Biological Malware Detector

Muntaha Samad Raytheon BBN muntaha.samad@rtx.com Dan Wyschogrod Raytheon BBN dan.wyschogrod@rtx.com Jacob Beal Raytheon BBN jakebeal@ieee.org

Motivation

Companies around the world have been synthesizing nucleic acids for the past several decades. With advances in recent years, vendors can produce long, low-cost, high-fidelity custom nucleic acid sequences. While this capability has helped advance biotechnology research in many positive ways, it has also created the potential for bad actors to attempt to synthesize hazardous sequences. To combat this issue, in 2009 a group of nucleic acid synthesis vendors founded the International Gene Synthesis Consortium (IGSC) to create guidelines for screening nucleic acid synthesis orders [1]. A similar set of voluntary guidelines was introduced by the US HHS in 2010 [2] The IGSC and HHS guidelines prescribe a common protocol for screening DNA sequences and customers, while promoting the beneficial use of nucleic acid synthesis, and have been adopted by most nucleic acid synthesis providers to screen orders for the presence of sequences of concern (SoCs).

While the development of screening guidelines was an important step towards increased biosecurity, exclusively relying on nucleic acid synthesis vendors introduces several problems. First, the primary mission of nucleic acid synthesis companies is not biosecurity, even though they are serving as a primary line of defense against bioterrorism and carelessness. This can leave these organizations under-resourced and vulnerable to methods for SoC concealment such as assembly from oligos or protein editing. Additionally, sequence screening happens only after customers have placed an order for biological materials, thereby leaving threat detection to latest possible point of intervention before a physical SoC is acquired. Finally, many nucleic acid vendors are using local alignment tools, such as BLAST against public reference databases, to identify potentially dangerous orders which can be slow (minutes per kilobase), computationally expensive (terabyte-scale databases, cluster-scale computation) [3], and prone to errors [4].

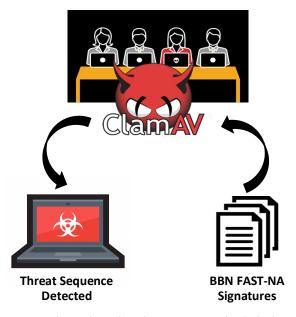


Figure 1: The Biological Malware Detector loads biological signatures for sequences of concern into standard signaturebased malware detection software to search the digital activities of individuals or organizations for biological threats.

Proposed Improvement: Biological Malware Detector

To address these shortcomings, we propose the Biological Malware Detector (BMD) a software tool that scans electronic systems for the sequences of biological threats in the same way anti-virus software currently scan systems for cyber-threats. Following the architecture shown in Figure 1, BMD combines biological threat signatures from BBN's FAST-NA Scanner [5] with ClamAV, a free and open-source host intrusion detection system to directly scan the digital activities of individuals or organizations for biological threats. BMD runs on a host device in a similar manner to anti-virus software except that instead of scanning for computer viruses, BMD scans for SoC's.

BMD has several significant use cases, each with substantial implications for enhancing biosecurity. First, BMD can serve as a precautionary measure to fortify facility biosecurity. For instance, the biosafety personnel of a laboratory could employ BMD on company laptops to monitor and detect insider threats or bio-error, by ensuring that all digital activities involving biological agents are in line with expectations. This use case would take some of the burden off nucleic acid synthesis vendors by catching bad actors before they place orders and before they have a chance to take steps towards concealment. Another use case made possible by BMD is targeted actor forensics, which would allow biosecurity and intelligence organizations to make preliminary assessments of the biological threat capabilities and intentions of persons and organizations. For instance, a security organization could use BMD to investigate potential malicious actors by obtaining their computers and directly examining their digital activities for SoC's. This sort of targeted forensics was not previously possible with BLAST-based methods because they are slow and computationally intensive, making a scan through of all the digital resources of an individual or organization infeasible.

Methods

The SoC signatures that we have tested in BMD are Raytheon BBN proprietary FAST-NA Scanner signatures, though signatures developed by other means could potentially be used as well. The FAST-NA Scanner library of signatures is comprehensive, including all potentially dangerous sequences listed by the IGSC, the United States Commerce Control List, as well as the Australia Group Control List. The FAST-NA Scanner signatures have undergone stringent tuning to exclusively identify SoCs. To ensure their accuracy, these signatures are validated against curated datasets from the National Center for Biotechnology Information (NCBI), encompassing all taxa covered by the signatures, as well as closely related nonthreatening sibling taxa. For instance, with this tuning process FAST-NA signatures can be used to correctly identify Bacillus anthracis as a significant threat while accurately classifying its close relative Bacillus cereus as a non-threat.

The anti-virus software utilized in BMD is ClamAV, a free and open-source signature-based detection system designed for detecting viruses and malware. ClamAV allows users to upload unique signatures and schedule scans to identify instances of the signatures within the digital resources stored on a device.

BMD combines the biological threat signatures from BBN's FAST-NA Scanner and ClamAV's signature-based detection to directly search the digital activities of individuals or organizations for biological threats.

BMD Deployment

BMD is designed to exclusively flag biological threats by leveraging finely tuned signatures. In a scan of an entire laptop (~20 GB) seeded with several files known to contain threat material, there were 0 false positives, 0 false negatives, and 10 true positives. Since nothing is reported other than signature matches, his low false positive rate can assure users that their privacy, particularly with respect to access to digital resources by lab administrators, will not be compromised by BMD flagging harmless files as potential threats. Additionally, BMD's computational load is low, only utilizing ~80% of a single CPU during scan. With this low computational load and fast scan time (Figure 2), use of BMD should not impede normal workflow for individuals or organizations.



Figure 2: Plot comparing the amount of data scanned vs the time it takes BMD to scan it.

Conclusion

The development of BMD is the beginning of a new era in digital biodefense, empowering biosafety personnel in laboratories to proactively address insider threats and enabling intelligence organizations to directly analyze digital activities for SoCs. By implementing BMD, early detection of intentions related to hazardous agents can become a reality, significantly mitigating the risks of biological accidents and bio-terrorism.

REFERENCES

- "Home: International Gene Synthesis Consortium." International Gene Synthesis Consortium | The Promotion of Biosecurity, 11 July 2023, genesynthesisconsortium.org/.
- HHS (Department of Health and Human Services). (2010) Screening Framework Guidance for Providers of Synthetic Double-Stranded DNA. Federal Register 75, 62820. October 13.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
- Beal J, Clore A, Manthey J. Studying pathogens degrades BLASTbased pathogen identification. Sci Rep. 2023 Apr 3;13(1):5390. doi: 10.1038/s41598-023-32481-z. PMID: 37012314; PMCID: PMC10068195.
- Dan Wyschogrod, et al., Adapting Malware Detection to DNA Screening, IWBDA, October, 2022.