Genomic and Synthetic Biology Digital Biosecurity

Corey M. Hudson and Nicholas D. Pattengale
Sandia National Laboratories
Albuquerque NM 87123, USA
Email: {cmhudso,ndpatte}@sandia.gov

Ravishankar K. Iyer and Zbigniew T. Kalbarczyk
University of Illinois at Urbana-Champaign
Champaign IL 61820, USA
Email: {rkiyer,kalbarcz}@illinois.edu

Nina Alli
Biohacking Village Executive Director & BIO-ISAC
Email: nina@villageb.io

Trends toward automation of synthetic biology and the individualization of biology and medicine raise varied and critical security issues. Digital biosecurity brings together researchers working in secure algorithms, vulnerability assessments, and emerging threat models. The fundamental goal of this digital biosecurity workshop is to identify and present distinct areas of research around making the next generation of biology safer and more secure. The workshop will include a panel overview of the field, including representatives from academia, industry, and non-profits. It will also include novel presentations from the research community. We expect that attendees will leave this workshop with a new appreciation of the research and implementation challenges in maintaining the digital aspects of biosecurity.

Keywords: Digital Biosecurity; Cybersecurity; Biosecurity

1. Introduction, Background, and Motivation

Fundamental decisions related to human health are being handled computationally. While doctors are ultimately charged with most final decisions, precision medicine fundamentally relies on pipelines that from extraction, to sequencing, to variant-calling, to genotype determination, never leave a computational environment. These systems must be secure. As sequencing replacing more straightforward human-driven medical and genetic tests, it is critical that these pipelines are not open to manipulation.[1]

Similarly, fundamental operations in synthetic biology, from sample intake and sample tracking through device orchestration, through testing, validation and verification that are increasingly being automated.[2,3] There are fundamental economic, quality and safety drivers that motivate this growing

field toward automation.[4] Concrete problems that emerge in synthetic biology as humans move out of the loop and major portions of the bioeconomy become reliant on automated systems.[5]

Finally, core bioinformatic operations, from imaging, to sequence assembly and alignment, to protein folding have benefited from higher accuracy and higher throughput, by implementing deep learning.[6,7] These methods benefit from large datasets and can often exceed the performance of state of the art in algorithmic development.[8] However, deep learning approaches are typically and notably resistant to audit and interpretation. It is also difficult to scrutinize imported models prior to their use. This makes them particularly prone to adversarial manipulation and presents in internal security threat to users.[9]

The principal objective of this workshop is to discuss, via invited talks and panel sessions, bio- and cyber-security challenges in genomics and synthetic biology. In particular, the presenters and attendees are expected to engage in so needed and timely conversations to address the cybersecurity and privacy problems posed by the increasing digitization and automation of synthetic biology processes and the use of omics data. This workshop will foster and promote cross-institutional dialog in on best practices to identify emerging risks in this area, determine research gaps, and recognize digital infrastructure issues.

2. Workshop Presenters

This workshop will consist of a short overview of the field, followed by four research presentations, and will end in a hosted panel.

2.1 Workshop Speakers

Michelle Holko (Google)

Jean Peccoud (Colorado State University)

Lisa Simirenko (U.S. Department of Energy Joint Genomics Institute)

Aaron Adler, Miles Rogers, Dan Wyschogrod (Raytheon BBN)

2.2 Panel Moderator

Nina Alli (Biohacking Village Executive Director & BIO-ISAC)

2.3 Panel Attendees

Jeff Moore (Draeger Medical)

Ravishankar Iyer (University of Illinois at Urbana-Champaign)

Alexander Titus (Google Cloud)

3 Speaker Abstracts

Securing the bioeconomy data ecosystem

Michelle Holko

Health and life sciences research domains are amassing data at an exponential rate - genomic projects will generate 40 exabytes of data in the next decade alone and genomic data is set to exceed the growth potential of Twitter, YouTube, and the entire field of astronomy. These data form a critical element of the bioeconomy, and infrastructure to promote and protect bioeconomy data are needed. As data and compute needs increase, many organizations are including cloud in their infrastructure. Cloud computing offers a number of benefits in terms of security, including managed updates/patches, anomaly detection, threat detection, and data protection at rest, in transit, and during compute. There are recent public-private partnerships to establish cloud infrastructure for healthcare and life sciences data. A thoughtful approach to expanding these partnerships, with an emphasis on security specific to bioeconomy data types, is needed.

Digital certificates for engineered DNA

Jean Peccoud

Synthetic biology relies on an ever-growing supply chain of synthetic genetic material. Technologies to secure the exchange of this material are still in their infancy. We are proposing to encode digital certificates in engineered DNA sequences to build a robust link between the DNA molecules circulating in the scientific community, the electronic records describing these molecules, and their developers. This technology is comparable to the VIN system used by the automobile industry. It is essential to support the development of a robust bioeconomy.

BLiSS and Biosecurity Sequence Screening for Synthetic Biology

Lisa Simirenko

In 2010, the U.S. Department of Health and Human Services (HHS) issued the Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA in response to concerns that individuals with malicious intent could exploit DNA synthesis technology to obtain genetic elements from pathogenic organisms. This Guidance outlines the U.S. government's voluntary recommendations to ensure that existing Select Agent Regulations (SAR) and Export Administration Regulations (EAR) are followed, and to encourage best practices in addressing biosecurity concerns.

In accordance with the HHS guidance, the U.S. Department of Energy Joint Genome Institute's (JGI) DNA Synthesis Science program has developed a DNA screening pipeline (BLiSS – Biosecurity List Sequence Screening) to screen all sequences that it synthesizes. BLiSS detects "sequences of concern" of at least 200 nucleotides in length on either DNA strand, including polypeptide translations using the three alternative reading frames on each DNA strand (six-frame translation). Sequences are aligned to GenBank's non-redundant nucleotide and protein databases. To minimize false positives from closely related organisms or highly conserved "house-keeping genes" which do not pose a biosecurity threat, a "Best Match" approach is used to determine whether any sequences, or sequence fragments, are unique to Select Agents or Toxins, or Commerce Control List agents.

We have added post processing to our pipeline to detect potential false positives (i.e. when a "sequence of concern" has a high likelihood of being a gene that is not involved in the pathogenicity of the Select Agent). Additionally, we screen all sequences against the DOE JGI

Integrated Microbial Genomes group's viral database (IMG VR) to detect if viral sequences are being requested, beyond what is required by the Guidance. This information is used to assist the follow-up required when sequences fail the screening process. This follow-up requires human interpretation and has been identified as the costliest aspect of implementing the Guidance by double-stranded DNA providers and a potential barrier to adoption.

Peering into the Cyberbio Threat Horizon

Aaron Adler, Jacob Beal, Partha Pal, Miles Rogers, Dan Wyschogrod

The history of cyber security provides both a cautionary tale and a potential roadmap for anticipating and mitigating digital threats in the expanding bioeconomy. Since the relatively primitive and low-consequence cyber attacks of the 1990s, threats and countermeasures have co-evolved. Synthetic biology and automated laboratory processes are now increasingly intertwined with the cyber world. As the varieties of attacks expanded from the purely digital into the realm of the "cyber-physical," attackers have held the advantage despite enormous investments in mitigations. By focusing exclusively on enhancing the capabilities of automated biology and ignoring the domain-specific threats, we risk repeating the past and handing attackers a long-lasting advantage as vulnerable systems become universal standards. It is therefore necessary to focus on a wide variety of threats and prioritize effective mitigations that respect the pace of innovation.

4 Conclusion

The principal objective of this genomic cybersecurity workshop is to bring together a consortium interested in genomic cybersecurity and the security of omics data in general. In particular, the attendees will address the cybersecurity and privacy problems posed by the increasing digitization and automation of the production, storage and processing of omics data. These include privacy and security concerns and identify area of research for maintaining the safety and security of the global omics-based medical system and bioeconomy.

References

1. D. DiEuliis, C. D. Lutes and J. Giordano, Biodata Risks and Synthetic Biology: A Critical Juncture, *J of Bioterrorism and Biodefense* **9**, 01 (2018).
2. J. Peccoud, J. E. Gallegos, R. Murch, W. G. Buchholz and S. Raman, Cyberbiosecurity: From Naïve Trust to Risk Awareness, *Tr in Biotechnol* **36**, 01 (2018).
3. R. S. Murch, W. K. So, W. G. Buchholz, S. Raman and J. Peccoud, Cyberbiosecurity: An Emerging New Discipline to Help Safeguard the Bioeconomy, *Front Bioeng Biotechnol* (2018).
4. D. DiEuliis, Parsing the Digital Biosecurity Landscape, *Georgtown J Internat Affairs* (2020).
5. J. C. Reed and N. Dunaway, Cyberbiosecurity Implications for the Laboratory of the Future, *Front Bioeng Biotechnol* (2019).
6. J. Jumper, et al., Highly Accurate Protein Structure Prediction with AlphaFold, *Nature* **596** (2021).
7. R. Poplin, et al., A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks, *Nature Biotechnol* **36** (2018).

8. J. Caswell, et al., Defending Our Public Biological Databases as a Global Critical Infrastructure, *Front Bioeng Biotechnol* (2019).
9. M. A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, Adversarial Examples – Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices, *IEEE Internet of Things Journal* **8,** 12 (2021).