AI-POWERED DIAGNOSTICS

A Sequence-based Pathogen-agnostic Detection/diagnostics Solution

Patrick Chain Los Alamos National Laboratory **Po-E Li** Los Alamos National Laboratory

While the detection of known pathogens is performed routinely and the addition of other targets can be rapidly implemented, the identification of any pathogen, and in particular the ability to identify putative novel or emerging pathogens is a problem that has not yet been fully solved. This work's prime objective is to develop a computational suite of tools designed to identify any pathogen and provide the evidence underlying this characterization. Our project entitled SPADES for a Sequence-based Pathogen-Agnostic DEtection Solution utilizes at its core the unique genomic signatures of any and all microbial lineages, generated using internally developed algorithms at LANL that perform all by all k-mer analysis. The unique genomic signatures enable highly robust and specific detection of genomic signatures in assembled or raw sequence data, across a wide range of pathogens. Our work further refines this approach by integrating traditional bioinformatics with machine learning methods for a faster, and more comprehensive identification of the putative pathogen and any signatures of antimicrobial resistance, virulence factors, or signatures of genetic manipulation. In incorporating large language models (LLMs) and training on key genetic signatures and associated pathogen labels, the identified putative pathogen will also have functional predictions.

We are working on constructing a curated database of unique organismal signatures, along with a full-fledged cross Kingdom taxonomic classification tool. This database leverages the GTDB taxonomy, as well as average nucleotide identity (ANI) and genome completeness metrics, to ensure a broad, high-quality selection of genomes. We aim to enhance compatibility with error-prone long-read data and assembled genomes and contigs to extend SPADES' applications. Integration with the EDGE Bioinformatics platform will provide accessibility for DTRA, its partners and other collaborators. Thorough testing will involve synthetic sequence datasets and mock metagenome communities. SPADES will thus offer an easy to use solution for the detection and analysis of pathogens, even within complex biological samples.

This work was funded by the Defense Threat Reduction Agency under CB11366.