INNOVATIVE APPROACHES TO ELUCIDATE OPTIMAL DEPLOYMENT OF CB SENSING ASSETS

Seeing The Trees For The Forest -- Identifying Trap-door Genomics Records Ahead Of Your Bioinformatics Workflow

Joseph Russell MRIGlobal

As the volume and complexity of genomic data increase, the integrity of data inputs becomes a pivotal concern for bioinformatics workflows, which are central to threat detection in contested biological threat environments. This study aims to address the issue of mislabeled or inaccurate genomic records, which can severely impact the operational readiness and accuracy of chemical-biological (CB) threat detection systems. By ensuring the reliability of data before integration into bioinformatics analyses, we enhance both preventative and responsive capabilities against CB threats. Incorrect genomic data can lead to flawed conclusions, affecting everything from medical diagnostics and and biosurveillance to countermeasure development. The methods we discuss here have implications for data integrity across biological scientific disciplines, demonstrating the interconnected nature of data validation and threat mitigation. Our approach utilizes a suite of computational tools that employ algorithms like sequence clustering based on Jaccard similarity and hierarchical clustering to identify discrepancies in data entries. Filters for sequence quality and metadata accuracy are applied to ensure robustness in data vetting processes. Early implementations of these tools in real-world databases have corrected significant numbers of erroneous records, thus preventing the propagation of these errors in downstream analyses. These corrections have substantially improved the reliability of data used in critical CB threat analysis and bioinformatics response systems. The effective identification and correction of erroneous genomic records significantly enhance the accuracy and reliability of subsequent analyses in bioinformatics workflows, crucial for accurate CB threat detection and remediation. By improving the foundational accuracy of genomic data, our tools directly support DTRA's mission to maintain and enhance warfighter readiness against biological threats. The integration of our tools into deployable bioinformatics workflows ensures that all levels of decision-making and response planning are based on accurate and reliable data, thus enhancing the operational effectiveness of the Joint Force in CB environments. The development and implementation of robust genomic data validation tools are essential for maintaining the integrity of bioinformatics databases used in forward-operating biological threat detection, as well as reach-back analyses. Our methods improve the efficiency of data processing and ensure that strategic and operational decisions are based on scientifically sound and accurate information, providing a critical enhancement to national and international biosecurity.