Applications and Implications of International Microbial Big Data Platforms (Postprint)

Authors: Liu Liu, Ma Juncai

Date: 2023-03-19T00:00:00+00:00

Abstract

Microbial resources and microbial big data constitute important national strategic resources, serving as essential material foundations for human survival and development and vital sources for biotechnology innovation. The open sharing of microbial resources and data is of great significance for the development and utilization of microbial resources. The Global Catalogue of Microorganisms (GCM) big data platform, constructed by the World Data Center for Microorganisms (WDCM), contains data information on 400,000 strains of microbial physical resources from 120 international microbial resource centers across 46 countries. This center provides information services on microbial strain resources to the global scientific community and industry through a unified data portal, fully participates in the formulation of international microbial data standards, and offers crucial support for the implementation of the Nagoya Protocol (TheNagoyaProtocol) and compliance efforts in the microbial domain. Building upon this foundation, in 2017, WDCM launched the Global Microbial Type Strain Genome and Microbiome Sequencing Collaboration Plan, thereby achieving a transition from microbial resource data sharing to the sharing and utilization of microbial physical resources. It is hoped that through the microbial big data platform, the development of the biological big data industry can be promoted, the implementation of the China Microbiome Initiative can be further advanced, international microbiome initiatives can be led, and China's discourse power in the microbial field can be enhanced.

Full Text

Applications and Insights from International Microbial Big Data Platforms

LIU Liu, MA Juncai*

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

Abstract

Microbial resources and microbial big data are important national strategic resources, serving as the material foundation for human survival and development and a crucial source of biotechnology innovation. The open sharing of microbial resources and data is of great significance for the development and utilization of microbial resources. The Global Catalogue of Microorganisms (GCM) big data platform, constructed by the World Data Center for Microorganisms (WDCM), contains physical resource information data for 400,000 microbial strains collected by 120 international microbial resource centers across 46 countries. The center provides information services on microbial strain resources to the global scientific and industrial communities through a unified data portal, fully participates in the formulation of international microbial data standards, and provides important support for the implementation of the Nagoya Protocol and its compliance efforts in the field of microbiology. Building upon this foundation, WDCM launched the Global Microbial Genome and Microbiome Sequencing Cooperation Plan in 2017, thereby realizing the transition from microbial resource data to the sharing and utilization of physical microbial resources. It is hoped that this microbial big data platform will promote the development of the biological big data industry, further drive the implementation of the China Microbiome Project, lead international microbiome initiatives, and enhance China's discourse power in the field of microbiology.

Keywords: microbial resources, microbial big data, big data platform, strategic resource

---

Construction of Microbial Resource Big Data Platform

Microbial resources and microbial big data are important national strategic resources, serving as the material foundation for human survival and development and a crucial source of biotechnology innovation. With the advent of the big data era in biology, microbial and genetic resource data are experiencing explosive growth. Microbiology research is gradually shifting from being data-supported to being data-centered. The organization, integration, and open sharing of massive data have become critical for the research and utilization of microbial resources, marking the entry of microbiology into the omics data era.

In China, numerous institutions have established nearly a hundred bioinformatics resource databases with total data volumes reaching the petabyte scale. Supported by the national "863" Program, China's bioinformation technology and platform management technology systems have matured. Beijing and Shanghai have established distributed basic public information sharing platforms for life sciences, which have done substantial and effective work in introducing international public databases, sharing China's basic biological science data, and developing secondary databases, laying a solid foundation for distributed basic public information sharing platforms for life sciences in China.

Under the big data background, future microbiology research will inevitably move toward forming a comprehensive network for microbial resource research, development, and application. The connections between various aspects of microbial research will become closer, while the depth of each aspect will continue to increase, inevitably placing higher demands on data applications. With the development of cloud technology, excellent solutions have been provided for large-scale data storage, computing, and diverse analyses. Therefore, utilizing cloud technology to provide scientists with platforms that include both integrated data and customizable data analysis services will also be an important trend in future microbiological data research.

WDCM's Establishment in China

The World Data Center for Microorganisms (WDCM) was established in 1966 under the World Federation for Culture Collections (WFCC) and the Global Biological Resources Centre Network (GBRCN) under UNESCO, representing the most important physical resource data center in the global microbial field. Following global competition, WDCM officially settled at the Institute of Microbiology, Chinese Academy of Sciences in 2010. This marked the first world data center established in China's life sciences field, signifying a substantial enhancement of China's influence in international microbiology research and bringing tremendous development opportunities for the research and utilization of China's microbial resources. To date, 755 microbial resource preservation centers from 76 countries have registered with WDCM.

WDCM has constructed and maintained a series of important databases related to microbial resources, including Culture Collections Information Worldwide (CCINFO), Global Catalogue of Microorganism (GCM), Reference Strain Catalogue (RSC), Analyzer of Bioresources citation (ABC), and others, making it the most important physical resource data platform in the global microbial field. In terms of big data integration technology research, the WDCM team has developed a biological resource citation platform system that uses advanced data mining methods to extract information on subsequent research and utilization of microbial resources from over 6 million published microbial-related literature, patents, nucleic acid sequences, and genomes worldwide, and has developed the Reference Strain Catalogue. As a cross-platform reference directory, this catalogue integrates ISO and other international standard strain unified numbering systems, promoting the high-standard application of global microbial resources. The WDCM team has also actively explored data integration and service mechanisms, enabling the platform to effectively integrate data resources globally and achieve sustainable development. Simultaneously, as a cooperation platform, WDCM allows Chinese scientists to organize and coordinate relevant forces from various countries from a global perspective, establish global cooperation frameworks, and provides China with opportunities to gradually occupy the international forefront and dominant position in microbial resource development, application, and data sharing. As of the end of July

2018, the platform's cumulative visits have exceeded 200,000.

Promoting the Sharing and Utilization of Microbial Data Resources

To promote the sharing and utilization of global microbial data resources and better integrate microbial-related data from different sources and formats, WDCM proposed the "Global Catalogue of Microorganism International Cooperation Plan" among global preservation centers on September 6, 2016. This initiative aims to provide a globally unified data portal for precious microbial resources currently scattered across various preservation centers and scientists worldwide. The portal system covers important microbial resources from major preservation centers and includes detailed information on microbial resource collection, identification, preservation, and application. This international cooperation plan has established a unified global microbial strain catalogue, standardized the catalogues of major preservation centers, and provided a unified search interface. Simultaneously, the catalogue integrates other knowledge resources such as literature, patents, sequences, and genomes obtained through automated knowledge mining methods, and develops various data retrieval tools as well as data push and customized data services.

This plan, led by the Microbial Resources and Big Data Center of the Institute of Microbiology, Chinese Academy of Sciences, is responsible for specific information platform construction, data standard establishment, and data integration and sharing implementation. Currently, 120 international microbial resource centers from 46 countries including the United States, France, Germany, and Japan have officially joined, with information on 400,000 microbial physical resources collected on the data platform developed by the Chinese team.

Full Participation in International Microbial Data Standard Formulation

For a long time, the use of different data formats for data management and sharing by various microbial resource centers has greatly hindered microbial data exchange and the efficiency of global resource sharing. Based on its work organizing the Global Catalogue of Microorganism (GCM) microbial data resource international cooperation plan and through discussions with the International Organization for Standardization's Biotechnology Committee (ISO/TC 276) and experts from various WDCM countries, the Microbial Resources and Big Data Center of the Institute of Microbiology, Chinese Academy of Sciences, and WDCM have gradually formed the "Standard for Data Management and Data Publication in Microbial Resource Centers (Draft)." After more than a year of preparation, in July 2017, as a joint project of ISO/TC 276's Biological Sample Bank and Biological Resources Working Group (WG2) and Biological Data Processing and Integration Working Group (WG5), the project passed the ISO new work item proposal vote and was officially initiated. The project has now been registered as a working draft and is expected to be officially released as an international standard within two years, which will become the

first ISO international standard in the microbial resource data field. The formulation and implementation of this standard will help ensure the quality of microbial resource data and improve the compatibility and interoperability of global microbial data, providing a foundation for efficient data sharing and big data analysis.

Providing Important Support for Nagoya Protocol Implementation in Microbiology

China has vast territory and is one of the world's 12 megadiverse countries, with extremely rich genetic resources. However, for a long time, China has been the primary target for developed countries to obtain genetic resources and related traditional knowledge. Foreign institutions and individuals have acquired China's rich biological genetic resources through various improper means, resulting in immeasurable loss in quantity and value, and the situation is very severe.

The Convention on Biological Diversity (CBD) aims to protect endangered plants and animals and maximize the protection of diverse biological resources on Earth for the benefit of current and future generations. China signed the convention on June 11, 1992, approved it on November 7, 1992, and deposited its instrument of accession on January 5, 1993. In October 2010, the 10th Conference of the Parties to the UN Convention on Biological Diversity (COP10) adopted the Nagoya Protocol (NP), which entered into force in October 2014. The Nagoya Protocol stipulates the protection of biodiversity through appropriate financial assistance and technical cooperation to achieve sustainable utilization of biological genetic resources, with the purpose of ensuring fair distribution of benefits from biological genetic resources.

The "Global Catalogue of Microorganisms System" (GCM) under the WDCM big data platform is a comprehensive database containing microbial resource retrieval, analysis, and visualization. GCM integrates more online catalogue data, associates strain resources with nucleic acid sequences, proteins, references, citation data, etc., and provides microbial strain resource information services to the global scientific and industrial communities through a unified data portal. GCM can provide effective data support for all aspects of physical microbial resources from collection, preservation, and transnational transfer to academic and commercial application and benefit sharing, providing the most important support for the implementation and enforcement of the Convention on Biological Diversity and the Nagoya Protocol (CBD/NP) in the field of microbiology. The GCM platform and its related guiding principles represent the first complete and operational information platform solution established internationally. WDCM's work on CBD/NP implementation also aligns with China's main direction in participating in CBD work. Currently, experts from the CBD Clearing-House Mechanism, the international microbial field, the legal community, and China's Ministry of Environmental Protection have highly recognized WDCM's related work and affirmed GCM platform's contribution to CBD/NP implementation.

Launching the "Global Cooperative Program on Type Strain Genome Sequencing, Data Mining, and Functional Analysis" (GCM 2.0)

Type strains are the strains preserved in pure culture (reproducible) state that serve as classification concept criteria when naming, classifying, recording, and publishing microorganisms. Due to their reference and uniqueness, type strains are crucial for microbial identification, functional research, and large-scale omics data analysis. Currently, known microbial type strains are widely distributed across global preservation centers, and there remain significant gaps in sequenced microbial genomes. Omics data analysis of type strains for all known species is of major scientific and strategic significance. With reduced sequencing costs and enhanced capabilities for massive data analysis, launching large-scale sequencing plans and conducting sequence analysis and function mining-based research has become the general trend.

In October 2017, led by the Institute of Microbiology, Chinese Academy of Sciences, and jointly with 12 countries worldwide, the "Global Catalogue of Microorganisms 10K Type Strain Sequencing Project" was launched. This plan will complete genome sequencing of over 10,000 type strains of bacteria, fungi, and archaea within five years, covering all currently known bacterial and archaeal type strains and important fungal type strains. It will establish a global cooperative network for microbial type strain genome and microbiome sequencing, covering more than 30 major preservation centers in over 20 countries, selecting currently unsequenced type microbial strains from global microbial resource preservation centers to complete genome sequencing of over 90% of total microbial type strains.

As an international big science program led by China, this initiative will establish a global scientific resource sharing network covering major partners worldwide, especially in developing countries, gathering top scientific talents and advantageous scientific resources in the global microbial field to help solve major fundamental and cutting-edge scientific problems in the field. It will also contribute Chinese wisdom and Chinese solutions to international cooperation on biological resource transnational transfer and benefit-sharing mechanisms in the Convention on Biological Diversity and the Nagoya Protocol, fully demonstrating China's scientific and technological innovation competitiveness and comprehensive international leadership capabilities in the microbial field.

Reflections and Recommendations

From Microbial Resource Data to Physical Resource Sharing and Utilization The "Global Catalogue of Microorganisms Cooperation Plan" (GCM 1.0) currently has 120 international microbial resource centers from 46 countries including the United States, France, Germany, Japan, China, India, Vietnam, and Brazil officially participating, with information on 400,000 microbial physical resources collected on the data platform developed by the Chinese team. Among these 120 international microbial resource centers, developed and devel-

oping countries each account for approximately half. The reason why so many microbial resource preservation centers from both developed and developing countries provide their data to WDCM for free is that the comprehensive microbial big data platform developed by WDCM can provide personalized value-added services needed by preservation centers in both developed and developing countries.

Taking the largest microbial resource preservation center in Europe—the Leibniz Institute DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ)—as an example, after they submitted data on their preserved catalogue of over 30,000 microbial resources to WDCM's data platform, we could clearly understand through our developed data mining system how many scientific papers worldwide scientists and industry have written using DSMZ-sold strains over the past 30 years, how many international patents they have applied for, and how much nucleic acid sequence data they have generated. In other words, we can inform DSMZ of their contributions to the international academic community over the past 30 years. The director of DSMZ used the data we provided when writing about their development process over the past decades. Taking the Vietnam Type Culture Collection (VTCC) as a representative of developing country preservation centers, by joining WDCM's global cooperation plan, they can conveniently use WDCM's big data platform to establish their own external homepage and online strain catalogue database, and can also display their strain information to the world through WDCM's global information platform to enhance their visibility and promote the sharing of global microbial resource data. GCM 1.0 is an international cooperation plan advocated and implemented by China in the international microbial resource field that has received widespread response, exploring an effective mechanism for integrating global microbial resource data, promoting the global sharing and utilization of microbial resource data, and establishing China's leading position in microbial resource data sharing.

Leveraging Existing Advantages to Lead International Type Strain Sequencing Plans: From Global Data Sharing to Physical Resource Cooperation Microbiomics is also a strategic scientific field that countries worldwide are competing to develop, with the United States, Japan, and other developed countries having already deployed national plans supporting microbiome research. On May 13, 2016, the United States announced the launch of the "National Microbiome Initiative," with relevant government departments and private institutions investing up to $500 million to conduct comprehensive and in-depth research on the microbiome and widely apply research results in key areas such as healthcare, food production, and environmental protection. We should seize the opportunity, using the type microbial genome sequencing plan as the starting point, relying on China's advantages in microbial resource research, sequencing technology, and microbial data comprehensive analysis capabilities, to promptly launch the "China Microbiome Project" key research and development special project covering human, agriculture, environment, tra-

ditional fermentation, new technologies, and other content. We should further utilize the international cooperation network established by this plan to launch a China-led international microbiome cooperation plan, seizing the strategic high ground in the microbial field. In October 2017, building upon existing global data cooperation, we launched the "Global Cooperative Program on Type Strain Genome Sequencing, Data Mining, and Functional Analysis" (GCM 2.0). Currently, 24 preservation centers from 14 countries have participated in this plan, providing corresponding type strains or DNA, enabling us to move from early-stage global microbial resource data sharing to physical resource cooperation.

Promoting the Development of Biological Big Data Industry Through Biological Big Data Platforms A report from BCC Research states: "In 2013, the global next-generation sequencing and data analysis market totaled $510 million, and by 2018, this market will grow to $7.6 billion, with a compound annual growth rate of 71.6%." Biological big data contains enormous industrial value and belongs to national strategic resources. China is a major country in biodiversity and biotechnology, with extremely rich biological species, biological resources, and biotechnology data that are closely related to the bioindustry. Future national core competitiveness will largely depend on the speed and ability to transform data into information and knowledge. Research and information discovery based on big data have become a new paradigm for life science research and an engine for scientific and technological innovation, which will change the bioindustry landscape and catalyze new industrial forms. Biological big data platforms are bridges for promoting industry development through science and technology. Through policy planning, scientific project layout, and other methods, we should guide the docking of big data research results with industrial applications, enhance enterprises' enthusiasm for participating in biological big data research and development, and promote the development of China's big data industry.

References

1. Colwell R R. Biodiversity amongst microorganisms and its relevance. Biodiversity and Conservation, 1992, 1(4): 342-345.
2. Senni K, Pereira J, Gueniche F, et al. Marine polysaccharides: a source of bioactive molecules for cell therapy and tissue engineering. Marine Drugs, 2011, 9(9): 1664-1681.
3. Prakash O, Shouche Y, Jangid K, et al. Microbial cultivation and the role of microbial resource centers in the omics era. Applied Microbiology and Biotechnology, 2013, 97(1): 51-62.
4. Wu L H, Sun Q L, Desmeth P, et al. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Research, 2017, 45(D1): D611–D618.
5. Kurtböke I. Microbial Resources: From Functional Existence in Nature to Applications. Cambridge: Academic Press, 2017.

6. Wu L H, McCluskey K, Desmeth P, et al. The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. GigaScience, 2018, 7(5): giy026.
7. Overmann J. Significance and future role of microbial resource centers. Systematic and Applied Microbiology, 2015, 38(4): 258-265.
8. 刘双江, 施文元, 赵国屏. 中国微生物组计划: 机遇与挑战. 中国科学院院刊, 2017, 32(3): 241-250.
9. The White House Office of Science and Technology Policy. FACT SHEET: Announcing the National Microbiome Initiative. Washington: OSTP, 2016.
10. Business Communications Company. Next-generation Sequencing: Emerging Clinical Applications and Global Markets, BIO126C. Wellesley: BCC Research, 2017.

---

*Corresponding author

Note: Figure translations are in progress. See original paper for figures.