Scientists show benefits of bioinformatics with PlasmidHawk tool

February 26 2021, by Mike Williams

Credit: Unsplash/CC0 Public Domain

Tracking the origin of synthetic genetic code has never been simple, but it can be done through bioinformatic or, increasingly, deep learning computational approaches.

Though the latter gets the lion's share of attention, new research by

computer scientist Todd Treangen of Rice University's Brown School of Engineering is focused on whether sequence alignment and pan-genome-based methods can outperform recent deep learning approaches in this area.

"This is, in a sense, against the grain given that deep learning approaches have recently outperformed traditional approaches, such as BLAST," he said. "My goal with this study is to start a conversation about how to combine the expertise of both domains to achieve further improvements for this important computational challenge."

Treangen, who specializes in developing computational solutions for biosecurity and microbial forensics applications, and his team at Rice have introduced PlasmidHawk, a bioinformatics approach that analyzes DNA sequences to help identify the source of engineered plasmids of interest.

"We show that a sequence alignment-based approach can outperform a convolutional neural network (CNN) deep learning method for the specific task of lab-of-origin prediction," he said.

The researchers led by Treangen and lead author Qi Wang, a Rice graduate student, reported their results in an open-access paper in *Nature Communications*.

The open-source software is available here: gitlab.com/treangenlab/plasmidhawk.

The program may be useful not only for tracking potentially harmful engineered sequences but also for protecting intellectual property.

"The goal is either to help protect intellectual property rights of the contributors of the sequences or help trace the origin of a synthetic

sequence if something bad does happen," Treangen said.

Treangen noted a recent high-profile paper describing a recurrent neural network (RNN) deep learning technique to trace the originating lab of a sequence. That method achieved 70% accuracy in predicting the single lab of origin. "Despite this important advance over the previous deep learning approach, PlasmidHawk offers improved performance over both methods," he said.

The Rice program directly aligns unknown strings of code from genome data sets and matches them to pan-genomic regions that are common or unique to synthetic biology research labs

"To predict the lab-of-origin, PlasmidHawk scores each lab based on matching regions between an unclassified sequence and the plasmid pan-genome, and then assigns the unknown sequence to a lab with the minimum score," Wang said.

In the new study, using the same dataset as one of the deep learning experiments, the researchers reported the successful prediction of "unknown sequences' depositing labs" 76% of the time. They found that 85% of the time the correct lab was in the top 10 candidates.

Unlike the deep learning approaches, they said PlasmidHawk requires reduced pre-processing of data and does not need retraining when adding new sequences to an existing project. It also differs by offering a detailed explanation for its lab-of-origin predictions in contrast to the previous deep learning approaches.

"The goal is to fill your computational toolbox with as many tools as possible," said co-author Ryan Leo Elworth, a postdoctoral researcher at Rice. "Ultimately, I believe the best results will combine machine learning, more traditional computational techniques and a deep

understanding of the specific biological problem you are tackling."

Rice graduate students Bryce Kille and Tian Rui Liu are co-authors of the paper. Treangen is an assistant professor of computer science.

The research was supported by the National Institutes of Health via the National Institute for Neurological Disorders and Stroke, the Office of the Director of National Intelligence and the Army Research Office. Addgene provided access to the DNA sequences of the deposited plasmids.

More information: Qi Wang et al. PlasmidHawk improves lab of origin prediction of engineered plasmids using sequence alignment, *Nature Communications* (2021). DOI: 10.1038/s41467-021-21180-w

Ethan C. Alley et al. A machine learning toolkit for genetic engineering attribution to facilitate biosecurity, *Nature Communications* (2020). DOI: 10.1038/s41467-020-19612-0

Provided by Rice University