INNOVATING CROSS-DOMAIN SOLUTIONS TO DETECT EMERGING BIOLOGICAL THREATS

BAT-NET : Bioinformatics Analysis Tool For Network Evaluation Of Threats

Megan Howard Battelle **Katie Liszewski** Battelle **Stuart Lambert** Battelle

Surveillance and tracking the risk posed by emerging and novel biothreats requires high-throughput assessment of large, unfiltered data sets and the ability to infer risk from unknown sequences. While advancements in data harmonization and sequence exploration has been made, these heavily rely on biological knowledge that is time intensive to obtain. While informatic approaches to characterize novel agents exist, utility diminishes as the degree of sequence novelty increases. BAT-NET aims to close this gap using computational linguistics, graph theory, network visualization, and data analysis, within an intuitive visualization interface. This is accomplished by linking pairwise analysis, bi-directional computational linguistics, and machine learning with graphical visualization. BAT-NET produces a matrix of pairwise tokenized sequence comparisons visualized as a graph of sequence relationships. This is advantageous over conventional approaches as it allows visualization of arbitrary, rather than just parent-child relationships. BAT-NET was developed through a marriage of virology and mathematics, leveraging tokenization, generative adversarial networks, and graphical neural networks to cluster sequences without knowledge of biological function. Testing used existing data sets, but additional data sets combined with SME-derived heuristics will explore predictive performance beyond current biological knowledge and sequence space. While developed for genomics data sets, the underlying principles are extensible across additional complex data sets (including multi-omics data). When applied to the body of all coronavirus sequences, BAT-NET identified specific notes representing a single accession (or sequence) from the larger data set; these notes connected sequences across differing host species. This, in combination with pairwise analysis, identified anomalous sequences. While several algorithms were tested, determining the set distance between closely related sequences enabled rapid analysis, and scaled better than the length of the genome. From a biological perspective, a higher risk sequence for host species jumping (higher pandemic risk) will align more ‘in between’ two host species, rather than clustering with one or the other. To analyze that risk, BAT-NET trained customized tokenizers on data exemplifying key traits (e.g., bat or human viruses) and identified vocabulary output aligned with key attributes (host species). In one example, tokenizer output trained with U.S.-SARS-CoV-2 samples compared with output from bat-SARS-CoV-2-like coronaviruses quickly identified a single bat coronavirus that shared longer and more ‘meaningful’ vocabulary with U.S. SARS-CoV-2 than other bat-SARS-CoV-2 coronaviruses. These results can identify the ‘closeness’ of a sequence to known pandemic viruses, enabling categorization (and data-driven assessment) of sequences across the continuum of pandemic risk. BAT-NET intends to inform decision makers and military commanders of risks and threats to forces. BAT-NET can be embedded into a high side server for widest distribution used through a web API or on stand-alone networks for more sensitive or geographically distributed use cases. Using BAT-NET analytics and visualization, specialists and regional combatant commanders can monitor incoming sequence data across their areas of responsibilities, enable translation of sequence data to risk assessment in the field and allow for resource deployment and allocation in response to data driven risk assessment.

We would like to acknowledge numerous colleagues at Battelle who supported this effort including Robert Murdoch, Danny Dervish, Adam Ronk, Rachel Hardison, Jacob Beaver, Rebekah Cross, Alex Mang, and numerous others who supported the development of the graphical user interface.