EasyChair Preprint

№ 9045

Adapting Malware Detection to DNA Screening

Dan Wyschogrod, Jeff Manthey, Tom Mitchell, Steven Murphy, Adam Clore and Jacob Beal

October 11, 2022

Adapting Malware Detection to DNA Screening

Dan Wyschogrod[1], **Jeff Manthey**[2], **Tom Mitchell**[1], **Steven Murphy**[1], **Adam Clore**[2], **Jacob Beal**[1]
[1]Raytheon BBN, Cambridge, MA USA, [2]Integrated DNA Technologies, Coralville, IA, USA
{dan.wyschogrod,tom.mitchell,steven.t.murphy}@raytheon.com,{aclore,jmanthey}@idtdna.com,jakebeal@ieee.org

1 MOTIVATION

As DNA synthesis becomes cheaper and more accessible, there is a corresponding increase in opportunities for synthesis of dangerous pathogenic sequences by either malicious or careless actors [2–4, 6, 8]. To mitigate this threat, major DNA synthesis providers screen sequence orders for pathogenic content, following guidance from the US Department of Health and Human Services [9] and the International Genome Synthesis Consortium (IGSC) [7].

Current methods for screening, however, have been unable to scale sufficiently to keep up. The current dominant method for screening is to evaluate sequence homology, using BLAST (or similar) to test if the sequence's best alignment is with a controlled pathogenic organism [2, 5, 8]. This approach produces a high rate of false positives, estimated at more than 4% from a survey of IGSC member companies [2], worsened by the fact that these methods generally search for all genes in an organism, including harmless "housekeeping" genes and others that have no functional relationship to pathogenesis. Moreover, the rate of false positives increases markedly as sequence length shortens [6]. Due to the cost of resolving false positives, synthesis providers thus typically only screen dsDNA sequences that are at least 200 bp long and do not screen oligonucleotides at all [2, 5].

We hypothesized that these challenges could be addressed by adapting methods for detection of malware in network traffic, which faces even greater challenges of scale. To this end, we adapted the Framework for Autogenerated Signature Technology (FAST) signature extraction method [10] for use with nucleic acid sequences, producing the FAST for Nucleic Acids (FAST-NA) method for DNA screening. Our resulting implementation of FAST-NA is able to detect DNA sequences far faster than BLAST-based methods, and with equivalent sensitivity and significantly improved specificity, even while reducing the minimum scanning window from 200bp to 50bp.

2 DEVELOPMENT OF FAST-NA

FAST begins by breaking collections of target and contrast material into small "signature" fragments. FAST stores the contrast signatures in a Bloom filter [1], a highly efficient data structure for testing set membership. The Bloom filter is then used to remove all target signatures that match any contrast signature, leaving only signatures that are diagnostic of threats. This proves highly effective for malware detection: even though polymorphic malware constantly mutates

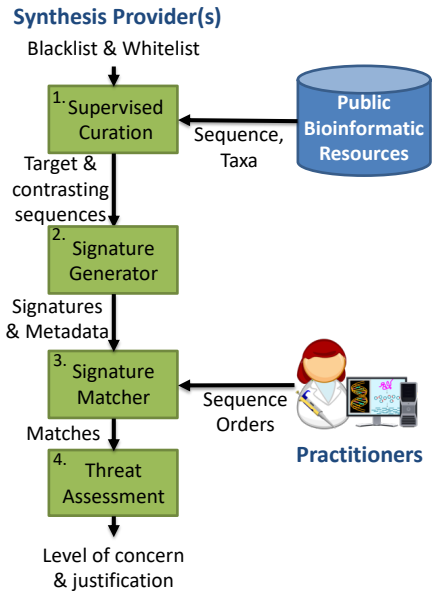
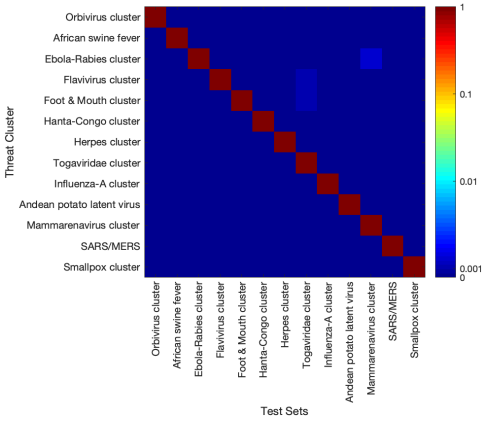


Figure 1: FAST-NA architecture: diagnostic signatures are identified by comparing target sequences to contrasting material, then applying these signatures in a matcher that scans sequence orders to assess their threat content.

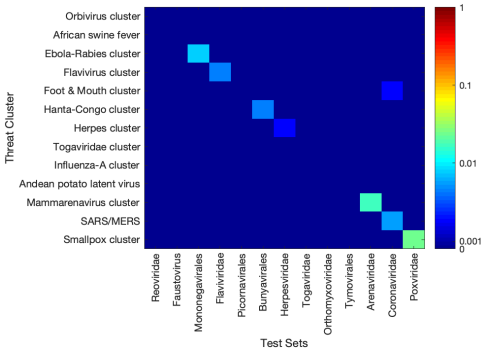
itself to try to evade detection, there are generally still some conserved sequences required for its function, which FAST is able to identify. Matching software using these signatures can then identify malware extremely rapidly and with high sensitivity and specificity.

Adapting this method for FAST-NA (Figure 1), we use nucleic acid and protein sequences from public databases such as NCBI as the source material, taking the target material from clusters of threat taxa to be detected and the contrasting material from other taxa that are closely related but not controlled. For example, SARS and MERS are in the coronavirus threat cluster, while the more benign human coronaviruses 229E and NL63 are in the coronavirus contrast collection. For signatures, we use k-mers, ranging from 26-42 base pairs for nucleic acids and 14-20 residues for amino acids.

Just as with malware, this process identifies signatures for conserved sequences defining the nature of a biological threat. These signatures, along with metadata on their origins, can then be given to a matcher that scans sequence orders to assess their threat content. With appropriate tuning and curation, this produces a signature collection that is both highly sensitive and highly specific.



(a) Threat Identification



(b) False Positives

Figure 2: Sensitivity and specificity of FAST-NA signatures for controlled viral pathogens: (a) probability of correct identification of threat sequences, (b) probability of false positives for closely related non-controlled sequences.

Figure 2 shows an example of FAST-NA performance, in this case for the set of all viral threats in the IGSC Regulated Pathogen Database. When comparing all 50+ bp viral threat sequences from NCBI and from close contrasting taxa, we find the signatures are highly sensitive, producing no false negatives. They are also highly specific: mean per-taxa likelihood that a threat is multiply identified is 0.039%, while the mean per-taxa likelihood of a false positive is 0.55%. Other kingdoms are not as clean as viruses—particularly the bacteria, which are highly prone to horizontal transfer—but the average all-threat rate for multiple identification and for false positives are both less than 2%, far lower than the typical 4% rate for BLAST-based screening despite the much-reduced screening window. Moreover, because it focuses only on diagnostic signatures, FAST-NA is able to scan >10 kilobases/second (orders of magnitude faster than BLAST) and with far less required computing resources.

The distribution of sequences in commercial synthesis orders is, of course, quite different than that found in sequences in NCBI. We have found, however, that the performance is maintained when applied to synthesis orders. A commercialized version of the system, named FAST-NA Scanner, is now deployed at IDT, and is seeing similar or better results when used against live customer data.

3 APPLICATIONS AND FUTURE DIRECTIONS

At present, the primary application of FAST-NA remains DNA synthesis order screening, with FAST-NA Scanner available from BBN as a commercial software product. In addition to the improvements in false positive rate, the high speed and low computational cost of FAST-NA can also enable other workflows that are impractical with BLAST-based scanning, such as online pre-order screening, secure on-site screening (e.g., in a benchtop synthesizer), and combinatorial screening of oligo assemblies. Finally, beyond synthesis order screening, we aim to further develop FAST-NA for other types of biosecurity applications, such as interpretation of sequencing data, incorporation of biosafety and biosecurity considerations into design tools, and threat scanning in information systems and laboratory management processes.

ACKNOWLEDGEMENTS

This work was partially supported by IARPA contract 2018-17110300002 and ARO grant W911NF-17-2-0092. Views and conclusions are of the authors and should not be interpreted as representing official policies, either expressed or implied, of ARO or the U.S. Government. Contains no technology or technical data controlled under U.S. ITAR or EAR.

REFERENCES

[1] Bloom, B. H. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM 13*, 7 (1970), 422–426.

[2] Carter, S. R., and Friedman, R. M. Dna synthesis and biosecurity: lessons learned and options for the future. *JCVI* (2015).

[3] Carter, S. R., and Warner, C. M. Trends in synthetic biology applications, tools, industry, and oversight and their security implications. *Health security 16*, 5 (2018), 320–333.

[4] DiEuliis, D., Berger, K., and Gronvall, G. Biosecurity implications for the synthesis of horsepox, an orthopoxvirus. *Health security 15*, 6 (2017), 629–637.

[5] Diggans, J., and Leproust, E. Next steps for access to safe, secure dna synthesis. *Frontiers in bioeng. and biotech. 7* (2019), 86.

[6] Garfinkel, M. S., Endy, D., Epstein, G. L., and Friedman, R. M. Synthetic genomics: options for governance. *Industrial Biotechnology 3*, 4 (2007), 333–365.

[7] IGSC (International Gene Synthesis Consortium). Harmonized screening protocol v2.0. Available at https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf, November 2017. Accessed October 20, 2020.

[8] NASEM. *Biodefense in the age of synthetic biology*. National Academies Press, 2018.

[9] US Department of Health and Human Services. Screening framework guidance for providers of synthetic double-stranded dna. *Fed Regist 75*, 197 (2010), 62820–62832.

[10] Wyschogrod, D., and Dezso, J. False alarm reduction in automatic signature generation for zero-day attacks. In *2nd Cyberspace Research Workshop* (2009), p. 73.