Toward a "Poisoner's Handbook" for Biological Databases

Tudor Dumitras
Electrical & Computer Engineering Department
University of Maryland

Mihai Pop
Computer Science Department
University of Maryland

***Abstract*—Modern biological research relies on public databases and, increasingly, on machine-learning based analytics. This introduces a critical vulnerability: the analytic tools may be misled by intentionally or unintentionally corrupted data—for instance, through a data poisoning attack. Such attacks manipulate training data to skew the learned models and can lead to spurious findings or to the suppression of real discoveries. Real-world examples, including taxonomic misclassification and sequence contamination, illustrate how small errors can propagate across databases and remain undetected, with potential consequences for research and clinical decision-making. Drawing on inter-disciplinary expertise in biology, computational biology, machine learning, and cybersecurity, we outline key challenges in detecting and mitigating the threat of data poisoning for biological databases, we present intriguing research opportunities at the intersection of these fields, and we report initial results toward understanding the resilience of bioinformatics systems to data corruption.**

1. Introduction

Modern biological research increasingly depends on advanced analytics, often powered by machine learning and artificial intelligence (AI), and on large public databases (e.g. DNA sequences, measurements of the abundance of molecules, or 3D structure information). This approach drives advances in a broad range of applications, such as food safety, biodefense, or pathogen diagnostics. Many papers describe novel biological insights derived from such analytics. At the same time, the combination of advanced analytics and public databases introduces a critical vulnerability: the bioinformatics algorithms may be misled by intentionally or unintentionally corrupted data—for instance, through a data poisoning attack.

In a data poisoning attack, an intelligent adversary is trying to identify and exploit opportunities to push the outputs of the analytic tool across the decision boundary and force the tool to make incorrect predictions. In a bioinformatics context, these incorrect predictions could lead to spurious findings or to the suppression of real findings from the data. Research in adversarial machine learning has demonstrated powerful poison-crafting algorithms that exploit the complexities of the decision boundary by generating a small amount of training data that would lead the learned model to fabricate or miss a discovery. For example, in a *label flipping* attack, the adversary changes the label of certain items in the training set in order to induce the bad prediction, e.g. assigning a new organism to the wrong family or mischaracterizing its antibiotic-resistance profile. In a *clean-label poisoning* attack [1], [2], the adversary modifies or reorders samples in the training set, retaining their correct labels. The contaminated database would then confound subsequent analyses in a targeted manner.

Compounding this threat, numerous errors make their way in public databases, despite significant investments in the curation of biomedical data. Modern biological technologies generate data at a speed that makes experimental validation and manual curation of the data impractical. In the context of biodefense applications, the current situations is poignantly highlighted in [3]: "*Approximately 60 percent of genomic records associated with microorganisms of relevance for biodefense lack source-of origin information, 18 percent have incorrect assignment to serotype or species and 36 percent lack descriptive information for other types of analytics (e.g. animal or in vitro passage)*". Similar concerns were raised in public health surveillance and food safety [4].

These examples relate to the situation for pathogens, which are typically relatively well studied. The quality of data related to other organisms or biological processes is likely much worse. Furthermore, it is not known whether intentionally-modified sequences could be detected when they are submitted to public databases, raising the potential of intentional poisoning of these critical resources.

These threats are not hypothetical. We provide a few examples that demonstrate the vulnerability of biological databases. *Gemmiger formicilis* is a bacterium first described in 1975 [5]. Its 16S rRNA gene sequence was added to the NCBI database in 2008 [6], and was incorrectly labeled as alpha-proteobacteria around 2012–2013 (*Gemmiger* is in fact closely related to anaerobic human gut bacteria). Other 16S rRNA databases then adopted the NCBI misclassification, leading to scientists commenting at scientific meetings about the surprising abundance of alpha-proteobacteria in human stool samples. Although the error was corrected before it affected published results, the outdated classification was still listed on itis.gov as of April 2026. This example illustrates the potential impact of label flipping attacks. As another example, extensive *Mycoplasma* contamination was identified in human genome sequences [7], leading

to potential errors in sequence-based diagnostic tests for *Mycoplasma*. Many assays initially screen out the host DNA by comparing the input to human reference sequence; if these sequences contain *Mycoplasma*, the screening process may also remove the targeted pathogen. Conversely, the contaminant may be mistaken for the host, confounding analyses of host DNA. This example illustrates the potential impact of clean-label poisoning attacks.

Thus, even small errors in biological databases can influence the output of analytical software. If this software is used for diagnostics or to guide decisions in the clinic, such errors could harm patients. Importantly, as we have shown above, such errors can occur despite quality control measures taken by public databases, and they may not even be apparent to experts (as was the case with *Gemmiger*), let alone to practitioners without deep bioinformatics expertise (e.g., the clinician trusting a *Mycoplasma* diagnostic system).

In this presentation, we explore the research challenges for systematically characterizing the impact of data errors on the accuracy of bioinformatics software. We build upon ideas from the field of adversarial machine learning and identify unique challenges that arise in the biological domain, presenting intriguing research opportunities. We also report on the initial results from a research project at the University of Maryland that aims to develop a framework for reasoning about the resilience of bioinformatics systems to errors, by leveraging inter-disciplinary expertise in biology, computational biology, machine learning, and cybersecurity.

Acknowledgments

This work was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM014698. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] O. Suciu, R. Marginean, Y. Kaya, H. D. III, and T. Dumitras, "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 1299–1316. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/suciu

[2] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 6106–6116, event-place: Montréal, Canada.

[3] J. B. Pettengill, J. Beal, M. Balkey, M. Allard, H. Rand, and R. Timme, "Interpretative Labor and the Bane of Nonstandardized Metadata in Public Health Surveillance and Food Safety," *Clinical Infectious Diseases*, vol. 73, no. 8, pp. 1537–1539, Oct. 2021. [Online]. Available: https://academic.oup.com/cid/article/73/8/1537/6317707

[4] A. B. Hall, M. Yassour, J. Sauk, A. Garner, X. Jiang, T. Arthur, G. K. Lagoudas, T. Vatanen, N. Fornelos, R. Wilson, M. Bertha, M. Cohen, J. Garber, H. Khalili, D. Gevers, A. N. Ananthakrishnan, S. Kugathasan, E. S. Lander, P. Blainey, H. Vlamakis, R. J. Xavier, and C. Huttenhower, "A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients," *Genome Medicine*, vol. 9, no. 1, p. 103, Dec. 2017. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0490-5

[5] J. Gossling and W. E. C. Moore, "Gemmiger formicilis, n.gen., n.sp., an Anaerobic Budding Bacterium from Intestines," *International Journal of Systematic Bacteriology*, vol. 25, no. 2, pp. 202–207, Apr. 1975. [Online]. Available: https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/00207713-25-2-202

[6] P. Yarza, M. Richter, J. Peplies, J. Euzeby, R. Amann, K.-H. Schleifer, W. Ludwig, F. O. Glöckner, and R. Rosselló-Móra, "The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains," *Systematic and Applied Microbiology*, vol. 31, no. 4, pp. 241–250, Sep. 2008. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S072320200800060X

[7] W. B. Langdon, "Mycoplasma contamination in the 1000 Genomes Project," *BioData Mining*, vol. 7, p. 3, 2014.