GENETIC ENGINEERING TECHNOLOGIES AND DETECTION OF GENE EDITING

Use Of A Biological Decompiler To Predict Engineered Phenotype From Genomic Sequence Data

Nicholas Roehner Raytheon BBN Technologies **Aaron Adler** Raytheon BBN Technologies **Brian Basnight** Raytheon BBN Technologies **John Grothendieck** Raytheon BBN Technologies **Tyler Marshall** Raytheon BBN Technologies **Helen Scott** Raytheon BBN Technologies **Allison Taggart** Raytheon BBN Technologies **Benjamin Toll** Raytheon BBN Technologies **Dan Wyschogrod** Raytheon BBN Technologies **Fusun Yaman** Raytheon BBN Technologies

In early warning or rapid response scenarios, forming a comprehensive risk assessment is essential to maintaining force readiness. Therefore, knowledge of the function and origin of an organism is critically impactful to this effort, much more so than sequence identification alone. At present, state-of-art approaches to predict an engineered organism's phenotype based on its genotype can require days to weeks of subject matter expert (SME) time to evaluate tens of samples. In addition, these approaches often require multi-omic data streams that can cost hundreds to thousands of dollars per sample, which limits their application to high-throughput screening. By enhancing phenotype prediction from DNA sequence data, we can constrain this cumbersome downstream work, leading to faster turnaround times. The development of a biological "decompiler" will inform early countermeasure development and integrate with multi-omic approaches to enhance the integrated layered defense mission.

As part of the IARPA Finding Engineering-Linked Indicators (FELIX) program, we demonstrated that it is possible to combine evidence from anomaly-based (is this sequence unnatural?) and signature-based (is this sequence engineered?) approaches to automatically detect DNA sequence inserts with 90% sensitivity and 95% specificity. During FELIX, we focused on aligning sequences to databases of vectors and other sequences indicative of genetic engineering instead of determining whether these sequences contained features that had known interactions or were organized into known design motifs. Now, we hypothesize that leveraging this domain knowledge would not only enable prediction of engineered phenotype from genotype, but also enhance the detection of subtler engineering signatures, such as native sequence re-inserts, deletions, and small edits, by account for their context as part of an engineering design.

We envision a biological decompiler that combines "bottom-up" and "top-down" approaches to infer possible design specifications from input DNA sequences (see Figure 1). Rather than rely solely on curated lists of threat sequences, our bottom-up approach infers engineered biological networks based on rules for composing interactions between the organism's engineered and natural sequence features. This approach makes it possible to predict phenotypes resulting from novel combinations of known sequence features and to identify how an engineered organism's phenotype may differ from a natural phenotype. By contrast, our top-down approach uses grammar-based models to infer design specifications based on the structural layout of putative engineered sequence features rather than their known interactions. This approach enables prediction of behavior for engineered organisms that include sequence features that have no known interactions, but are organized into known design motifs. Ultimately, recognizing that a sequence is similar to a known threat is no longer sufficient - we must be able to predict threat level and mechanism by fusing bottom-up and top-down engineering context, and in doing so inform threat characterization to maintain force readiness while enabling development of effective countermeasures.

Figure 1. In this example of biological decompilation, a sensor design is inferred “bottom-up” from interactions between known DNA features and “top-down” based on the structural layout of these features into a known sensor motif. Finally, these specifications are fused to provide context for threat characterization.

This work was supported by the IARPA Finding Engineering- Linked Indicators (FELIX) award HR0011-15-C-0084. This document does not contain technology or technical data controlled under either U.S. International Traffic in Arms Regulation or U.S. Export Administration Regulations.