AI-POWERED DIAGNOSTICS

Predicting Sequences Of Concern With Machine Learning

Krista Ternus Signature Science, LLC **Gene Godbold** Signature Science, LLC **Advait Balaji** Rice University **Michael Nute** Rice University **Yunxi Liu** Rice University **Anthony Kappell** Signature Science, LLC **Danielle LeSassier** Signature Science, LLC **Matthew Scholz** Signature Science, LLC **Joseph Orton** Signature Science, LLC **Curt Hewitt** Signature Science, LLC **Todd Treangen** Rice University

Background Information and Objective: The U.S. Government is moving away from a strict taxonomic definition of a biothreat to one that seeks to understand sequences of concern (SOCs), which contribute to pathogenicity or harm if introduced into new genetic frameworks. The availability of published experimental evidence to describe the function of a given sequence and the time-intensive nature of manual literature searching for these functions are limiting factors in cataloging SOCs through traditional annotation processes. The objective of this study was to scale the identification and annotation of SOCs with machine learning algorithms.

Methods: We manually reviewed thousands of papers in microbial pathogenesis, annotating more than 3000 virulence factors from more than 140 bacterial species, 85 viruses, and 25 eukaryotic pathogens based on functional experimental evidence. This gold standard, curated dataset was used for training and testing machine learning approaches that were integrated into our open-source SeqScreen software for classifying functions of sequences of concern.

Preliminary Results and Conclusions: Out of more than ten different machine learning models, the top three performing machine learning models were selected to inform our predictions. These were (i) a binary neural network model with support vector classifiers for feature selection, (ii) two-stage multi-class multi-label neural network, and (iii) a two-stage binary support vector classifier. The two-stage networks consisted of architectures that were trained for detection and classification tasks sequentially. The binary predictions of each of the classifiers over each function of sequence of concern were combined in a majority voting scheme to predict the final labels. Ultimately within the SeqScreen software, each query sequence was assigned a binary label indicating the presence or absence of each of the 32 FunSoCs. A primary focus during the development of the machine learning models was to make the feature selection and classification strategies as explainable as possible instead of applying it as “black box” techniques. The interpretability of the models was also imperative for iterative curation where these features and labels could be passed on to the manual biocurators to potentially curate and refine more examples of proteins belonging to the respective features. More details about our methods and results are available at https://doi.org/10.1186/s13059-022-02695-x and https://doi.org/10.1145/3584371.3612960.

Impact to the DTRA JSTO Mission: This work represents a first step toward building software to meet DTRA JSTO’s requirements for a field-operable tool that will enable general purpose forces to identify “known and validated" biothreat or biohazard quickly and accurately. A key innovation is our machine learning-enabled threat identification capability, which goes beyond state-of-the-art taxonomic identification tools by including a functional threat assessment of sequencing data in real-time. Transitioning this rapid threat identification capability to general purpose forces aligns with DTRA JSTO’s goals for addressing emerging biological threats.

All of the co-authors were either fully or partially supported by the Fun GCAT program from the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

Advait Balaji, Bryce Kille, Anthony D. Kappell, Gene D. Godbold, Madeline Diep, RA Leo Elworth, Zhiqin Qian, Dreycey Albin, Daniel J. Nasko, Nidhi Shah, Mihai Pop, Santiago Segarra, Krista L. Ternus, Todd J. Treangen. 2022. "SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning." Genome Biology 23, no. 1 (2022): 133. https://doi.org/10.1186/s13059-022-02695-x

Advait Balaji, Yunxi Liu, Michael G. Nute, Bingbing Hu, Anthony D. Kappell, Danielle S. Lesassier, Gene D. Godbold, Krista Ternus, and Todd Treangen. 2023. "SeqScreen-Nano: a computational platform for streaming, in-field characterization of microbial pathogens." In Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '23). Association for Computing Machinery, New York, NY, USA, Article 17, 1–10. https://doi.org/10.1145/3584371.3612960