Toward Autonomous Anomaly Detection within Biological Ecosystems

Craig W. Howser, Kristina K. Zudock, Thomas S. Mehoke, Daniel S. Berman, Brian B. Merritt, and Joseph P. Bernstein

ABSTRACT

By integrating artificial intelligence with next-generation sequencing technology, autonomous surveillance of ecosystem health is possible. This article describes the work a Johns Hopkins University Applied Physics Laboratory (APL) team is doing toward autonomous anomaly detection within biological ecosystems.

The biological world is complex, not well understood, and constantly changing. Native populations change naturally within ecosystems, often because of healthy evolutionary competition or seasonal variation. However, stress placed on ecosystems from external sources, whether by design or an unwitting byproduct, can also cause measurable perturbations in community structure that often go unnoticed until it is too late. A system capable of autonomously characterizing baseline biological communities could detect anomalous changes in the environment at their onset. It is possible that such systems could even track and characterize the source of the interference, thereby enabling indirect surveillance of human activity or the mitigation of environmental tipping points, such as those caused by climate change.

Improvements in high-throughput sequencing technologies, such as increased speed and fidelity and reduced size, have led to the broad pursuit of deploying this technology as a tool for ubiquitous biological surveillance. We are developing a lightweight and generalized framework for rapid biological community characterization and anomaly detection for use as a field forward analysis tool (Figure 1). Using unsupervised deep learning methods such as feature embedding, the goal is to transform an

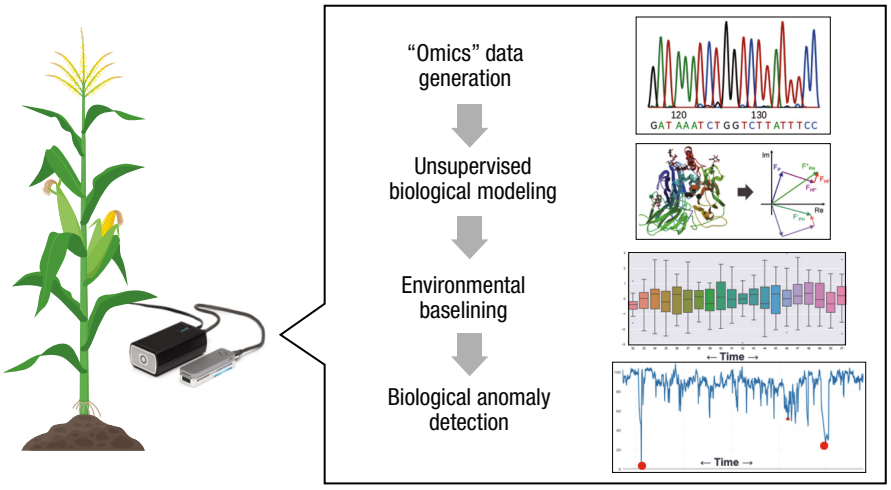


Figure 1. Research concept. The project seeks to develop a framework capable of autonomous anomaly detection within biological ecosystems.

environment's metagenomic makeup into numerical vectors as quickly as sequences are read from the sequencer and deliver them as input to an anomaly detection model. This model uses a deep autoencoder architecture to learn the natural state of a biological community as it varies through time and to detect abnormal deviations from that baseline.

Analysis of biological unknowns traditionally requires large reference databases and high-compute hardware for string comparisons between the millions of nucleotide sequences generated by the sequencer. However, field forward community analysis does not have the luxury of large computational resources. A satisfactory solution must be capable of running on a battery pack and providing actionable information in a timeline short enough for response. By transforming the problem into one that is learnable by a deep autoencoding model, the problem becomes tractable on a graphics processing unit (GPU), thereby enhancing energy efficiency of analysis and eliminating the need for string comparisons or reference databases.

Preliminary results demonstrate that our encoding methods can meaningfully transform biological sequences into numerical vectors, which can then be used to perform taxonomic clustering and sequence reconstruction. To mimic the sequencer as if it were deployed as a real-time environmental sensor and test our anomaly detector, we are generating a series of longitudinal experiments composed of simulated high-throughput sequencing data mimicking one of two vastly different genomic environments: a marine metagenome or a plant's transcriptome.

A simple proof-of-concept experiment was provided to our anomaly detection algorithm for method validation. Sequences from full-length mitochondrial genes of two aquatic organisms (sea spider and alpine bullhead) were generated using DeepSimulator, an externally developed simulator of Oxford MinION data.[1] To create our simplified "ecosystem," we simulated the community transformation from 100% sea spider to 100% alpine bullhead, at varying speeds of growth and decline. Figure 2 visualizes the performance of the classifier in one of these simulations. As expected, it detected the ecosystem's transition period with 100% accuracy (Figure 2b, black).

Future research will ramp up the complexity of the ecosystem simulations to stress-test the performance of the encoding and anomaly detection framework, working with increased diversity in genes within the simulated ecosystem as well as increased levels of variation in population measurements as a means of mimicking sampling bias. Because each component of the framework was designed to be organism and data agnostic, this concept has the potential to provide insight into a variety of research areas of national interest, including agricultural surveillance, effects of climate change on marine ecosystems, and the detection of human activity in austere environments. Looking even further, if

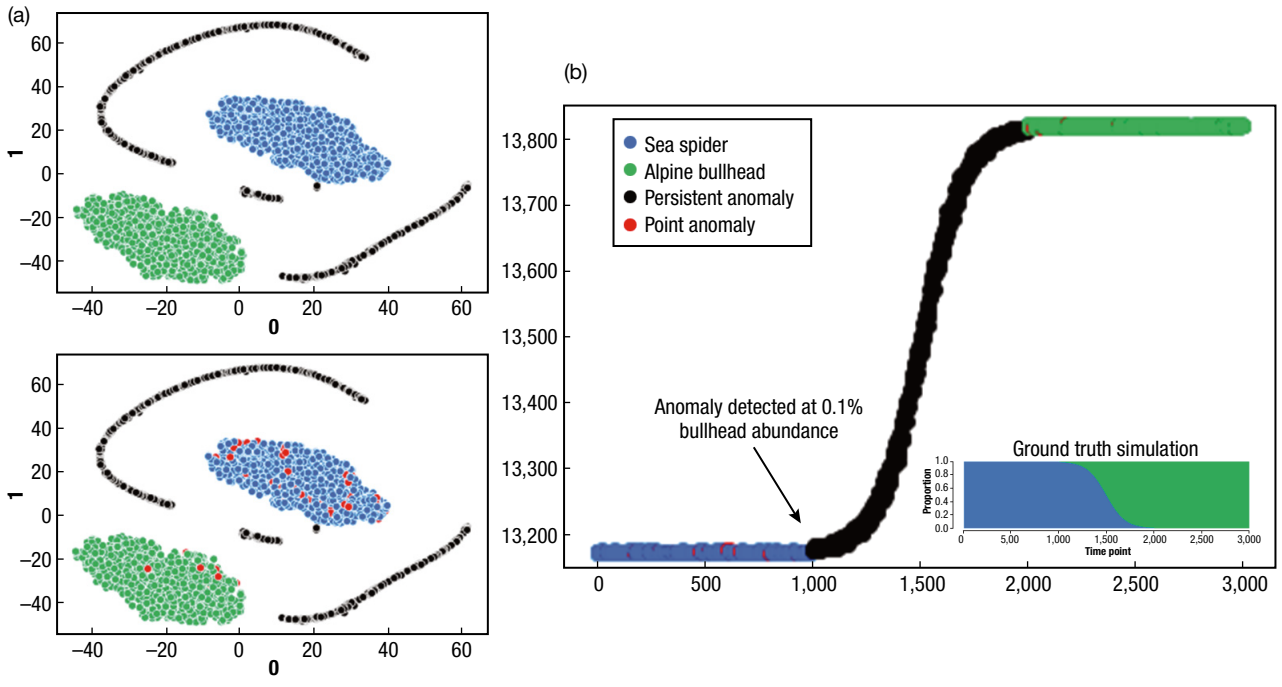
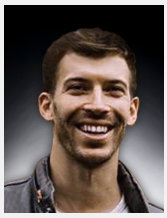


Figure 2. Validation of the anomaly detection pipeline with simple simulation of a marine ecosystem. (a) The embedded representation of each time point's data set was visualized using the t-test stochastic neighbor embedding algorithm (t-SNE).[2] Each point was colored by ground truth labels (top) and predicted labels (bottom). (b) The change in signal across the entirety of the data set given only the first dimension of embedded representation of each time point's data set.

protein sequencing becomes a viable high-throughput method of ecosystem interrogation, our anomaly detection framework can be rapidly adapted to handle these new data types.

REFERENCES

[1]Y. Li, R. Han, C. Bi, M. Li, S. Wang, and X. Gao, "DeepSimulator: A deep simulator for nanopore sequencing," *Bioinformatics*, vol. 34, no. 17, pp. 2899–2908, 2018, https://doi.org/10.1093/bioinformatics/bty223.
[2]L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.



Craig W. Howser, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Craig W. Howser is a bioinformatician and the assistant section supervisor of APL's Bioanalytics and Modeling Section. He has a BS in biotechnology from James Madison University and an MS in computer science from Johns Hopkins University. Craig has a strong background in molecular biology and genomics. Much of his current research focuses on merging artificial intelligence/machine learning techniques with APL's genomic research to inform pathogen evolution and personalized health. He is continuously expanding his expertise in data analytics, statistical modeling, big data management, and next-generation sequencing (NGS) laboratory techniques. Craig served as the project manager on this effort. His email address is craig.howser@jhuapl.edu.

Kristina K. Zudock, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Kristina K. Zudock is a bioinformatician in APL's Research and Exploratory Development Department. She has a BS in biology from Washington University and is pursuing an MS in computer science at Johns Hopkins University. Kristina has a strong background in genomics and computational biology and experience in marine and plant science. Much of her current work centers on developing pipelines to automate analysis of large environmental, infectious disease and microbiome data sets to produce intuitive visualizations providing actionable information. Specifically, Kristina provides data management, processing, and analysis for projects involving biosecurity, infectious disease surveillance, microbiome characterization, pathogen detection and characterization, and environmental sense and detect. On this project, Kristina was the lead technical contributor on the data management and genomics simulation tasks. Her email address is kristina.zudock@jhuapl.edu.

Thomas S. Mehoke, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Thomas S. Mehoke is a project manager and supervisor of APL's Bioanalytics and Modeling Section. He has a BS in biology from Duke University and an MS in computer and information systems from Johns Hopkins University. His research background is in virology, metagenomics, and microfluidics, with an emphasis on the creation of easy-to-use exploratory data visualization tools. Thomas is currently a co-principal investigator on an APL internal project focused on disease transmission, is the lead bioinformatician in several other efforts, and is working on adapting machine learning tools and algorithms to bioinformatics problems. He has experience with a variety of open-source tools, with an emphasis on developing custom UNIX scripts incorporating complex shell scripting techniques for rapid data analysis as well as using Python- and JavaScript-based web interfaces for user interaction. On this effort, Thomas served as the chief bioinformatician, where, in addition to his contributions in data simulation and modeling, he advised in study design and analytical method development. His email address is thomas.mehoke@jhuapl.edu.

Daniel S. Berman, Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Daniel S. Berman is a data scientist in APL's Asymmetric Operations Sector. He has a BA in psychology, a BS in physics, and an MS in applied physics, all from Johns Hopkins University. Daniel has a strong background in solving physics- and psychology-based problems, with research experience including some original research. He has worked on projects involving systems engineering; test and evaluation; feasibility assessment; statistical and mathematical analysis, especially with Bayesian statistics; metric development; programming using R, MATLAB, and Python; cyber systems; and market research. Daniel served as the lead deep learning engineer on this effort. His email address is daniel.berman@jhuapl.edu.

Brian B. Merritt, Research and Exploratory Development Department, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Brian B. Merritt is an associate professional staff member in APL's Research and Exploratory Development Department. He has a BS in biochemistry and molecular biology from the University of Georgia and an MS in bioinformatics from the Georgia Institute of Technology. Brian served as a contributor in data management on this effort. His email address is brian.merritt@jhuapl.edu.

Joseph P. Bernstein, Air and Missile Defense Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Joseph P. Bernstein is an associate professional staff member in APL's Research and Exploratory Development Department. He has a BS in computer science and mathematics from the University of Maryland, Baltimore County. Joseph was the lead software developer for the anomaly detection task on this project. His email address is joseph.bernstein@jhuapl.edu.