Leveraging Biological Data

Introduction

Biological data describe a wide range of biological systems and organisms and are essential to develop bio-based solutions and products. Without a strong biological data ecosystem, the United States risks ceding global leadership in scientific innovation and biomanufacturing to competitors like China. This white paper identifies opportunities to strengthen U.S. biological data generation, collection, and sharing to maintain global leadership in biotechnology innovation.

Improving U.S. Data Infrastructure

The United States lacks a coordinated and sustainable data infrastructure. Without fit-for-purpose, persistent, and well-maintained U.S. data storage, some research programs that generate biological data may not be able to sustain these data, limiting their usefulness. Data collection efforts also suffer without sustained infrastructure, leading to a lower overall quantity of biological data to work with.

The United States has historically funded biological database development and management, largely through the National Institutes of Health (NIH).[1] Some of these databases are housed centrally at NIH's National Center for Biotechnology Information (NCBI), while others are created and housed at research institutions and nonprofit organizations. These databases, however, are relatively small when compared with the larger data assets of other countries like China.[2]

The U.S. biotechnology industry needs sustainable data infrastructure to maximize the value and reusability of biological data. Like infrastructure for water, electricity, and roads across the country, the United States needs a persistent biological data infrastructure to support biotechnology advancements. This includes hardware and software to store the data, mechanisms to move data among collaborators, and people to establish and maintain these systems. Without this infrastructure, researchers may not be able to store the data they generate, continue to analyze existing information in the context of new data, share data with collaborators, or transfer data to entities for use in product development.

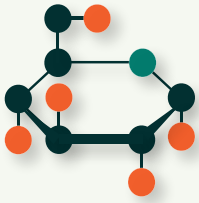
Biological 'Omics Data Defined

While 'omics data are only one subset of biological data, 'omics are relevant to nearly all sectors of biotechnology, including biomanufacturing, medicine, and agriculture.

Genomic data describe all or part of a biological sample's DNA.

Transcriptomic data describe all or part of a biological sample's sequences that are read to make proteins, also called messenger RNAs.

Proteomic data describe the proteins in a biological sample.



Metabolomic data describe the small molecule biochemical intermediates, also called metabolites, in a biological sample, such as carbohydrates and fats.

Epigenomic data describe the reversible chemical modificatons to the DNA, or to the proteins that bind DNA, and contribute to its 3-D structure. Note that epigenomics does not encompass the DNA base pair sequences that constitute the DNA "code." [3]

Standardizing U.S. Biological Databases

Not only are U.S. biological data assets fragmented, but researchers currently generate and store biological data without considering interoperability, making it difficult to easily combine data across databases.[4] Data quality can also vary widely across databases, further limiting their

usefulness. Researchers can maximize the full potential of data by recording, reporting, and storing data in the same way, even in different databases. When data are standardized, practitioners can better search, find, and use data resources. For biotechnology specifically, there currently are no consistent standards for how data and contextual information, called metadata, are reported and stored for access.[5]

Types of biological metadata include:

- Source: biological material, animal/organism, organ, environmental sample details, etc.
- Process: process used to obtain the biological sample, protocol used to prepare the biological sample for analysis, bioreactor conditions, and any other information about how the data was generated
- Project: goal of the project, what was compared, and what was measured
- Technical: file format of the output data, instrument type and model used to create the data, and the software used for creating the data and turning the instrument's signals into biological data
- Terms of Use: describes how to use the data, including rights for use, reanalysis, sharing, and monetization

For biological metadata, there are different standards according to the type of experiment and the type of database.[5] Ideally, researchers use the metadata associated with a project to find relevant data and use it in new analyses. For example, if a dataset of genomic sequence information from cancerous lung tissue is only described as "cancer," rather than "lung carcinoma," it may not be used by other lung cancer researchers.

Updating Data Use Terms and Licenses

Current data use policies do not adequately balance data sharing, re-use of data, and improved collaboration with protecting intellectual property, resulting in overly restrictive and inconsistent data use terms.[6] Owners of different databases and data sources often create highly variable data use terms or licenses that can restrict researchers from using and publishing results generated from the data, forcing them to spend more time managing compliance than transforming data into new ideas.[7] These cumbersome terms and licenses can also restrict users from manipulating and re-sharing the data, even in the absence of privacy and security concerns.[8] Although the terms for federally funded research data may promote analysis, those same terms can restrict users from generating intellectual property from the data, which some researchers believe hampers innovation and commercialization of biotechnology products.[9]

A Path Forward for U.S. Biological Data

Although the biotechnology ecosystem has widely sought to make data findable, accessible, interoperable, and reusable, clear and consistent federal guidelines, policies, and tools would codify those principles and help move the field of biotechnology forward.

By strengthening U.S. biological data infrastructure, developing data standards to support interoperability, and ensuring that data use terms are fit-for-purpose, the United States can better position itself to promote biological data as a strategic national asset and ensure U.S. biotechnology leadership.

At the same time, the misuse of biological data poses risks to individual privacy and national security. The NSCEB will explore how the United States can best protect its biological data resources from both intentional and inadvertent exploitation in our forthcoming report to Congress.

Sources

1 National Counterintelligence and Security Center. "China's Collection of Genomic and Other Healthcare Data from America: Risks to Privacy and U.S. Economic and National Security."
2 National Center for Biotechnology Information. "Our Mission."
3 The National Academy of Medicine. "Omics-Based Clinical Discovery: Science, Technology, and Applications."
4 National Research Council. "Barriers to the Use of Databases."
5 Gonçalves, Musen. "The variable quality of metadata about biological samples used in biomedical experiments."
6 Grabus, Greenberg. "The Landscape of Rights and Licensing Initiatives for Data Sharing."
7 Oza et al. "Ten simple rules for using public biological data for your research."
8 National Research Council. "Protecting Privacy and Confidentiality: Sharing Digital Representations of Biological and Social Data."
9 Carbon et al. "An analysis and metric of reusable data licensing practices for biomedical resources."

For any questions about this white paper, or related work at the National Security Commission on Emerging Biotechnology, please contact us at ideas@biotech.senate.gov.

Staff at the National Security Commission on Emerging Biotechnology authored this paper with input from the expert Commissioners. The content and recommendations of this white paper do not necessarily represent positions officially adopted by the Commission.