质量的数据是实现个性化治疗方案和新药开发的基础。此外，规范的生命科学数据格式还能促进数据的开放共享，加速科学发现和技术创新。开放共享的数据可以被更多的研究者和开发者利用，从而产生更多的创新应用和商业机会。这对于推动生物技术产业的发展，尤其是在诊断学和制药行业，具有重要意义。开放的数据还能促进国际合作，提升我国在全球生命科学领域的影响力和竞争力。从社会效益的角度来看，统一的数据格式有助于提升公众健康水平。通过标准化的数据管理和分析，可以更快地识别疾病风险因素，开发新的治疗方法，提高疾病预防和治疗的效率。这对于提高国民健康水平，减少医疗成本具有重要作用。同时，国家标准的建立还能够促进相关法规和伦理规范的发展。随着生命科学数据的增长，涉及隐私保护、数据安全和伦理使用的问题日益突出。统一的数据格式和标准可以为制定相关法规提供技术基础，确保数据的合法合规使用，保护个人隐私和权益。最后，国家标准的制定还能够推动相关技术和服务市场的发展。随着数据格式的统一，将催生对数据管理和分析工具的需求，促进相关软件和硬件技术的发展。同时，也会出现新的服务模式，如数据托管、分析咨询等，为经济增长提供新的动力。综上，建立生命科学数据格式的国家标准，不仅能够推动科学研究和技术创新，还能够提升公众健康水平，促进经济发展，具有重大的产业发展和社会效益。

六、采用国际标准和国外先进标准的程度，以及与国际、国外同类标准水平的对比情况，或与测试的国外样品、样机的有关数据对比情况

本标准等同采用 ISO 20691:2022 《Biotechnology — Requirements for data formatting and description in the life sciences》，与 ISO 20691:2022 无技术差异。

七、在标准体系中的位置，与现行相关法律、法规及标准，特别是强制性标准的协调性



本标准属于生化检测通用标准中 2-00 的数据及数据库标准类 23-00 的数据结构标准 23-02 小类。

本标准与现行相关法律、法规、规章及相关标准协调一致。

本标准与现行有效的标准没有冲突，配套使用。

八、重大分歧意见的处理经过和依据

无重大分歧意见。

九、标准性质的建议

建议本标准为推荐性国家标准发布。

十、贯彻标准的要求和措施建议

十一、废止现行有关标准的建议

无

十二、其他应予说明的事项

无