BBN FAST-NA Final Report

Raytheon BBN Technologies,
Integrated DNA Technologies

Raytheon BBN Technologies

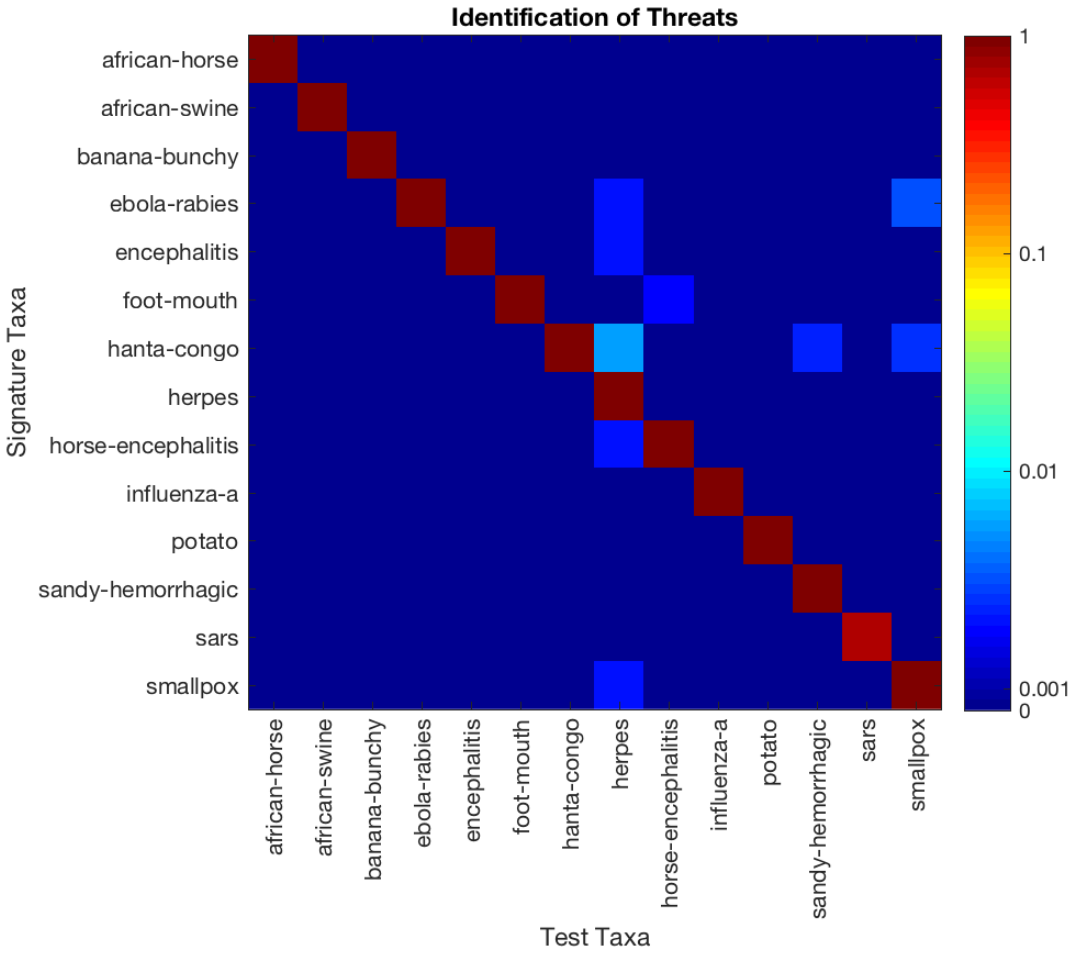
IDT INTEGRATED DNA TECHNOLOGIES

Applicability of Malware Signature Extraction to Nucleic Acid Screening

Applicability of FAST-NA to Nucleic Acid Screening

Jacob Beal, Dan Wyschogrod, Tom Mitchell, Susan Katz,
Jeff Manthey and Adam Clore

February 2019



Applicability of FAST-NA to Nucleic Acid Screening

Jacob Beal, Dan Wyschogrod, Tom Mitchell and Susan Katz

Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138

Jeff Manthey and Adam Clore

Integrated DNA Technologies
1710 Commercial Park
Coralville, Iowa 52241I

Final Report

Sponsored by:
Intelligence Advanced Research Projects Activity
Contract No. 2018–17110300002

Prepared for Intelligence Advanced Research Projects Activity
Maryland Square Research Park
5850 University Research Court, Riverdale Park, MD 20737

Abstract: We evaluate the effectiveness of applying the Framework for Autogenerated Signature Technology (FAST) signature extraction method to the detection of small nucleic acid sequences of concern in samples submitted for synthesis. This approach (FAST-NA) aims to address shortcomings in existing methods for detecting sequences of concern that are either maliciously or inadvertently included in nucleic acid synthesis orders. In particular, FAST-NA comparison of threat and contrasting data should reduce false positives without increasing false negatives by focusing detection on only those sequence fragments that are actually diagnostic of threat, rather than irrelevant portions shared with other organisms.

Our results indicate that the FAST malware screening technology can be effectively adapted for screening of viral nucleic acid sequences. It appears likely that FAST-NA can use publicly curated data to identify short sequences diagnostic of viral threat potential in a nucleic acid sequence, significantly reducing false positives in screening for viral threats without introducing false negatives. FAST-NA also appears extensible beyond the viral domain for screening against bacterial and eukaryotic threats as well. Finally, FAST-NA should be able support an effective CONOPS for biosecurity screening with a reasonable resource budget.

Given the potential for significant improvement over the current state of the art in nucleic acid synthesis screening, **in the interest of national security, we thus recommend funding further development of FAST-NA in support of transition into widespread industrial usage.**

Table of Contents

Figures and Tables ..... **iv**

1 Task Objectives **1**

2 Technical Problems **3**

3 General Methodology **4**

3.1 Development of FAST-NA 4
3.1.1 Mathematical Properties of FAST 4
3.1.2 Adaptation for Nucleic Acid Screening 5
3.2 Curation of Training and Test Data 7
3.3 Experimental Pipeline 10

4 Technical Results **13**

4.1 Identification of Diagnostic Sequences from Publicly Curated Data 13
4.2 Assessment and Modulation of FAST-NA Performance 15
4.2.1 Failure Modes of FAST-NA 18
4.2.2 Estimation of Appropriate Signature Length 20
4.2.3 Systematic Tuning of Parameters 23
4.3 Full-Scale Viral Application of FAST-NA 24
4.3.1 Signature Development for All Viral Threats 24
4.3.2 Time and Signature Scale for All Viral Threats 26
4.3.3 Cross-Taxa False Positives and Threat Misidentification 29
4.3.4 Detection of Threats in Very Short Sequences 33
4.3.5 Enhancing Threat Detection 34
4.4 Generalization to Bacterial and Eukaryotic Threats 38
4.5 Realism of CONOPS and Resource Requirements 39
4.5.1 Evaluation Against Realistic Sequence Distribution 39
4.5.2 Scaling of Resource Requirements 41
4.6 Opportunities for Generalization of FAST-NA 43
4.7 Recommendations for Control of Information Related to FAST-NA Technology 44

5 Summary and Discussion **46**

5.1 Progress Against Waypoints 46
5.2 Important Findings and Conclusions 46
5.3 Special Comments 46
5.4 Implications for Further Research 47
5.5 Commercial/Proprietary/Third-Party Material in Deliverables 47

References **48**

Figures and Tables

Figures

Figure 1. FAST-NA signature-based screening CONOPs. 1
Figure 2. Architecture of FAST-NA: white boxes are FAST-NA applications, blue cylinders are data collections, and wavy-bottom boxes are configuation files. 6
Figure 3. Example of output from `sig-diagrammer` tool. 7
Figure 4. Viral threat taxa form 14 clusters at the Order or Family level. 8
Figure 5. Distribution of viral threat information with respect to total NCBI viral sequence information. Values are based on total number of nucleotide records; for each cluster this includes both threat and contrasting. 9
Figure 6. Amount of sequences and base pairs for threat and contrasting data in each viral threat collection. 9
Figure 7. Architecture of k-fold cross-validation in experimental pipeline. 11
Figure 8. False positives and false negatives for FAST-NA applied to all clusters with 14bp signatures and all contrasting sequences. 14
Figure 9. False positives are power-law correlated with contrasting data size across all threat clusters. 14
Figure 10. Parameter surveys for five viral threat taxa: (a) H5N6 influenza, (b) classical swine fever virus (an encephalitis), (c) ebolavirus, (d) SARS-related coronavirus, and (e) banana-bunchy. Color indicates signature length, shading in hue from n=10 (orange) to n=36 (red). 16
Figure 11. Number of signatures generated in parameter surveys for five viral threat taxa: (a) H5N6 influenza, (b) classical swine fever virus (an encephalitis), (c) ebolavirus, (d) SARS-related coronavirus, and (e) banana-bunchy. Color indicates signature length, shading in hue from n=10 (orange) to n=36 (red). 17
Figure 12. Examples of FAST-NA failure modes: (a) Training on encephalitis with very short sequences (stars) has much greater false negatives (red) than eliminating very short sequences (circles), while false positives (blue) are essentially unaffected. (b) False negatives rise sharply as false positives fall when contrasting data is too close, as in this example of H5N6 influenza contrasted against all non-threat orthomyxoviridae including non-threat strains of influenza A. Color indicates signature length, sharing in hue from n=10 (orange) to n=36 (red). (c) Too much contrasting data can eliminate useful signatures, causing false negatives to rise sharply, as in this example with encephalitis. 19
Figure 13. Conservative estimation of random signature match rate for (a) pruning signatures with contrasting data, (b) matching a single 10^3 bp gene, (c) matching a 10^5 bp viral genome, or (d) matching a 10^8 bp eukaryotic chromosome. 21
Figure 14. Random signature match estimates can be applied to amino-acid signatures as well, finding shorter lengths are required due to the larger alphabet, as in this example for matching amino acid signatures at the 10^8 bp chromosome scale. 23
Figure 15. Example of landscape evaluation for protein-based signatures with the smallpox threat cluster, showing tunable response to signature length and contrasting data percentage. 24
Figure 16. Results of 10-fold cross-validation with tuned parameters for all viral threats. 25
Figure 17. (a) Number of signatures per cluster for full-scale viral threat detection, and (b) fraction of sequence used in signatures. 27
Figure 18. Distribution of number of threat sequences sharing each signature for selected threat clusters. 28
Figure 19. Distribution of overlap between signatures for selected threat clusters. 29

Figure 20. Signature coverage appears biologically correlated, as in this example of Ebola virus, which shows signature coverage (red) focused most heavily on the genes for replication and cell docking. 30
Figure 21. Rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures. 31
Figure 22. (a) No significant correlation is observed between signature count and false positive rate, but (b) errors go down with increased contrasting sequence data..... 32
Figure 23. Distribution of lengths for short (< 50 bp) sequences in viral threat and contrasting sequence data. 34
Figure 24. Nucleic acid screening with reverse complement and false-positive reduction: rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures. 35
Figure 25. Protein screening rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures. 37
Figure 26. Results of subset testing for a single small taxon of threat and contrast samples for a variety of eukaryotic ("abrin" through "snail") and bacterial ("argt-potato-ring-rot" through "typhus-spotted-fever") threat clusters. Results are highly variable, but indicate likely applicability of FAST-NA to these cellular threats as well. 38
Figure 27. Results of applying full-scale viral threat signatures to IDT customer-related sequences. Color indicates rate of detection, grey square indicates IDT data set had no matches to be tested..... 40
Figure 28. Time scaling for FAST-NA components. 42

Tables

Table 1. Limits of FAST for nucleic acid screening addressed in developing FAST-NA..... 5
Table 2. Contrasting taxa for each threat cluster 8
Table 3. Scale of threat taxa, including bacterial and eukaryotic threats and both nucleic acid and protein sequences. 10
Table 4. Expected interpretation of FAST-NA results based on comparison with IDT's current biosecurity screening system and human expert judgement..... 12

1 Task Objectives

The project "Applicability of Malware Signature Extraction to Nucleic Acid Screening" aimed to evaluate the effectiveness of applying the Framework for Autogenerated Signature Technology (FAST) signature extraction method [1, 2], developed by BBN for the detection of malware in network traffic, to the detection of small nucleic acid sequences of concern in samples submitted for synthesis. Detection of small sequences provides the advantages that a) small segments of pathogenic DNA inserted into otherwise benign sequences are more likely to be detected, b) sequence variations are less likely to lead to missed detection, and c) artificially engineered sequences of concern may be detected by their similarity to small natural segments of pathogenic DNA required to achieve the pathogenic functionality. Specifically, we aimed to evaluate: 1) efficacy and scalability of these techniques for viral pathogens, and 2) the likely potential for applicability to other classes of pathogens and toxins.

The concept of operations for the application of FAST to nucleic acid screening (which we refer to as FAST-NA), is shown in Figure 1 and consists of the following steps: (1) Blacklist and whitelist data is integrated with public bioinformatic resources to obtain large volumes of target and contrasting sequence data; (2) Sequences are compared to generate diagnostic signatures for threats; (3) Signatures are matched against sequence orders to find possible areas of concern; and (4) Matches are collated and assessed to determine

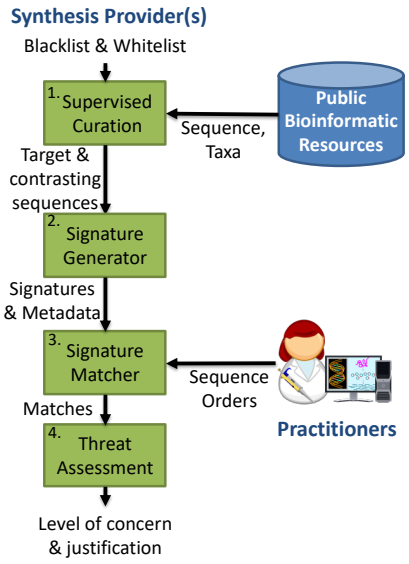


Figure 1. FAST-NA signature-based screening CONOPs.

threat level, justifying judgments using the metadata associated with matching signatures.

If successful, the FAST-NA application of this method to nucleic acid screening will address shortcomings in existing methods for detecting sequences of concern that are either maliciously or inadvertently included in nucleic acid synthesis orders. Comparison of threat and contrasting data should reduce false positives (compared to the current rate of $\sim$2%) without increasing false negatives, by focusing detection on only those sequence fragments that are actually diagnostic of threat, rather than irrelevant portions shared with other organisms.

The tasks executed in pursuit of this objective were:

1. Adapt FAST for sequence screening, including handling sequence data at large scale, organizing curated datasets, and providing appropriate software interfaces and experimental testbeds.

2. Evaluation of efficacy and scalability of FAST-NA for viral screening and comparison against current methods, including determining applicable taxa, scalability of threat and contrasting data, false positive and false negative rates vs. signature size and training corpus size, and computational cost of training and execution.

3. Evaluation of potential for generalization of FAST-NA to detection of threats from taxa other than viruses.

This report summarizes all progress against these tasks during the execution of this project.

2 Technical Problems

Through this investigation, we aimed to answer five core questions regarding the applicability of FAST malware signature extraction to nucleic acid screening:

- Are there short diagnostic DNA sequences that can distinguish threats and non-threats?
- Can the FAST-NA approach identify such diagnostic sequences reliably?
- Can effective training data be reasonably curated from available data sources?
- How confident can we be in the results of FAST-NA screening?
- Can FAST-NA support a realistic CONOPS with a reasonable resource budget?

In support of this investigation, we organized our technical effort around four main strands of work:

- adaptation of the FAST software in a new biology-focused FAST-NA implementation (Section 3.1),
- curation of training and test data from public data sources and IDT proprietary data sources (Section 3.2),
- construction and operation of an experimental pipeline for evaluation of FAST-NA (Section 3.3), and
- analysis of experimental results to answer core questions (Section 4).

To maximize efficiency and maintain integration across these thrusts, we made use of agile software engineering tools and methods, notably the GitLab repository manager and GitFlow development workflow. This combination provides source code control and test data management based on git, issue tracking for management of development progress, code review in support of effective development, and continuous integration and regression testing to ensure continuous functionality.

3 General Methodology

In this section, we discuss the methods taken for development of the new FAST-NA screening software, curation of training and test data, and development of our experimental pipeline.

3.1 Development of FAST-NA

The Framework for Auto-Generated Signature Technology (FAST) [1, 2] approach was originally designed for detection of malware in network traffic. Its use begins with network traffic of concern being forwarded by an anomaly detector (a separate network screening system). At the same time, contrasting benign network traffic is also accumulated from the network of interest. The Automatic Signature Generator (ASG) compares the two to find unique segments in the traffic of concern, which are then exported as signatures formatted for the SNORT network traffic filter.

3.1.1 Mathematical Properties of FAST

The efficacy of the FAST signature generation approach is primarily regulated by the nature and volume of contrasting data and the length of signatures. The effect of both signature length and contrasting data can be predicted from the tendency of network traffic to be semi-structured, comprised of a mix of highly structured information (e.g., packet headers or HTML protocol) and effectively random data (e.g., compressed or encrypted information). This results in a power-law distribution in the frequency of occurrence of sub-sequences. As nucleic acid sequences evince similar semi-structured patterns and appear to follow similar power-law distributions (e.g., [3, 4, 5]), it is reasonable to predict that the efficacy of nucleic acid screening will follow similar principles.

With respect to contrasting data, as the size n of the contrasting data set increases, the distribution assumption predicts that the rate of false positives should decrease following a power-law. As a baseline, one might expect a $1/n$ decrease, but in practice the slope may vary. If the volume of contrasting data is too high, however, random information may begin to remove critical signature information, causing the rate of false negatives to rise.

Signature length similarly needs to be neither too short nor too long. If signatures are too short, then random traffic will tend to contain many false positives, which cannot be removed without also creating many false neg-

Technique	Shortcomings
Embed DNA sequences in TCP/IP packets	Limits size of sample and doesn't allow for carrying metadata. Requires fictional TCP port information, protocols, etc.
Use STIX files for describing attacks	STIX is a cyber security based standard not designed for carrying biological information
Use pcap files for collecting benign traffic	Requires assembling contrasting DNA sequences into packets and adding fictitious time and other information
Use SNORT as matcher	Poor at carrying labeling and metadata information for matches, size limitations
Signature analysis	Primitive in FAST, performed mostly by hand

Table 1. Limits of FAST for nucleic acid screening addressed in developing FAST-NA

atives. If signatures are too long, however, then they will tend to contain irrelevant "flanking" information alongside the true sequence of concern, meaning the signature will only match in the exact same context and increasing the chance of false negatives.

3.1.2 Adaptation for Nucleic Acid Screening

In preliminary work for this project, the FAST software was used for a proof-of-concept distinguishing 1918 influenza from other strains by literally putting DNA sequences into network packets. There are a number of limitations to this approach, however, that make this inappropriate for general use for nucleic acid screening, as detailed in Table 1. Accordingly, we have developed FAST-NA, a new collection of tools based on the original FAST software but adapted for DNA screening.

FAST-NA is implemented in C++, using the original speed optimizations from FAST and adding more as needed. Rather than STIX and pcap files capturing network traffic, FAST-NA takes FASTA sequence files as its input. Biological metadata is associated with each signature and match: sequence offsets and (when available), sequence accession number and taxon ID. SNORT is replaced with a custom matcher for nucleic acid sequences, and new tools have been created for signature evaluation.

Our implementation of FAST-NA comprises six applications, linked together in the architecture shown in Figure 2. First, the `makebloom` application digests FASTA files of contrasting data into a Bloom filter [6] used for pruning

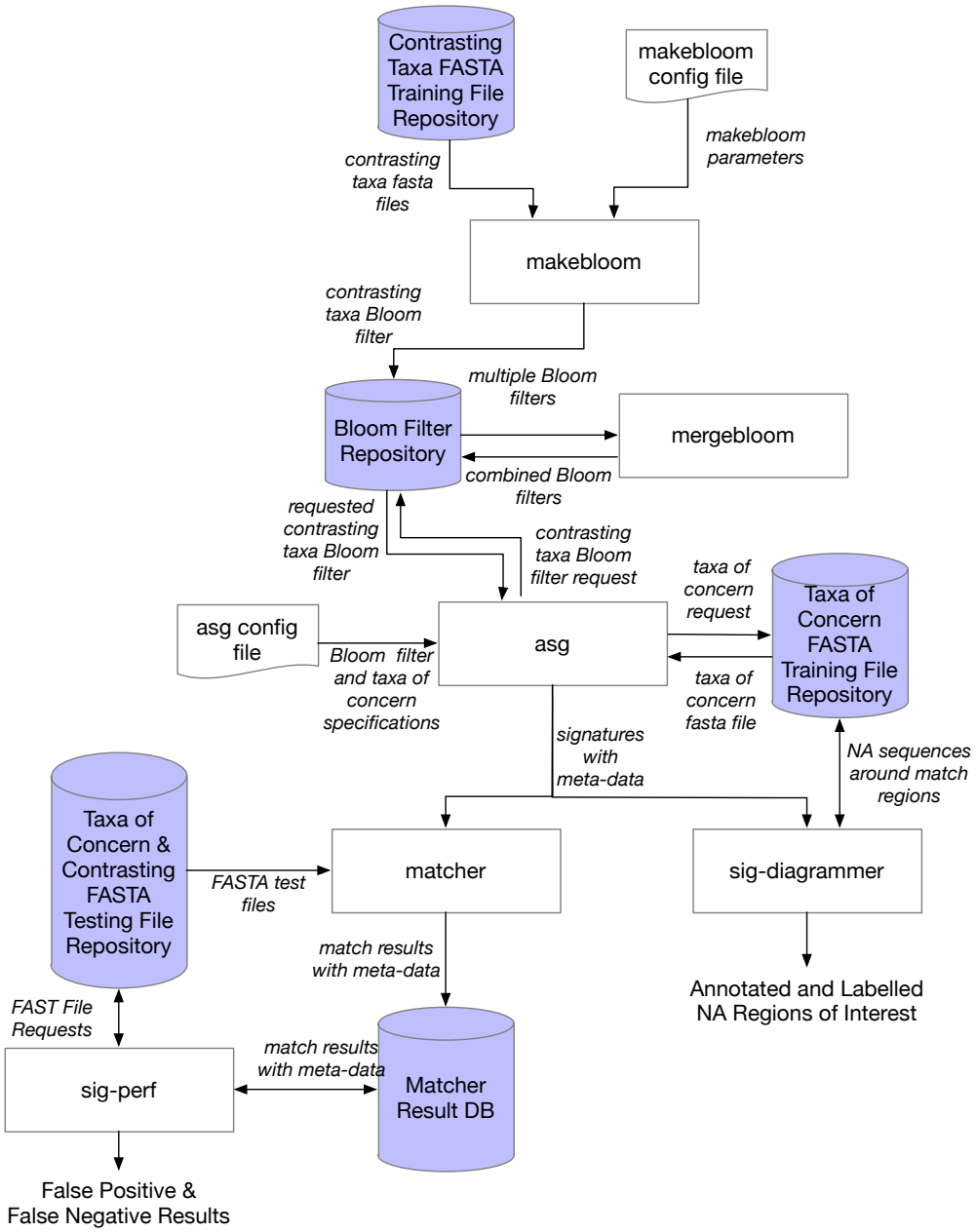


Figure 2. Architecture of FAST-NA: white boxes are FAST-NA applications, blue cylinders are data collections, and wavy-bottom boxes are configuation files.

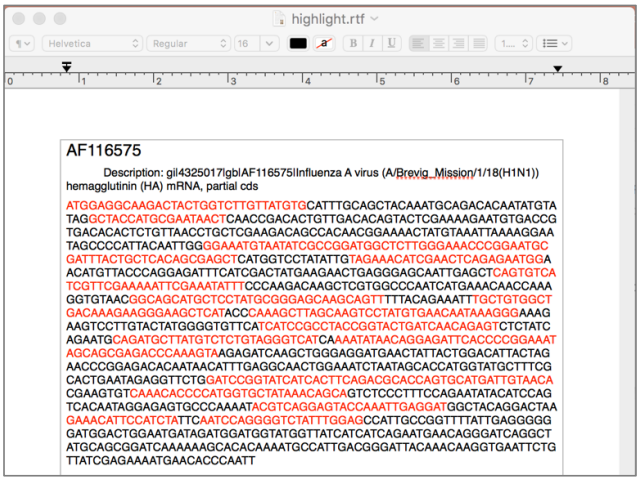


Figure 3. Example of output from `sig-diagrammer` tool.

potential signatures, and Bloom filters from multiple contrasting data sets can be joined using `mergebloom`. Automated signature generation is performed using the `asg` tool on samples of concern presented as FASTA files and a Bloom filter of contrasting samples. These signatures can be inspected using the `sig-diagrammer` tool, which provides information about signature coverage of samples of concern as well as origin of signatures in multiple samples—sample output from this tool is illustrated in Figure 3. Signatures are applied for threat detection using the `matcher` tool, which finds occurrences of signatures in unknown samples presented as FASTA files. Finally, the `sig-perf` tool evaluates matches to decide whether a sequence is a threat—though currently this is a trivial implementation where any match is considered a threat.

3.2 Curation of Training and Test Data

Our initial subject of evaluation was the applicability of these methods to detection of viral threats. As such, we have curated the collection of all viral threats currently screened for in IDT's biosecurity system into a collection of 14 threat sequence training sets, comprising approximately 334 thousand threat nucleic acid sequences and 1.1 million contrasting nucleic acid sequences. Collectively, these threat collections cover approximately half of all threat taxa currently screened for by IDT, though only ∼4% of all threat nucleotide records. The threat collections are assembled from public records retrieved from NCBI's GenBank using its E-Utilities web interface. These records also contain taxonomic information: NCBI's Taxonomy Database is organized by taxonomic rank (Kingdom to Species), and we find that the viral threat taxa form 14 clusters at the Order/Family level (Figure 4), from which we take all sequences from non-threat taxa as contrasting data (Ta-

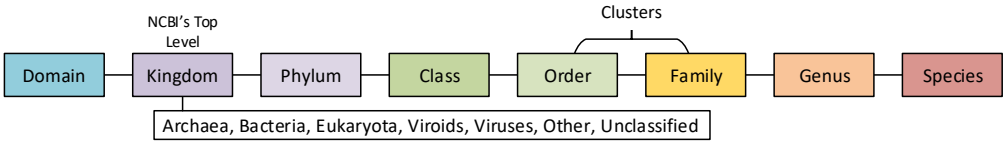


Figure 4. Viral threat taxa form 14 clusters at the Order or Family level.

Common Name	As of	Constrast		Alternate	
		TaxonId	Order/Family	TaxonId	Order/Family
Influenza A	3/2/2018	11308	Orthomyxoviridae		
Encephalitis	3/9/2018	11050	Flaviviridae		
Ebola_Rabies	4/2/2018	11157	Mononegavirales		
SARS	4/2/2018	11118	Coronaviridae		
SmallPox	4/2/2018	10240	Poxviridae		
Herpes	4/2/2018	10292	Herpesviridae		
Hanta_Congo	4/2/2018	1980410	Bunyavirales		
Foot_Mouth	4/2/2018	464095	Picornavirales		
AfricanHorse_Bluetongue	4/2/2018	10880	Reoviridae		
HorseEncephalitis	4/2/2018	11018	Togaviridae		
AfricanSwine	4/2/2018	137992	Asfarviridae	1477405	Faustovirus
SandyHemorrhagic	4/2/2018	11617	Arenaviridae		
BananaBunchy	4/2/2018	251095	Nanoviridae		
Potato	4/2/2018	675063	Tymovirales		

Table 2. Contrasting taxa for each threat cluster

ble 2), AfricanSwine being the exception as its Family level is also in the threat list. Figure 5 shows the distribution of threats with regards to the overall taxonomy and distribution of viral sequence information in GenBank, and Figure 6 shows the total amount of sequences and base pairs for each training cluster. Note that since threats are not hand-curated, there are a number of curation problems that can be encountered, particularly around misclassification of sequences, some of which are discussed in our technical results below.

In addition, we have curated a collection of de-identified customer-related sequence segments, to provide a more realistic distribution of sequences for CONOPS evaluation. These are sequences run through IDT's biosecurity screening during the period of July 2017 to March 2018. They comprise a collection of 69,835 "no hit" sequences that were automatically cleared by IDT, 15,879 "false positive" sequences (4,521 of which matched viral threats), and 1753 "true positive" sequences that had to be cleared manually (629 of which matched viral threats).

Beyond viral nucleic acid sequences, we have also curated collections of protein sequences, organized into the same taxonomic divisions. For evaluation of the generalizability of FAST-NA beyond viruses, to include bacterial and eukaryotic threats as well, we have further curated collections of threat and contrasting sequences, both nucleic acid and protein, for taxonomic groups in these threats as well. Figure 3 summarizes the division and size of threat

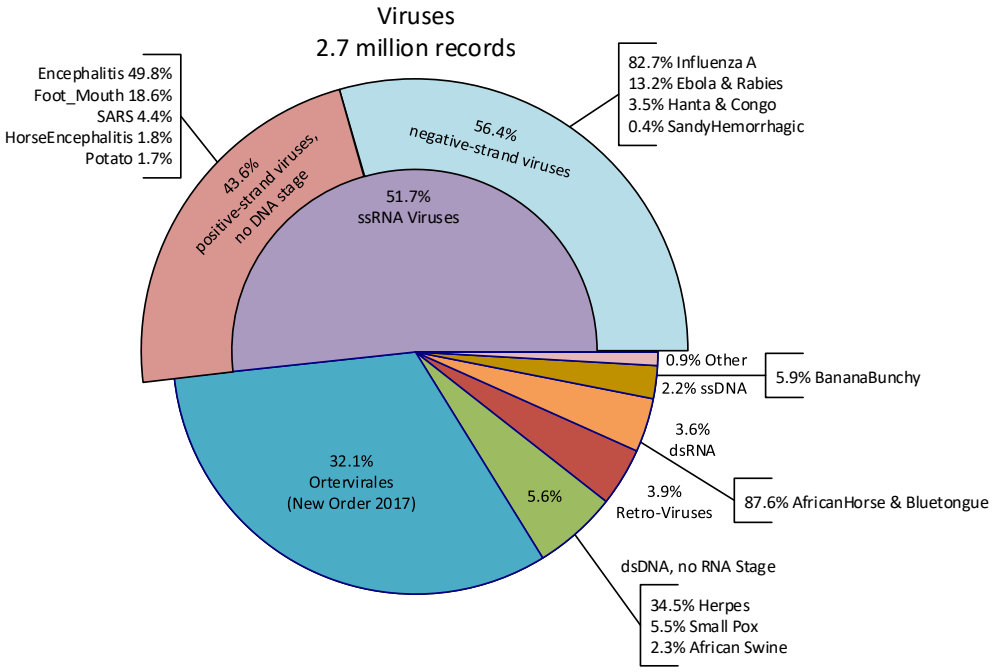
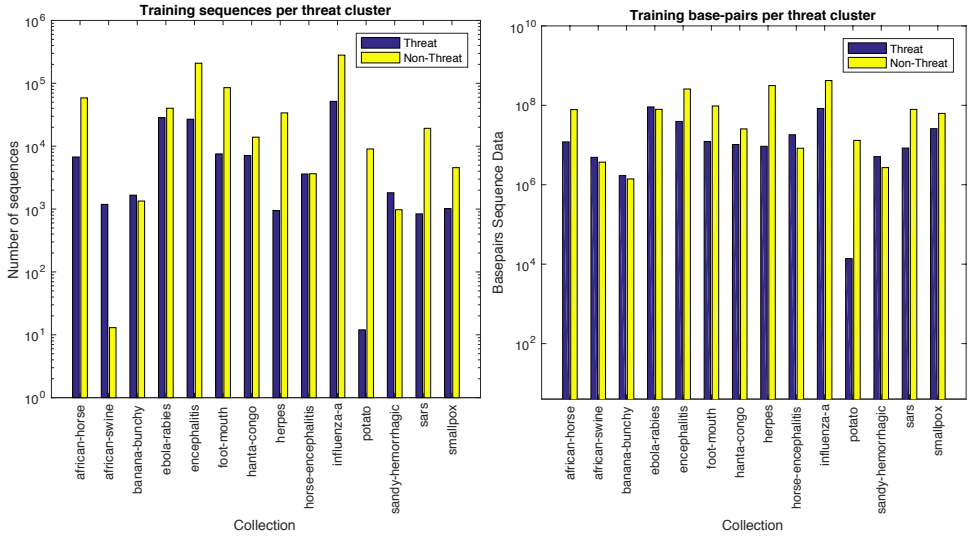


Figure 5. Distribution of viral threat information with respect to total NCBI viral sequence information. Values are based on total number of nucleotide records; for each cluster this includes both threat and contrasting.



(a) Training Sequences (b) Training Sequence Base Pairs

Figure 6. Amount of sequences and base pairs for threat and contrasting data in each viral threat collection.

		Nucleotide				Protein			
		Records	MegaBasepairs	% Records	% MegaBasepairs	Records	MegaBasepairs	% Records	% MegaBasepairs
Bacteria		4369578	159752.5	90.82%	98.15%	56513609	17221.6	98.78%	98.30%
	Bacillales	1564563	70599.4	35.81%	44.19%	23377741	6941.0	41.37%	40.30%
	Enterobacterales	1367983	39042.8	31.31%	24.44%	14253556	4270.5	25.22%	24.80%
	Xanthomonadales	458294	6907.2	10.49%	4.32%	1838688	593.8	3.25%	3.45%
	Burkholderiales	328706	16655.0	7.52%	10.43%	5813410	1924.5	10.29%	11.17%
	Clostridiales	241950	8639.0	5.54%	5.41%	4233128	1301.9	7.49%	7.56%
	Vibrionales	229321	8074.1	5.25%	5.05%	3244361	1020.2	5.74%	5.92%
	Legionellales	74221	4568.0	1.70%	2.86%	1578571	519.6	2.79%	3.02%
	Thiotrichales	41074	1146.3	0.94%	0.72%	450902	131.9	0.80%	0.77%
	Rhizobiales	34964	3550.4	0.80%	2.22%	1457832	435.0	2.58%	2.53%
	Chlamydiales	16919	203.4	0.39%	0.13%	96497	31.1	0.17%	0.18%
	Rickettsiales	5551	102.7	0.13%	0.06%	69556	20.1	0.12%	0.12%
	Mycoplasmatales	4068	78.5	0.09%	0.05%	40774	13.8	0.07%	0.08%
	Micrococcales	1964	185.8	0.04%	0.12%	58593	18.4	0.10%	0.11%
Eukaryota (Fungi)		272375	2617.2	5.66%	1.61%	434756	175.3	0.76%	1.00%
	Basidiomycota	91146	831.7	33.46%	31.78%	161799	64.0	37.22%	36.53%
	Embryophyta	90904	768.6	33.37%	29.37%	61044	24.1	14.04%	13.77%
	Ascomycota	76993	1009.7	28.27%	38.58%	200144	85.5	46.04%	48.78%
	Bilateria	13259	7.13	4.87%	0.27%	11752	1.60	2.70%	0.91%
	Peronosporaceae	62	0.025	0.02%	0.00%	17	0.003	0.00%	0.00%
	Chytridiomycetes	11	0.022	0.00%	0.00%	0	0.000	0.00%	0.00%
Viruses		167242	387.7	3.48%	0.24%	252295	119.0	0.44%	0.68%
	Orthomyxoviridae	67407	107.1	40.31%	27.62%	89589	36.1	35.51%	30.31%
	Mononegavirales	38990	105.1	23.32%	27.10%	63295	28.7	25.09%	24.14%
	Flaviviridae	14748	44.4	8.82%	11.46%	14693	15.5	5.82%	13.05%
	Bunyaviridae	9953	13.0	5.98%	3.35%	10391	4.03	4.12%	3.39%
	Picornavirales	8924	14.0	5.34%	3.60%	8620	4.24	3.42%	3.57%
	Reoviridae	8109	14.2	4.85%	3.67%	8626	4.69	3.42%	3.94%
	Togaviridae	7030	24.8	4.20%	6.40%	8670	7.92	3.40%	6.66%
	Asfarviridae	3490	6.56	2.09%	1.69%	6904	1.70	2.74%	1.43%
	Arenaviridae	2173	5.78	1.30%	1.49%	2894	1.73	1.15%	1.46%
	Nanoviridae	1957	1.92	1.17%	0.49%	1646	0.30	0.65%	0.25%
	Poxviridae	1743	28.8	1.04%	7.44%	26845	7.63	10.64%	6.41%
	Herpesviridae	1496	10.6	0.89%	2.72%	6039	2.84	2.39%	2.39%
	Coronaviridae	1132	11.5	0.68%	2.96%	4164	3.57	1.65%	3.00%
	Tymovirales	14	0.02	0.01%	0.01%	19	0.01	0.01%	0.01%
unclassified sequences		1859	1.33	0.04%	0.00%	11545	3.58	0.02%	0.02%
Viroids		397	0.14	0.01%	0.00%	0	0.00	0.00%	0.00%
synthetic viruses		1	0.03	0.00%	0.00%	13	0.01	0.00%	0.00%

Table 3. Scale of threat taxa, including bacterial and eukaryotic threats and both nucleic acid and protein sequences.

taxa across all of these classes. In particular, note that the scale of cellular threat data is much much larger than for viral threat data.

3.3 Experimental Pipeline

In order to evaluate FAST-NA against the curated training and test data, we have set up an automated experimental pipeline. This pipeline is designed to produce reproducible and deterministic results, be configurable to support many experiments, run unattended, make good use of compute cycles, and record all information necessary to support useful results and analysis.

One instance of our current experimental pipeline is set up primarily for k-fold cross-validation experiments, following the architecture in Figure 7. The pipeline is designed to run a batch of experiments, iterating over a directory of experiment configuration files. These configuration files are simple and can be programmatically created, so experiments with many conditions and combinations can be scripted relatively simply. The experiment pipeline runs as a series of small programs (either FAST or FAST-NA applications), each of which takes files as inputs and emits files as outputs. This makes debugging and manual inspection of intermediate states simple, since individual steps can readily be run again on the same inputs. Debugging output is captured in log files, and data is gathered after each k-fold run and for the

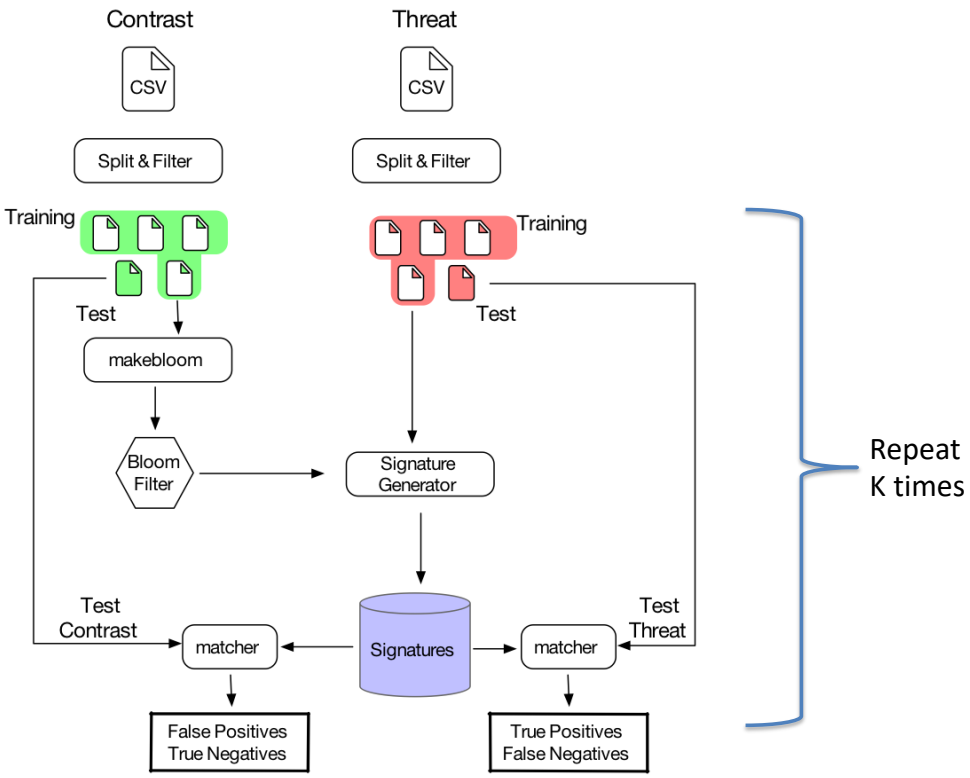


Figure 7. Architecture of k-fold cross-validation in experimental pipeline.

experiment as a whole. In particular, the key information that is captured is:

- Counts of threat and contrast sequences
- Counts of threat and contrast alerts
- Counts of signatures generated
- FASTA file of potential false negatives (non-alerted threat sequences)
- FASTA file of potential false positives (alerted contrast sequences)

A second instance of our experimental pipeline, set up for cross-taxa testing and CONOPS evaluation, is identical except that the threat and contrasting sequences are not split into training and test subsets. Instead, the full collection of threat and contrasting sequences is used for creating signatures, and these are then matched against one or more separately provided collections of test sequences.

Testing for protein-based signatures uses a third variant of this pipeline, mostly identical to the cross-taxa testing pipeline except for two modifications: 1) except the training data is amino acid sequences, 2) nucleic acid

FAST-NA	IDT	Expert	Interpretation
Threat	Threat	~	Baseline
Threat	Non-threat	Threat	Improvement
Threat	Non-threat	Non-threat	False positive
Non-threat	Non-threat	~	Acceptable
Non-threat	Threat	Threat	False negative
Non-threat	Threat	Non-threat	Improvement

Table 4. Expected interpretation of FAST-NA results based on comparison with IDT's current biosecurity screening system and human expert judgement.

sequences are converted into amino acid sequences (in all possible reading frames) to be run in the matcher, and 3) there is no need for cross-validation with this pipeline since protein training data and nucleic-acid test data do not overlap.

Finally, unified protein and nucleic acid screening is done by fusing the results of protein-based screening and nucleic acid screening—at present, simply by taking the union of sequence alerts from both sources.

Results from any of these pipelines are evaluated against the current state of the art by comparison of each potential false negative with IDT's current biosecurity screening system: Table 4 shows the expected interpretation of FAST-NA results based on comparison with the IDT system and/or expert judgement. In particular, we have focused on the potential false negatives, i.e., any threat sequences for which no alert was raised by FAST-NA, as any case in which FAST-NA misses a true threat detected by the current system is of major concern for the value of this approach. IDT thus runs the collection of potential false negatives through its screening system to determine whether it is a judged a threat (omitting potential matches against the test data itself), and evaluating each into one of three categories: "threat", "non-threat", or "too short" for those sequences that FAST-NA can be applied to but IDT's current biosecurity system cannot. We thus compute a final number of false negatives for each test as the number of non-alerted threat sequences that are judged as threats by IDT's current biosecurity system. For well-tuned usage of FAST-NA, the number of such false negatives generally ranges from low to zero, as will be seen in the next section, so the need for involvement of human experts in quantitative evaluation of test results has been only with regards to a relatively small number of specific sequences.

4 Technical Results

Here we report on results from applying FAST-NA to the collection of curated training and testing data using our experimental pipeline, as well as comparison with IDT customer-related sequences and CONOPS evaluation. We begin with an assessment of the ability of FAST-NA to identify diagnostic sequences from publicly available data and how its performance can be assessed and modulated by parameter adjustment. We then evaluate the operation of FAST-NA at scale, across the entire current collection of viral threats, and finally its ability to support a realistic CONOPS with a reasonable resource budget.

4.1 Identification of Diagnostic Sequences from Publicly Curated Data

For an initial assessment of the potential of FAST-NA, we considered a "smoke test" in which the method is used without any tuning of either signature size or contrasting data volume. The goal of this "blind" assessment is to determine whether FAST-NA results are within a reasonable distance of our target rates of <1% false positives and no false negatives. If both of these are generally close, then it is reasonable to expect that the desired performance may be able to be achieved by tuning signature length and contrasting data; if they are generally far, then it may be much more difficult.

Accordingly, we conducted 10-fold cross-validation for each cluster, using all available contrasting data for the cluster and signatures 14 base pairs (bp) long, this being the value used in our preliminary work with 1918 influenza. The rates of false positives and false negatives for all 14 clusters are shown in Figure 8. As desired, false negatives are generally zero. Only two datasets have any false negatives at all, and both of these are at quite a low rate. This indicates that it is reasonable to expect that it generally will be possible for FAST-NA to achieve zero false negatives for viral taxa with appropriate tuning.

False positives are somewhat higher, mostly ranging between 1-10%. Critically, however, the rate of false positives is strongly correlated with the number of contrasting sequences, as shown in Figure 9. This is the behavior that is expected for FAST-NA and indicates that we may reasonably expect false positives to be tunable to radically decreased rates with either increased signature length or increased amounts of contrasting data.

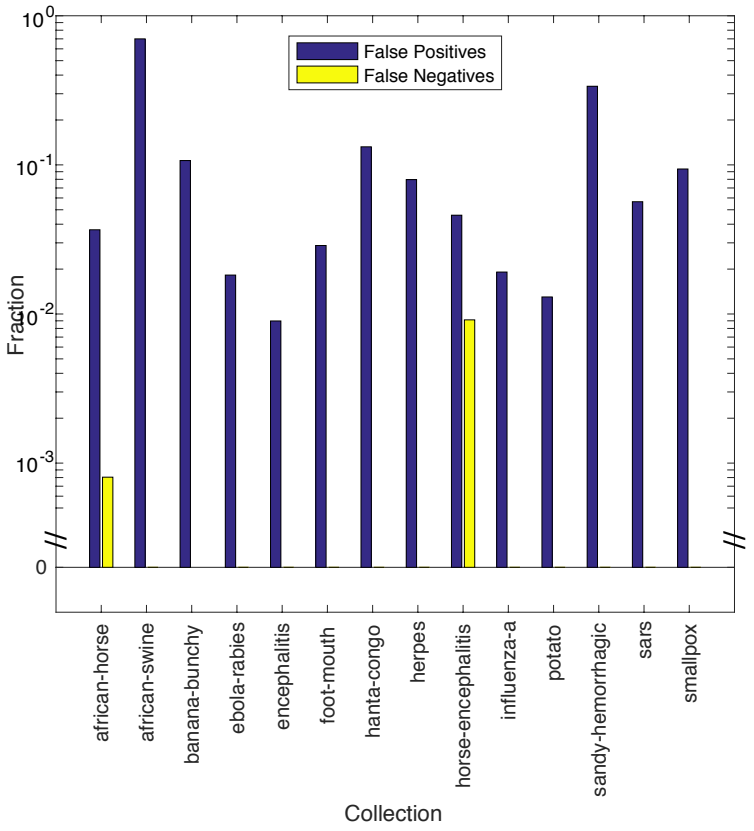


Figure 8. False positives and false negatives for FAST-NA applied to all clusters with 14bp signatures and all contrasting sequences.

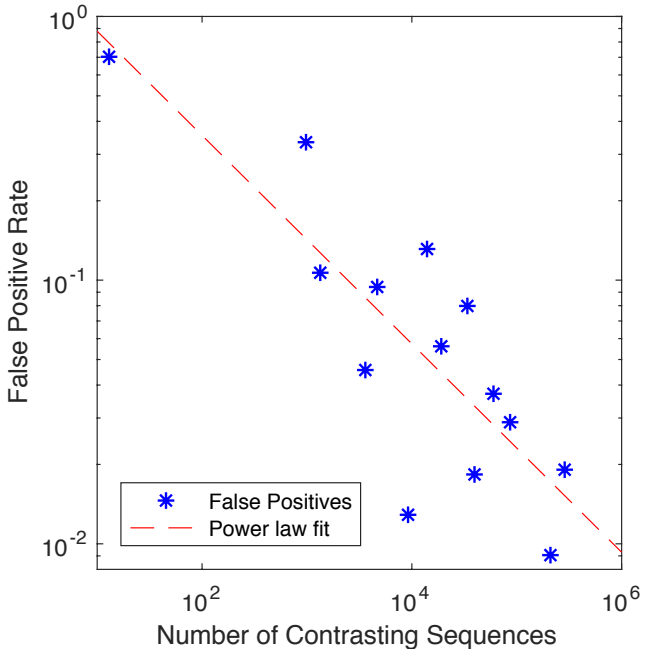


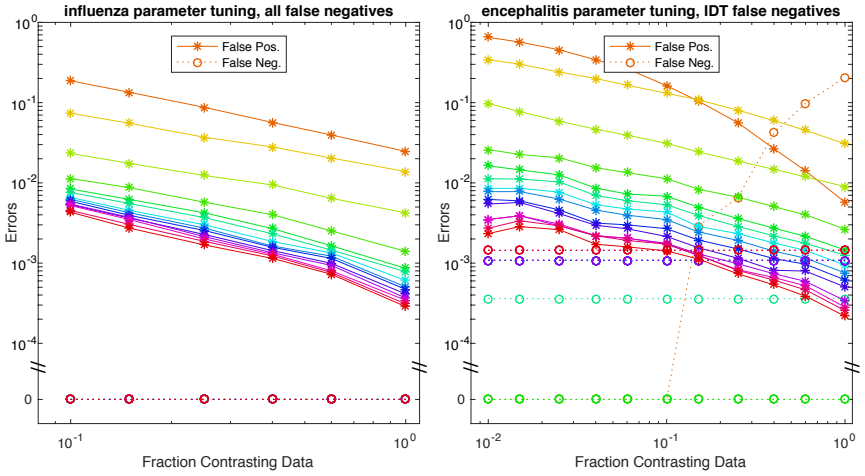
Figure 9. False positives are power-law correlated with contrasting data size across all threat clusters.

We evaluated the potential for tuning in detail with parameter surveys for several viral taxa. For taxa to survey, we selected the sars and banana-bunchy clusters in their entirety, plus sub-taxa of approximately 2000 sequences each from the clusters influenza-a (H5N6 influenza), encephalitis (classical swine fever virus), and ebola-rabies (ebolavirus). For each taxon, we conducted 10-fold cross-validation at all combinations of signature length ranging from 10 to 36 base pairs in steps of 2 (i.e., 10, 12, 14, ...) and fractions of training data ranging from 1% to 100% logarithmically at 5 steps/decade (i.e., 0.01, 0.015, 0.02, 0.04, 0.06, 0.1, 0.15, 0.2, 0.4, 0.6, and 1.0).[1]

Results of these surveys are shown in Figure 10. We find that the rate of false negatives remains consistently zero for all of these except for the encephalitis taxon, which is still consistently zero for most parameter values. The rate of false positives, on the other hand, responds strongly to both signature length and volume of contrasting data. The response is generally stronger for contrasting data, which helps to eliminate shared sequence structure at any length, while the benefit of increasing signature length tends to decrease in benefit after around 20 bp. Critically, note that for every taxa we can achieve a false positive rate of less than 1% (often far less) without introducing any false negatives.

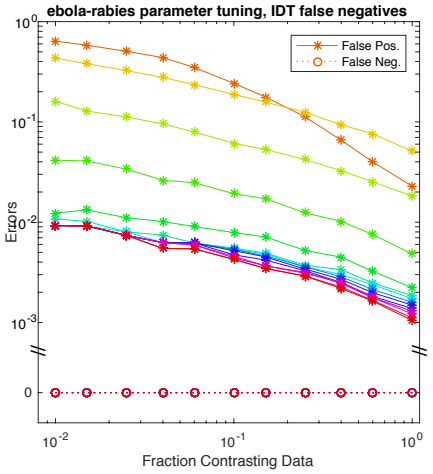
Figure 11 shows the number of signatures generated for each condition of the surveys in Figure 10. For most conditions, we find that the primary determinant in number of signatures is the length of the signatures. The amount of contrasting data does affect the number of signatures, but generally quite weakly except for the shortest signatures. Those shortest signatures, 10 bp or 12 bp long, trend to a power-law decrease in signature count, consistent with the hypothesis that such signatures are too short and will experience a progressive loss of significant sequence information to random matches. Interestingly, this indicates that most (but not all) of the benefit of comparing threat and contrasting data appears to be obtained very quickly through the use of even a small amount of contrasting data. In other words, it is very easy for FAST-NA to pick out the highly conserved sequences most likely to cause false positive hits. Complementarily, this also suggests that one path to further improvements might be through the use of more sophisticated and biology-focused representations of conserved sequence information.

[1] For H5N6 influenza, the minimum fraction was set to 0.1 due to constraints in our experimental system at the time that this test was run.

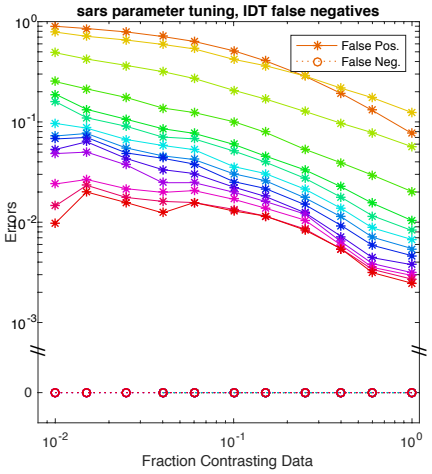


(a) H5N6 influenza

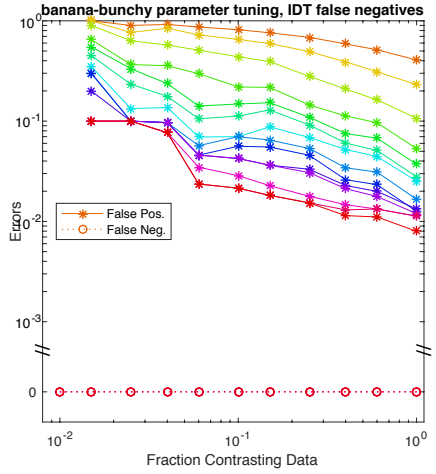
(b) Classical swine fever virus



(c) Ebolavirus

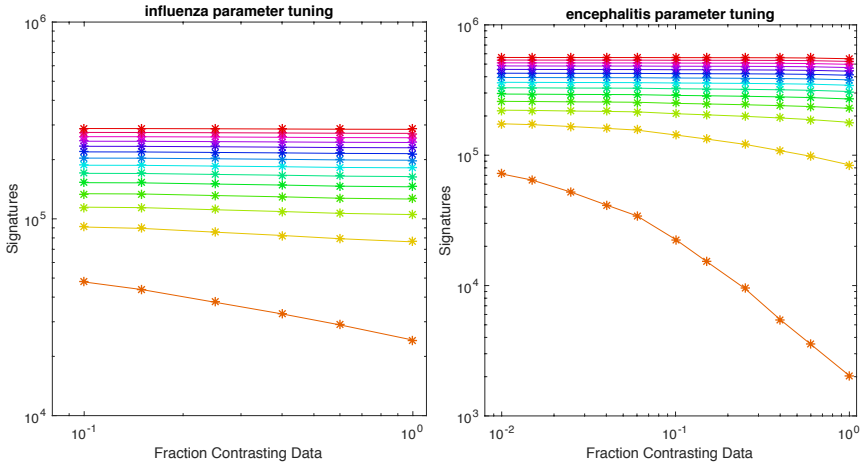


(d) SARS-related coronavirus



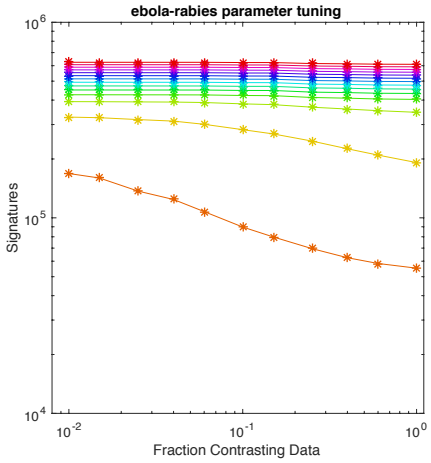
(e) banana-bunchy

Figure 10. Parameter surveys for five viral threat taxa: (a) H5N6 influenza, (b) classical swine fever virus (an encephalitis), (c) ebolavirus, (d) SARS-related coronavirus, and (e) banana-bunchy. Color indicates signature length, shading in hue from n=10 (orange) to n=36 (red).

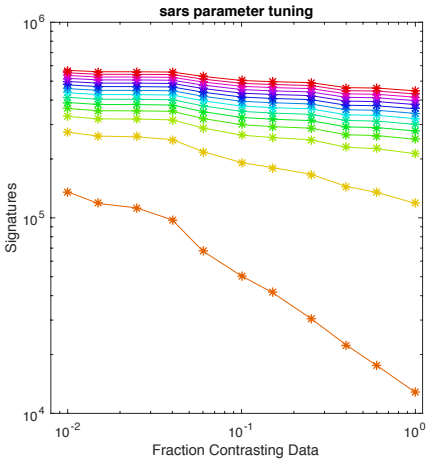


(a) H5N6 influenza

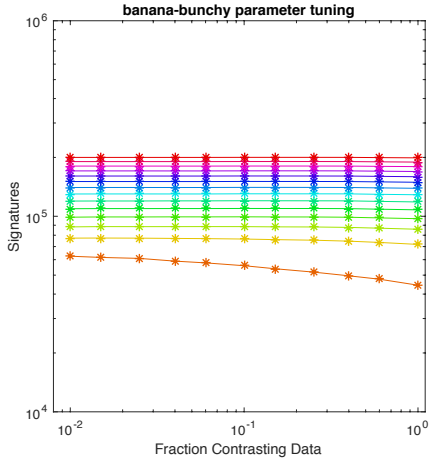
(b) Classical swine fever virus



(c) Ebolavirus



(d) SARS-related coronavirus



(e) banana-bunchy

Figure 11. Number of signatures generated in parameter surveys for five viral threat taxa: (a) H5N6 influenza, (b) classical swine fever virus (an encephalitis), (c) ebolavirus, (d) SARS-related coronavirus, and (e) banana-bunchy. Color indicates signature length, shading in hue from n=10 (orange) to n=36 (red).

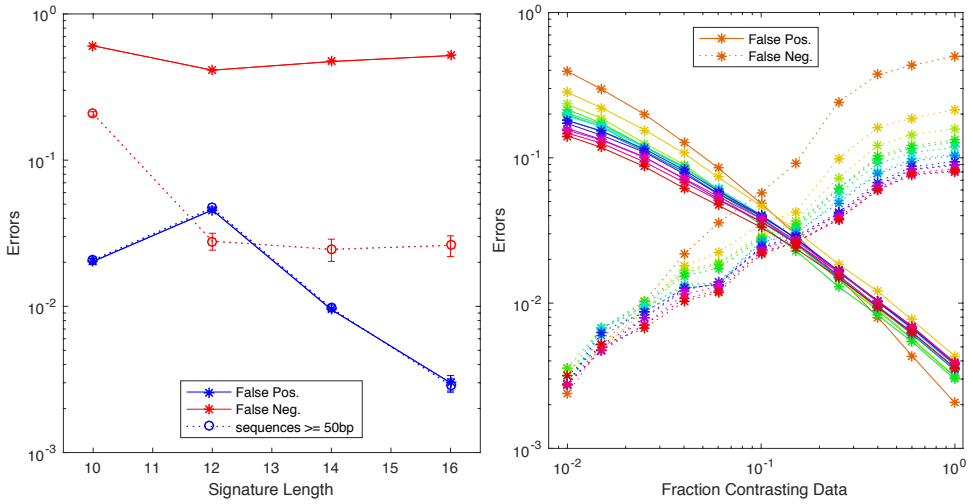
4.2 Assessment and Modulation of FAST-NA Performance

In order to confidently deploy FAST-NA in place of existing systems, we need to know not only that it is working, but also why it is working. Some of this we can infer from the expected scaling behaviors of FAST-NA, as discussed in the previous section, and as further developed below through analyses in this section. Complementarily, it is also valuable to identify failure modes and their causes, in order to be able to diagnose and correct potential misbehaviors.

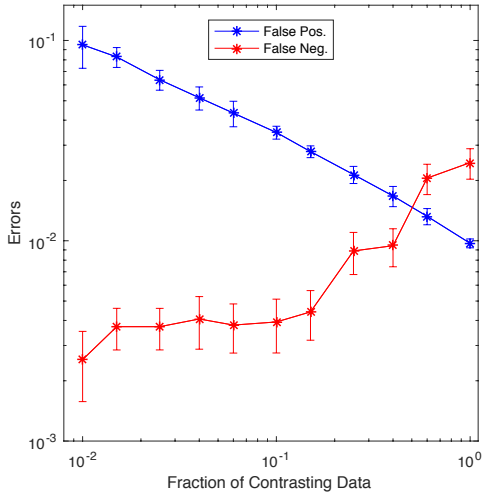
4.2.1 Failure Modes of FAST-NA

From experimentation with FAST-NA to date, we have identified five failure modes, each with a characteristic diagnostic behavior and a recommended means of correction:

- **Short sequence pollution:** Questionable curation decisions cause some taxa in NCBI GenBank to contain large numbers of extremely short sequences, such as lists of primers. These cannot and should not be reliably matched, and can be addressed simply by excluding extremely short sequences from training and testing. For example, the encephalitis threat cluster contains many thousands of 14 bp sequences, mostly from a single exhaustive list in a particular patent filing. Figure 12(a) shows an example of the effect of ignoring sequences of <50 bp from the encephalitis test set; we have applied this filtering for all of the results presented in the prior section.
- **Contrasting sequences too closely related:** If contrasting sequences are too closely related to the threat taxa, then false positives cannot be reduced without simultaneously increasing false negatives. For example, the Influenza A taxa includes many extremely closely related sub-taxa, some of which are categorized as threats and others as non-threats. This failure mode can be recognized from a consistent power-law increase in false negatives vs. contrasting data, as shown in Figure 12(b). It can be addressed by designating the most closely related taxa in the contrasting data as "neutral," for which either threat or non-threat judgements are reasonable (e.g., there is no particular reason to assume H7N5 influenza is a non-threat, even though it has not been specifically designated as a threat like H1N1 influenza). We have done this with both the influenza-a and sars clusters in producing the results in Figure 10.
- **Signatures too short:** When signatures are too short, either false positive rates or false negative rates will be too high (or even both too high



(a) Short sequence pollution (b) Contrasting material too close



(c) Too much contrasting data

Figure 12. Examples of FAST-NA failure modes: (a) Training on encephalitis with very short sequences (stars) has much greater false negatives (red) than eliminating very short sequences (circles), while false positives (blue) are essentially unaffected. (b) False negatives rise sharply as false positives fall when contrasting data is too close, as in this example of H5N6 influenza contrasted against all non-threat orthomyxoviridae including non-threat strains of influenza A. Color indicates signature length, sharing in hue from n=10 (orange) to n=36 (red). (c) Too much contrasting data can eliminate useful signatures, causing false negatives to rise sharply, as in this example with encephalitis.

at once). This failure mode can be diagnosed by raising the signature length incrementally: if there are marked improvements in performance, then it is likely that signature length was too short; if not, then signature length is not the problem. The parameter surveys in Figure 10 all demonstrate this relation.

- **Too much contrasting data:** When there is too much contrasting data, it may produce false negatives due to random sequences eliminating critical signatures. When this failure mode pertains, reducing contrasting data will reduce the number of false negatives to zero faster than false positives rise. An example of this failure mode is shown for the encephalitis cluster in Figure 12(c).

- **Too little contrasting data:** When there is too little contrasting data, false positives will tend to be too high, but there will be no false negatives. This failure mode can be diagnosed by adjusting the amount of contrasting data, which should result in a power-law decrease of false positives while false negatives remain zero. The parameter surveys in Figure 10 all demonstrate this relation.

4.2.2 Estimation of Appropriate Signature Length

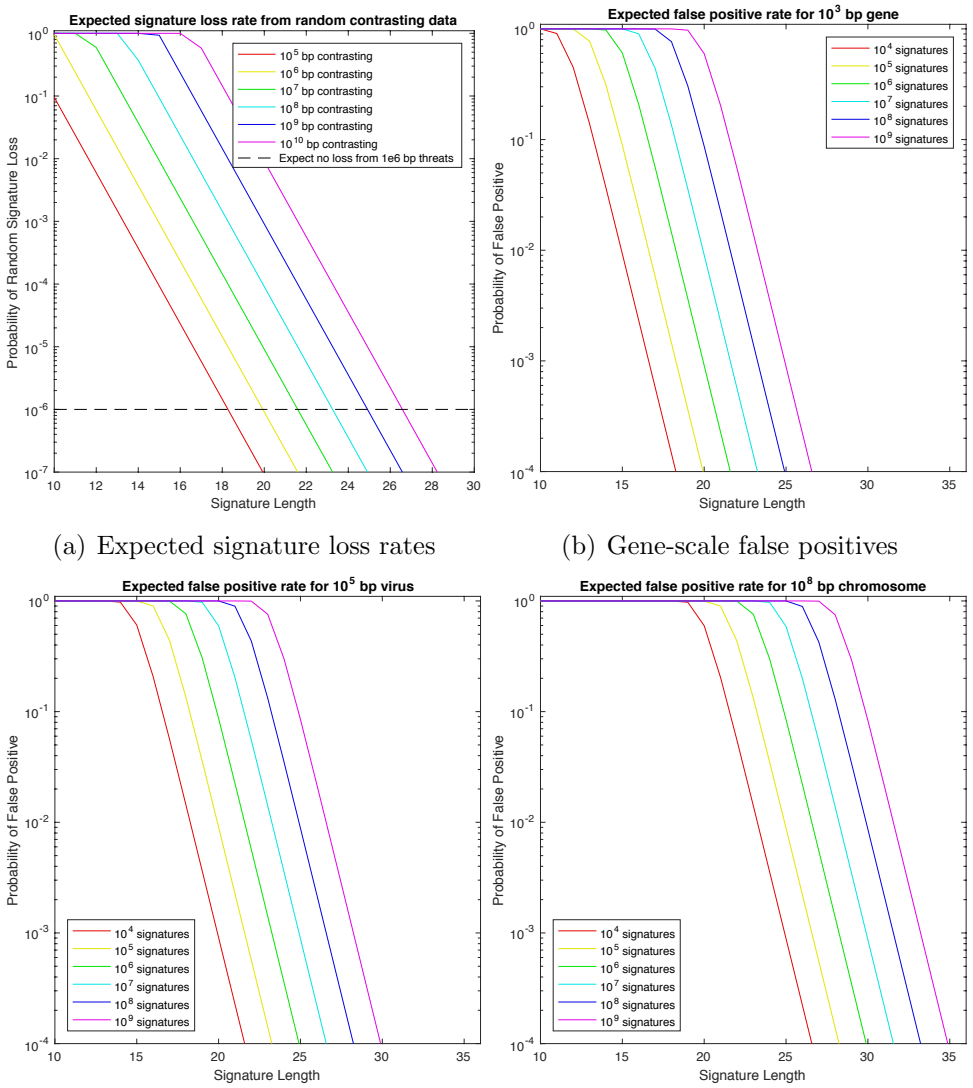
As signature length is a critical parameter for FAST-NA without an obvious default value, we analyzed the mathematical relationship between signature length and false positive rate, in order to be able to predict values for signature length that are likely to provide good performance.

As seen in our results above, signature length matters for FAST-NA in two ways. First, false negatives can be produced when signatures are eliminated by effectively random sequences when they collide with signatures. Second, false positives can be produced when signatures match against an effectively random sequence.

For analysis of both of these cases, we consider a conservative approximation estimating the probability of a match between two random sequences:

$$p_{random\_hit} = \frac{sequence\_size}{4^{signature\_length}}$$

In other words, the likelihood of a hit is approximately the number of opportunities for a signature to hit a sequence divided by the number of possible signatures (an approximation valid when numerator is significantly smaller than the denominator). The actual probability of match will, of course, be



(a) Expected signature loss rates

(b) Gene-scale false positives

(c) Virus-scale false positives

(d) Chromosome-scale false positives

Figure 13. Conservative estimation of random signature match rate for (a) pruning signatures with contrasting data, (b) matching a single 10^3 bp gene, (c) matching a 10^5 bp viral genome, or (d) matching a 10^8 bp eukaryotic chromosome.

different, as sequences are not random, signatures and match opportunities overlap, and edges are not accounted for, but this conservative estimate provides a good baseline to work from for recommending a reasonable initial value for signature length.

Applying this relation to the potential for false negatives, one obtains:

$$p_{lost\_signature} = \min(1, \frac{contrasting\_bp}{4^{signature\_length}})$$

Representative values for the likelihood of losing a critical signature to random pruning is shown in Figure 13(a). considering signatures from 10 to 30 bp and a volume of contrasting data from 10^5 to 10^{10} bp. From this, one can

see that although signatures in the 10-16 bp range are expected to be badly ablated by even relatively small amounts of contrasting data, the likelihood of randomly losing signatures drops so sharply with increasing length that with these volumes of contrasting data, by the time signatures are 18 to 27 bp, not even one in a million signatures should be expected to be randomly lost.

Complementarily, the same relation may be applied to estimate the false positive rate:

$$p_{false\_positive} = 1 - ((1 - p_{random\_hit})^{sequence\_length})$$

In other words, the chance of getting a false positive is the complement of the chance that every chance of a random hit on a sub-sequence will fail. Figures 13(b)-(d) show these probabilities for various signature lengths and threat signature counts against representative sizes for the sequence of a typical gene (10^3 bp), large virus (10^5 bp), or eukaryotic chromosome (10^8 bp), respectively. As with the likelihood of random signature loss, very short signatures are expected to have an unacceptably high likelihood of false positives, but even moderately longer signatures can set the probability of random hits to be very low indeed. For example, even with very large numbers of signatures, large viruses have a random hit rate of less than 1 in 1000 with only 28 bp signatures.

Considering both false positives and false negatives, we may thus conservatively estimate an appropriate signature length as the maximum signature length required to achieve the desired performance for either false positives or false negatives. In the case of this investigation, we chose to consider false negatives less than 10^{-6} with 10^{10} bp of contrasting data, giving a length of 27 bp, and false positives less than 10^{-3} with 10^9 signatures on a 10^5 bp virus, giving a length of 28 bp. The initial signature length for developing signatures for each viral taxa was thus selected to be 28 bp.

The same system of estimation may be applied to potential amino acid signatures as well, with one modification: since amino acids have an alphabet of 20 options for amino acids as opposed to 4 options for nucleic acids, the equations need to be re-written with 20 instead of 4. The consequence is that amino acid signatures can be expected to have a dramatically reduced rate of random matches, as illustrated in Figure 14: in general, an amino acid signature will have the same probability of random matches as a nucleic acid signature a little over twice as long. Ultimately, however, since amino acids are defined using codons of 3 nucleic acids per amino acid, this means

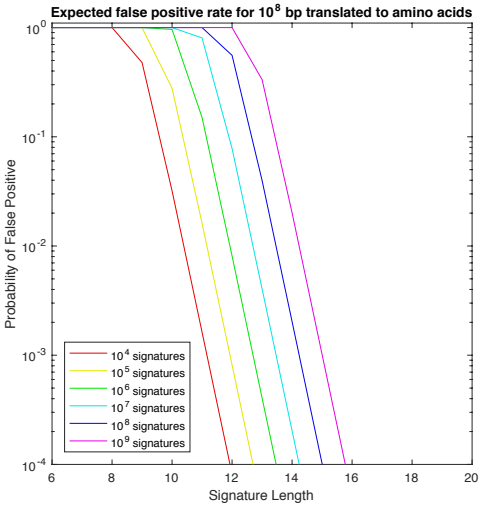


Figure 14. Random signature match estimates can be applied to amino-acid signatures as well, finding shorter lengths are required due to the larger alphabet, as in this example for matching amino acid signatures at the 10^8 bp chromosome scale.

that amino acid signatures should generally be expected to be effective when matching against slightly longer regions of nucleic acids than purely nucleic acid signatures.

4.2.3 Systematic Tuning of Parameters

From these investigations of failure modes and random hits vs. signature length, we have developed an initial approach to systematic tuning of parameters for a viral taxa:

1. Train using 10-fold cross-validation with the computed conservative signature length (28 bp) and 100% of contrasting sequence data.
2. If there are any false negatives, then modulate signature length and percent contrasting data (independently) to diagnose the likely cause. The specific values that we chose for this implementation were signature length at 22 bp to 27 bp and contrasting data on a logarithmic scale: 10%, 15%, 25%, 40%, and 60%, for a total of 11 additional training runs.
3. From the results of these two modulations, we obtain one of three cases:
   (a) Modulation identifies a functional length or contrasting fraction.
   (b) Modulation identifies that contrasting sequences are too closely related, implying the need to designate "neutral" taxa. Once these are identified, start over again with 28 bp signatures and 100% contrasting data (less the neutral taxa).

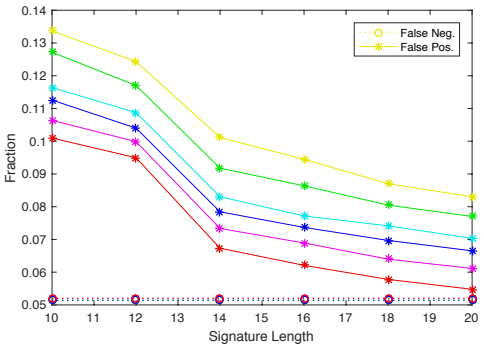


Figure 15. Example of landscape evaluation for protein-based signatures with the smallpox threat cluster, showing tunable response to signature length and contrasting data percentage.

(c) Otherwise, the problem likely cannot be resolved by tuning, and hand-inspection of failures is required for diagnosis.

Note that high false positives (>1%) are not addressed by this tuning procedure. The modulation of fraction of contrasting data may, in fact, indicate that additional contrasting data is needed, per the diagnosis suggested above. In such cases, we experimented with the possibility adding extra contrasting data (necessarily from less related taxa) as a mechanism for addressing high false positives. Our initial experimentation with this approach, however, showed little or no benefit (consistent with our results on random signature matches), and we have thus excluded it for now from our recommended tuning approach.

We have also validated that this approach can work for the tuning of protein-based signatures. Evaluation of the tuning landscape of protein-based signatures for select taxa shows a similar tunable response to signature length and contrasting data percentage, indicating that the same systematic tuning approach should be expected to be effective for protein-based signatures. Figure 15 shows an illustrative example with the smallpox threat cluster.

4.3 Full-Scale Viral Application of FAST-NA

To evaluate the potential of FAST-NA for application to viral screening at full scale, we have applied it to create signatures to screen for all viral threats currently screened for in IDT's biosecurity system. We then evaluated the efficacy of these signatures both for detecting threat sequence fragments of at least 50 bp and also for detecting very short sequences excluded from training data.

4.3.1 Signature Development for All Viral Threats

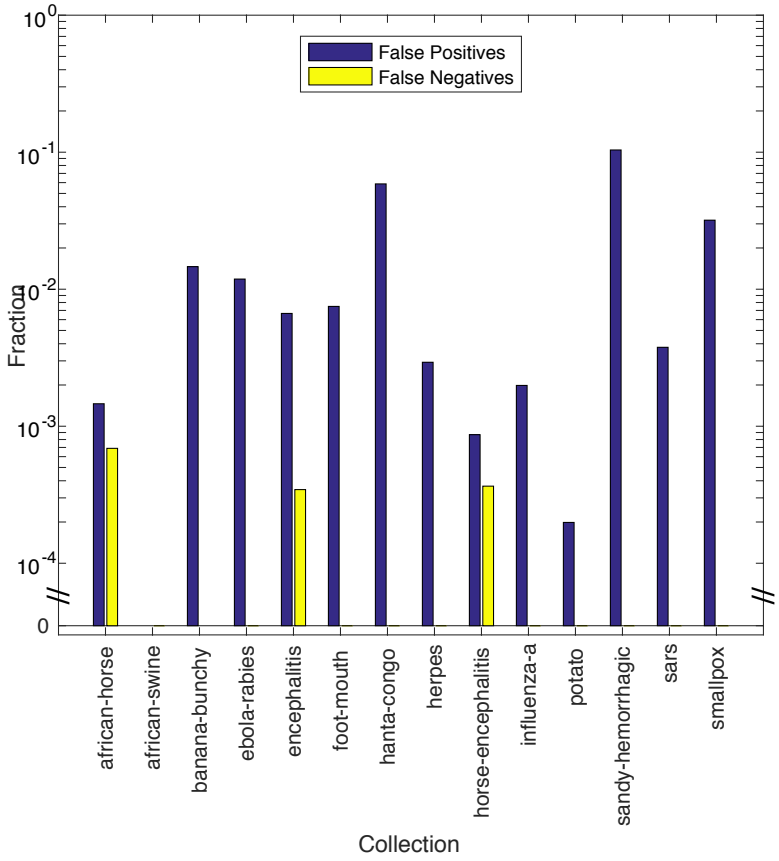


Figure 16. Results of 10-fold cross-validation with tuned parameters for all viral threats.

For each of the 14 threat clusters, we applied the systematic tuning procedure presented in Section 4.2.3. The majority of threat clusters (8 of 14) produced no false negatives on the first pass with 28 bp signatures and 100% contrasting sequence data. Of the six that did not, the tuning procedure produced final results with the following adjustments:

- african-horse: 26 bp signatures
- ebola-rabies: 15% of contrasting sequences
- encephalitis: 15% of contrasting sequences
- hanta-congo: 25 bp signatures
- horse-encephalitis: 24bp signatures + neutral taxa
- influenza-a: neutral taxa

Figure 16 shows the results of 10-fold cross-validation testing with the final tuned parameters for each of the clusters of viral threats. Overall, the rate

of false positives is acceptably low: 9 of the 14 threat clusters have under 1% false positives, and the mean rate of false positives, weighted by test sequence count, is only 0.72%.

Out of 163,130 threat sequences, there were 13 false negatives in three taxa, each of which we analyzed individually. We find that these false negatives break down into three cases:

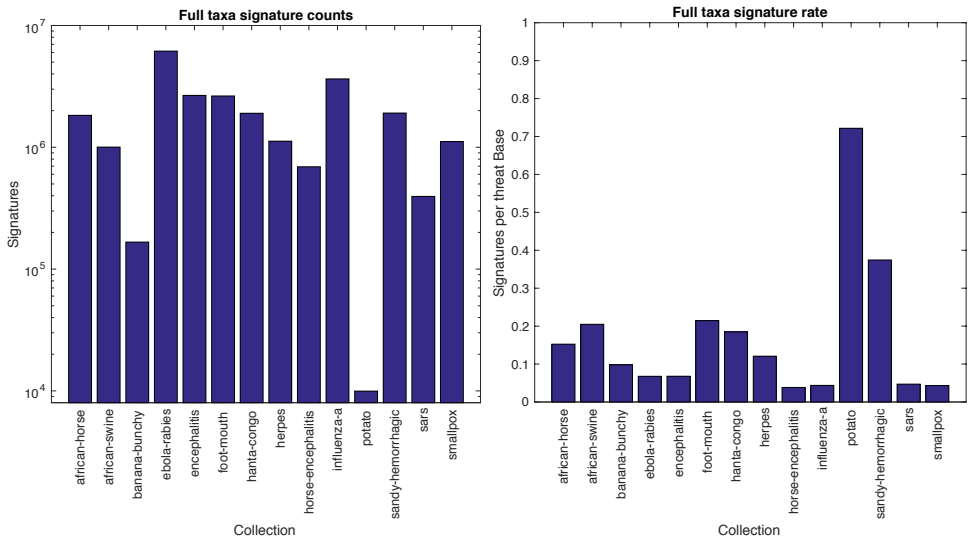
- Five sequences (1H1K_J, 1H1K_K, 1H1K_L, 1H1K_M, and 1H1K_N) are meaningless BLAST matches between a likely spurious sequence of 200+ A or 200+ T bases, matching against a long unspecified ("N") sequence representing an unknown portion of a rice fungus genome. Here we judge these to be clear false alarms and FAST-NA correct to not trigger on these sequences.
- Two sequences (KF022090.1 and KF022091.1) are reverse-complement matches that would likely be caught if we were generating signatures for reverse-complement sequences as well as original sequences.
- Six sequences (HQ719213.1, AF004437.1, AF004436.1, EU303181.1, KJ624719.1, and AF196534.1) are matches with poor alignment in nucleic acid sequence, but long matching amino-acid sequences that would likely be caught if we were generating signatures for encoded amino acids as well as nucleic acids.

As expected, we thus find that FAST-NA is highly effective for screening against threats, but may be unable to detect some mutated, codon-optimized, or reverse-complement threats until reverse-complement and amino-acid signatures are added.

4.3.2 Time and Signature Scale for All Viral Threats

The total time to develop all signatures was acceptably short. The first pass took approximately two days to complete, followed by an additional six days of tuning and adjustment for the five threat clusters that did not produce acceptable results on the first pass, for a total of approximately eight days to produce signatures with acceptable performance for detection of all current viral threats.

Figure 17 shows the number of signatures produced from each cluster, as well as the fraction of sequence data used in signatures, computed in approximation by dividing number signatures by total base-pairs of threat data.

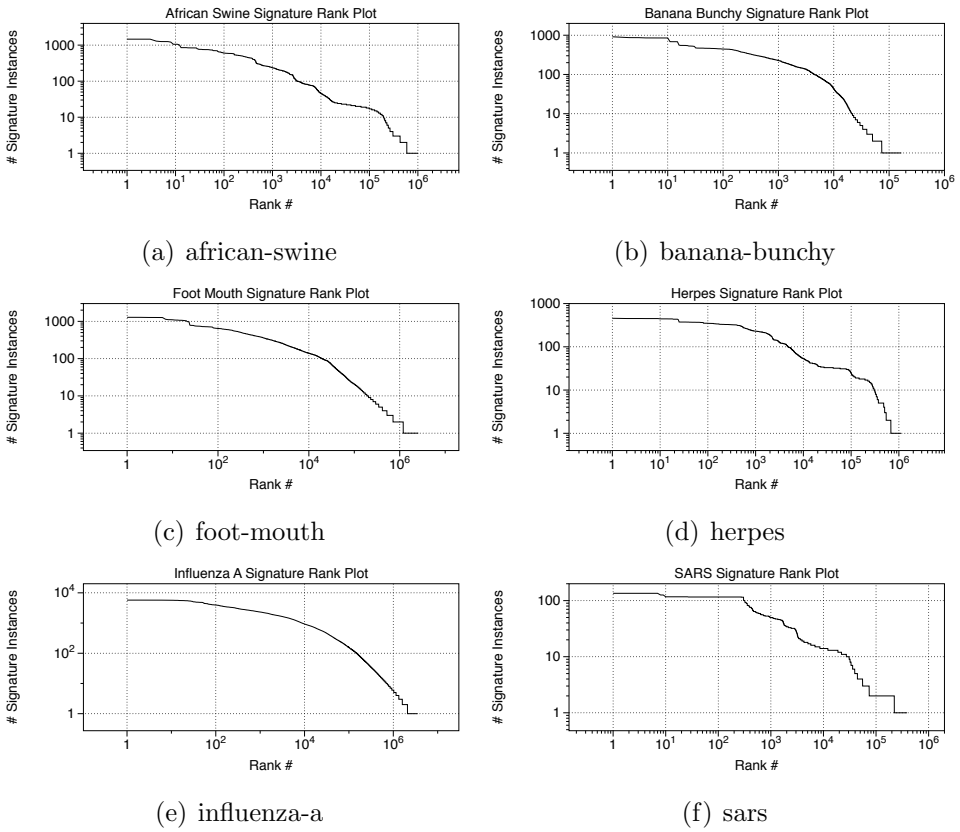


(a) Signature Counts (b) Fraction of Sequence

Figure 17. (a) Number of signatures per cluster for full-scale viral threat detection, and (b) fraction of sequence used in signatures.

Most taxa fall into a band of approximately 5% to 20% of sequence used in signatures, and thus have total counts that correlate with the amount of threat data provided. The only exceptions are the potato and sandy-hemorrhagic clusters. While this might indicate that both of these clusters have insufficient contrasting data, their outcomes are quite different: the potato cluster appears to be part of a highly diverse and poorly explored taxa of mostly plant-affecting viruses, and thus to be so taxonomically distinct as to have little difficulty with false positives. The sandy-hemorrhagic cluster, on the other hand, does suffer in performance, finding many false positives. Indeed, overall there appears to be no strong correlation between fraction of sequence in signatures and false positive rate.

Looking more deeply into the nature of the signatures extracted from threat files by FAST-NA, we analyzed the relationship between signatures for a variety of threat clusters. Figure 18 shows the degree to which signatures are shared between different threat sequences for six threat clusters: african-swine, banana-bunchy, foot-mouth, herpes, influenza-a, and sars. For each signature, we compute the number of threat sequences in which an instance of the signature appears, then visualize the distribution of instance counts by sorting by rank (most common first) and plotting rank versus count. We find that every distribution begins with a significant collection of widely shared signature instances, which we believe likely correspond to highly conserved aspects of the threats in the threat cluster. The curve does not maintain a consistent slope, however, but undergoes numerous bends that likely correlate with some mixture of both meaningful structure (e.g., dif-



(a) african-swine (b) banana-bunchy

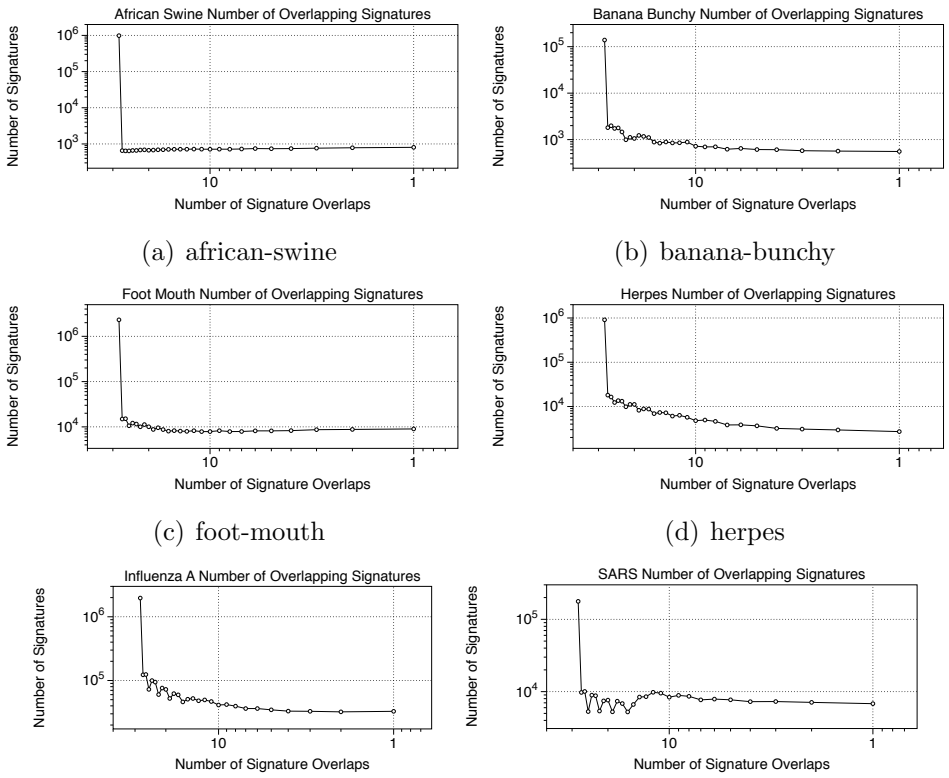
(c) foot-mouth (d) herpes

(e) influenza-a (f) sars

Figure 18. Distribution of number of threat sequences sharing each signature for selected threat clusters.

ferent taxonomic clusters) and non-informative happenstance (e.g., the degree of sequencing that investigators happen to have done regarding particular subjects). At the lower end of the curve, most signatures appear in only one or two threat sequences. The diagnostic value of such signatures is dubious, as for the most part they appear much more likely to capture non-significant variations rather than any conserved quantity regarding a sequence of concern—as well as possible curation errors. An important question for future investigation is whether false-positive rates can be reduced without affecting false negative rates by removal of low-replication signatures.

Figure 19 plots another key statistic over signatures: the number of signatures from a sequence that overlap with the first base of each signature. This computation includes the signature itself, so for 28 bp signatures the values plotted range from 1 to 28. Here, the critical thing to observe is that in the case of every collection analyzed, the frequency of signatures with maximum overlap is much higher than any other length of overlap—in fact, by nearly two orders of magnitude. What this implies is that most signatures are part of long segments of a threat sequence, rather than isolated



(a) african-swine (b) banana-bunchy

(c) foot-mouth (d) herpes

(e) influenza-a (f) sars

Figure 19. Distribution of overlap between signatures for selected threat clusters.

fragments, which in turn implies that the results we have obtained are likely to be relatively tolerant of changes in signature length.

Finally, we performed an in-depth spot-check on the biological information associated with signatures by performing sequence-level analysis of signature coverage on selected viral genomes. Figure 20 shows an example of these results, in particular for a specific instances of the Ebola virus. Signature coverage for Ebola focus heavily on the NP gene, which encodes viral replication, and the GP genes, which is involved with cell docking and penetration, both of which tend to be heavily implicated in the danger of a pathogen. Other areas of interest seem to be the sequence caps (important for viral initiation and stability), while most other locations have scattered signatures likely representing small non-significant mutations differentiating this virus from its non-pathogenic relatives. These and similar spot-check results give further evidence that FAST-NA appears to be preferentially selecting meaningfully diagnostic features of pathogens.

4.3.3 Cross-Taxa False Positives and Threat Misidentification

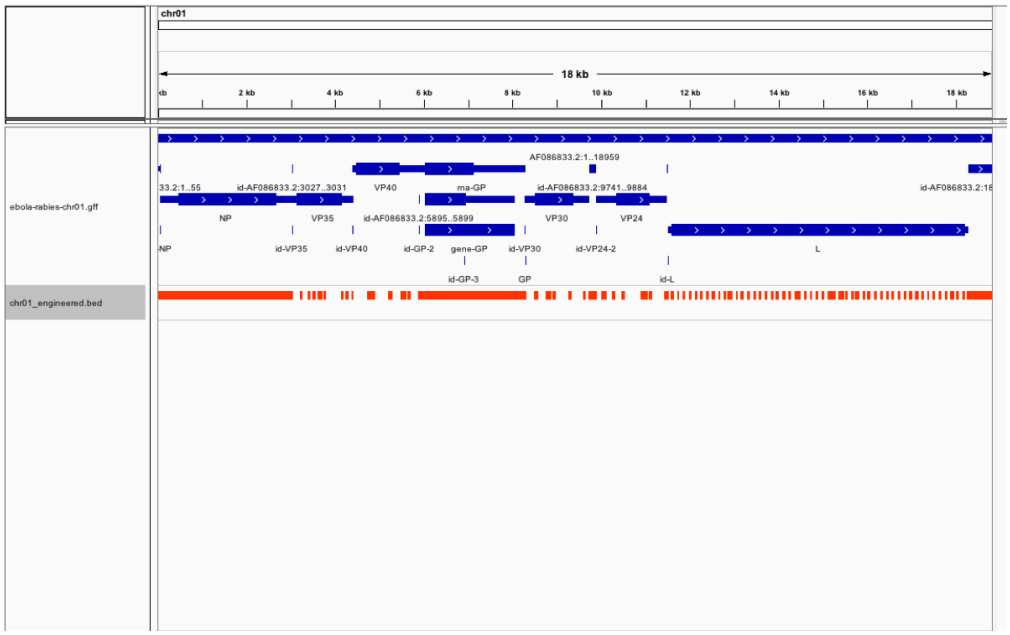
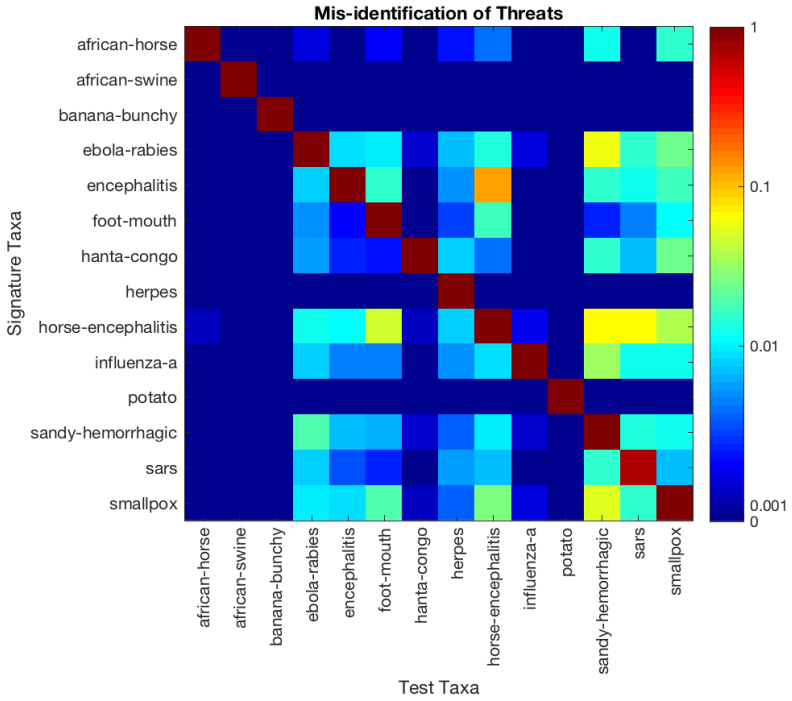


Figure 20. Signature coverage appears biologically correlated, as in this example of Ebola virus, which shows signature coverage (red) focused most heavily on the genes for replication and cell docking.

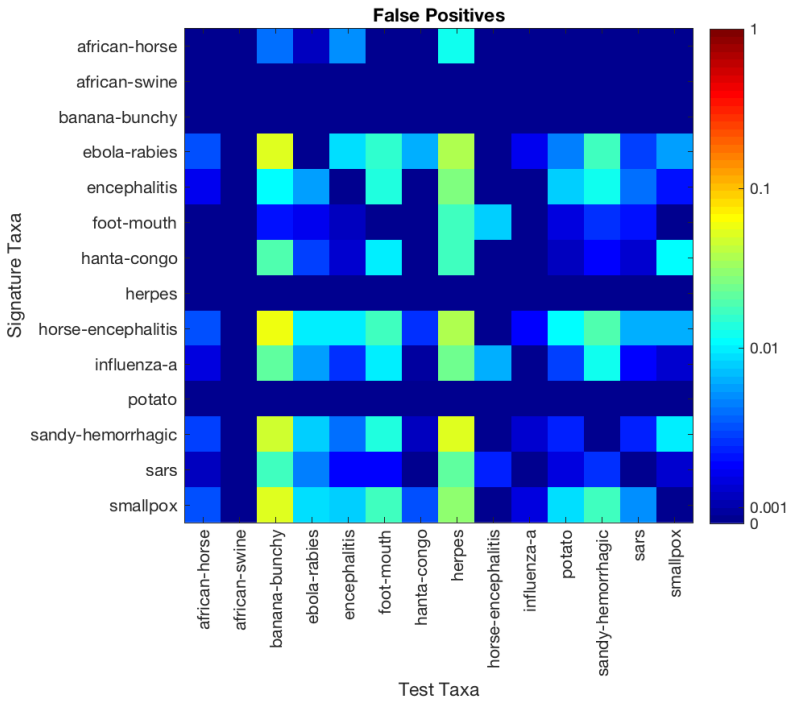
We next evaluated the interaction of signatures across taxa, applying signatures generated from the entire training set for each cluster (nothing held back for cross-validation) to the threat and contrasting sets of each cluster. The ideal match rate across clusters should be zero: every match of a non-threat from any cluster is a false positive, and every match of a threat from a different cluster is a mis-identification of a threat.

Figure 21(a) shows the rate at which threats are identified as belonging to a particular cluster by the signatures for a cluster, i.e., testing signatures against threat sequences. As indicated by the previous section, the diagonal shows that the rate of false negatives is negligible. Off the diagonal axis, the rates are in general low, with a weighted mean of 0.35% mis-identifications. Similarly, Figure 21(b) shows the rate of false positives across taxa, i.e., testing signatures against contrasting sequences. Here again, off the diagonal axis the rates are similarly low, with a weighted mean of 0.45% false positives.

The distribution of mis-identifications and false positives, however, is quite uneven, with a relatively small fraction of interactions accounting for a large percentage of errors. For mis-identifications, errors are highest for sequences from horse-encephalitis, sandy-hemorrhagic, sars, and smallpox, and for signatures from encephalitis and horse-encephalitis. We might expect the rates of mis-identification to be highest within related taxa, e.g., for it to be eas-

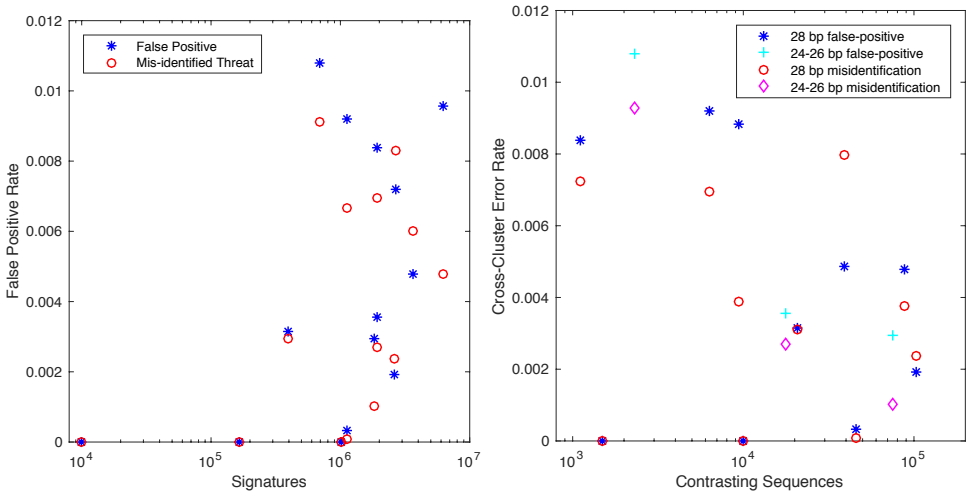


(a) Threat Identification



(b) False Positives

Figure 21. Rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures.



(a) Signatures vs. errors (b) Contrasting sequences vs. errors

Figure 22. (a) No significant correlation is observed between signature count and false positive rate, but (b) errors go down with increased contrasting sequence data.

ier to mis-identify a dsDNA viruses as another dsDNA virus, rather than to mis-identify a dsDNA virus as an ssRNA virus. There is no evidence for such a relationship, however: there are 11 sequence/signature pairs with a mis-identification rate of higher than 2%, and of these six are in the same taxonomic grouping per Figure 5 and one in the same top-level grouping (horse-encephalitis and sandy-hemorrhagic both being ssRNA viruses). The remaining five all involve smallpox, either as signature or test sequence.

For false positives, results are similarly uneven and without obvious pattern. Errors are highest for sequences from banana-bunchy and herpes and for signatures from ebola-rabies, horse-encephalitis, and smallpox. Here there are 12 sequence/signature pairs with a false positive rate of higher than 2%, every one of which involves sequences from either banana-bunchy (five) or herpes (seven). There is no clear relationship with taxonomic closeness here either, but all of the clusters with higher than 2% false positive for banana-bunchy also have higher than 2% false positive for herpes.

Digging deeper into the sources of false positives and misidentifications, we might suspect that having more signatures would lead to more errors, simply due to the increased number of opportunities for error. In fact, however, Figure 22(a) shows that we find no such correlation. This is consistent with the fact that false positive and misidentification rates are well above the predicted baseline for random hits, suggesting that false positives are indeed due to some sort of related structural properties between sequences.

There does, however, appear to be a correlation between the number of closely related contrasting sequences available for training and the rate of false positives and misidentifications, as shown in Figure 22(b). This suggests that error rates could be further reduced if additional contrasting sequence data can be obtained, either through identification of taxa likely to have enough relationship to provide useful contrast or through the growth of sequence collections over time.

In fact, however, deep inspection of signatures and signature sources determined that the vast majority of false positives stemmed from three sources:

- A small number of artificial hybrid sequences that cross taxonomic boundaries, thus confusing threat taxa. For example, consider the high rate of threat mis-identification in which horse-encephalitis cluster threat sequences are identified as belonging to the encephalitis cluster. Of those, 99.5% are caused by a set of six engineered sequences from a single research project that hybridized an encephalitis sequence with VEEV, which belongs to the horse-encephalitis cluster.
- Signatures with long sequences of unknown ("N") nucleotides. Since gaps in knowledge have no information content and no taxonomic correlation, these effectively shorten the signature length, resulting in more non-diagnostic matches.
- Signatures with long poly-A or poly-T sequences. Long poly-A sequences are a frequently used signal in mRNA processing, and long poly-T sequences are their necessary complement, so such sequences are also not expected to contain any significant diagnostic content. As with long sequences of unknown nucleotides, long poly-A and poly-T sequences effectively shorten signature length, resulting in more non-diagnostic matches.

Accordingly, we enhanced FAST-NA to add the ability to filter out signatures with more than a certain number of repeated N, A, or T nucleotides in sequence, as well as the ability to omit a list of specific threat sequences from training.

4.3.4 Detection of Threats in Very Short Sequences

Although we have tuned the full-scale viral signatures to detect sequences of at least 50 bp, training only on signatures of that length or more, it is worth asking whether these signatures can also effectively detect shorter sequences. Across all of the clusters of viral threat and contrasting sequence data, there

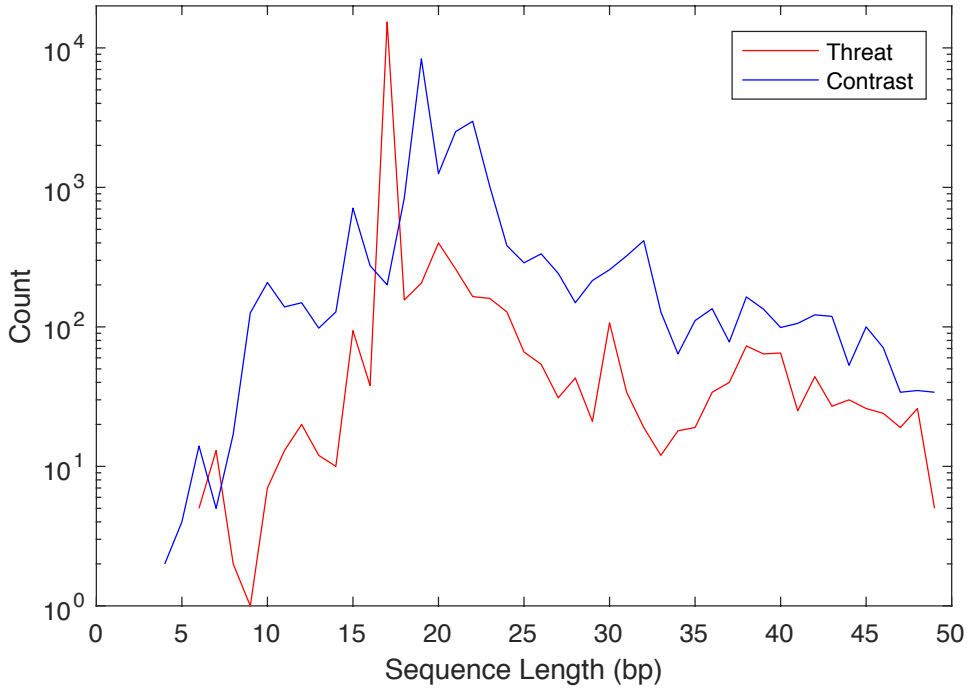


Figure 23. Distribution of lengths for short (< 50 bp) sequences in viral threat and contrasting sequence data.

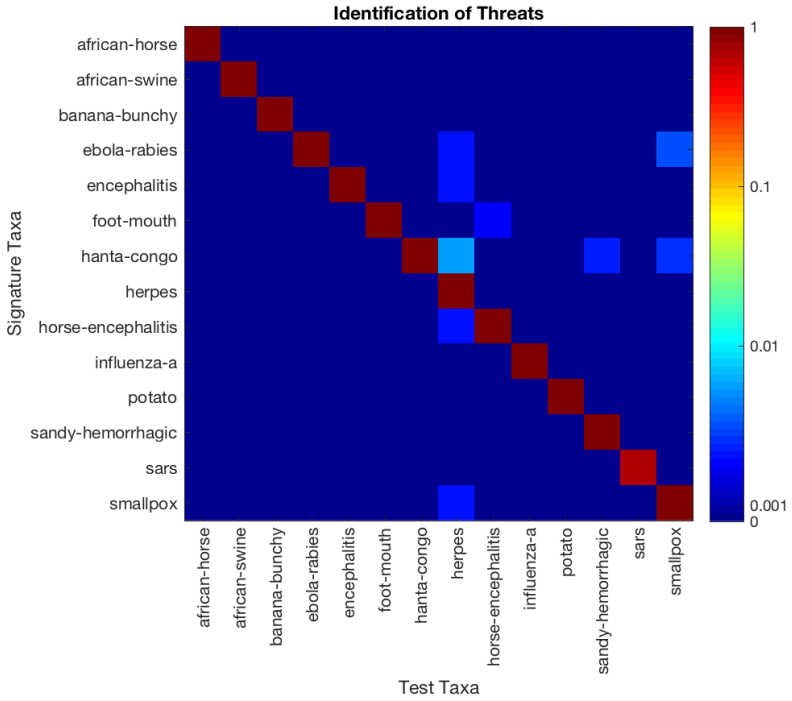
are a total of 17,962 threat sequences and 23,218 contrasting sequences excluded for being less than 50bp in length. Figure 23 shows the distribution of lengths across these sequences.

The bulk of these sequences are, in fact, shorter than the length of our signatures, and thus by definition cannot be detected by these signatures. For all sequences at least 28 bp in length, however, we find the rate of detection to be remarkably high. In total, 52.9% of all threats are identified correctly, with only 1.0% misclassifications. Interestingly, false positive rates are higher as well for these extremely short sequences at a rate of 8.9%.

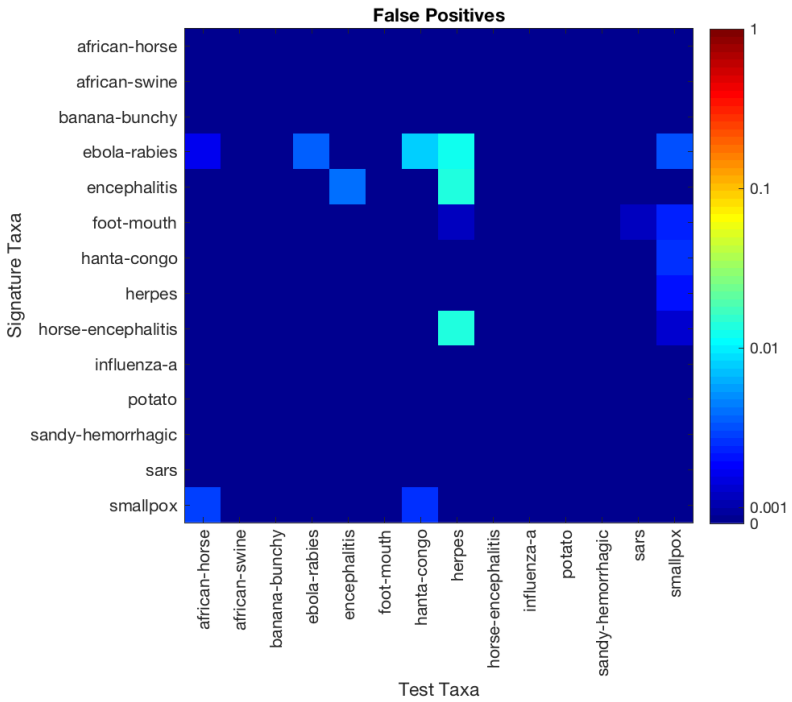
While this is a much lower efficacy than for the longer sequences that we have tuned FAST-NA to detect, these results are still good enough that it may be worth further investigation of whether FAST-NA can be applied to at least reduce the potential threat from very short sequences.

4.3.5 Enhancing Threat Detection

Per the analyses above, we enhanced FAST-NA with the addition of reverse complement screening and false-positive reduction techniques. Reverse complement screening was implemented by training against both forward and reverse complement contrasting sequences, then screening using both the forward and reverse complement instances of a sequence. False positive re-



(a) Threat Identification



(b) False Positives

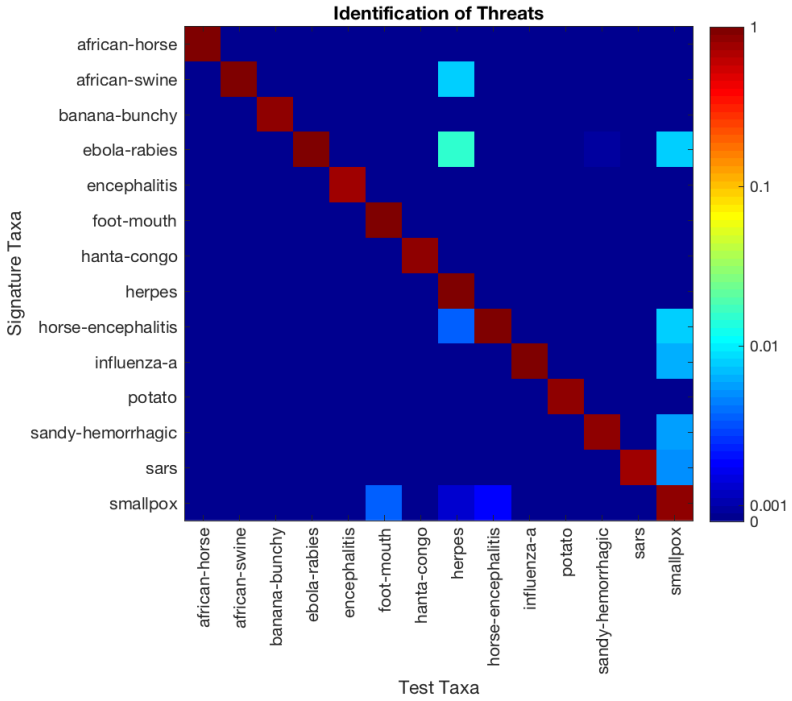
Figure 24. Nucleic acid screening with reverse complement and false-positive reduction: rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures.

duction was conducted by addressing the three primary sources of false positives identified above. Analysis of false positive sources identified 16 "submission clusters" (that is, sets of sequences with closely related identifiers, typically all submitted as part of a single publication) that contained at least one problematic hybrid sequence. Removing these clusters from training removed a total of 75 out of 163101 sequences, or less than 0.05%, a negligible fraction. In addition, we removed all signatures with any unknown ("N") nucleotide and any containing a sequence of 15 or more consecutive A or T bases. Figure 24 shows the improvement in performance brought by the combination of false positive reduction and reverse complement screening: threat mis-identifications drop to a cumulative 0.06% across all taxa (weighted mean: 0.005%), while false positives drop to a cumulative 0.62% (weighted mean: 0.03%).

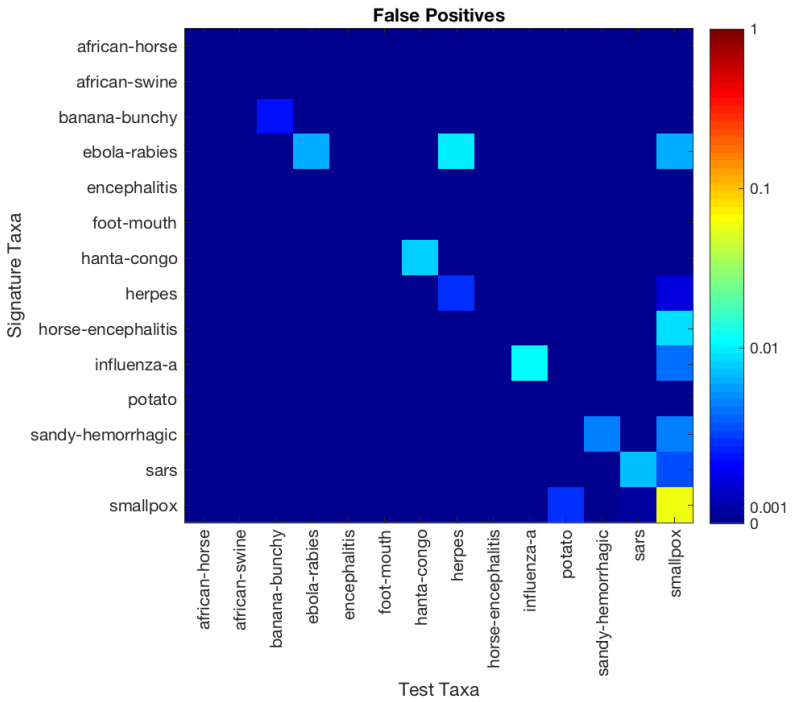
Similarly, we also enhanced FAST-NA for protein-based screening against amino acid sequences. Since most pathogenic mechanisms of action involve proteins, it is typically not the genetic sequence per se but the protein sequence that it can be translated into that is actually of concern. Since 64 three-nucleic-acid codons map onto only 20 amino acids, the same protein sequence can be encoded in many different ways. This means that matching in amino acid space rather than nucleic acid space is useful for avoiding mismatches caused by neutral mutations or codon optimization.

Thus, in order to screen against nucleic acid sequences with FAST-NA, signature generation is done against collections of protein sequences rather than nucleic acid sequences. Screening is then preceded by a translation stage in which a sequence is translated into amino acids following all three potential codon frames, and an alert is raised if a signature matches any of the three potential translations. Figure 25 shows the results for protein-based screening of nucleic acid sequences. Overall, the rate of false negatives is low but well above zero (though expected to be covered by nucleic acid screening), while the rate of mis-identifications and false positives remains extremely low (respectively 0.10% and 0.45% cumulative across all taxa).

Results can be further improved by combining the results of nucleic acid screening and protein-based screening. Preliminary results indicate that unified screening is indeed likely to be effective at removing false negatives: of the biologically meaningful false negatives reported in Section 4.3.1, every single one is caught by either reverse-complement nucleic acid screening (KF022090.1, KF022091.1, HQ719213.1, KJ624719.1, EU303181.1), protein-based screening (AF004437.1, AF004436.1) or both (AF196534.1). Furthermore, the baseline rates of false positives are quite low for both nucleic acid



(a) Threat Identification



(b) False Positives

Figure 25. Protein screening rate of (a) threat identification, i.e., testing signatures against threat sequences across all viral clusters, and (b) false positives, i.e., testing signatures against contrasting sequences across all viral clusters. The X-axis shows the cluster of test sequences, while the Y-axis shows the cluster of origin for signatures.

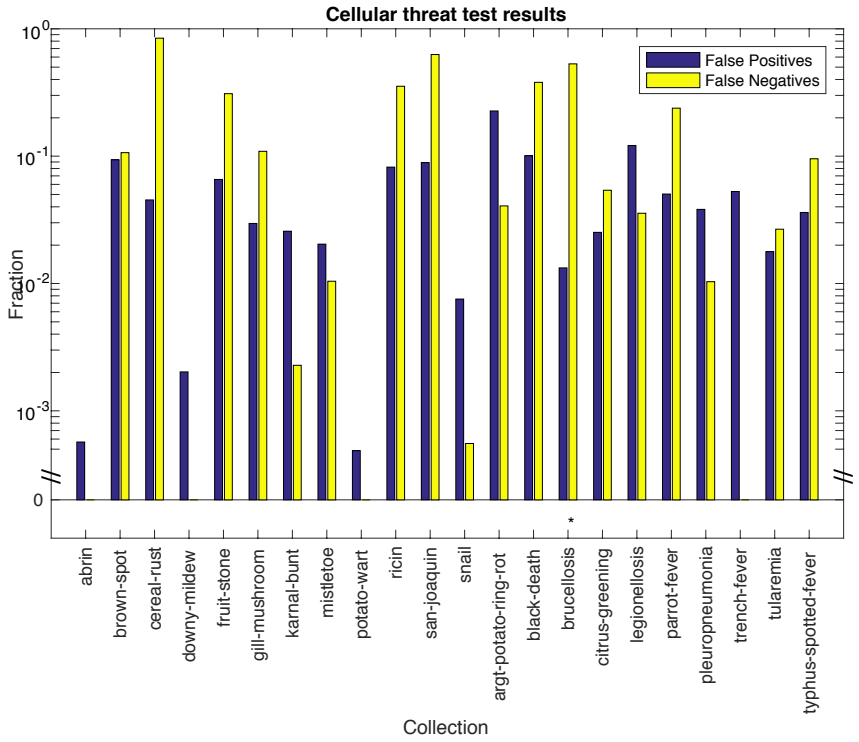


Figure 26. Results of subset testing for a single small taxon of threat and contrast samples for a variety of eukaryotic ("abrin" through "snail") and bacterial ("argt-potato-ring-rot" through "typhus-spotted-fever") threat clusters. Results are highly variable, but indicate likely applicability of FAST-NA to these cellular threats as well.

and protein-based screening and there is significant overlap between the two in the set of false positives that raise alerts, so the increase in false positives from unified screening remains manageable.

4.4 Generalization to Bacterial and Eukaryotic Threats

In addition to the in-depth application analysis of FAST-NA with respect to viral threats reported above, we also conducted preliminary analysis of the applicability of these techniques to threats from bacterial and eukaryotic taxa. Figure 26 shows the results of a preliminary small-scale test in 12 eukaryotic threat clusters and 10 bacterial threat clusters. For each threat cluster, a single taxon of threats were selected (the closest available to 100 sequences in size) and a single taxon of contrast (the closest available to 1000 sequences in size). While there is a high variability in the results obtained, our analysis shows these initial results to be highly promising in the likelihood that FAST-NA can be effectively applied to cellular threats as well as viral threats.

False positives ranged from less than 0.1% to 23%, with a mean of 5.2%. These numbers are similar to initial tests with viral taxa. Accordingly, with use of full scale contrasting data and tuning, it appears likely that false positives for cellular threats should be able to be similarly reduced to very low ultimate rates.

False negatives have an even higher variability. Four threat clusters have zero false negatives, while three other clusters have more than 50% false negatives. Examination of the reasons for these misses, however, reveal that they appear to primarily be due to the number of samples being less than the number of independently reported genes for these organisms. Unlike viruses, which uniformly have a very small number of coding sequences, cellular threats typically have anywhere from hundreds to tens of thousands. These are often reported individually, rather than in entire genomes, and thus the small threat sequence sample size used in this preliminary experiment was too small for a reliable evaluation of false negatives: for many organisms, there was simply not enough chance of a gene showing up in both the training and test sets. When genes did show up in both, however, they appeared to be caught effectively, as we would expect based on the results of applying FAST-NA to viral threats. Our preliminary assessment is thus that FAST-NA is likely to generalize beyond viral threats such that it can also apply effectively to bacterial and eukaryotic threats.

4.5 Realism of CONOPS and Resource Requirements

To further evaluate the realism of the proposed application of FAST-NA to detection of viral sequences, we evaluated the performance of FAST-NA against customer-related data supplied by IDT, and also evaluated the scaling of resource requirements in operation of the current key components of the FAST-NA pipeline.

4.5.1 Evaluation Against Realistic Sequence Distribution

Nucleic acid synthesis orders are likely to be distributed differently across taxa than the public sequence databases on which we trained FAST-NA, so in order to evaluate the performance of FAST-NA in a more realistic environment, we applied the signatures generated from the full-scale viral threat detection test to the collection of customer-related sequence data supplied by IDT.

This data is segmented into five collections: one in which the IDT system detected nothing suspicious, and four “threat” collections where the threat

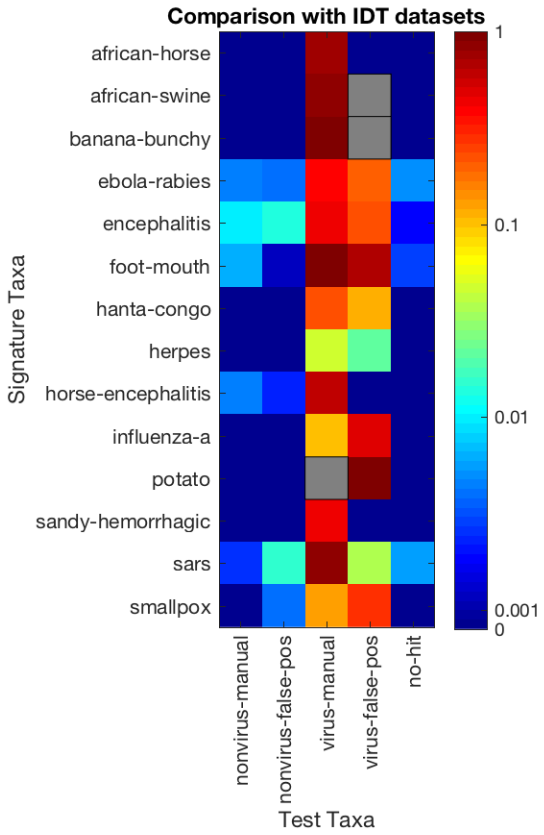


Figure 27. Results of applying full-scale viral threat signatures to IDT customer-related sequences. Color indicates rate of detection, grey square indicates IDT data set had no matches to be tested.

was either viral or non-viral and where it either required human judgement to resolve or did not (labelled "manual" and "false-pos" respectively). Of these five collections, we would thus desire FAST-NA to have a 100% match rate on the virus-manual collection, and a 0% match rate on all of the others.

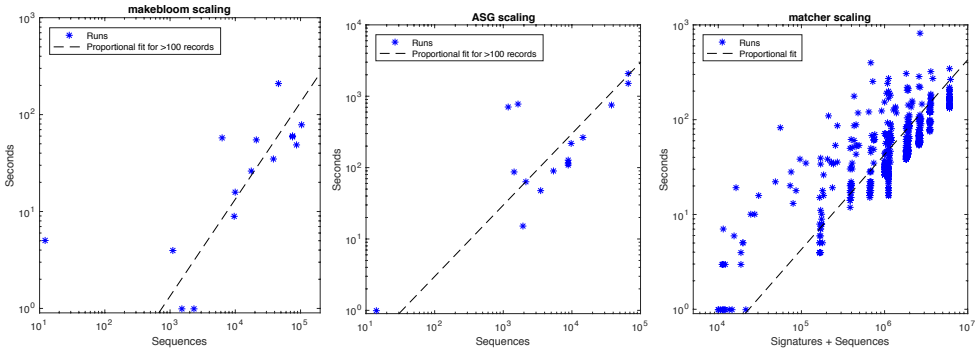
Figure 27 shows the results of comparison with the current FAST-NA system. The no-hit collection has an overall false positive rate of 1.7%, while the non-viral manual and false-positive collections have misidentification rates of 2.7% and 4.2%. The rate of hits in the IDT false-positive collection was higher, at 20.7% of the number of IDT detections, and the rate of successful threat detection was 75.9% of IDT threat calls.

Critically, hand inspection of a selection of missed detections found every such instance to either have reverse-complement matches or else to have a poor nucleotide match but long amino-acid sequence matches. We thus conclude that, as with the false negatives found during tuning of the full-scale viral signature set, false negatives on IDT customer-related data should be expected to resolve with the inclusion of reverse-complement and amino-acid signatures into FAST-NA.

We thus conclude that the FAST-NA is likely to be effective for screening of synthesis orders, once upgraded to avoid the current false negatives. While the false positive rates on synthesis order data are higher than the ideal rates, these may be able to be lowered through additional tuning as discussed above. Furthermore, FAST-NA can also markedly improve current CONOPS by being coupled with current screening techniques to serve as a lightweight first-pass filter, such that more costly BLAST-based inspections are performed only on the small fraction of sequences where FAST-NA matches a signature.

4.5.2 Scaling of Resource Requirements

While viral threats comprise approximately half of the current threat taxa, bacterial and eukaryotic threats have both larger sequences and more sequences. Application of FAST-NA to these threats will thus require scaling up by approximately two orders of magnitude. With regards to this challenge of scaling, three resources are likely to be limiting factors in the deployment and scaling of FAST-NA: working memory (RAM), disk space, and execution time.



(a) makebloom (b) ASG (c) matcher

Figure 28. Time scaling for FAST-NA components.

Space: RAM and disk Space requirements—both RAM and disk—are currently dominated by the Bloom filters used for signature generation. With current settings, these are approximately 2 GB per filter, but no more than two ever need to be loaded into memory at the same time. They are also highly compressible: we have observed reductions of approximately 40-fold under standard compression, meaning that even storing many Bloom filters on disk is likely to be quite manageable.

When scaling up operations, signatures may come to surpass Bloom filters for space requirements, entering a range of 10^9 or 10^{10} signatures in total. Currently, the largest signature collection, ebola-rabies with 6.2 million signatures, requires 916 MB of storage space, i.e., approximately 150 bytes per signature. The space requirements are expected to scale proportional to the number of signatures, so as with Bloom filters, operations at scale will require care in space management, but these scales are still well within reach of current systems, particularly since signatures can readily be split into multiple smaller collections that can each be applied independently.

We do note, however, that the current pipeline implementation is not thoroughly optimized for scaling, and thus some aspects are likely to be problematic until scale-aware choices are made for every aspect of the pipeline. We have already observed this in one instance, in which the matcher ran out of memory while executing the full-scale ebola-rabies signature system against its own sequences on a machine with far more memory available than should have been needed, likely due to inefficiency in the JSON-parsing library currently being used. There are no anticipated barriers to addressing such issues, however, which are routine for contemporary big data systems.

Execution Time In order to evaluate time requirements, we recorded the times required to execute each step of the experimental pipeline for the cross-taxa and IDT customer-related data experiments reported above, executing on a machine with 240 GB of memory. The components of FAST-NA that dominate time requirements are, as expected, generation of contrasting data Bloom filters with the `makebloom` component, generation of signatures with the `ASG` component, and matching signatures with the `matcher` component.

All three of these are expected to have a linear time scaling. Both `makebloom` and `ASG` are expected to execute linearly in the number of sequences, as each component applies each signature independently against a fixed-size data structure. The `matcher` is expected to be linear in the base-pairs of information to be matched against and also in the number of signatures to be loaded into memory to prepare the matcher. Figure 28 shows the observed execution times for each component against the expected linear scaling parameter(s). As can be seen, in all cases the observed times fit well with the expected scaling and moreover the proportionality constants are within a reasonable range to be scaled up two orders of magnitude, particularly with the aid of additional parallelization.

4.6 Opportunities for Generalization of FAST-NA

We have demonstrated the efficacy of applying FAST-NA to both the DNA and protein sequences of viral threats, as well as demonstrating the likely potential for generalization of FAST-NA to apply to non-viral threats.

There are also a number of potential opportunities for improvement of performance by including more biological knowledge or more sophisticated processing into the implementation of FAST-NA, including:

- automatic assistance for contrasting sample selection,
- multi-length signatures,
- signature generation from clustered consensus threat sequences, removing high-variability non-diagnostic regions,
- pre-screening signatures for detection utility,
- multi-segment signatures, and
- non-binary detection values.

4.7 Recommendations for Control of Information Related to FAST-NA Technology

As the FAST-NA technology is intended for defense against potential construction of dangerous biological agents, it is worth considering if access to this technology should be restricted. On the one hand, freely disseminating information may make it easier for bad actors to learn about potential threats or to attempt to evade detection by FAST-NA. On the other hand, the technology is more effective the more widely it is deployed, and may be made yet more effective by having more good actors readily able to innovate to improve the technology and its opportunities for deployment.

As a starting point, we consider an analogy to the computer security world in which the FAST technology was originally developed. There, signature matchers like SNORT and signature construction software like FAST are free and open source, as are older sets of signatures, while information restrictions are applied only to early access to emerging threats. In short, the value is not in the methodology, but in the careful collection and curation of particular signatures. We find this to be in general the case of FAST-NA as well: the software and methodology are relatively simple to derive from existing open source artifacts, while the primary costs are instead found in curation and training. It is also important to note that the signatures of FAST-NA are sensitive to the particulars of curation and training, and even a single relevant signature unknown to a bad actor can spoil an evasion attempt.

It is also important to consider the types of actors that FAST-NA is intended to defend against. High-capability actors with bad intentions, such as nation-states and large private organizations, will in general not be affected by FAST-NA since such actors will generally be able to readily create their own *de novo* nucleic acid synthesis capabilities. The target is instead to defend against carelessness and accidents ("bio-error") by well-intentioned actors and against low-capability bad actors, such as individuals or terrorist organizations, who will generally need to depend on external organizations as their source of nucleic acids. It is thus advantageous to have FAST-NA readily disseminated internationally to all of the companies and similar organizations that can serve as sources of nucleic acids.

Following these principles, we thus recommend the following with regards to dissemination of the elements of FAST-NA. The following information should be disseminated freely and openly:

- The FAST-NA algorithm and methodology, including refinements such as the decision to exclude signatures with poly-A/T, N, and certain individual accession IDs.
- Performance data and results
- The software implementation of FAST-NA and associated workflow tooling
- Tuning information for viral taxa.
- Illustrative examples of threat signatures.

The following information should be distributed in aggregate form, but full details withheld except from qualified organizations such as members of the IGSC and government agencies:

- Tuning information for cellular taxa.
- List of threat and contrasting sequence accession IDs
- List of omitted accession IDs

The following information should be restricted to qualified organizations such as members of the IGSC:

- Large-scale collections of threat signatures.

Note that the last may be adjusted if a way can be developed to distribute signature information without danger of simple reverse engineering.

5 Summary and Discussion

5.1 Progress Against Waypoints

Our progress against key waypoints for this project is as follows:

- Curation of training and testing data: **complete**
- Construction of prototype FAST-NA software: **complete**
- Setup of experimental pipeline: **complete**
- Evaluation of FAST-NA potential for screening viral threat taxa: **complete**
- Evaluation of FAST-NA potential to generalization to other taxa: **complete**

5.2 Important Findings and Conclusions

Our findings in this report are as follows:

- FAST malware screening technology can be effectively adapted for screening of viral nucleic acid sequences.
- FAST-NA can use publicly curated data to identify short sequences diagnostic of viral threat potential in a nucleic acid sequence.
- FAST-NA can significantly reduce false positives in screening for viral threats, without introducing false negatives.
- FAST-NA can likely be extended to apply with similar efficacy to bacterial and eukaryotic threats.
- FAST-NA can support an effective CONOPS for biosecurity screening with a reasonable resource budget.

5.3 Special Comments

None.

5.4 Implications for Further Research

Our results indicate that FAST-NA can enable a significant improvement over the current state of the art in nucleic acid synthesis screening for viral threats, and likely to bacterial and eukaryotic threats as well. **In the interest of national security, we thus recommend funding further development of FAST-NA in support of transition into widespread industrial usage.**

Specifically, we recommend investment in transition of FAST-NA for deployment to screen for viral threats in commercial and government environments, likely in some sort of cooperation with organizations such as the International Gene Synthesis Consortium (IGSC). We further recommend development of FAST-NA for detection of bacterial and eukaryotic threats. Such efforts will involve addressing biological differences between taxa, such as the decreased density of pathogen-specific sequences in cellular genomes, as well as differences in the curation of these taxa. There will also need to be adjustments in the operation of FAST-NA to handle the increased scale of data involved.

Finally, we note that there are likely to be other applications of interest for FAST-NA as well, particularly from a biosecurity perspective, and recommend further investigation of such possibilities.

5.5 Commercial/Proprietary/Third-Party Material in Deliverables

None.

References

[1] Raytheon BBN Technologies. Framework for Auto-Generated Signature Technology (FAST). Final Report on award HSHQDC-14-C-B0031, September 2015.

[2] Daniel Wyschogrod and Jeffrey Dezso. False alarm reduction in automatic signature generation for zero-day attacks. In *2nd Cyberspace Research Workshop*, page 73, 2009.

[3] Nicholas M Luscombe, Jiang Qian, Zhaolei Zhang, Ted Johnson, and Mark Gerstein. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome biology*, 3(8):1, 2002.

[4] RN Mantegna, SV Buldyrev, AL Goldberger, S Havlin, C-K Peng, M Simons, and HE Stanley. Systematic analysis of coding and noncoding dna sequences using methods of statistical linguistics. *Physical Review E*, 52(3):2939, 1995.

[5] Jiang Qian, Nicholas M Luscombe, and Mark Gerstein. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of molecular biology*, 313(4):673–681, 2001.

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