Center for Health Security

Discussion on the Future Science and Technology of Biological Attribution

Summary of 6 December 2022 meeting organized by the Office of Science and Technology Policy

January 24, 2023

Johns Hopkins Bloomberg School of Public Health
Center for Health Security

Meeting summary prepared by:

Matthew E. Walsh
PhD Student, Johns Hopkins Bloomberg School of Public Health

Gigi Kwik Gronvall, PhD
Senior Scholar, Johns Hopkins Center for Health Security

Acknowledgements

We would like to express our gratitude to all the participants in the meeting and for the White House Office of Science and Technology Policy for organizing and facilitating the meeting. The authors would like to thank Tom V. Inglesby for his valuable feedback and support; and Alyson Browett, Julia Cizek, Cagla Giray, and Prarthana Vasudevan for their editing, design, and publication support.

The views expressed in this publication and/or made by meeting attendees do not necessarily reflect the official policies of the US government, nor does mention of trade names, commercial practices, or organizations imply endorsement by the US government.

Suggested citation: Walsh ME, Gronvall GK. Discussion on the Future Science and Technology of Biological Attribution. Baltimore, MD: Johns Hopkins Center for Health Security; 2023.

Contents

Introduction

After a biological incident—whether it is natural, deliberate, accidental, or undetermined—there is an imperative to investigate and identify the cause of the incident, and attribute who, if anyone, is responsible. The ability to attribute responsibility for a biological incident (bioattribution) helps to ensure that the deliberate use of biological weapons may be fully prosecuted and those responsible are held accountable. Bioattribution capabilities may also serve as a deterrent for use of biological weapons. Such a capability is the result of an attribution investigation that integrates multiple data sources, including information collected by law enforcement and public health officials, intelligence information, and technical information about the biological agent and other biological and environmental samples collected. The process is complicated; it relies on technical methodology and social systems (ie, the ability to get samples and to have a trusted process) to produce the technical information and sampling for attribution. It is important to routinely evaluate the state of the science available for bioattribution to ensure that investigations may leverage state-of-the-art technology and that efforts are being made to overcome technical challenges.

Summary

On 6 December 2022, the Office of Science and Technology Policy (OSTP) hosted an unclassified, not-for-attribution roundtable discussion on the future of science and technology of biological attribution, including ~15 technical experts and US government (USG) stakeholders. The purpose of the daylong meeting was to provide OSTP and other USG stakeholders an opportunity to obtain information and viewpoints from individual subject matter experts from industry, academia, and national laboratories on the technical aspects—largely, laboratory analysis—of bioattribution. The technical experts came from a diverse range of backgrounds covering genomics, proteomics, bioanalytical chemistry, immunology, bioinformatics, virology, and synthetic biology. Discussions in the morning session focused on the current state of bioattribution technical capabilities with an emphasis on laboratory analysis of biological samples and ideal operating scenarios, and the afternoon discussion focused on pragmatic steps for the bioattribution field in the future. Early on, there was a discussion focused on whether an effort to exhaustively sequence all biological agents of interest to create a reference database was feasible and/or worthwhile. It was recognized that such an effort to exhaustively sequence everything of interest was not practical and that the future of technical bioattribution would need to operate without such a resource.

Significant discussion was dedicated to sample analysis techniques and identifying mid-term (5-10 years) technology development goals. Sample analysis methods generate significant amounts of data and rely on even greater amounts of public data.

Considering how that data is generated, processed, stored, shared, and represented was a common theme throughout the meeting, as it is the underpinning of bioattribution. The Genetic Engineering Attribution Challenge was discussed as an example of how public competitions could be used to make rapid advancements in the field as well as a case study for understanding data needs for building machine learning models for effective bioattribution. Machine learning methods are likely to gain prevalence and popularity in coming years, and it was discussed that the selection of a machine learning model will need to consider the intended use of the output information. Given the accepted lack of an exhaustive reference database, there was discussion on how to maximize the value of multiple pieces of data that each provide some unique insight. Lastly, experts thought that the role of the USG in bioattribution science and technology should be clarified and expanded—it was thought that the government could play a catalytic role in advancing bioattribution technology.

Dedicated research and development efforts are needed to overcome technical challenges in bioattribution, and it was noted that current incentive structures do not support developing a workforce to pursue careers in bioattribution. The technical experts agreed that continued conversation is needed and that the field needs to have more advancement as a community, and the experts expressed enthusiasm in continuing to work together. There was a positive sense in the room in support of future meetings, roundtable discussions, conferences, and community challenges to strengthen bioattribution capabilities.

Meeting Themes

The following themes were present in discussion throughout the day:

Methods: Laboratory analysis of biological samples was categorized into 3 fields of study: genomics, proteomics, and metabolomics. Analysis methods from these fields of study are needed to characterize complex mixtures/samples that may or may not contain living organisms. Capabilities within the field of genomics generally exceed those of proteomics and proteomics capabilities far exceed those of metabolomics. As opposed to PCR-based methods, today's genomic methods focus on sequencing the whole genome. A shortfall of current proteomic methods is the throughput, owing to the time required to run the analysis and the time required to reconfigure and prepare instrumentation between samples. It was noted that multiple independent measures providing the same result would be particularly helpful for attribution, and the ability to identify connectedness among samples from separate events would be valuable in identifying networks of individuals with malintent. Validated methods and core technologies in the public domain would provide an additional element of trust in the results.

Reference samples, databases, and big data: Much of the work surrounding bioattribution relies on matching the analytical output of an unknown sample to a previously collected reference sample or information in an existing database. However, it will not be possible to a priori categorize all of biology to create a database expansive enough to adequately address all future needs. There was discussion about making this problem tractable by investing in understanding smaller, representative subsets of different genera of organisms, for example, to develop a general understanding the genus. Some large databases do exist within industry but are the proprietary information of the companies that own them and should not be considered an available resource to others. It was noted that criminal prosecution relies on publicly available data.

There was general agreement that researchers should endeavor to publish any collected data in a reproducible and transparent manner. In addition to the data itself, there is a desire to include metadata in a standardized fashion. The conversation did not progress to the specificity of exactly what data and metadata would be most valuable in this context. However, some data repositories are growing unsustainably fast and are on pace to become less useful in the coming 2–5 years. Such efforts could be supported by the National Institute of Standards and Technology (NIST) and the National Center for Biotechnology Information (NCBI), and it was suggested that representatives from NIST and NCBI be included in future attribution conversations. There was discussion about cloud-based solutions in academia and industry, but, due to security practices, these solutions may not be feasible for all USG stakeholders. Dual use concerns surrounding what data is collected and aggregated, and how that information could be misused, will also need to be considered.

Genetic Engineering Attribution: One of the more notable activities in the field of bioattribution in recent years is the Genetic Engineering Attribution Challenge that occurred in 2020.[1] This public competition was intended to build upon an earlier academic publication in which the authors demonstrate an ability to predict the lab-of-origin of an engineered DNA plasmid.[2] Prize money was awarded to teams with the highest accuracy in predicting the lab-of-origin. This challenge served as a case study that was referenced during discussion throughout the day. This challenge used data from the nonprofit organization AddGene. The characteristics of the dataset that made it well suited for the challenge were 1) its size, 2) its public availability, 3) its standardized metadata, and 4) the distribution of entries across many academic laboratories. Competitors produced machine learning models that were marked improvements from the earlier publication. There are practical limitations to this work as the concept of operations relies on a bad actor having published their work, deposited their information in a public database, like AddGene, or someone having a priori knowledge of that actor's prior genetic engineering history. Additionally, this work is predicting who designed a sequence and not necessarily who made the sequence.

"Black box" machine learning methods: There are differences between technical and policy experts in their expectations for bioattribution data.[3] Some users of bioattribution data need and expect a rationale for why a machine learning algorithm produced a specific result, something that remains an inherent challenge of using deep learning based methods. One interesting finding from the Genetic Engineering Attribution Challenge was that neural networks perform well on attribution but that traditional machine learning methods also perform well. This suggests that there may not be a meaningful tradeoff in accuracy and explainability, and that technology development should proceed with the needs of the end users in mind. The use of deep learning methods may still provide value in pointing investigators in the right direction but likely would be insufficient as a standalone method of bioattribution. While noted as important, there was limited discussion as to the ideal level of human involvement in the operation of the machine learning algorithms.

Partial solutions: While there was a sense that a perfect solution will remain elusive, there was discussion on how helpful information can be generated from a sample. Such information includes if the pathogen had characteristics of being grown in a laboratory setting, if it underwent directed evolution, if the evolutionary chronometry aligns with what would be expected in nature, if there are abnormalities in the epidemiological data, and sometimes the function of the organism (or molecule). To support these goals, there was a desire to better understand how much variability exists in nature (ie, a baseline) and how much of the knowledge space is unknown. Although none of these processes will individually and conclusively link a biological weapons attack to the responsible party, the collective set of information may be able to.

Role of government: There does not appear to be a single office within the USG that "owns" the challenge of bioattribution. Having a dedicated responsible USG entity would be beneficial to technology research and development. There was a similar roundtable discussion held by the UK government several weeks prior to the USG meeting and intergovernmental collaboration would be beneficial. There are limited incentives for industry and academia, particularly early career scientists, to operate in this space; government can play a role to catalyze careers in bioattribution.

Moving Forward

This roundtable discussion will be the start of continued discussion and engagement. Moving forward, USG, industry, and academia all have roles to play:

Technological development: One clear gap identified was the throughput of proteomics assays. With such shortcomings being known and success metrics easily defined, the USG should invest in a program to develop technologies to more rapidly or cost

effectively generate data required for investigations. Additionally, there was some discussion about exploring federated learning, a method that would allow one entity to use another entity's data to train a machine learning model without exchanging the data, to overcome expressed concerns about disclosing propriety data. Work has been started in this space[4] and additional conversations among the technology developers (bioinformatics and cryptographic experts) and government and industry stakeholders would be required to determine if this is a viable path toward a generalizable and acceptable means for the USG to leverage industry-owned data in support of bioattribution.

Partial solutions: Given the acceptance that an exhaustive reference database will not be available, focus should be on how to maximize the contributions of information that answers questions tangential to identifying a specific individual or entity responsible for a biological event. These methods should be developed with the intent on integrating them into a generalized workflow and efforts should simultaneously be made on maximizing the value of the integration. The USG should consider funding such efforts in industry and academia.

Standardization: Future conversation will need to become more specific with regards to what data is collected, how it is processed, annotated, stored, and shared. This work could be coordinated through NIST or NCBI.

Conferences: The American Society for Microbiology (ASM) has previously hosted ASM Biothreats, an annual scientific conference dedicated to emerging research in the field of biothreats. The 2023 meeting could include a session on bioattribution to inspire broader audience engagement.

Community challenges: The Genetic Engineering Attribution Challenge demonstrated the ability to engage with individuals outside of the biology community and to make technical progress on defined problems in exchange for the possibility of winning a relatively small monetary prize. Future challenges could be developed and conducted to be more realistic of bioattribution activities by including less-than-perfect data sources. Additionally, such a challenge could require participants to curate and publicize data resources for future bioattribution work.

References

1. Crook OM, Warmbrod KL, Lipstein G, et al. Analysis of the first genetic engineering attribution challenge. *Nat Commun*. 2022;13(1):7374. doi:10.1038/s41467-022-35032-8

2. Nielsen AAK, Voigt CA. Deep learning to predict the lab-of-origin of engineered DNA. *Nat Commun*. 2018;9(1):3135. doi:10.1038/s41467-018-05378-z

3. Warmbrod KL, Gronvall GK. Attitudes and Expectations of Investigations and Evidence for Biological Attribution. Published online October 24, 2022. doi:10.20944/preprints202210.0365.v1

4. Titus AJ, Flower A, Hagerty P, et al. SIG-DB: Leveraging homomorphic encryption to securely interrogate privately held genomic databases. *PLOS Comput Biol*. 2018;14(9):e1006454. doi:10.1371/journal.pcbi.1006454