Control Number: 2022-A-20-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: Importance of biological sequence similarity in datasets for machine learning methods

Author Block: **A. Ferrer Florensa**[1], P. Lanken Conradsen Clausen[1], J. Almagro Armenteros[2], H. Nielsen[1], F. M. Aarestrup[1];
[1]Technical University of Denmark, Copenhagen, DENMARK, [2]Stanford University, Palo Alto, CA.

Abstract Body: The use of machine learning methods on biological sequential data has largely increased during the last years. Computational microbiology is not an exception, with methods using whole genome sequences, as host or pathogenicity prediction, to methods applied only to DNA regions or proteins, as subcellular location prediction. Although several techniques developed for other fields than computational microbiology can be used on biological sequences, there are certain particularities of this type of data that should always be considered when building a machine learning method. Homology between the sequences of a dataset is one of those particularities. Not only the similarity between data points in a raw database vary greatly, but most of the phenotypes between two homologous organisms will be usually more similar. Because of that, if homology reduction or splitting is not performed on the dataset, machine learning methods can tend to learn to predict based on similarities with sequences on the training set, instead of learning to perform the task based on the function underlying the data. In fact, the machine learning methods using sequence similarity as prediction can be specially damaging in microbiology, where the amount of annotated data can vary a lot through the tree of life. Moreover, the effect of the distribution of homologous sequences on the database on the accuracy of the algorithm can be difficult to detect, as it compromises of the test set. In this work, it has been studied the overfitting effect that can cause homology in datasets. And, subsequently, we have developed a pipeline to split or reduce the homology in the dataset, able to handle long sequences and large databases at a competitive performance level.

---PAGEBREAK---

Control Number: 2022-A-67-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: A Highly Generalizable Semi-supervised DeepSVDD Methodology for Detecting Anomalous Metagenomic Samples

Author Block: **C. W. Price**, J. Russell;
MRIGlobal, Falls Church, VA.

Abstract Body: The composition of metagenomic communities within the human body often reflects localized medical conditions such as upper respiratory diseases and gastrointestinal diseases. Fast and accurate computational tools to flag anomalous metagenomic samples from typical samples are desirable to understand different phenotypes. Deep learning, specifically the semi-supervised Deep Support Vector Data Description (DeepSVDD) algorithm, can reliably learn a typical one-class metagenome representation that generalizes well to any metagenomic anomaly detection problem. Here we present a methodology that utilizes two types of DeepSVDD models, one trained on taxonomic feature space output by the Pan-Genomics for Infectious Agents (PanGIA) taxonomy classifier and one trained on kmer counts. Utilizing multiple feature spaces adds robustness to the model, as the taxonomy feature space is strong in cases where a single key taxon may be anomalously absent or present and the kmer feature space is strong in cases where taxonomic classification may not change meaningfully, but the nucleotide composition has generally changed. We demonstrate the generalizability of this methodology to three diverse data sets. The first data set is retrieved from a publication hosted on the NIH Sequence Read Archive (SRA) which contains a set of nasopharyngeal swabs from healthy and COVID-19 positive patients (n=1062). DeepSVDD is able to learn a typical healthy nasopharyngeal swab sample and reliably flag the COVID-19 positive samples in both feature spaces. The second data set is also retrieved from SRA, and is made up of gut microbiome samples from normal controls and from patients with slow transit constipation (STC) (n=2262). Again, DeepSVDD is able to reliably flag STC samples from the typical control as anomalous. The final data set is a synthetic metagenome data set created by CAMISIM. PanGIA identified 50 common taxa present in a set of laboratory sequencing blanks, which were used to create a control laboratory blank data set. The experimental conditions included 12 different spiked-in contaminants that are taxonomically similar to taxa present in the laboratory blank sample ranging from one strain distance away to one family distance away. DeepSVDD was again able to flag the contaminant inserts as anomalous. We believe this methodology is highly generalizable to a variety of metagenomic anomaly detection problems such as identification of diseased samples, improving laboratory confidence in background samples, monitoring drug affects on the metagenome, and more.

---PAGEBREAK---

Control Number: 2022-A-74-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: Deep Learning to Guide Outbreak and Pandemic Response

Author Block: **B. Hu**, M. Babinski, P. Chain;
Los Alamos National Laboratory, Los Alamos, NM.

Abstract Body:
A significant fraction of pathogens known to infect humans originate in non-human (zoonotic) hosts and new and emerging pathogens continue to spill over into the human population more frequently at an alarming rate (e.g., SARS, MERS, Cholera, etc.). The recent outbreaks of Ebola virus in West Africa and the ongoing SARS-CoV-2 pandemic demonstrate the need for rapid and reliable assessments of viral phenotype information to help inform scientists and policy makers how best to control the spread of disease. Further understanding of the virus pathogenic evolutionary space and potential trajectory could guide appropriate control measures to limit the spread of a new virus throughout the local and global human population.

Once a viral disease begins to circulate, reliable diagnostics, protective vaccines and therapeutic antibodies are essential tools for preventing, monitoring, and managing disease spread. However, the efficacy of these tools can be diminished by mutations in viral genomes, and the delay between the emergence of new viral strains and redesign of vaccines and diagnostics allows for continued viral transmission. Given the combinatorial explosion of potential mutations that could enable a virus to "escape" diagnostics, vaccines and antibodies, and the high cost of biomedical research, it is essential to focus countermeasure development efforts only on viral strains that pose the highest risk to society. Towards this end, the questions we ask are: Is it possible to predict the most likely evolutionary trajectory of circulating genomes and anticipate novel variants before they emerge? Is it possible to assess the risk of future variants by computationally predicting key virulence determinants and exploring the evolutionary space for pathogenicity?

To address these questions using the machine learning approach, we have developed several neural networks using deep mutational scanning (DMS) data. This simple model was able to predict fairly accurately the RBD expression and binding to ACE2 ($R^2 = 0.76$). It only takes less than a second for the model to predict the effect of an arbitrary mutation, regardless of the combinatorial complexity, on a consumer PC. Recently, principal component analysis of amino acid biochemical properties and graph neural network (GNN) to learn protein properties have been combined to predict antibody binding and enzyme activities in five proteins (Gelman et al. 2021). Using a similar approach, we have developed a GNN model to study RBD and are currently evaluating the model. Combining DMS and deep learning, we can predict the mutational effects of SARS-CoV-2 RBD, both in its expression and binding to the ACE2 receptor. We are evaluating different ML models and will deploy either the best model or an ensemble of models to our existing SARS-CoV-2 sequence monitoring workflow. The upgraded workflow will be able to rank the latest mutations based on potential threat level.

---PAGEBREAK---

Control Number: 2022-A-90-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: Prediction of Bacterial Transcripts from Direct RNA sequencing

Author Block: **J. S. Mattick**[1], R. E. Bromley[1], J. F. Lebov[2], K. Watson[1], T. S. Tyson[1], B. C. Sparklin[3], D. A. Rasko[1], J. C. Dunning Hotopp[1];
[1]Institute of Genome Sciences at the University of Maryland Baltimore, Baltimore, MD, [2]Personal Genome Diagnostics, Baltimore, MD, [3]AstraZeneca, Colombia, MD.

Abstract Body:
While eukaryotic annotation allows for quantification of transcripts, the lack of experimental data defining bacterial transcript structures precludes this in bacterial analyses. Instead, bacterial differential expression analyses often use CDSs, despite there being numerous well-documented issues with using CDSs. Advances in Oxford Nanopore Technology (ONT) have allowed for the direct sequencing of long RNA reads. The application of this technology to bacterial transcriptomes has allowed us to observe the diversity of bacterial transcriptional structures in several bacteria including *Escherichia coli*. Here, we use ONT direct RNA sequencing to predict transcript structure in *E. coli* K12 and *E. coli* E2348/69, the latter of which has a key pathogenicity island called the LEE operon. Transcript structures were predicted using cultures grown under a variety of conditions including low nutrient DMEM and high nutrient LB media. We use a three-phase modeling approach in the prediction of transcriptional start and stop sites. First, regions are identified with continuous sequencing depth, which may include numerous overlapping transcripts, particularly in genomes with a high coding density. Next, sites enriched for read start and stop sites are identified, followed by identification of putative full-length RNA sequencing reads that span from an enriched start site to an enriched stop site. Then, all complete and incomplete reads that may have arisen from those transcripts are filtered out, and the remaining reads are used to predict longer transcripts that occur at lower frequencies. Applying this model to the entire E2348/69 genome identified 114 regions with sufficient depth for transcript prediction, with a total of 701 transcripts predicted over these regions (approximately 6.1 distinct transcripts per region). Work is ongoing to implement this algorithm into publicly available software. Overall, our goal is to develop a novel tool to predict bacterial transcripts to improve bacterial transcriptomics research, including differential expression analyses.

---PAGEBREAK---

Control Number: 2022-A-102-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: The ARETE Pipeline: Tracking antimicrobial resistance on the move

Author Block: A. Manuele[1], F. Maguire[1], H. Sanderson[1], R. C. Fink[1], K. L. Gray[2], F. Brinkman[2], **R. G. Beiko**[1];
[1]Dalhousie University, Halifax, NS, CANADA, [2]Simon Fraser University, Vancouver, BC, CANADA.

Abstract Body: The spread of antimicrobial resistance (AMR) has been heavily influenced by within- and between-species transmission through recombination and lateral gene transfer. No single inference tool can generate the entire picture of AMR transmission, and multiple lines of evidence must be integrated in order to obtain a comprehensive view. With tens of thousands of genomes available for multiple pathogen-containing species, automation and reproducibility are essential.

We introduce ARETE (https://github.com/beiko-lab/arete), a new software pipeline designed to automate the inference of evolutionary dynamics of AMR and associated genes, and mobile genetic elements (MGEs). ARETE proceeds in three steps: (i) Genome assembly and quality control; (ii) Annotation of genes and MGEs, pan-genome inference, and phylogenomic analysis; and (iii) Inference of transmission patterns using phylogenetic tree comparisons, recombination detection, and coevolutionary analysis. Each of these steps can also be run independently to enable compatibility with other related workflows such as Bacass or Bactopia. ARETE is implemented using the Nextflow pipeline framework and has been tested in multiple high-performance computing environments using the Slurm scheduler. Nextflow implementation allows for easy extensibility using Conda packages and Docker / Singularity containers.

We used the first two steps of ARETE to analyze 1766 genomes of Enterococcus faecium collected from multiple habitats in Alberta, Canada and the United Kingdom. Genomes were re-assembled consistently using Unicycler, and relevant genes (AMR, heavy-metal resistance genes, and virulence factors) and MGEs (plasmids, genomic islands, and prophages) were predicted in the 1273 genomes that passed QC. We found a negligible effect of geographic location on the distribution of most genes and MGEs: by contrast, strong associations between AMR genes, virulence factors, and MGEs vs. habitat were observed. We identified multiple plasmid clusters and genomic islands that were associated with genes conferring resistance to vancomycin, erythromycin, tetracycline, and other antimicrobials. Coevolutionary models identified positive and negative associations between copper-resistance genes, among others.

We are currently implementing the third step of ARETE, which will identify and integrate distinct lines of evidence for AMR transmission within and between species. ARETE will combine evidence from phylogenetic-tree comparisons using the rSPR software; patterns of recombination using Gubbins; and coevolutionary analysis using the recently published Community Coevolution Model. While the predictions made by these methods overlap, each has “blind spots” that can be addressed by the other approaches. The resulting aggregated predictions will allow us to provide a more-detailed view of the patterns and risks of AMR transmission in relation to phylogenetic similarity and habitat.

---PAGEBREAK---

Control Number: 2022-A-198-NGS

Session Title: **Pipe Dreams I**

Topic 1: Pipe Dreams: Analytical Methods, Bioinformatic Tools and Pipelines

Publishing Title: High-throughput screening of sequences that promote proteolysis in bacteria

Author Block: **P. Beardslee**, K. R. Schmitz;
University of Delaware, Newark, DE.

Abstract Body: **All bacteria possess multiple ATP-dependent proteases that degrade cytosolic proteins. These enzymes help maintain protein homeostasis and regulate discrete pathways, including the expression of virulence phenotypes in pathogenic bacteria, and have emerged as attractive antibacterial targets. ATP-dependent proteases are able to selectively recognize substrate proteins and ignore non-substrate proteins, minimizing harmful or wasteful off-target proteolysis. Many substrates are recognized directly by short, unstructured terminal sequences, termed degrons. While a small number of degrons have been identified, there is little known about the overarching rules that allow proteases to effectively discriminate between valid degrons and the millions of other possible terminal sequences.To address this gap in our knowledge, we have developed a cell-based screening platform that will allow us to interrogate global degron specificity and define the sequence-based rules that govern recognition of protein substrates by ATP-dependent proteases. Our method incorporates a novel selection-based screen, in which a library of protein toxin bearing a randomized terminal tag is expressed in host bacteria. Accumulation of toxin in host cells causes cell death. However, toxins bearing bona fide degrons are proteolyzed by endogenous proteases, allowing cell survival. Bacteria expressing valid degron sequences are enriched over time, and identified by Next-Generation Sequencing. Here we describe the efficacy of our method in *E. coli*, supported by NGS data from screening experiments**

performed in multiple proteolytic deletion strains. In addition to our toxin-based approach, we demonstrate a complementary method that instead utilizes a fluorescent protein substrate as a reporter for proteolysis, coupled with fluorescence-activated cell sorting (FACS) to isolate cells expressing valid degrons. The information gathered from these methods will ultimately help us understand the roles that ATP- dependent proteases play in individual pathogenic bacteria.

---PAGEBREAK---

Late Breakers

Microreact and Data-flo: flexibly enhancing the linkage and visualisation of data for genomic epidemiology

Georgina Haines-Woodhouse on behalf of the Centre for Genomic Pathogen Surveillance, University of Oxford, UK

ABSTRACT

Bringing data together from different systems within different locations and intuitive visualisations are key to addressing important genomic epidemiology questions. Data-flo and Microreact are interoperable, web-based applications that enable easy, flexible, data processing and visualisation without the need for expertise in informatics.

Data-flo [https://data-flo.io/] is an open source, modular tool for data integration and manipulation. The tool allows the creation of drag and drop pipelines that can be utilised for the joining of disparate data. Reading data from many different formats enhances the potential of combining epidemiological, genomic, laboratory and meta-data.

Microreact [https://microreact.org/] is a web-based tool with a focus on interactive data visualisation and exploration. Microreact allows the rapid rendering and linkage of phylogenetic trees, maps, networks, charts and timelines with metadata. The flexibility and utility provided by the Microreact empowers the public health sector enabling faster processing and reporting of data.

Utilising Data-flo for institutional agnostic system linkage to tailor and deliver data to Microreact can be applied to numerous applications; pathogen-agnostic and system agnostic. Here we will demonstrate the tools, their direct application, and flexibility to address genomic epidemiology questions across multiple pathogen domains in both high-income and low-income countries.