INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

Biophysics- And Biochemistry-informed Machine Learning To Predict The Effect Of Nucleotide Mutations On Molecular Detection Assays

Taylor Moehling Sandia National Laboratories **Robert Meagher** Sandia National Laboratories **Raga Krishnakumar** Sandia National Laboratories **Steve Verzi** Sandia National Laboratories **Drew Levin** Sandia National Laboratories **Daniel Krofcheck** Sandia National Laboratories **Kismet Kairon** Sandia National Laboratories **Jesse Aubin** Sandia National Laboratories **Kelsey Jorgensen** Sandia National Laboratories

All pathogens are subject to genetic drift, which raises the possibility that primer-based molecular assays that were designed against historic isolates of a pathogen may be less efficient at detecting lineages with mutations in the priming regions. The effect of signature erosion on PCR has been studied empirically, whereas there is minimal research into the effect of nucleotide modifications on isothermal methods such as loop-mediated amplification (LAMP). We experimentally explored this phenomenon for a SARS-CoV-2 LAMP assay, but the small size of the dataset was insufficient to make robust predictions. Online tools that track pathogen variation and primer performance over time exist; however, these platforms flag primers based on basic primer-target mismatches rather than experimental data. The objective of this work is to collect a large experimental dataset to develop a supervised machine learning (ML) model that can predict the probability of successful amplification for a primer set when tested with an imperfectly matched template.

We are developing a training dataset first for a SARS-CoV-2 template followed by Burkholderia thailandensis. We are modifying every base position in which the six LAMP primers bind, as well as producing a substantial set of sequences with multiple substitutions, insertions, and deletions. We are measuring the effect of these modifications on LAMP assay speed and sensitivity through experimentation. This data is being used to build a supervised ML model that can predict which mutations are detrimental for LAMP-based detection of the target. The model also considers important biophysics and biochemistry principles (i.e., free energy of primer binding to mismatched template) and incorporates sample meta data (i.e., type of modification, priming region of each mutation, exact base change).

We have designed the first 90 samples representing the RdRP gene of SARS-CoV-2 with substitutions every eighth base (group 1 rules: C to G, G to C, T to A, A to T; group 2 rules: C to A; G to T, T to C, A to G). While empirical testing is underway, we are using results from our 69 preliminary samples along with 523 data points from a published article to assess the effectiveness of various ML approaches (large language models, random forest, gradient boosting, linear regression) in predicting amplification success. Based on sequence input alone, the large language model grouped samples that amplify at similar rates, which suggests that sequences are somewhat predictive of amplification success. Using a gradient boosting approach, the model predicted sample time to positivity (Ct value) for the testing dataset with a mean square error value on the same order as the training dataset value. While these early results are promising, we will continue to iterate and train the ML models as we collect more experimental data in hopes of continuously improving the prediction accuracy.

A biophysics- and biochemistry-informed ML model that can predict the effect of nucleotide mutations on primer-based assays would greatly improve the speed at which we can develop and adapt detection technologies for emerging, evolving, and engineered biothreats.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.