Predicting Sequences of Concern with Machine Learning

Vernon McIntosh[1] (Presenter) • Krista Ternus[1] • Gene Godbold[1] • Advait Balaji[2] • Michael Nute[2] • Yunxi Lu[2] • Anthony Kappell[1] • Danielle LeSassier[1] • Matthew Scholz[1] • Joseph Orton[1] • Curt Hewitt[1] • Todd J Treangen[2]

[1] Signature Science, LLC • [2] Rice University

Background 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.

Leveraging advanced machine learning techniques for rapid threat identification, this work is part of a larger integrated strategy that not only improves detection capabilities but also provides opportunities for collaborative, real-world responses to emerging challenges. By adopting a proactive, data-driven framework, our approach paves the way for more robust chemical and biological risk mitigation.

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. We then tested eleven machine learning models based on three strategies that used different feature selection criteria, as well as a two-step pipeline to filter proteins not associated with any functions of sequences of concern (FunSoCs). Our software was further modified for use with MinION sequencing data, where batches of sequences were analyzed in real time as they were released from the sequencer and each open reading frame in a long read was evaluated for FunSoCs.

SEQSCREEN™

Acknowledgements

All of the coauthors 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.

Results

Out of more than ten different machine learning models, the top three performing machine learning models were selected to inform our predictions. These were:

- a binary neural network model with support vector classifiers for feature selection,
- two-stage multi-class multi-label neural network, and
- 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 is assigned a binary label indicating the presence or absence of each of the 32 FunSoCs (**Figure 1**). 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 (**Figure 2**).

Scan for more details about our methods and results

Top Performing Models: Positive Label Precision and Recall per FunSoc

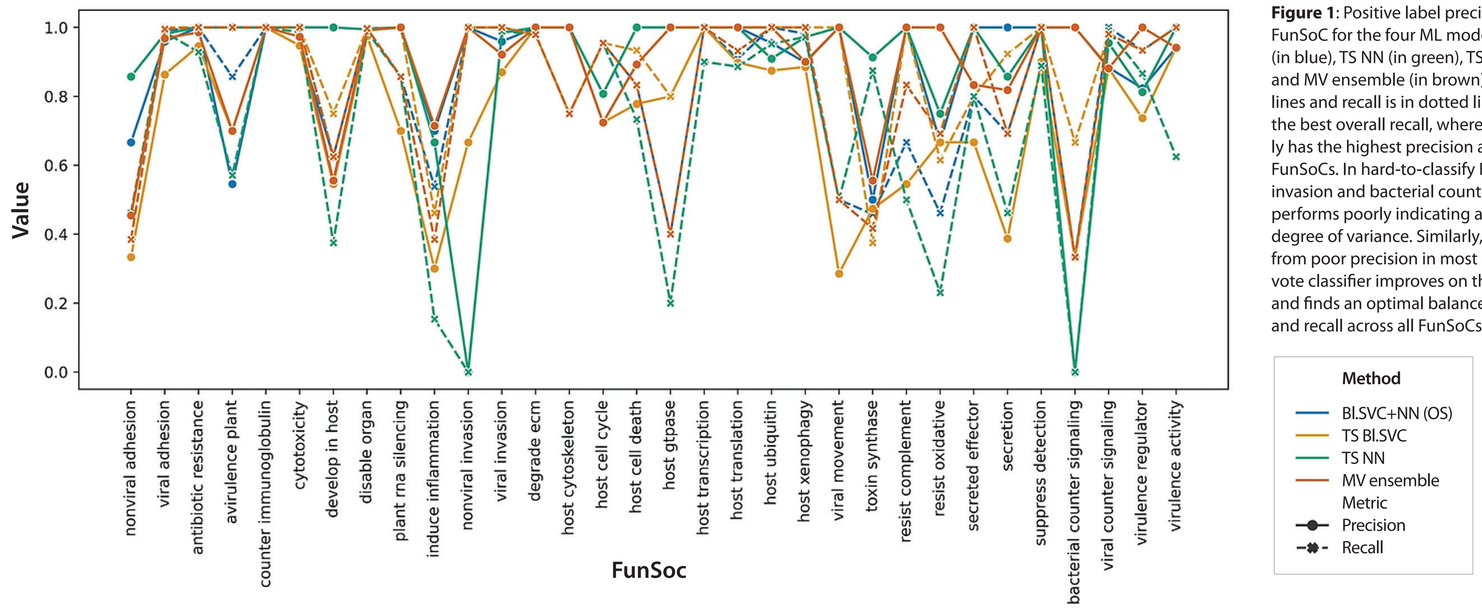

Figure 1: Positive label precision and recall per FunSoC for the four ML models BI. SVC + NN (OS) (in blue), TS NN (in green), TS BI. SVC (in yellow), and MV ensemble (in brown). Precision is in solid lines and recall is in dotted lines. TS BI. SVC shows the best overall recall, whereas TS NN consistently has the highest precision across most of the 32 FunSoCs. In hard-to-classify FunSoCs like nonviral invasion and bacterial counter signaling, TS NN performs poorly indicating a model with a high degree of variance. Similarly, TS BI. SVC suffers from poor precision in most cases. The majority vote classifier improves on the BI. SVC + NN (OS) and finds an optimal balance between precision and recall across all FunSoCs.

Analysis of Simulated Novel Pathogens

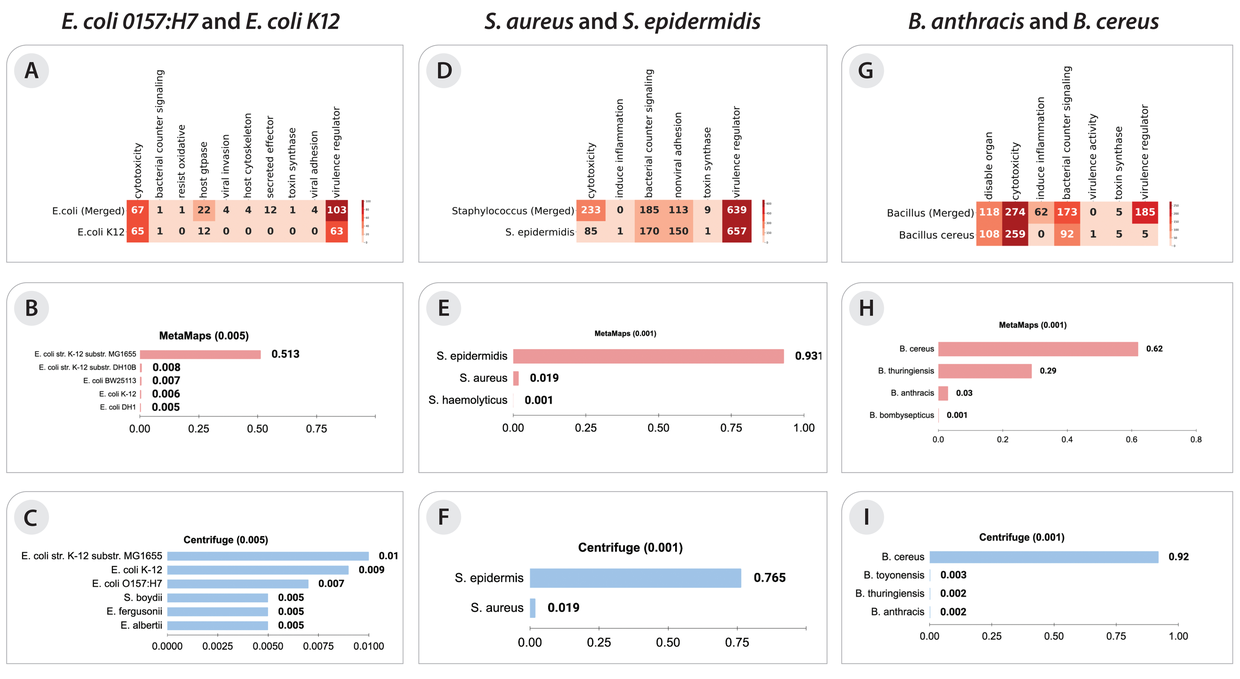

Figure 2: Analysis of simulated novel pathogens using SeqScreen-Nano, MetaMaps, and Centrifuge.

A. FunSoCs for merged *E. coli* O157:H7 and *E. coli* K12 strains vs. *E. coli* K12. Numbers represent the number of ORFs detected with that FunSoC.
B. MetaMaps results for merged *E. coli* genomes
C. Centrifuge results for merged *E. coli* genomes
D. FunSoCs for merged *S. aureus* and *S. epidermidis* vs. *S. epidermidis*
E. MetaMaps results for merged Staphylococcus genomes
F. Centrifuge results for merged Staphylococcus genomes
G. FunSoCs for merged *B. anthracis* and *B. cereus* vs. *B. cereus*
H. MetaMaps results for merged Bacillus genomes
I. Centrifuge results for merged Bacillus genomes.

signature science LLC
Signature Science, LLC
8501 North Mopac Expressway
Austin, TX 78759

RICE UNIVERSITY
Rice University
6100 Main Street
Houston, TX 77005