

## 微生物大数据资源：资源支撑发展，数据驱动创新

马俊才，蔡磊

中国科学院微生物研究所，北京 100101

### **Microbial resources big data: resource supports development, data drives innovation**

Juncai Ma, Lei Cai

Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

微生物资源在支撑前沿基础研究，及发展先进产业技术中起着至关重要的作用。来源于微生物的大数据资源是微生物资源的重要组成部分。随着微生物在生命健康、绿色农业、可持续发展、生物技术、生态文明等领域越来越多的应用，微生物资源及其组学数据呈现爆炸性增长，围绕微生物大数据平台的建设与数据的挖掘与开发，使得海量数据得以整理、整合并开放共享。建立微生物战略生物资源从保存到利用的实物资源库及信息平台，对于打通资源的收集保藏-大数据分析-功能评价-技术应用全链条，创新生物资源研发体系至关重要。为了更加系统地报道我国微生物大数据领域的优秀资源平台，促进数据资源的积累与应用。国家微生物科学数据中心、中国科学院战略生物资源计划与《微生物学报》合作，策划出版了这一期“微生物大数据资源”专题，集中展示了近期我国在微生物资源大数据平台、数据挖掘与数据服务方向的一些重点成果。

国家微生物科学数据中心 (NMDC) 是由科技部、财政部支持建设，承担我国微生物领域的科学

数据汇交管理和共享服务等职能的数据中心。在这个专栏中范国梅等首先介绍了 NMDC 的数据资源体系、数据汇交和共享应用方式，并专门介绍了 NMDC 的特色数据库和服务，以利于广大的微生物领域专家学者更有效地了解和利用微生物数据资源。

生物元件是具有一定功能的 DNA 序列，也是合成生物学组装的基本单位。合成生物学的发展改变了生物数据的应用方式，也对数据的准确性和数据库的全面性提出了新的要求。刘婉等人介绍了国内外生物元件库、标准和组织，着重介绍了国内的合成生物学元件与数据库，并指出了该领域目前存在的问题和发展方向。酶是一切生命活动的催化剂。孙璠原等人分五大类介绍了与微生物有关的国内外常用的酶学数据库，并讨论了如何利用这些数据库进行相关的研究或开发应用。数据需要有唯一性的通用标识符才能被检索。扩增子测序引物标准化命名数据库 (DPSN) 是一个引物数据库 probeBASE 库的二级数据库，通过建立一个不易引起混淆的全新命名系统对

probeBASE 中的所有引物及探针序列的短命名做出了人工修正, 并让其映射回 probeBASE 中所对应的条目, 解决了源数据库中命名错误和不唯一的问题。DPSN 也对源数据库中的一些其它错误做出了修正, 并支持全文模糊搜索 (谭宇翔等)。

微生物的数据也能提供很多更直接的应用。针对有害生物的检疫是进出口贸易和国家生物安全的重要组成部分, 但传统的检疫方法存在诸多局限, 经常导致检测结果不够准确和快速。植物检疫性菌物数据库针对检疫性菌物物种建立了一个全面的“知识库”, 并在此基础上针对部分物种开发了能快速、有效地鉴定这些有害菌物的一站式解决方案 (赵鹏)。“环境微生物资源信息库”是一个提供可用于治理环境污染的微生物信息的网站, 它提取出了部分可用于环境治理的菌株的信息, 将它们按照用途分类, 并提供了一个可以通过环境或污染物的关键词比对来推荐可用菌株的“群体合成系

统”(王臣)。邹伟等建立了一个白酒微生物的专题数据库, 为使用微生物相关的技术对酿酒工艺进行改进提供了一定帮助。刘柳等对近 20 年来全球专利微生物菌种的保藏和分发进行了分析, 介绍了国内外在专利菌种保藏方面的现状和问题。

微生物作为世界上遗传多样性最丰富、分布范围最广且最易被工业化利用的生物资源, 也形成了极为重要的海量的数据资源, 在不断革新的技术突破加持之下, 未来的数据增长量将更为显著。由本期专题也可看出微生物数据资源整合和挖掘的发展方向, 加强对微生物资源数据的系统整合, 利用多组学数据进行微生物功能元件的挖掘与利用、面向行业与领域应用建立多功能的“一站式”数据服务平台都将产生巨大的科学价值和社会价值, 也将逐步改变科研范式, 支持微生物资源在内的生物资源的高效挖掘和利用。

(本文责编: 张晓丽)



**马俊才**, 博士, 中国科学院微生物研究所研究员、国家微生物科学数据中心主任、科技部人类遗传资源管理专家委员会委员、世界微生物数据中心 (WDCM) 主任、世界微生物菌种保藏联合会理事。主要从事于微生物领域大数据体系建设, 超大规模全文检索技术、远程异构数据库的关联融合技术、超级计算环境和超高速学术网络环境的综合利用等方面的研究。在国内微生物界率先建立了国际互连网络 WWW 服务器, 使中国微生物资源数据库逐步从字符型单机模式向多媒体网络化发展, 实现了中国微生物资源数据库在世界范围内的信息共享。与国家疾控中心国家病原微生物资源库合作, 共同建设新型冠状病毒国家科技资源服务系统, 入选 2020 乌镇世界互联网大会《世界互联网领先科技成果》。作为科技部重点研发项目首席, 主持国家食源性致病微生物全基因组数据库及溯源网络建设。作为 ISO TC276 国际标准中“微生物资源中心数据管理和数据发布标准 (ISO 21710)”工作组主席, 牵头制定 ISO TC276 中微生物数据资源的管理和发布国际标准, 促进微生物资源保藏中心的数据获取、保存、分发、认证和全球数据集成与共享。与国际生物多样性公约秘书处及国际菌种保藏中心合作, 构建《名古屋议定书》的惠益共享框架内的微生物资源信息平台。



**蔡磊**, 博士, 中国科学院微生物研究所研究员。研究方向为真菌多样性与系统进化, 在真菌新生命发现, 重要类群系统演化与多样性形成机制, 有害真菌检疫等领域取得一定成绩。担任基金委“杰青”、“优青”项目负责人; 获中国科协“全国优秀科技工作者”称号、国际菌物学会 (IMA) “杰出青年菌物学家”奖、亚洲菌物学会“杰出青年奖”、中国菌物学会“戴芳澜优秀青年奖”; 担任国际真菌分类委员会委员、亚洲菌物学会副主席、中国菌物学会副秘书长、荷兰国家科学院真菌研究所客座研究员; *Studies in Mycology*, *Mycological Progress*, *Journal of Fungi* 等国际期刊的副主编或编委; 在本领域国际核心期刊发表论文 160 篇, 总引用 11000 多次, h-index(去除自引) = 49, 多篇论文入选 ESI 高被引, 入选 2019、2020 科睿唯安“全球高被引科学家”。