SeqScreen: a biocuration platform for robust taxonomic and biological process characterization of nucleic acid sequences of interest

Dreycey Albin*[1], Dan Nasko*[3], R. A. Leo Elworth[2], Jacob Lu[2], Advait Balaji[2], Christian Diaz[2], Nidhi Shah[3,4], Jeremy Selengut[3], Chris Hulme-Lowe[5], Pravin Muthu[6], Gene Godbold[7], Mikael Lindvall[8], Madeline Diep[8], Adam Porter[4,8], Mihai Pop[3,4], Krista Ternus†[5], and Todd J. Treangen†[1,2]

[1]Program in Systems, Synthetic and Physical Biology, Rice University, Houston, TX.
[2]Department of Computer Science, Rice University, Houston, TX
[3]Center for Bioinformatics and Computational Biology, University of Maryland College Park.
[4]Department of Computer Science, University of Maryland, College Park, MD.
[5]Signature Science, LLC, Austin, TX.
[6]Signature Science, LLC, Arlington, VA.
[7]Signature Science, LLC, Charlottesville, VA.
[8]Fraunhofer Center for Experimental Software Engineering, College Park, MD.
*These authors have contributed equally to this work.
†These authors have equally supervised this work.

I. Abstract

Rapid advancements in synthetic biology and nucleic acid synthesis, in particular concerns about its intentional or accidental misuse, call for more sophisticated screening tools to identify genes of interest within short sequence fragments. One major gap in predicting genes of concern is the inadequacy of current tools and ontologies to describe the specific biological processes of pathogenic proteins. The objective of this work is to design software that sensitively assigns taxonomic classifications, functional annotations, and biological processes of interest to short nucleotide sequences of unknown origin (50bp-1,000bp). The overarching goal is to perform sensitive characterization of short sequences and highlight specific pathogenic biological processes of interest (BPoIs). The SeqScreen software executes these tasks in analytical workflows with Nextflow and outputs results in a tab-delimited report. Local and global alignments differentiate hits to taxonomically-related sequences from similar but unrelated sequences, and an ensemble approach leverages multiple tools and databases to assign a variety of functional terms to each query sequence. Final biological process assessments are made from the predicted functional annotations, which leverage information in pre-existing databases, as well as new custom biocurations. Machine learning models predict each biological process of interest on large protein databases before incorporation into the SeqScreen framework to streamline computational efficiency, ensure reproducible results, allow for version control, and facilitate the review of the automated predictions by expert biocurators. The SeqScreen source code is available at https://gitlab.com/treangenlab/seqscreen.

II. Introduction

A. Background

The cost of synthesizing customized nucleic acid sequences has dropped significantly in the last two decades, and the number of individuals ordering designer oligonucleotides has skyrocketed. While these advances deliver many benefits, they also introduce the risk of inadvertently or intentionally creating a potential biothreat (1; 2). This concept is commonly referred to as the "dual-use dilemma", where research could be used for both good and bad. This is particularly important when considering applications of synthetic biology. One specific example is the genetic engineering of gene drives. Gene drives aim to expand a specific trait within a population, thus allowing for designer phenotypes to be spread throughout a local community. Many gene drives have focused on the regulation of mosquito populations to prevent the spread of mosquito-borne diseases like malaria (3; 4); however, gene drives may also be used, for example, to select for mosquito populations that transport bacterial toxins (5). This represents only one particular example, yet related dual use cases have been seen throughout synthetic biology and other biotechnological fields.

The use of methods stemming from bioinformatics to screen potentially dangerous sequences offers a promising route for biodefense. The 2018 Winter Mid-Atlantic Microbiome Meet-up (M^3) demonstrated different use cases for bioinformatics tools in biodefense and pathogen surveillance, and furthermore, the conference helped to shed light on current pitfalls of such approaches (6). One major limitation is that public

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

sequence databases and standardized ontologies, such as the Gene Ontology (7; 8), offer a limited set of features to describe pathogenic functions. This has opened the door for critical bioinformatic advancements to enhance functional predictions from potentially dangerous sequences.

While guidelines and best practices have been previously proposed (9), accurate and complete characterization of synthetic DNA sequences remains an open challenge that requires novel approaches (10; 11). A study by Mahé and Tournoud investigated the ability of using regression models to select for k-mers involved in resistance to certain antibiotics. This effectively created a probabilistic model for predicting resistance from the DNA sequence alone (12). Other research has focused on the danger of short peptides, which are increasingly being explored for their promising therapeutic uses. One such study used support vector machines to discriminate toxic peptides from non-toxic peptides. While this research offered little explanation to the mechanisms of the toxicity, it showcased the ability of computational approaches to quickly screen proteins (13). A related study focused on creating a computational model for predicting the hemolytic capabilities of a given peptide, further illustrating the benefits of computational approaches in predicting possible dangers using only the sequence (14). There also exist specific toxin prediction tools, such as those targeted for cone snail toxin (conotoxin) identification (15). While these research endeavors illustrate the use of bioinformatics for biodefence, there is an existing technical gap in efficiently predicting the functional potential danger of a given sequence.

SeqScreen is a computational pipeline developed to leverage expert biological knowledge and machine learning approaches to better curate pathogenic proteins. As a proof of concept, a selection of biological processes of interest (BPoI) were compiled in house to further annotate and highlight proteins with important pathogenic features. SeqScreen also offers a "fast mode" option to rapidly find common sequences matches, as well as a "default mode" to more sensitively determine taxonomic and functional characterizations. The outputs of either approach can then be used by machine learning models to predict the BPoIs for each input query sequence. In summary, SeqScreen's novel pipeline fills a current gap in software needed for determining the pathogenic features of interest from an unknown sequence.

B. SeqScreen High-Level Design and Components

The SeqScreen pipeline was constructed as a means of sensitively characterizing short, unknown oligonucleotide sequences. SeqScreen was built using a durable, lightweight and portable workflow manager called Nextflow (16), which allows for modular workflows to be combined into a single systematic pipeline for characterizing input sequences (Figure 1). The pipeline accepts nucleotide FASTA files as input. The initialization workflow converts all ambiguous nucleotides in this input to their corresponding unambiguous possibilities, and also performs six-frame translations of nucleotide to amino acid sequences for input into downstream modules. After initialization, the many included modules work together to output a wide variety of functional and taxonomic information that aids in predicting BPoIs.

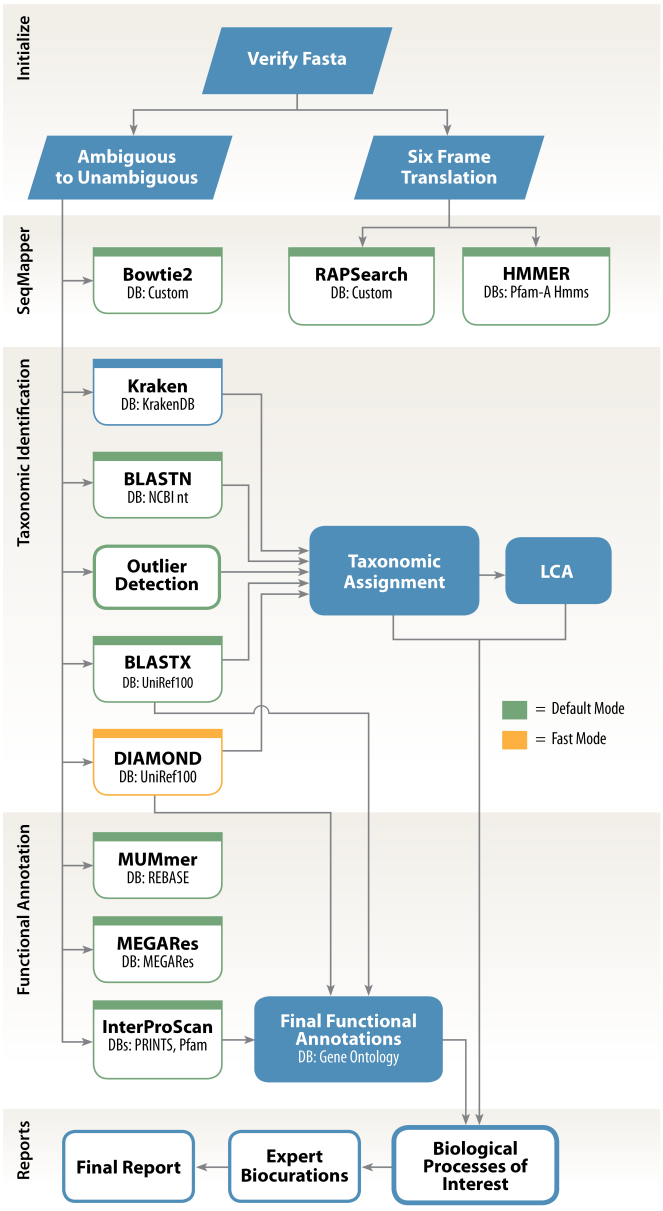

Fig. 1. Illustration of SeqScreen workflows and modules. Sequences are processed through multiple characterization workflows, followed by the generation of an automated final report to highlight biological processes of interest. The blue modules are those used in both default and fast modes of SeqScreen. Green modules are used exclusively in default mode and the yellow module is only used in fast mode.

1) SeqMapper: The SeqMapper workflow of the SeqScreen pipeline is focused on traditional detection of Biological Select Agents and Toxins (BSAT) sequences through nucleotide (Bowtie2 (17)) and amino acid (RAPSearch2 (18)) alignments to BSAT reference genomes, as well as the detection of a variety of other features. While this workflow reports hits to

BSAT genes and proteins, it does not determine whether or not a gene is of interest at a functional level (e.g., BSAT housekeeping and toxin genes have equivalent hits in this portion of the SeqScreen pipeline with no functional differentiation between the two). Downstream workflows are used to capture and collate this functional information and identify biological processes of interest. Other features of interest captured by the SeqMapper workflow include HMMs identified by HMMER (19) from Pfam (20) proteins.

2) Taxonomic Classification: The "default" and "fast" modes of SeqScreen execute slightly different taxonomic classification workflows. Both modes first run Kraken (21) and KronaTools ktGetLCA (22) to quickly assign the lowest common ancestor (LCA) to a given input sequence. Both options also leverage alignment methods to known organisms classified with NCBI taxonomic IDs, but fast mode uses DIAMOND ((23)) rather than BLAST (24) to speed up the process. Default mode runs the more computationally intensive steps of BLASTN with NCBI nt database (24) and BLASTX (24) (25) with UniRef100 database (26) to sensitively query expansive public nucleotide and protein databases. The cutoff for top hits within the BLASTN results is determined by an outlier detection method, where Bayesian Integral Log-Odds scores (27) split the sequence alignments (28).

After determining the most likely taxa for a given input sequence, an ensemble approach is used for the final taxonomic assignments. Top hits are independently selected from the outputs of Kraken and BLASTN/BLASTX (default mode) or DIAMOND (fast mode) and merged together in a single file. The relative likelihood for each taxonomic assignment is provided in the final output file, and an optional LCA step is available if users would like to report a single taxon for each sequence from its array of top hits.

3) Functional Characterization: The functional characterization workflow of the SeqScreen pipeline uses high performance tools to identify a broad range of functions in the input sequences. Similar to the taxonomic classification workflow, default and fast mode execute different tools and databases. In default mode, the unambiguous nucleotide sequences are queried for restriction sites through MUMmer (29) alignments to REBASE (30), a restriction enzyme database, and screened for antimicrobial resistance genes through BLASTN alignments to MEGARes (31), an antibiotic resistance database. Default mode also runs InterProScan (32) for protein motif detection and functional predictions. InterProScan is equipped with a database that integrates several member protein databases together, which enables queries into multiple rich sources of protein information (33). Fast mode does not run these modules and relies solely on DIAMOND (23), a module exclusive to fast mode, to retrieve functional information.

4) Compiling Functional Annotations: Both default and fast modes compile functional annotations with a final assignment script that combines all of the module outputs and reports molecular function and biological process GO terms. This script identifies GO terms found with BLASTX or DIAMOND in the taxonomic classification workflow, as well as GO terms detected by other tools run in default mode. The GO terms for each sequence, out of all possible GO terms, are generated and added to a unique set to prevent duplicates. The GO term assignment is comprehensive in that it also includes obtaining parent GO IDs. The output report from this approach includes an expectation value (E-value), a predicted gene name, predicted UniProt ID, and the selected GO terms, leveraging the strengths of all programs used in the SeqScreen functional characterization workflow.

C. Machine Learning Methods for Predicting Biological Processes of Interest

1) Truth Set Data from Biocurator Classifications: Manual classifications by biocurators are essential for many reasons, but they are also time consuming and susceptible to bias from an individual's experience or opinions. Machine learning offers tools that can serve as an accelerator and aid to biocurators to identify new features and capture overall method performance. Biocurators can review information output by machine learning algorithms to verify that results are valid, and provide additional training data and feedback to predictive models to improve them over time. Such a process expedites and scales the biocuration process beyond what is currently possible with limited staffing and funding (34). With this in mind, the proteins identified by the biocurator definitions in Table I were used as a gold standard truth data set for training a machine learning model to determine if more examples could be curated for the BPoIs in SeqScreen. Although a large number of GO terms and keywords were correlated with each BPoI, these were often not specific enough to identify pathogenic functions because many GO terms and keywords apply to both host and pathogen biological processes. To improve the precision in identifying only pathogenic proteins, all biocurator-defined BPoIs definitions excluded proteins that were derived from taxa that did not contain pathogens (e.g., archaea (taxonomy:2157), S. cervisiae (taxonomy:4932), C. elegans (taxonomy:6239), mammals (taxonomy:40674), fruit fly (taxonomy:7215), bony fishes (taxonomy:7954), birds (taxonomy:8782), rice (taxonomy:4527), or Arabidopsis (taxonomy:3701)).

2) Connection to SeqScreen: A set of highly rated (4 and 5 star) SwissProt proteins were annotated according to the biocurator-defined BPoI classifications in Table I, and subsequently used as the truth data set to train and evaluate all machine learning analyses described here. In addition to BPoIs, other biocurated features were included in this truth data set, such as whether a protein targets humans, primates, livestock, or crops. After training, the models were used to generate BPoI assessments for the entire UniProt databank, thus creating a pre-computed lookup table that can be used by SeqScreen. After SeqScreen assigns UniProt IDs to input sequences, this lookup table maps each UniProt ID to the set of BPoIs predicted to be associated with each protein. Figure 2 shows a breakdown of the BPoIs predicted for all proteins in UniProt.

TABLE I

AS A PROOF OF CONCEPT TO BETTER CHARACTERIZE SEQUENCES OF INTEREST, BIOCURATORS DEFINED A SMALL SET OF RELEVANT BPOIS FROM PROTEINS WITH WELL-DOCUMENTED NEGATIVE EFFECTS ON HOST CELL BIOLOGY. BIOCURATORS EVALUATED HOW EXISTING GO TERMS AND KEYWORDS COULD BE LEVERAGED TO QUERY REVIEWED SWISS-PROT PROTEINS FOR EXAMPLES OF EACH BPOI.

BPoI	# Proteins*	% SwissProt	Biocurator Definition
Attachment to Target Cell	1,742	0.31	(KW-1168 or KW-1161 or KW-1234 or KW-1233) or (KW-0130 and KW-0843)
Specialized Bacterial Secretion System	279	0.05	(GO:30253 or GO:15628 or GO:30254 or GO:30255 or GO:46819 or GO:33103 or GO:44315)
Manipulation of Host Programmed Cell Death	32	0.01	(KW-1119 or KW-1081)
Antibiotic Resistance	1,784	0.31	((KW-0046) and not KW-0698 and not KW-0804 and not KW-0689 and not KW-0030 and not KW-0547)
Invasion into Host Cell	2,920	0.52	(GO:39665 or GO:99008 or GO:19060 or GO:35635 or KW-1160 or KW-1162 or KW-1164 or KW-0916)
Active Manipulation of Host Immune Effectors	2,195	0.39	(KW-0899 or KW-1190 or KW-1100 or KW-0922 or KW-1130 or GO:0052018 or GO:0020012 or GO:0052059)
Cytotoxicity	2,201	0.39	(KW-1213 or KW-1172 or GO:31640)
Degradation of Extracellular Matrix	759	0.14	(((KW-0645 and KW-0843) not KW-0528) or KW-1217 or KW-1061)
Disabling of Host Organ System	3,634	0.65	(KW-0123 or KW-0260 or KW-0528 or KW-0959 or KW-0872)
Total	15,546	2.77	-

**Proteins from non-pathogen taxa were eliminated from the query results.*

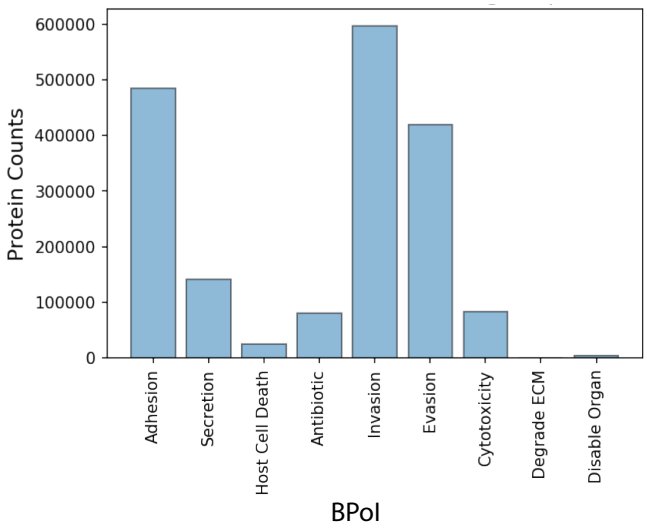

Fig. 2. Proteins assigned to each BPoI within UniProt. Overall, the number of assignments amounted to roughly 1% of the entire UniProt database.

3) Support Vector Classifier Feature Selection: To begin analyzing the impact of machine learning on SeqScreen, two simple machine learning models were evaluated. The first model was the support vector classifier for both feature selection and BPoI classification. The second method ran neural networks to classify BPoIs based on the selected features from the support vector classifier. The overall goal for the machine learning portion of SeqScreen was two fold. The first aim was to improve the curated queries by selecting GO terms and keywords not in the original UniProt queries. The second aim was to utilize these machine learning algorithms for automating BPoI classification. To satisfy the first aim, a support vector classifier (SVC) was used because of the large number of input features. In addition, SVCs may be used with an L1 penalty, ensuring that only important features are selected. Solving the second aim, an SVC with binary classification was used to indicate the presence or absence of an individual BPoI.

When classifying BPoIs, or when selecting the neural network features, we used a Support Vector Classifier (35) with a linear kernel from the sklearn.svm.LinearSVC module within the Scikit-learn library (36). The LinearSVC module uses the LIBLINEAR library for implementing the SVC, which uses a coordinate descent method for training (37; 38). The default squared hinge loss was used for the loss function and an L1 regularization penalty was selected to cause the coefficients of the classifier to become sparse (and go to zero). Separate classifiers were trained for each BPoI. The L1 regularization of the classifiers searched for GO terms and keywords that were common to a given BPoI.

To train these classifiers, the truth data set was randomly split into testing and training subsets. First, a test set of 25% (the amount was an adjustable parameter that defaulted to 25%) of the data was set aside to test the performance of

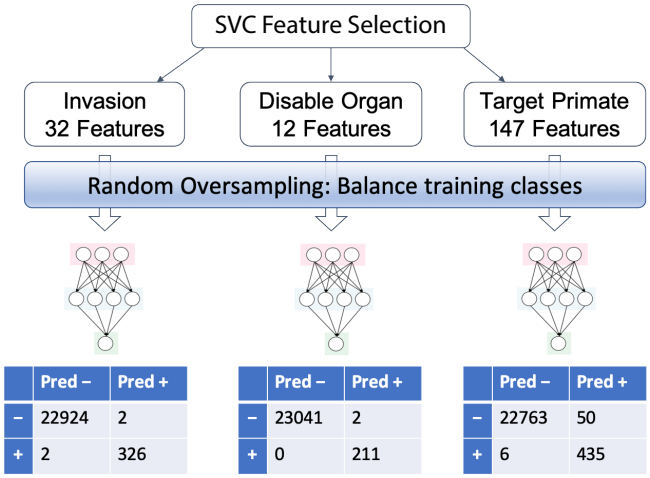

Fig. 3. Confusion matrices for three different predictions using a feed forward neural network. Shown here are the resulting confusion matrices for *Invasion*, *Disable Organ*, and *Target Primate*. Features with non-zero coefficients were used as input for the neural networks. The datasets used for discovering the non-zero coefficients were the automatic queries from UniProt. For this figure, the datasets resulted from the queries for *Invasion*, *Disable Organ*, and *Target Primate* (as shown in Table 1). Oversampling was used for class balancing, because there were a varying number of curated SwissProt sequences for each prediction.

the models after training was completely finished. In the case of the neural network training, the remaining data used to train the model was further divided into another subset (default size of 25%, but also an adjustable parameter) used as validation data to evaluate the performance of the model after each training epoch. Random oversampling is available as an option within the code base and was applied to balance the training classes (has-BPoI and has-not).

4) Prediction of BPoIs Using a Neural Network: While fitted SVC models along with the selection of relevant features can directly generate BPoI predictions, an alternative approach can use neural network learning. In this case, the SVC models are used as feature selection tools, where for each BPoI, the set of features associated with non-zero coefficients in the SVC model are used to train a neural network that acts as an individual binary classifier for the BPoI, as shown in Figure 3. These neural networks were constructed with densely connected, forward-feeding sequential layers and employed Adam (39) optimization for parameter training.

III. Results

The unification of multiple characterization workflows allows for an all-inclusive pipeline for determining the underlying BPoIs, as well as thoroughly annotating the functional profile of an unknown, potentially very short sequence. SeqScreen can be run in default mode to capture as many functional annotations as possible for BPoIs predictions, or in fast mode to use a subset of the programs with a focus on optimizing run time. Based on internal benchmarking, SeqScreen assigns the correct taxon to test sequences (50bp - 1,000bp) among its reported top hits at greater than 90 percent accuracy. GO terms are assigned with an average precision rate of greater than 70 percent and recall of greater than 80 percent (data not shown).

TABLE II
Support vector classifier selected features (GO terms and keywords) for the BPoI *Adhesion*. The features are ordered in descending order of coefficients.

New	Feature ID	Biological Process of Interest	Weight
NO	kw-1161	Viral attachment to host cell	1.3847
NO	kw-0130	Cell adhesion	1.3625
NO	kw-0843	Virulence	1.1773
NO	kw-1168	Fusion to host membrane	0.946
YES	kw-1160	Virus entry into host cell	0.9058
YES	go:0055036	virion membrane	0.4351
YES	go:0019031	viral envelope	0.4291
YES	go:0019062	virion attachment to host cell	0.2933
YES	kw-0167	Capsid protein	0.1895
YES	kw-1162	Viral penetration into host cytoplasm	0.1884
YES	kw-1188	Viral release from host cell	-0.4117

TABLE III
Features selected by using the L1 regularization in the support vector classifier training. The "Query" column gives the number of features used for the original mappings. The "SVC" column shows the total count of features used by the support vector classifier.

BPoI	Query	SVC	Intersection	Union
Adhesion	6	11	4	13
Secretion	7	4	2	9
Host Cell Death	2	2	2	2
Antibiotic	6	8	4	10
Invasion	8	4	3	9
Evasion	9	7	4	12
Cytotoxicity	3	11	3	11
ECM degradation	5	7	4	8
Disable Organ	5	5	5	5

1) Machine Learning Results: As a first step in analyzing the results of the machine learning models, BPoI features selected by SVCs were analyzed. Biocurators manually reviewed these non-zero weight BPoI features and determined them to be reasonable. An example showing features selected for one BPoI and their overlap with the original manually curated features is shown in Table II. Statistics for the overlap between manual biocurations and machine learning-derived features for all BPoIs are shown in Table III. Additional testing on larger data sets will confirm how well this model performs on predicting BPoIs outside of the truth set.

Next, the performance for each of the two methods was evaluated for BPoI classification. Performance metrics were obtained for using the SVC alone for classification, as well as the SVC and neural network classification scheme. The input for both classification schemes is a binary vector indicating the presence or absence of each unique possible feature for a given protein (GO terms and keywords), and the binary output indicates whether each BPoI has been predicted or not. The metrics evaluate the performance of the two different methods on the test set, which is left out at the beginning and only used after the completion of all epochs of the training process (an adjustable parameter with a default value of 4). Results

TABLE IV
SCORING METRICS FOR THE ACCURACY, PRECISION, AND SENSITIVITY OF THE DIFFERENT APPROACHES USED FOR BPOI PREDICTION FROM GO TERMS AND KEYWORDS.

Method	Accuracy	Precision	Sensitivity
Support Vector Classifier	99.72%	99.52%	94.60%
Feed-forward NN	99.39%	93.75%	98.09%

for the two methods are shown in Table IV. To gauge a model's performance, its accuracy, precision, and sensitivity was quantified with the following equations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$
$$Precision = \frac{TP}{TP+FP}$$
$$Sensitivity = \frac{TP}{TP+FN}$$

Based on these results, it was found that the neural network classification scheme generally trades a loss in precision for an increase in sensitivity over scheme using only the SVC.

IV. DISCUSSION

While these advancements demonstrate the benefits of using a hybrid biocuration platform by combining expert curation, leading bioinformatics tools and databases, and ML-based approaches for robust screening of sequences of interest, there are still many areas that could be optimized. For one, tools in computational biology could benefit from more standardization and benchmarking data, showing how tools comparatively perform on a given test set. This allows for community members to choose the tools that have the highest relative performance and are most likely to be helpful for a specific use case (6). SeqScreen was developed in a modular way, so its components can each be updated or replaced in the future if new bioinformatics tools and databases are shown to outperform its current modules and workflows. As a whole, SeqScreen would certainly benefit from comparisons to other hybrid biocuration platforms, but to our knowledge, no alternate tools for characterizing these kinds of BPoIs exist in the open source community.

Another bottleneck of this approach is scalability, since it relies on expert biocurations, GO terms, and keywords for training and test data sets. Given that less than 1% of all proteins contained in UniProt have high quality annotations, and that Gene Ontology (GO) terms have limited reach into BPoIs, there is a critical need for additional curation efforts or functional feature sets to scale this approach. Manual biocurations have provided an invaluable resource for this project, allowing expert-level curations of each protein to be used for BPoI validation. Accurately scaling biocurations will improve the quality of data in predictive models, and may also serve as a competitive advantage in saving time and money (34). Likewise, crowdsourcing (40) efforts have shown promise in data retrieval and may represent a promising approach for expanding, reviewing, and scaling future biocurations.

A focal point in SeqScreen development was explainability, since machine learning approaches can be applied without any understanding of the reasoning behind the underlying "black box" model, leaving end users without any justification for their results (41). In some cases, machine learning results may also not be reproducible or consistent on the same query sequence from one run to the next. To address this, machine learning-based predictions are pre-computed on all of UniProt and integrated in SeqScreen as a static lookup file. This allows end users to view all possible BPoI predictions to assess their biological accuracy, and it ensures that SeqScreen will reproducibly produce the same result for every protein detected in every run. To increase explainability further in the future, the machine learning models used for BPoIs could be solely focused on retrieving the relevant features. These features could then be used in an end user-curated classification scheme for determining potential BPoIs.

Future work can also include additional time spent looking into how best to merge the machine learning methods with manual curation in a harmonious way. Currently, both machine learning classification schemes used in this paper use a final feature set determined by the SVC coefficient weights. While it was found that these selected features overlap well with the previously supplied manually curated features, as well as add new and useful features, the final feature sets often exclude a few features that had been previously manually curated (Table III). Alternative strategies could include taking the union set of both manual and ML based features or a hybrid approach that iteratively refines features based on alternating rounds of ML and expert curation.

The pipeline presented within SeqScreen is generalizable to a number of different questions arising in synthetic biology, biodefense, functional genomics, metagenomics, public health, and pathogen biosurveillance programs. Its novel ensemble taxonomic classification approach, along with its extremely thorough functional characterization capabilities, yields a detailed report on all genes within a data set, as well as highlights particular genes of interest or pathogenic importance. As a pipeline aimed at the detailed functional characterization of microbes from short sequence fragments, there are many potential applications for SeqScreen, including a tool for evaluating the functional profile of metagenomes or discovering new antimicrobial resistance elements. Most biological questions could benefit from a deeper understanding of the relevant functional mechanisms involved, giving SeqScreen broad applicability and potential for further development.

V. ACKNOWLEDGEMENTS

We would like to acknowledge Letao Qi and Chris Jermaine of Rice University for insightful discussions specific to the machine learning algorithms. We would also like to thank Signature Science staff members Nicolette Albright and Katharina Weber for their assistance in developing the UniProt queries, Isaac Mayes and Don Bowman for creating a database for our internal test data sets, and Chris Grahlmann for his feedback on the software design.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1842494. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. All of the co-authors were either fully or partially supported by the FunGCAT 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.

REFERENCES

[1] T. W. Drew and U. U. Mueller-Doblies, "Dual use issues in research a subject of increasing concern?" *Vaccine*, vol. vol. 35, no. 44, pp. pages 5990–5994, 2017.

[2] B. Erickson, R. Singh, and P. Winters, "Synthetic biology: regulating industry uses of new biotechnologies," *Science*, vol. 333, no. 6047, pp. 1254–1256, 2011.

[3] A. Hammond, R. Galizi, K. Kyrou, A. Simoni, C. Siniscalchi, D. Katsanos, M. Gribble, D. Baker, E. Marois, S. Russell *et al.*, "A crispr-cas9 gene drive system targeting female reproduction in the malaria mosquito vector anopheles gambiae," *Nature biotechnology*, vol. 34, no. 1, p. 78, 2016.

[4] K. Kyrou, A. M. Hammond, R. Galizi, N. Kranjc, A. Burt, A. K. Beaghton, T. Nolan, and A. Crisanti, "A crispr–cas9 gene drive targeting doublesex causes complete population suppression in caged anopheles gambiae mosquitoes," *Nature biotechnology*, vol. 36, no. 11, p. 1062, 2018.

[5] L. J. Getz and G. Dellaire, "Angels and Devils: Dilemmas in Dual-Use Biotechnology," *Trends in Biotechnology*, vol. vol. 36, no. 12, pp. pages 1202–1205, 2018.

[6] J. S. Meisel, D. J. Nasko, B. Brubach, V. Cepeda-Espinoza, J. Chopyk, H. Corrada-Bravo, M. Fedarko, J. Ghurye, K. Javkar, N. D. Olson *et al.*, "Current progress and future opportunities in applications of bioinformatics for biodefense and pathogen detection: report from the Winter Mid-Atlantic Microbiome Meet-up, College Park, MD, January 10, 2018," 2018.

[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, p. 25, 2000.

[8] G. O. Consortium, "The gene ontology resource: 20 years and still going strong," *Nucleic acids research*, vol. 47, no. D1, pp. D330–D338, 2018.

[9] M. Fischer and S. M. Maurer, "Harmonizing biosecurity oversight for gene synthesis," *Nature biotechnology*, vol. 28, no. 1, p. 20, 2010.

[10] J. Diggans and E. Leproust, "Next Steps for Access to Safe, Secure DNA Synthesis," *Frontiers in bioengineering and biotechnology*, vol. 7, p. 86, 2019.

[11] D. DiEuliis, S. R. Carter, and G. K. Gronvall, "Options for synthetic DNA order screening, revisited," *mSphere*, vol. 2, no. 4, pp. e00 319–17, 2017.

[12] P. M. . M. Tournoud, "Predicting bacterial resistance from whole-genome sequences using k-mers and stability selection," *BMC Bioinformatics*, vol. vol. 383, no. 19, 2018.

[13] S. Gupta, P. Kapoor, K. Chaudhary, A. Gautam, R. Kumar, G. P. Raghava, O. S. D. D. Consortium *et al.*, "In silico approach for predicting toxicity of peptides and proteins," *PloS one*, vol. 8, no. 9, p. e73957, 2013.

[14] K. Chaudhary, R. Kumar, S. Singh, A. Tuknait, A. Gautam, D. Mathur, P. Anand, G. C. Varshney, and G. P. Raghava, "A web server and mobile app for computing hemolytic potency of peptides," *Scientific reports*, vol. 6, p. 22843, 2016.

[15] Y.-X. Fan, J. Song, X. Kong, and H.-B. Shen, "PredCSF: an integrated feature-based approach for predicting conotoxin superfamily," *Protein and peptide letters*, vol. 18, no. 3, pp. 261–267, 2011.

[16] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature biotechnology*, vol. 35, no. 4, p. 316, 2017.

[17] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[18] Y. Zhao, H. Tang, and Y. Ye, "RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data," *Bioinformatics*, vol. 28, no. 1, pp. 125–126, 2011.

[19] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W29–W37, 2011.

[20] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry *et al.*, "Pfam: the protein families database," *Nucleic acids research*, vol. 42, no. D1, pp. D222–D230, 2013.

[21] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome biology*, vol. 15, no. 3, p. R46, 2014.

[22] B. D. Ondov, N. H. Bergman, and A. M. Phillippy, "Interactive metagenomic visualization in a web browser," *BMC bioinformatics*, vol. 12, no. 1, p. 385, 2011.

[23] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nature methods*, vol. 12, no. 1, p. 59, 2015.

[24] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[25] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Pa-

padopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," *BMC bioinformatics*, vol. 10, no. 1, p. 421, 2009.
[26] B. E. Suzek, H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," *Bioinformatics*, vol. 23, no. 10, pp. 1282–1288, 2007.
[27] S. F. Altschul, J. C. Wootton, E. Zaslavsky, and Y.-K. Yu, "The construction and use of log-odds substitution scores for multiple sequence alignment," *PLoS computational biology*, vol. 6, no. 7, p. e1000852, 2010.
[28] N. Shah, S. F. Altschul, and M. Pop, "Outlier detection in blast hits," *Algorithms for Molecular Biology*, vol. 13, no. 1, p. 7, 2018.
[29] G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin, "MUMmer4: a fast and versatile genome alignment system," *PLoS computational biology*, vol. 14, no. 1, p. e1005944, 2018.
[30] R. J. Roberts, T. Vincze, J. Posfai, and D. Macelis, "REBASEa database for DNA restriction and modification: enzymes, genes and genomes," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D234–D236, 2009.
[31] S. M. Lakin, C. Dean, N. R. Noyes, A. Dettenwanger, A. S. Ross, E. Doster, P. Rovira, Z. Abdo, K. L. Jones, J. Ruiz *et al.*, "MEGARes: an antimicrobial resistance database for high throughput sequencing," *Nucleic acids research*, vol. 45, no. D1, pp. D574–D580, 2016.
[32] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez, "InterProScan: protein domains identifier," *Nucleic acids research*, vol. 33, no. suppl_2, pp. W116–W120, 2005.
[33] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka *et al.*, "InterProScan 5: genome-scale protein function classification," *Bioinformatics*, vol. 30, no. 9, pp. 1236–1240, 2014.
[34] I. S. for Biocuration, "Biocuration: Distilling data into knowledge," *PLoS biology*, vol. 16, no. 4, p. e2002846, 2018.
[35] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.
[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[37] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.
[38] C.-J. Hsieh, K.-W. Chang, C.-J. Lin, S. S. Keerthi, and S. Sundararajan, "A dual coordinate descent method for large-scale linear svm," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 408–415.
[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[40] D. P. Renfro, B. K. McIntosh, A. Venkatraman, D. A. Siegele, and J. C. Hu, "GONUTS: the gene ontology normal usage tracking system," *Nucleic acids research*, vol. 40, no. D1, pp. D1262–D1269, 2011.
[41] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in *ESANN*, vol. 12. Citeseer, 2012, pp. 163–172.